The Journal of Chemical Physics

RESEARCH ARTICLE | JUNE 01 2023

Understanding the sources of error in MBAR through asymptotic analysis

Xiang Sherry Li ; Brian Van Koten ; Aaron R. Dinner ; Erik H. Thiede

J. Chem. Phys. 158, 214107 (2023)
https://doi.org/10.1063/5.0147243

AIP Publishing

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Understanding the sources of error in MBAR through asymptotic analysis

Cite as: J. Chem. Phys. 158, 214107 (2023); doi: 10.1063/5.0147243
Submitted: 20 February 2023 • Accepted: 8 May 2023 •
Published Online: 1 June 2023

Xiang Sherry Li,[1] Brian Van Koten,[2] Aaron R. Dinner,[1] and Erik H. Thiede[3,a)]

AFFILIATIONS

[1]Department of Chemistry and James Franck Institute, University of Chicago, Chicago, Illinois 60637, USA
[2]Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003, USA
[3]Center for Computational Mathematics, Flatiron Institute, New York, New York 10010, USA

[a)]**Author to whom correspondence should be addressed:** ehthiede@flatironinstitute.org

ABSTRACT

Many sampling strategies commonly used in molecular dynamics, such as umbrella sampling and alchemical free energy methods, involve sampling from multiple states. The Multistate Bennett Acceptance Ratio (MBAR) formalism is a widely used way of recombining the resulting data. However, the error of the MBAR estimator is not well-understood: previous error analyses of MBAR assumed independent samples. In this work, we derive a central limit theorem for MBAR estimates in the presence of correlated data, further justifying the use of MBAR in practical applications. Moreover, our central limit theorem yields an estimate of the error that can be decomposed into contributions from the individual Markov chains used to sample the states. This gives additional insight into how sampling in each state affects the overall error. We demonstrate our error estimator on an umbrella sampling calculation of the free energy of isomerization of the alanine dipeptide and an alchemical calculation of the hydration free energy of methane. Our numerical results demonstrate that the time required for the Markov chain to decorrelate in individual states can contribute considerably to the total MBAR error, highlighting the importance of accurately addressing the effect of sample correlation.

I. INTRODUCTION

Molecular dynamics simulations provide a means to compute key quantities in statistical mechanics, typically in the form of ensemble averages of certain observables. In principle, it is possible to estimate ensemble averages by running a long, unbiased simulation of a system and averaging over the resulting trajectory. However, in practice, this can be inefficient. To estimate many ensemble averages, the simulation must explore multiple, well-separated regions of a system's configuration space. For instance, estimates of the free energy difference between two states are notoriously slow to converge.[1,2]

A common strategy for addressing this problem is to sample from multiple states. The resulting data are then combined with appropriate weights. This approach is the basis of multiple simulation strategies, such as umbrella sampling, parallel tempering, and alchemical free energy simulation. Umbrella sampling ensures that a broad range of molecular conformations is seen by running a series of independent simulations, each biased to sample a different region of a physical system's phase space.[3,4] Parallel tempering simulates several replicas of a system with the same Hamiltonian but at different temperatures to allow the system to cross energetic barriers.[5,6] Alchemical free energy calculations of molecular systems estimate the free energy difference between two molecular states by interpolating between their Hamiltonians.[7–9] In all of these cases, data sampled from multiple states must be combined to estimate the averages of interest.

A popular algorithm for performing the reweighting is the Multistate Bennett Acceptance Ratio (MBAR). Popularized in Ref. 10, the algorithm has been independently derived in multiple contexts[11,12] and is closely related to the weighted histogram method (WHAM).[13,14] For statistically independent samples, MBAR combines the data across multiple states in a statistically optimal manner. Although molecular dynamics simulations are correlated, MBAR nevertheless achieves good results in practice. However, solving for the MBAR estimate involves solving a nonlinear fixed point problem, which complicates error analysis. Previous attempts to construct error estimates have explicitly assumed that samples are statistically independent.[10,15] Again, we stress that this is typically not true for molecular dynamics data.

In this paper, we build on Refs. 16 and 17 to derive a formal expression for the asymptotic variance of the MBAR estimator that explicitly accounts for correlation in sampled data. As such, it helps theoretically justify the use of MBAR in practical situations. Our work also leads to an asymptotic error estimator that can be decomposed into error contributions from individual states. This can potentially give practitioners insight into how sampling in individual states affects global error and may lead to adaptive sampling strategies to accelerate convergence. As such, it helps theoretically justify the use of MBAR in practical situations. Our work also leads to an asymptotic error estimator that can be decomposed into error contributions from individual states. This can potentially give practitioners insight into how sampling in individual states affects error and may lead to adaptive sampling strategies to accelerate convergence.

II. BACKGROUND ON MONTE CARLO AND ASYMPTOTIC ERROR

Many fundamental quantities in statistical mechanics take the form of high-dimensional integrals. Physical theories often require estimating averages over a physical system's Boltzmann distribution,

$$\langle g \rangle = \frac{\int g(x) e^{-H(x)/k_B T} dx}{\int e^{-H(x)/k_B T} dx}, \tag{1}$$

where H is the system's Hamiltonian and x is its configuration in $\mathbb{R}^n$. Alternatively, they may require estimates of the free energy difference between two regions of phase space,

$$\Delta G_{A \to B} = -k_B T \log \frac{\int \mathbb{1}_B(x) e^{-H(x)/k_B T} dx}{\int \mathbb{1}_A(x) e^{-H(x)/k_B T} dx}, \tag{2}$$

$$= -k_B T \log \frac{\langle \mathbb{1}_B \rangle}{\langle \mathbb{1}_A \rangle}, \tag{3}$$

where $\mathbb{1}_D(x)$ is 1 if the configuration x is in a region labeled D and 0 otherwise. Similar equations are used to estimate free energy surfaces. Letting σ be a function that maps a configuration to the value of a collective variable, the free energy surface along the collective variable obeys

$$G(s) = -k_B T \log \int \delta(s - \sigma(x)) e^{-H(x)/k_B T} dx + C, \tag{4}$$

where C is an unknown constant. In practice, the free energy at each value of s is approximated by the free energy of a small histogram bin centered at s. If all histogram bins used are of the same size, we can approximate

$$G(s) \approx -k_B T \log \int \mathbb{1}_S(x) e^{-H(x)/k_B T} + C', \tag{5}$$

where we have defined S to be the set of configurations that map to the histogram bin centered at s in collective variable space.

Alternatively, rather than estimating free energy differences between conformations, we may wish to estimate the free energy difference between two Hamiltonians, which is given by

$$\Delta G_{\alpha \to \beta} = -k_B T \log \frac{\int e^{-H_\beta(x)/k_B T} dx}{\int e^{-H_\alpha(x)/k_B T} dx}. \tag{6}$$

For most systems, these integrals are too complex to be evaluated analytically, and the dimension of x is too high to use quadrature. Instead, they are typically treated by Monte Carlo methods.

A. Markov chain Monte Carlo

Assume that we are given a probability distribution with an unnormalized density q over the Lebesgue measure on $\mathbb{R}^n$. We can then write the average of a function $g : \mathbb{R}^n \to \mathbb{R}$ over this distribution as

$$\langle g \rangle = \frac{\int g(x) q(x) dx}{c}, \qquad c = \int q(x) dx. \tag{7}$$

Particular choices of g allow us to rewrite key quantities in statistical mechanics as such *ensemble averages*. For instance, in statistical mechanics, q is typically the Boltzmann factor,

$$q(x) = e^{-H(x)/k_B T}, \tag{8}$$

and substituting into our expression for an average recovers (1). The free energy difference between regions of phase space in (3) is merely the ratio between two averages. Similarly, to estimate the free energy difference between two Hamiltonians, we can set q to $\exp(-H_\alpha/k_B T)$ and rewrite (6) as

$$e^{-\Delta G_{\alpha \to \beta}/k_B T} = \left\langle e^{-(H_\beta(x) - H_\alpha(x))/k_B T} \right\rangle. \tag{9}$$

Monte Carlo methods approximate ensemble averages by drawing a sequence of N samples $\{X_t\}$ from the probability distribution and averaging over them. If the sampling procedure is chosen appropriately, we expect sample averages to converge to the true (ensemble) average over the distribution associated with q,

$$\bar{g} = \frac{1}{N} \sum_{t=0}^{N-1} g(X_t) \xrightarrow{a.s.} \langle g \rangle. \tag{10}$$

Here, *a.s.* denotes almost sure convergence, a strong form of probabilistic convergence (specifically, the probability of $\bar{g}$ not converging to $\langle g \rangle$ is zero). If the samples are statistically independent, we say that our samples are *independent and identically distributed* (IID), and (10) is guaranteed to hold by the Law of Large Numbers.[2] However, in practice, it is often impossible to generate IID samples, and we must instead generate samples by running a Markov chain that has q as the density associated with its stationary distribution: a practice known as *Markov chain Monte Carlo* (MCMC). Then, (10) still holds if the Markov chain is ergodic.[2]

B. Asymptotic variance of Monte Carlo estimates

While (10) guarantees that the error goes to zero as the number of samples increases, it says nothing about how quickly this happens. A common method to quantify how the sampling error decreases as the sample size increases is to use a *central limit theorem* (CLT): a theorem showing that a sequence of random variables converges to a known normal distribution.[2] Specifically, assume that we wish to evaluate the ensemble average of several functions, each denoted g_i. Concatenating our sample means and ensemble averages into vectors that we denote by $\bar{\mathbf{g}}$ and $\langle \mathbf{g} \rangle$, respectively, we can often show

that the error between the two converges to a multivariate normal (Gaussian) distribution when appropriately scaled

$$\sqrt{N}(\tilde{\mathbf{g}} - \langle \mathbf{g} \rangle) \xrightarrow{d} \mathbf{N}(0, \Sigma). \tag{11}$$

Here, $\mathbf{N}$ denotes a normal (Gaussian) random variable with mean vector μ and covariance matrix Σ, known as the *asymptotic covariance*. The symbol $\xrightarrow{d}$ denotes convergence in distribution (a weaker form of probabilistic convergence than almost sure convergence). For IID samples, (11) holds for all functions with finite variance, and Σ_{ij} is simply the covariance between g_i and g_j over the distribution associated with q. If samples are instead gathered from a Markov chain X_t, proving a CLT requires more technical conditions on the nature of the Markov Chain and $\mathbf{g}$.[18–21] However, for most commonly used Markov chains and most reasonable applications, we can expect (11) to hold. In this case, if the Markov chain is time-homogeneous (i.e., the rule for updating the Markov chain is the same at all times) and stationary (i.e., at each time the Markov chain has the same distribution), and has the distribution associated with q as its ergodic distribution, the asymptotic covariance matrix is given by

$$\Sigma_{ij} = \operatorname{cov}\{g_i(X_t), g_j(X_t)\} + 2\sum_{k=1}^{\infty} \operatorname{cov}\{g_i(X_t), g_j(X_{t+k})\}. \tag{12}$$

The CLT and the asymptotic covariance help diagnose the error and convergence of a Markov chain Monte Carlo simulation. For example, under mild technical conditions (specifically geometric ergodicity and bounded g), Σ_{ii}/N is asymptotic to $\operatorname{var}\{\tilde{g}_i\}$.[21] Consequently, we can treat $\sqrt{\Sigma_{ii}/N}$ as a rough estimate for the error associated with using $\tilde{g}_i$ to estimate $\langle g_i \rangle$. The sampling efficiency of the Markov chain relative to IID sampling from the distribution associated with q can be quantified by the *autocorrelation time*,

$$\tau_{g_i} = \frac{\Sigma_{ii}}{\operatorname{var}\{g_i\}}. \tag{13}$$

Since $\operatorname{var}\{g_i\}/N$ is the variance for IID sampling, we can interpret the autocorrelation time as how many MCMC samples are required to achieve the same reduction in error as a single IID sample.[2,8]

III. THE MBAR EQUATIONS

In Sec. II B, we considered sampling from a single distribution. However, we may often have samples collected from multiple, related probability distributions. For concreteness, assume we have L probability distributions, each with an unnormalized probability density q_i. We refer to these distributions as *states*. The ensemble average of an observable $g(x)$ in each state is given by

$$\langle g \rangle_i = \frac{\int g(x) q_i(x) dx}{c_i}, \quad c_i = \int q_i(x) dx. \tag{14}$$

Here, the constant c_i is the normalization constant for q_i. If q_i is a Boltzmann distribution, then c_i is the corresponding partition function. Next, we assume that for every state, we have collected a set of N_i samples, denoted $\{X_t^i\}$. We can then approximate $\langle g \rangle_i$ by the sample average,

$$\langle g \rangle_i \approx \frac{1}{N_i} \sum_{t=1}^{N_i} g(X_t^i).$$

However, if the states have shared regions with high probability, we can construct improved estimates of (14) by using data from all states, not just state i. This is the aim of the MBAR algorithm.[10,11] Following the treatment in Refs. 22 and 23, we observe that we can view the union of the samples from the states as samples from a combined distribution known as a *mixture distribution*. Let

$$N = \sum_{i=1}^{L} N_i$$

be the total sample size, and let

$$\kappa_i = \frac{N_i}{N}$$

be the fraction of sample points collected in state i. To simplify the presentation, we assume that κ_i is constant and always greater than zero (a version of our main result that relaxes this assumption is given in the supplementary material). We define the mixture distribution to be

$$\pi_{\text{mix}}(x) = \sum_{i=1}^{L} \kappa_i q_i(x)/c_i. \tag{15}$$

We can then write

$$\langle g \rangle_i = \int \frac{g(x) q_i(x)/c_i}{\pi_{\text{mix}}(x)} \pi_{\text{mix}}(x) dx = \int \frac{g(x) q_i(x)/c_i}{\sum_{k=1}^{L} \kappa_k q_k(x)/c_k} \sum_{j=1}^{L} \kappa_j q_j(x)/c_j dx, \tag{16}$$

$$= \sum_{j=1}^{L} \kappa_j \left\langle \frac{g q_i/c_i}{\sum_{k=1}^{L} \kappa_k q_k/c_k} \right\rangle_j. \tag{17}$$

In general, the normalization constants for the states are not known. We, therefore, rewrite (17) in terms of the states' relative free energies, which we denote by f_i. We arbitrarily set the average free energy to be zero, so

$$\frac{1}{L} \sum_{i=1}^{L} f_i = 0, \tag{18}$$

and, therefore, the free energies are defined by

$$f_i = -\log c_i + \frac{1}{L} \sum_{j=1}^{L} \log c_j. \tag{19}$$

Here and subsequently, we give the free energy as a ratio to $k_B T$ so that it is dimensionless.

Dividing both the numerator and denominator of (17) by $\exp\left(-(1/L)\sum_{i=1}^{L} f_i\right)$, after a few manipulations we have

$$\langle g \rangle_i = \sum_{j=1}^{L} \kappa_j \left\langle \frac{g q_i e^{f_i}}{\sum_{k=1}^{L} \kappa_k q_k e^{f_k}} \right\rangle_j. \tag{20}$$

This equation can be used to estimate $\langle g \rangle_i$ if we are given an estimate of the free energies, $\tilde{f}$. Replacing each ensemble average on the right-hand side with a Monte Carlo estimate, we have

$$\tilde{g}_i = \sum_{j=1}^{L} \frac{\kappa_j}{N_j} \sum_{t=1}^{N_j} \frac{g(X_t^j) q_i(X_t^j) e^{\tilde{f}_i}}{\sum_{k=1}^{L} \kappa_k q_k(X_t^j) e^{\tilde{f}_k}}. \tag{21}$$

This estimator uses data from every state, not just state i. Moreover, we can also use (20) to estimate the free energies themselves. Since the ensemble average of the function $g(x) = 1$ is always 1,

$$1 = \sum_{j=1}^{L} \kappa_j \left\langle \frac{q_i e^{f_i}}{\sum_{k=1}^{L} \kappa_k q_k e^{f_k}} \right\rangle_j, \tag{22}$$

$$\Rightarrow \quad f_i = -\log \sum_{j=1}^{L} \kappa_j \left\langle \frac{q_i}{\sum_{k=1}^{L} \kappa_k q_k e^{f_k}} \right\rangle_j. \tag{23}$$

One can thus estimate the free energy by defining $\tilde{f}$ to be the solution to

$$\tilde{f}_i = -\log \left(\sum_{j=1}^{L} \frac{\kappa_j}{N_j} \sum_{t=1}^{N_j} \frac{q_i(X_t^j)}{\sum_{k=1}^{L} \kappa_k q_k(X_t^j) e^{\tilde{f}_k}} \right). \tag{24}$$

Not only can this equation be solved using standard root-finding methods, such as Newton–Raphson and gradient descent,[10] but also there exist algorithms for solving it through a succession of estimation tasks.[16,17,24] Equations (21) and (24) are the MBAR estimates of the ensemble average and the free energies, respectively.[10] With sufficient overlap between the samples from different states, (23) uniquely determines f. Specifically, if the matrix $M_{ij} = \langle q_i \rangle_j$ is irreducible, then by Theorem 1 in Ref. 22 or Proposition 1.1 in Ref. 25, the f_i is uniquely specified by (23). An analogous statement holds for $\tilde{f}$. When M is irreducible, Theorem 1.1 in Ref. 25 implies that Eq. (24) almost surely has a unique solution $\tilde{f}$ when the total sample size N is sufficiently large. Moreover, the estimates of the free energies and the ensemble averages converge to f_i and $\langle g \rangle_i$ as N increases. To be precise, $\tilde{f}_i \xrightarrow{a.s.} f_i$ and $\tilde{g}_i \xrightarrow{a.s.} \langle g \rangle_i$ by Theorem 1 in Ref. 22.

A. Estimating chemical quantities using MBAR

Specific manipulations of state free energies and ensemble averages allow us to reconstruct quantities of interest in a broad range of contexts. Here, we discuss the analysis of data from three common algorithms: parallel tempering, alchemical free energy simulations, and umbrella sampling.

In parallel tempering, we seek to estimate ensemble averages for a system with unnormalized probability density,

$$e^{-H(x)/k_B T}. \tag{25}$$

However, this density may be highly multimodal, making the probability distribution difficult to sample. Parallel tempering addresses this by running multiple copies of the system with the same Hamiltonian but different temperatures.[5,6] We write their distributions as

$$q_i(x) = e^{-H(x)/k_B(T+\delta T_i)}. \tag{26}$$

One copy, here arbitrarily chosen to have index 1, is set to be at the original temperature (i.e., $\delta T_1 = 0$), and all other copies have $\delta T_i \neq 0$. The copies then periodically swap molecular configurations via Monte Carlo moves on the space of copies. In principle, one can estimate averages over (25) using only configurations drawn from q_1. However, using the MBAR estimator (21) allows one to use data from all states, giving a more accurate answer.

In alchemical free energy simulations, we seek to estimate the free energy difference between two Hamiltonians as in (6).[7–9] However, rather than sampling only the state with H_α, we sample a set of L states that interpolate between H_α and H_β. A simple choice would be to set

$$-k_B T \log q_i = H_\alpha + \lambda\left(\frac{i-1}{L-1}\right)(H_\beta - H_\alpha), \tag{27}$$

where $\lambda : [0,1] \to [0,1]$ is a monotonic function such that $\lambda(0) = 0$ and $\lambda(1) = 1$, although, in practice, more complex interpolations are often required.[26–28] With this set of state definitions, the (unitless) free energy difference between the two Hamiltonians is simply the difference between the free energies of the first and last states,

$$-\log \frac{\int e^{-H_\beta(x)/k_B T} dx}{\int e^{-H_\alpha(x)/k_B T} dx} = f_L - f_1. \tag{28}$$

Consequently, we can solve (24) and estimate the free energy difference as $\tilde{f}_L - \tilde{f}_1$.

In umbrella sampling,[3,4] we construct a collection of states,

$$q_i(x) = \psi_i(x) q(x), \tag{29}$$

by multiplying an unnormalized density q with a biasing function ψ_i. We then aim to estimate averages of observables over the distribution associated with q, such as those in (1) and (3). To estimate these averages using MBAR by steps similar to those used to derive (17), we write

$$\begin{aligned}
\frac{\int g(x) q(x) dx}{\int q(x) dx} &= \frac{\int g(x) q(x) (\pi_{\text{mix}}(x)/\pi_{\text{mix}}(x)) dx}{\int q(x) (\pi_{\text{mix}}(x)/\pi_{\text{mix}}(x)) dx} \\
&= \frac{\int g(x) q(x) \frac{\sum_j \kappa_j q_j(x) e^{f_j}}{\sum_l \kappa_l q_l(x) e^{f_l}} dx}{\int q(x) \frac{\sum_k \kappa_k q_k(x) e^{f_k}}{\sum_m \kappa_m q_m(x) e^{f_m}} dx} \\
&= \frac{\sum_{j=1}^{L} \kappa_j \left\langle g q / \left(\sum_{l=1}^{L} \kappa_l q_l e^{f_l} \right) \right\rangle_j}{\sum_{k=1}^{L} \kappa_k \left\langle q / \left(\sum_{m=1}^{L} \kappa_m q_m e^{f_m} \right) \right\rangle_k} \\
&= \frac{\sum_{j=1}^{L} \kappa_j \left\langle g / \left(\sum_{l=1}^{L} \kappa_l \psi_l e^{f_l} \right) \right\rangle_j}{\sum_{k=1}^{L} \kappa_k \left\langle 1 / \left(\sum_{m=1}^{L} \kappa_m \psi_m e^{f_m} \right) \right\rangle_k}.
\end{aligned} \tag{30}$$

We can also use umbrella sampling to estimate the difference in free energy between two states. Comparing to (3) and setting q to be the Boltzmann factor, we have

$$\begin{aligned}
\Delta G_{A \to B} &= -k_B T \log \frac{\int \mathbb{1}_B(x) e^{-H(x)/k_B T} dx}{\int \mathbb{1}_A(x) e^{-H(x)/k_B T} dx} \\
&= -k_B T \log \frac{\sum_{j=1}^{L} \kappa_j \left\langle \mathbb{1}_B q / \left(\sum_{l=1}^{L} \kappa_l q_l e^{f_l} \right) \right\rangle_j}{\sum_{k=1}^{L} \kappa_k \left\langle \mathbb{1}_A q / \left(\sum_{m=1}^{L} \kappa_m q_m e^{f_m} \right) \right\rangle_k} \\
&= -k_B T \log \frac{\sum_{j=1}^{L} \kappa_j \left\langle \mathbb{1}_B / \left(\sum_{l=1}^{L} \kappa_l \psi_l e^{f_l} \right) \right\rangle_j}{\sum_{k=1}^{L} \kappa_k \left\langle \mathbb{1}_A / \left(\sum_{m=1}^{L} \kappa_m \psi_m e^{f_m} \right) \right\rangle_k},
\end{aligned} \tag{31}$$

by steps similar to those for (30).

These examples show how MBAR can be used to construct estimates from algorithms that collect data in multiple states. When IID samples are collected from each state, then MBAR gives the maximum likelihood estimate[11] and achieves the best possible mean-squared error in the large-sample limit.[10] As such, MBAR is commonly treated as an algorithm that operates on statistically independent samples, and practitioners often subsample trajectories to attempt to construct a dataset of statistically independent samples. However, evaluating the criteria used for independence typically involves estimating one or more autocorrelation times, a notoriously difficult statistical task.[29] Moreover, subsampling runs the risk of discarding too much data, decreasing the statistical power of the method. Consequently, once the burn-in period has been removed from the trajectory, we believe it preferable to apply the MBAR estimator without subsampling the data. While MBAR does not give the maximum likelihood estimate for correlated samples, it is still a consistent estimator as shown in Ref. 22. Moreover, applying MBAR to correlated data has been observed to give good results in practice.[22] Indeed, in Sec. V, we find that applying the MBAR estimator without subsampling the data performs as well or better than with subsampling.

IV. ASYMPTOTIC VARIANCE FOR THE MBAR EQUATIONS

However, quantifying the uncertainty in MBAR estimates has proved challenging. In previous work, Kong *et al.* constructed an estimator for the asymptotic covariance using the Cramer–Rao lower bound of the variance.[15] When samples are uncorrelated, MBAR achieves this lower bound, but, when samples are correlated, this estimator underestimates the asymptotic error. Moreover, while it gives an estimate of the total error, it does not give immediate insight into how sampling in the individual states affects the overall error. This makes it difficult to determine how one should tune simulation hyperparameters and/or allocate additional simulations.

In this work, we pursue an alternate approach by constructing a CLT for MBAR estimates. As discussed in Sec. II B, CLTs are able to capture the effect of the dynamics on sampling error. Moreover, previous work on closely related algorithms for recombining data from multiple states[16,17] showed that CLTs can be used to connect the sampling of individual states to the total error of the estimate. Our CLT gives detailed insight into how the parameters of multistate simulations contribute to the total error.

The approach taken in this section builds upon the work of Geyer.[22] Our contribution is essential to fill in missing details and to correct errors. Most importantly, the formula for the asymptotic variance of observable averages $\bar{g}$ is not correct in Ref. 22.

Our discussion has two parts. In the first part, we note that all of the estimates described in Sec. III are functions of estimated state free energies and ensemble averages. Consequently, we show that a CLT holds for these quantities. In the second part, we show that this CLT can then be extended to error estimates for arbitrary functions of ensemble averages.

A. CLTs for the raw output of MBAR

MBAR estimates of observables require calculating the values of $\bar{f}$ as well as one or more empirical averages of the form,

$$\bar{\omega} = \sum_{j=1}^{L} \frac{\kappa_j}{N_j} \sum_{t=1}^{N_j} \frac{w(X_t^j)}{\sum_{k=1}^{L} \kappa_k q_k(X_t^j) e^{\bar{f}_k}}, \tag{32}$$

for some function $w : \mathbb{R}^n \to \mathbb{R}$. For instance, in (21), we set $w = q_i g$ and subsequently multiply by $e^{\bar{f}_i}$. The presence of $\bar{f}$ in (30) means that errors in estimated state free energies can propagate to averages of observables. Consequently, we must consider the asymptotic covariance of the free energies and our observables jointly.

To do so, we rewrite (24) and (32) as a single root finding problem. We concatenate the vector of estimated free energies and empirical averages into a single vector,

$$\bar{\nu} = \left(\bar{f}_1, \ldots, \bar{f}_L, \bar{\omega}_1, \ldots, \bar{\omega}_M\right). \tag{33}$$

Under the assumptions discussed in Sec. II, we expect $\bar{\nu}$ to converge to

$$\nu = (f_1, \ldots, f_L, \omega_1, \ldots, \omega_M), \tag{34}$$

where we have defined

$$\omega_i = \sum_{j=1}^{L} \kappa_j \left\langle \frac{w_i}{\sum_{k=1}^{L} \kappa_k q_k e^{f_k}} \right\rangle_j. \tag{35}$$

Concatenating the free energies and empirical averages together allows us to derive a CLT for both the free energies and empirical averages with a single proof. Rearranging (24) and (32), we see that the vector $\bar{\nu}$ is the root of the function $\bar{F} : \mathbb{R}^{L+M} \to \mathbb{R}^{L+M}$, where if $i \le L$,

$$\bar{F}_i(y) = \kappa_i - \sum_{j=1}^{L} \kappa_j \frac{1}{N_j} \sum_{t=1}^{N_j} \frac{\kappa_i q_i(X_t^j) e^{y_i}}{\sum_{k=1}^{L} \kappa_k q_k(X_t^j) e^{y_k}}, \tag{36}$$

and if $i > L$,

$$\bar{F}_i(y) = y_i - \sum_{j=1}^{L} \kappa_j \frac{1}{N_j} \sum_{t=1}^{N_j} \frac{w_{i-L}(X_t^j)}{\sum_{k=1}^{L} \kappa_k q_k(X_t^j) e^{y_k}}. \tag{37}$$

We remind the reader that the first L entries in $\bar{\nu}$ correspond to the MBAR estimates of the L states' relative free energies and that subsequent entries of $\bar{\nu}$ correspond to estimates of ensemble averages. Writing the MBAR estimates as the roots of $\bar{F}$ suggests a strategy for proving a CLT. For any fixed y, each element in $\bar{F}(y)$ is a sum of sample averages over our states. It is, therefore, reasonable to assume the existence of a CLT for each of the sample averages. If we can then convert a CLT for each average into a CLT for the *roots* of $\bar{F}$, then we have proven a CLT for MBAR estimates. Indeed, this is precisely the strategy we pursue. A full proof of the CLT is given in Sec. I of the supplementary material. Here, we introduce the key quantities and state our results.

We first discuss the asymptotic covariance structure of each of the averages in (36) and (37). For convenience, we define

$$\xi_i(x, y) = \frac{\kappa_i q_i(x) e^{y_i}}{\sum_{k=1}^{L} \kappa_k q_k(x) e^{y_k}} \tag{38}$$

for $i \le L$ and

$$\xi_i(x, y) = \frac{w_{i-L}(x)}{\sum_{k=1}^{L} \kappa_k q_k(x) e^{y_k}} \tag{39}$$

for $i > L$. We can then write $\bar{F}_i(y)$ using a κ_j-weighted sum of sample averages of the form

$$\bar{\xi}_i^j(y) = \frac{1}{N_j}\sum_{t=1}^{N_j} \xi_i(X_t^j, y).$$

In the limit as $N \to \infty$, $\bar{\xi}_i^j(y)$ converges to $\langle \xi_i(\cdot, y)\rangle_j$ and $\bar{F}$ converges to

$$F_i(y) = \kappa_i - \sum_{j=1}^{L} \kappa_j \langle \xi_i(\cdot, y_i)\rangle_j \tag{40}$$

for $i \le L$, and to

$$F_i(y) = y_i - \sum_{j=1}^{L} \kappa_j \langle \xi_i(\cdot, y_i)\rangle_j. \tag{41}$$

for $i > L$.

We assume that a central limit theorem holds for the sample averages $\bar{\xi}_i^j(y)$. This assumption is likely to hold in practice: One could use, for example, the results in Chapter 17 of Ref. 30 to verify our CLT assumption (42). See Ref. 2 for a more detailed discussion of the CLT in the context of molecular dynamics.

To phrase our assumption precisely, we assume that for any fixed y,

$$\sqrt{N}\left(\bar{\xi}(y) - \langle \xi(\cdot, y)\rangle\right) \xrightarrow{d} \mathbf{N}(0, \Xi(y)). \tag{42}$$

Here, we have written all our sample averages as a single vector,

$$\bar{\xi}(y) = (\bar{\xi}_1^1(y), \ldots, \bar{\xi}_{L+M}^1(y), \bar{\xi}_1^2(y), \ldots, \bar{\xi}_{L+M}^L(y)).$$

The vector $\bar{\xi}$ has $L \times (L + M)$ elements: the first $L + M$ elements correspond to all of the sample averages required by MBAR that are estimated in the first state, the second $L + M$ elements correspond to all of the sample averages that are estimated in the second state, and so forth. The covariance matrix $\Xi(y)$ can be written in block form as

$$\Xi(y) = \left[\begin{array}{c|c|c|c} \Xi^{11}(y) & \Xi^{12}(y) & \cdots & \Xi^{1L}(y) \\ \hline \Xi^{21}(y) & \Xi^{22}(y) & \cdots & \Xi^{2L}(y) \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \Xi^{L1}(y) & \Xi^{L2}(y) & \cdots & \Xi^{LL}(y) \end{array}\right] \tag{43}$$

Here, we have written the covariance matrix using a block structure consisting of L^2 blocks. Each block Ξ^{lm} is the covariance matrix between the averages in state l and those in state m and is thus a real matrix of size $(L + M) \times (L + M)$.

Given a fixed value of y, the matrix Ξ gives us the asymptotic covariance of $\bar{F}$. We now convert this into an expression for the asymptotic covariance of the roots of $\bar{F}$.

Theorem 4.1. *Assume that when $y = v$, the central limit theorem in (42) holds. Let $A \in \mathbb{R}^{(L+M)\times(L+M)}$ be the matrix with entries,*

$$A_{jl} = \sum_{m=1}^{L}\sum_{n=1}^{L} \kappa_m \kappa_n \, \Xi_{jl}^{mn}(v). \tag{44}$$

Under some technical assumptions (given in Sec. I of the supplementary material),

$$\sqrt{N}(\bar{\nu} - \nu) \xrightarrow{d} N(0, \Gamma A \Gamma^T), \tag{45}$$

where $\Gamma \in \mathbb{R}^{(L+M)\times(L+M)}$ is a matrix that can be expressed in block form as

$$\Gamma = \begin{bmatrix} H^{\#} & 0 \\ \beta H^{\#} & I \end{bmatrix}, \tag{46}$$

where I is the $L \times L$ identity matrix and the matrices $H \in \mathbb{R}^{M\times M}$ and $\beta \in \mathbb{R}^{L\times M}$ are given by

$$H_{ij} = \kappa_i\left(\delta_{ij} - \left\langle \frac{\kappa_j q_j(x) e^{f_j}}{\sum_k \kappa_k q_k(x) e^{f_k}} \right\rangle_i\right),$$

$$\beta_{ij} = \kappa_j \left\langle \frac{w_i}{\sum_k \kappa_k q_k(x)/z_k} \right\rangle_j,$$

and $H^{\#}$ is the group inverse of H.

In (45), we have used the group inverse, a type of matrix pseudoinverse. A numerical recipe for estimating the group inverse can be found in Ref. 31.

A proof of 4.1 is given in the supplementary material; here, we give the result and a quick informal sketch of the proof. The core concept is that, since both $\bar{\nu}$ and $\bar{F}(\nu)$ converge to ν and $F(\nu) = 0$ with an increasing number of samples, with enough samples, we only need to consider small deviations from ν and $F(\nu)$. In this case, we can employ a Taylor expansion of F around ν and truncate to obtain a linear relationship between their deviations (the matrix Γ). Consequently, for small deviations, $\bar{\nu}$ converges to a linear function of a Gaussian random variable. We can, therefore, scale the asymptotic covariance matrix of $\bar{F}(\nu)$ by Γ to get the asymptotic covariance of $\bar{\nu}$.

For many applications, the asymptotic variance in (45) can be further simplified by observing that the structure of $\Xi(y)$ depends on precisely how the states are sampled. We are interested primarily in two particular cases: (1) The X_t^j are independent Markov chains, and the sample fractions κ_j may differ but do not vary with N. (2) The sample fractions $\kappa_j = 1/L$ are equal, and $(X_t^1, \ldots, X_t^L)$ is a Markov process. The first case covers umbrella sampling or alchemical calculations performed without replica exchange. The second case covers parallel tempering and replica-exchange umbrella sampling.

In the first case, since the processes sampling the different states are independent, all off-diagonal blocks of $\Xi(y)$ are zero. The diagonal blocks can be expressed as

$$\Xi_{ij}^{ll}(y) = \frac{1}{\kappa_l}\left(\mathrm{cov}\left\{\xi_i(X_t^l, y), \xi_j(X_t^l, y)\right\} + 2\sum_{k=1}^{\infty} \mathrm{cov}\left\{\xi_i(X_t^l, y), \xi_j(X_{t+k}^l, y)\right\}\right), \tag{47}$$

where, here, we assume that the process X_t^l is stationary as in (12). The factor of $1/\kappa_l$ arises since in (42), we scale by $\sqrt{N} = \sqrt{N_l/\kappa_l}$ instead of $\sqrt{N_l}$.

In the second case, the processes sampling the states are correlated, so off-diagonal blocks may be nonzero. In this case, we have

$$\Xi_{ij}^{lm}(y) = L\Bigg(\mathrm{cov}\Big\{\xi_i(X_t^l, y), \xi_j(X_t^m, y)\Big\} \times 2\sum_{k=1}^{\infty} \mathrm{cov}\Big\{\xi_i(X_t^l, y), \xi_j(X_{t+k}^m, y)\Big\}\Bigg), \tag{48}$$

where, here, we assume that the joint process $(X_t^1, \ldots, X_t^L)$ is stationary. The factor of L arises since in (42), we scale by $\sqrt{N} = \sqrt{LN_l}$ instead of $\sqrt{N_l}$.

B. CLTs and the delta method

For most applications, practitioners are not interested in the values of $\bar{\nu}$ directly but instead wish to evaluate nonlinear combinations of these terms. To construct a CLT for these combinations, we employ the Delta method.

Lemma 4.2 (The Delta method; Proposition 6.2 in Bilodeau and Brenner[32]). Let θ_N be a sequence of random variables taking values in $\mathbb{R}^d$. Assume that a central limit theorem holds for θ_N with mean $\mu \in \mathbb{R}^d$ and an asymptotic covariance matrix $\Sigma \in R^{d\times d}$, i.e.,

$$\sqrt{N}(\theta_N - \mu) \xrightarrow{D} \mathbf{N}(0, \Sigma). \tag{49}$$

Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a function that is differentiable at μ. We then have the central limit theorem,

$$\sqrt{N}(\Phi(\theta_N) - \Phi(\mu)) \xrightarrow{D} \mathbf{N}\Big(0, \nabla\Phi(\mu)^T \Sigma \nabla\Phi(\mu)\Big), \tag{50}$$

for the sequence of random variables $\Phi(\theta_N)$.

If we have a CLT for a certain variable, we can use the Delta method to derive a CLT for any differentiable function of that variable. For our particular case, we set μ and θ_N to be ν and $\bar{\nu}$, respectively, and set Φ to be the function taking ν to our quantity of interest. For instance, to construct a CLT for an estimate of the free energy between two alchemical states $\Delta G_{\alpha\to\beta}$, as defined in (28), we set $\Phi(y) = y_L - y_1$, and $\nabla\Phi(\mu)$ is given by

$$\nabla\Phi(\mu) = (-1, 0, \ldots, 0, 1, 0, \ldots)^T. \tag{51}$$

Here, the non-zero entries correspond to the first and Lth elements in the vector. These are the only entries in μ that contribute to the free energy and correspond to states α and β, respectively. We then substitute into (50) to get the asymptotic variance of our estimate of $\Delta G_{\alpha\to\beta}$. Similarly, for the ensemble average in (30), we set $w_{L+1} = gq$ and $w_{L+2} = q$, and (30) is recovered by setting $\Phi(y) = y_{L+1}/y_{L+2}$. Then, $\nabla\Phi(\mu)$ is zero apart from the $(L+1)$th and $(L+2)$th entries, which are given by

$$\begin{aligned} \nabla\Phi(\mu)_{L+1} &= 1/\omega_{L+2}, \\ \nabla\Phi(\mu)_{L+2} &= -\omega_{L+1}/\omega_{L+2}^2, \end{aligned} \tag{52}$$

respectively. As a final example, we consider the construction of error estimates of free energy differences estimated using umbrella sampling. We set $w_{L+1} = \mathbb{1}_A q$ and $w_{L+2} = \mathbb{1}_B q$, and $\Phi(\mu) = -\log(w_{L+2}) + \log(w_{L+1})$. Then, $\nabla\Phi(\mu)$ is again zero apart from the $(L+1)$th and $(L+2)$th entries, which are

$$\begin{aligned} \nabla\Phi(\mu)_{L+1} &= -1/\omega_{L+1}, \\ \nabla\Phi(\mu)_{L+2} &= 1/\omega_{L+2}. \end{aligned} \tag{53}$$

Consequently, we can combine Lemma 4.2 with Theorem 4.1 to have a CLT for MBAR estimates.

Theorem 4.3. *Let $\mathscr{G}$ be an observable whose MBAR estimate $\tilde{\mathscr{G}}$ is constructed by applying a function $\Phi : \mathbb{R}^{L+M} \to \mathbb{R}$ to the vector $\bar{\nu}$ and assume that Φ is differentiable at ν. The estimate $\tilde{\mathscr{G}}$ then obeys*

$$\sqrt{N}\big(\tilde{\mathscr{G}} - \mathscr{G}\big) \xrightarrow{D} N(0, \mathscr{A}), \tag{54}$$

where the asymptotic covariance matrix $\mathscr{A}$ is given by

$$\mathscr{A} = \nabla\Phi^T(\nu)\Gamma A \Gamma^T \nabla\Phi(\nu). \tag{55}$$

Proof. The proof follows immediately by applying 4.2 to Theorem 4.1. □

C. Computationally estimating the asymptotic variance

In principle, one could directly estimate asymptotic variances for observables by individually estimating each of the matrices and vectors in (54). However, directly evaluating A would require first populating the covariance matrix Ξ, which would, in turn, require evaluating as many as $L^2(L+M)^2$ correlation functions. Consequently, we provide simplified formulas for evaluating the asymptotic variance of observables in the specific case where sampling is performed independently in every state. In Sec. II of the supplementary material, we give analogous formulas for schemes, such as parallel tempering and replica exchange umbrella sampling that sample all states jointly using a single Markov chain.

If each state is sampled independently, then Ξ^{lm} is zero for $l \neq m$, eliminating one of the sums in (44). In Sec. II of the supplementary material, we show that by moving the remaining sum to the outside and bringing the remaining terms inside the expectation, we can rewrite each term in the integrated covariance matrix in (54) as

$$\mathscr{A}_{ij} = \sum_{k=1}^{L} \mathrm{cov}\Big\{\chi_i(X_t^k), \chi_j(X_t^k)\Big\} + \sum_{k=1}^{L} 2\sum_{\tau=1}^{\infty} \mathrm{cov}\Big\{\chi_i(X_t^k), \chi_j(X_{t+\tau}^k)\Big\}, \tag{56}$$

where

$$\chi_j(x) = \sum_{i=1}^{L+M} \sqrt{\kappa_i}\xi_i(x, \nu)\Big(\Gamma^T\nabla\Phi\Big)_{ij}(\nu). \tag{57}$$

To construct an estimate of the asymptotic variance, we first replace ν in (57) with the MBAR estimate $\bar{\nu}$ from sampled data and then estimate the integrated autocovariance of the resulting trajectory. In this work, we employ the ACOR algorithm to estimate the integrated autocovariance.[33] Moreover, since each summand in (56) depends

only on the sampling in state i, we can interpret the integrated autocovariance of χ_i as accounting for how much state i contributes to the total error. A Python code implementing this algorithm for estimating asymptotic error can be found in the EMUS repository.[34]

V. APPLICATION OF THE ERROR ESTIMATOR

We demonstrate our error estimator on two test cases: an alchemical free energy calculation and an umbrella sampling calculation.

A. Alchemical calculation of the free energy of solvating methane in water

The solvation free energy of methane can be determined via an alchemical simulation process in which the interaction between methane and surrounding water molecules is introduced gradually. We interpolate between the two states using (27), setting H_α (the Hamiltonian of the first state) to the Hamiltonian where the methane molecule and the water do not interact, and H_β (the Hamiltonian of the second state) to the Hamiltonian where they interact fully. We then estimate the free energy difference between the two states using (28) and estimate the asymptotic variance as described in Sec. IV C.

We performed 20 independent alchemical simulations at 298 K using GROMACS version 2019.4,[35] the OPLS-AA force field,[36] and the TIP3P water model.[37] A total of 21 equidistant λ values from 0 to 1 (endpoints included) were chosen. Each state was equilibrated at constant volume and then at a constant pressure of 1 bar for 100 ps using the Parinello–Rahman barostat with a time constant of 1 ps. The state was then further sampled at constant pressure for 1 ns to generate 1000 data points. The P-LINCS algorithm was used to constrain bonds to hydrogen atoms.[38,39] In all simulations, a stochastic Langevin dynamics integrator with a time step of 2 fs and a time constant of 1 ps was used to maintain a constant temperature of 300 K. In Fig. 1, we plot the cumulative free energy change between states as well as the free energy difference between successive states.

The total asymptotic standard deviation in the solvation free energy is estimated to be 0.0221 ± 0.0007 kcal/mol using (56) with $\nabla\Phi$ given by (51). This is close to the standard deviation over all 20 replicate simulations, which we estimate to be 0.0250 kcal/mol.

Often, practitioners subsample the trajectory into a collection of uncorrelated samples. This runs the risk of introducing additional error to an MBAR estimate: if the subsampling frequency is chosen to be too large, useful data may be discarded. However, subsampling has the advantage that it reduces the dataset's size. Moreover, it allows error estimates to be constructed using the asymptotic error estimator from Refs. 10 and 15, which requires statistically independent samples in each state. When subsampling is performed well, this error estimator gives results of comparable quality to the estimator presented in this work. For example, subsampling each trajectory with a period equal to the autocorrelation time of the state's potential energy divided by k_BT, the error estimator given in Refs. 10 and 15 gives an asymptotic standard deviation of 0.0212 ± 0.0006 kcal/mol. We note, however, that ensuring the subsampling is performed well is crucial. As an extreme example, when no subsampling is performed, this approach estimates the asymptotic standard deviation to be 0.003 91 ± 0.000 02 kcal/mol, nearly an order of magnitude too low.

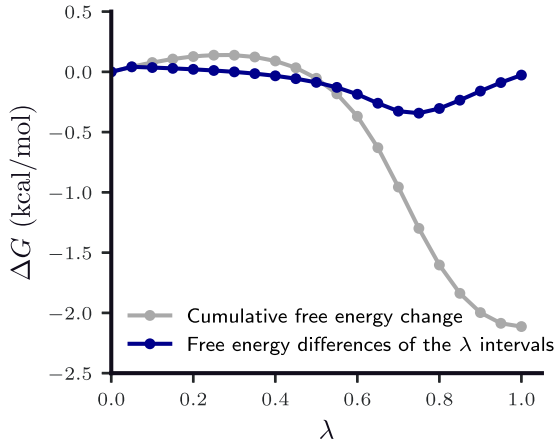

FIG. 1. The free energy of solvating methane in water as computed from alchemical simulations. The free energy difference was estimated using the MBAR estimator (28). The blue line indicates free energy differences between neighboring states, and the gray line is the cumulative free energy changes. The total free energy of solvation, i.e., the cumulative ΔG at $\lambda = 1$, is estimated to be 2.13 kcal/mol.

In Fig. 2(a), we give the error contributions from all states. As the error contributions for different states can vary by more than two orders of magnitude, we depict them on a logarithmic scale. Moreover, comparing with Fig. 1, we see that the error contributions correlate with the magnitudes of the free energy differences between neighboring states. The fact that different states' error contributions differed by orders of magnitudes suggests that the error in alchemical free energy simulations may be dominated by a few states. However, authoritatively establishing this hypothesis would require further investigation over many alchemical simulations in a variety of settings, with schedules typical of practical applications.

To further examine the source of the errors in our simulation, we attempt to disentangle the effect of the dynamics used to sample the state from the choice of λ values. Recalling the definition of the integrated autocorrelation time in (13) and combining it with (56), we can write the integrated autocovariance of each state as a product of the integrated autocorrelation time and a sampler-independent factor, namely, $\mathrm{var}\{\chi\}$ [c.f. (57)]. In Fig. 2(b), we plot both of the error components on a logarithmic scale: the logarithm of a state's total contribution is a sum of the two curves. Our results show that both the sampler-independent component of the error and the integrated autocorrelation time are important for the total error.

The fact that a small number of alchemical states dominate the contributions to the error suggests that it may be possible to use the error estimates provided here to help tune simulation parameters to achieve dramatic reductions in the error of MBAR estimates. Indeed, finding better simulation parameters and design principles for free energy methods has been the subject of considerable prior work,[40–46] which has shown that better allocation of computational resources can substantially reduce the error in alchemical free energy simulations. However, most of this work has focused purely on the static properties of the states, omitting the dynamic effects that arise from correlation within the states. For instance, Refs. 43 and 44 used information-geometric distances between states. Our results suggest

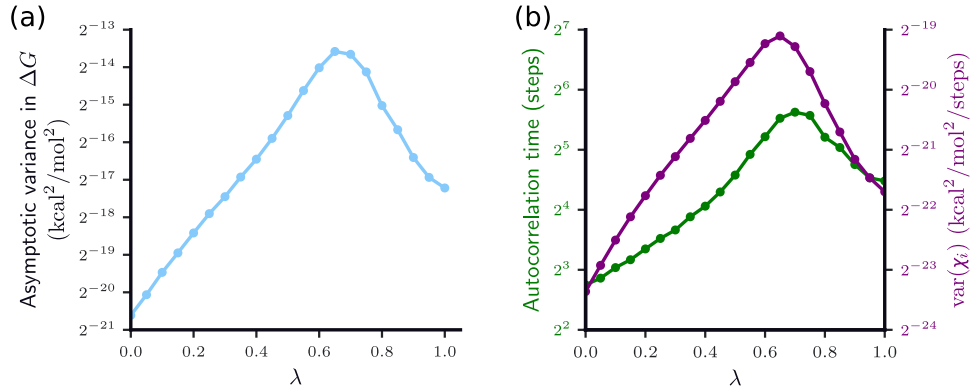

FIG. 2. Analysis of the estimated error in the free energy of solvation for methane. (a) State contributions to the asymptotic variance of the free energy difference between the initial ($\lambda = 0$) and final states ($\lambda = 1$). The contributions were estimated according to (56), with $\nabla\Phi$ given by (51). (b) Decomposition of the error contributions into the integrated autocorrelation times and the variance of χ_i, the trajectory encoding the state's contribution to the error. Each state's total contribution to the error is the product of these two quantities. For ease of comparison, quantities are plotted on a logarithmic scale.

that to fully capture all sources of error, such approaches must also take into account kinetic effects from the specific choice of sampler used. This corroborates previous work,[42,45] which attempted to optimize the choice of alchemical states by minimizing objective functions that measured the error that would arise from sampling states IID. In both works, it was observed that alchemical states that minimized these objective functions might not be optimal, in practice, due to the resulting states having exceedingly large correlation times. We hope that our analysis, which more fully incorporates the effect of time correlation, will help overcome these difficulties.

B. Umbrella sampling simulation of the alanine dipeptide

We also applied the error estimator to a two-dimensional umbrella sampling simulation of the alanine dipeptide (*N*-acetyl-alanyl-*N*′-methylamide) in vacuum. We performed ten independent umbrella sampling calculations for the free energy as a function of the ϕ and ψ dihedral angles. Simulations were run at 300 K using OpenMM version 7.6[47] with harmonic restraints applied to ϕ and ψ. The molecule was represented by the AMBER force field with bonds to hydrogen atoms constrained by the SHAKE algorithm.[48] The force constant for the harmonic restraints was 0.007 605 kcal mol^{-1} degree^{-2}, which corresponds to a Gaussian bias function with a standard deviation of ~9° in the absence of the molecular potential. We partitioned each dihedral angle into 30 intervals and placed the centers of the harmonic restraints at the centers of the cells of the resulting 30 × 30 grid; the resulting grid ranged from $(-174°, -174°)$ to $(174°, 174°)$. Each state was sampled independently using the BAOAB integrator for Langevin dynamics[49,50] with a time step of 1 fs and a time constant of 0.1 ps. Each state was equilibrated for 10 ps and then sampled for 100 ps, with ϕ and ψ values output every 0.1 ps.

In Fig. 3, we show the free energy surface over the ϕ and ψ dihedral angles. To construct the free energy surface, we constructed an evenly spaced grid of 50 × 50 histogram bins and evaluated MBAR estimates of (5). We then estimate the free energy difference between the C_{7ax} and C_{7eq} basins. We define the C_{7ax} basin as the region in the $\phi\psi$-space enclosed by a circle of radius 10° centered at $(65°, -40°)$ and the C_{7eq} basin as the space enclosed by the circle of radius 10° centered at $(-75°, 50°)$. The free energy between the basins can then be obtained by estimating the logarithm of the ratio of averages of two indicator functions as in (3). Note that the precise definition of the sets can affect the value of the free energy and the corresponding asymptotic variance of its estimate.

In Table I, we give our estimate of the error in the free energy difference evaluated as the asymptotic standard deviation (square root of the asymptotic error), as evaluated by (56), with $\nabla\Phi$ given by (53). We compare the estimated error with an estimate of the standard deviation calculated over 10 identical replicates. Our error estimator gives results that are about 2.6 times smaller than the

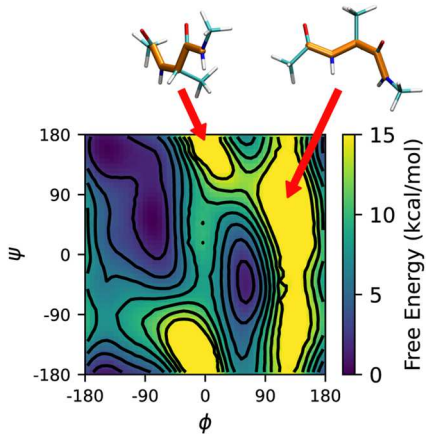

FIG. 3. Free energy surface over the ϕ and ψ dihedral angles (measured in degrees) of the alanine dipeptide. The scale bar indicates free energy values in kcal/mol, and the contour spacing is 2 kcal/mol. Structures are representative of configurations at free energy maxima that are sterically strained and complicated sampling.

TABLE I. Asymptotic standard deviation of the free energy between the C_{7ax} and C_{7eq} basins of the alanine dipeptide compared with the standard deviation over 10 independent simulations. The top row uses all states in the dataset, while the bottom row uses only states whose relative free energies are within 20 k_BT of the lowest free energy state.

States used	Estimated asymptotic SD (kcal/mol)	SD over replicates (kcal/mol)
All	0.049 ± 0.007	0.126
Low-FE	0.054 ± 0.007	0.054

standard deviation of the free energy estimate calculated over statistical replicates. While this constitutes reasonable agreement, we show below that the discrepancy comes from a few states, which, if removed, improve the quality of a free energy estimate and its uncertainty.

In Figs. 4(a) and 4(b), we give each state's contribution to the total asymptotic variance on linear and logarithmic scales. To further understand the source of each state's contribution, in Figs. 4(c) and 4(d), we depict the autocorrelation time and the variance of χ for each state. We see that states with large contributions to the error estimator can be divided into two categories. The first category includes states that are located on the transition pathway connecting the two C_{7ax} and C_{7eq} basins. While these states do not have large autocorrelation times, they have comparatively large values for the variance of χ. Our interpretation is that while these states are important for getting a good estimate of the free energy, they are not intrinsically difficult to sample. The second category includes states located in high free energy regions, such as the free energy peaks near $(0°, 180°)$ and $(135°, 90°)$. While these states have large autocorrelation times, they have small values for the variance of χ. Indeed, we were surprised to find that these high-energy states contributed so much to the total error despite being far from the minimum energy path connecting the metastable basins. Our belief is that because these states are bifurcated by a peak in free energy, converging their statistics requires observing slow barrier-crossing events. This makes them act as "bad apples" that spoil the accuracy of the scheme. For instance, crossing the peak near $(0°, 180°)$ requires moving two carbonyl oxygens that are undergoing a steric clash past each other. Similarly, moving across the peak $(120°, 50°)$ requires moving two oxygens through a clash with the methyl group. In Fig. 3, we show representative configurations for these states.

Our analysis suggests that we should be able to remove these high free-energy states that give large contributions to the error estimate without reducing the quality of the estimated free energy. To validate this hypothesis, we remove all states from our dataset where the state's unitless free energy is at least 20 k_BT larger than the lowest free energy state. We repeat this process for each statistical replicate of the umbrella sampling calculation; the precise locations of the removed windows are shown in Sec. III of the supplementary material. For each replicate, we then recalculate the estimate of the free energy as well as the corresponding error estimate. The standard deviation of our free energy estimate, as well as the average estimated asymptotic standard deviation, is given in Table I in the row titled "Low-FE." Despite using less data, our curated dataset has a lower standard deviation by more than a factor of two. Additionally, our estimate of the asymptotic standard deviation now agrees with the empirically calculated standard deviation. This suggests that the previously observed discrepancy between the predicted asymptotic standard deviation and the empirically observed standard deviation is due to the difficulty in estimating the autocorrelation time for these slowly decorrelating states.

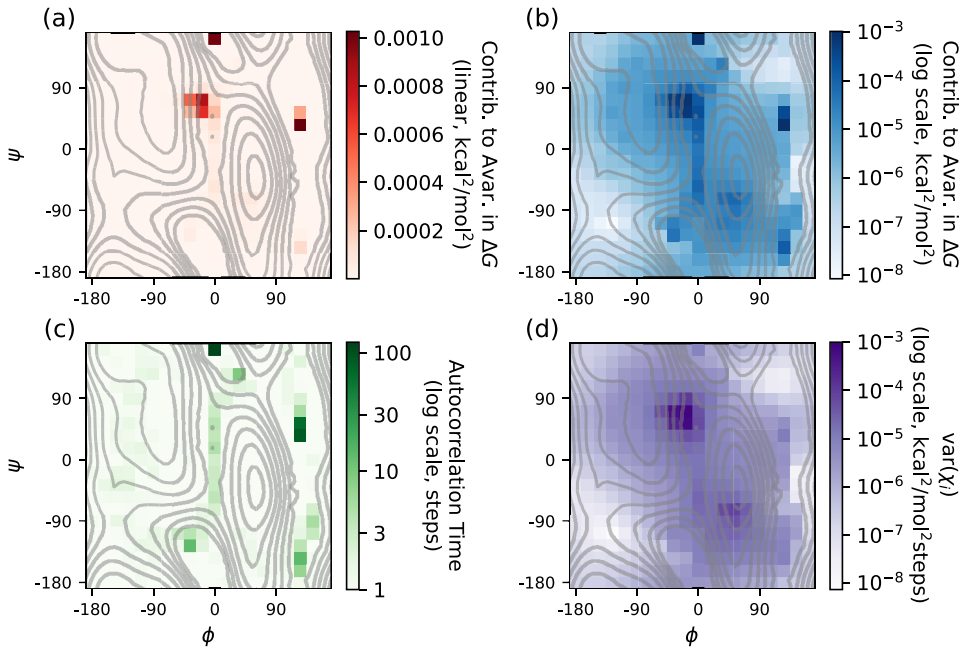

FIG. 4. Analysis of the estimated error in the free energy difference between the C_{7ax} and C_{7eq} states of the alanine dipeptide. For each state in the 20×20 grid of states, we give (a) the state contributions to the asymptotic variance of the free energy calculated using all states, depicted on a linear scale, (b) the state contributions to the asymptotic variance of the free energy calculated using all states, depicted on a logarithmic scale, (c) the integrated autocorrelation times of χ_i for the free energy estimate using every state, and (d) the variance of χ_i for the free energy estimate using every state. In each plot, we give the underlying free energy surface on the ϕ and ψ dihedral angles in light gray for reference.

To conclude our analysis, we compare the error estimator presented in our work with the error estimator presented in Refs. 10 and 15. As this error estimator requires statistically independent samples, we subsample each trajectory by either the autocorrelation time of the ϕ or the ψ dihedral angle, whichever is larger. Applying this error estimator gives an asymptotic standard deviation of 0.049 ± 0.006 when using all states, and 0.054 ± 0.007 when using only the low free energy states. In contrast to the alchemical example, we see a much smaller difference between the two error estimators. However, this is expected: for the majority of states considered here, the autocorrelation time is already close to the period at which the data are recorded. Consequently, subsampling the trajectory has a much smaller effect than in the alchemical example.

We leave a systematic procedure for using the error estimator to refine the sampling for future work. However, our preliminary results, along with our earlier work on the Eigenvector Method for Umbrella Sampling and its application,[16,17,51] demonstrate that our error estimator has the potential to improve the error of umbrella sampling and other multistate methods.

VI. CONCLUSIONS

We derive a central limit theorem for estimates of both the normalization constants and function averages of the MBAR estimator. In contrast to previous work, our treatment directly accounts for the effect of correlation in sampled trajectories, further justifying the use of MBAR when samples are not drawn independently. The central limit theorem allows us to devise a computational procedure for estimating the asymptotic error for arbitrary observables calculated through MBAR. In particular, it allows us to estimate the asymptotic error in free energy calculations. Notably, if states are sampled independently, the analytical expression of the total asymptotic error takes the form of a sum of contributions from all states. This enables us to trace how the errors in sampling each state contribute to the total error. For both of the examples we study, we observe that the autocorrelation times of the states contribute strongly to the total asymptotic error. These results highlight the importance of error analysis that explicitly accounts for correlation in a sampled trajectory when attempting to tune free energy calculations.

We demonstrate the error estimator for an alchemical calculation of the solvation free energy of methane and a two-dimensional umbrella sampling calculation of the free energy of isomerization of the alanine dipeptide. In both cases, the asymptotic error estimates agree reasonably well with the standard deviation over all replicates, giving results that are equally good or better than competing approaches. Moreover, the fact that we can decompose the error estimate into contributions from each state allows us to probe which states contribute most to the overall error. Analyzing these contributions for our umbrella sampling calculation, we observe that a substantial fraction of the error comes from high free-energy states that decorrelate slowly. We find that removing these states gives a more precise error free energy estimate, despite using less data overall. This initial analysis suggests that our error estimator could be the basis for adaptive strategies for tuning free energy simulations. We hope to investigate adaptive sampling strategies based on our error estimates in future work.

SUPPLEMENTARY MATERIAL

In the supplementary material, we give a detailed derivation of the asymptotic variance for the MBAR estimator and discuss how to estimate it from data. Additionally, we depict the precise locations of the states pruned during the umbrella sampling calculation performed in Sec. V B.

ACKNOWLEDGMENTS

We wish to thank Jonathan Weare for advice and for pointing out a simplification in our error analyses, and Adam Antoszewski for helpful discussions. This work was supported by the National Institutes of Health under Award No. R35 GM136381 and the National Science Foundation under Award Nos. DMS-2054306 and DMS-2012207. The Flatiron Institute is a division of the Simons Foundation.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Xiang Sherry Li: Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Brian Van Koten**: Conceptualization (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Aaron R. Dinner**: Conceptualization (equal); Funding acquisition (lead); Methodology (equal); Project administration (equal); Writing – review & editing (equal). **Erik H. Thiede**: Conceptualization (equal); Methodology (equal); Software (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in Error in MBAR Data, at https://drive.google.com/drive/u/1/folders/1BpeSdypdhxngLY9ANHV231AL1zyJ5Mmm.

REFERENCES

[1] C. Chipot and A. Pohorille, *Free Energy Calculations*, Springer Series in Chemical Physics (Springer, 2007), Vol. 86.
[2] T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations* (Imperial College Press, London, 2010).
[3] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," J. Comput. Phys. **23**, 187 (1977).
[4] C. Pangali, M. Rao, and B. J. Berne, "A Monte Carlo simulation of the hydrophobic interaction," J. Chem. Phys. **71**, 2975–2981 (1979).
[5] R. H. Swendsen and J.-S. Wang, "Replica Monte Carlo simulation of spin-glasses," Phys. Rev. Lett. **57**, 2607 (1986).
[6] C. J. Geyer, "Markov chain Monte Carlo maximum likelihood," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (American Statistical Association, 1991).
[7] B. L. Tembre and J. A. Mc Cammon, "Ligand-receptor interactions," Comput. Chem. **8**, 281–283 (1984).

[8]D. Frenkel and B. Smit, *Understanding Molecular Simulation from Algorithms to Applications*, *Computational Science* (Academic Press, San Diego, CA, 2002), p. 638.
[9]J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande, "Alchemical free energy methods for drug discovery: Progress and challenges," Curr. Opin. Struct. Biol. **21**, 150–160 (2011).
[10]M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105 (2008).
[11]Y. Vardi, "Empirical distributions in selection bias models," Ann. Stat. **13**, 178–203 (1985).
[12]C. Bartels, "Analyzing biased Monte Carlo and molecular dynamics simulations," Chem. Phys. Lett. **331**, 446–454 (2000).
[13]S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," J. Comput. Chem. **13**, 1011–1021 (1992).
[14]M. Souaille and B. Roux, "Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations," Comput. Phys. Commun. **135**, 40–57 (2001).
[15]A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan, "A theory of statistical models for Monte Carlo integration," J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **65**, 585–604 (2003).
[16]E. H. Thiede, B. Van Koten, J. Weare, and A. R. Dinner, "Eigenvector method for umbrella sampling enables error analysis," J. Chem. Phys. **145**, 084115 (2016).
[17]A. R. Dinner, E. H. Thiede, B. V. Koten, and J. Weare, "Stratification as a general variance reduction method for Markov chain Monte Carlo," SIAM/ASA J. Uncertainty Quantif. **8**, 1139–1188 (2020).
[18]K. S. Chan, "On the central limit theorem for an ergodic Markov chain," Stochastic Processes Their Appl. **47**, 113–117 (1993).
[19]L. Tierney, "Markov chains for exploring posterior distributions," Ann. Stat. **22**, 1701–1728 (1994).
[20]C. Geyer and K. Chan, "Discussion of the paper by Tierney," Ann. Stat. **22**, 1747–1758 (1994).
[21]G. L. Jones, "On the Markov chain central limit theorem," Probability Surveys **1**, 299–320 (2004).
[22]C. J. Geyer, "Estimating normalizing constants and reweighting mixtures," Technical Report 568, University of Minnesota, 1994.
[23]M. R. Shirts, "Reweighting from the mixture distribution as a better way to describe the multistate Bennett acceptance ratio," arXiv:1704.00891 (2017).
[24]X.-L. Meng and W. H. Wong, "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration," Stat. Sin. **6**, 831–860 (1996).
[25]R. D. Gill, Y. Vardi, and J. A. Wellner, "Large sample theory of empirical distributions in biased sampling models," Ann. Stat. **16**, 1069–1112 (1988).
[26]T. Simonson, "Free energy of particle insertion: An exact analysis of the origin singularity for simple liquids," Mol. Phys. **80**, 441–447 (1993).
[27]T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren, "Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations," Chem. Phys. Lett. **222**, 529–539 (1994).
[28]T. Steinbrecher, D. L. Mobley, and D. A. Case, "Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations," J. Chem. Phys. **127**, 214108 (2007).
[29]D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "emcee: The MCMC hammer," Publ. Astron. Soc. Pac. **125**, 306 (2013).
[30]S. Meyn, R. L. Tweedie, and P. W. Glynn, *Markov Chains and Stochastic Stability* (Cambridge University Press, 2009), p. 624.
[31]G. H. Golub and C. D. Meyer, Jr., "Using the QR factorization and group inversion to compute, differentiate, and estimate the sensitivity of stationary probabilities for Markov chains," SIAM J. Algebraic Discrete Methods **7**, 273–281 (1986).
[32]M. Bilodeau and D. Brenner, *Theory of Multivariate Statistics* (Springer Science & Business Media, 2008).
[33]J. Goodman and D. Foreman-Mackey, ACOR 1.1.1, https://pypi.org/project/acor/1.1.1, 2014.
[34]E. H. Thiede, EMUS, https://github.Com/ehthiede/EMUS, 2022.
[35]E. Lindahl, M. J. Abraham, B. Hess, and D. van der Spoel, GROMACS 2019.4 Source Code, 2019.
[36]M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen, "Improved peptide and protein torsional energetics with the OPLS-AA force field," J. Chem. Theory Comput. **11**, 3499–3509 (2015).
[37]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys. **79**, 926–935 (1983).
[38]B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem. **18**, 1463–1472 (1997).
[39]B. Hess, "P- LINCS: A parallel linear constraint solver for molecular simulation," J. Chem. Theory Comput. **4**, 116–122 (2008).
[40]F. A. Escobedo, "Optimized expanded ensembles for simulations involving molecular insertions and deletions. II. Open systems," J. Chem. Phys. **127**, 174104 (2007).
[41]F. A. Escobedo and F. J. Martínez-Veracoechea, "Optimized expanded ensembles for simulations involving molecular insertions and deletions. I. Closed systems," J. Chem. Phys. **127**, 174103 (2007).
[42]F. J. Martínez-Veracoechea and F. A. Escobedo, "Variance minimization of free energy estimates from optimized expanded ensembles," J. Phys. Chem. B **112**, 8120–8128 (2008).
[43]D. K. Shenfeld, H. Xu, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Minimizing thermodynamic length to select intermediate states for free-energy calculations and replica-exchange simulations," Phys. Rev. E **80**, 046705 (2009).
[44]T. T. Pham and M. R. Shirts, "Identifying low variance pathways for free energy calculations of molecular transformations in solution phase," J. Chem. Phys. **135**, 034114 (2011).
[45]T. T. Pham and M. R. Shirts, "Optimal pairwise and non-pairwise alchemical pathways for free energy calculations of molecular transformation in solution phase," J. Chem. Phys. **136**, 124120 (2012).
[46]C. Predescu, M. Snarski, A. Robinson-Mosher, D. Sritharan, T. Szalay, and D. E. Shaw, "Times square sampling: An adaptive algorithm for free energy estimation," arXiv:2112.05109 (2021).
[47]P. 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 *et al.*, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," PLoS Comput. Biol. **13**, e1005659 (2017).
[48]J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes," J. Comput. Phys. **23**, 327–341 (1977).
[49]B. Leimkuhler and C. Matthews, "Rational construction of stochastic numerical methods for molecular sampling," Appl. Math. Res. Express **2013**, 34–56.
[50]B. Leimkuhler and C. Matthews, "Robust and efficient configurational molecular sampling via Langevin dynamics," J. Chem. Phys. **138**, 174102 (2013).
[51]A. Antoszewski, C.-J. Feng, B. P. Vani, E. H. Thiede, L. Hong, J. Weare, A. Tokmakoff, and A. R. Dinner, "Insulin dissociates by diverse mechanisms of coupled unfolding and unbinding," J. Phys. Chem. B **124**, 5571–5587 (2020).