Accuracy of free-energy perturbation calculations in molecular simulation. II. Heuristics

Nandou Lu and David A. Kofke
Department of Chemical Engineering, University at Buffalo, State University of New York, Buffalo, New York 14260-4200

(Received 9 April 2001; accepted 2 August 2001)

We examine issues involved in applying and interpreting free-energy perturbation (FEP) calculations in molecular simulation, with the aim to develop simple heuristics that can guide their use and warn when a result is likely to be inaccurate. We build on the accuracy model developed in the first paper of this series [N. Lu and D. A. Kofke, J. Chem. Phys. **114**, 7303 (2001)], which emphasized the sign of the entropy difference (ΔS) between the target and reference systems as an essential indicator for the correct implementation of FEP calculations: such calculations must be performed in the ''insertion'' direction, for which $\Delta S<0$, or else they are very likely to be systematically incorrect (i.e., inaccurate). We describe here an extended analysis for insertion FEP calculations, and identify the group $M\exp(\Delta S/k)$, where M is the number of independent FEP samples taken and k is Boltzmann's constant, as a relevant quantity for characterizing the accuracy of FEP result. We find that if $M\exp(\Delta S/k)$ is of order 100 or larger, then one can expect the FEP calculation to yield a result of minimally acceptable accuracy; for a margin of safety a value of 1000 or greater is preferable for this group. Although the FEP-measured ΔS is required to apply this heuristic, it is ''safe'' in that any inaccuracy in this ΔS will be such that the group $M\exp(\Delta S/k)$ is even smaller than it is for the true ΔS, and will therefore still warn of an inaccurate result. The analysis is demonstrated for a very wide range of ΔS values, considering a model FEP calculation, a hard-sphere insertion calculation, and a diameter-change FEP in the Lennard-Jones model. We apply the results of this analysis, and earlier work, to consider the question of the optimal number of intermediate stages to use in a staged FEP calculation. The analysis shows that, for optimal accuracy, stages should be selected such that the entropy difference per stage satisfies $\Delta S/k=-1$; however, consideration of the precision instead prescribes that $\Delta S/k=-2$. Inasmuch as the precision is the larger concern once accuracy reaches an acceptable level, the latter criterion forms our recommendation for selecting the number of intermediate stages. © *2001 American Institute of Physics.* [DOI: 10.1063/1.1405449]

I. INTRODUCTION

Free-energy perturbation (FEP) is a method for calculating free-energy differences by molecular simulation.[1] The basic technique is widely used as the foundation of analyses of phase and reaction equilibria, solvation, binding affinity, stability, and kinetics.[2–5] The approach is very popular because it is very convenient to apply. In its simplest form it gives the free-energy difference between one system (the *reference*) and another (the *target*) while performing a simulation of only the reference.[6] However, FEP calculations are highly prone to systematic errors, and simple countermeasures that are sometimes applied to remedy these inaccuracies often do not improve the outcome. The evidence for an incorrect result is found by applying the FEP calculation twice, taking each of the two systems of interest as the reference. Usually these results differ systematically, and a simple average is taken in an attempt to get the correct value.[7–17] In the best case the inaccuracies are removed by applying more computation, usually by introducing very many intermediate stages in the FEP calculation, until the difference between the forward and reverse averages for each stage is brought to an acceptable level. Such a solution is effective but very inefficient.

The usual methods for characterizing the quality of a simulation result consider the variance of block averages of the simulation data, and thus consider the *precision* of the calculation, not its *accuracy*. Completely different methods are needed to characterize the accuracy of the calculation.[18–20] The difficulty is to identify inaccuracy in the FEP result using simulation data that might themselves be inaccurate. In the first paper of this series[21] (hereafter referred to as paper I), we performed an analysis of the accuracy of FEP calculations. The aim of the work in paper I is to provide a formal basis for developing simple but effective methods for recognizing and compensating for systematic errors in FEP calculations. In the present work we apply this formalism to develop an easy-to-use heuristic for identifying systematic errors in FEP calculations. We review in Sec. II the primary results obtained in paper I, and in Secs. III through V we develop our accuracy heuristics. Simulation tests for verifications are given in Sec. VI and concluding remarks in Sec. VII.

II. REVIEW

Entropy plays a central role in the analysis of the accuracy and precision of FEP calculation, so we characterize the

two systems of interest according to their relative entropies.[19,20] We use a subscript "H" to denote a property of the higher-entropy system, and we use "L" to describe a property of the system having lower entropy. The free energy difference can be calculated via simulation in which either the H or the L system is used as the reference (i.e., the system that governs the sampling of configurations), while the other serves as the target. If the H system is the reference, we say that the FEP calculation is a (generalized) *insertion*, while if the L system forms the reference, we have a *deletion* calculation. In paper I we concluded that deletion FEP calculations were hopelessly prone to inaccuracy, and should be avoided entirely, so in the present work we consider inaccuracy in insertion calculations only.

Insertion free-energy perturbation calculations are based on the following exact formula:[22,23]

$$\exp(-\beta\Delta A)=\langle\exp(-\beta u)\rangle_H, \tag{1}$$

where $u=U_L-U_H$ is the difference in energy for a given configuration in the H and L systems, and $\Delta A=A_L-A_H$ is the corresponding free-energy difference; β is the reciprocal temperature, $1/kT$, with k the Boltzmann's constant and T the absolute temperature. The angle brackets indicate an ensemble average, and the H-subscript thereon indicates sampling of the high-entropy system. A corresponding formula applies to the deletion calculation[24]

$$\exp(+\beta\Delta A)=\langle\exp(+\beta u)\rangle_L. \tag{2}$$

It is useful to write the ensemble averages as one-dimensional integrals

$$\exp(-\beta\Delta A)=\int du\,\exp(-\beta u)f(u), \tag{3}$$

and

$$\exp(+\beta\Delta A)=\int du\,\exp(+\beta u)g(u), \tag{4}$$

where the integrals run from $-\infty$ to $+\infty$, and $f(u)$ and $g(u)$ are distribution functions that describe the probability of observing a configuration having energy difference u when sampling the higher-entropy and lower-entropy systems, respectively. These distributions are related[24,25]

$$f(u)\exp(\beta\Delta A)=g(u)\exp(\beta u). \tag{5}$$

We model the inaccuracy of FEP calculations by assuming that all error arises from inadequate sampling of the tail of the distribution where the exponential in the integrals in Eqs. (3) and (4) are large. We identify a lower-limit energy u_f for the insertion calculation, or upper-limit u_g for the deletion calculation, and we assume that sampling is perfect for energies up to this value, but that no sampling occurs beyond it. The unsampled region is the source of the inaccuracy in the calculation. It happens that the fractional or relative inaccuracy of $\exp(\beta\Delta A)$, δe, in the insertion average (subscript "ins") is simply the area under the g distribution lying below u_f, while the error in the deletion average (subscript "del") is the area under the f distribution lying above u_g

$$\delta e_{\rm ins}\equiv\frac{\exp(-\beta\Delta A_{\rm ins}^{\rm sim})-\exp(-\beta\Delta A^{\rm exact})}{\exp(-\beta\Delta A^{\rm exact})}=-\int_{-\infty}^{u_f}g(u)du, \tag{6}$$

and

$$\delta e_{\rm del}\equiv\frac{\exp(+\beta\Delta_{\rm del}^{\rm sim})-\exp(+\beta\Delta A^{\rm exact})}{\exp(+\beta\Delta A^{\rm exact})}=-\int_{u_g}^{+\infty}f(u)du, \tag{7}$$

where the superscripts "sim" and "exact" of ΔA indicate the simulation result and the true value of the free-energy difference, respectively.

The absolute error in $\beta\Delta A$ can be calculated using

$$\delta(\beta\Delta A_{\rm ins})=\beta\Delta A_{\rm ins}^{\rm sim}-\beta\Delta A^{\rm exact}=-\ln\left(1-\int_{-\infty}^{u_f}g(u)du\right), \tag{8}$$

for insertion calculation, and

$$\delta(\beta\Delta A_{\rm del})=\beta\Delta A_{\rm del}^{\rm sim}-\beta\Delta A^{\rm exact}=\ln\left(1-\int_{u_g}^{+\infty}f(u)du\right) \tag{9}$$

for deletion calculation. For small error, the fractional error δe is approximately the absolute error in $\beta\Delta A$, as one can see from the series expansion

$$\delta(\beta\Delta A_{\rm ins})=-\ln(1+\delta e_{\rm ins})=-\delta e_{\rm ins}+O(\delta e_{\rm ins}^2), \tag{10}$$

and

$$\delta(\beta\Delta A_{\rm del})=\ln(1+\delta e_{\rm del})=\delta e_{\rm del}+O(\delta e_{\rm del}^2). \tag{11}$$

From Eqs. (6) and (7), or (8) and (9), one can see that any error in the insertion calculation results in overestimate in ΔA, while error in the deletion calculation results in underestimate in ΔA. As defined here, δe is always a negative number. For convenience, in the following discussion when we mention the value of δe, we actually refer to its absolute value.

We further model the inaccuracy by considering the most likely outcome from a simulation in which the FEP average is sampled M times independently. Applying a probabilistic argument, we can develop expressions for the most likely values of the limit energies (u_f^*,u_g^*) in terms of the energy distributions, thus

$$\left.\frac{\partial\ln f(u)}{\partial u}\right|_{u=u_f^*}=Mf(u_f^*), \tag{12}$$

and

$$\left.\frac{\partial\ln g(u)}{\partial u}\right|_{u=u_g^*}=-Mg(u_g^*). \tag{13}$$

We have argued that a FEP calculation performed in the insertion direction will be much more reliable than one performed as a deletion. One way to approach this conclusion using the accuracy model reviewed here involves consideration of the width of the f and g distributions. The limit energy of the broader distribution can more easily come to lie on the other side of the narrower distribution—the sampled region of the broad distribution is better able to overlap the entire narrow distribution, whereas much more sampling is required to have the sampled region of the narrow distribution encompass the broad distribution. Since the error is based on the area of the conjugate distribution lying outside the sampled region of the reference-system distribution, sampling that uses the narrow distribution as the reference will have a larger error. We expect that this is the low-entropy distribution, which is the one sampled in a deletion calculation.

III. SAMPLING THRESHOLD FOR 50% ERROR

To develop a heuristic, we consider first a means to gauge the order of magnitude of the sample size M (the number of independent FEP samples contributing to the ensemble average) needed to approach a reasonable result in an insertion calculation. To this end we develop an expression for the value of M that yields a fractional error of 50% in the free energy. Of course a practical FEP calculation would aim for a much better accuracy than this, and consequently would carry on much more sampling. But it turns out that the 50% threshold is easily characterized using our formalism, and it does provide an order-of-magnitude gauge for the amount of sampling that is necessary for a good result, as well as hinting the path to follow for further analysis.

According to our accuracy model, a 50% error in the insertion free energy occurs when the most likely limit energy u_f lies at the median of $g(u)$. For this purpose we can approximate g as a symmetric distribution, which is appropriate for many cases. Then the median becomes the mode, or the peak value, of g. To apply this criterion, we need to reformulate the expression for the most likely u_f in terms of the g distribution. This is easily accomplished by combining Eqs. (5) and (12), with the result

$$\left.\frac{\partial \ln g(u)}{\partial u}\right|_{u=u_f^*} + \beta = M g(u_f^*) e^{-\beta \Delta A} e^{+\beta u_f^*}. \tag{14}$$

We want to use this to get an expression for the value of M for which u_f^* lies at the maximum of $g(u)$. Thus, we may drop the derivative, since it is zero at this point. Also, we break the free-energy difference into its energy and entropy components, yielding

$$\beta = M_{1/2} \exp(\Delta S/k) g(u_f^*) \exp[+\beta(u_f^* - \Delta U)], \tag{15}$$

where the 1/2 subscript on M indicates it is the value for 50% accuracy. We showed in paper I that the energy difference ΔU should in many cases be approximately equal to the mean of $g(u)$. If we equate the mean to the median/mode u_f^*, then the exponential in the energy will drop out, and Eq. (15) becomes

$$\beta = M_{1/2} \exp(\Delta S/k) g(u_f^*). \tag{16}$$

The maximum value of $g(u)$, i.e., $g(u_f^*)$, can be characterized by its width, or standard deviation σ_g—since g is normalized to unity, its maximum should vary inversely with σ_g. The advantage in making this change is that σ_g can be more easily calculated in a simulation; it does not require histogramming. Introducing these considerations, we arrive at a simple, compact expression for the 50%-accuracy sample size

$$M_{1/2} \exp(\Delta S/k) \sim \beta \sigma_g. \tag{17}$$

Clearly, the 50%-accuracy sample size, $M_{1/2}$ depends on the entropy difference between the perturbation systems, and the width of the g distribution. Note that for insertion $\Delta S < 0$. More sampling is required as the entropy difference increases, and (much less so) as the distribution of energies in the target system widens.

The primary observation that we take from this analysis is the importance of the group $M \exp(\Delta S/k)$ in characterizing the expected accuracy of an insertion calculation. The relevance of $M \exp(\Delta S/k)$ is consistent with the common understanding that the accuracy of a FEP calculation is affected by both the sampling size and the magnitude of the perturbation between target and reference, which is closely related to the entropy difference ΔS. The present result quantifies this view, and shows how M and ΔS come together to influence the accuracy. Contrary to some popular views, the magnitude of ΔA is not itself the primary quantity affecting the accuracy.

In the remainder of this work, we will develop this observation and test it with some simple example FEP calculations. We examine the relevance of $M \exp(\Delta S/k)$ first for a model based on reasonable forms for the energy distributions, then we consider the hard-sphere insertion FEP calculation, and finally some simulation results for a FEP calculation involving the Lennard–Jones model.

IV. TWO MODEL SYSTEMS

A. Energy-distribution model

The most likely inaccuracy model enables us to perform a full error analysis, provided detailed knowledge of the f and g distributions. We begin our development by selecting a simple but reasonable form for these functions, and examine the expected accuracy of a FEP calculation performed on a system that exhibits these distributions. Since f and g are related exactly via Eq. (5), it is sufficient to specify a form for just one of them; we will use g. An appropriate form of the function should be able to characterize the g distribution well, especially for the distribution tails which are important for the error analysis. It is especially important that the high-energy range of $g(u)$ vanish no faster than $\exp(-u)$ (so, for example, a Gaussian form would be inappropriate). Accordingly, we consider the following form

$$g(x) = \kappa x^{\lambda} \exp(-\alpha \beta x), \tag{18}$$

where $x = u - U_0$ with U_0 is the minimum possible energy difference between the perturbation systems; $g(x)$ is defined to be zero for $x < 0$. The parameters λ and α are positive

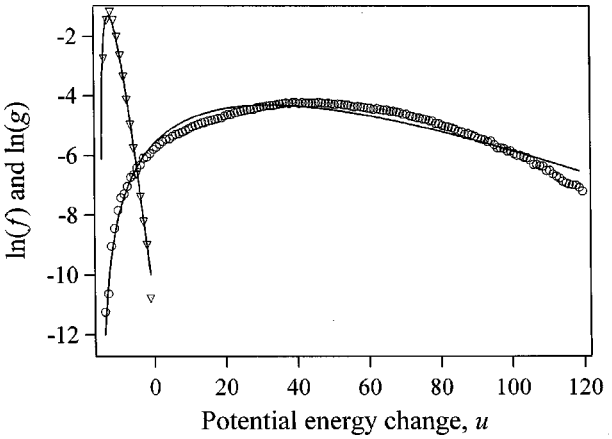

FIG. 1. The fitting of f and g histograms from a Monte Carlo simulation to the algebraic functions used in the energy-distribution model in Sec. IV A. On the plot, $\ln(f)$ and $\ln(g)$ are presented as a function of potential-energy change u. The triangles and circles represent the histograms of f and g obtained in the simulation, respectively. The continuous curves are fitting results. Note that the variable x in Eq. (18) is taken as $x=u-U_0$ for this fitting. U_0 is the minimum energy change that could be encountered in a deletion calculation, and was taken as the energy of a LJ particle in a perfect FCC lattice at the same density. The Monte Carlo simulation is conducted at $\rho=0.9$ and $\beta=0.5$ (in units of the LJ potential parameters). The high entropy system contains 107 LJ particles with reduced diameter 1 and one LJ particle with reduced diameter 0.8; the low entropy system contains 108 LJ particles with reduced diameter 1. Parameters in Eq. (18), $\lambda=2.820$ and $\alpha=2.346$, provide a good fit for both f and g histograms.

numbers (β is the reciprocal temperature, as above), and κ is the normalization constant, which is given in terms of the gamma function,

$$\kappa=\frac{(\beta\alpha)^{\lambda+1}}{\Gamma(\lambda+1)}. \tag{19}$$

The f distribution function is derived using the relationship given by Eq. (5):

$$f(x)=\kappa x^{\lambda}\exp[-(\alpha-1)\beta x]\exp(-\beta\Delta A). \tag{20}$$

Thus, for f to remain well-defined, we require the parameter α to be greater than the unity, i.e., $\alpha-1>0$.

The function used above mimics well the behavior of the distributions, at least for simple particle-insertion FEP calculations. As an example, in Fig. 1 we show the f and g histogram from a simulation of the Lennard–Jones (LJ) system and their fit to this form. The simulation conditions, as well as the fitting parameters, are given in the caption of the figure. As we can see, both f and g distributions are well represented by this simple form.

Relevant thermodynamic quantities can be described in terms of the function parameters λ and α. The exact free-energy difference, ΔA, is obtained from Eq. (4),

$$\beta\Delta A=-(\lambda+1)\ln(1-\omega), \tag{21}$$

where $\omega=1/\alpha$ and has value less than 1. Approximating ΔU as the mean of g,[21] the average energy difference between the target and reference is given by

$$\beta\Delta U=\beta\int ug(u)du=\omega(\lambda+1). \tag{22}$$

Then the entropy difference is

TABLE I. List of conditions used in the analysis of energy-distribution model. The values of parameters, λ, α, and β, are pre-given for each series. The entropy difference $\Delta S/k$ is computed using the approach described in the text of Sec. IV A. Note that two series pairs, A1 and A2, A8 and A9, are included in the study. Each pair has about the same value of $\Delta S/k$.

Series	λ	α	β	$\Delta S/k$
A1	2.2	1.5	1.0	−1.382
A2	3.4	1.667	0.9	−1.392
A3	4.5	1.5	1.0	−2.378
A4	7.0	1.5	1.0	−3.456
A5	3.0	1.2	1.5	−3.834
A6	10.0	1.5	1.0	−4.751
A7	4.5	1.2	1.0	−5.271
A8	2.2	1.034	1.45	−7.790
A9	3.0	1.06	1.7	−7.791
A10	18.0	1.5	1.0	−8.207
A11	8.2	1.125	1.6	−12.037
A12	5.0	1.052	3.24	−12.314
A13	5.0	1.012	3.3	−20.509

$$\Delta S/k=\beta\Delta U-\beta\Delta A=(\lambda+1)[\omega+\ln(1-\omega)], \tag{23}$$

and is always negative, consistent with our definition of an insertion FEP calculation.

With this simple g distribution function in hand, a full accuracy analysis can be easily conducted. Given a simulation length M, the most likely limit energy u_f^*, and u_g^* can be calculated numerically from Eqs. (12) and (13), respectively, using the given forms of the g or f distributions. The most likely error in the free-energy difference then can be computed using Eqs. (6) or (7). Various FEP conditions can be simulated by choosing different sets of parameters λ, α, and β, since the shape of the distribution functions, as well as the value of the thermodynamic quantities, depends on the choice of these parameters. The parameter sets used in this study, together with the corresponding thermodynamic quantities, are summarized in Table I. We cover a wide range of values of $\Delta S/k$, and also include different parameter sets having the same $\Delta S/k$.

It is worth noting that the ratio of variances of f and g is given by

$$\frac{\sigma_f^2}{\sigma_g^2}=\frac{\alpha^2}{(\alpha-1)^2}>1, \tag{24}$$

indicating that f is a wider distribution than g. According to our argument in paper, I, sampling on the wider distribution will provide greater accuracy in the free energy. The numerical results, discussed in the following, bear out this expectation.

B. Hard-sphere insertion model

To develop our heuristic, it is worthwhile also to analyze the most likely error for a particle-insertion FEP (Widom insertion) in the simple hard sphere (HS) system. This perturbation has only two discrete energy changes, either zero or infinity. The insertion calculation is the only meaningful FEP calculation for the HS system (deletion is always inaccurate, regardless of sampling length). The purpose of this analysis is to look for any common inaccuracy behavior between this

TABLE II. List of success rates (p) used in the analysis of hard-sphere system discussed in Sec. IV B, and the corresponding values of entropy difference, $\Delta S/k$.

Series	Success rate for insertion, p	$\Delta S/k$
H1	0.2211	−1.509
H2	1.1×10^{-5}	−11.408
H3	8.12×10^{-6}	−11.721
H4	1.23×10^{-7}	−15.907

calculation and the one modeled with the continuous energy distributions of the previous section. Trends found to apply in these systems having very different energy distributions and characterizations may be anticipated to hold generally.

Only the successful insertion trials, which result in no overlap between particles, have nonzero contribution to the exponential of free-energy difference, $\exp(-\beta\Delta A)$. The exact free-energy difference, which is equivalent to $\Delta S/k$ because ΔU is zero, is simply related to the success probability p of an insertion trial

$$\exp(-\beta\Delta A)=\exp(\Delta S/k)=p. \tag{25}$$

The method for inaccuracy analysis of this discrete system must be different from that for the systems with continuous energy distributions.[26] For an insertion calculation with finite sampling size M, the number of successful insertions k would differ from Mp because k must have an integer value, whereas Mp does not. Because the free-energy measurement is given in terms of k/M, this discretization effect gives rise to inaccuracy in ΔA.

The probability that there are k successful trials out of M attempts is given by the binomial distribution,

$$P_k=\frac{M!}{k!(M-k)!}p^k(1-p)^{M-k}. \tag{26}$$

The most likely number of successful insertion trials, k^*, can be obtained by maximizing the probability P_k; since k must be discrete, we perform this maximization numerically, by simply scanning all values of integer k near the real value Mp. Then the most likely free-energy difference, ΔA^*, is given by

$$\exp(-\beta\Delta A^*)=\frac{k^*}{M}, \tag{27}$$

and the most likely fractional error is

$$\delta e=\frac{\exp(-\beta\Delta A^*)-\exp(-\beta\Delta A^{\text{exact}})}{\exp(-\beta\Delta A^{\text{exact}})}=\frac{k^*}{Mp}-1. \tag{28}$$

For the tests we choose different success rates p, and thereby change the value of the entropy difference; these are listed in Table II.

V. HEURISTICS

A. Analysis of models

For comparison, we compute the most likely inaccuracy in both the insertion and deletion calculations for the continuous energy-distribution model. We present a typical plot of results in Fig. 2, using series A9 as an example. The figure, where the absolute error in ΔA is plotted as a function of sample size M, shows a picture of most likely errors very similar to that obtained from actual simulations in paper I. From Fig. 2, one can see that the free-energy errors for the insertion and deletion calculations have opposite sign, and both decrease with the increasing sample size M. However, the magnitude of error for the insertion calculation differs markedly from that for the deletion calculation. With the same sample size, the insertion error is much smaller. One also can see that the insertion inaccuracy decreases faster than its deletion counterpart, with increasing sampling size. This clearly shows that the insertion calculation has higher efficiency in improving the simulation accuracy when increasing the sample size.

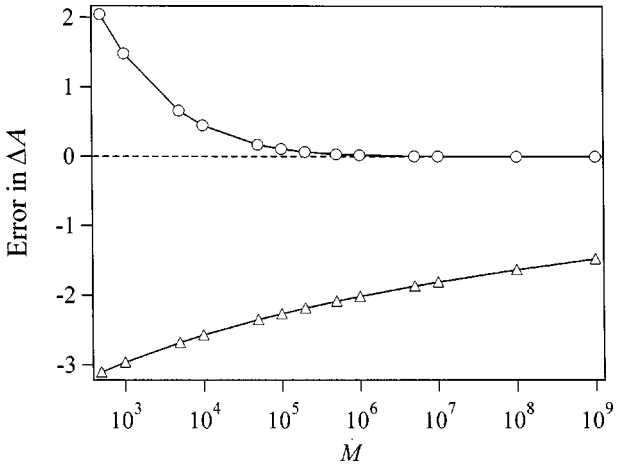

FIG. 2. Results of the most likely inaccuracy from the analysis of energy-distribution model. The error in ΔA of series A9 is plotted as a function of sample size M. The curve with open circles represents the most likely error of the insertion calculation, and that with open triangles is for the deletion calculation. The dashed horizontal line indicates zero inaccuracy.

Numerical results for the fractional error δe for the models described in Sec. IV are presented in Fig. 3. In Fig. 3(a), δe is plotted as a function of sampling length M, where one can note that for different systems the error is non-negligible for M spanning nine orders of magnitude in range (and this could be made yet broader by selecting systems with even larger entropy differences). The same set of inaccuracy data are presented in plot (b) instead as a function of $M\exp(\Delta S/k)$. Clearly, the inaccuracy is a decreasing function of both the sampling size M and the group parameter $M\exp(\Delta S/k)$. Equally clear is the collapse of the curves when presented in terms of the group $M\exp(\Delta S/k)$, where the error across different systems is non-negligible over, at most, two orders of magnitude in this group. Importantly, the curves from both models, the continuum energy-distribution model, and the hard-sphere insertion model, collapse onto a single, nearly universal form. This behavior convincingly demonstrates the important effect of the entropy difference on the FEP inaccuracy, and confirms that the group $M\exp(\Delta S/k)$ is an appropriate quantity to generalize the common behavior of the inaccuracy in the FEP calculations. Figure 3(c) shows that the data, when presented as suggested by Eq. (17) in terms of $M\exp(\Delta S/k)/\beta\sigma_g$ differ from each other even less than they do in Fig. 3(b) (this plot excludes the hard-sphere model data, for which the representation

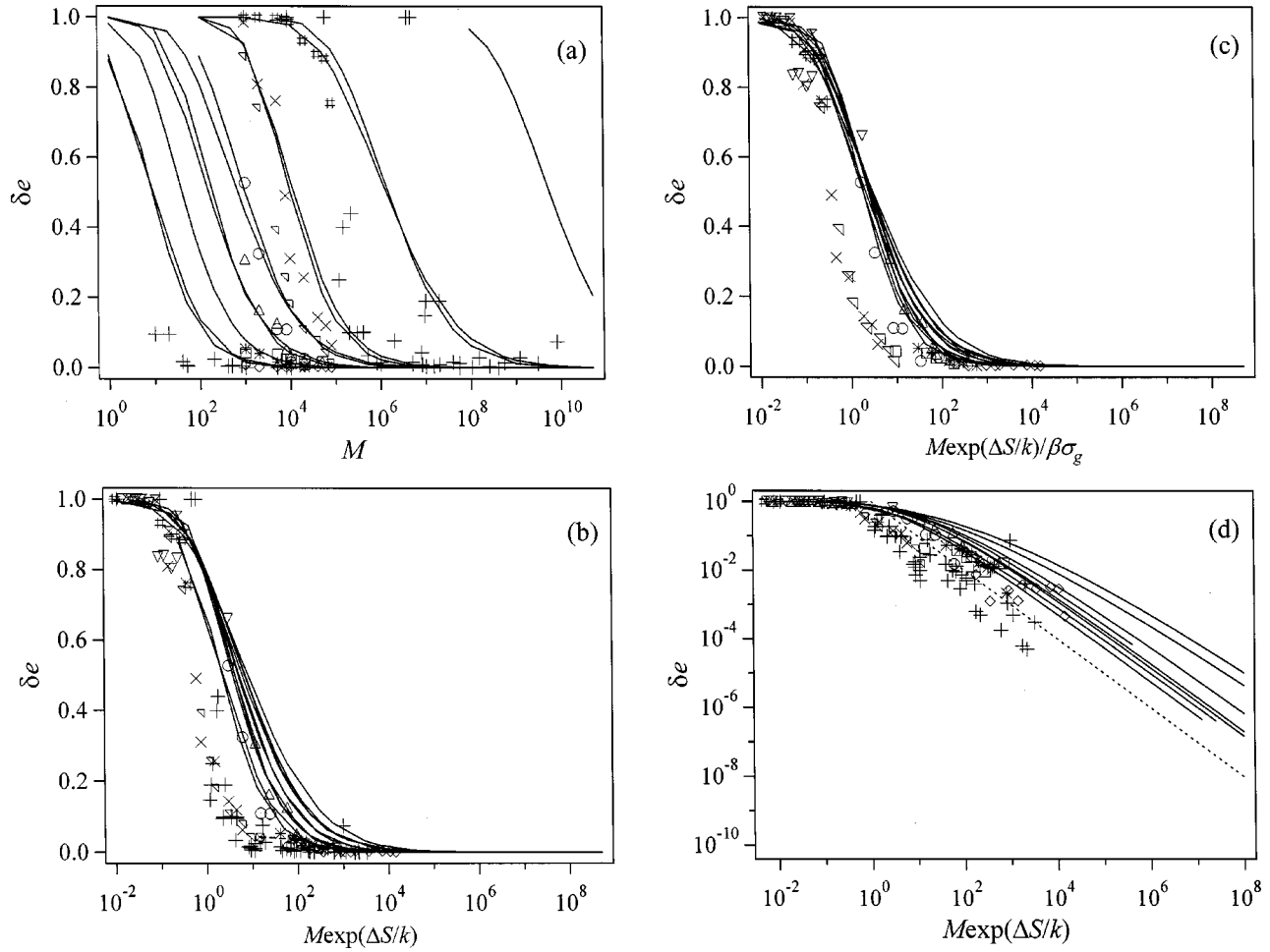

FIG. 3. Numerical results for the inaccuracy of δe observed in energy-distribution model analysis, hard-sphere system analysis, and simulation tests over different conditions. In plot (a), the inaccuracy (in the linear scale) is presented as a function of simulation length M (in the logarithmic scale); in plot (b) and (c), the same set of inaccuracy data (in the linear scale) are plotted as a function of the group parameter $M\exp(\Delta S/k)$ and $M\exp(\Delta S/k)/\beta\sigma_g$ (in the logarithmic scale), respectively; plot (d) shows δe and $M\exp(\Delta S/k)$ both in the logarithmic scale, highlighting the large-M behavior. Note all the plots have the same range of orders of magnitude in x-axis. The solid curves are results from the energy-distribution model analysis; the (+) symbols are those of hard-sphere system analysis, and the other markers indicate inaccuracy observed in simulation tests. In plot (c) the hard-sphere data are not included. The dotted line in plot (d) has a slope of -1, and is presented as a reference. Conditions of the model analysis and simulation tests are given in Tables I to III.

does not apply). However the improvement to the universality of the behavior is not so great to warrant this complication, so we do not include it in what follows.

The decay rate of inaccuracy can be described in terms of the slope of the log-scale curves [Fig. 3(d)], and is given by the exponent γ in the relationship $\delta e \sim[M\exp(\Delta S/k)]^{-\gamma}$. The results show that γ is not constant, but increases as $M\exp(\Delta S/k)$ becomes larger. At very small $M\exp(\Delta S/k)$, the error decays slowly, and γ is close to zero. From this perspective the behavior of interest is instead at very large $M\exp(\Delta S/k)$, where γ approaches 1 for all tests. This indicates a relatively rapid decay in the inaccuracy at sufficiently large M, and should be compared with the $M^{-1/2}$ decay in the imprecision. The unsurprising, but nevertheless important, conclusion is that for sufficiently large M, inaccuracy (systematic error) eventually becomes washed out by imprecision (noise).

For comparison, in Fig. 4 we also plot δe from the continuous model analysis as a function of $M\exp(-\beta\Delta A)$ instead of $M\exp(\Delta S/k)$. Clearly $M\exp(-\beta\Delta A)$ lacks the feature of collapsing the curves nicely, as $M\exp(\Delta S/k)$ has done in Fig. 3(b). This again indicates that ΔA is not itself the primary quantity affecting the accuracy.

B. Heuristic for estimating inaccuracy from simulation results

It is of great interest to know how much FEP sampling is required to yield a result to an acceptable level of accuracy, say 95%, or at least somewhere within the precision of the calculation. The results of the model-system analyses, as presented in Fig. 3(b), suggest an appropriate heuristic to address this question. There we see that the inaccuracy for many cases diminishes to something less than 5% when the value of $M\exp(\Delta S/k)$ reaches about 100. However, this result is not clear cut, inasmuch as some of the continuous-energy distributions do not reach this level of accuracy until $M\exp(\Delta S/k)$ reaches 1000 or more. In Fig. 5 we consider the behavior of the fractional error in ΔA, which for some researchers may be of greater interest than the fractional error in $\exp(-\beta\Delta A)$ (or absolute error in $\beta\Delta A$), used in Fig. 3(b). By this measure $M\exp(\Delta S/k)=100$ presents a reasonably

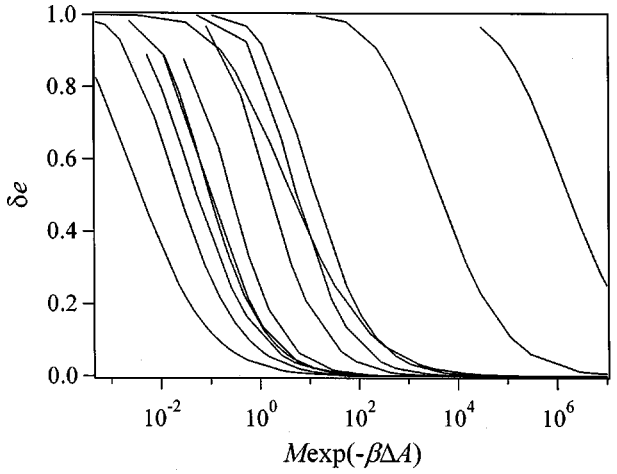

FIG. 4. Plot of inaccuracy δe from the continuous energy-distribution model against $M\exp(-\beta\Delta A)$. The inaccuracy data are exactly the same as those presented in Fig. 3. The same number of decades are used for the x-axis for a better comparison with Figs. 3(a) and 3(b).

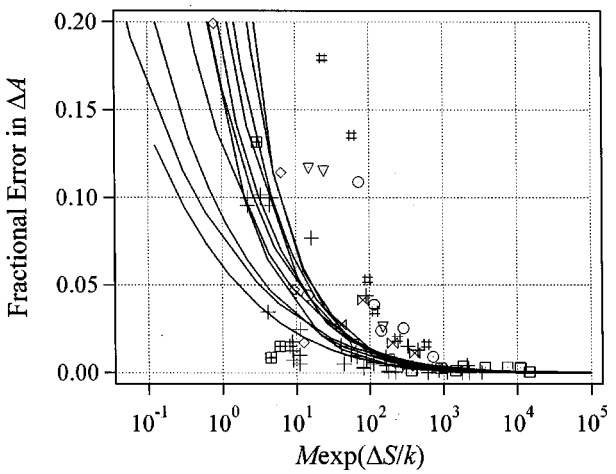

FIG. 5. Results of inaccuracy from model analysis and simulation tests. The inaccuracy is given in terms of fractional error in ΔA as a function of group parameter $M\exp(\Delta S/k)$. Conditions of model analysis and simulation tests are given in Tables I to III.

safe threshold for proclaiming a free-energy result to be accurate. Consequently, given the behavior of the error as demonstrated by both Fig. 3(b) and Fig. 5, we propose $M\exp(\Delta S/k)=100$ as the standard for gauging the quality of any FEP calculation. If the sampling length M does not reach this threshold, then sampling is insufficient and the calculated free-energy difference is likely to be in error. If it reaches this level but not much more, the result may be accurate, but should still be regarded with some suspicion. If this measure reaches 1000 or more, the free energy can with good confidence be considered to be an accurate measurement. The group $M\exp(\Delta S/k)/\beta\sigma_g$ [cf. Eq. (17)] presents a better gauge of the accuracy of the result; if available, this quantity can instead be compared to the threshold of 100, and if it passes this value, the measurement can be accepted with even more confidence.

Application of this heuristic requires knowledge of ΔS, but evaluation of ΔS is the primary outcome of the FEP calculation. So naturally we ask how can this ΔS, which is of suspect accuracy, be trusted to verify its own correctness?

The entropy difference ΔS can be computed using formula

$$\Delta S/k=\beta(\Delta U-\Delta A), \tag{29}$$

where ΔU and ΔA are simulation results of the potential-energy difference and free-energy difference for the perturbation systems. Obviously, this ΔS can contain inaccuracies from both ΔU and ΔA. However, the potential-energy difference ΔU is a mechanical quantity, and can be measured with quite good accuracy in a simulation. In some cases, ΔU can be obtained from simulation of only the reference system; for example, in a particle-insertion FEP calculation it is well-described by the average energy of one of the molecules present in the system. In the worst case, it can be measured accurately by separate simulations of the reference and target systems, which perhaps is being done anyway as part of a staged FEP calculation. Compared to that of ΔA, the inaccuracy of ΔU is small and therefore will be ignored. Now, it has been shown by our most likely accuracy model, as well as by common observation, that an insertion FEP calculation normally overestimates ΔA. In turn, the value of ΔS given by Eq. (29) is underestimated (more negative). Consequently with the relationship $\delta e\sim[M\exp(\Delta S/k)]^{-\gamma}$, an underestimated ΔS will result in an overestimate of δe. In other words, the value of $M\exp(\Delta S/k)$ computed using simulation results will indicate a larger error than its true value. Therefore, application of the heuristic, using the simulation result for ΔS, provides a ''safe'' indicator of the accuracy of the FEP calculation.

Importantly, all of the above assumes that the FEP calculation is proceeding in the insertion direction, such that $\Delta S<0$. For many perturbations, the relative magnitude of the entropy of the systems can be correctly predicted by considering their degrees of freedom qualitatively. However, in certain situations we may be unable to judge which of two systems has the lower entropy without actually performing simulations on them. Three cases can then arise:

(1) We correctly select the high-entropy system as the reference, and the entropy difference is non-negligible (say $-\Delta S/k>1$). But if we simulate appropriately long to get an accurate free energy, and thus find that $M\exp(\Delta S/k)$ is sufficiently large, we can trust the result.
(2) We incorrectly select the low-entropy system as the reference, and the true entropy difference (unknown to us) is non-negligible. The inaccuracy of the deletion calculation is such that the measured entropy change will be close to zero, and perhaps with the noise in the calculation $\Delta S/k$ may be found to be slightly negative. The reasoning considers that the sampling of narrow g distribution would result in $\exp(-\beta\Delta A)=\langle\exp(-\beta u)\rangle_g$, which is about the same as $\exp(-\beta\Delta U)$. Although $M\exp(\Delta S/k)$ will indicate a good value, the result is incorrect, and we must suspect it on the basis of the small $\Delta S/k$.
(3) The true entropy difference is in fact small, $\Delta S/k\approx 0$. Unbeknown to us, we can safely choose either system as the reference, and we get a correct result for the free energy and entropy. However, we cannot distinguish this

outcome from case (2), which also yields a small (but incorrect) entropy difference. Thus, we need to perform the FEP calculation in another direction, and see that the computed entropy difference remains a small number.

We anticipate that the third case will not arise often in practice. A more likely problem is connected to the implicit supposition that the target system, of smaller entropy, has important configurations in phase space that form a wholly contained subset of the important configurations of the reference. This issue was discussed in some detail in paper I. One must apply some physical intuition and qualitative reasoning, specific to the systems of interest, to judge if this situation holds. At present we do not have a quantitative means to verify if two systems are related this way. If they are not, it becomes necessary to apply staging methods to construct a set of reference-target pairs that each meet this condition.

C. Heuristic for multistage FEP calculations

We now consider how our insight developed previously can be used to better understand the formulation of multistage FEP calculations. The idea of a multistage method is to break a large free-energy difference into a collection of smaller ones, formed using a set of one or more intermediate systems that interpolate the reference and target systems. For example, the intermediates for a molecule-insertion FEP calculation could define a set of stages in which the molecule is inserted one atom at a time. Whereas the single-stage FEP calculation may be prohibitively long to achieve an acceptable level of accuracy, the FEP calculations between each of the substages may be quite tractable. There are endless choices associated with the use of multistage methods: Should they be used at all? How many intermediate stages is best? How should the intermediates be formulated? In previous work[20] we considered this issue from the standpoint of the precision of the calculation, and concluded that the intermediates should be selected such that the entropy difference between them was the same for all stages, but we did not consider how many stages should be used, so we turn to this topic now.

We consider a FEP calculation between a given target and reference, having thermodynamic differences ΔA, ΔS, ΔU, etc. which are of course independent of the number of FEP stages used to compute them. We consider using n intermediate FEP stages, and following the prescription developed previously, we define interpolating systems such that the entropy difference for all intermediates stages is the same, and equal to $\Delta S/n$. Consequently the sampling length m required to reach some prescribed level of accuracy is the same for all stages, and satisfies

$$m \exp(\Delta S/nk) = c, \tag{30}$$

where c is a constant. The total number of FEP samples M is the sampling length per stage, times the number of stages

$$M = mn = nc \exp(-\Delta S/nk). \tag{31}$$

We can minimize M with respect to n, and find that the optimal number of intermediates is such that

$$\frac{-\Delta S/k}{n_{\text{opt}}} = 1. \tag{32}$$

That is, the optimal number of stages corresponds to a unit entropy difference per stage.

We can also consider the same question from the point of view of the precision of the calculation. In previous work we showed that the variance of the free-energy calculation is related to the entropy difference and sampling length as follows

$$\sigma^2_{\Delta A} \sim M^{-1} \exp(-\Delta S/k). \tag{33}$$

For independent stages of a multistage calculation, the total variance is the sum of those for the substages. With the stages formulated to have equal entropy differences, and with equal allocation of sampling to each stage, the total variance is

$$\sigma^2_{\Delta A} \sim n(M/n)^{-1} \exp(-\Delta S/nk) = \frac{n^2}{M} \exp(-\Delta S/nk). \tag{34}$$

Minimization of the variance with respect to the number of intermediate stages now yields

$$\frac{-\Delta S/k}{n_{\text{opt}}} = 2, \tag{35}$$

which differs slightly from the accuracy-based heuristic, Eq. (32). We note that if we approach this by instead minimizing the overall sample size for fixed precision, we get this same criterion for the optimal number of stages.

As mentioned above, beyond a certain amount of sampling, the precision of the calculation is of greater concern than its accuracy, so in general Eq. (35) is the heuristic to follow in deciding if, and how many, intermediate stages should be used. Regarding accuracy, one need only take care that the amount of sampling devoted to each stage is sufficient to ensure an accurate result. Using our heuristic for sufficient sampling, each stage should perform at least $100 \times \exp(2)$, or about 1000 independent FEP samples. It would be wise to pad this substantially to add a margin of safety; the degree of course depends on the expense associated with obtaining independent FEP measurements, and the computational budget.

We reiterate the stipulation mentioned previously, that all of this presupposes stages that have been formulated to ensure the ''subset'' relation of the important regions of configuration space. Depending on the relation between the reference and target systems, the intermediates may be constructed in different ways to ensure that this relationship holds for all stages. The choices lead to methods commonly known as umbrella sampling,[27,28] Bennett's method,[29] and staged insertion.[18,19] The reader is referred to paper I for a more detailed discussion of this issue.

VI. SIMULATION TESTS

We finish by examining the arguments and heuristics developed previously using actual FEP simulation data. We conducted a set of Monte Carlo simulations in the canonical ensemble for this purpose. In our simulation tests, the refer-

TABLE III. Conditions for the simulation tests conducted in Sec. VI. The "potential" and "diameter" columns define the interaction and size of the special particle in the reference system. IG means ideal gas, and LJ refers to Lennard-Jones potential; the diameter indicates the size of the special particle relative to a normal LJ sphere (diameter 1). $\Delta S_t/k$ is the "exact" entropy difference of the perturbation system computed by very long simulation with full f and g analysis, as described in the text.

Series	Potential	Diameter	Density, ρ	β	$\Delta S_t/k$
S1	IG	0	0.8	1.0	−8.743 4
S2	IG	0	0.9	1.0	−12.178 7
S3	IG	0	0.9	1.111	−12.727 6
S4	LJ	0.65	0.8	1.0	−4.450 5
S5	LJ	0.9	0.9	0.5	−1.702 6
S6	LJ	0.8	0.9	0.5	−3.202 0
S7	LJ	0.72	0.9	0.5	−4.250 5
S8	LJ	0.7	0.9	1.0	−5.799 0
S9	LJ	0.3	0.9	1.0	−9.503 8
S10	LJ	0.2	0.9	1.0	−12.413 2

ence system contains 107 Lennard-Jones (LJ) particles, as well as another particle which can be an ideal gas particle, or a LJ particle with smaller diameter than a normal one. All target systems have 108 LJ particles. We change the system conditions (both thermodynamic state and perturbation conditions) from test to test in order to cover a wide range of entropy difference ΔS between the reference and target systems.

For each condition, the "true" free-energy difference is computed using a very long simulation ($M \geqslant 5\times 10^6$) with full f and g analysis, as described in paper I. The corresponding "true" ΔS, referred to as ΔS_t, is also calculated according to Eq. (29). Test simulation conditions, together with the values of "true" ΔS_t, are listed in Table III. Note that all ΔS_t are negative, and we deal only with insertion FEP calculations.

In the test simulations, the free-energy difference is computed using the standard FEP formula, Eq. (1). We vary the number of FEP samples M to study the dependence of accuracy on the sampling length. Each sample is taken for configurations separated by 540 MC trials, so we consider each to represent an independent contribution to the ensemble average. We repeat simulations independently up to 200 times for each condition and simulation length, and collect the free-energy differences. We then find the median for these results and use it as the most likely outcome of the free-energy difference, since the median typically is closer than the average to the peak (mode) of a distribution of measured ΔA values. The difference between this median and the "true" ΔA under the same conditions is taken as the most likely inaccuracy. The entropy change for each finite-length simulation is also computed based on the median of ΔA, and referred as ΔS_f. Both $M\exp(\Delta S_t/k)$ and $M\exp(\Delta S_f/k)$ are computed for comparison. Note that all the data used to compute the group parameter $M\exp(\Delta S_f/k)$ is exclusively based on the simulation results, and makes no use of the additional simulations performed to determine the true ΔA.

We superimpose the simulation results on Figs. 3 and 5. Note in Fig. 3(b) and Fig. 3(c), $M\exp(\Delta S_t/k)$ is used for the x-axis. Figures 3 and 5 clearly show that the curves of simulation results are very similar to those of the model analysis. This gives us the confidence that (a) the analysis conducted is meaningful and (b) the heuristics are appropriate. Once again, the simulation results in plot (a) show that the simulation length itself is not enough to reveal any internal common characteristic of inaccuracy of FEP calculations. In contrast, the group quantity $M\exp(\Delta S/k)$ well characterizes the common inaccuracy behavior of FEP calculations across a wide variety of simulation conditions, as one can see from Fig. 3(b). In Fig. 5, we can find that a value of 10^2 for the group quantity $M\exp(\Delta S/k)$ corresponds to an appropriate accuracy in the free-energy results.

VII. CONCLUSIONS

The formalism developed in paper I provides powerful tools to investigate the inaccuracy problem in FEP calculations. The entropy difference, ΔS, of the perturbation plays a central role in determining the extent of systematic error due to inadequate sampling. It works together with the simulation length, M, as a group quantity, $M\exp(\Delta S/k)$, in determining the inaccuracy of free-energy difference. Our study reveals that $M\exp(\Delta S/k)$ well characterizes the common behavior of inaccuracy, and can be used as a good indicator for identifying the accuracy level of the FEP calculation.

To ensure an acceptable level of accuracy in the FEP calculation, two conditions must hold:

(1) The important configurations of the target must form a wholly contained subset of the configurations important to the reference; accordingly, the entropy of the target must be less than the reference-system entropy.
(2) The value of $M\exp(\Delta S/k)$ should be of the order of 10^2 or greater.

Assuming condition 1 is satisfied, condition 2 can be checked as follows:

(1) compute ΔA from the FEP calculations in the usual way, according to Eq. (1);

(2) compute the "exact" potential energy difference ΔU between the reference and target during the simulation, by whatever means is convenient;

(3) compute entropy difference according to Eq. (29);

(4) compute $M\exp(\Delta S/k)$;

(5) if $\Delta S/k$ is negative and significantly different from zero, and if $M \exp(\Delta S/k)$ is about 10^2 or larger, the calculated free energy can be trusted; if $\Delta S/k$ is near zero, conduct the FEP calculation in the reverse direction (exchanging the roles of the target and reference). If consistent results are obtained, they can be trusted; otherwise the result corresponding to the insertion FEP calculation (for which $\Delta S < 0$) is the one to believe.

In future work we will examine these heuristics in the context of other FEP calculations, considering other types of perturbations, and other, more complex, systems. It would also be of value to develop a quantitative measure that indicates the degree of overlap of the important regions of configuration space for the reference and target systems.

ACKNOWLEDGMENT

Acknowledgment is made to the Donors of the Petroleum Research Fund, administrated by the American Chemical Society, for the support of this research.

[1] A. E. Mark, in *Encyclopedia of Computational Chemistry*, Vol. 2, edited by P. v. R. Schleyer (John Wiley & Son, New York, 1998).
[2] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, New York, 1996).
[3] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon, Oxford, 1987).
[4] K. E. Gubbins, in *Foundations of Molecular Modeling and Simulation, AIChE Symp. Ser*., Vol. 97, edited by P. Cummings and P. Westmoreland (2001), pp. 26–34.
[5] P. Kollman, Chem. Rev. **32**, 2395 (1993).
[6] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[7] W. L. Jorgensen and C. Ravimohan, J. Chem. Phys. **83**, 3050 (1985).
[8] W. L. Jorgensen, J. K. Buckner, S. Boudon, and J. Tirado-Rives, J. Chem. Phys. **89**, 3742 (1988).
[9] D. A. Pearlman and P. A. Kollman, J. Chem. Phys. **90**, 2460 (1989).
[10] R. H. Wood, W. C. F. Mühlbauer, and P. T. Thompson, J. Phys. Chem. **95**, 6670 (1991).
[11] R. H. Wood, J. Phys. Chem. **95**, 4838 (1991).
[12] C. Chipot, C. Millot, B. Maigret, and P. A. Kollman, J. Phys. Chem. **98**, 11362 (1994).
[13] D. A. Pearlman, J. Phys. Chem. **98**, 1487 (1994).
[14] C. Chipot, P. A. Kollman, and P. D. A., J. Comput. Chem. **17**, 1112 (1996).
[15] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[16] R. J. Radmer and P. A. Kollman, J. Comput. Chem. **18**, 902 (1997).
[17] D. A. Pearlman and B. G. Rao, in *Encyclopedia of Computational Chemistry*, edited by P. v. R. Schleyer (John Wiley & Son, New York, 1998).
[18] D. A. Kofke and P. T. Cummings, Mol. Phys. **92**, 973 (1997).
[19] D. A. Kofke and P. T. Cummings, Fluid Phase Equilib. **150**, 41 (1998).
[20] N. Lu and D. A. Kofke, J. Chem. Phys. **111**, 4414 (1999).
[21] N. Lu and D. A. Kofke, J. Chem. Phys. **114**, 7303 (2001).
[22] B. Widom, J. Chem. Phys. **39**, 2808 (1963).
[23] J. L. Jackson and L. S. Klein, Phys. Fluids **7**, 228 (1964).
[24] K. S. Shing and K. E. Gubbins, Mol. Phys. **46**, 1109 (1982).
[25] M. P. Allen, in *Proceedings of the Euroconference on "Computer Simulation in Condensed Matter Physics and Chemistry,"* Vol. 49, edited by K. Binder and G. Ciccotti (Como, Italy, 1996), pp. 255–284.
[26] N. Lu and D. A. Kofke, in *Foundations of Molecular Modeling and Simulation, AIChE Symp. Ser*., Vol. 97, edited by P. Cummings and P. Westmoreland (2001), pp. 146–149.
[27] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[28] J. P. Valleau and D. N. Card, J. Chem. Phys. **57**, 5457 (1972).
[29] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).