

A Partition Function Estimator

Ying-Chih Chiang,^{1,2} Frank Otto,³ and Jonathan W. Essex⁴

¹*Kobilka Institute of Innovative Drug Discovery, School of Medicine,
The Chinese University of Hong Kong, Shenzhen, 2001 Longxiang Boulevard, 518172,
Shenzhen, China*

²*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen,
2001 Longxiang Boulevard, 518172, Shenzhen, China*

³*Advanced Research Computing Centre, University College London, WC1H 9BT,
UK*

⁴*School of Chemistry and Chemical Engineering, University of Southampton, SO17 1BJ,
UK*

(*Electronic mail: f.otto@ucl.ac.uk)

(*Electronic mail: chiangyc@cuhk.edu.cn)

(Dated: 16 May 2025)

We propose an estimator that allows us to calculate the value of a simple system's partition function using finite sampling. The core idea is to neglect the contribution from high energy microstates, which are difficult to be sampled properly, and then calculate a volume correction term to compensate for this. As proof of concept, the estimator is applied to calculate the partition function for several model systems, ranging from a simple harmonic oscillator to a Lennard-Jones fluid with hundreds of particles. Our results agree well with the numerically exact solutions or reference data, demonstrating that efficiently estimating partition functions for the studied example cases is possible and computationally affordable.

I. INTRODUCTION

The partition function of a system is of fundamental interest in statistical physics because many quantities can be derived from it. For instance, the free energy F can be calculated from the partition function Z via $F = -k_B T \ln Z$, where $k_B T$ is the Boltzmann constant multiplied with the temperature. The energy as well as the entropy can also be derived, once the partition function and its temperature dependence are known. Therefore, it has long been an important question how to calculate a system's partition function. In the past two decades, this question has been addressed with various methods, such as the Wang-Landau algorithm¹ and nested sampling²⁻⁴. Both methods compute the (cumulative) density of states of a system which can then be multiplied with the Boltzmann factor and directly integrated to yield the partition function as well as other quantities of interest. In particular, the nested sampling algorithm has demonstrated its power in studying thermodynamic properties in material science⁵. The tested systems include Lennard-Jones clusters⁶, metallic systems⁷⁻⁹, alloys¹⁰, and small water clusters¹¹. In these examples, the partition function is rarely the sole goal of the study, but rather a basic property that one can calculate alongside the heat capacity or a phase diagram.

Another practically useful quantity is the ratio between two partition functions, as it indicates the free energy difference between two states. For instance, comparing the partition function of a system with a protein and a free ligand to the partition function of the protein-ligand complex yields the binding free energy between the protein and the ligand. Such information can guide the design of drug molecules during lead optimization. However, assessing the ratio between two partition functions via their direct evaluation is computationally too expensive. Instead, it is common practice to calculate this ratio using theories tailored for this purpose. For example, importance sampling can turn the ratio between two partition functions into an ensemble average, which can then be computed via a single sampling on one of the states. This method is known as Zwanzig's equation¹² or free energy perturbation (FEP)¹³. Together with post-processing methods such as Bennett's acceptance ratio (BAR)¹⁴ or multi-state Bennett acceptance ratio (MBAR)¹⁵ to combine data from sampling over several intermediate states, multi-step free energy perturbation (mFEP)¹⁶ has become a standard approach for binding free energy calculations. Other commonly used free energy methods include Kirkwood's thermodynamic integration¹⁷, the Jarzynski equality¹⁸, etc. A comprehensive review can be found in the literature¹⁹. These free energy methods enable e.g. the calculation of the binding free energy between a protein and a ligand²⁰⁻²³, between two pro-

teins²⁴, or for evaluating the permeability of ligands through a lipid bilayer²⁵. Today, free energy calculations have become an indispensable tool in computer-aided drug design^{26–29}.

The aforementioned achievements highlight not only advances in theory but also the power of ensemble averages obtained via finite sampling. Thus, one may wonder whether the partition function itself can also be obtained from an ensemble average. To answer this question, we propose a Partition Function Estimator (PFE) that can estimate the value of a single partition function using a finite sampling. Our idea is rather straightforward: As the high energy microstates are difficult to sample, can we ignore their contribution to the partition function, and then calculate a term to correct the impact of neglecting those microstates? As a proof of concept, we apply this theory to several model examples, including the one-dimensional harmonic oscillator potential, the one-dimensional double-well potential, the two-dimensional Müller-Brown potential, and Lennard-Jones particles in a three-dimensional box. These examples are chosen because their references can be easily obtained: For model potentials, exact numerical solutions are available via brute-force integration, while the partition functions of Lennard-Jones particles can be obtained using nested sampling or standard methods such as mFEP-MBAR.

II. THEORY AND METHODS

A. The Partition Function Estimator

Consider a system composed of N identical particles of mass m . The Hamiltonian reads,

$$H(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p}^2}{2m} + U(\mathbf{q}) \tag{1}$$

where \mathbf{p} and \mathbf{q} denote the momenta and the spatial coordinates of the particles, respectively. $U(\mathbf{q})$ denotes the potential of the system. The canonical partition function Z of the system is defined as,

$$Z = \frac{1}{N! h^{3N}} \int e^{-\beta H(\mathbf{p}, \mathbf{q})} d\mathbf{p} d\mathbf{q}, \tag{2}$$

where h is Planck’s constant, dividing the integral by h^{3N} renders Z dimensionless. β denotes the inverse temperature T multiplied with the Boltzmann constant k_B ($\beta = 1/k_B T$). Z is a function of β , but this dependency is dropped for ease of notation. The integration over phase space in Eq. 2 is separable, and the integration over the momenta can be performed analytically. The partition function of the system hence can be rewritten as⁴,

$$Z = \frac{1}{N!} \left(\frac{2\pi m}{\beta h^2} \right)^{3N/2} \int e^{-\beta U(\mathbf{q})} d\mathbf{q}. \tag{3}$$

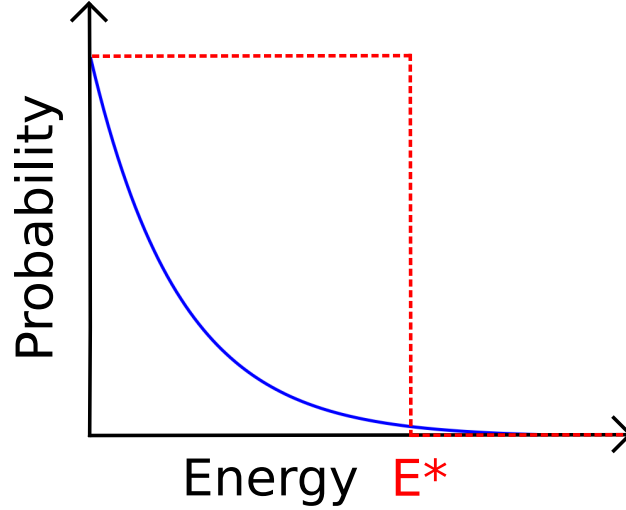


FIG. 1. A sketch of the energy distribution for the canonical ensemble of a finite system. Owing to the exponential nature of the Boltzmann factor, the probability for sampling microstates with high energy will be very low. Above a certain energy E^* the sampling is insufficient and cannot reproduce the exact distribution correctly.

With the integration over momentum space carried out, we can focus on the integration over coordinate space, and thus define the configurational integral

$$Q = \int e^{-\beta U(\mathbf{q})} d\mathbf{q}, \quad (4)$$

to denote the configurational part of the partition function Z , which is sometimes also referred to as the configurational partition function³⁰. It has a dimension of configuration space volume. In the same vein, the expectation value for any function f that depends solely on the coordinates \mathbf{q} , can be expressed as

$$\langle f \rangle = \int f(\mathbf{q}) \frac{e^{-\beta U(\mathbf{q})}}{Q} d\mathbf{q}.$$

Our discussion below will focus on how Q can be estimated from a finite sampling of the system's canonical ensemble.

When sampling a system at finite temperature, the probability for finding a microstate i with energy E_i is proportional to its Boltzmann factor $\exp(-\beta E_i)$, and consequently the probability for finding *any* microstate with energy E is proportional to $g(E) \exp(-\beta E)$ where $g(E)$ denotes the density of states. Due to the exponential nature of the Boltzmann factor, the energy distribution of the system eventually decreases with increasing energy, such that the sampling becomes insufficient or even completely fails for microstates with energies above a certain energy level E^* . See

Fig. 1 for an illustration. In practice, insufficient sampling of the high energy microstates does not affect the sample average of physical quantities much, as long as they do not grow exponentially with E .

Yet, in the extreme case one could be interested in evaluating the sample average of the inverse Boltzmann factor $b(E) = \exp(+\beta E)$. Its theoretical expectation value is given by

$$\langle b(E) \rangle = \int e^{+\beta U(\mathbf{q})} \frac{e^{-\beta U(\mathbf{q})}}{Q} d\mathbf{q} = \frac{L^{3N}}{Q}, \quad (5)$$

where L^3 is the volume of a cubic box that the particles are confined to. According to Eq. 5, if one could estimate $\langle b(E) \rangle$ via sampling, then Q is readily available. However, in practice this idea never works, because a finite sampling fails to provide a correct estimation of $\langle b(E) \rangle$: In a sample of size n with energies E_i , the sample average of b is given by

$$\bar{b} = \frac{1}{n} \sum_{i=1}^n e^{+\beta E_i}, \quad (6)$$

which is extremely sensitive to the high-energy samples. Namely, as the energy distribution follows the Boltzmann weight, i.e. $p(E) \propto \exp(-\beta E)$, the insufficient sampling of the high energy tail leads to a very large fluctuation of \bar{b} , making its value deviate from the exact $\langle b \rangle$. Obtaining Q from this poor estimate of $\langle b \rangle$ then also fails. Here we would like to stress the distinction between the expectation value $\langle b \rangle$ and the sample average \bar{b} . The former is an exact (theoretical) value, while the latter is empirically obtained from a finite sample and will fluctuate if the sampling is repeated. It might even fail to converge in case of insufficient sampling.

Since it is impractical to obtain enough samples to reproduce the actual probability distribution in the high energy tail, why don't we try excluding their contribution in the expectation value? With this idea in mind, let us now consider another function $f(E; E^*)$ that reads,

$$f(E; E^*) = e^{+\beta E} \theta(E^* - E), \quad (7)$$

where $\theta(E^* - E)$ denotes the Heaviside step function, which is 1 for $E \leq E^*$ and 0 for $E > E^*$. In short, $f(E; E^*)$ is the inverse Boltzmann factor, but truncated to zero for energies E larger than some chosen parameter E^* . With E set to the potential energy, i.e. $E = U(\mathbf{q})$, the expectation value of f is then given by,

$$\langle f(E; E^*) \rangle = \int e^{+\beta E} \theta(E^* - E) \frac{e^{-\beta U(\mathbf{q})}}{Q} d\mathbf{q} = \frac{V(E^*)}{Q}, \quad (8)$$

with $V(E^*)$ defined as,

$$V(E^*) = \int \theta(E^* - U(\mathbf{q})) d\mathbf{q}, \quad (9)$$

which is the volume of the coordinate space with the potential energy smaller than E^* . Consequently, Q can be obtained via,

$$Q = \frac{V(E^*)}{\langle f(E; E^*) \rangle}. \quad (10)$$

Eq. 10 is the proposed estimator. This equation itself is exact and holds for any value of E^* . However, to make use of this equation, we have to find a “good” value of E^* , such that the expectation value of $\langle f(E; E^*) \rangle$ can be well estimated by a sample average. We also need to determine values for $\langle f(E; E^*) \rangle$ and $V(E^*)$.

B. Finding E^*

Let us first concentrate on the expectation value $\langle f(E; E^*) \rangle$. We wish to use the sample average of f , i.e.

$$\bar{f} = \frac{1}{n} \sum_{i=1}^n e^{+\beta E_i} \theta(E^* - E_i), \quad (11)$$

as an estimate for $\langle f(E; E^*) \rangle$. In contrast to Eq. 6, this sample average is not sensitive to the high-energy tail, provided that E^* is chosen such that energies below E^* are all sufficiently sampled, cf. Fig. 1. Indeed, if E^* were chosen to be vary large, then Eq. 11 would suffer from strong fluctuation due to insufficient sampling. On the other hand, if E^* were chosen to be very small, then the number of samples that contribute effectively to Eq. 11 would be very small, which again increases its error. This suggests that there is an optimal choice for E^* that can minimize the relative error in \bar{f} .

According to Eq. 10, $\ln Q$ is given by the difference between $\ln V(E^*)$ and $\ln \langle f(E; E^*) \rangle$. We are now using $\ln \bar{f}$ as an estimate for the second term. Since this is the term that comes from the sample average (we will discuss $\ln V(E^*)$ in the next section), it makes sense to choose an E^* that can minimize the error of $\ln \bar{f}$. This is given by,

$$\sigma_M = \sqrt{\frac{1}{n} \frac{\langle f(E; E^*)^2 \rangle - \langle f(E; E^*) \rangle^2}{\langle f(E; E^*) \rangle^2}}, \quad (12)$$

which is just the relative standard error of $f(E; E^*)$, with n being the number of samples. To minimize this error, we find E^* such that

$$\frac{\partial \sigma_M^2}{\partial E^*} = \frac{\partial}{\partial E^*} \left[\frac{1}{n} \left(\frac{\langle f(E; E^*)^2 \rangle}{\langle f(E; E^*) \rangle^2} - 1 \right) \right] = 0. \quad (13)$$

Expressing the expectation values in the above equation as integrals over energy space by utilizing the density of states $g(E)$, we have

$$\langle f(E; E^*) \rangle = \int e^{+\beta E} \theta(E^* - E) g(E) \frac{e^{-\beta E}}{Q} dE = \frac{1}{Q} \int \theta(E^* - E) g(E) dE, \quad (14)$$

and

$$\langle f(E; E^*)^2 \rangle = \int e^{+2\beta E} \theta(E^* - E)^2 g(E) \frac{e^{-\beta E}}{Q} dE = \frac{1}{Q} \int e^{+\beta E} \theta(E^* - E) g(E) dE. \quad (15)$$

Note that in the above equation, we have used the property of the Heaviside function that $\theta^2 = \theta$. Further, the derivative of the Heaviside function gives the Dirac delta function,

$$\frac{\partial}{\partial E^*} \theta(E^* - E) = \delta(E^* - E), \quad (16)$$

which is only non-zero when $E = E^*$. Using Eqs. 14-16, Eq. 13 results in the condition

$$\begin{aligned} 0 &= \frac{1}{\langle f(E; E^*) \rangle^2} \frac{\partial}{\partial E^*} \langle f(E; E^*)^2 \rangle - 2 \frac{\langle f(E; E^*)^2 \rangle}{\langle f(E; E^*) \rangle^3} \frac{\partial}{\partial E^*} \langle f(E; E^*) \rangle \\ &= \frac{1}{\langle f(E; E^*) \rangle^2} \left[\frac{g(E^*)}{Q} e^{+\beta E^*} - 2 \frac{\langle f(E; E^*)^2 \rangle g(E^*)}{\langle f(E; E^*) \rangle Q} \right], \end{aligned}$$

or after simplification,

$$e^{+\beta E^*} = 2 \frac{\langle f(E; E^*)^2 \rangle}{\langle f(E; E^*) \rangle}. \quad (17)$$

According to Eq. 17, E^* depends on the expectation values of $f(E; E^*)$ and $f(E; E^*)^2$. Both can be estimated from their respective sample averages. However, as both also depend on E^* , this equation must be solved iteratively: First, one collects n samples e.g. via Monte Carlo sampling, and saves the trajectory and the associated energies E_i . Then, one calculates \bar{f} via Eq. 11 and similarly $\overline{f^2}$, using an initial guess for E^* , e.g. the maximum energy encountered during sampling. Afterwards, E^* is updated via

$$E^* = \frac{1}{\beta} \left(\ln 2 + \ln \overline{f^2} - \ln \bar{f} \right), \quad (18)$$

and this process is repeated until E^* converges. Notably, each iteration can employ the same trajectory; one merely has to update E^* and recalculate the sample averages, which is computationally cheap. Finally, one can use the optimized E^* to estimate $\langle f(E; E^*) \rangle$ and σ_M from \bar{f} and $\overline{f^2}$.

C. Calculating $V(E^*)$

With E^* determined, we can turn to the calculation of the volume term $V(E^*)$. According to Eq. 9, $V(E^*)$ is the volume of the coordinate space with $E < E^*$. For simple low-dimensional models, such as a harmonic oscillator or Müller-Brown potential, this volume can be calculated directly via “binning”, i.e. dividing the space into small bins, assigning each sample to its corresponding bin, and simply counting the number of occupied bins. However, as the dimensionality increases, no trajectory could cover the entire admissible coordinate space. Calculating $V(E^*)$ by this or other naïve integration approaches is hence not possible. Second, $V(E^*)$ soon becomes very small in comparison with the total volume of the coordinate space. Thus a simple Monte Carlo integration would also be insufficient to address this issue.

Fortunately, the nested sampling algorithm^{2,4} was designed to tackle such a problem. Here, we employ a modified version of nested sampling, making it more suitable for the purpose of Eq. 9. Briefly, the implemented algorithm is as follows:

0. Randomly generate N “walkers” with potential energies smaller than a starting energy ceiling E_0 . Each walker is an independent copy of the system that will be propagated in the course of the algorithm. As the coordinates of the initial walkers are randomly generated (each coordinate uniformly distributed between 0 and L), all possible configurations are equally probable to be chosen. The starting energy ceiling E_0 is chosen to be extremely large (e.g. 10^{12} or larger), such that the initial volume V_0 can be set to the volume of the entire coordinate space, viz. $V_0 = L^{3N}$. Alternatively one can calculate the success rate of the initial walker generation and adjust the initial volume accordingly, e.g. 0.5 success rate implies an initial volume $V_0 = L^{3N}/2$.
1. Select a fraction p ($0 < p < 1$). This determines the new energy ceiling for the current (i -th) iteration, via $(E_i - E_{\min}) = p \cdot (E_{i-1} - E_{\min})$, where E_{\min} is the minimum potential taken from the energy samples, to ensure that the scaling is performed over a positive number. Use a simple Monte Carlo integration to calculate the volume $V(E_i)$, i.e. the volume of coordinate space with energy smaller than E_i . If n_i is the number of walkers with energy below E_i , we have $V(E_i) = (n_i/N) \cdot V(E_{i-1})$. In practice, we choose $p = 0.99$ or $p = 0.999$. This guarantees that E_i is fairly close to E_{i-1} , so that most walkers fall within E_i , and the volume $V(E_i)$ can be easily evaluated.

2. Relax the walkers whose energies are larger than E_i via rejection sampling. That is, only moves that lower the potential energy will be accepted. Afterwards, propagate those relaxed walkers freely within the energy boundary E_i , until they are equilibrated. This means that moves will be accepted as long as the walker's energy E does not exceed E_i upon moving. If the propagation is sufficiently long, this equilibration generates walkers with equal probability over the configuration space with $E \leq E_i$.
3. Repeat Steps 1 and 2 until $E_i \leq E^*$ is reached. As it is unlikely for E_i to exactly hit E^* , one can assign $E_i = E^*$ when detecting $E_i < E^*$. This avoids using an extra interpolation step for determining $V(E^*)$ from neighboring energy levels.

Compared to the original nested sampling, our implementation is more akin to performing multiple Monte Carlo integrations iteratively. Our approach differs from nested sampling in a few aspects. First and most importantly, nested sampling throws away a fixed number of walkers in each iteration, leading to a fixed ratio of volume reduction. For instance, if one throws away the walker with the highest energy in each iteration, then at the i -th iteration, nested sampling yields a volume $V(E_i) = ((N - 1)/N)^i \cdot V_0$, where the energy E_i is determined by the energy of the walker that is currently thrown away. In contrast, our algorithm would yield a volume $V(E_i) = \prod (n_i/N) \cdot V_0$. Thus, any sampling fluctuation would reflect on E_i in nested sampling, whereas it reflects on $V(E_i)$ in our implementation.

Second, nested sampling does not "relax" the walkers as we did in Step 2. Rather, it randomly duplicates a walker with energy smaller than E_i and propagates the cloned walker until it becomes uncorrelated from the original one. This operation might be more efficient than ours, but as our algorithm terminates at E^* rather than proceeding to the lowest potential energy, the number of iterations required is much smaller. Third, when applying nested sampling to integrate over the Boltzmann factor, more walkers are needed in order to better resolve the density of states in energy. Since PFE does not require any such integration, we can use much less walkers or even more aggressive energy truncation (via the choice of fraction p) to calculate $V(E^*)$.

III. RESULTS AND DISCUSSION

A. Harmonic Oscillator

Our first example is the one-dimensional harmonic oscillator potential, $U(x) = kx^2/2$. Here we set the force constant as $k = 300$ and $k_B T = 0.59616$. The potential energy curve and the corresponding distributions are depicted in Fig. 2(a). In this system, we employ a simple Metropolis Monte Carlo algorithm for sampling, utilizing a total of 10^6 steps and a step size of 0.1. The trajectory and sampled potential energy are recorded every 10 steps. Hundreds of independent samplings are conducted and processed using PFE (Eq. 10) to derive $\ln Q$. The corresponding average and standard deviation (fluctuation) are then depicted in Fig. 2(b). Here, different E^* are selected to demonstrate their impact on PFE. The term $a\%$ indicates the criterion for selecting E^* : It is chosen such that for the top $a\%$ of samples in energy, the corresponding value of $f(E; E^*)$ is zero.

The outcomes obtained from the optimal E^* (calculated using Eq. 18) are labeled as “opt” (red curve). As expected, utilizing PFE (Eq. 10) with the optimal E^* (Eq. 18) leads to a converged result after approximately 10^4 steps, while the sampling fluctuation diminishes rapidly with the step count. Conversely, the $\ln Q$ calculated using the maximum energy sampled as E^* (black curve, 0%) exhibits substantial fluctuation and fails to converge. However, upon applying the cutoff through the Heaviside function, PFE results converge towards the exact solution (dashed line), albeit with slightly inferior performance compared to using the optimal E^* . This behavior aligns with the notion that high-energy microstates are often undersampled, leading to convergence challenges.

The σ_M of $\ln\langle f(E; E^*) \rangle$ can be calculated using Eq. 12 and compared with the standard deviation of $\ln Q$ across the 100 sampling repetitions. The former is labeled as standard error in Fig. 2(c), while the latter is labeled as fluctuation. With the exception of the 0% cutoff, the predicted standard error closely matches the fluctuation, and increasing the sample size results in diminished error. This intriguing finding indicates that, in this simple example, the fluctuation primarily arises from inaccuracies in computing $\ln\langle f(E; E^*) \rangle$, as the $\ln V(E^*)$ can be fairly accurately computed from the trajectory histogram using 100 bins.

As PFE demonstrates good convergence, one may wonder whether this good performance comes from the sampling itself being “sufficient.” To investigate whether the sampling is “suf-

cient,” we process the same trajectories to obtain Boltzmann weights for a naive summation. That is, we reconstruct the density of states $g(E)$ through the energy histogram $h(E)$, i.e. $h(E) \propto g(E)\exp(-\beta E)$. This requires a normalization of $g(E)$, the integral of which gives the volume of configuration space that can be accessed. In theory, this refers to the entire configuration space. In practice, owing to the limitation of finite sampling, this volume is computed from the trajectory. The $g(E)$ computed from one of the trajectories and a comparison to its analytical solution can be found in Appendix 1. Then we calculate $\ln Q$ via,

$$Q = \int g(E)e^{-\beta E} dE . \quad (19)$$

The result is depicted in Fig. 2(d). One can see that, while PFE already converges at 10^4 steps, the naive summation still struggles to converge, and it also exhibits a much larger fluctuation. This result is not surprising: Reconstructing the density of states involves calculating $h(E)\exp(+\beta E)$, but the inverse Boltzmann factor is extremely relevant in the high energy region, where the sampled histogram $h(E)$ fails to reproduce the true probability distribution. Additionally, for energies above the highest sampled energy, $g(E)$ gets estimated as zero. These errors then enter $g(E)$ through the normalization, and eventually prevent an accurate estimation of $\ln Q$. In contrast, PFE addresses the issue of insufficient sampling of the high energy regions by terminating their influence via the step function and by introducing the volume term as a correction, which in combination let PFE outperform the naive summation. Interestingly, the naive summation gives a result very close to PFE with 0% data cutoff, because both use the volume that was accessed by the trajectory.

Finally, a note about the meaning of E^* : It is the optimal choice that minimizes the fluctuation of $\ln\langle f(E; E^*) \rangle$. In the present example, it roughly corresponds to cutting off the top 13% of the energy samples. As Fig. 2(d) also shows, $\ln Q$ calculated using a 13% cutoff agrees excellently with the result from the optimal E^* . This suggests that microstates with energy higher than the top 13% energy samples are not well sampled. A discussion on the same system with a 10 times higher temperature is provided in Appendix 2.

B. Double Well Potential

Next we consider a double well potential defined as follows,

$$U(x) = \frac{16h}{x_0^4} x^2 (x - x_0)^2 , \quad (20)$$

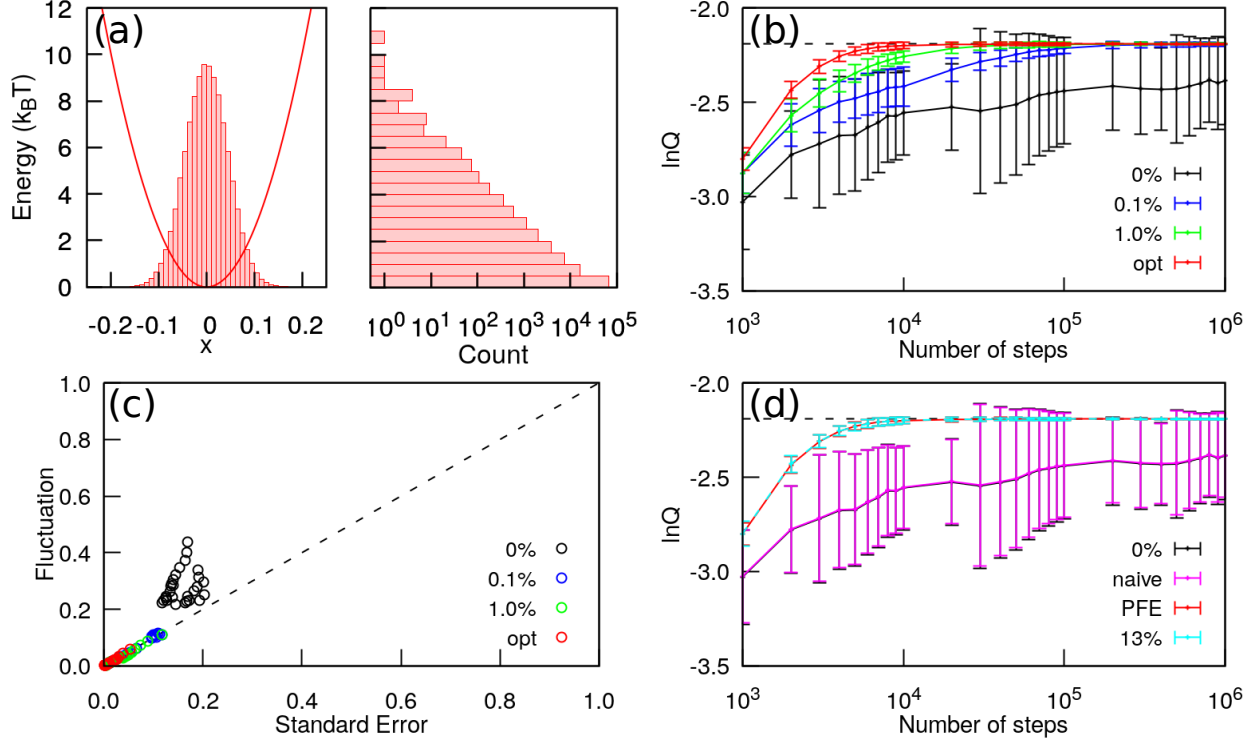


FIG. 2. (a) The potential energy curve of a harmonic oscillator and the corresponding spatial and energy distributions. As anticipated with finite sampling, the spatial probability distribution follows a Gaussian function (left figure), while the histogram diminishes as the energy level rises (right figure), demonstrating that the sampling in the higher energy range is insufficient. (b) $\ln Q$ computed using PFE (Eq. 10). Shown are results obtained using different E^* , including those resulting in the removal of 0%, 0.1%, 1.0% of data, as well as E^* determined using Eq. 18 (labeled as “opt”). The dashed line indicates the exact solution. (c) A comparison between the standard error of $\ln\langle f(E; E^*) \rangle$ and the fluctuation (standard deviation) of $\ln Q$ observed across 100 independent samplings. Each data point represents the fluctuation and the standard error data, obtained at one specific number of steps from panel (b). A linear relationship is observed in nearly all datasets, except for the black ones. (d) A comparison between $\ln Q$ calculated via PFE (opt) and via a naive summation over the Boltzmann weights, using the same trajectories. PFE converges after 10^4 steps, while the naive summation fails to converge even after 10^6 steps.

where h is the barrier height and x_0 is the position of the second minimum. In this example, $k_B T$ is maintained at 0.59616 , while h and x_0 are configured to $10k_B T$ and 3 , respectively, to introduce some sampling difficulties. For this system, a total of 10^6 steps are sampled using both Monte Carlo and replica-exchange (RE)³¹ techniques. The latter is a typical enhanced sampling

method that simulates multiple replicas with various temperatures simultaneously. The coordinates between different replicas can be swapped to enhance the sampling efficiency. In the current calculation, 10 replicas are employed in total. The temperatures are linearly scaled, with the lowest and highest $k_B T$ values set to 0.59616 and 1.9872, respectively.

The potential curve and the corresponding distributions are depicted in Fig. 3(a). Due to the substantial barrier, Monte Carlo sampling encounters difficulties in traversing the barrier, as evident in the blue distribution. In contrast, RE sampling facilitates frequent barrier crossings, resulting in a symmetric spatial distribution even for the replica with the lowest temperature, as observed in the orange distribution. The energy histograms exhibit similarities between both sampling methods, except in the high-energy region. Shown in Fig. 3(b) are $\ln Q$ calculated using PFE. With the exception of the curve labeled “RE”, all results are obtained from an extensive Monte Carlo simulation. Once again, the optimal E^* (opt) provides the closest estimation of $\ln Q$, although the result is not yet converged after 10^6 sampling steps. Owing to the sampling difficulty posed by the high barrier, the $\ln Q$ plot displays a plateau around 10^5 steps ($\ln Q = -0.9$), as depicted by the red curve. This plateau arises because the simulation struggles to surmount the barrier (Boltzmann factor $e^{-10} = 4.54 * 10^{-5}$) in shorter sampling durations. However, with longer sampling durations, simulations eventually overcome the barrier, exploring the second well. Consequently, the fluctuation increases, and $\ln Q$ approaches the exact solution.

In contrast, $\ln Q$ calculated with RE sampling and the optimal E^* shows an early convergence well before 10^5 steps (orange curve). This observation underscores how sampling difficulties can influence the convergence rate of calculations. Because detailed documentation of enhanced sampling methods can be found in the literature^{31,32}, our discussion will focus on the behavior of PFE. Interestingly, when the sampling remains confined to one well (reflected in the plateau-like $\ln Q$ values), the deviation from the exact solution is around 0.7. This magnitude is about the same as $\ln(2)$, implying that the primary source of error stems from halving the size of $V(E^*)$ when the simulation is trapped in one well. Introducing enhanced sampling techniques here primarily aids in the accurate computation of $V(E^*)$, rather than to $\langle f(E; E^*) \rangle$.

In Fig. 3(c), a comparison is made between the standard error of $\ln \langle f(E; E^*) \rangle$ and the fluctuation (standard deviation) of $\ln Q$. Unlike the previous harmonic oscillator example, here the fluctuation surpasses the standard error, suggesting that calculating $\ln V(E^*)$ can introduce a significant error, even when directly assessed from the trajectory. Nonetheless, increasing the number of sampling steps can reduce the fluctuation, as demonstrated by the orange curve in panel (b), where the

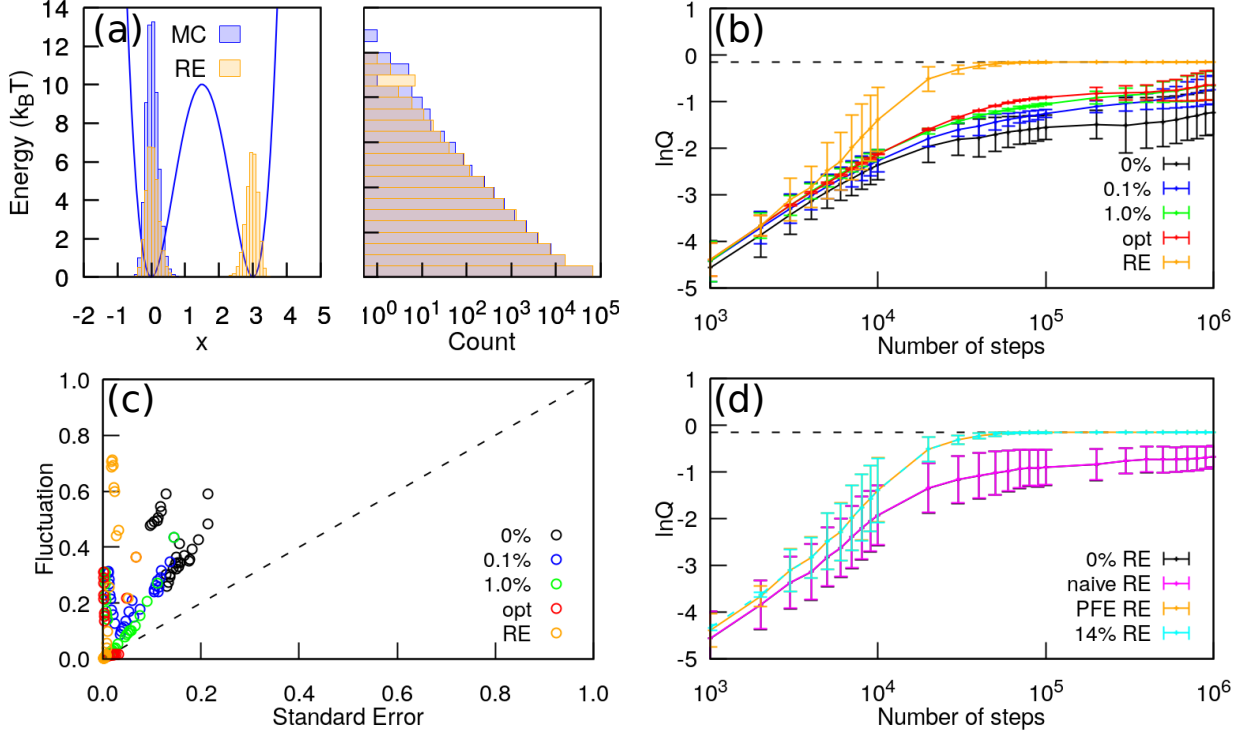


FIG. 3. (a) The double-well potential energy curve and the corresponding spatial distributions (left) and energy histograms (right). Results from a basic Monte Carlo sampling are represented in blue, while those from replica-exchange (RE) sampling are shown in orange. (b) $\ln Q$ for the double-well potential, computed using PFE and Monte Carlo sampling. Reported are the average and standard deviation from 100 independent calculations. Labels such as 0%, 0.1%, 1.0% denote the percentages of data removal, owing to the choice of E^* . The outcome derived from RE sampling and the optimal E^* is also depicted (orange). The numerically exact value (-0.15) is indicated by the dashed line. (c) A comparison is made between the anticipated standard error of $\ln \langle f(E; E^*) \rangle$ and the fluctuation of $\ln Q$ across 100 independent samplings. (d) Comparison between results obtained via PFE and via the naive summation over Boltzmann factors, using the same RE trajectories. Again, PFE converges nicely as the number of steps increases, but the naive summation struggles to converge. The optimal E^* corresponds to an energy that truncates the top 14% energy samples.

fluctuation eventually dwindles to insignificance. Comparing PFE with the naive summation over Boltzmann factors again shows how challenging it is to reconstruct the density of states correctly (see Appendix 1 for its plot), even when the RE sampling is employed. See Fig. 3(d) for $\ln Q$ calculated using the two methods. Interestingly, the optimal E^* here corresponds to cutting off the

top 14% energy samples' contribution. This value is very close to the previous harmonic oscillator example. Both these one-dimensional cases indicate that the sampling of the high energy regime is worse than expected.

C. Müller-Brown Potential

To investigate the impact of temperature on $\ln Q$ calculations, we employ the Müller-Brown potential, which is defined as follows:

$$U(x,y) = \sum_{k=1}^4 A_k \exp [a_k(x - x_k^0)^2 + b_k(x - x_k^0)(y - y_k^0) + c_k(y - y_k^0)^2] + U_0, \quad (21)$$

with $A = (-200, -100, -170, 15)$, $a = (-1, -1, -6.5, 0.7)$, $b = (0, 0, 11, 0.6)$, $c = (-10, -10, -6.5, 0.7)$.

In the original Müller-Brown potential, $U_0 = 0$. Here we set $U_0 = 147.70$ to ensure the potential minimum is at zero. The potential energy surface is shown in Fig. 4(a). The potential features three local minima, with the highest energy barrier approximately at 107. Three scenarios are considered: First, the surface is well explored, and all minima are well sampled ($k_B T = 100$). Second, the two local minima with higher energies are barely sampled at the given length of simulation ($k_B T = 10$). Third, the sampling is trapped at the global minimum of the surface ($k_B T = 2$). The trajectory histograms associated with these three scenarios are depicted in Fig. 4(b-d). In these panels, the number of bins employed in each direction is 100, but the bin width is automatically adjusted, resulting in varying count scales across the panels.

Monte Carlo simulations are employed again to compute $\ln Q$. The sampling is conducted for 10^7 steps, and the energy is saved every 10 steps. The volume $V(E^*)$ is determined based on the 2-dimensional trajectory histogram, where the area of the populated bins is summed to give $V(E^*)$. Results are illustrated in Fig. 5. Like before, 100 independent trajectories are processed to derive the mean and standard deviation (fluctuation) of $\ln Q$. Various E^* values are employed to showcase the effect of incorporating the Heaviside function to truncate the inverse Boltzmann factor: PFE with the optimal E^* (opt) consistently yields the best results, while PFE with the highest sampled energy (0%) performs the poorest. That said, PFE with the optimal E^* steadily converges to the exact solution, exhibiting smaller fluctuation compared to other choices of E^* . This underscores the significance of selecting E^* to minimize error (as per Eq. 18). Most importantly, PFE (opt) yields the correct $\ln Q$ in all three scenarios, suggesting that its performance is not affected much by the temperature.

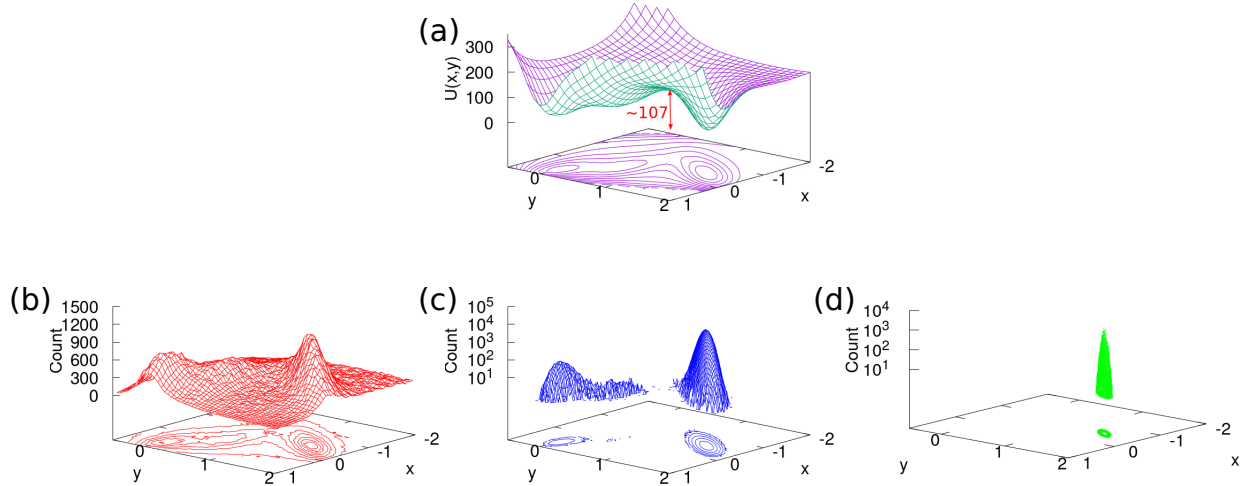


FIG. 4. (a) The surface plot of the Müller-Brown potential, with a vertical shift to align the global minimum at 0. The primary barrier height measures around 107, while the secondary barrier height is approximately 34. Both the potential energy and $k_B T$ are represented in the same arbitrary unit. (b) Histogram of a single sampled trajectory at $k_B T = 100$. The population effectively covers all three minima, indicating adequate sampling. (c) Histogram of a trajectory sampled at $k_B T = 10$. The population at the two local minima is non-negligible, but considerably lower than that at the global minimum. Sampling is marginally sufficient in this scenario. (d) Histogram of a trajectory sampled at $k_B T = 2$. The population is distinctly confined to the deepest well, showcasing entrapment at the global minimum.

Noteworthy observations arise when contrasting the standard error of $\ln\langle f(E; E^*) \rangle$ with the fluctuation of $\ln Q$: At the high temperature limit (panel d), the fluctuation and standard error align off-diagonally. This demonstrates that calculating $\ln V(E^*)$ via binning the trajectory entails inherent errors, which eventually diminish when the number of steps increases, as exemplified by the red curve in panel (a). At the low temperature limit (panel f), the fluctuation and the standard error align diagonally, because the sampling is trapped in one spot so that binning the trajectory can yield highly accurate $\ln V(E^*)$, and calculating $\ln V(E^*)$ contributes no error to the $\ln Q$ calculation. The good performance for these two cases echoes the central idea of PFE: Only microstates with relatively low energy (compared with $k_B T$) matter, because their Boltzmann weights are larger. Other relatively high energy microstates only introduce noise and should be truncated and accounted for via the correction term $V(E^*)$. However, a problem arises at the middle temperature regime (panel e), where the sampling just has few chances to venture beyond barriers and explore other local minima. Depending on whether the other minima are sampled or not during the simulation, the

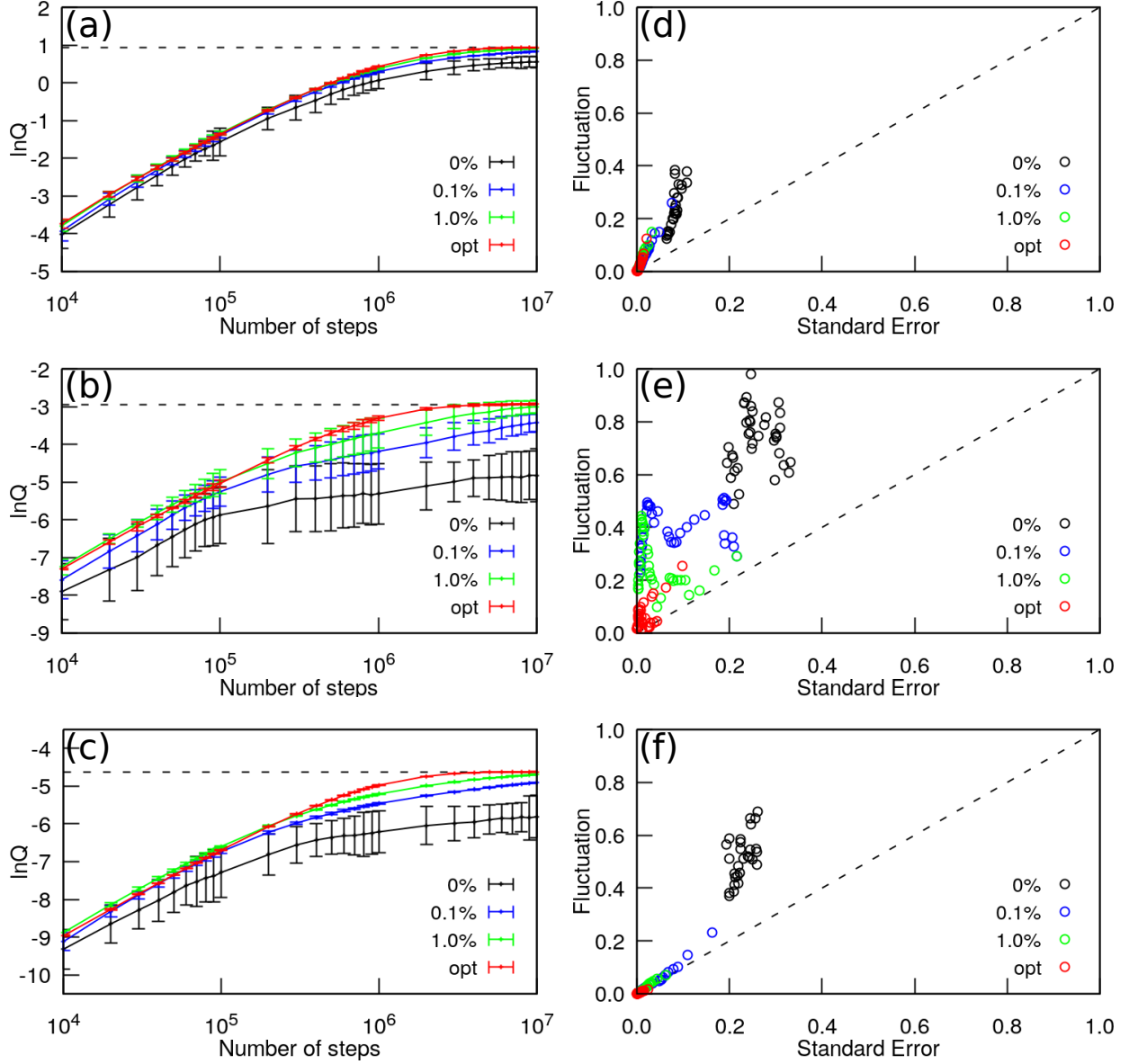


FIG. 5. (a) - (c): $\ln Q$ for Müller-Brown potential, computed using PFE for $k_B T = 100$, $k_B T = 10$, and $k_B T = 2$, respectively. The mean and standard deviation from 100 independent calculations are reported. The dashed line represents the exact solution obtained from a numerical integration. PFE with the optimal E^* exhibits minimal fluctuation, which is further reduced with an increase in the number of steps. (d) - (f): the standard error of $\ln \langle f(E; E^*) \rangle$ compared with the fluctuation of $\ln Q$ for $k_B T = 100$, 10, and 2, respectively. Owing to sampling challenges, the fluctuation is notably larger than the standard error for $k_B T = 10$. See the text for further discussion.

value of $\ln V(E^*)$ obtained from binning the trajectory can fluctuate widely. Although eventually the calculation will converge correctly with increasing number of samples, the fluctuation introduced by $V(E^*)$ suggests that alternative methods to binning trajectories must be used for more complicated systems.

D. Lennard-Jones Particles

The last example to be discussed is a more realistic system: Lennard-Jones particles in a 3-dimensional box with periodic boundary conditions. In this example, the box length is fixed at 25 Å and Lennard-Jones particles are sequentially inserted into the box. The parameters are taken from the OpenMM example (Ar atom)³³: $\epsilon = 0.238$ kcal/mol, $\sigma = 3.4$ Å, with the potential shifted to zero at 3σ for cutoff. The temperature is set to 120 K. The dispersion correction is disabled for simplicity. Once again $\ln Q$ is calculated via PFE with the optimal E^* . Metropolis Monte Carlo sampling with a step size of 1.0 Å is employed for data collection. The simulation consists of a 50000-step equilibration and a 10^6 -step production run. The potential energy is saved every 1000 steps. The volume $V(E^*)$ is determined using the modified nested sampling algorithm as described in Section II C. For the present example, we employed 200 walkers, 2000 equilibration steps, and an energy scaling fraction of 0.99 at each iteration. The relaxation steps and the number of walkers to be relaxed are dynamically determined on the fly. As a reference, the modified nested sampling algorithm employing the same parameters can be used to calculate the density of states and consequently $\ln Q$ via a direct integration.

The standard mFEP-MBAR method^{15,16} is also applied to calculate another reference for $\ln Q$. Briefly, mFEP-MBAR computes the free energy difference between two states, although the $N!$ correction for N indistinguishable particles is not taken into account. That is, it calculates $\ln(Q_{N+1}/Q_N Q_1)$, where N is the number of particles. Since Q_1 is simply the box volume, subsequent values $\ln Q_2, \ln Q_3, \dots, \ln Q_N$ are derived by accumulating the mFEP-MBAR outcomes. The mFEP-MBAR calculation is conducted using molecular dynamics and alchemical functions implemented in OpenMMTools. The interaction between the chosen (alchemical) Ar particle and the other N particles is turned off linearly according to $U_\lambda = U_{N+1} + \lambda(U_N - U_{N+1})$, where $\lambda = 0, 0.05, \dots, 1$. The potential reduces to the $(N+1)$ -particle potential U_{N+1} at $\lambda = 0$, while at $\lambda = 1$ the alchemical particle becomes decoupled from the rest. In total 21 windows are used in each calculation, and for each window 1000 samples are collected after a 10000-step equilibration

run. Between each data collection step, the simulation length is chosen to be increased linearly with the number of particles from 160 steps ($N = 1$) to 1000 steps ($N = 29$). This leads to a maximum number of steps of 21,210,000 and provides a comparable level of precision to PFE for the $\ln Q$ calculation.

Results for $\ln Q$ are depicted in Fig. 6(a). This time, 10 independent calculations are carried out for each method to determine the mean and the corresponding standard deviation (fluctuation). All three methods yield identical $\ln Q$, which increases with the number of particles. Compared to the mean of $\ln Q$, the fluctuation is negligible, suggesting that a single calculation can already provide a reliable value for $\ln Q$. In particular, when comparing the fluctuation of $\ln Q$ encountered in the PFE calculations with the standard error of $\ln\langle f(E; E^*) \rangle$, one finds that the largest fluctuation here is merely 0.67, as indicated by the outlier in panel (c). In contrast, the chemical potential μ , defined as $-k_B T \ln(Z_{N+1}/Z_N)$, reveals the downside of PFE. As shown in Fig. 6(b), the chemical potential calculated via PFE fluctuates greatly around $N = 13$ and $N = 14$. NS data also show some fluctuation, whereas the reference mFEP-MBAR result is smooth and has very small fluctuation. Upon examination, the chemical potential at both $N = 13$ and $N = 14$ requires calculating $\ln Z_{14}$, whose configurational contribution $\ln Q_{14}$ has the largest fluctuation among all PFE calculations, as shown by the red curve in panel (d). This large fluctuation comes from one unfortunate sampling run where the high energy samples are not properly truncated. Removing this run from the statistics immediately stabilizes the mean and fluctuation, as shown by the orange curve in panel (b) and (d). This demonstrates that searching for the optimal E^* may fail occasionally, but one can easily identify this issue and fix the problem by employing a suboptimal E^* , e.g. by cutting off the top 10% of energy samples. Clearly, PFE is not well suited for evaluating the chemical potential in this way as it requires evaluating the difference between two adjacent $\ln Z$. While both $\ln Z_{N+1}$ and $\ln Z_N$ have very small relative errors, this can still be a large number in comparison with their difference.

Considering performance metrics, a comparison is made based on the number of steps required to complete the calculations. Because different sampling techniques are used, viz. Monte Carlo for PFE and molecular dynamics for mFEP-MBAR, we simply count the number of sampling steps required for one single calculation. Depicted in Fig. 6(e) is the average step count over 10 $\ln Q$ calculations. While a direct numerical integration over the Boltzmann factor (NS, blue) requires running the NS algorithm down to very small energies in order to obtain the complete density of states and is therefore expected to be computationally expensive, it seems odd that mFEP-MBAR

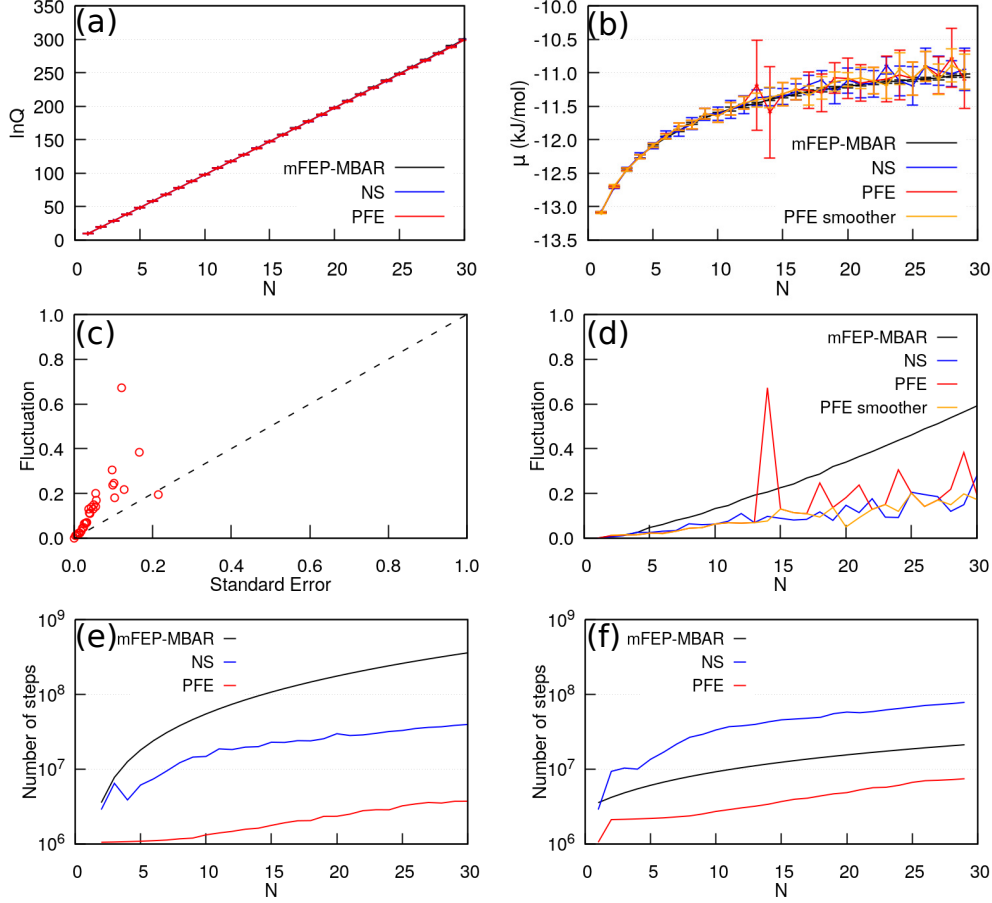


FIG. 6. (a) $\ln Q$ for various numbers of particles (N), calculated using different methods. The standard mFEP-MBAR served as the reference, and NS denotes data obtained from a numerical integration via the modified nested sampling algorithm. Reported are the averages and the standard deviations from 10 independent calculations. It is evident that the standard deviation is negligible in comparison with the mean value, suggesting that a single calculation is sufficient to obtain an accurate $\ln Q$. (b) The chemical potential calculated using the same methods. The reference mFEP-MBAR data are the most stable with very small fluctuation. PFE and NS data in principle follow the reference, but their fluctuations are large. (c) The fluctuation of $\ln Q$ vs. the standard error of $\ln \langle f(E; E^*) \rangle$ observed in the PFE calculations. Each data point corresponds to one N . (d) Fluctuation of $\ln Q$ depicted in panel (a). The fluctuation of mFEP-MBAR is larger but smoother than those from NS and from PFE, while PFE's fluctuation peaks at $N = 14$. See the text for detailed discussion. (e) The average number of steps required to calculate $\ln Q$. As mFEP-MBAR evaluates $\ln(Q_{N+1}/Q_N Q_1)$ and Q_1 is known, $\ln Q_N$ is obtained by accumulating data over N . (f) The average number of steps required to calculate μ . This is the typical operation for which mFEP-MBAR is designed, although surprisingly PFE does not perform worse.

demands even more steps. This discrepancy arises because mFEP-MBAR is designed to efficiently compute the free energy difference between two states. To get $\ln Q$, the free energy differences need to be accumulated over N , resulting in a higher total step count. What is really surprising is that PFE does not require many sampling steps to compute the chemical potential in this instance. In comparison with the mFEP-BAR reference, PFE even requires less steps, as indicated by the red and black curves in Fig. 6(f), although its result has a much larger fluctuation. Nonetheless, this intriguing outcome highlights that calculating $\ln Q$ at a given temperature can really be achieved with a computational effort similar to traditional free energy calculations.

Extending the calculations to a higher number of particles, i.e. into the liquid regime, is possible. Ar fluid has a boiling point of 87.303 K and a density of 1.396g/ml, which corresponds to 328 particles in our cubic box (size 25^3 \AA^3). Hence, we consider systems composed of N and $N + 1$ Ar particles at 87 K, with $N = 300, 310, 320, 330, 340, 350$. We calculate the associated $\ln Q$ and μ to demonstrate PFE’s performance. Since PFE does not require the potential energy samples to be collected strictly from a Monte Carlo simulation, we turn to the molecular dynamics simulations using OpenMM to collect the potential energy samples at equilibrium. Totally 10 independent simulations are conducted for each N . Each simulation is 20 ns long, and the potential energy is saved every 2 ps, yielding 10000 energy samples for each simulation. Similar to our previous examples, the mean and the standard deviation of $\ln Q$ obtained from the 10 samples are depicted in Fig. 7(a). Results obtained from using different numbers of energy samples (i.e. different simulation lengths) are compared. According to the plot, the calculation of $\ln Q$ is again converged with a negligible fluctuation. Here we do not calculate $\ln Q$ using other references. Instead, we assess the quality of PFE’s $\ln Q$ via the chemical potential, a reference for which can be easily computed using mFEP-MBAR. Briefly, the mFEP-MBAR calculations are conducted using OpenMM, and 10 repeats are performed to obtain the mean and standard deviation (fluctuation). For a single run, 31 windows are employed. The simulation in each window is 1 ns long for equilibration and 2 ns long for data collection (1000 samples for each window). As expected, the chemical potential computed via PFE does not behave as well as the mFEP-MBAR reference. Fig. 7(b) shows that PFE’s chemical potential calculation has a much larger fluctuation than mFEP-MBAR, although the average μ still follows the reference closely. This good agreement suggests that $\ln Q$ calculated by PFE is indeed reliable, as we already know that the large fluctuation is the consequence of changing the baseline for the relative error: For $\ln Q_{N+1}$ and $\ln Q_N$, their relative error is negligible because the associated values are huge. For $\ln Q_{N+1} - \ln Q_N$, the error is large because their

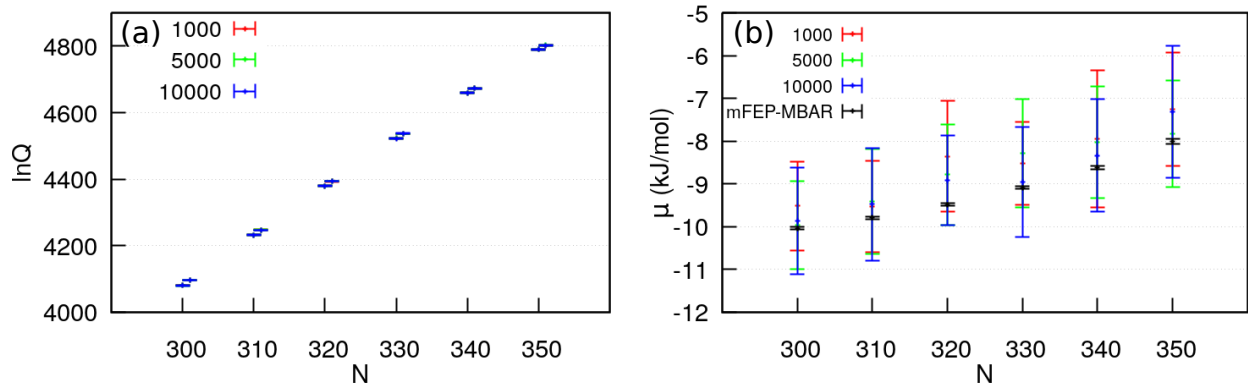


FIG. 7. (a) $\ln Q$ for a Lennard-Jones fluid (Ar, 87 K) calculated using PFE. Depicted are the average and standard deviation of $\ln Q$, obtained from 10 independent calculations. The numbers 1000 (red), 5000 (green), and 10000 (blue) refer to the number of energy samples used to calculate $\ln\langle f(E; E^*) \rangle$. More details are provided in the text. (b) Chemical potential μ computed using $\ln Q$ from panel (a) and using mFEP-MBAR, which is more accurate than PFE, as the fluctuation for PFE is rather large. Nevertheless, PFE’s mean value follows the reference closely, demonstrating that the $\ln Q$ s reported in panel (a) are reliable.

difference is small.

Last but not least, we discuss the difficulties encountered in the calculation of the volume $V(E^*)$, which gets more and more challenging as N increases. We still employ 200 walkers to conduct the modified nested sampling. However, the energy scaling fraction is increased to 0.999. In other words, we can only afford a minor energy reduction, so that the volume does not shrink too fast within each Monte Carlo integration. At each new energy level, outliers will be relaxed via rejection sampling, until their energies go below the new energy level. Following the relaxation, a long equilibration run with 8000 steps and a step size of 0.5 \AA is conducted via rejection sampling. Yet, even with this choice of parameters, the fluctuation of $V(E^*)$ is still large, and contributes more to the error of $\ln Q$ than the fluctuation of $\ln\langle f(E; E^*) \rangle$. In principle one can increase the number of walkers to reduce the volume integration error, but here we simply address this issue by executing the integration 5 times and taking the average value for numerical stability.

IV. CONCLUSION

In this article we addressed the question whether the partition function of a system can be evaluated via sampling. Our proposed partition function estimator (PFE) combines the expectation

value of the inverse Boltzmann factor, truncated at a suitable energy E^* via a Heaviside step function, with a correction term $V(E^*)$, which represents the volume of the configuration space with energy below E^* . Due to the energy truncation, the expectation value of the proposed function can be evaluated accurately via sampling, e.g. using Monte Carlo methods or molecular dynamics simulations, without suffering from undersampling of the high energy microstates. The volume term then accounts for the energy cutoff in configuration space. The energy cutoff parameter E^* can in theory be chosen freely, but we also outlined a strategy for determining E^* such that the sampling error is minimized. We demonstrated that the volume term can be directly evaluated from the sampling trajectory for 1- and 2- dimensional model systems, while a modified nested sampling integration method is proposed for systems of higher dimensionality.

As a proof of concept, the performance of PFE was tested using model examples, including the harmonic oscillator potential, the double-well potential, the Müller-Brown potential, and Lennard-Jones particles in a 3D box. Good agreement with the numerically exact solution and with reference calculations is found across all the test cases. While this result is not surprising, in the case of Lennard-Jones particles, we observe that the performance of PFE is comparable to the standard free energy perturbation (FEP) method. We also extended the Lennard-Jones example into the fluid regime, with up to 350 particles in a $(25\text{\AA})^3$ box. Our results show that PFE can provide a reliable estimation of $\ln Q$ and hence of the partition function as well. However, when calculating the chemical potential from the PFE results, the fluctuation of μ is still relatively large because a small relative error in a large value of $\ln Z$ is still significant when considering $\ln Z_{N+1} - \ln Z_N$. On the other hand, if one is interested in $\ln Z_M - \ln Z_N$ with very different numbers of M and N , then PFE may provide a more accurate result. This corresponds to calculating the free energy difference between two rather distinct states, which could be challenging for the standard FEP method, as it requires a certain overlap between the distributions of the two states. PFE is free from this requirement, as it yields the value of $\ln Z$ directly.

So far, PFE is not yet ready to handle large biological systems like the standard mFEP-MBAR method can, as the configuration space volume calculation is an unsolved problem for such systems. Nor can PFE currently calculate the partition function at multiple temperatures in a single run like nested sampling can, though we note that changing the temperature does not influence the volume calculation so that most of the data can be readily reused. Yet, the proposed method remains intriguing and warrants further development as it offers a fresh perspective on addressing a long-standing challenging task.

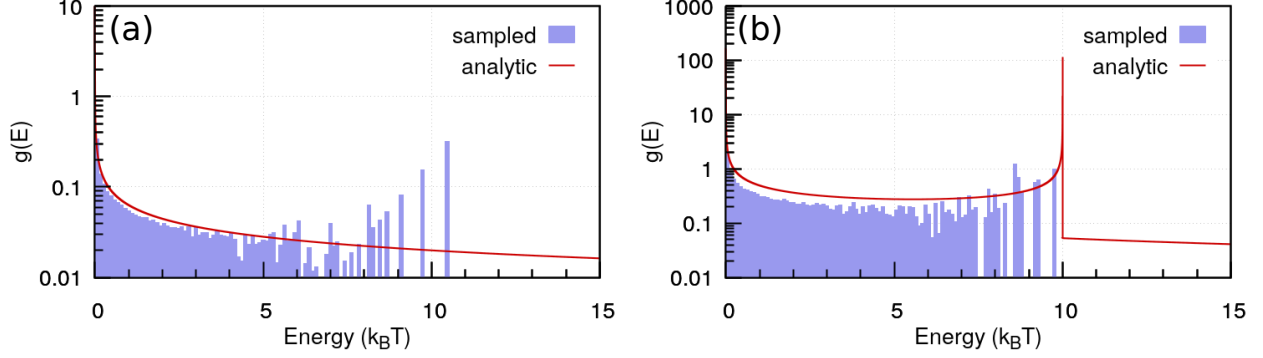


FIG. 8. (a) The density of states for the harmonic oscillator model. The blue bar chart depicts the density of states obtained from a trajectory. The analytical $g(E)$ is shown by the red curve. (b) The density of states for the double well potential model. The color code is the same as in panel (a).

APPENDIX

1. Density of states for 1D models

Fig. 8 depicts the density of states for the harmonic oscillator model (a) and for the double well potential model (b). The analytical $g(E)$ for the harmonic oscillator is given by,

$$g(E) = \sqrt{2/kE} \quad (22)$$

where k is the force constant. For the double well potential, the analytical solution reads,

$$g(E) = \begin{cases} \frac{x_0}{4h} \left[\left(\frac{E}{h} \left(1 + \sqrt{\frac{E}{h}} \right) \right)^{-1/2} + \left(\frac{E}{h} \left(1 - \sqrt{\frac{E}{h}} \right) \right)^{-1/2} \right] & \text{for } E \leq h \\ \frac{x_0}{4h} \left(\frac{E}{h} \left(1 + \sqrt{\frac{E}{h}} \right) \right)^{-1/2} & \text{for } E > h, \end{cases} \quad (23)$$

where x_0 and h are defined as in Eq. 20. As expected, the high energy region is not well sampled, so that $g(E)$ reconstructed from the trajectory fails to reproduce the exact solution.

2. Harmonic oscillator at high temperature

We also investigated the impact of temperature on PFE's performance for the harmonic oscillator potential. When the temperature is increased tenfold ($k_B T = 5.9616$), the sampling can reach even higher energies, but the sampling quality at the high energy tail is still bad, as shown in Fig. 9(a). Under the high temperature condition, PFE can still provide accurate $\ln Q$, as long as the contribution from high energy microstates is nullified, as evidenced by the red curve in panel

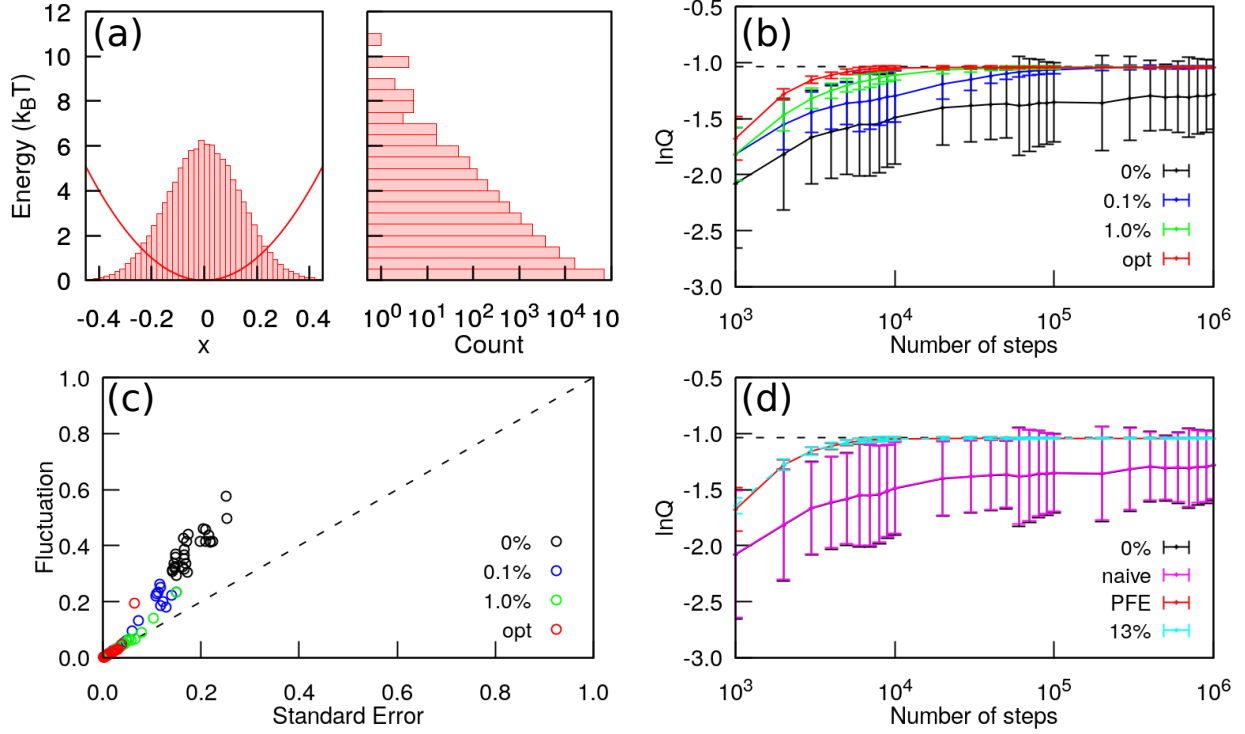


FIG. 9. (a) The potential energy curve (unit in $k_B T$) of the harmonic oscillator as well as the spatial and energy distributions. The system is the same as the one described in Fig. 2, but the temperature is now ten times higher. (b) $\ln Q$ computed using PFE. Shown are results obtained using different E^* . (c) A comparison between the standard error of $\ln \langle f(E; E^*) \rangle$ and the fluctuation of $\ln Q$ observed across 100 independent samplings. (d) A comparison between $\ln Q$ calculated via PFE (opt) and via a naive summation over the Boltzmann weights. The same trajectories are employed for data processing.

(b). The impact of a higher temperature reflects more on the fluctuation: Now, some data points show a larger fluctuation in $\ln Q$ (computed from 100 sampling repetitions) than the standard error of $\ln \langle f(E; E^*) \rangle$, see the blue points in Fig. 9(c). However, these are data corresponding to fewer energy samples (shorter trajectories). If sufficient energy samples are collected, the error becomes negligible again, and the observed fluctuation aligns with the standard error. Fig. 9(d) shows that a naive summation over Boltzmann factors again struggles to recover the correct $\ln Q$, while cutting off the top 13% energy samples again reproduces the PFE results (with optimal E^*) well.

3. Impact of the choice of p on $V(E^*)$ calculations

PFE calculates the volume term $V(E^*)$ via a modified version of nested sampling. At each iteration, the algorithm chooses a new energy level and calculates its volume as a ratio to the current one. In our current implementation, the new energy level is determined by reducing the distance between the previous energy level and the minimum energy (as encountered in the sampling) based on a chosen fraction p . Choosing different values of p can affect the performance of PFE.

Here we use the same systems as in Fig. 6 to discuss the effect of the fraction p . For each value of p we perform 10 independent calculations of $V(E^*)$ to obtain mean values and fluctuation for $\ln Q$. As shown in Fig. 10(a), the mean of $\ln Q$ is nearly identical for all values of p , but the fluctuation behaves very differently. The larger p will generate a new level closer to the old one, leading to a smaller fluctuation in $V(E^*)$, as shown in panel (b). This test reveals a few points for applying the modified nested sampling algorithm: First, using larger p can improve the accuracy from the Monte Carlo integration. Second, the result of the integration is identical for all choices of p , as long as the calculation is converged. Third, the required computational effort increases moderately with increasing p , see Fig. 10(c).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. The software code that has been used to produce these data will be made available on GitHub shortly after this work has been accepted for publication.

ACKNOWLEDGMENTS

This work is supported by the Kobilka Institute of Innovative Drug Discovery (KIIDD), School of Medicine, Chinese University of Hong Kong (Shenzhen), and the Shenzhen Science, Technology, and Innovation Commission (SZSTI), grant number JCYJ20230807114206014. YCC acknowledges the Royal Society for their support of the precursor to this work (NF171278). Special thanks to Prof. Guanglian Li and Prof. Ye Mei for their valuable discussions, to Prof. Livia B. Pártay for a fruitful discussion on nested sampling, as well as to Dr. Lantian Yao, Prof. Tzong-Yi Lee, and staff at KIIDD for providing a supportive working environment that nurtured this work.

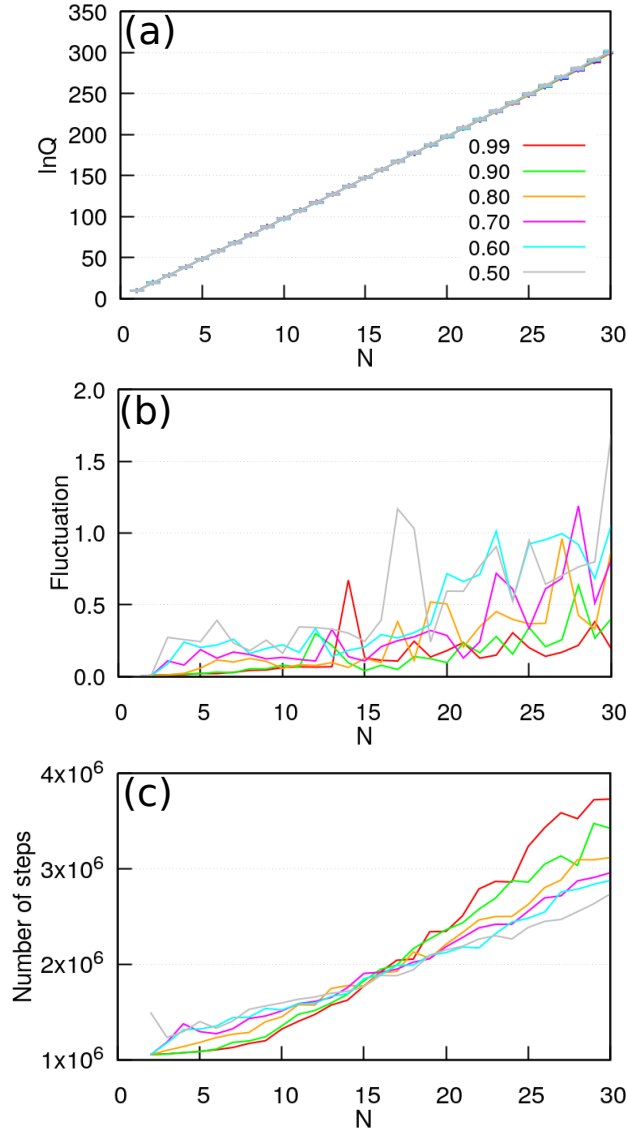


FIG. 10. (a) $\ln Q$ calculated using different fractions p for computing $V(E^*)$. For each choice of p , shown are the average and the standard deviation obtained from 10 independent calculations. The same color scheme is also used in panel (b) and (c). (b) Sampling fluctuation observed across 10 independent calculations. Employing a more aggressive fraction (such as $p = 0.5$) induces a larger fluctuation in $V(E^*)$. (c) The computational effort required to calculate $\ln Q$ using different choices of p .

REFERENCES

¹F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

²J. Skilling, AIP Conference Proceedings **735**, 395 (2004), https://pubs.aip.org/aip/acp/article-pdf/735/1/395/11702789/395_1_online.pdf.

- ³J. Skilling, *Bayesian Analysis* **1**, 833 (2006).
- ⁴G. Ashton, N. Bernstein, J. Buchner, X. Chen, G. Csányi, A. Fowlie, F. Feroz, M. Griffiths, W. Handley, M. Habeck, E. Higson, M. Hobson, A. Lasenby, D. Parkinson, L. B. Pártay, M. Pitkin, D. Schneider, J. S. Speagle, L. South, J. Veitch, P. Wacker, D. J. Wales, and D. Yallup, *Nature Reviews Methods Primers* **2**, 39 (2022).
- ⁵L. B. Pártay, G. Csányi, and N. Bernstein, *The European Physical Journal B* **94**, 159 (2021).
- ⁶L. B. Pártay, A. P. Bartók, and G. Csányi, *The Journal of Physical Chemistry B* **114**, 10502 (2010), pMID: 20701382, <https://doi.org/10.1021/jp1012973>.
- ⁷L. B. Pártay, *Computational Materials Science* **149**, 153 (2018).
- ⁸J. Dorrell and L. B. Pártay, *The Journal of Physical Chemistry B* **124**, 6015 (2020), pMID: 32543865, <https://doi.org/10.1021/acs.jpccb.0c03882>.
- ⁹G. A. Marchant and L. B. Pártay, *Physical Sciences Forum* **5** (2022), 10.3390/psf2022005005.
- ¹⁰K. Rossi, L. B. Pártay, G. Csányi, and F. Baletto, *Scientific Reports* **8**, 9150 (2018).
- ¹¹J. Dorrell and L. B. Pártay, *Phys. Chem. Chem. Phys.* **21**, 7305 (2019).
- ¹²R. W. Zwanzig, *The Journal of Chemical Physics* **22**, 1420 (1954), https://pubs.aip.org/aip/jcp/article-pdf/22/8/1420/18805749/1420_1_online.pdf.
- ¹³A. Pohorille, C. Jarzynski, and C. Chipot, *The Journal of Physical Chemistry B* **114**, 10235 (2010), pMID: 20701361, <https://doi.org/10.1021/jp102971x>.
- ¹⁴C. H. Bennett, *Journal of Computational Physics* **22**, 245 (1976).
- ¹⁵M. R. Shirts and J. D. Chodera, *The Journal of Chemical Physics* **129**, 124105 (2008), https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.2978177/15418484/124105_1_online.pdf.
- ¹⁶J. P. Valleau and D. N. Card, *The Journal of Chemical Physics* **57**, 5457 (1972), https://pubs.aip.org/aip/jcp/article-pdf/57/12/5457/18882977/5457_1_online.pdf.
- ¹⁷J. G. Kirkwood, *The Journal of Chemical Physics* **3**, 300 (1935), https://pubs.aip.org/aip/jcp/article-pdf/3/5/300/18788537/300_1_online.pdf.
- ¹⁸C. Jarzynski, *Phys. Rev. Lett.* **78**, 2690 (1997).
- ¹⁹C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology* (Springer, 2007).
- ²⁰S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus, *The Journal of Physical Chemistry B* **107**, 9535 (2003), <https://doi.org/10.1021/jp0217839>.
- ²¹H.-J. Woo and B. Roux, *Proceedings of the National Academy of Sciences* **102**, 6825 (2005), <https://www.pnas.org/doi/pdf/10.1073/pnas.0409005102>.

- ²²D. L. Mobley, J. D. Chodera, and K. A. Dill, *Journal of Chemical Theory and Computation* **3**, 1231 (2007), pMID: 18843379, <https://doi.org/10.1021/ct700032n>.
- ²³L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, and R. Abel, *Journal of the American Chemical Society* **137**, 2695 (2015), pMID: 25625324, <https://doi.org/10.1021/ja512751q>.
- ²⁴J. C. Gumbart, B. Roux, and C. Chipot, *Journal of Chemical Theory and Computation* **9**, 3789 (2013), <https://doi.org/10.1021/ct400273t>.
- ²⁵J. Comer, K. Schulten, and C. Chipot, *Journal of Chemical Theory and Computation* **13**, 2523 (2017), pMID: 28475319, <https://doi.org/10.1021/acs.jctc.7b00264>.
- ²⁶Z. Cournia, B. Allen, and W. Sherman, *Journal of Chemical Information and Modeling* **57**, 2911 (2017), pMID: 29243483, <https://doi.org/10.1021/acs.jcim.7b00564>.
- ²⁷I. Muegge and Y. Hu, *ACS Medicinal Chemistry Letters* **14**, 244 (2023), <https://doi.org/10.1021/acsmchemlett.2c00541>.
- ²⁸D. M. York, *ACS Physical Chemistry Au* **3**, 478 (2023), <https://doi.org/10.1021/acspchemau.3c00033>.
- ²⁹J. H. Moore, C. Margreitter, J. P. Janet, O. Engkvist, B. L. de Groot, and V. Gapsys, *Communications Chemistry* **6**, 82 (2023).
- ³⁰G. Cole, in *An Introduction to the Statistical Theory of Classical Simple Dense Fluids*, edited by G. Cole (Pergamon, 1967) pp. 40–64.
- ³¹Y. Sugita and Y. Okamoto, *Chemical Physics Letters* **314**, 141 (1999).
- ³²A. Laio and M. Parrinello, *Proceedings of the National Academy of Sciences* **99**, 12562 (2002), <http://www.pnas.org/content/99/20/12562.full.pdf>.
- ³³P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, *PLOS Computational Biology* **13**, e1005659 (2017).