Appropriate methods to combine forward and reverse free-energy perturbation averages

Nandou Lu,[a] Jayant K. Singh, and David A. Kofke
Department of Chemical Engineering, University at Buffalo, The State University of New York, Buffalo, New York 14260-4200

(Received 29 July 2002; accepted 19 November 2002)

We consider the accuracy of several methods for combining forward and reverse free-energy perturbation averages for two systems (labeled 0 and 1). The practice of direct averaging of these measurements is argued as not reliable. Instead, methods are considered of the form $\beta(A_1-A_0) = -\ln[\langle w(u)\exp(-\beta u/2)\rangle_0 / \langle w(u)\exp(+\beta u/2)\rangle_1]$, where A is the free energy, $\beta=1/kT$ is the reciprocal temperature, $u=U_1-U_0$ is the difference in configurational energy, $w(u)$ is a weighting function, and the angle brackets indicate an ensemble average performed on the system indicated by the subscript. Choices are considered in which $w(u)=1$ and $1/\cosh[(u-C)/2]$; the latter being Bennett's method where C is a parameter that can be selected arbitrarily, and may be used to optimize the precision of the calculation. We examine the methods in several applications: calculation of the pressure of a square-well fluid by perturbing the volume, the chemical potential of a high-density Lennard-Jones system, and the chemical potential of a model for water. We find that the approaches based on Bennett's method weighting are very effective at ensuring an accurate result (one in which the systematic error arising from inadequate sampling is less than the estimated confidence limits), and that even the selection $w(u)=1$ offers marked improvement over comparable methods. We suggest that Bennett's method is underappreciated, and the benefits it offers for improved precision and (especially) accuracy are substantial, and therefore it should be more widely used. © *2003 American Institute of Physics.* [DOI: 10.1063/1.1537241]

I. INTRODUCTION

Knowledge of free energy is central to understanding of the behavior of many thermophysical and chemical processes, including phase and reaction equilibria, molecular association, thermodynamic stability, rare-event kinetics, binding affinity, and so on.[1] Free-energy calculations almost always involve computation of free-energy *differences*, measured between two systems that differ in thermodynamic state, Hamiltonian, or in the form of a constraint. The free-energy difference can be computed in many ways, some closely related. Categories include free-energy perturbation (FEP) and other nonequilibrium methods,[2] thermodynamic integration,[3] parameter-hopping,[4] histograms,[5–7] and adiabatic switching.[8] The options are many, yet FEP remains a popular choice because it is very simple to apply, and in its basic form it involves no extra calculations on systems that are otherwise of no intrinsic interest. However, it is prone to inaccuracy, and if applied carelessly it can give results that are highly reproducible but incorrect.[9,10] Symptoms of the problem are seen by performing the calculation in two directions, arbitrarily designated ''forward'' and ''reverse.'' The two calculations should in principle yield the same result, but usually they differ.[11]

A way to improve accuracy involves staging the FEP calculation so that the overall difference is computed as the sum of two or more smaller differences. Popular two-stage versions[10,12] of this approach include umbrella sampling[13] and Bennett's method.[5] Because multistage methods have FEP as their elementary component, they can suffer from the same problems as single-stage FEP if they are not used wisely. One practice calls for using as many stages as possible to minimize the observed forward/reverse asymmetry in each. This procedure can yield a correct overall difference, but it is very inefficient.[14] Another common practice simply reports the free-energy difference of each stage as the average of the forward and reverse results,[12,15,16] based on the assumption that the systematic errors in these two directions of calculation are of the same magnitude but opposite sign. We have argued that such an assumption is not reliable in general, and consequently simple averaging is not a good practice because it is liable to yield an incorrect result.[14,17,18]

The working equation for a single-stage FEP calculation can usually be put in the form

$$e^{-\beta(A_1-A_0)}=\langle e^{-\beta(U_1-U_0)}\rangle_0. \tag{1}$$

The '0' and '1' subscripts indicate the two systems of interest; A is the Helmholtz free energy, U is the configurational energy, and $\beta=1/kT$ with k Boltzmann's constant, and T the absolute temperature. The angle brackets indicate an ensemble average performed on the ''0'' system, which we call the *reference*; the ''1'' system we call the *target*. Simulation is performed to sample configuration space with a limiting distribution proportional to $e^{-\beta U_0}$ (for a canonical NVT ensemble). Selection of one or the other system as the reference gives rise to the forward and reverse implementations of the FEP calculation.

[a]Present address: Department of Physiology, Johns Hopkins University, Baltimore, MD 21205-2185.

If one takes the arithmetic average of the free-energy differences from the forward and reverse calculations to obtain an ''improved'' estimate for the free-energy difference, in effect the following formula is being used for the measurement

$$e^{-\beta(A_1-A_0)}=\left(\frac{\langle e^{-\beta(U_1-U_0)}\rangle_0}{\langle e^{-\beta(U_0-U_1)}\rangle_1}\right)^{1/2}. \tag{2}$$

The inappropriateness of this formula can be illustrated through an extreme but nevertheless relevant example. The chemical potential of the hard-sphere model is computed as a FEP between two systems differing in the presence of a single ''test'' sphere (or more precisely, in one system the test sphere does not interact with the others, and in the other system it interacts as a regular sphere). If the ''0'' system is taken as the one in which the test sphere does not interact with the others, then the numerator in Eq. (2) is simply the fraction of the configurations in which the test sphere by chance does not overlap one of the other spheres (the quantity being averaged is zero in the case of overlap, and unity for nonoverlap). A simulation average of the denominator will always give a value of unity, because in this case the test sphere samples configurations in which it interacts with the others, and thus it will never overlap another sphere (but if it were to sample one of these zero-probability configurations, the contribution to the average would be infinite, so something significant is being missed). Consequently, in effect this forward/reverse averaging scheme computes the chemical potential according to

$$e^{-\beta(A_1-A_0)}=\langle e^{-\beta(U_1-U_0)}\rangle_0^{1/2}, \tag{3}$$

which of course is simply incorrect. The error is so obvious that no one makes the mistake of using forward/reverse averaging in this application. But in more complex systems the problem is much more subtle, and averaging is routinely applied without realizing that the same type of error is being introduced. Part of the problem is the great reproducibility of the incorrect result. The calculation can have good precision but poor accuracy.

The remainder of this paper is organized as follows. In the next section we develop an alternative approach to combining forward and reverse FEP averages, one which does not exhibit the deficiencies of simple averaging. The method works through an intermediate system that is formulated such that its configurations are important to both the 0 and 1 systems. We generalize the approach, and then in Sec. III show how it connects to Bennett's method for free-energy calculations. In Sec. IV we demonstrate the method with two types of applications, one in which a volume perturbation is used to calculate the pressure, and another for calculation of the chemical potential. We conclude in Sec. V.

II. OVERLAP SAMPLING

We have argued that there are situations in which the better practice is not to mix the results of forward and reverse averaging, but to use only one of them.[9,17] Then the direction that gives the correct result is the one that perturbs from the system of higher entropy to the system of lower entropy.[14] For the chemical potential calculation, this is the direction in which the test sphere is ''inserted,'' i.e., goes from noninteracting to interacting. We adopt this picture when generalizing to other perturbations, and say that the higher-to-lower-entropy direction is ''insertion,'' while ''deletion'' is the direction from lower to higher entropy. We have hypothesized that FEP should always be performed in the insertion direction, and that deletion calculations are likely to be very inaccurate.[14,17,18]

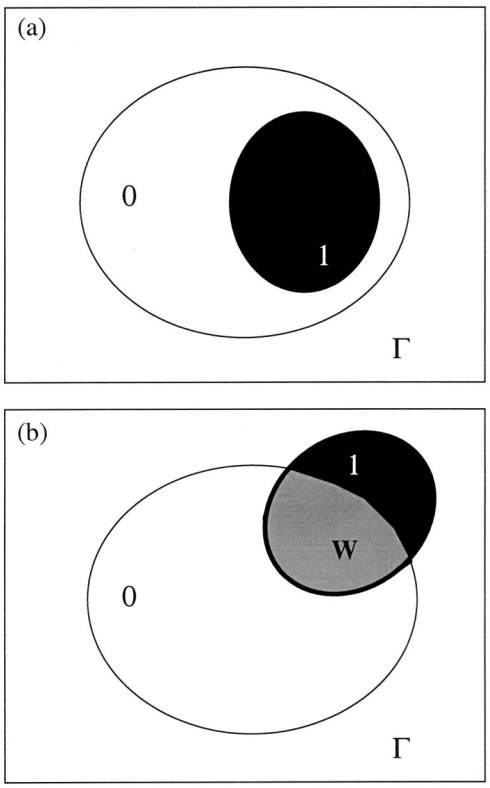

FIG. 1. A schematic depiction of the configuration space Γ. The oval shapes show (abstractly) the important configurations for the 0 and 1 systems, as indicated. (a) The important configurations of the lower-entropy system is wholly contained in the higher-entropy system; (b) the important configurations of the 0 and 1 systems overlap, but one does not form a subset of the other; the intermediate system *W* is formulated such that its important configurations are formed from the overlap of the configurations important to 0 and 1.

A schematic is helpful to illustrate the asymmetry of the insertion/deletion calculations.[10,17] Figure 1(a) presents a cartoon depiction of the configuration space. Each point in the two-dimensional square represents a configuration of the system, a point in the $3N$-dimensional configuration space (considering a three-dimensional system of N monatomic particles, and ignoring momentum). Configurations important to target and reference systems are enclosed in the simple oval shapes (again, a highly schematic depiction). The higher-entropy system is the one with the larger set of important configurations. For a FEP calculation to yield an accurate result, representative configurations of both important regions must be sampled in the simulation. In the insertion calculation, sampling is performed among the configurations in the large oval, and contributions to the average are made whenever configurations in the small oval are encountered

by chance (e.g., the test sphere happens not to overlap another sphere). The barrier to sampling these configurations is entropic; it is a matter of when they are encountered by chance. In the deletion calculation, sampling is performed among the configurations in the small oval. The barrier to sampling the other configurations is energetic; moreover, the contribution from the unsampled configurations increases inversely with sampling probability (e.g., an infinite contribution from the zero-probability configurations in which the interacting test-sphere overlaps another sphere).

The argument for using FEP in only the insertion direction assumes that the important configuration spaces relate to each other as shown in Fig. 1(a). That is, configurations important to the low-entropy system form a *subset* of the configurations important to the high-entropy system. This situation holds for the hard-sphere chemical-potential calculation, and is probably true for many other perturbation systems as well. However, it is difficult to know for sure whether two systems relate this way. Even in situations where the subset relation holds, application of FEP is complicated by the need to know which is the higher-entropy and which is the lower-entropy system.[14] It would be valuable if one could proceed with a calculation without having to be so careful in analyzing the nature of the two systems.

To motivate the direction we take, consider the configuration-space diagram presented in Fig. 1(b). Here the two systems of interest do not exhibit the subset relation for their set of important configurations. Consequently there is no single-stage FEP calculation that will yield an accurate result for the difference in free energies of these systems. However, their important configurations do overlap. Of course, the configurations in the overlap region are a subset of the important configurations of both systems. Consequently we can expect a single-stage FEP calculation to be effective in calculating the free-energy difference between the reference and a system W in which only these overlap configurations are important. We can reasonably expect to construct such a system by defining its Hamiltonian as the average of the Hamiltonians of the two systems of interest. Considering just the configurational energy, we define

$$U_W = \tfrac{1}{2}(U_1 + U_0). \tag{4}$$

We then compute the overall free-energy difference by staging two intermediate FEP calculations, $0 \rightarrow W$ and $1 \rightarrow W$, thus

$$e^{-\beta(A_1 - A_0)} = \frac{e^{-\beta(A_W - A_0)}}{e^{-\beta(A_W - A_1)}}. \tag{5}$$

Then combining Eqs. (1), (4), and (5) we have the final working formula

$$e^{-\beta(A_1 - A_0)} = \frac{\langle e^{-\beta(U_1 - U_0)/2} \rangle_0}{\langle e^{-\beta(U_0 - U_1)/2} \rangle_1}. \tag{6}$$

The proposed formula has a strong similarity to the flawed formula given in Eq. (2), but the differences are crucial. To illustrate, we need consider only the hard-sphere chemical potential calculation. In Eq. (6) the numerator again will yield the fraction of configurations in which the test sphere by chance does not overlap another sphere, while the denominator is again unity. But the factor of $\frac{1}{2}$ has already been applied, so when the averages are combined the correct result is recovered.

We can apply an analysis of the accuracy to better understand the behavior of this calculation, and to investigate whether the W system can be defined differently to improve the accuracy further. To this end let us generalize the definition of Eq. (4) by defining a weighting function $w(u)$ in terms of the energy difference $u \equiv U_1 - U_0$ such that $U_W = -kT \ln w(u) + (U_1 + U_0)/2$; it is simple to show that the free-energy difference can be given generally

$$e^{-\beta(A_1 - A_0)} = \frac{\langle w(u) e^{-\beta u/2} \rangle_0}{\langle w(u) e^{+\beta u/2} \rangle_1} \tag{7}$$

for which Eq. (6) obviously arises as the special case $w(u) \equiv 1$.

We define accuracy as the difference between the most likely outcome (the mode of the distribution of measured values obtained by many independent ensemble averages) and the correct outcome. For evaluating the reliability of FEP calculation results, accuracy is of greater concern than the precision, which is analyzed in terms of the variance of the distribution of measured values. If the result of a FEP calculation is inaccurate, usually the measurement of the variance is inaccurate too, indicating a smaller error—greater confidence—than warranted. Bennett's method (discussed below) follows a line similar to the one we have taken so far, except it optimizes the free-energy calculation by minimizing the variance with respect to $w(u)$. We will consider instead an optimization of the accuracy.

FEP averages can be written in terms of one-dimensional integrals of the energy difference u. Distribution functions $f(u)$ and $g(u)$ are defined as the normalized probability densities for observing the energy difference u when simulating the 0 and 1 systems, respectively. These distributions are related[6,19]

$$g(u) e^{\beta u} = f(u) e^{\beta(A_1 - A_0)}. \tag{8}$$

To model FEP inaccuracy, we assume that the simulation samples the f and/or g distributions perfectly, but only between the maximum and minimum values of u encountered in the simulation.[17] Inaccuracy arises from the neglect of the contributions coming from the tails of the distribution. As the simulation proceeds, the extreme values of u move further out into the tails, the neglected region becomes smaller, and the accuracy improves.

For single-direction, single-stage FEP calculations, we have shown that this (fractional) error is given simply by the area under the conjugate distribution above or below the limit energy. For example, if sampling the g distribution, some maximum value u_g is encountered in a finite-length simulation. Then the fractional error in the measurement is the integral of f for $u > u_g$. If f and g do not have a large amount of overlap, this error is much or all of the area under f, and is substantial. In contrast, the error from poor sampling of the tails when applying Eq. (6) is, approximately

$$\beta(A_1-A_0)_{\mathrm{err}}=\frac{\int_{-\infty}^{u_f}duw(u)[f(u)g(u)]^{1/2}-\int_{u_g}^{\infty}duw(u)[f(u)g(u)]^{1/2}}{\int_{-\infty}^{\infty}duw(u)[f(u)g(u)]^{1/2}}. \tag{9}$$

Thus the error is instead given in terms of the area of the *product* of the distributions. This is an important distinction from the one-way average. Naturally, one of the distributions will always be small beyond its most extremely sampled energy so the error integrals have an automatic tendency to be small. Moreover, Eq. (9) shows that errors from inadequate sampling in the two ensemble averages will tend to cancel each other [which is also the tendency when using Eq. (2)]. Nevertheless, the method does have limitations, and like all FEP techniques, it cannot produce a reliable result if the distributions do not have some degree of overlap (insufficient overlap leads the denominator in the error term to become small).

Within this model, the simplest way to minimize the inaccuracy is to set the weighting function $w(u)$ to zero for $u<u_f$ and $u>u_g$. Then the error vanishes completely. However, the only way to do this is to investigate the energy distributions, which puts the method into a different class of techniques. Alternatively we can apply a Gaussian-like weighting function that emphasizes the contributions from the region of overlap and diminishes that from the tails. This in fact is what is done by Bennett's method.[5]

III. BENNETT'S METHOD

Beginning from an equation very similar to Eq. (7), Bennett selected $w(u)$ to minimize the variance of the FEP average. Bennett is able to take the analysis to completion, specifying exactly the form of $w(u)$, because the weighting function does not influence the sampling. Optimization of $w(u)$ for methods such as umbrella sampling, where the weighting function affects the sampling of configurations, is much more difficult and cannot be done with the generality of Bennett's optimization.

Without using the same language as employed here, Bennett also addressed the issue of accuracy of FEP calculations, where he considers the ''small sample regime.'' He points out that his algorithm provides a useful estimate of the free-energy difference in cases in which the tails of the distributions (using the present language) are not well sampled. He also notes that in this case the confidence limits on the average are not adequately represented by the spread among independent estimates, i.e., accuracy is more of a concern than precision.

We can make the connection to Eq. (7) by recognizing that in Bennett's method the weighting function is given by a hyperbolic secant function:

$$w(u)=1/\cosh[\beta(u-C)/2], \tag{10}$$

where C is a constant selected to minimize the variance of the free-energy measurement, which prescribes that it equal the free-energy difference being measured: $C=\Delta A$. This choice puts the zero of the cosh argument at the value of u where f and g are equal, i.e., where they cross each other [from Eq. (8), $f=g$ when $u=\Delta A$]. This of course is the region of greatest overlap and, appropriately, Bennett's method gives it the maximum weight. An illustration of the f and g distributions and the Bennett weighting is given in Fig. 2.

As is well known, Bennett's method can be applied using any value of C, and in principle will give a correct result regardless. Selection of the optimal value requires knowledge of the free energy being measured, which implies iteration, or more simply that a survey of averages for different C values be taken, and the optimum selected self-consistently. Bennett's requirement to tune a single scalar quantity is a modest imposition. Still, this complication seems to hinder the broad application of the technique, and among the quick-and-dirty approaches to FEP calculations, the ill-advised forward-and-reverse averaging [Eq. (2)] sees much wider use. Thus the advantage of Eq. (6), which combines forward and reverse averages in a much more appropriate way, is that it can be applied with the same effort that is used to collect other ensemble averages—it abandons attempts to optimize for minimum variance and thereby removes any prescription for tweaking the implementation. It is likely that Eq. (10) will do even better, but the question is, how important is the selection of C to the quality of the result? Must we apply the full optimization routine to get a result that improves on single stage insertion, or on Eq. (6)? We are interested in this question from the standpoint of the accuracy of the calculation.

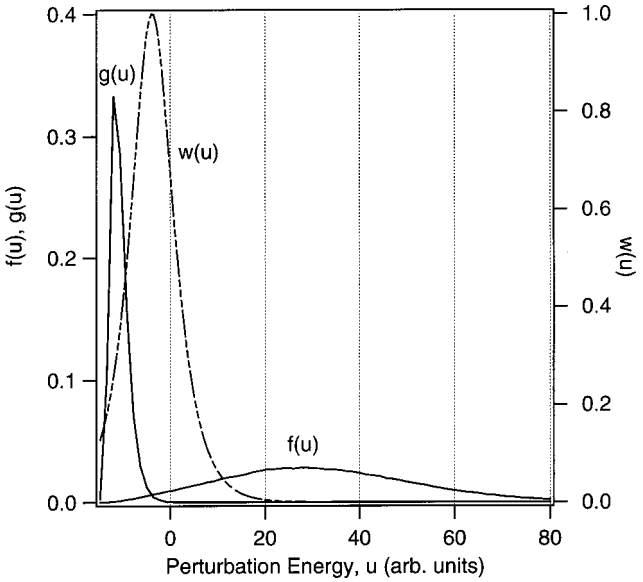

FIG. 2. Typical energy distributions $f(u)$, $g(u)$ for forward and reverse perturbations, superimposed with the optimal Bennett weighting function, $w(u)$. The peak in w is located at the value of u where $f=g$.

IV. APPLICATIONS

A. Calculating the pressure

Free-energy perturbation can be used as a method for calculation of the pressure. Harismiadis *et al.*[20] described a method of this type, which is based on the relation $\beta P = -(\partial \beta A/\partial V)_{T,N} \approx -[\beta A(V+\Delta V) - \beta A(V)]/\Delta V$, where P and V are the pressure and volume of the system, respectively. Thus we perform FEP calculations to evaluate the change in free energy with a small perturbation in the volume. In principle, we need to perform two simulations, one of the system of volume V perturbed to volume $V+\Delta V$, and a complementary one for a system of volume $V+\Delta V$ perturbed to volume V. We consider an approximation in which the $V+\Delta V \rightarrow V$ calculation is given by a simulation of a system of volume V perturbed to one of volume $V-\Delta V$. In this case the pressure is given by

$$\beta P \approx \frac{N}{V} + \frac{1}{\Delta V} \ln \left[\frac{\langle \exp[-\beta(U(V+\Delta V) - U(V))/2] \rangle_V}{\langle \exp[-\beta(U(V-\Delta V) - U(V))/2] \rangle_V} \right]. \tag{11}$$

Here, we have applied the usual scaling of coordinates, such that a change in the volume causes all molecule positions to be scaled proportionately, giving rise to the dependence of the configurational energy U on volume indicated by the formula. The transformation also gives rise to the additive ideal-gas contribution N/V.

We consider application to the square-well model, which is interesting because the FEP calculation is asymmetric. Perturbations that expand the system volume will not register the contribution to the pressure from the repulsive core. However, it will be effective in gauging the attractive contributions to the pressure, because it will cause molecules near the outer edge of the well separation to come apart, giving rise to a measurable change in the energy. Compression perturbations will measure the repulsive contribution, but will be less effective at getting the attractive part because fewer spheres will lie just outside the well cutoff and thus there will not be as much to sample.

In this manner we calculated the pressure of a $\lambda = 1.75$ square-well system, where λ is the diameter of the attractive well (all quantities and results are given in units of the repulsive core diameter σ and well depth ε). Simulations of $N = 256$ particles were performed in the canonical (NVT) ensemble. We selected conditions over a range of temperatures corresponding to saturated liquid and vapor phases, as reported by Vega *et al.*[21] Simulations proceeded beyond a period of equilibration for approximately 0.7×10^6 Monte Carlo cycles, with one volume perturbation in each direction (compression and expansion) attempted in each cycle. The free-energy volume change perturbation was 0.05%.

The data of Vega *et al.* were taken using Gibbs ensemble simulations, while our results are measured in independent NVT simulations of each phase. Consequently the comparison with the literature data is imperfect, a situation further complicated by recent data of del Rio *et al.*[22] which indicates some imperfections in the Vega *et al.* results. Our point in this study is to compare the performance of the methods under consideration. We use the literature data only to ensure the plausibility of our calculations. In the figure caption we describe how we examined possible sources of discrepancies in our calculation.

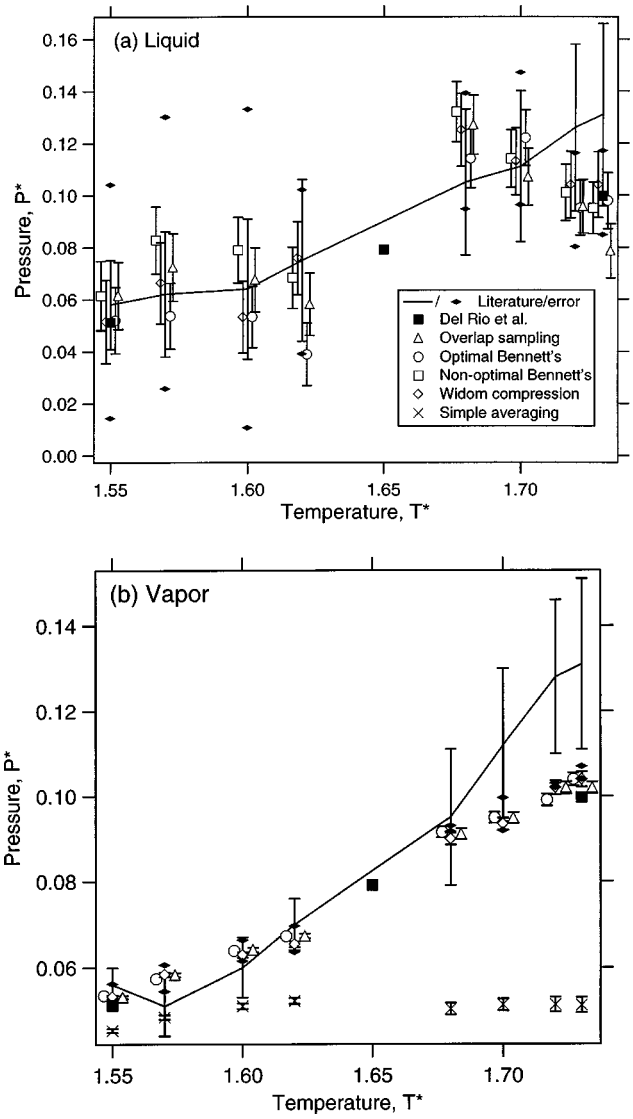

FIG. 3. Pressure P^* (in reduced units, defined as $P\sigma^3/\varepsilon$) measured by free-energy volume perturbation in a system of 256 square-well particle of well diameter $\lambda = 1.75$, given as a function of temperature T^* (defined as kT/ε). Data are measured at the saturated liquid (a) and vapor (b) densities for the given temperature (as given in Ref. 21). The methods used are as follows: single-stage Widom Eq. (1) for a small compression of the system (diamonds) (all other methods combine compression and expansion trials); overlap sampling, Eq. (6) (triangles); nonoptimal Bennett's method, Eqs. (7) and (10), with $C=0$ (squares); optimal Bennett's method (circles) and simple averaging, Eq. (2) (crosses) [for (a) these data are at about $P^* = -1.0$, and are not visible on the plot]. The error bars represent a 67% confidence limit based on the variance of block averages. Confidence limits on the literature values of Vega *et al.* (Ref. 21) are shown in two ways. The reported error bars from Ref. 21 are indicated; also, filled flat diamonds show pressures computed here (using the nonoptimal Bennett's method) using densities at the top and bottom of the confidence limits of density reported in Ref. 21. The latter calculation shows how the imprecision in the density results of Ref. 21 contributes to the uncertainty in our pressure comparison. Some of the data are shifted left or right a small amount to permit the error bars to be discerned—each cluster of points is measured at the same temperature (literature-data series is not shifted and indicates temperature of surrounding cluster). Finally, the recent vapor-pressure data of del Rio *et al.* (Ref. 22) are shown.

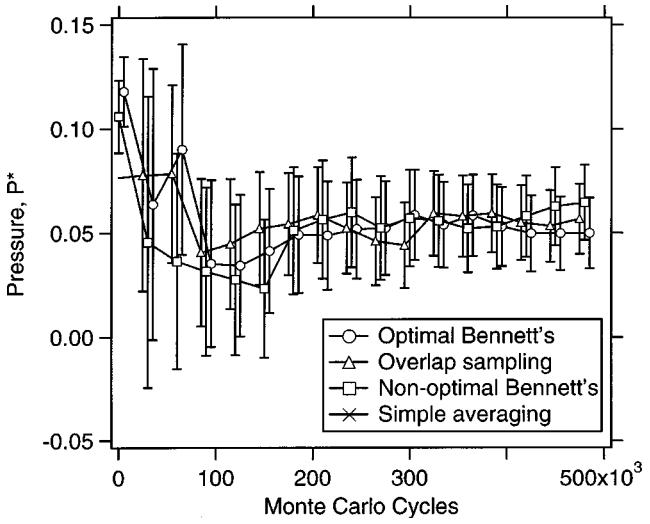

FIG. 4. Cumulative average of the pressure of a $\lambda = 1.75$ square-well system at state conditions $\rho^* = 0.537$, $T^* = 1.55$. Some of the data are shifted left or right a small amount to permit the error bars to be discerned. The data for simple averaging are not visible on the plot.

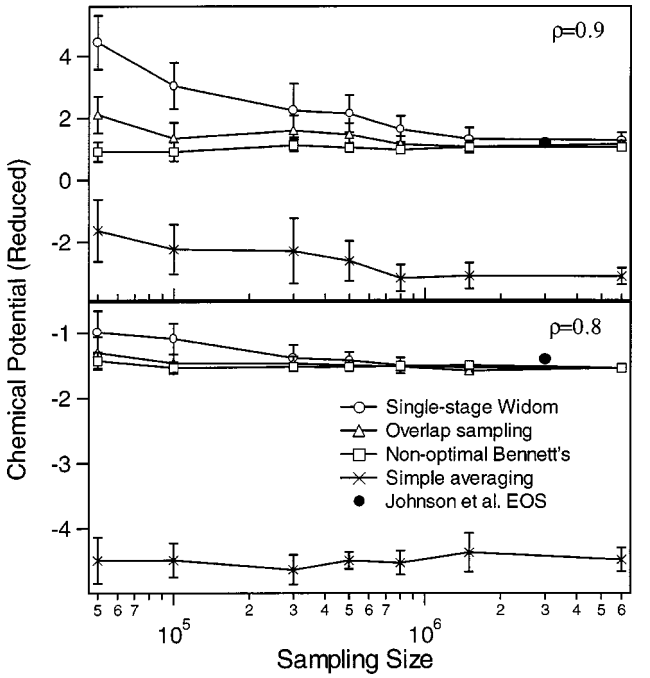

FIG. 5. Cumulative average of the chemical potential of Lennard-Jones system computed using different methods as a function of sampling size (number of perturbation trials). The reduced density (in LJ units) of the systems are 0.9 and 0.8 as indicated. The methods used are as follows: single-stage Widom, Eq. (1) for insertion of a particle (circles); overlap sampling, Eq. (6) (triangles); nonoptimal Bennett's method, Eqs. (7) and (10), with $C=0$ (squares); and simple averaging, Eq. (2). The error-bar represents 67% a confidence limit based on the variance of block averages. The chemical potential according to the equation of state (EOS) of Johnson *et al.* (Ref. 23), shifted to remove the long-range contribution for comparison with the simulation data, is presented at the (arbitrary) abscissa value of 3×10^6. The small discrepancy between the EOS and the converged values can arise from system-size effects and the limitations of the semiempirical EOS.

Results are presented in Figs. 3 and 4. All methods perform comparably well, with the important exception of the simple averaging technique [Eq. (2)], which not surprisingly gives an extremely poor result. This combining method gives inadequate weight to the contribution from the repulsive core, and consequently the pressure is consistently low by a large amount. Among the other methods, none stand out as exceptionally better than the others, while all give confidence limits smaller than the literature data (though perhaps not significantly so). The rough equivalence of these methods might be connected to the small size of the perturbation being performed. Even though the free-energy change is amplified when dividing by the volume change, this magnification does not bring out any differences in the quality of the data. Such a small change does not require the accuracy- or variance-enhancing features of Bennett's method. This outcome is further reinforced in examination of the convergence of the averages, Fig. 4. Simple averaging is always off the scale, but the other methods are indistinguishable. We note that the failure of our ''saturation'' pressure to consistently increase with temperature is an indication of possible problems in the saturation density data of Vega *et al.*

B. Calculating the chemical potential

The chemical potential calculation was discussed in the Introduction. It is a FEP in which the perturbation systems differ in the presence of a single molecule. In principle, the overlap-sampling and Bennett's methods should be applied by performing a simulation of $N+1$ molecules perturbed by deleting one of them, and performing another simulation of N molecules and perturbing by adding another at random, thus

$$e^{-\beta\mu_r} = \frac{\langle w(u)e^{-\beta u/2}\rangle_N}{\langle w(u)e^{+\beta u/2}\rangle_{N+1}}, \tag{12}$$

where $\mu_r = \mu - kT\ln\rho$ is the residual chemical potential (above an ideal gas of density ρ), and u is the energy of the test molecule when it interacts with all the others. It is possible to proceed as we did with the pressure calculation, and approximate the two stages by inserting and deleting a molecule from the N-particle system (as long as the molecule is not too large and the density not too high); we did not do that in this study, and instead performed separate simulations of systems of N and $N+1$ particles, as prescribed by Eq. (12).

We used the methods described above to compute the residual chemical potential of a Lennard-Jones (LJ) system. The simulations are conducted at the NVT ensemble with system densities $\rho\sigma^3 = 0.9$ and 0.8, and $kT/\varepsilon = 1.0$ (where σ and ε are the LJ parameters). In both densities $N = 108$ is used and no long-range correction is applied. The free-energy perturbation sampling is conducted once at the end of each MC simulation cycle, which contains N random translational displacement moves. An equilibration run of 2×10^6 cycles is carried out before starting the FEP sampling. We examine the convergence of the different methods in Fig. 5, and compare with the value given by the equation of state of Johnson *et al.*[23] The simple averaging method Eq. (2) produces clearly unacceptable results, showing no sign of converging to the correct value, while presenting error estimates

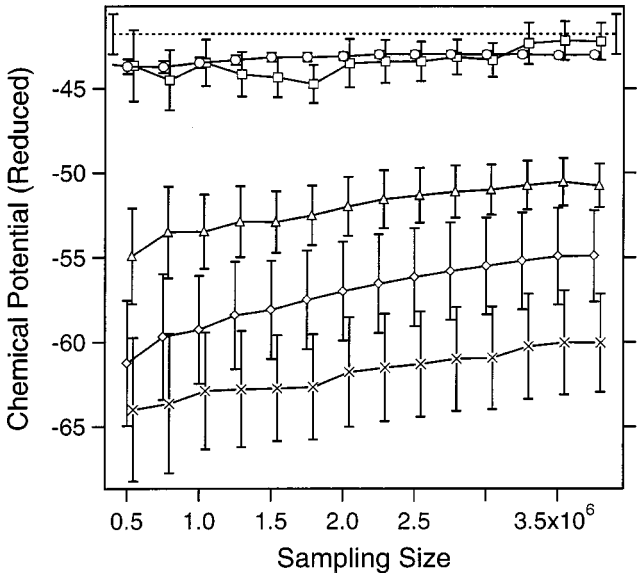

FIG. 6. Cumulative average of the chemical potential of SPC water at 298 K, as a function of sampling length (simulation cycles); results are made dimensionless with the LJ epsilon for the O-O interactions (650.4 J/mol), and are in excess of an ideal gas at the same density. The methods used are indicated by the symbols as follows: overlap sampling, Eq. (6) (triangles); nonoptimal Bennett's method, Eqs. (7) and (10), with $C=0$ (squares); optimal Bennett's method with $C=-41.77$ (circles); standard, single-stage Widom insertion (diamonds); and simple averaging, Eq. (2) (crosses). The reported value (-41.77) in the previous studies (Refs. 26, 27) is indicated by the dashed line. The error-bars represent 67% confidence limits (they are barely visible—about the same size as the symbols—for the optimal Bennett's method); the reported confidence limits from the previous studies is indicated on either end of the dashed line. Some data series are shifted slightly to the right to permit error bars to be discerned.

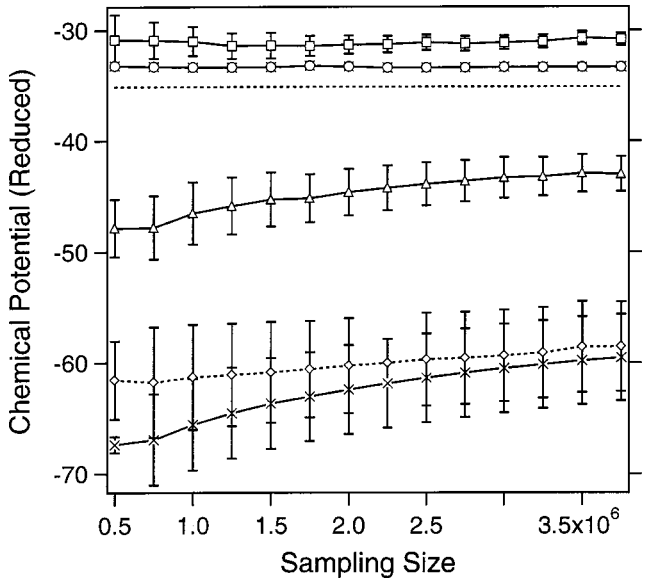

FIG. 7. Same as Fig. 6, except temperature is 373 K. Literature value (Ref. 27) of chemical potential under similar simulation conditions is -35.2 in reduced units.

that indicate (falsely) an increasingly precise result. Even single-stage Widom's method,[24] early in the simulation, presents error bars that are smaller than the inaccuracy of the calculation; but it also displays an acceptable degree of convergence to the correct value as the simulation proceeds. Overlap sampling Eq. (6) and Bennett's method Eq. (10) (using a nonoptimal value $C=0$), both show good results, with Bennett's yielding a correctly converged result very quickly, already at the beginning of the plot. The free-energy difference is rather close to zero, relative to the extremes of energy that are observed in the insertion and deletion, so in this case the arbitrary selection of $C=0$ is not too far from the optimum value.

Now we consider the chemical potential of a water model, which is a much more difficult average to measure due to the large entropy change of the perturbation (roughly $7k$; the free-energy change is about $10kT$, but this is of less relevance to the difficulty of the calculation). We expect a hard time for the direct FEP measurement and want to use this calculation to examine the performance of the simple overlap and Bennett's methods. We choose the SPC model[25] for water and apply a 6 Å cutoff for interaction potential, with no long-range correction of any type. The simulation is conducted in cubic simulation box with periodic boundary condition applied. The system density is $\rho=1.0$ g/cm^3, and two temperatures are examined: $T=298$ K and 373 K. These settings are similar to those of Hermans *et al.*[26] and Quintana and Haymet.[27] In these previous studies, thermodynamic integration and much lengthier multistage methods were applied. The reported chemical potentials are -41.77 for 298 K and -35.2 for 373 K. We make these and our results dimensionless with the LJ ε for the O-O interactions (650.4 J/mol), and they are reported in excess of an ideal gas at the same density. In this study we use $N=256$ water molecules, which is different from the literature; this difference would slightly offset the chemical potential, but the literature values should still provide a reasonable reference. The definition of sampling size is the same as that of LJ simulation, except that each simulation cycle also contains N random rotational MC moves. The configuration is pre-relaxed with 8×10^6 MC cycles before proceeding with the FEP sampling.

Figures 6 and 7 present data from the different methods at $T=298$ K and 373 K, respectively, similar to that given in Fig. 5 for LJ. Bennett's method was applied in both optimally ($C=-41.77,-35.2$,resp.) and nonoptimally ($C=0$), and these applications provide the only acceptable results. All other methods—Widom's, simple overlap sampling, and simple averaging—provide values that differ from the correct chemical potential by an amount significantly greater than indicated by their confidence limits. In contrast, Bennett's method, particularly with the optimal C, has converged correctly almost by the beginning of the plot. There remains a small but significant difference between optimal and nonoptimal Bennett's method in Fig. 7. This outcome highlights the insidious nature of the inaccuracy of these calculations. It is likely that the optimal-Bennett's method is providing the more accurate result. In principle it is a better method, and it is in better agreement with the literature value. The nonoptimal form is not bad, and in particular it shows pretty good accuracy, at least in comparison to the other methods—only the precision of the calculation is noticeably compromised by the use an a less-than-optimal value of C. The

accuracy and precision of the optimized Bennett's method is outstanding.

V. CONCLUSIONS

The practice of simple forward-and-reverse averaging [Eq. (2)] is, at best, highly inefficient. It may provide accurate results only if the perturbation can be made very small through the introduction of numerous intermediate stages in the overall free-energy difference calculation. Yet even then, when used with all its inefficiencies, the method is prone to inaccuracy that is hard to detect. As standard practice it really should be considered undesirable—perhaps even unacceptable.

Equation (6) is no more complicated than the widely used FEP ensemble averages [Eq. (1) or (2)], yet it is much safer to use. It does not demand prior knowledge of the relative entropies of the systems of interest, and it does not even require that the important regions of phase space satisfy a subset relation. The only disadvantage in Eq. (6) is that it does require FEP averages sampled in both of the systems of interest. In practice this is hardly an issue. Unless one is absolutely sure that Eq. (1) is being applied in the insertion direction, and that the subset relation [Fig. 1(a)] is satisfied [and not Fig. 1(b)], then one should perform both averages to be sure the result is not inaccurate. As shown here, in some instances both averages can be obtained—albeit approximately—in a single simulation. In many cases one is traversing a range of values of the perturbation parameter, and the opportunity to perform both averages arises naturally. The double-wide sampling method[15] attempts to exploit this situation, but it suffers from the same flaw as the forward-and-reverse averaging, and it too can be easily improved using the combining methods advocated here.

Going further, application of Bennett's method, even without full optimization, produces marked improvement over all other methods of combining forward and reverse averages. The use of a suboptimal value of C seems to be less detrimental to the accuracy of the calculation than it is to the precision. Given that the accuracy is very hard to gauge without detailed calculations, any low-cost step taken to ensure a higher-accuracy calculation is worthwhile. Moreover, with just a bit of thought, reasonable bounds can be placed on the free-energy difference, and this can be used to guide in the selection of an appropriate (near optimal) value of C. Bennett's method applied in optimal form produces very impressive results, and should be used if at all possible. Barring this, we advocate strong consideration of overlap sampling, Eq. (6), or some other nonoptimal form of Bennett's method as methods of choice for basic FEP calculations.

ACKNOWLEDGMENTS

This work is supported by the Division of Chemical Sciences, Office of Basic Energy Sciences, Office of Energy Research of the U.S. Department of Energy (Contract No. DE-FG02-96ER14677). Computational resources were provided by the University at Buffalo Center for Computational Research.

[1] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon, Oxford, 1987); D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, New York, 1996); P. Kollman, Chem. Rev. **32**, 2395 (1993).
[2] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954); C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[3] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[4] I. Nezbeda and J. Kolafa, Mol. Simul. **5**, 391 (1991); A. P. Lyubartsev, A. A. Marsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1992).
[5] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[6] K. S. Shing and K. E. Gubbins, Mol. Phys. **46**, 1109 (1982).
[7] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989); S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[8] M. Watanabe and W. P. Reinhardt, Phys. Rev. Lett. **65**, 3301 (1990).
[9] D. A. Kofke and P. T. Cummings, Fluid Phase Equilib. **150**, 41 (1998); N. Lu and D. A. Kofke, J. Chem. Phys. **111**, 4414 (1999).
[10] D. A. Kofke and P. T. Cummings, Mol. Phys. **92**, 973 (1997).
[11] A. E. Mark, in *Encyclopedia of Computational Chemistry*, edited by P. v. R. Schleyer (Wiley, New York, 1998), Vol. 2, pp. 1070.
[12] R. J. Radmer and P. A. Kollman, J. Comput. Chem. **18**, 902 (1997).
[13] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977); J. Valleau, Adv. Chem. Phys. **105**, 369 (1999).
[14] N. Lu and D. A. Kofke, J. Chem. Phys. **115**, 6866 (2001).
[15] W. L. Jorgensen and C. Ravimohan, J. Chem. Phys. **83**, 3050 (1985).
[16] D. A. Pearlman and P. A. Kollman, J. Chem. Phys. **91**, 7831 (1989); J. F. Blake and W. L. Jorgensen, J. Am. Chem. Soc. **112**, 7269 (1990); R. H. Wood, W. C. F. Mühlbauer, and P. T. Thompson, J. Phys. Chem. **95**, 6670 (1991); D. A. Pearlman, *ibid.* **98**, 1487 (1994); C. Chipot, L. G. Gorb, and J.-L. Rivail, *ibid.* **98**, 1601 (1994); J. E. Eksterowicz, J. L. Miller, and P. A. Kollman, J. Phys. Chem. B **101**, 10971 (1997); R. H. Henchman and J. W. Essex, J. Comput. Chem. **20**, 499 (1999).
[17] N. Lu and D. A. Kofke, J. Chem. Phys. **114**, 7303 (2001).
[18] N. Lu and D. A. Kofke, AIChE Symp. Ser. **97**, 146 (2001).
[19] M. P. Allen, in Proceedings of the Euroconference on "Computer simulation in condensed matter physics and chemistry," edited by K. Binder and G. Ciccotti (Como, Italy, 1996), Vol. 49, pp. 255.
[20] V. I. Harismiadis, J. Vorholz, and A. Z. Panagiotopoulos, J. Chem. Phys. **105**, 8469 (1996).
[21] L. Vega, E. de Miguel, L. F. Rull, G. Jackson, and I. A. McLure, J. Chem. Phys. **96**, 2296 (1992).
[22] F. del Rio, E. Avalos, R. Espindola, L. F. Rull, G. Jackson, and S. Lago, Mol. Phys. **100**, 2531 (2002).
[23] J. K. Johnson, J. A. Zollweg, and K. E. Gubbins, Mol. Phys. **78**, 591 (1993).
[24] B. Widom, J. Chem. Phys. **39**, 2808 (1963).
[25] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, in *Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry*, edited by B. Pullman (Reidel, Dordrecht, The Netherlands, 1981), pp. 331.
[26] J. Hermans, A. Pathiaseril, and A. Anderson, J. Am. Chem. Soc. **110**, 5982 (1988).
[27] J. Quintana and A. D. J. Haymet, Chem. Phys. Lett. **189**, 273 (1992).