Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration

Michael R. Shirts and Vijay S. Pande
Department of Chemistry, Stanford University, Stanford, California 94305-5080

(Received 23 November 2004; accepted 25 January 2005; published online 14 April 2005)

Recent work has demonstrated the Bennett acceptance ratio method is the best asymptotically unbiased method for determining the equilibrium free energy between two end states given work distributions collected from either equilibrium and nonequilibrium data. However, it is still not clear what the practical advantage of this acceptance ratio method is over other common methods in atomistic simulations. In this study, we first review theoretical estimates of the bias and variance of exponential averaging (EXP), thermodynamic integration (TI), and the Bennett acceptance ratio (BAR). In the process, we present a new simple scheme for computing the variance and bias of many estimators, and demonstrate the connections between BAR and the weighted histogram analysis method. Next, a series of analytically solvable toy problems is examined to shed more light on the relative performance in terms of the bias and efficiency of these three methods. Interestingly, it is impossible to conclusively identify a "best" method for calculating the free energy, as each of the three methods performs more efficiently than the others in at least one situation examined in these toy problems. Finally, sample problems of the insertion/deletion of both a Lennard-Jones particle and a much larger molecule in TIP3P water are examined by these three methods. In all tests of atomistic systems, free energies obtained with BAR have significantly lower bias and smaller variance than when using EXP or TI, especially when the overlap in phase space between end states is small. For example, BAR can extract as much information from multiple fast, far-from-equilibrium simulations as from fewer simulations near equilibrium, which EXP cannot. Although TI and sometimes even EXP can be somewhat more efficient in idealized toy problems, in the realistic atomistic situations tested in this paper, BAR is significantly more efficient than all other methods. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1873592]

I. INTRODUCTION

Finding the free energy difference between different states of a physical system is of great general interest in many scientific fields, from drug design,[1] to basic statistics,[2] to even nonperturbative quantum chromodynamics.[3] It is of interest to the experimental community as well as the theoretical and computational communities.[4] Recently, there has been increased interest in determining the uncertainty and bias in any attempt to extract free energies from a suitable sets of data.[2,5–15] As these calculations or experiments are in general expensive, requiring significantly more effort than simply measuring a single ensemble-averaged observable, there is also great interest in maximizing the efficiency of such measurements.

There are essentially two disjoint problems that must be solved in order to calculate free energy differences precisely and accurately. First, we must generate a number n of statistically uncorrelated measurements of the system, the particular measurement depending on the type of free energy estimation performed. Second, we must extract a free energy estimate from these n measurements, ideally along with estimates for the statistical bias and variances of our estimate. The generation of accurate estimates for statistical uncertainty and statistical bias is vital for any such free energy estimate to be of use. This first problem reduces to sampling the system in the proper (usually Boltzmann) manner, and will not be discussed further in this study. We will assume that we are already in possession of a set of n uncorrelated measurements of the proper observable for our method, and deal only with the statistical issues related to the extraction of free energy estimates from these measurements.

In many cases, the two systems have so little overlap in phase space that useful estimation of the free energy becomes impossible with data solely from the end states. It is simple, however, to define a series of intermediate states, determine the energies between these intermediate states, and sum these intermediate energies to obtain an overall free energy. If the data is being collected by simulation, the intermediate states need not be physically realizable, only amenable to simulation. Since the free energy is a state variable, only the end states matter. However, the choice of a pathway of intermediate states can also greatly influence the precision and accuracy of the free energy obtained. Although we touch on the question of pathway choice to some extent, this study is not intended as a full exploration of the choice of maximal efficient pathway for arbitrary problems for either equilibrium and nonequilibrium free energy calculations. We instead examine the efficiency assuming a suitable pathway has been found.

0021-9606/2005/122(14)/144107/16/$22.50

There are several commonly used methods for finding the free energy of a physical change in a system. In “multicanonical” thermodynamic integration,[16] now usually referred to simply as thermodynamic integration (TI), a parametrization (in some variable λ) from the initial state to the final state is introduced. The equilibrium ensemble average of the derivative of the Hamiltonian H with respect to this parametrization in λ is then computed at a number of points along this pathway, and integrated numerically to obtain the free energy difference.

“Slow growth,” in which the numeric integral of $\langle dH/d\lambda\rangle$ over different equilibrium states is approximated by a single integral in which λ goes from 0 to 1 in a time dependent manner, has been shown to have high intrinsic biases and generally yields very poor results,[16,17] although the “free energies” obtained from these simulations can be treated as measurements from nonequilibrium work distributions and exponentially averaged to obtain correct free energies.[6] This can be done by means of the recently discovered Jarzynski equality, which specifies that if the nonequilibrium work W to take a number of systems in thermal equilibrium from an initial Hamiltonian to a different final Hamiltonian is averaged over the entire Boltzmann-weighted initial ensemble, the free energy (or equilibrium work) between the two states is given by $\Delta F = -\beta^{-1}\ln\langle\exp(-\beta W)\rangle$.[18] This remarkable result is independent of the path and depends only on the thermal equilibrium of the initial state, and the process moving the states from one Hamiltonian to the other obeying detailed balance. Using this relationship, other methods such as “fast growth,”[5,6] and thermodynamic perturbation theory (TPT) or free energy perturbation (FEP) (Ref. 19) reduce to finding the free energy difference between two equilibrium states given a distribution of nonequilibrium work differences between the states by exponential averaging.[18,20] FEP was originally developed in the context of estimating free energy differences by exponentially averaging potential energy differences between a reference state sampled at equilibrium and a target state.[19] However, this potential energy difference is simply equivalent to the work of an infinitely fast adiabatic transition between the two states, and thus can be interpreted under Jarzynski’s relationship as well. FEP and TPT are sometimes used to describe all methods of finding differences between systems with “perturbed” Hamiltonians, though most commonly they refer to finding free energies through exponential averaging.

Recent work has shown that a neglected method, Bennett’s acceptance ratio method,[21] is rigorously always more efficient than exponential averaging in computing the free energy given a set of work values in both the forward and reverse directions.[13,22] Working from different principles, an alternate minimum variance algorithm called the weighted histogram analysis method (WHAM) has been developed.[23,24] WHAM is usually used to compute potentials of mean force or other thermodynamic observables along a reaction coordinate from a histogram of intermediate simulations. However, in order to compute these observables, the free energies between the states must first be estimated. The expression for the free energies in WHAM reduces to the Bennett acceptance ratio method in the case of determining the free energy between two states. However, despite the theoretical advantages, many researchers do not use either of these methods in computation of free energies for chemical processes, most likely not because they are not aware of the existence of these methods, but because they do not realize the advantages of these methods in practice.

In this paper, we will compare both theoretical and experimental variances and biases of exponential averaging, the Bennett acceptance ratio and thermodynamic integration. We will apply these methods to examine their variances and biases in practice, first to a series of illustrative toy models, many of which yield analytical results, and more importantly in a variety of computations of free energies of sample molecular systems of types that may be of practical use for chemical and biochemical experiments.

In doing so, we will introduce some new derivations of these variances that are significantly simpler than previous derivations, and illustrate connections to other important methods, such as the weighted histogram analysis method. We will sometimes use the abbreviations EXP for exponential averaging, TI for thermodynamic integration, BAR for the Bennett acceptance ratio, and WHAM for the weighted histogram analysis method.

II. COMPARISON OF THEORETICAL VARIANCE AND BIAS ESTIMATES

A. Underlying theory of free energy estimates

Assume there are two states defined by energy functions on a phase space, $U_A(\vec{q})$ and $U_B(\vec{q})$. Let ΔF be the free energy between the states, defined as the log of the ratio of the partition functions associated with $U_A(\vec{q})$ and $U_B(\vec{q})$. We can associate a work with the process of changing energy functions from U_A to U_B or visa versa, while the system is maintained in temperature equilibrium with the surroundings. By sampling initial conditions from equilibrium, we obtain a distribution in either direction of such work values. For infinitely fast switching, these distributions are simply of $\pm\Delta U = \pm(U_B - U_A)$ canonically sampled from the initial state.

It has long been known that the free energy difference between two states can be computed by taking the exponential average of the potential energy differences,[19] where the exponential average of a set of data, $X=\{x_i,\ldots,x_n\}$, is defined as $-(1/\beta)\ln\langle\exp(-\beta X)\rangle$, where $\beta=1/kT$. Jarzynski demonstrated that distribution of nonequilibrium work values over the canonical ensemble of stating states also yields an equilibrium free energy,[18] and indeed that the equilibrium exponential average is a special case in the limit of infinitely fast changes.

However, the exponential average over a distribution is a statistic that is both inherently noisy and biased, even if the spread of the data is only moderately larger than kT. Exponential averaging depends a great deal on the behavior at the tails of the distribution, which, by definition, are not well sampled, and the results of exponential averaging will therefore will have both high variance and bias. Previous studies have explored and demonstrated the poor behavior of exponential averaging for small sample sizes.[7–13]

Bennett showed that the value of ΔF which satisfied the equation

$$\sum_{i=1}^{n_F} \frac{1}{1+\exp(\beta(M+W_i-\Delta F))} - \sum_{j=1}^{n_R} \frac{1}{1+\exp(-\beta(M+W_j-\Delta F))} = 0, \qquad (1)$$

where $M=kT\ln n_f/n_r$, with n_f and n_r being the number of values from the forward and reverse distributions of work, respectively, minimized the variance in the estimation of the free energy among free energy estimates of the form

$$\exp(-\beta\Delta F) = \frac{\langle f(W)\rangle_F}{\langle f(-W)\exp(-\beta W)\rangle_R}, \qquad (2)$$

where $f(W)$ is an arbitrary function and the averages are over the two end states. The left side of Eq. (1) is a monotonically decreasing function for all ΔF, and is unbounded for both positive and negative ΔF, so we are guaranteed that we have one unique root for the free energy.

The overall variance of this estimate can be written as

$$\frac{1}{\beta^2 n_{tot}}\left\{\left[\left\langle \frac{1}{2+2\cosh(\beta(M+W_i-\Delta F))}\right\rangle\right]^{-1} - \left(\frac{n_{tot}}{n_f}+\frac{n_{tot}}{n_r}\right)\right\}, \qquad (3)$$

where the average in the above equations is over all work measurements, both forward and reverse.[22]

We have shown[22] that the Bennett acceptance ratio can be understood as the maximum likelihood estimator of the free energy given a set of forward and reverse nonequilibrium work measurements, starting from the basic relation[20]

$$\ln\left(\frac{P_F(W)}{P_R(-W)}\right) = \beta(W-\Delta F), \qquad (4)$$

where $P_F(W)$ and $P_R(W)$ are probability distributions for the work of nonequilibrium processes from the two states in opposing directions, arbitrarily labeled as F (“forward”) and R (“reverse”). It is therefore the minimum variance estimator among all asymptotically unbiased estimators, i.e., those estimators that become unbiased in the limit of an infinite number of observations. No other asymptotically unbiased estimator can extract a more precise value for ΔF information from a given set of data.

We also demonstrated that EXP in one direction is the maximum likelihood estimator in the limit of no information about the work distribution in other direction.[22] However, this lack of knowledge about the other direction greatly limits the precision of this method, and adding measurements from the opposing distribution usually yields improved results.[21,22] Various studies have explored and demonstrated the poor behavior of exponential averaging in the limit of small numbers of measurements,[7–13] and at least one has confirmed the superiority of the Bennett acceptance ratio to EXP in a limited number of physically relevant cases.[13]

It can also be shown that WHAM between only two states can be reduced to the formula for BAR. To observe this, we take the iterative equations for WHAM for a two state problem with two Hamiltonians

$$\begin{aligned} \exp(-\beta F_0) &= \left\langle \frac{1}{1+\frac{n_1}{n_0}\exp(\beta F_1-\beta(H_1-H_0))}\right\rangle_0 + \left\langle \frac{1}{1+\frac{n_1}{n_0}\exp(\beta F_1-\beta(H_1-H_0))}\right\rangle_1, \\ \exp(-\beta F_1) &= \left\langle \frac{1}{1+\frac{n_0}{n_1}\exp(\beta F_0-\beta(H_0-H_1))}\right\rangle_0 + \left\langle \frac{1}{1+\frac{n_0}{n_1}\exp\ (\beta F_0-\beta(H_0-H_1))}\right\rangle_1. \end{aligned} \qquad (5)$$

Eliminating F_0 and F_1 by replacement with $\Delta F=F_1-F_0$ and applying the relationship Eq. (4) yields the Bennett acceptance ratio formula Eq. (2). It appears that the general formulation for WHAM for more than two states is not applicable for general “fast-growth” simulations, as it does not take into account the special relationships between paired forward and reverse work distributions expressed in Eq. (4), but it is very possible that later research will discover a WHAM-like relationship between multiple work distributions, as the subject has not been sufficiently explored.

B. Limiting moment bias and variances

Although complicated methods have been presented for the asymptotic bias of EXP,[8] we provide a extremely simple method that gives the dominant terms for both the variance and bias of EXP, known in statistics as the limiting moment approach.[25] This method is trivially generalizable to a wide class of other estimators.

Suppose that we have a random variable X, sampled n times, such that $\langle X\rangle=\xi$ and var $(X)=\sigma^2$. Here, $\langle X\rangle$ is the expectation value over all possible values, not just the average over n values. We instead define $\overline{X}=n^{-1}\Sigma_{i=1}^{n}X_n$. $\langle\overline{X}\rangle$ is then the average of $\overline{X}$ over all possible selections of n samples. It has been shown[25] that if we have a function h, given certain weak constraints on h, the finite n bias in a function of the mean of X is

$$\langle h(\overline{X})\rangle - h(\xi) = \frac{\sigma^2}{2n}h''(\xi) + O(n^{-2}). \qquad (6)$$

In the case that $h=\ln(X)$, all the constraints on $h(X)$ required are satisfied, and we obtain

$$\langle \ln(\overline{X})\rangle - \ln(\xi) = -\frac{\sigma^2}{2n}\frac{1}{\xi^2} + O(n^{-2}). \qquad (7)$$

These equations can be obtained by examining the Taylor expansion of $h(\overline{X})$. In the case of exponential averaging, our random variable is not the work W, but instead the exponential of the work $\exp(-\beta W)$. Let us define $\hat{\sigma}^2$ as the variance of $X=\exp(-\beta(W-\Delta F))$ for the forward case and $X=\exp(\beta(W-\Delta F))$ for the reverse case. This is different from exponential averaging only by an additive constant, and thus does not affect the variance. With this choice, $h''(X)=(h'(X))^2=1$ when X is evaluated as above, at $W=\Delta F$. We can then write Eq. (7) as

$$\langle \ln(\overline{\exp(-\beta W)})\rangle - \beta\Delta F = \frac{\hat{\sigma}^2}{2n} + O(n^{-2}). \tag{8}$$

The term $\hat{\sigma}^2/2n$ is the large n limit of the bias in the exponential average. This is the same large n limit for the bias term derived by Zuckerman and Woolf.[8] We note that although the method presented here does not give information about higher powers in n^{-1} as does their previous derivation, it is significantly simpler and trivially generalizable. Almost always, statistics will be unreliable unless we have collected enough data so that the n^{-2} and lower terms are negligible, so the fact that this process gives only the n^{-1} term is not a serious limitation of the large n limit approximation.

The variance from the mean of this same function can be expressed as

$$\mathrm{var}(h(\overline{X})) = \frac{\sigma^2}{n}[h'(\xi)]^2 + O(n^{-2}). \tag{9}$$

In the case of $h=\ln(X)$, and $X=\exp(\beta(W-\Delta F))$, this reduces to

$$\mathrm{var}(h(\overline{X})) = \mathrm{var}(\ln(\overline{\exp(-\beta W)})) = \frac{\hat{\sigma}^2}{n} + O(n^{-2}). \tag{10}$$

We again obtain a result identical to that of Zuckerman and Woolf,[8] and we similarly see that, to order n, the bias of the exponential average is half that of the variance. We can observe that the reason for this particularly simple result is that $h''(\xi)=[h'(\xi)]^2$ for the case that $h(x)=\ln x$. The uncertainty, or standard error, is $O(n^{-1/2})$ while the bias is $O(n^{-1})$. This brings out an important point—in most cases of asymptotically unbiased estimators, the variance will be a much greater source of error in the calculations than the bias, as the uncertainty will be of lower order.

C. Averages of the forward and reverse simulations

From Eqs. (4) and (8), we note that since the exponential averages of the forward and reverse distributions have opposite signs, the bias will have opposite sign for the two distributions. Occasionally, the free energy from forward and reverse EXP simulations are averaged in order to remove some of the bias from the two individual values. Although the biases are usually in opposite directions, since the biases are directly proportional to the variances (which can be drastically different for the forward and reverse simulations), they will therefore not in general cancel.

The bias and variance of the average of the forward and reverse cases, ΔF_{sum} can thus be computed as

$$\mathrm{bias}(\beta\Delta F_{\mathrm{ave}}) = \frac{\hat{\sigma}_F^2}{8n_F} - \frac{\hat{\sigma}_R^2}{8n_R} + O(n^{-2}), \tag{11}$$

$$\mathrm{var}(\beta\Delta F_{\mathrm{ave}}) = \frac{\hat{\sigma}_F^2}{4n_F} + \frac{\hat{\sigma}_R^2}{4n_R} + O(n^{-2}). \tag{12}$$

As has been noted,[10–12] the simple average of the exponential average is not the ideal combination of the forward and reverse work. Is there a way to combine the forward and reverse exponential averages which will give smaller variance? If we estimate the free energy by the sum of the forward variance times a factor a and the reverse variance by the factor $1-a$ (such that the two weighting factors sum to 1), we find that the variance will be minimized by

$$a = \frac{n_F\hat{\sigma}_R^2}{n_F\hat{\sigma}_R^2 + n_R\hat{\sigma}_F^2} = \frac{\mathrm{var}(\beta\Delta F)_R}{\mathrm{var}(\beta\Delta F)_R + \mathrm{var}(\beta\Delta F)_F}. \tag{13}$$

This is, therefore, an improved combination compared to the simple average. We note that summing together the biases for the forward and reverse cases using this weighting and Eq. (8), the bias of order n^{-1} vanishes. So this weighting is also the minimum square error linear combination of the forward and reverse exponential averages.

Given $\hat{\sigma}_F$ and $\hat{\sigma}_R$, what is the best choice of n_F and n_R given the constraint that $n_F+n_R=n$? In other words, what is the best way to maximize computational efficiency in choosing to run simulations in the forward or in the reverse direction? Substituting the weighting faction in Eq. (13) into the expression for the variance and substituting $n-n_F$ for n_R, we find an expression for the variance

$$\mathrm{var}(\beta\Delta F) = \frac{\hat{\sigma}_F^2\hat{\sigma}_R^2}{n_F\hat{\sigma}_R^2 + (n-n_F)\hat{\sigma}_F^2} \tag{14}$$

to minimize with respect to n_F. There are no extrema in the interval $(0,n)$ and therefore the variance is minimized when all the trial runs are of the direction with the smallest variance $\hat{\sigma}^2$. The ideal division between forward and reverse distribution measurement in exponential averaging is therefore to sample both directions sufficiently to roughly estimate the variances $\hat{\sigma}^2$, and then sample only from the distribution with the smallest variance. Kofke and co-workers have demonstrated that this sampling is usually worse when sampling from the region of lower entropy,[10,11,26] though a rigorous understanding of this phenomena has not yet been reached.

D. Bennett acceptance ratio

However, as is noted earlier in this paper, the best way to utilize a number of forward and reverse work measurements is usually not to exponentially average them and weight these averages but to use the same data in BAR. In some cases, it may not be possible or reasonable to run free energy simulations in both directions—for example, where a large number of different ligand free energies are computed simultaneously, in one step, from a nonphysical intermediate state whose binding free energy is precisely known.[27] In this case, using only EXP is clearly preferable if it is possible to converge accurately.

A limiting moment analysis of the Bennett acceptance ratio method is not directly possible, as it is an implicit function of ΔF. Of course, we already have a perfectly serviceable estimate for the variance, but we have no such estimate for the bias. It is important, of course, to note that the bias is *not* simply 1/2 times the variance, since the estimate is not simply the logarithm of an average. But there is a close relationship to the variance. We present an expression for the leading $1/n$ term of the bias of BAR, which can be written in the form

$$\text{bias}(\beta\Delta F_{\text{BAR}}) = \frac{K}{2}\text{var}(\beta\Delta F_{\text{BAR}}), \tag{15}$$

where K is of order $O(n^0)$ in the number of measurements (see the Appendix for derivation). Indeed, in most situations (as demonstrated in our sample problems), this K is much less than 1. Whenever the forward and reverse distributions are symmetric around $W=\Delta F$, for example, K is exactly zero.

E. Thermodynamic integration

What about the bias and variance of thermodynamic integration? TI is significantly harder to compare theoretically with the other methods presented here, as the variance and bias that are obtained are not easily related to the variance and bias estimates for either EXP or for BAR. In TI, we first choose some pathway (here parametrized by the variable λ) between our initial state and final state, which we denote by $U(\lambda)$, where $U(0)$ is the initial state and $U(1)$ is the final state. There are an infinite number of ways to define the path. Once we define a path, regardless of the choice, we can compute the free energy as

$$\Delta F = \int_0^1 \left\langle \frac{dU(\lambda)}{d\lambda} \right\rangle d\lambda \tag{16}$$

and the variance of this free energy estimate, computing our averages at fixed λ, as

$$\begin{aligned}\text{var}(\Delta F) &= \int_0^1 \text{var}\left(\frac{dU(\lambda)}{d\lambda}\right) d\lambda \\ &= \int_0^1 \left\langle \left(\frac{dU(\lambda)}{d\lambda}\right)^2 \right\rangle - \left\langle \frac{dU(\lambda)}{d\lambda} \right\rangle^2 d\lambda .\end{aligned} \tag{17}$$

Inherently, there is no bias in this equation, if we could truly sample from the equilibrium distribution along λ continuously. But to compute fixed-λ averages with a finite number of samples, we must select a finite number of values of λ at which to measure.

Most chemical and biological simulations are therefore performed at fixed values of λ, and a free energy is generated by some sort of numerical integration. It is in this step that the bias is introduced. The variance of simple averages is well behaved, in that relatively few uncorrelated measurements are needed to get a decent estimate of the averages. Cursory analysis would therefore indicate that TI would be preferable to the other methods presented here, if we could somehow sample along the entire “reaction coordinate” λ profile. However, if we are running just two endpoint simulations or running just a few intermediates, it is also easy to see how curvature along this pathway could result in extremely biased free energy estimates, meaning other methods might be preferable.

Though it is not commonly known in the chemical and physics community, it is also correct to compute this average as an expectation value over *both* the physical ensemble and λ.[2] This also gives the free energy without bias; however the variance will then be

$$\begin{aligned}\text{var}(\Delta F) &= \int_0^1 \left\langle \left(\frac{dU(\lambda)}{d\lambda}\right)^2 \right\rangle d\lambda - \left(\int_0^1 \left\langle \frac{dU(\lambda)}{d\lambda} \right\rangle \right)^2 d\lambda \\ &= \int_0^1 \left\langle \left(\frac{dU(\lambda)}{d\lambda}\right)^2 \right\rangle d\lambda - \Delta F^2,\end{aligned} \tag{18}$$

which will always be somewhat larger. There have been some attempts to sample along λ in the chemical literature.[28,29] However, there are several difficulties in this. Such methods are nontrivial to implement into various chemical and biomolecular codes, and not particularly well understood in comparison to other sampling methods of a single Hamiltonian. The correlation times for reequilibration for most methods of sampling along λ are usually extremely long, and it is difficult to devise schemes that sample along all the λ range for the relevant Hamiltonians. Additionally, this second variance is always larger than the fixed λ variance, so it may not be an ideal method, though it may bear further investigation in some cases.

The bias of EXP and BAR are inversely proportional to the number of samples collected at each intermediate state. In the continuous λ case of TI, the bias will always be zero, whereas in the case of discrete λ values, the bias of TI is of order 1 in the number of samples and can only be reduced by running simulations at additional intermediate points. An alternate way to decrease the bias for discrete λ sampling is to use a higher order integration algorithm that the trapezoidal rule usually used—for example, Simpson’s rule or Gaussian quadrature.[30] While these algorithms scale better in the number of intermediates, they perform more poorly when the variance is high, as the numerical error depends on derivatives that are much less well determined by the data than the function itself.[31] Gaussian quadrature also requires the variances, or at least estimates for the variances be known beforehand in order to determine at which values of λ to collect data, which is usually not practical. Simpson’s rule requires intervals that are equally spaced, a requirement that may be at odds with functions where the curvature is concentrated primarily in one region. For these reasons, we will only briefly touch on the use of these alternate numerical integration methods for the use in TI for general free energy calculations, though in some specialized cases they may be appropriate.

III. ANALYTICALLY SOLVABLE MODELS

A. Offset harmonic wells

The use of simple, analytically solvable models allows researchers to investigate the behavior of the various methods over a fairly wide range of parameter values; this would be computationally expensive to accomplish with more com-

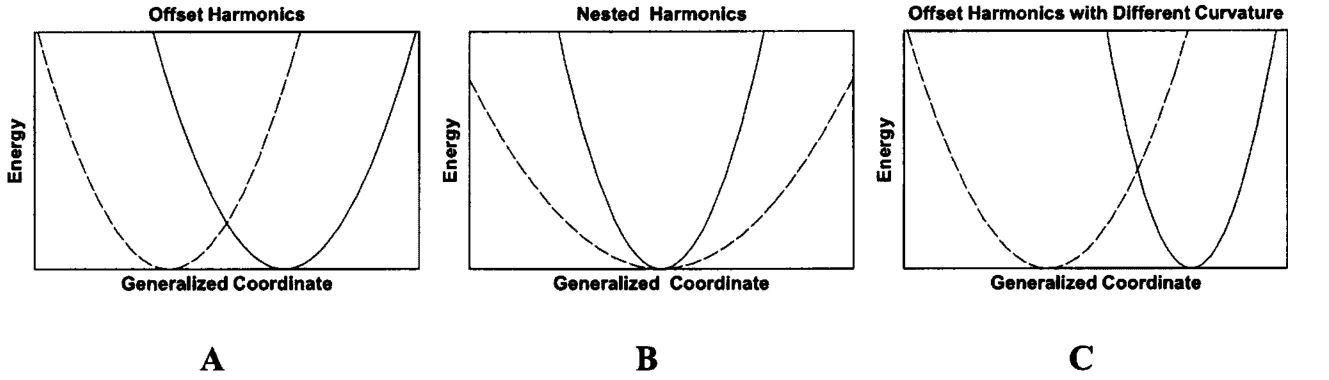

FIG. 1. Three analytical problems from this paper: (A) offset harmonic wells where $U_A(x)=a(x+c/2)^2$ and $U_B(x)=a(x-c/2)^2$, (B) nested harmonic wells where $U_A(x,y)=a(x^2+y^2)$ and $U_B(x,y)=b(x^2+y^2)$, with $a>b$ (cross section shown), and (C) offset harmonics with different curvatures, with $U_A(x)=a(x-c)^2$ and $U_B(x)=bx^2$, again, with $a>b$.

plicated (or realistic) models. What results do these estimation methods yield for model systems? For example, in perhaps the simplest case, where the distribution of work in both directions is Gaussian, what might the behavior be?

Such a distribution can be generated by actual potential energy functions; for example, two harmonic wells with identical curvature a separated by some constant b will result twin distributions $P_R(W)$ and $P_F(W)$ which are both Gaussians, where W is the potential energy difference between the two wells. If we have potential energy functions $U_A=a(x+c/2)^2$ and $U_B=a(x-c/2)^2$ [see Fig. 1(a)], the energy difference W at any point x will be simply $2acx$. Assume we are sampling from the $a(x-c/2)^2$ surface, the normalized probability of any state will be simply

$$P(x)dx=\sqrt{\frac{a\beta}{\pi}}\exp(-\beta a(x-c/2)^2)dx. \tag{19}$$

As there is a one-to-one correspondence in this case between x values and W values, making the substitution $W=2acx$ yields

$$P(W)=\sqrt{\frac{\beta}{4\pi ac^2}}\exp\left(-\beta\left(\frac{(W-ac^2)^2}{4ac^2}\right)\right). \tag{20}$$

$P_R(W)$ and $P_F(W)$ will be Gaussians with variance $2kTac^2$, separated by $2kTac^2$, symmetric across zero.

Generally, then, let us assume that $P_F(W)$ is a normalized Gaussian, with mean μ and variance σ^2 (σ having units of β^{-1}). The free energy is therefore

$$\Delta F=-\beta^{-1}\ln\int\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(W-\mu)^2}{2\sigma^2}\right)\times\exp(-\beta W)dW=\mu-\beta\sigma^2/2. \tag{21}$$

Remembering Eq. (4), we see that if $P_F(W)$ is Gaussian, with the given mean and variance, then $P_R(-W)$ must also be Gaussian, with identical variance, with the means between the two distribution differing by σ^2. Without loss of generality, we can set the free energy to be zero, and set the means of the two forward and reverse Gaussians to be $-\beta\sigma^2/2$ and $\beta\sigma^2/2$, respectively.

What are the variance and biases of the free energy of a Gaussian distribution with exponential averaging? With the theory developed earlier, we find them by evaluating the function $\hat{\sigma}^2$ described above. With symmetric work distributions, the free energy must be zero, so

$$\langle[\exp(-W)]^2\rangle=\langle\exp(-2W)\rangle=\exp(\sigma^2) \tag{22}$$

and the limiting term in the variance of EXP with n samples will be

$$\mathrm{var}(\beta\Delta F)\approx\frac{\hat{\sigma}^2}{n}=\frac{\exp(\sigma^2)-1}{n} \tag{23}$$

with, as usual, bias equal to half the variance. As we can see, this has particularly poor precision for $\sigma>1$. If we were to average results from both the forward and reverse directions, the variance would be of the same form (as both forward and reverse give the same variance), but the bias would be zero, as the two estimates are perfectly symmetric.

What are the variance and bias of BAR in this case? We have an equation for the leading term of the variance [Eq. (3)], but it is not particularly amenable for analytical work. Gelman and Meng derive the variance for BAR between two Gaussian distributions as approximately proportional to $\sigma^2(\exp(\sigma^2/8)-1)$ for large σ.[2] The limiting behavior of the bias, in the case of the twin Gaussians symmetric around zero, is identically zero in the case of equal numbers of forward and reverse simulations, much different from the sometimes significantly large bias of EXP for Gaussians.

Next we examine the bias and variance of TI. For now, we will assume the most "natural" case that sampling can be done at any point along the λ path—later, we will address the case of intermediate points. One simple path definition is simply the linear interpolation between potential functions

$$U(\lambda)=\lambda U_A+(1-\lambda)U_B=\lambda a(x+c/2)^2+(1-\lambda)a(x-c/2)^2. \tag{24}$$

In this case, it is simple to compute that $\langle dH/d\lambda\rangle=\langle dU/d\lambda\rangle=ac^2(1-2\lambda)$, and the variance (where $\sigma^2=2kTac^2$) is $\sigma^2+\sigma^4/12$, leading to a total variance of $n^{-1}(\sigma^2+\sigma^4/12)$.

There are, of course, an infinite number of other possible paths. Perhaps the simplest natural path is for the constant c to be λ dependent, so that $U(\lambda)=a((x-(1-2\lambda)c)/2)^2$, shifting the location of the harmonic well but not its curvature at

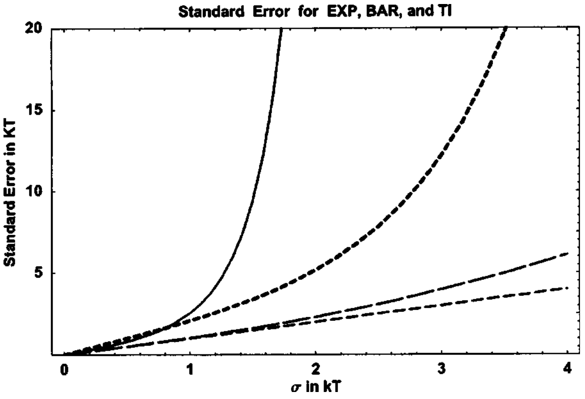

FIG. 2. Standard error $[\sqrt{n}\,\mathrm{var}(\Delta F)]$ vs σ $(2kTac^2)$ of offset harmonic wells for EXP (solid), BAR (dotted), linear TI (long dashed), and $c \propto \lambda$ TI (short dashed). BAR is significantly better than EXP for values of $\sigma > 1$, and both TI pathways are in turn better than BAR, although TI takes advantage of certain simplicities in this toy problem. There is only one EXP result, as the probability distribution is symmetric in the forward and reverse direction.

each intermediate state. In this case, each intermediate distribution is a Gaussian with the same curvature and hence same free energy, so $\langle dH/dl \rangle = \langle dU/d\lambda \rangle = \Delta F = 0$. Then the variance is $2kTac^2 = \sigma^2$ for all λ, leading to an overall variance of $(1/n)\sigma^2$. It is actually possible to construct pathways that have even lower variance for the Gaussian case,[2] but even in the simplest cases like this one these highly optimal paths are difficult to derive.

In Fig. 2, we compare, as a function of σ in units of kT, the variance of EXP, BAR, and TI. We plot the standard error (i.e., the square root of the variance) for both EXP (just one curve, as forward and reverse are symmetric) and BAR, as a function of σ in units of kT, multiplied by $\sqrt{n}$ (to be independent of the number of samples). BAR requires data from two states, so we assume that $n/2$ measurements are performed at each state. Although for very small values of σ, EXP is slightly better than BAR, clearly, if $\sigma > 1$, BAR becomes much more efficient than EXP. Both continuously sampled TI pathways are even better than BAR in this simplest model case.

B. Nested harmonic wells

Next we examine a model yielding slightly more complicated, nonsymmetric behavior. We take two nested two-dimensional harmonic potential energy functions, $U_A = a(x^2 + y^2)$ and $U_B = b(x^2 + y^2)$ [See Fig. 1(b)]. The free energy between these wells can be easily computed for arbitrary d-dimensional harmonic wells to be $(2\beta)^{-1} d \ln(a/b)$, yielding $\beta^{-1} \ln(a/b)$ in this particular two dimensional case.

Let us compute the forward and reverse differences in potential energy between these two distributions, assuming $a > b$ for simplicity. We will call the direction in which we move to a higher energy surface "forward" $(P_F(W))$, and the opposite direction "reverse" $(P_R(W))$.

If we are simulating on the lower surface b, then energy difference $W = (a-b)(x^2+y^2)$ corresponds to a state with energy $b(x^2+y^2)$, which has normalized probability (from the Boltzmann distribution),

$$P(r)dr = \frac{\beta b}{\pi} \exp(-\beta b(x^2+y^2))dxdy$$
$$= \frac{\beta b}{\pi} \exp(-\beta b(x^2+y^2)) 2\pi r dr. \qquad (25)$$

Since there is still a one-to-one correspondence between differences in energy and states, we can substitute $W = (a-b)r^2$ and $dW = 2(a-b)rdr$ to obtain

$$P_F(W) = \frac{\beta b}{a-b} \exp\left(\frac{-\beta b W}{a-b}\right). \qquad (26)$$

Similarly,

$$P_R(W) = \frac{\beta a}{a-b} \exp\left(\frac{-\beta a W}{a-b}\right). \qquad (27)$$

Applying EXP to these equations yields the correct free energy difference from b to a, $\beta^{-1} \ln(a/b)$.

1. Variance and bias with EXP

Evaluating the variance of $\langle \exp \beta(W - \Delta F) \rangle$ over the two distributions, we obtain

$$\mathrm{var}(W)_{P_F} = \frac{1}{n\beta^2} \frac{(a-b)^2}{2ab - b^2}, \qquad (28)$$

$$\mathrm{var}(W)_{P_R} = \frac{1}{n\beta^2} \frac{(a-b)^2}{2ab - a^2}, \qquad (29)$$

with the biases one half of this variance. Surprisingly, the variance and bias are only defined for $P_R(W)$ if $a < 2b$. If the upper well is more than twice as narrow as to the lower well, the spread of values leads to a variance that does not converge. This does not mean that no information on the error is obtainable, as there is still a perfectly good probability density; we could still compute confidence intervals, for example. But the uncertainty will not have the simple n^{-1} behavior, and will always scale worse than n^{-1} when $a \geq 2b$.

This demonstrates some serious problems of the use of EXP in determining free energies, even in simple toy problems. This also demonstrates a clear example of the fact, as noted before, that the distribution from the higher entropy state (the well of lower curvature) to lower entropy state yields a better result using EXP than the reverse distribution. Not surprisingly, BAR is still too complicated for an analytical form; a graph of the standard error of all methods is shown in Fig. 3, again accounting for the fact that BAR must have $n/2$ simulations at each end state.

2. Thermodynamic integration

Using the simple linear interpolation, we have the potential energy function $U(\lambda) = \lambda U_A + (1-\lambda)U_B = \lambda a r^2 + (1-\lambda) b r^2$ and $dH/d\lambda = dU/d\lambda = (a-b)r^2$. Then

$$\langle dH/d\lambda \rangle = \frac{a-b}{\beta(\lambda a + (1-\lambda)b)}, \qquad (30)$$

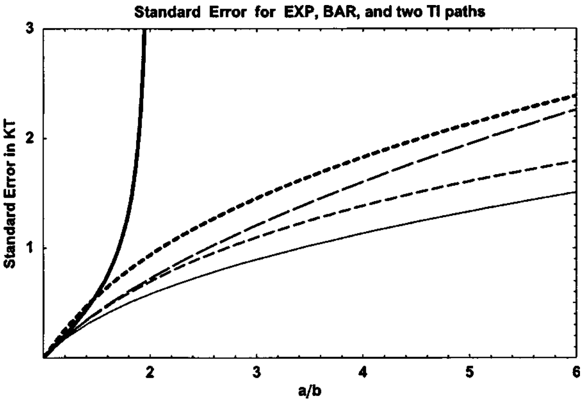

FIG. 3. Standard error $[\sqrt{n\,\mathrm{var}(\Delta F)}]$ vs a/b, the ratio of the curvatures of nested harmonic wells for reverse EXP (thick solid), forward EXP (solid), linear TI (long-dashed), logarithmic interpolation TI (short-dashed), and BAR (dotted). Surprisingly, forward EXP is slightly better than BAR, and contrary to the previous example, both methods are more efficient than TI.

$$\mathrm{var}(dH/d\lambda) = \frac{(a-b)^2}{\beta^2(\lambda a + (1-\lambda)b)^2} \tag{31}$$

yielding an overall total variance (integrating uniformly along λ) of

$$\frac{(a-b)^2}{\beta^2 ab}. \tag{32}$$

Of course, there is nothing to prevent us from sampling more in areas of higher variance, and less in areas of low variance—TI is still unbiased in this case. The minimum variance is obtained by choosing a density of sampling that is proportional to $\mathrm{var}(dH/d\lambda)^{-1/2}$ at each λ value, yielding a final variance of $\beta^{-2}\ln(a/b)^2$. Of course, this requires knowing beforehand the variance as a function of λ! One could construct a situation, of course, where the sampling is dynamically reweighted to different values of λ as data is collected. Care must be taken to ensure that this dynamic sampling is done over long enough time intervals or some regions of high variance may not be sufficiently sampled. In practice, therefore, this ideal weighting is not achievable, though significant improvement from equal weighting may be made.

There are of course other pathways, for example, $U(\lambda) = a^L b^{1-L} r^2$. With this pathway, we obtain $\mathrm{var}(dU/d\lambda) = \beta^{-2}\ln(a/b)^2$ for all λ, yielding a total variance of $\beta^{-2}\ln(a/b)^2$, the same as the reweighted variance above.

In Fig. 3, we compare the total variance as a function of a/b (with $b=kT$), of forward and reverse EXP, BAR, and the two pathways for TI. We see that reverse EXP is extremely poor, with variance going to infinity when $a>2b$. Forward EXP is the most efficient free energy method in this case, a surprising fact, confirmed with numerical simulations. BAR finds the minimum for all asymptotically unbiased estimators for a given set of forward and reverse measurements. But it appears that the data from the low-information content reverse sampling is making BAR worse than the estimate achieved from the high information forward sampling. In any case, the efficiency of BAR is very close to forward EXP with both TI pathways close behind.

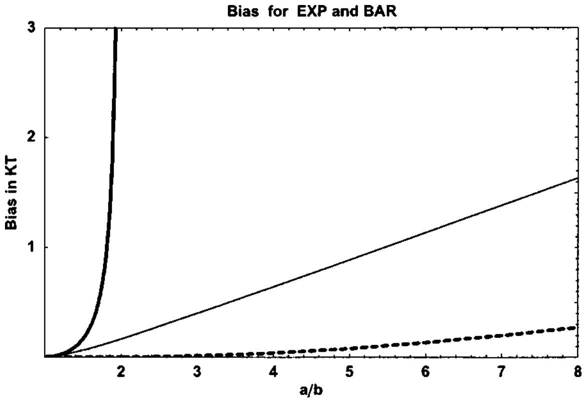

FIG. 4. Finite-sample bias times n (number of samples) vs a/b, the ratio of the curvatures of nested harmonic wells for reverse EXP (thick solid), forward EXP (thin solid), and BAR (dotted). The plotted reverse and BAR biases are −1 times the actual bias, for comparison purposes. The bias for BAR is close to zero for all a/b, with the bias of forward EXP increasing linearly with increasing a/b, and the bias of reverse EXP diverging to ∞ as $a/b \rightarrow 2$.

In Fig. 4, we compare the total bias as a function of a/b (with $b=kT$) of the forward and reverse EXP and BAR. We defer the discussion of the bias of TI to a later discussion of the role of intermediates, as continuous sampling for TI in λ is free of bias for any pathway. We note that the bias of BAR is extremely low over all the entire interval. The bias of both EXP are simply 1/2 the variance, so they have identical relative behavior as in the previous graph.

C. Offset harmonic wells of different curvature

It may seem from the above cases that it may be relatively easy to construct such potentials. We give an example of one such other very simple analytically solvable problem that demonstrates the difficulties of constructing such distributions in general, and show some of the nonintuitive nature of these probability densities.

Take $U_A(x)=a(x-c)^2$ and $U_B(x)=bx^2$. This represents two harmonic wells, offset, and with different curvature [Fig. 1(c)]. Again, we assume $a>b$ and that call the direction in which we move to the higher curvature surface "forward" $(P_F(W))$, and the opposite direction "reverse" $(P_R(W))$. The free energy can easily be computed to be the same as non-offset wells, namely, $\beta^{-1}\ln(a/b)$.

Solving a quadratic equation for x, the energy difference $W=(a-b)x^2-2ac+c^2$ now corresponds to two possible states. Both states are equally valid. Transforming the probability density as done previously, we end up with normalized probability densities

$$P_F(W) = \left(\frac{\beta b}{4\pi\sqrt{abc^2+(a-b)W}}\right)^{1/2} \times \left[\exp\left(-\beta b\left(-c+\frac{ac-\sqrt{abc^2-(a-b)W}}{a-b}\right)^2\right) + \exp\left(-\beta b\left(-c+\frac{ac+\sqrt{abc^2-(a-b)W}}{a-b}\right)^2\right)\right], \tag{33}$$

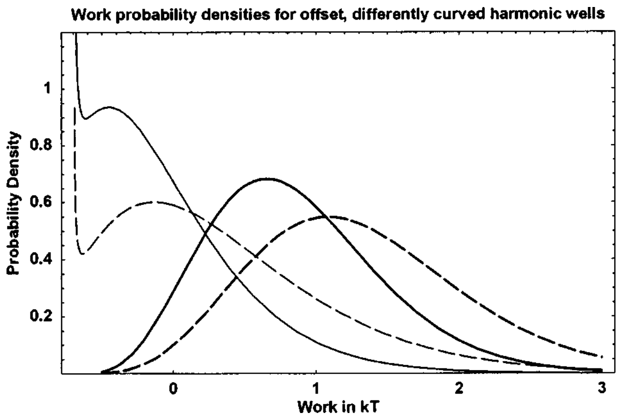

FIG. 5. Probability of forward (dashed) and reverse (solid) work distributions, for (thick lines) $a=1.5$, $b=1.3$, and $c=0.4$, and (thin lines) $a=1.3$, $b=1$, and $c=0.4$ for nested Gaussians of the form $U_A(x)=a(x-c)^2$ and $U_B(x)=bx^2$. Relatively small changes in a and b drastically change the shape of the distributions, roughly interpolating between the 1D logistic curves of the second toy problem, and the Gaussians of the first toy problem.

$$P_R(W) = \left(\frac{\beta a}{4\pi\sqrt{abc^2+(a-b)W}}\right)^{1/2} \times\left[\exp\left(-\beta a\left(-c+\frac{ac-\sqrt{abc^2-(a-b)W}}{a-b}\right)^2\right) + \exp\left(-\beta a\left(-c+\frac{ac+\sqrt{abc^2-(a-b)W}}{a-b}\right)^2\right)\right],$$

which is, obviously, much more complicated than any previous distribution equation derived in this paper. Looking at Fig. 5, the probability curves interpolate between the Gaussians of the first toy problem, and the one dimensional variation of the logarithmic functions of the second toy problem.

In general, we need to find a reverse function to map from each sampled state to the corresponding work value. As the complexity of the system increases, the number of energy states which correspond to a given work difference increases dramatically. With biomolecules in the condensed phase, any such equations become completely unusable at best, and most cases unwritable. For nonequilibrium work distributions, the computation of these work distributions cannot even be attempted with the methods presented here, as there are no equilibrium weightings of the individual states to work from.

D. Using intermediate states to improve efficiency

The above analyses assumed either simulations were sampled with no intermediate states sampled between the end states, or in the case of TI, that all intermediate states are accessible to sampling. These situations are frequently either far from ideal or simply not possible. In the case of BAR and EXP, the states of interest frequently have too little overlap in phase space to be accessible from each other with sampling from endpoints. In TI, the equilibration times to switch between intermediate states are far too long, and so we must run multiple simulations at a specified number of fixed λ intermediate states.

We will make these analyses assuming that we are collecting our n samples at over m states (intermediates plus the initial and final state). We will use the same pathways as are described for TI in the examples above. For simplicity, we will use equally spaced intermediates, although other spacings may prove preferable in practice. In general, for well-chosen paths, using multiple intermediate states will result in the measurement of free energies between states with more overlap, and that therefore have better convergence properties. However, since with m intermediates, we will need to measure $m+1$ free energies, whose variance and bias will generally be additive, and each of these states will be sampled with only $(m+2)^{-1}$ as many data points. If the statistic of interest converges too slowly with increasing numbers of intermediate states, then more intermediates can actually result in a greater variance or bias.

1. Variance and bias of harmonic wells with intermediates

For simplicity, we first approach the case of harmonic potential wells with the simple linear interpolation of the potential function, as in Eq. (24). If we assume we are measuring an energy difference between states with λ_A and λ_B, using the methods developed earlier in this paper, we obtain $W=2acx(\lambda_B-\lambda_A)$, and normalized distribution

$$P_A(W) = \sqrt{\frac{\beta}{4\pi ac^2(\lambda_a-\lambda_b)}} \times\exp\left(-a\beta\left(\frac{W-ac^2(1-2\lambda_A)(\lambda_B-\lambda_A)}{4ac(\lambda_A-\lambda_B)}\right)^2\right). \quad (34)$$

As this is still a Gaussian, but now with variance $2kT\pi ac^2(\lambda_A-\lambda_B)^2$, and average $ac^2(\lambda_A-\lambda_B)(1-2\lambda_A)$, we can apply the same equations we did before. With m equally spaced intermediates, the variance of each distribution will be simply $(m+1)^{-1}$ the original σ^2 variance, meaning that the overall variance of EXP will be (with $n/(m+1)$ observations at each of $m+1$ total states yielding $m+1$ free energies with additive variances)

$$\mathrm{var}(\Delta F_m) = \frac{(m+1)^2}{n}(\exp(\sigma^2/(m+1)^2)-1). \quad (35)$$

In the $c \propto \lambda$ pathway, then

$$W = a(x-(1-2\lambda_A)c/2)^2 - a(x-(1-2\lambda_B)c/2)^2 = 2acx(\lambda_A-\lambda_B) + ac^2(\lambda_B-\lambda_A)(1-(\lambda_B+\lambda_A)). \quad (36)$$

However, the normalized distribution again ends up being a Gaussian with different mean than in the previous example, but the same variance, $2kTac^2(\lambda_A-\lambda_B)^2$. This yields the same overall variance in the free energy between intermediates as in the linear pathway, shown in Eq. (35).

Increasing the number of intermediates will always improve the precision for EXP in the Gaussian case, although the efficiency falls off relatively quickly. In the limit of large m, the variance converges to the variance in the TI case, which can be simply computed using the previously presented theory to be σ^2 for all values of λ (see Fig. 6). The

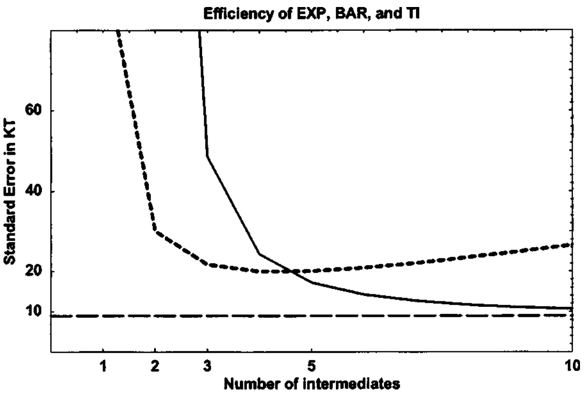

FIG. 6. Standard error of EXP (solid), BAR (dotted), and TI (dashed) estimates of the free energy, as a function of number of intermediates, from Gaussian probability distributions with variance $\sigma^2 = 80$. EXP converges to the TI result, but BAR reaches a minimum before increasing in variance. For a large portion of the range, however, BAR is still significantly more efficient than EXP.

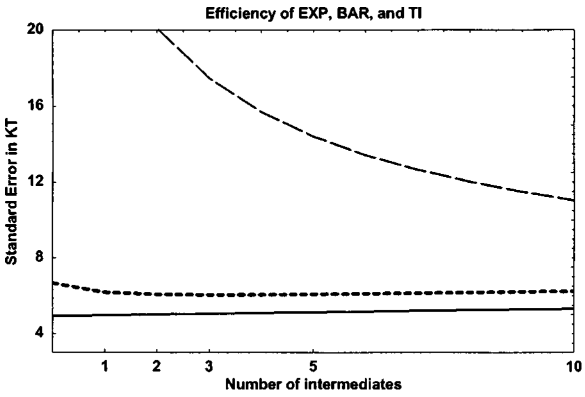

FIG. 7. Standard error of TI (dashed), BAR (dotted), and forward EXP (solid) estimates of the free energy, as a function of number of intermediates, from nested harmonic wells with $a/b = 50$, using a linear path. Error from forward EXP and BAR increase slightly with increase in intermediates, while TI initially starts high but eventually decreases to a similar value as the other methods.

variance and bias of BAR, again, are not possible to express analytically. Because the variance using BAR scales slightly greater than exponential in σ^2 as discussed earlier (though an exponential with a much slower rate in increase than EXP), it will actually reach a minimum with number of intermediates, and then increase (Fig. 6).

We note that with no intermediates, we will need to sample from both end states for BAR, but only one for EXP. The ratio of the number of total states we need to sample with BAR versus EXP with m is $(m+1)/m$, so the more intermediates need to be run, the smaller this advantage of EXP. If double wide sampling is used for EXP,[32] where EXP is performed in the forward and reverse direction from intermediate n to intermediates $n-1$ and $n+1$ but not at intermediates $n+1$ and intermediate $n-1$ themselves, this 2:1 advantage remains. However, this only is relevant if the variances of forward and reverse EXP are comparable, as in the case of the Gaussian distributions. In other cases, such an improvement is unlikely, as the results from one of the directions will be much less accurate, and we do not assume double wide sampling in Fig. 6.

The variance for TI in the case of Gaussian distributions will be independent of the number of intermediates, since the variance is the same at all points. In the linear case, bias would be identically zero for any number of intermediate states as long as the states $\lambda = 0$ and $\lambda = 1$ are sampled from. In the $c \propto \lambda$ case, the bias will be zero for any number of samples. This is, of course, a special, even degenerate case, as the variance is the same at all λ and $\langle dH/d\lambda \rangle$ is a straight line (Fig. 6).

2. Intermediates in the case of nested harmonic wells

We now turn to the more complicated case of intermediates in nested harmonic wells discussed earlier. Using a linear interpolation with λ, and writing $k(\lambda) = b + \lambda(a-b)$ we compute the distributions from λ_A to λ_B as

$$P_F(W) = \frac{\beta k(\lambda_A)}{(\lambda_A - \lambda_B)(a-b)} \exp\left(\frac{-\beta k(\lambda_A) W}{(\lambda_A - \lambda B)(a-b)} \right). \quad (37)$$

Similarly,

$$P_R(W) = \frac{\beta k(\lambda_B)}{(\lambda_A - \lambda_B)(a-b)} \exp\left(\frac{-\beta k(\lambda_A) W}{(\lambda_A - \lambda B)(a-b)} \right) \quad (38)$$

yielding an overall variance of

$$\mathrm{var}(\Delta F(\lambda_A \to \lambda_B)) = \frac{(a-b)(\lambda_A - \lambda_B)}{k(\lambda_A)(k(\lambda_A) - 2(a-b)(\lambda_A - 2\lambda_B))}. \quad (39)$$

There is no simple analytical result as a function of the number of states m. The formulas for BAR and TI are similarly nonanalytical, so we will examine the graphical comparisons for all methods. We use a ratio of a/b of 50, and note that the variance will only be defined for reverse EXP when $m > 50$, and is therefore not shown.

In Fig. 7, we see that only in the case of TI and reverse EXP does adding intermediates significantly improve the variance. Forward EXP and BAR are virtually the same, with TI needing significantly more intermediates to converge to a similar answer. This is because in the linear path, most of the variance is concentrated between the end state with curvature a and its nearest intermediate. Adding equally spaced intermediates does not necessarily help reduce the variance of most of the intermediate free energies, and the variance from these states actually increases as they are sampled with fewer measurements, canceling out the improvement in the uppermost states.

Examining the bias of EXP, BAR, and TI (Fig. 8), we see that the bias of BAR is small for all numbers of intermediates, and additionally converges quickly to zero. TI bias starts out quite high, but rapidly converges to zero as well. The bias of forward EXP, however, like the variance, increases slightly from the minimum obtained with only the end states.

We also examine the second, logarithmic interpolation pathway between nested harmonic wells. In Fig. 9, we see

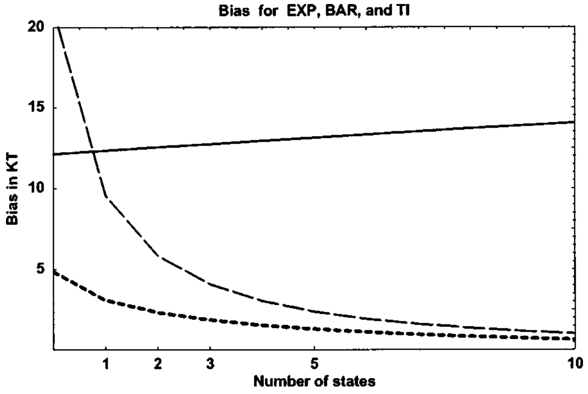

FIG. 8. Bias of forward EXP (solid), BAR (dotted), and TI (dashed) estimates of the free energy, as a function of number of intermediates, from nested harmonic wells with $a/b=50$, using a linear path. Error from forward EXP increases slightly with number of intermediates, while TI decreases quickly from relatively high values, and BAR remains low in all cases.

that the variance of TI is independent of the number of states sampled from, as the variance is the same for all states. Forward EXP and BAR dip slightly below the TI variance, and then plateau, while reverse EXP asymptotically approaches a similar variance as well. All methods perform significantly better for the logarithmic pathway than the linear pathway.

For the bias of the logarithmic path (no figure), we remember that the bias of both the forward and reverse EXP is just $1/2$ of the variance, and that the logarithmic interpolation for TI has no bias, as the slope of $\langle dH/d\lambda \rangle$ is a straight line. The bias of BAR goes to zero with increasing intermediate states even more quickly than in the linear interpolation case.

The improved behavior of the logarithmic interpolation pathway makes sense, as neighbors will tend to be more evenly spaced than with linear interpolation. For example, if $a=50$ and $b=1$ as above, in the logarithmic case a single intermediate will have prefix $\sqrt{50}$, whereas in the linear case, the intermediate will have prefix $21/2$. Since the free energy is $kT\ln(a/b)$, the logarithmic intermediates have free energies equally spaced between the end states, where in the linear case the intermediate is much close to the a end state.

E. Summary of results with toy models

This review of only these few simple models demonstrates the complexity of this comparison. Each of the three methods appear to be best in at least some situations. Where $\langle dH/d\lambda \rangle$ is smooth, TI appears to be almost uniformly best. Surprisingly, in the case of nested harmonic wells, forward EXP is not only better than TI but it is better than BAR. This is because BAR is the best estimate from a *specified* set of sampled work values. There is nothing that says that another set of sampled work values might give an improved estimate. In this case, it is apparently because the extremely poor properties of reverse EXP, which actually has a divergent variance, and essentially contributes “negative” information to the BAR estimate. BAR has been shown to be the best estimate under previously published toy problems.[21] These observations should serve as fair warning to carefully evaluate the applicability of various free energy methods, as methods that work in some situations may be much less than optimal in some cases.

Adding a few intermediates can be extremely helpful in many situations. In many cases, the ideal number of intermediates can be quite low, even for extremely large free energy differences. Using too many intermediates will actually decrease the efficiency of free energy estimates, as each additional intermediate result worsens the statistical sampling for each intermediate free energy. We can see by the drastic differences between the uncertainties of logarithmic and linear pathways between nested harmonic wells the importance of a good choice of pathway. This choice can be even more important than the choice of method used to extract these energies.

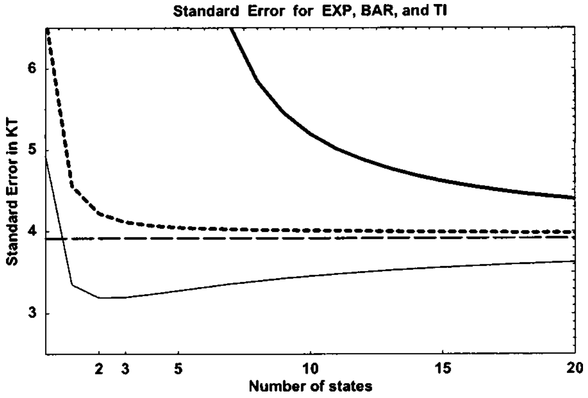

FIG. 9. Standard error of reverse EXP (thick solid), forward EXP (thin solid), BAR (dotted), and TI (dashed horizontal) estimates of the free energy, as a function of number of intermediates, from nested harmonic wells with $a/b=50$, using a logarithmic path. All values go to their asymptotic limit (approximately the same in all cases) much more quickly than the linear path. Reverse EXP is by far the least efficient, but increases steadily with number of intermediates, while BAR and forward EXP quickly reach a minimum and then gradually increase.

IV. ATOMISTIC FREE ENERGY CALCULATIONS ON SYSTEMS RELEVANT TO BIOMOLECULES

Theoretical analysis and toy problems are not sufficient to elucidate all the subtleties of free energy computations for more complex systems, so we turn to data collected from simulations for more guidance and examples. We will first look at the free energy of a Lennard-Jones sphere solvated in explicit water, first using work distributions generated from slow- and fast-growth simulations, and then with work distributions generated from EXP data with a number of different intermediates.

Next, we will examine the deletion/insertion of a model of 3-methylindole, the small molecule analog of the side chain of the amino acid tryptophan. This system was chosen primarily because data had already been collected in a previous study[15]—the primary interest for this molecule in this study is simply its large size, with 19 atoms, making it a challenging test case, and rigid structure, making it easier to remove the question of sampling uncorrelated measurements.

TABLE I. Free energies of hydration computed by both fast and slow growth of solvation of OPLS-UA methane in TIP3P water. The first column is the length of the simulation runs in picoseconds used to generate the distribution of work values, and the second column is the number of simulations run in the forward and backward directions, for a total constant aggregate time of 20 ns simulation in all cases. The third through seventh columns are the free energy estimates and analytic uncertainty estimates from BAR, forward EXP, reverse EXP the direct average of forward and reverse EXP, and the optimally weighted average of forward and reverse EXP, as per Eq. (14). All values are in kcal/mol. The free energy of solvation using exhaustive TI (71 total lambda values, 1 ns simulation at each value of λ, four copies, for ≈280 ns total) is 2.654±0.007 kcal/mol. We see that BAR does well even with many very short simulations, where as the other methods break down in this regime.

Time	n	BAR	Forward EXP	Reverse EXP	Average	Optima EXP average
500	20	2.67±0.03	2.70±0.04	2.65±0.05	2.68±0.03	2.68±0.02
200	50	2.64±0.03	2.67±0.04	2.65±0.06	2.66±0.04	2.67±0.02
100	100	2.58±0.04	2.65±0.06	2.58±0.10	2.61±0.06	2.63±0.04
50	200	2.69±0.04	2.69±0.08	2.67±0.08	2.68±0.06	2.68±0.04
20	500	2.61±0.04	2.67±0.09	2.36±0.07	2.52±0.06	2.49±0.04
10	1 000	2.67±0.04	2.49±0.11	2.24±0.08	2.37±0.07	2.32±0.05
5	2 000	2.62±0.04	2.75±0.13	2.69±0.36	2.72±0.19	2.74±0.10
2	5 000	2.61±0.05	2.59±0.20	2.51±0.41	2.55±0.23	2.58±0.13
1	10 000	2.63±0.05	3.04±0.12	1.49±0.30	2.26±0.16	2.82±0.09
0.5	20 000	2.64±0.05	2.63±0.19	0.14±0.10	1.38±0.11	0.64±0.06
0.2	50 000	2.73±0.05	2.58±0.28	−0.39±0.11	1.10±0.15	0.01±0.08
0.1	100 000	2.66±0.05	2.33±0.22	−0.81±0.12	0.76±0.12	−0.05±0.08

A. Free energy of methane solvation from nonequilibrium work

Our first test physical system is a Lennard-Jones sphere using the united atom methane parameters of OPLS ($\epsilon = 0.294$ kcal/mol, $\sigma = 3.73$ Å)[33] inside a box of 216 TIP3P water molecules.[34] It is modified slightly in that inside of 0.8σ, the Lennard-Jones term is replaced by a quadratic function, chosen such that it is continuous and has a derivative at 0.8σ.[6] For $kT = 0.592$ kcal/mol, (corresponding to 298 K) the region inside 0.8σ is almost never occupied, so it is very similar to the original Lennard-Jones. This formulation of Van der Waals attractions was chosen so that insertions from zero interaction do not have numerical instabilities resulting from the r^{-12} term as r goes to zero. Separate simulations estimated that the solvation free energy difference between this modified potential and the original Lennard-Jones potential in TIP3P water to be about -0.02 ± 0.02 kcal/mol.

The pathway in λ between the two states (here, the presence and absence of the modified Lennard-Jones particle) is simply λ^2 times the intramolecular energy between the solute and the water.[6] We will determine the free energy of solvation of this Lennard-Jones sphere using the methods presented above. For a reference, we first used exhaustive TI (71 total lambda values, 1.0 ns at each λ value repeated in fourfold replica, for ≈280 ns total sampling time) to obtain a free energy of solvation of 2.654 ± 0.007 kcal/mol.

We compute the work required to insert and remove the Lennard-Jones sphere from the box of water over a range of times from 500 to 0.1 ps. We fix the total simulation time at 10 ns in each direction. We then compute the free energy from BAR, EXP in the forward direction, EXP in the reverse direction, average of the two EXP calculations, and the optimally weighted EXP (using the weighting in Eq. (13)) presented in Table I.

Uncertainties for all measurements were determined both by the bootstrap method,[35] using 1000 bootstrap samples of the initial data set, and the analytical variance and bias estimates presented in the theory section. We find that the bootstrap and the analytical uncertainty estimates agree to within 15% for all measurements, usually with significantly better agreement. In Table I, we therefore present only the analytic uncertainties. The forward and reverse EXP estimates use only half the data than the BAR, the average EXP and the optimal EXP averages do.

We see in Table I that BAR is uniformly good over all sets of data, although the variance grows slightly as the work distributions are taken further from equilibrium (i.e., from shorter simulations). Forward EXP is relatively good over most of the range, but the uncertainties are higher using faster growth simulations. Reverse EXP is also good for slow growth work distributions, but becomes very poor for fast growth simulations, indicating that we are not sampling well from the reverse distribution. This is in line with the previously noted observation that simulating from a low to high entropy state is less efficient than in the other direction. The average and optimal exponential average combination have good behavior for slow growth, with values close to the actual value and with low variances. However, as the reverse average becomes poor, the average and optimal combination values suffer as well.

B. Free energies of solvation of methane from equilibrium simulations

To augment this data, a series of equilibrium simulations were performed, each simulating at a different set of fixed λ along the same pathway described in the nonequilibrium work section. 20 ns total time was run for each series, divided among the different λ values equally, so that sets of more closely spaced λ values were run for less time at each

TABLE II. Free energies of hydration computed with various methods. The first column (N) is the total number of states simulated. The number of $\Delta\lambda$ intervals is therefore $N-1$. Columns two through five are the free energy estimate using BAR, forward EXP, reverse EXP, and TI. Uncertainties are from the estimates presented in the paper. The answer using exhaustive TI (71 total lambda values, 1 ns simulation at each λ value, four copies, for ≈ 280 ns total) is 2.654 ± 0.007 kcal/mol. All free energies were estimated by running 20 ns total of each simulation, divide equally among the states used. The potential energy difference and value of $dH/d\lambda$ was output every 0.1 ps.

N	BAR	EXP forward	EXP reverse	TI
2	2.66 ± 0.08	2.61 ± 0.26	-1.65 ± 0.06	-2.91 ± 0.003
3	2.63 ± 0.04	2.66 ± 0.09	-0.10 ± 0.12	-1.48 ± 0.008
4	2.60 ± 0.02	2.54 ± 0.05	1.34 ± 0.14	-0.25 ± 0.01
5	2.68 ± 0.01	2.62 ± 0.04	2.76 ± 0.13	1.24 ± 0.02
6	2.67 ± 0.01	2.73 ± 0.02	2.47 ± 0.05	1.77 ± 0.02
8	2.67 ± 0.01	2.68 ± 0.02	2.48 ± 0.08	2.41 ± 0.02
10	2.68 ± 0.01	2.69 ± 0.02	2.52 ± 0.03	2.40 ± 0.02

λ. We ran 2, 3, 4, 5, 6, 8, and 10 values of λ, using equal spacing from 0.0 to 1.0, inclusive. Results from BAR and forward and reverse EXP are shown in Table II. We see that BAR is essentially correct over all the intervals, even including one-step insertion/deletion. Only forward EXP (insertion) gets close to the right value in one step, but with a much higher variance. TI has a strong numerical bias that results in incorrect free energies over all the numbers of intervals. The curvature of $\langle dH/d\lambda \rangle$ near $\lambda = 0$ (full deletion) is such that a reallocation of λ sampling near the origin would have given a lower overall bias, but clearly BAR is much more effective over a much larger range of allocation of independent measurements.

We also reanalyze the EXP data for the single step insertion/deletion (the first line in Table II to illustrate the nature of the variance and bias of BAR compared to EXP, using bootstrap sampling, as shown in Table III. We compare only to forward EXP, as the reverse distribution is not well sampled enough in this case.

It is clear from Table III that BAR estimate is uniformly better than the forward EXP. Comparable results (both with respect to bias and variance) can be obtained with only $10\,000 \times 2$ measurements using the acceptance ratio estimate as with all 100 000 samples with exponential averaging, indicating that in this case, the acceptance ratio is five times more efficient for an equivalent precision. Since we know for the agreement for large n that the analytical and bootstrap bias and variances agree, the deviations for small n seem to indicate that we are now getting into the range that the terms beyond the n^{-1} term in the Taylor series expansion in Eq. (7) are significant.

C. Larger molecules

We now examine the case of the solvation of 3-methylindole, an analog of the side chain of the amino acid tryptophan. This is the largest molecule examined in an earlier study of force fields and small molecule solvation.[15] The molecule was solvated in TIP3P water, and the free energy of hydration was computed. The complete methodological details are given in that paper, but we review the key features here. A pathway was constructed that involved first turning off the charges with a linearly in λ, and the Lennard-Jones intermolecular interactions were turned off with a "soft-core" interaction,[36] which smoothes out the infinity produced by the r^{-12} term as $\lambda \rightarrow 0$. This pathway was previously shown to be more efficient than pathway linear in λ or powers of λ. We used a total of 21 states along the Coulomb pathway and 41 states along the Lennard-Jones pathway (both numbers

TABLE III. Sample number dependence on the variance and bias of free energy calculations. The first column is n, the number of samples pulled from the 100 000 sample set in Table I used to estimate the free energy. The second column is b, the number of bootstrap samples used to compute the average and variance. Thus, all work values are used, on average, an equal number of times in each row. The third and eight columns (ΔF) are the average of BAR or EXP on n samples selected from the full set b times. The fourth and ninth columns are the analytic variance estimates from the full set (Var_A), and the sixth and eleventh columns are analytic bias estimates (Bias_A). The fifth and tenth columns are the bootstrap variance estimates (Var_B), and the seventh and twelfth columns are the bootstrap bias estimates (Bias_B). Close agreement between the two in the limit of large n indicates that the analytical estimates are accurate.

		BAR					Forward EXP				
n	b	ΔF	Var_A	Var_B	Bias_A	Bias_B	ΔF	Var_A	Var_B	Bias_A	Bias_B
100 000	1 000	2.67	0.08	0.08	0.004	0.003	2.68	0.26	0.28	0.03	0.07
80 000	1 250	2.67	0.09	0.09	0.01	0.004	2.70	0.30	0.33	0.07	0.09
50 000	2 000	2.67	0.11	0.11	0.01	0.003	2.73	0.37	0.43	0.12	0.12
25 000	4 000	2.68	0.16	0.16	0.02	0.01	2.88	0.53	0.66	0.24	0.27
20 000	5 000	2.68	0.18	0.18	0.02	0.02	2.97	0.59	0.75	0.29	0.36
10 000	10 000	2.71	0.25	0.26	0.04	0.04	3.36	0.84	1.23	0.59	0.75
8 000	12 500	2.72	0.28	0.30	0.05	0.05	3.57	0.93	1.48	0.74	0.96
5 000	20 000	2.74	0.36	0.38	0.08	0.08	4.12	1.18	2.03	1.18	1.51
2 500	40 000	2.80	0.51	0.52	0.17	0.14	5.50	1.67	3.11	2.35	2.89
2 000	50 000	2.83	0.57	0.57	0.21	0.17	6.11	1.87	3.55	2.94	3.50
1 000	100 000	2.97	0.80	0.72	0.42	0.31	8.63	2.64	5.17	5.89	6.01
800	125 000	3.03	0.90	0.75	0.52	0.36	9.66	2.95	5.78	7.36	7.05
500	200 000	3.15	1.14	0.79	0.84	0.49	12.38	3.73	7.36	11.77	9.77
250	400 000	3.32	1.61	0.77	1.67	0.65	17.64	5.28	10.03	23.55	15.03
200	500 000	3.37	1.80	0.75	2.09	0.70	19.69	5.90	10.90	29.44	17.08
100	1 000 000	3.48	2.54	0.65	4.18	0.82	27.12	8.35	13.64	58.87	24.51

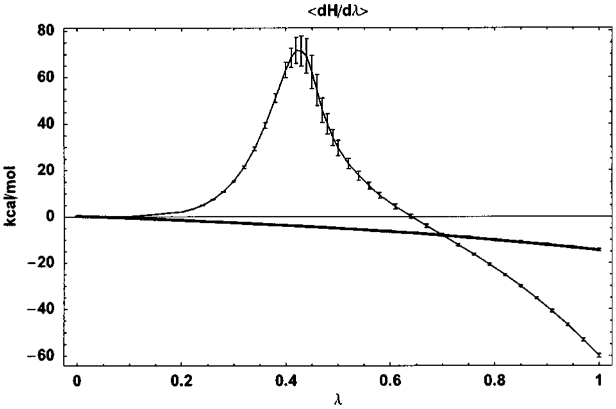

FIG. 10. $\langle dH/d\lambda\rangle_C$ (thick line) and $\langle dH/d\lambda\rangle_{LJ}$ (thin line) as a function of λ for 3-methylindole, the small molecule amino acid side chain analog of tryptophan, using the OPLS-AA force field. Uncertainties are multiplied by 5 in order to guide the eye, as they would otherwise be difficult to distinguish from the lines themselves. The coulombic decharging pathway is very smooth, and is very amenable to both TI and BAR. The LJ desovlation pathway, on the other hand, has high curvature, and the free energy difference can be sampled much more efficiently with BAR.

TABLE IV. Free energies of hydration of 3-methylindole, in kcal/mol. Methodology and further description of simulations are as described in a previous study (Ref. 15). States is total number of intermediate plus end states simulated. Columns labeled Coul and LJ are the free energies of the charging and uncharged solvation processes, respectively.

Method	States	Coul	LJ	Total
TI	61	-5.77 ± 0.01	2.08 ± 0.02	-3.69 ± 0.03
TI	8	-6.11 ± 0.04	3.17 ± 0.25	-2.94 ± 0.25
FEP ($A\rightarrow0$)	8	-5.57 ± 0.07	2.13 ± 0.07	-3.44 ± 0.09
FEP ($0\rightarrow A$)	8	-5.68 ± 0.04	-0.32 ± 0.18	-6.01 ± 0.19
FEP (Ave)	8	-5.63 ± 0.04	0.90 ± 0.10	-4.72 ± 0.11
BAR	8	-5.69 ± 0.01	2.01 ± 0.02	-3.68 ± 0.02

including end states), and sampled five runs of 1 ns each using molecular dynamics to compute the free energy using TI.

Taking the OPLS-AA model of the molecule (one of several force field parametrizations examined this previous study), the final value of the solvation energy was -3.69 ± 0.03 kcal/mol.[37] We neglect any long-range correction for Lennard-Jones attractive energy,[15] which is independent of the free energy method used. Analysis after the fact revealed that approximately the 61 intermediate states was required to eliminate bias larger than the uncertainty using TI. For example, using TI with only eight total states (spaced to be sample areas of high curvature more than areas of low curvature),[38] we obtain a free energy of -2.94 ± 0.26 kcal/mol, significantly different than our previous, more accurate, estimate. The importance of a low curvature is especially evident in this very realistic example. The charging free energy using only three points is -6.11 ± 0.04 versus -5.77 ± 0.02 for 21 points over a relatively flat curve, a much smaller error than in the free energy the desolvation of the uncharged molecule, where we obtain 3.17 ± 0.25 using seven states versus 2.08 ± 0.04 using 41 states. In the desolvation case, this is a free energy over a function with much higher curvature (see Fig. 10). We note that using Simpson's rule or Gaussian quadrature to integrate the Coulombic part yields -5.76 ± 0.05, within uncertainty of the correct answer, however, it may be somewhat of a coincidence in this case. For sampling along smoother curves where equal spacing is reasonable, therefore, an increased order algorithm like Simpson's rule may sometimes prove useful.

How do BAR and forward and reverse EXP compare in efficiency and accuracy? Using exactly these same eight states, insertion EXP gives a free energy of -3.44 ± 0.09, whereas deletion EXP yields -6.01 ± 0.18. BAR gives -3.68 ± 0.02 kcal/mol, almost indistinguishable to the TI result achieved with more than seven times as much simulation data. Although there may be some cancellation of error to achieve results so close to the original TI result, the results from BAR are still clearly better for the simulation time used than any other method. These numbers are presented for comparison in Table IV. In this case, BAR provides a clear advantage to TI and to either version of EXP.

V. DISCUSSION

It appears that in many realistic atomistic situations, BAR estimate is significantly better than EXP. For near-equilibrium measurements, this difference may be marginal, but for larger differences in free energy and lower phase space overlaps, this advantage can become appreciable, as seen in all three atomistic examples presented here. However, the example of nested Gaussians demonstrates that occasionally obtaining the free energy using EXP can be more efficient. Care must be taken to identify such situations, and more study may need to be done to understand when this can occur. Additionally, we have seen that in toy problems, TI can be the best behaved estimator. For charging free energies, which have a relative smooth profile, TI may compete with BAR for efficiency. However, in realistic chemical and biomolecular situations, pathways with moderate amounts of curvature can make TI very inefficient. In some cases, EXP may end up being slightly better than other methods, but these cases may need to be investigated more closely; in any case, BAR appeared to be almost as good in this particular toy problem.

There are a number of additional questions, beyond the scope of this study, that must be addressed in order to effectively apply these methods to quantitative calculation of free energies. First, although we have presented formulas for estimating the variance and bias due to limited sampling of the free energy methods, these estimates themselves are subject to sample size bias and thus cannot always be depended upon to provide accurate uncertainty estimates. We note that although the free energy depends on logarithms of exponential averages, the variance of the free energy depends on averages of exponentials and are therefore much more dependent on good sampling than the free energy estimates themselves.

Recent work[14] has used bias estimates to correct the free energy from EXP. However, since bias correction is inherently error prone, it remains unclear what the advantage may actually be. Of course, for extremely large energy differences, this type of block averaging bias correction could be applied to BAR as well.

Another important question for applications is the nature of the distributions $P_F(W)$ and $P_R(W)$. There is, of course, only one independent distribution, as the two are exactly related by Eq. (4). This distribution will depend on the underlying process and cannot usually be computed or currently even estimated *a priori*. However, a better understanding of what features of the switching process lead to narrow distributions in the work could greatly facilitate the design of ideal pathways and intermediates between the end states being studied. Some work has been performed at understanding this difference in the case of EXP,[10,11] but the problem is still unresolved in any sort of general case more complicated than the toy problems presented here.

Additionally, the fact that BAR is the asymptotically unbiased estimator with the minimum variance does not necessarily guarantee that is the best estimator by all measures. For example, in many cases, it is possible to find an estimator that is more inherently biased, but has significantly smaller variances, resulting in a much smaller mean squared error. Statistically, the most important figure of merit is the mean square error, the sum of the variance and the square of the bias. Preliminary evidence indicates, for example, that the probability distributions $P_F(W)$ and $P_R(W)$ can be smoothed by such methods as convolution with a kernel and used in Eq. (4) to obtain significantly smaller mean squared errors in some cases.[39] The choice of a path is also vitally important for efficiently estimating large free energies. In a previous studies,[15,36] and in the toy problems in this study, we have seen that improved choice of path significantly decreases the bias and variance of all methods. However, there are many more questions left to be asked about improved pathways for given biomolecular simulations, and optimizing this path is a subject that has not been sufficiently addressed in the literature.

VI. CONCLUSIONS

We have presented a comprehensive comparison of the major methods for free energy computation, using theory, toy problems, and atomistic simulations. Most importantly, the Bennett acceptance ratio appears to be significantly more efficient than other methods such as exponential averaging and thermodynamic integration in realistic atomistic simulations, although in some interesting special cases these other methods may be somewhat better.

We have found in a variety of sample atomistic simulations relevant to biomolecular and chemical studies that BAR is always better than EXP in computing free energies, frequently significantly better when the free energies to be measured are moderately large compared with kT. Unless the curvature in $\langle dH/d\lambda\rangle$ is extremely low (which can be the case in electrostatic charging), BAR is much more efficient than TI as well. BAR is actually simpler to use than TI, as it does not require code to compute analytical derivatives to be implemented, merely that the potential energy be evaluated for the mutable part of the system at a variety of λ values. Because of this, BAR (or an equivalent method like WHAM) appears to be recommended as the default method for computing free energies of biomolecular and chemical processes, especially when the overlap in phase space between end states is particularly small.

ACKNOWLEDGMENTS

The authors thank Eric Bair and Giles Hooker for general advice on statistical methods, and Chris Jarzynski and Xiao-li Meng for useful discussion. M.R.S. acknowledges funding from the Fannie and John Hertz foundation.

APPENDIX: BIAS OF BAR

This derivation will follow loosely from a more general derivation provided in Mardia *et al.*[40] Let us express the free energy estimate obtained from a finite number of samples n_f and n_r from Eq. (1) as ΔF_n, and the infinite estimate as ΔF. The bias from finite sampling will be $\langle \Delta F_n - \Delta F\rangle$, where the average is over all possible realizations of the n_f and n_r samples.

We will express the finite n_f, n_r expression for the derivative of the log likelihood as $L_n(W)$, with $L(W)$ defined as the infinite n limit. By construction, $L_n(\Delta F_n)=0$, and $L(\Delta F)=0$.

Because BAR is an asymptotically efficient estimator of the free energy,[25] the distribution of estimates of the free energy at large n tends to a normal distribution with variance as given in Eq. (3). Following Mardia *et al.*[40] we can estimate to first order

$$\Delta F_n - \Delta F \approx \mathrm{var}(\Delta F) L_n(\Delta F) + O(n^{-2}). \tag{A1}$$

We then express $L_n(\Delta F_n)$ as a Taylor series in ΔF_n,

$$\begin{aligned} 0 = L_n(\Delta F_n) = L_n(\Delta F) + (\Delta F_n - \Delta F)L_n'(\Delta F) + \tfrac{1}{2}(\Delta F_n \\ - \Delta F)^2 L_n''(\Delta F) \\ + O((\Delta F_n - \Delta F)^3). \end{aligned} \tag{A2}$$

Solving for $\Delta F_n - \Delta F$, we get

$$\begin{aligned} \Delta F_n - \Delta F \approx \mathrm{var}(\Delta F)\big((\Delta F_n - \Delta F)L_n'(\Delta F) + \tfrac{1}{2}(\Delta F_n \\ - \Delta F)^2 L_n''(\Delta F)\big) + O(n^{-2}). \end{aligned} \tag{A3}$$

We could solve directly for $\Delta F_n - \Delta F$, but this will make our average more difficult to compute. So we will estimate again by substituting $\mathrm{var}(\Delta F)L_n(\Delta F)$ for $\Delta F_n - \Delta F$ as per Eq. (A1), yielding

$$\begin{aligned} \Delta F_n - \Delta F \approx \mathrm{var}(\Delta F)\big(\mathrm{var}(\Delta F)L_n(\Delta F)L_n'(\Delta F) \\ + \tfrac{1}{2}(\Delta F_n - \Delta F)^2 L_n'' df\big) + O(n^{-2}). \end{aligned} \tag{A4}$$

Taking the expectation value of both sides over all realizations of n_f and n_r, $\mathrm{var}(\Delta F)$ is a constant, and will go through the expectation operator. Additionally, the covariance of $(\Delta F_n - \Delta F)^2$ and $L_n''(\Delta F)$ will be $O(n^{-2})$, so we can take the expectations separately and multiply, yielding

$$\langle \Delta F_n - \Delta F\rangle \approx \mathrm{var}(\Delta F)\Big(\mathrm{var}(\Delta F)\langle L_n(\Delta F)L_n'(\Delta F)\rangle + \tfrac{1}{2}\mathrm{var}(\Delta F)\langle L_n'' df\rangle\Big) + O(n^{-2}), \tag{A5}$$

$$\langle \Delta F_n - \Delta F\rangle \approx \tfrac{1}{2}\mathrm{var}(\Delta F)^2\big(2\langle L_n(\Delta F)L_n'(\Delta F)\rangle + \langle L_n'' df\rangle\big) + O(n^{-2}), \tag{A6}$$

$\mathrm{var}(\Delta F)$ is of order n^{-1} as is $\langle L_n''(\Delta F)\rangle$. Since $L_n(\Delta F)=0 + O(n^{-1})$, the dominant term in the $L_n(\Delta F)L_n'(\Delta F)$ term is the covariance of $L_n(\Delta F)L_n'(\Delta F)$, which is proportional n^{-1}. We can therefore rewrite this as

$$\langle \Delta F_n - \Delta F\rangle \approx \tfrac{1}{2}\mathrm{var}(\Delta F)[\mathrm{var}(\Delta F)(2\langle L_n(\Delta F)L_n'(\Delta F)\rangle + \langle L_n''(\Delta F)\rangle)] + O(n^{-2}), \tag{A7}$$

which can be written as

$$\langle \Delta F_n - \Delta F\rangle \approx \tfrac{1}{2}\mathrm{var}(\Delta F)K, \tag{A8}$$

where

$$K = \mathrm{var}(\Delta F)(2\langle L_n(\Delta F)L_n'(\Delta F)\rangle + \langle L_n''(\Delta F)\rangle) \tag{A9}$$

and where K is of $O(n^0)$ by the analysis above. We then note that in the case of BAR,

$$L_n(\Delta F) = \sum_{n_f} \frac{1}{1+\exp(\beta(M+W-\Delta F))} - \sum_{n_f} \frac{1}{1+\exp(-\beta(M+W-\Delta F))}, \tag{A10}$$

$$L_n'(\Delta F) = \sum_{n_r,n_f} \frac{1}{2+2\cosh(\beta(M+W-\Delta F))}, \tag{A11}$$

$$L_n''(\Delta F) = \sum_{n_r,n_f} \frac{\sinh(M+W-\Delta F)}{2+2\cosh(\beta(M+W-\Delta F))^2}. \tag{A12}$$

We note that $L_n(\Delta F)$ and $L_n''(\Delta F)$ is antisymmetric around $M-\Delta F$, whereas $L_n'(\Delta F)$ is symmetric. If $n_f = n_r$ and therefore $M=0$, then if $P_R(W)$ and $P_F(W)$ are symmetric (as in the case of Gaussian distributions), K and therefore the leading term in the bias will be zero.

[1] *Free Energy Calculations in Rational Drug Design*, edited by M. R. Reddy and M. D. Erion (Kluwer Academic, Dordrecht, MA, 2001).
[2] A. Gelman and X. L. Meng, Stat. Sci. **13**, 163 (1998).
[3] T. Schafer and E. V. Shuryak, Rev. Mod. Phys. **70**, 323 (1998).
[4] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante, Science **296**, 1832 (2002).
[5] G. Hummer, Mol. Simul. **28**, 81 (2002).
[6] D. A. Hendrix and C. Jarzynski, J. Chem. Phys. **114**, 5974 (2001).
[7] D. A. Kofke and P. T. Cummings, Fluid Phase Equilib. **151**, 41 (1998).
[8] D. M. Zuckerman and T. B. Woolf, Phys. Rev. Lett. **89**, 602 (2002).
[9] D. M. Zuckerman and T. B. Woolf, Chem. Phys. Lett. **351**, 445 (2002).
[10] N. D. Lu and D. A. Kofke, J. Chem. Phys. **114**, 7303 (2001).
[11] N. D. Lu and D. A. Kofke, J. Chem. Phys. **115**, 6866 (2001).
[12] N. D. Lu and D. A. Kofke, J. Chem. Phys. **111**, 4414 (1999).
[13] N. D. Lu, J. K. Singh, and D. A. Kofke, J. Chem. Phys. **118**, 2977 (2003).
[14] F. M. Ytreberg and D. M. Zuckerman, J. Chem. Phys. **120**, 10876 (2004).
[15] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande, J. Chem. Phys. **119**, 5740 (2003).
[16] T. P. Straatsma and J. A. McCammon, J. Chem. Phys. **95**, 1175 (1991).
[17] D. A. Pearlman and P. A. Kollman, J. Chem. Phys. **91**, 7831 (1989).
[18] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[19] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[20] G. E. Crooks, Phys. Rev. E **61**, 2361 (2000).
[21] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[22] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).
[23] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[24] M. Souaille and B. Roux, Comput. Phys. Commun. **135**, 40 (2001).
[25] E. L. Lehmann and G. Casella, *Theory of Point Estimation* (Springer, New York, 1998).
[26] D. A. Kofke, J. Chem. Phys. **117**, 6911 (2002).
[27] C. Oostenbrink and W. F. van Gunsteren, Proteins **54**, 237 (2004).
[28] X. J. Kong and C. L. Brooks, J. Chem. Phys. **105**, 2414 (1996).
[29] S. Banba, Z. Y. Guo, and C. L. Brooks, J. Phys. Chem. B **104**, 6903 (2000).
[30] H. Resat and M. Mezei, J. Chem. Phys. **99**, 6052 (1993).
[31] P. E. Smith and W. F. Van Gunsteren, J. Chem. Phys. **100**, 577 (1994).
[32] W. L. Jorgensen and C. Ravimohan, J. Chem. Phys. **83**, 3050 (1985).
[33] W. L. Jorgensen, J. D. Madura, and C. J. Swenson, J. Am. Chem. Soc. **106**, 6638 (1984).
[34] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[35] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, Boca Raton, FL, 1993).
[36] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren, Chem. Phys. Lett. **222**, 529 (1994).
[37] In the original paper, an uncertainty of 0.05 kcal/mol was reported as an average of the uncertainty over five simulations. A more accurate uncertainty of the average of these five simulations would therefore be the original uncertainty divided by $\sqrt{5-1}=2$ or 0.03, as reported here.
[38] Specifically, we use the states along the published Lennard-Jones softcore pathway (Ref. 15), with $\lambda=0$, 0.32, 0.42, 0.50, 0.64, and 1.0, and the states along the linear Coulombic pathway $\lambda=0$, 0.5, and 1.
[39] Authors (unpublished).
[40] K. V. Mardia, H. R. Southworth, and C. C. Taylor, J. Stat. Plan. Infer. **76**, 31 (1999).