THE JOURNAL OF CHEMICAL PHYSICS **123**, 054103 (2005)

Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation

Di Wu and David A. Kofke[a)]
Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, New York 14260-4200

(Received 6 May 2005; accepted 13 June 2005; published online 5 August 2005)

We consider ways to quantify the overlap of the parts of phase space important to two systems, labeled A and B. Of interest is how much of the A-important phase space lies in that important to B, and how much of B lies in A. Two measures are proposed. The first considers four total-energy distributions, formed from all combinations made by tabulating either the A-system or the B-system energy when sampling either the A or B system. Measures for A in B and B in A are given by two overlap integrals defined on pairs of these distributions. The second measure is based on information theory, and defines two relative entropies which are conveniently expressed in terms of the dissipated work for free-energy perturbation (FEP) calculations in the $A \rightarrow B$ and $B \rightarrow A$ directions, respectively. Phase-space overlap is an important consideration in the performance of free-energy calculations. To demonstrate this connection, we examine bias in FEP calculations applied to a system of independent particles in a harmonic potential. Systems are selected to represent a range of overlap situations, including extreme subset, subset, partial overlap, and nonoverlap. The magnitude and symmetry of the bias ($A \rightarrow B$ vs $B \rightarrow A$) are shown to correlate well with the overlap, and consequently with the overlap measures. The relative entropies are used to scale the amount of sampling to obtain a universal bias curve. This result leads to develop a simple heuristic that can be applied to determine whether a work-based free-energy measurement is free of bias. The heuristic is based in part on the measured free energy, but we argue that it is fail-safe inasmuch as any bias in the measurement will not promote a false indication of accuracy. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1992483]

I. INTRODUCTION

An important application of molecular simulation[1] is the calculation of free-energy differences, which are required for phase and reaction equilibria, solvation, binding affinity, stability, kinetics, and so on. Free-energy methods can be broadly categorized into density-of-states approaches and work-based methods,[2] though the line between these categories is not sharp. Both approaches are in wide use, and as no single technique is recognized as demonstrably superior for all applications, the choice of which to apply to a given problem is often a matter of personal preference or experience. All methods have their own idiosyncrasies and pitfalls, and despite decades of development[3–6] there remains a strong need for more efficient and robust methods. Indeed, the development and application of free-energy calculation methods are expanding as a topic of great interest.[7–10]

The free energy of a system characterizes an unnormalized distribution that is defined on the phase space of the system. The domain of non-negligible parts of this distribution constitutes a subspace, and the free energy depends on the "size" of this subspace, and on the weight accorded to each point in it. Thus the difference in free energies between two systems involves two subspaces of phase space, and to measure this free-energy difference it is necessary to consider both. It is not feasible to measure absolute free energies directly (at least in the general case), so it is not possible to consider these subspaces independently. While roundabout ways to get a desired free energy may introduce other systems that circumvent the direct comparison, at some point a free-energy difference between some two systems must be evaluated, and at that time two phase-space subspaces must be considered at once. Regardless of the method, the challenge in these calculations always stems from the fact that these two spaces may be wildly different, perhaps, for example, many orders of magnitude different in size, yet it is necessary to perform a single simulation in which both are adequately represented.

To address this difficulty it can be helpful to have a clear conception of the nature of the phase spaces and how they relate to each other in a geometric sense. The free-energy difference provides some such information, but it is not intrinsically a statistic for the pair of subspaces taken together. Rather it is formed simply as the difference between two terms, each relating to just one of the subspaces. A free-energy difference of zero, for example, could result when the phase spaces are mutually exclusive just as when they fully coincide. To understand the performance of a free-energy calculation method it is necessary to know something about how much overlap there is between the subspaces of the two systems. The free-energy difference really says nothing about this. The nature of this overlap is important because it determines how easily both spaces can be sampled at once,

---

[a)] Electronic mail: kofke@buffalo.edu

and it can provide guidance when introducing constructs that aim to bridge them or to promote sampling of both.

While we think that all free-energy methods can benefit from consideration of phase-space overlap, our focus in the present study is on work-based methods. Work-based methods are synthesized by the nonequilibrium work (NEW) formalism of Jarzynski,[9] and encompass free-energy perturbation[4] (FEP) and thermodynamic integration,[3] as well as other standard methods.[6,8,11] The primary pitfall in NEW methods generally is that they can be highly prone to systematic errors. Consequently, in applying them one normally sacrifices efficiency to gain a result that can be considered accurate.

To provide context for our discussion we briefly review the FEP method.[4] FEP can be conducted in either of two directions, and the working equations for each are

$$\begin{aligned} \exp(-\beta\Delta F) &= \overline{\exp(-\beta W_{A\to B})} \quad A\to B \text{ direction}, \\ \exp(+\beta\Delta F) &= \overline{\exp(-\beta W_{B\to A})} \quad B\to A \text{ direction}. \end{aligned} \tag{1}$$

Here $\Delta F = F_B - F_A$ is the free-energy difference between two systems A and B, $\beta = 1/kT$ is the reciprocal temperature in energy units, and $W_{A\to B} = U_B - U_A$ is the work involved in perturbing a given configuration from system A (for which the energy is U_A) to system B (of energy U_B); likewise, $W_{B\to A} = U_A - U_B$. The overbar denotes the ensemble average sampled in the corresponding reference system A or B. The accuracy of a FEP calculation can depend greatly on the direction in which the perturbation is performed. Any significant inaccuracies in the results obtained from the two directions will be of opposite sign, with the calculation overestimating the free-energy difference when performed in one direction, while underestimating it in the other.[12] Sometimes the bias is symmetric, but often the magnitude of the inaccuracy in one direction is very different from the other. It is not unusual to be able to obtain a very good result with a reasonable amount of sampling in one direction, while obtaining a result that is consistently inaccurate for any feasible amount of sampling in the other direction. For more general NEW methods, a systematic error is diminished to the extent that the work process is performed reversibly.

In previous studies[2,8,11,13,14] we examined the problem of bias in NEW calculations and showed that the key consideration influencing the accuracy is the overlap of the important parts of phase space for the systems of interest. Typically three overlap relations, subset, partial overlap, and nonoverlap, correspond to different levels of difficulty when calculating the free energies and should be treated with different staging methods.

Regarding direction, perturbations should always be performed from a system of larger phase space to a system of smaller phase space, and this rule should also be followed by the successive stages during multistage perturbations or general NEW calculations. Otherwise important contributions will be omitted from the ensemble average, and inaccuracies will occur. Above all, it is essential that any individual stage of a work-based calculation be performed from one system into another that is a phase-space subset of it.

In this paper, we attempt to quantify phase-space overlap relations and use these measures to understand and identify bias in the free-energy calculations. In Sec. II we review phase-space overlap relations, and use FEP calculation as an example to explain why they can help us understand inaccuracy in free-energy calculations. In Sec. III we introduce two approaches to quantify phase-space overlap relations. In Sec. IV we examine a model system to develop a quantitative connection between phase-space overlap and bias in free-energy calculations. We then conclude in Sec. V.

II. QUALITATIVE PHASE-SPACE OVERLAP RELATIONS AND FEP CALCULATIONS

We first consider phase-space relations qualitatively. *Phase space* Γ is (assuming a three-dimensional physical space) the $6N$-dimensional space formed from the $3N$ configuration and $3N$ momentum components of all N atoms. Free-energy calculations are usually concerned with only the $3N$-dimensional configuration space (assuming momentum contributions can be included analytically), and we will present most of the discussion with this in mind. Thus each point in phase space represents a unique configuration. The *important* phase space of a system is the subset of the entire phase space that makes significant contributions to the partition function. A specific definition of important phase space is not needed in what follows, but to fix ideas it can be helpful to present a definition. In the canonical ensemble the weight of a configuration $\gamma \in \Gamma$ depends only on its energy via the Boltzmann factor $\exp[-\beta U(\gamma)]$, so it is possible to compare and order the weight (importance) of any two configurations. If we consider for system A some subset of configurations Γ_A, such that all configurations in Γ_A have greater weight than those configurations not in Γ_A, then we could define the important phase space Γ_A, such that

$$\frac{\int_{\Gamma_A} \exp[-\beta U(\gamma)]d\gamma}{\int_{\Gamma} \exp[-\beta U(\gamma)]d\gamma} = f, \tag{2}$$

where f is a value close to unity. Clearly the energy defines the importance, such that all configurations of energy below a cutoff will be important, and all above it will not.

Introducing the normalized total-energy distribution $p(U)$, the cutoff $U_A^{\max}$ would satisfy $\int_{-\infty}^{U_A^{\max}} p(U)dU = f$, and any configuration with energy less than this would be considered part of the important phase space. The distribution of energies is the product of the degeneracy $\Omega(U)$ and the Boltzmann factor $\exp(-\beta U)$, so a given configuration might be "important" because it has a substantial weight (large Boltzmann factor), or because it is one of the relatively many configurations having a low but non-negligible weight [large $\Omega(U)$]. The point of the important phase space is that it identifies those configurations that must be sampled well by a free-energy algorithm in order to produce accurate results.

We compare two systems' important phase spaces by mapping them together in a single (schematic) phase-space diagram. If the two systems differ in their Hamiltonian or

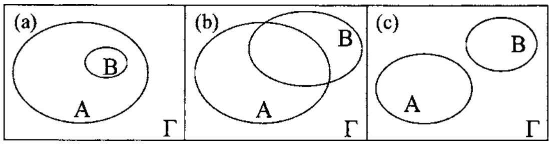

FIG. 1. Schematic depiction of phase-space overlap relations. The large square represents all of phase space, and the circles indicate the regions important to the *A* and *B* systems, as indicated. The cases shown are (a) the subset relation, (b) partial overlap relation, and (c) nonoverlap relation.

some other constraints, but have an equivalent set of coordinates (and, in particular, the same number of atoms), then their phase spaces are of the same dimension, and we can in principle easily compare their important phase-space regions. If two systems differ in their number of atoms, for example, system A contains N molecules and system B has $(N+1)$ molecules (so that the calculated free energy is exactly the chemical potential), then the A configuration space has $3N$ dimensions but B configuration space has $3(N+1)$ dimensions (for a monatomic molecule). To enable them to be mapped together we can instead view the two systems as having different Hamiltonians, one in which the $(N+1)$th molecule interacts as any of the others, and the other in which it behaves as an ideal-gas particle with no interactions with the other molecules. In both cases the molecules are "there," so the phase-space dimensions can be the same for the two systems.

Without loss of generality we stipulate that the A-system important phase space is always larger than the B-system phase space if they are not equal (for brevity we will term these the A phase space and B phase space). Then there are three qualitative categories of relation that can be observed for the two systems' important phase-space regions, and they are depicted in Fig. 1. We consider each in turn.

A. Subset relation

If system B's important phase space constitutes a wholly contained subset region of system A, we call it *subset* relation. Figure 1(a) gives a schematic picture of such a case. We notice that the important phase-space regions of the two systems are related asymmetrically, and this is reflected in the calculations performed in two directions.

1. $A \rightarrow B$ direction

When sampling in system A and perturbing into system B, all the configurations that are important to system B can be reached when sampling the important phase space of A. This is the easiest case for free-energy calculations, and in general calculations applied to it do not require special schemes to connect A and B phase-space regions. In this case FEP calculations can, with a feasible amount of sampling, yield results that are limited more by precision than by accuracy.

An exception for the $A \rightarrow B$ direction is the extreme subset case. This describes the situation in which the B phase space forms only a tiny subset region of the A phase space. This case is often seen in particle-insertion free-energy calculations, where it can be very hard for the randomly inserted molecule to be fitted into the phase without an overlap (e.g., inserting a chain molecule into a dense phase). Generally it is possible for the A system to accommodate the new molecule without a significant increase in energy, but this requires a concerted movement of many molecules to make room for the new one. This means that the configuration is not outside of the A phase space, rather it is a part of it (albeit a very small part). The amount of sampling required to obtain an accurate result is inversely proportional to the fraction that the B phase space occupies in the larger A phase space, so in principle with sufficient sampling an accurate result can be obtained in all cases. We refer to this as an *entropic* sampling barrier.

2. $B \rightarrow A$ direction

If FEP calculation is performed in this direction, results will be consistently inaccurate (almost) regardless of how much sampling is performed. When sampling in the smaller B phase space, the important configurations in the A phase space outside the B oval will not be sampled, and these unreachable configurations can introduce large inaccuracies—indeed, their contribution is *inversely* proportional to their likelihood of being sampled. We refer to this as an *energetic* sampling barrier.

B. Partial-overlap relation

Figure 1(b) gives a picture of an overlap relation. Here there is only a partial overlap of the A and B phase spaces. This means there will always be some amount of configurations that are important to the other system but which are not sampled, regardless of the direction taken for the perturbation. Therefore results from both directions are not accurate. The symmetry of the bias (for $A \rightarrow B$ vs $B \rightarrow A$ perturbation directions) will reflect the symmetry of the relation between the phase spaces. It is possible for both methods to give results that are equally biased (in magnitude) and of opposite sign, in which case averaging the two results will be an estimate better than either. On the other hand, the systems may have a symmetry that approaches that of the subset relation discussed above. In this case one result will be better than the other, and averaging of the results is not prescribed. If, for example, the A phase space encompasses most of the B phase space, such that only a small amount of B region is outside A but a large amount of A region is outside B, then for an equivalent amount of sampling the $A \rightarrow B$ result will be much closer to the correct value.

C. Nonoverlap relation

In this case there is no overlap region between the two systems' important phase spaces, as depicted in Fig. 1(c). The problems are severe for FEP calculations, and single-stage perturbations completely fail when conducted in either direction. With energetically disfavored perturbations, almost every configuration sampled in one system is not important to the other system, thus none of them contributes to the calculations. This is also the case in which histogram-based free-energy calculation methods are prone to fail.

In summary, for successful perturbations the systems of interest must have a subset relation, and the perturbation must be performed in the direction from the system of a larger phase space to a system of a smaller phase space (which we call the insertion direction and label it $A \rightarrow B$). Subset relations are associated with entropic barriers (extreme subset case) and can in principle be overcome with enough sampling; overlap and especially nonoverlap relations are discouraged by energetic barriers and must be treated by advanced work-based methods (multistage methods) instead of simple FEP calculations.

III. QUANTIFYING PHASE-SPACE OVERLAP

Phase-space overlap relations provide a useful conceptual framework to guide the application of free-energy calculations. These ideas help us not only to understand the difficulties of the calculations, but also to choose and design the appropriate methods to perform free-energy calculations. For this framework to be effective, it is essential to know the phase-space overlap relation for a given system of interest. In this section, we present two methods to quantify the systems' phase-space overlap relations.

A. Total-energy distribution method

Energy is the key quantity characterizing the importance of a configuration to a given system, so it is appropriate to consider energy distributions when trying to understand phase-space relations. Most obvious are the total-energy distributions, which we denote here by $p_{AA}(U_A)$ and $p_{BB}(U_B)$, respectively, defined such that $p_{AA}(U_A)dU_A$ is the probability that a system-A energy within dU_A of U_A would be observed in a Boltzmann-weighted sampling of system A; likewise for $p_{BB}(U_B)$. A comparison of $p_{AA}(U_A)$ and $p_{BB}(U_B)$ is not very informative, as they do not provide any connection between the A and B systems.

One way to bring the systems together is to introduce the cross distributions, $p_{BA}(U_A)$ and $p_{AB}(U_B)$, which describe the likelihood to observe a configuration important to one system, *when sampling the other*. More specifically, $p_{AB}(U_B)dU_B$ is the probability that a configuration with system-B energy within dU_B of U_B would be observed in a Boltzmann-weighted sampling of system A; likewise $p_{BA}(U_A)$ is the probability density for observing a configuration of energy U_A (when defined according to system A) while sampling system B.

To aid the discussion we introduce an alternative notation for the energy distributions: $\langle A\rangle_A \equiv p_{AA}(U_A)$, $\langle A\rangle_B \equiv p_{BA}(U_A)$, $\langle B\rangle_A \equiv p_{AB}(U_B)$, and $\langle B\rangle_B \equiv p_{BB}(U_B)$, which is slightly more concise, and makes clear which of the systems is sampled and which system defines the dependence of the tabulated energy on the configuration.

With these new distributions, we consider the two pairs

$$\begin{aligned} &(a) \quad B \text{ in } A \quad \langle A\rangle_A \leftrightarrow \langle A\rangle_B, \\ &(b) \quad A \text{ in } B \quad \langle B\rangle_B \leftrightarrow \langle B\rangle_A. \end{aligned} \tag{3}$$

Pair (a) compares the A-energy distribution sampled in two systems, and characterizes whether system B's important phase-space region is inside or outside that of system A. To see this, consider a configuration generated when sampling B. We are interested in knowing if this configuration is important to system A, and in effect we ask "is the system-A energy of this configuration typical of the system-A energies encountered when sampling system A itself?" To answer, we compare this system-A energy to the distribution $\langle A\rangle_A$. The distribution of such energies U_A for configurations sampled in B then gives an overall picture of how much of system-B phase space coincides with that of system A. Figure 2 shows one such case, where the $\langle A\rangle_B$ distribution is on the right of the $\langle A\rangle_A$ distribution. The overlap region between the two distributions indicates some set of configurations that are encountered when sampling both systems, and thus this describes the overlap region of their important phase space. The higher-energy levels that lie above the $\langle A\rangle_A$ distribution but within the energy range of $\langle A\rangle_B$ distribution represent configurations that are unimportant to the A system, but are still in the B phase space. Thus the degree to which $\langle A\rangle_B$ lies on or to the left of $\langle A\rangle_A$ indicates (loosely) the amount of the B system in the A system. When $\langle A\rangle_B$ lies well to the left of $\langle A\rangle_A$, then the B system preferentially accesses a small set of configurations that are energetically favorable to A, but which are entropically disfavored.

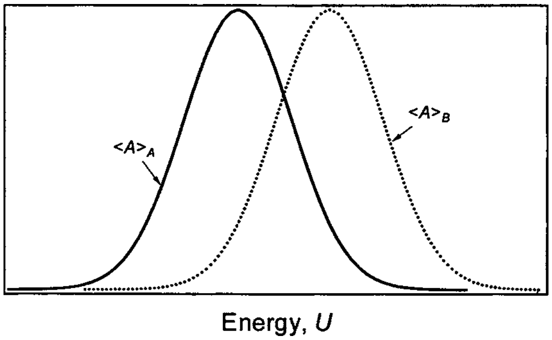

FIG. 2. Schematic illustration of total-energy distribution. The curve labeled $\langle A\rangle_A$ is the distribution of system-A energies observed when sampling system A, while $\langle A\rangle_B$ is the distribution of system-A energies observed when sampling system B.

Comparison of the $\langle A\rangle_A$ and $\langle A\rangle_B$ distributions does not provide sufficient information to determine the conjugate question, regarding how much of the A system is in the B system—just because system B accesses some configurations with energies important to A does not mean it accesses all such configurations. To tell the amount of A in B requires comparison of the other pair, denoted (b) in Eq. (3). If $\langle B\rangle_A$ is on the left of or coincides with $\langle B\rangle_B$, then it tells that the A phase space is inside the B phase space; and if $\langle B\rangle_A$ is on the right of $\langle B\rangle_B$, then A is outside B. Thus in combining the information from two pair comparisons we can tell a system's overall phase-space overlap relation. For example, if B is inside A and A is outside B, it is a subset relation; if B is outside A and A is outside B, it is an overlap or nonoverlap relation. The degree of overlap can be ascertained qualitatively by looking at how far away $\langle A\rangle_A$ and $\langle A\rangle_B$ or $\langle B\rangle_B$ and $\langle B\rangle_A$ are separated from each other.

We can define metrics that more quantitatively characterizes the degree to which the complementary energy distributions coincide. Considering that the order of the distributions on the energy axis is significant, an appropriate metric will indicate if one distribution is to the right or left of the other if they do not fully overlap. The following overlap integrals satisfy this requirement:

$$(a)\ B \text{ in } A \quad K_{BA} = 2\int_{-\infty}^{\infty} dU_{A1} p_{AA}(U_{A1}) \int_{-\infty}^{U_{A1}} dU_{A2} p_{BA}(U_{A2}), \tag{4}$$

$$(b)\ A \text{ in } B \quad K_{AB} = 2\int_{-\infty}^{\infty} dU_{B1} p_{BB}(U_{B1}) \int_{-\infty}^{U_{B1}} dU_{B2} P_{AB}(U_{B2}).$$

A picture describing one of the pair distributions is plotted in Fig. 3, where the solid distribution represents the $\langle A\rangle_A$ (or $\langle B\rangle_B$) distribution and the dotted lines represent the $\langle A\rangle_B$ (or $\langle B\rangle_A$) distribution. We note three special or limiting cases: (a) the two distributions exactly coincide, $K=1$; (b) the dotted distribution is to the left of the solid distribution, $1 \leqslant K \leqslant 2$, with $K=2$ when the distributions are completely separated this way (such as distributions 1 and 2 shown in Fig. 3); and (c) the dotted distribution is to the right of the solid distribution, $0 \leqslant K < 1$, with $K=0$ when the distributions are completely separated this way (such as distributions 3 and 2 shown in Fig. 3). Thus, for K_{BA} approaching zero, less of B phase space lies inside the A phase space, while for it approaching (and perhaps exceeding) unity, we conclude that the B phase space is a wholly contained subset of the A phase space. A FEP calculation $A \rightarrow B$ can give accurate results only if K_{BA} is not too close to zero. Analogous statements apply to K_{AB}.

B. Relative-entropy measurements

An alternative metric can be defined using ideas from information theory,[15] which provides a definition for the "distance" between two distributions. Our fundamental interest is in the distributions defined on phase space, $p_A(\gamma)$ and $p_B(\gamma)$, which are the probability densities for phase-space points ($\gamma \in \Gamma$) sampled in the A and B systems, respectively. Information theory defines the relative entropy, actually two such quantities, which we designate s_A and s_B

$$s_A = \int_{\Gamma} d\gamma p_A(\gamma) \ln\left[\frac{p_A(\gamma)}{p_B(\gamma)}\right],$$
$$s_B = \int_{\Gamma} d\gamma p_B(\gamma) \ln\left[\frac{p_B(\gamma)}{p_A(\gamma)}\right]. \tag{5}$$

Although simply expressed and similar in form to the Gibbs entropy formula, such quantities are not often encountered in a statistical-mechanics context.[16]

It is not convenient to measure the systems' multidimensional phase distributions. Fortunately these relative entropies can be expressed in terms of more accessible quantities. Given that $p_A(\gamma)=e^{-\beta U_A(\gamma)}/Q_A$, where Q_A is the partition function of the A system, and likewise for system B, Eq. (5) can be expressed as

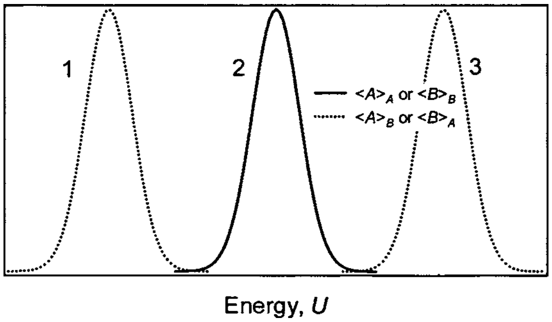

FIG. 3. Illustration of cases that might be observed for the total-energy distributions, which correspond to three boundary values examined by the overlap integral method. The method is based on the comparison of the distributions $\langle A\rangle_A$ and $\langle A\rangle_B$, or $\langle B\rangle_B$ and $\langle B\rangle_A$.

$$s_A = \langle \beta W\rangle_A - \beta\Delta F,$$
$$s_B = -\langle \beta W\rangle_B + \beta\Delta F, \tag{6}$$

for which the difference on the right of each equation is the dissipated work. It is also easy to show that these exact quantities are obtained when the relative-entropy metric is defined in terms of the work distributions:

$$s_A = \int dW p_A(W) \ln\left[\frac{p_A(W)}{p_B(W)}\right],$$
$$s_B = \int dW p_B(W) \ln\left[\frac{p_B(W)}{p_A(W)}\right], \tag{7}$$

from which Eq. (6) follows upon introducing the relation[10] $p_A(W)e^{-\beta W}=p_B(W)e^{-\beta\Delta F}$.

The quantities s_A and s_B are non-negative, and equal to zero only when the two distributions are identical. Larger values of s_A and s_B mean more difference or more distance between the two distributions, corresponding to an increasing degree of nonoverlap of the two systems and indicating larger energetic barriers to free-energy sampling. As the ratio of s_A to s_B increases, the two systems' phase spaces become more and more asymmetric, and entropic barriers will dominate the samplings. Here the system with the smaller value of s has a smaller important phase space, and will be a subset of the other as long as the values are not both very large. As the value of s_A approaches the value of s_B, the two distributions become more and more symmetric, which will result in the symmetric bias for FEP calculations performed in the two directions.

Thus knowledge of the dissipated work is sufficient to determine the relative entropies, and from them the nature of the phase-space overlap. Unfortunately the dissipated work requires knowledge of the free-energy difference that is the aim of the calculation. We must consider the usefulness of this overlap measure with consideration to the fact that its measurement is subject to some degree of inaccuracy.

IV. OVERLAP MEASURES AND BIAS IN FREE-ENERGY CALCULATIONS

We have no rigorous connection between the overlap measures introduced above, and the inaccuracy that should

TABLE I. Parameter sets for the systems of N independent particles in a harmonic potential. In all cases, $N=10$, $\omega_A=1$, and $\beta=1$.

Cases	ω_B/ω_A	x_0	$\beta\Delta F$	s_A	s_B	K_{AB}	K_{BA}
a	1	0	0	0	0	1	1
b	1	1	0	10	10	0.05	0.05
c	1	3	0	90	90	0	0
d	5	0	8.05	12	4	0.02	1.98
e	5	1	8.05	62	14	10^{-5}	0.06
f	5	3	8.05	462	94	0	0
g	20	0	14.98	80	10	5×10^{-5}	2
h	20	1	14.98	280	20	0	0.06
i	20	2	14.98	880	50	0	0

be expected in a free-energy calculation applied to a given system. So to employ these measures for practical purposes, we must calibrate and test them in some simple applications. To this end we examine the multiharmonic model, which is defined as a system of N independent particles in a harmonic potential.[14]

Specifically we define the systems A and B according to the potential-energy functions,

$$U_A=\sum_{i=1}^{N}\omega_A x_i^2,$$
$$U_B=\sum_{i=1}^{N}\omega_B(x_i-x_0)^2, \tag{8}$$

where x_i is the coordinate of particle i, and x_0, ω_A, and ω_B are model parameters. This is a simple but nontrivial system for which many properties can be derived analytically. In particular, the free-energy difference is independent of x_0: $\Delta F=1/2NkT\ln(\omega_B/\omega_A)$, and the relative entropies are

$$s_A=-\frac{1}{2}N\ln R+NRX+\frac{1}{2}N(R-1),$$
$$s_B=\frac{1}{2}N\ln R+NX+\frac{1}{2}N(1/R-1), \tag{9}$$

where $R\equiv\omega_B/\omega_A$ and $X=\beta\omega_A x_0^2$. In previous work[14] we introduced the "neglected-tail model" for the FEP bias and showed that it gives a very good description of the inaccuracy of a FEP calculation for the multiharmonic model. Here we choose to study nine cases of the model, for which the parameters and properties are listed in Table I. The choices span the range of systems depicted in Fig. 1.

The dimension of the phase space for this one-dimensional system is equal to N, but because the particles are uncorrelated we can get a picture of the A-B phase-space relation by plotting the distributions for a single-particle coordinate. This is done in Fig. 4.

Distributions in the first row are symmetric. In subsequent rows the two distributions become less symmetric, with the B-system distribution becoming more and more narrow than the A-system distribution. From left to right, the columns arrange systems to increasing separation of the A and B phase spaces.

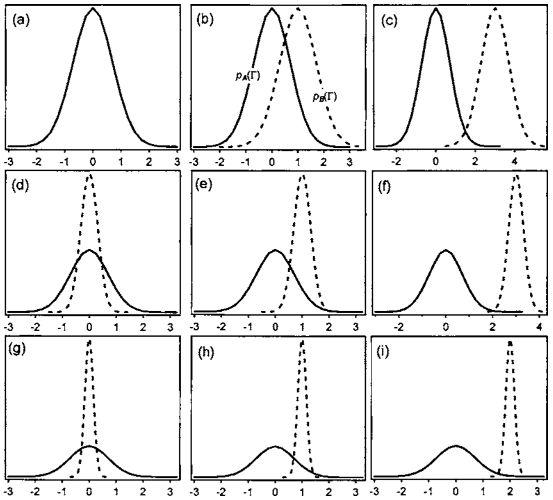

FIG. 4. One-particle phase-space distributions for the multiharmonic model, given for the nine cases listed in Table I. The solid and dashed lines are for a particle influenced by the A and B potentials, respectively, as defined in Eq. (8). Ordinate scales are such that the curves are normalized to unity.

Figure 5 shows the work distributions that would be expected for FEP calculations performed in the two directions for each of the nine system pairs; these distributions can be calculated using analytic formulas developed previously.[14] The layout of the plots corresponds to those in Fig. 4 and give the relative-entropy and overlap integral measurements in Table I. These overlap measurements are consistent with the phase-space overlap relations illustrated in Fig. 4.

Now we calculate the free energies for these nine cases from two directions $A\rightarrow B$ and $B\rightarrow A$, and plot the bias as a function of the number of samplings M in Fig. 6. Data were obtained by Monte Carlo simulation of the multiharmonic

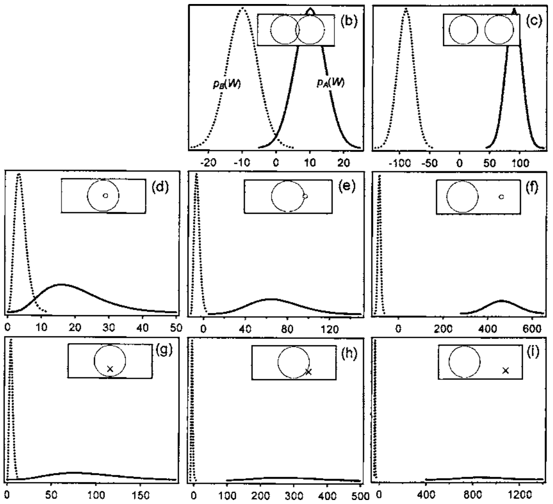

FIG. 5. Work distributions expected for free-energy perturbations applied to the multiharmonic model for the nine cases listed in Table I [work is always zero for case (a) and is not shown]. The solid and dashed lines are for perturbations $A\rightarrow B$ and $B\rightarrow A$, respectively. Ordinate scales are such that the curves are normalized to unity. The insets provide a semiquantitative illustration (roughly corresponding to the K_{AB} values) of the A-B phase-space relation, in the manner shown in Fig. 1. The A-important space is represented by the larger circle, and the B-important space by the smaller one (or by an x if the space is very small). The relative size and amount of overlap of the circles are indicative of the relation of the important phase spaces.

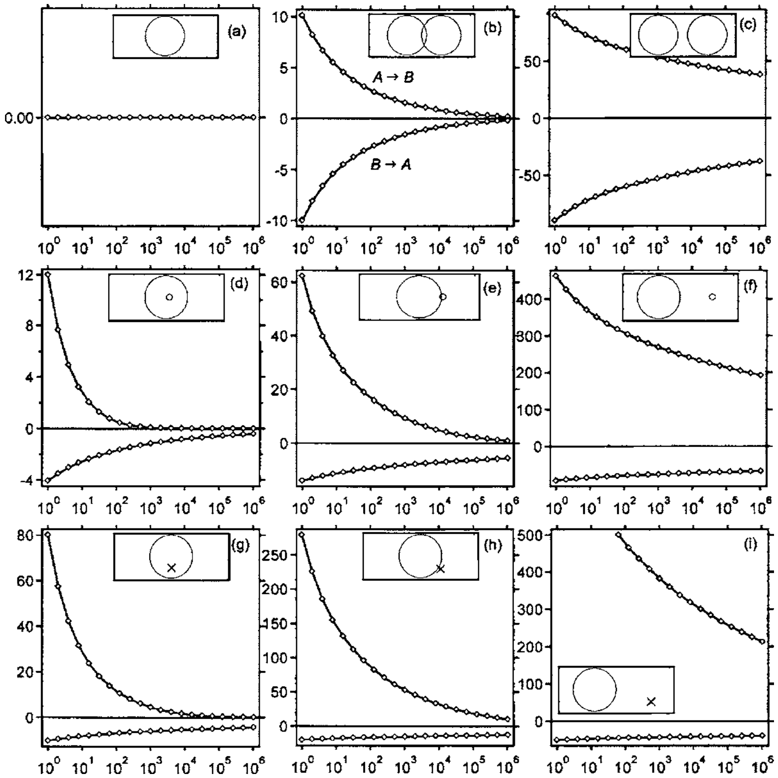

FIG. 6. Expected bias in the free energy for the nine cases of the multiharmonic model listed in Table I. Bias is that expected for FEP calculations in the directions $A \rightarrow B$ (upper curve in each figure) and $B \rightarrow A$ (lower curve). Bias in ΔF (in units of kT) is plotted against the number of perturbation samples M. The inset illustrations are as described in Fig. 5.

system, performing M FEP measurements to obtain the free energy. This free-energy measurement was performed 8000 times for each case, and the bias is given as the average of these measurements, minus the correct free energy. In the first row of the figure, the bias is symmetric with respect to the perturbation direction—the sign of the inaccuracy changes, but the magnitude of the error does not. This is easily understood from the symmetry of the phase-space and work distributions. Moving to the second and third rows we see progressively increasing asymmetry in the bias with respect to the perturbation direction. This reflects the relative effects of energetic versus entropic barriers, wherein the bias in one direction owes more to entropic barriers and is more easily overcome by added sampling, while in the other direction the barrier is increasingly energetic and the opposite holds.

Moving from left to right in the columns, we see the effect on the calculations of increasing energetic barriers, which are introduced as the important phase spaces move apart. The qualitative picture does not change across a row, but the magnitude of the bias increases considerably (taking note of the scale of the ordinate in each plot).

In our study of bias in Gaussian-distributed work distributions,[13] we proposed the following quantity as an effective measure of the amount of sampling:

$$\Pi = \sqrt{\mathbf{W}_L\left[\frac{1}{2\pi}(M-1)^2\right]} - \beta\sigma. \tag{10}$$

Here $\mathbf{W}_L(x)$ is the Lambert W function, defined as the solution for w in $x = we^w$. M is the number of work trials used for the free-energy calculation, and σ is the standard deviation of the Gaussian work distribution. In our previous study[13] we showed that when plotted as a function of this quantity the dependence of the free-energy bias on the amount of sampling collapses onto a single curve, such that the bias can be considered unimportant for Π approximately equal to (or greater than) zero. This can be a very useful result, in that it can be applied to tell us whether a given amount of sampling is sufficient to render an accurate free energy. The outcome has two limitations. First, application of this heuristic requires (through σ) knowledge of the free energy being measured, and inaccuracies in the measured free energy could potentially invalidate the application of a heuristic based on the value of Π. Fortunately we could demonstrate that any expected free-energy bias would not be sufficient to cause Π to take on a value that would indicate sufficient sampling when sampling was, in fact, insufficient. In this sense the heuristic is "fail-safe." More limiting then is the restriction of this result to Gaussian work distributions, for which the bias is symmetric. We now consider how to extend the definition of Π to handle the more general case. We find that the relative entropies s_A and s_B to be appropriate quantities to use for this extension.

Figure 6 clearly shows that the amount of sampling needed to obtain an accurate free energy depends upon the direction of the perturbation, $A \rightarrow B$ vs $B \rightarrow A$, so the generalized definition of the scaled sampling amount must reflect this asymmetry, and reduce to a direction-independent form in the Gaussian-work limit. The quantity σ in Eq. (10) is related to the dissipated work, as are s_A and s_B shown in Eq. (6). Indeed, in the Gaussian-work limit $s_A = s_B = (\beta\sigma)^2/2$, suggesting that we can obtain a direction-dependent generalization of Π by substituting for σ in Eq. (10) a term involving s_A or s_B. We can obtain further guidance by considering earlier studies[17] that examined highly asymmetric FEP calculations on systems obeying the phase-space subset rela-

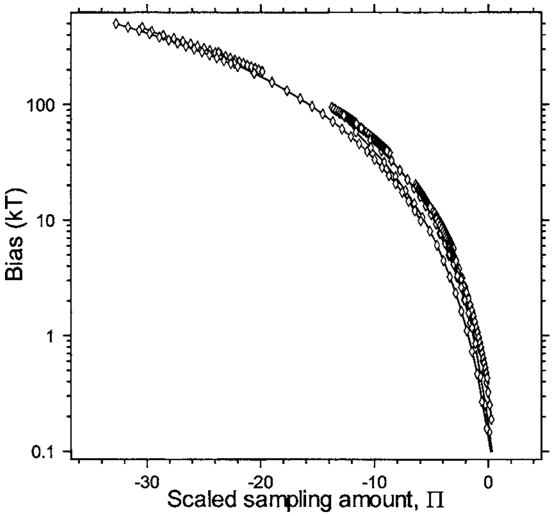

FIG. 7. Expected bias (in absolute value) in the free energy for the nine cases of the multiharmonic model listed in Table I, plotted against the scaled sampling amount defined in Eq. (11). All results presented in Fig. 6 are given here, including both $A \rightarrow B$ and $B \rightarrow A$ directions.

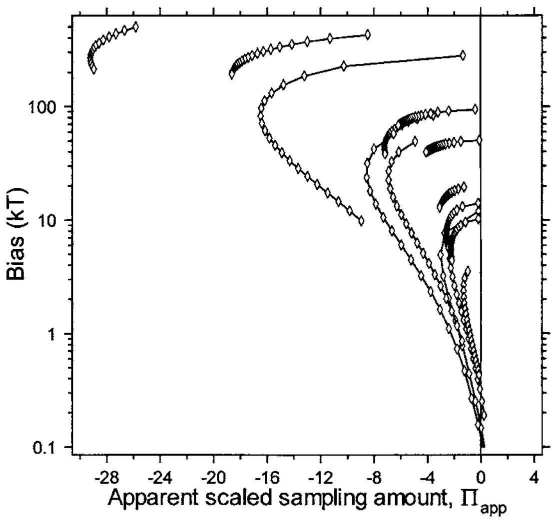

FIG. 8. The same expected biases in Fig. 7 are plotted against the apparent scaled sampling amount Π_{app} defined in Eq. (13).

tion. There it was shown that the group $M \exp(\Delta S/k)$ provided a good characterization of the amount of sampling, where $\Delta S = S_B - S_A < 0$ is the (thermodynamic) entropy difference between the B and A systems, and the perturbation is performed in the insertion direction $A \rightarrow B$. For the types of systems examined in Ref. 17, $\Delta S \approx -s_B$. If we note also that for large x, $\mathbf{W}_L(x) \sim \ln x$, we find that to connect to the present analysis, the generalized definition of Π should depend on the group $\mathbf{W}_L / s_B$.

A definition of Π that is consistent with these two special cases (viz., Gaussian work and phase-space subset) is

$$\begin{aligned}\Pi_{A\rightarrow B} &= \sqrt{\frac{s_A}{s_B}\mathbf{W}_L\left[\frac{1}{2\pi}(M-1)^2\right]} - \sqrt{2 s_A},\\ \Pi_{B\rightarrow A} &= \sqrt{\frac{s_B}{s_A}\mathbf{W}_L\left[\frac{1}{2\pi}(M-1)^2\right]} - \sqrt{2 s_B}.\end{aligned} \tag{11}$$

In Fig. 7 we plot the inaccuracies observed for the nine multiharmonic test cases introduced above, considering perturbations in both directions. We see that the data presented in Fig. 6 to a very good approximation all collapse onto a single curve. The bias drops off steeply as the sampling parameter Π goes through zero, indicating that the sign of Π provides a good indicator of sufficient sampling. We note that a factor of $s_A^{1/2}$ (or $s_B^{1/2}$ for $\Pi_{B\rightarrow A}$) could be separated out of the definition, and the sign of Π will be determined by the remaining term. However, except for the drop-off near zero on the abscissa, this term does not exhibit the apparently universal form demonstrated when using Eq. (11), so it is not preferred except in cases where s_A is effectively infinite (such as for hard-sphere insertion).

Finally, we consider the practical question of whether bias in the free energy could lead to inaccuracy in Π sufficient to give a false indicator of inaccuracy. If the perturbation is performed in the direction $A \rightarrow B$, insufficient sampling will yield a biased free energy that *overestimates* the true free energy (cf. Fig. 6). Thus, according to Eq. (6) s_B will also be overestimated (provided that $\langle \beta W \rangle_B$ is measured accurately, which can be reasonably assumed). After factoring out $s_A^{1/2}$, the sign of $\Pi_{A\rightarrow B}$ is determined by $\sqrt{\mathbf{W}_L / s_B} - \sqrt{2}$. An overestimate of s_B will tend to make this difference more negative, and therefore will not promote a false conclusion of sufficient sampling. A similar argument can be made for perturbation in the reverse direction. Thus we conclude that the rule $\Pi > 0$ provides a heuristic for sufficient sampling that is fail-safe with respect to the influence on it of bias in the free energy.

As a further demonstration of the fail-safe nature of the heuristic, we define

$$\begin{aligned}s_{A,A\rightarrow B} &= \langle \beta W \rangle_A - \beta(\Delta F + B_{A\rightarrow B}),\\ s_{B,A\rightarrow B} &= -\langle \beta W \rangle_B + \beta(\Delta F + B_{A\rightarrow B}),\\ s_{A,B\rightarrow A} &= \langle \beta W \rangle_A - \beta(\Delta F + B_{B\rightarrow A}),\\ s_{B,B\rightarrow A} &= -\langle \beta W \rangle_B + \beta(\Delta F + B_{B\rightarrow A}).\end{aligned} \tag{12}$$

Here $B_{A\rightarrow B}$ and $B_{B\rightarrow A}$ are the corresponding bias obtained when sampling system A and system B, respectively. The "apparent" relative entropies given by Eq. (12) are not the true relative entropies, but are meant to describe the values one would compute when using the simulation data itself, and thus include the effects of any bias in the free energy. Similar to our previous study,[13] we define the apparent scaled sampling amount

$$\begin{aligned}\Pi_{A\rightarrow B,\text{app}} &= \sqrt{\frac{s_{A,A\rightarrow B}}{s_{B,A\rightarrow B}}\mathbf{W}_L\left[\frac{1}{2\pi}(M-1)^2\right]} - \sqrt{2 s_{A,A\rightarrow B}},\\ \Pi_{B\rightarrow A,\text{app}} &= \sqrt{\frac{s_{B,B\rightarrow A}}{s_{A,B\rightarrow A}}\mathbf{W}_L\left[\frac{1}{2\pi}(M-1)^2\right]} - \sqrt{2 s_{B,B\rightarrow A}}.\end{aligned} \tag{13}$$

We plot in Fig. 8 the bias observed for the nine multiharmonic test cases as a function of the apparent Π value. There we see a behavior that is similar to that we observed[13] for the symmetric-bias special case in which the work distribution is

Gaussian. The important feature of those results also seen here: application of the bias-detection heuristic *using the most likely (biased) free energy itself*, does not give a false indication of a bias-free result ($\Pi > 0$) when bias is, in fact, present; the apparent Π does not cross zero except where the bias becomes small. One should note though that in the limit of very small sampling the apparent Π approaches zero, and one should take care not to construe this case as giving evidence of an unbiased result. The situation arises only for M less than about 4, and a signal for it is the rapid *decrease* of the apparent Π with added sampling. Application of the bias-detection heuristic to identify a bias-free result should of course take note that sampling is not in this regime.

V. CONCLUSIONS

We have presented two methods to quantify overlap of the important phase spaces of a pair of systems, and we have shown how one of these measures can be used in a heuristic to detect bias in free energies calculated by molecular simulation. Although development has been formulated in terms of phase-space overlap, the relative-entropy metrics can be expressed purely in terms of the work distributions. Consequently we expect the heuristic to apply both to single-stage FEP calculations as well as to more general NEW calculations.

To summarize, the "recipe" for determining if a work-based free-energy measurement is free of bias is as follows.

- Perform M nonequilibrium work measurements, collecting work values $\{W_{i,A\rightarrow B}\}$; do the same in the direction $B\rightarrow A$, collecting M work values $\{W_{i,B\rightarrow A}\}$ (M may be different for the two directions, but to simplify notation we will take them to be equal).
- Evaluate the average work, $\bar{W}_{A\rightarrow B}=M^{-1}\Sigma_{i=1}^{M}W_{i,A\rightarrow B}$, and likewise for $B\rightarrow A$.
- Evaluate the free-energy difference, $\beta\Delta F_{A\rightarrow B}=-\ln[M^{-1}\Sigma_{i=1}^{M}\exp(-\beta W_{i,A\rightarrow B})]$, and $\beta\Delta F_{B\rightarrow A}=\ln[M^{-1}\Sigma_{i=1}^{M}\exp(-\beta W_{i,B\rightarrow A})]$ (for both cases ΔF is defined F_B-F_A and W_i is defined as described in Eq. (1), for configuration i).
- Evaluate $s_A=\beta\bar{W}_{A\rightarrow B}-\beta\Delta F_{B\rightarrow A}$, and $s_B=-\beta\bar{W}_{B\rightarrow A}+\Delta F_{A\rightarrow B}$. In principle either free-energy estimate can be employed in the calculation of s_A and s_B, but use of the indicated estimates is recommended to guard against adverse effects of free-energy bias on the heuristic.
- Evaluate $\Pi_{A\rightarrow B}$ and $\Pi_{B\rightarrow A}$ according to Eq. (11). For each case, a positive value indicates that the corresponding free-energy estimate is free of bias; otherwise more sampling is needed. Of course, an accurate result from only one direction is required.

One could compute the Π values using the procedure outlined in Eqs. (12) and (13), but we find that the simpler approach outlined here for evaluating s_A and s_B has some advantage.

Application of the heuristic highlights the necessity to obtain work results from both directions to ensure that the free energy is measured accurately by at least one of them. Careful practitioners have long been aware that the accuracy of a NEW free-energy calculation cannot be ascertained without information from calculations performed in the "reverse" direction. However, the common practice has been either to average differing results, or to increase sampling until results from the two directions agree. Examples from the multiharmonic model clearly show that the former practice is still highly prone to error, while the latter can be very inefficient. With the proposed heuristic, one can determine which result from the two directions is more accurate, and whether one is sufficiently accurate even though it disagrees with the other.

As a practical note, it is important to remember that another manifestation of inadequate sampling is poor precision. It is possible for a fluctuation in an average to result in a positive value of Π, falsely indicating a bias-free result. Thus a conclusion about whether Π is non-negative should be made with consideration of its confidence limits. These are easily gauged in the usual manner, examining the scatter in several independent measurements.

Also one should remember that the heuristic $\Pi > 0$ is only a rule of thumb, and to be safe one should attempt to apply sampling sufficient for Π to be well above zero. Indeed our study of Gaussian work suggested $\Pi > 0.5$ as an appropriate rule, showing that there is some slack in the specification. The heuristic is not derived as a rigorous result, but it has now been demonstrated to be effective for a wide variety of model cases. As it is applied more in practice, we can develop a clearer picture of its effectiveness for a greater variety of situations, and in what manner it can be best used.

The present study has considered only single-stage NEW calculations. In many cases accurate results cannot be obtained from single-stage calculations for any feasible amount of sampling, in either direction. In these cases multistage methods are needed. Multistage methods are of course built from single-stage calculations, and the accuracy of each stage of the calculation can (and should) be evaluated using these methods. Moreover, the overlap relations can be useful in other ways in this context. In subsequent work we will consider how knowledge of phase-space overlap can be applied to help choose an appropriate staging method. Overlap relations may also prove useful to the design of pathways for free-energy calculations.

ACKNOWLEDGMENTS

This work is supported by the Division of Chemical Sciences, Office of Basic Energy Sciences of the U.S. Department of Energy (Contract No. DE-FG02-96ER14677). Computer resources have been provided by the University at Buffalo Center for Computational Research. We also acknowledge Chris Jarzynski for suggesting the use of information-theoretic measures to quantify phase-space overlap.

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic, San Diego, 2002).
[2] D. A. Kofke, Fluid Phase Equilib. **228–229**, 41 (2005).
[3] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).

[4] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).

[5] Z. W. Salsburg, J. D. Jacobson, W. Fickett, and W. W. Wood, J. Chem. Phys. **30**, 65 (1959); I. R. McDonald and K. Singer, Discuss. Faraday Soc. **43**, 40 (1967); W. G. Hoover and F. H. Ree, J. Chem. Phys. **49**, 3609 (1968); J. P. Valleau and D. N. Card, *ibid.* **57**, 5457 (1972); S. Romano and K. Singer, Mol. Phys. **37**, 1765 (1979); K. S. Shing and K. E. Gubbins, *ibid.* **46**, 1109 (1982); D. Frenkel and A. J. C. Ladd, J. Chem. Phys. **81**, 3188 (1984); T. P. Straatsma and H. J. C. Berendsen, *ibid.* **89**, 5876 (1988); A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988); **63**, 1195 (1989); J. I. Siepmann, Mol. Phys. **70**, 1145 (1990); B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992); S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992); A. P. Lyubartsev, A. A. Marsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1992); W. P. Reinhardt and J. E. Hunter III, *ibid.* **97**, 1599 (1992); P. Kollman, Chem. Rev. (Washington, D.C.) **93**, 2395 (1993); J. Lee, Phys. Rev. Lett. **71**, 211 (1993); B. Smit, S. Karaborni, and J. I. Siepmann, Macromol. Symp. **81**, 343 (1994); T. C. Beutler, D. R. Beguelin, and W. F. van Gunsteren, J. Chem. Phys. **102**, 3787 (1995); A. Dejaegere and M. Karplus, J. Phys. Chem. **100**, 11148 (1996).

[6] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976); G. M. Torrie and J. P. Valleau, *ibid.* **23**, 187 (1977); M. Watanabe and W. P. Reinhardt, Phys. Rev. Lett. **65**, 3301 (1990).

[7] R. J. Radmer and P. A. Kollman, J. Comput. Chem. **18**, 902 (1997); A. D. Bruce, N. B. Wilding, and G. J. Ackland, Phys. Rev. Lett. **79**, 3002 (1997); G. E. Crooks, J. Stat. Phys. **90**, 1481 (1998); Phys. Rev. E **61**, 2361 (2000); W. P. Reinhardt, M. A. Miller, and L. M. Amon, Acc. Chem. Res. **34**, 607 (2001); F. G. Wang and D. P. Landau, Phys. Rev. E **64**, 056101 (2001); Phys. Rev. Lett. **86**, 2050 (2001); G. Hummer, J. Chem. Phys. **114**, 7330 (2001); G. Hummer and A. Szabo, Proc. Natl. Acad. Sci. U.S.A. **98**, 3658 (2001); P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002); T. Simonson, G. Archontis, and M. Karplus, Acc. Chem. Res. **35**, 430 (2002); J. S. Wang and R. H. Swendsen, J. Stat. Phys. **106**, 245 (2002); M. Athenes, Phys. Rev. E **66**, 046705 (2002); C. Jarzynski, *ibid.* **65**, 046122 (2002); W. F. Van Gunsteren, X. Daura, and A. E. Mark, Helv. Chim. Acta **85**, 3113 (2002); D. M. Zuckerman and T. B. Woolf, Phys. Rev. Lett. **89**, 180602 (2002); S.-T. Lin, M. Blanco, and W. A. Goddard III, J. Chem. Phys. **119**, 11792 (2003); A. D. Bruce and N. B. Wilding, AIP Conf. Proc. **127**, 1 (2003); S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten, J. Chem. Phys. **119**, 3559 (2003); D. J. Evans, Mol. Phys. **101**, 1551 (2003); J. R. Errington, J. Chem. Phys. **118**, 9915 (2003); M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos, *ibid.* **119**, 9406 (2003); S. X. Sun, *ibid.* **118**, 5769 (2003); Q. L. Yan and J. J. de Pablo, Phys. Rev. Lett. **90**, 035701 (2003); J. Gore, F. Ritort, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A. **100**, 12564 (2003); M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003); D. Rodriguez-Gomez, E. Darve, and A. Pohorille, J. Chem. Phys. **120**, 3563 (2004); M. Athenes, Eur. Phys. J. B **38**, 651 (2004); F. M. Ytreberg and D. M. Zuckerman, J. Comput. Chem. **25**, 1749 (2004); D. M. Zuckerman and T. B. Woolf, J. Stat. Phys. **114**, 1303 (2004); R. Radhakrishnan and T. Schlick, J. Chem. Phys. **121**, 2436 (2004); C. Peter, C. Oostenbrink, A. van Dorp, and W. F. van Gunsteren, *ibid.* **120**, 2652 (2004); S. Wan, R. H. Stote, and M. Karplus, *ibid.* **121**, 9539 (2004); R. P. White and H. Meirovitch, *ibid.* **121**, 10889 (2004); D. Wu and D. A. Kofke, J. Chem. Phys. **122**, 204104 (2005).

[8] D. A. Kofke and P. T. Cummings, Mol. Phys. **92**, 973 (1997).

[9] C. Jarzynski, Phys. Rev. E **56**, 5018 (1997); Phys. Rev. Lett. **78**, 2690 (1997).

[10] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[11] D. A. Kofke and P. T. Cummings, Fluid Phase Equilib. **151**, 41 (1998); N. D. Lu and D. A. Kofke, J. Chem. Phys. **111**, 4414 (1999); **114**, 7303 (2001); D. A. Kofke, Mol. Phys. **102**, 405 (2004).

[12] N. Lu, J. Adhikari, and D. A. Kofke, Phys. Rev. E **68**, 026122 (2003).

[13] D. Wu and D. A. Kofke, J. Chem. Phys. **121**, 8742 (2004).

[14] D. Wu and D. A. Kofke, Phys. Rev. E **70**, 066702 (2004).

[15] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Illinois, 1963); R. M. Gray, *Entropy and Information Theory* (Springer, New York, 1990); *C.E. Shannon: Collected Papers*, edited by N. J. A. Sloane and A. D. Wyner (IEEE, New York, 1993).

[16] H. Qian, arXiv.org/abs/math-ph/0007010/ (2000).

[17] N. D. Lu and D. A. Kofke, J. Chem. Phys. **115**, 6866 (2001).