Targeted free energy perturbation

C. Jarzynski*
Complex Systems, T-13, MS B213 Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 4 June 2001; published 3 April 2002)

In this paper generalization of the free energy perturbation identity is derived, and a computational strategy based on this result is presented. A simple example illustrates the efficiency gains that can be achieved with this method.

DOI: 10.1103/PhysRevE.65.046122 PACS number(s): 05.70.Ce, 05.10.−a

The development of efficient methods for the numerical estimation of free energy differences remains an outstanding problem in the computational sciences [1], with applications as diverse as rational drug design [2], *ab initio* prediction of material properties [3], and the study of condensates in nonperturbative QCD [4]. Many schemes for estimating free energy differences trace their origins to the *perturbation identity* [5]

$$\langle e^{-\Delta E/kT}\rangle_A = e^{-\Delta F/kT}. \tag{1}$$

Here, $\Delta F = F_B - F_A$ is the Helmholtz free energy difference between two equilibrium states *A* and *B*, of a system, defined at a common temperature *T* but different settings of external parameters. The variable $\mathbf{x}$ (and later $\mathbf{y}$) denotes a microstate of the system, e.g., a point in configuration space or phase space; $E_A(\mathbf{x})$ and $E_B(\mathbf{x})$ denote the internal energy as a function of microstate, for the two parameter settings; and

$$\Delta E(\mathbf{x}) \equiv E_B(\mathbf{x}) - E_A(\mathbf{x}) \tag{2}$$

is the energy difference associated with changing the external parameters from one setting to the other, while holding fixed the microstate. Finally, $\langle \cdots \rangle_A$ denotes an average over microstates sampled from the canonical distribution representing state *A*.

The traditional perturbation approach to estimating ΔF is a direct application of Eq. (1): the quantity $\exp(-\Delta E/kT)$ is averaged over microstates sampled from ensemble *A* [6]. However, this method converges poorly when there is a little overlap in configuration space between ensembles *A* and *B*. Intuitively, this makes sense: if there is a little overlap, then we very slowly accumulate information about state *B* by generating microstates typical of state *A*.

The aim of this paper is to present a generalization of Eq. (1), as well as a computational method, *targeted free energy perturbation*, based on this result. The practitioner of this method must attempt to construct an invertible transformation $\mathcal{M}$, under which ensemble *A* gets mapped onto an ensemble A' that overlaps significantly with *B* [see Eqs. (13) and (14)]. The more successful this attempt, the more rapidly the method converges. Indeed, if A' and *B* overlap perfectly, then convergence is immediate. This strategy thus provides a mechanism for taking advantage of prior knowledge about states *A* and *B* (used to construct the mapping $\mathcal{M}$) in order to speed up the estimation of ΔF.

While, to the best of our knowledge the central result and method derived below are new, the use of invertible mappings to enhance the efficiency of free energy calculations has precedents. For simple displacements, $\mathbf{x} \to \mathbf{x} + \mathbf{d}$, the method proposed herein is closely related to one developed years ago by Voter [7], for energy functions E_A and E_B, which resemble one another apart from a spatial translation. Bruce *et al.* [8] have proposed the use of invertible transformations as collective Monte Carlo moves—"lattice switches"—to enable the sampling of disparate regions of configuration space. Finally, our method is similar in spirit to the metric scaling scheme developed by Miller and Reinhardt [9], whereby one attempts to "guide" the system in question through a continuous sequence of equilibrium states, by dynamically, and linearly, distorting the space in which the constituent particles evolve.

We now derive our central result, Eq. (10), below.

Consider an invertible transformation of configuration space onto itself:

$$\mathcal{M}: \mathbf{x} \to \mathbf{y}(\mathbf{x}). \tag{3}$$

This might be a displacement as in Ref. [7], or perhaps a scaling transformation as in Ref. [9], or it might be a considerably more complicated, nonlinear mapping. Now imagine microstates $\mathbf{x}_1, \mathbf{x}_2, \ldots$ sampled from some (so far, arbitrary) *primary* ensemble, represented by a distribution $\rho(\mathbf{x})$; and construct their images under the transformation: $\mathbf{y}_1, \mathbf{y}_2, \ldots$, where $\mathbf{y}_n = \mathcal{M}(\mathbf{x}_n)$. The $\mathbf{y}$'s are, effectively, sampled from a *secondary* ensemble, which is the image of the primary ensemble under $\mathcal{M}$. This secondary ensemble is represented by a distribution $\eta(\cdot)$, related to the primary distribution $\rho(\cdot)$ by

$$\eta(\mathbf{y}) = \rho(\mathbf{x})/J(\mathbf{x}), \tag{4}$$

where $J(\mathbf{x}) = |\partial \mathbf{y}/\partial \mathbf{x}|$ is the Jacobian of the mapping $\mathcal{M}$. Here and henceforth, when the variables $\mathbf{x}$ and $\mathbf{y}$ appear together, it will be understood that they are related by $\mathbf{y} = \mathcal{M}(\mathbf{x})$.

Next, define a function

$$\Phi(\mathbf{x}) \equiv E_B(\mathbf{y}) - E_A(\mathbf{x}) - kT \ln J(\mathbf{x}), \tag{5}$$

let the primary ensemble be the canonical distribution corresponding to state *A*

*Email address: chrisj@lanl.gov

$$\rho(\mathbf{x}) = \frac{1}{Z_A} e^{-E_A(\mathbf{x})/kT}, \tag{6}$$

and evaluate the average of $\exp(-\Phi/kT)$ over points $\mathbf{x}$ sampled from this ensemble

$$\langle e^{-\Phi/kT}\rangle_A = \int d\mathbf{x}\rho(\mathbf{x}) e^{-\Phi(\mathbf{x})/kT} \tag{7}$$

$$= \frac{1}{Z_A}\int d\mathbf{x} J(\mathbf{x}) e^{-E_B(\mathbf{y})/kT} \tag{8}$$

$$= \frac{1}{Z_A}\int d\mathbf{y} e^{-E_B(\mathbf{y})/kT} = \frac{Z_B}{Z_A}, \tag{9}$$

where Z_A and Z_B are partition functions. (Note the change in the variable of integration: $\int J\, d\mathbf{x} \cdots = \int d\mathbf{y} \cdots$.) Invoking the relation $F = -kT \ln Z$, we finally obtain

$$\langle e^{-\Phi/kT}\rangle_A = e^{-\Delta F/kT}. \tag{10}$$

Let us now turn our attention to the application of this result to the problem of estimating ΔF.

Equation (10) generalizes the free energy perturbation identity, reducing to the latter in the special case $\mathcal{M}:\mathbf{x}\to\mathbf{x}$. However, Eq. (10) is valid for arbitrary invertible transformations $\mathcal{M}$. *It is plausible that one can take advantage of this generality to enhance the efficiency of computing* ΔF. That is, there may exist mappings $\mathcal{M}$ for which the average of $\exp(-\Phi/kT)$ converges more rapidly than the average of $\exp(-\Delta E/kT)$.

To investigate this possibility, consider $p(\phi|\mathcal{M})$, the distribution of values of $\phi = \Phi(\mathbf{x})$, for $\mathbf{x}$ sampled from A [Eq. (6)]. Equation (10) asserts that

$$\int d\phi p(\phi|\mathcal{M}) e^{-\phi/kT} = e^{-\Delta F/kT}, \tag{11}$$

for any choice of $\mathcal{M}$. In practice, we estimate ΔF by averaging $\exp(-\Phi/kT)$ over a finite number of sampled microstates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

$$\frac{1}{N}\sum_{n=1}^{N} e^{-\phi_n/kT} \approx e^{-\Delta F/kT}, \tag{12}$$

where $\phi_n \equiv \Phi(\mathbf{x}_n)$. This approximation becomes an equality as $N\to\infty$, but the rate of convergence depends strongly on the choice of $\mathcal{M}$; roughly speaking, the narrower the distribution $p(\phi|\mathcal{M})$, the faster the convergence. Therefore, we are faced with the practical question, how do we choose $\mathcal{M}$ so as to maximize the rate of convergence of the left side of Eq. (12)?

Recall that a poor convergence of the usual perturbation method is a symptom of too little overlap between A and B in configuration space: we then learn little about B when sampling from A. Now note that, when generating the sequence of ϕ_n's, we are effectively harvesting information from *two* ensembles; the points $\mathbf{x}$ sample the primary ensemble A, while the points $\mathbf{y}$ sample the secondary ensemble A', which is the image of A under the transformation $\mathcal{M}$,

$$A \overset{\mathcal{M}}{\to} A'. \tag{13}$$

Intuition suggests that, since the primary ensemble represents state A (by construction), we ought to attempt to maximize the overlap between the secondary ensemble and state B

$$A' \approx B \tag{14}$$

so as to speedily gain information about both of the equilibrium states (A and B) that interest us.

Pursuing this line of intuition, let us first consider the extreme case, and define a *perfect* transformation $\mathcal{M}^*$ to be one under which A maps exactly onto B,

$$\rho(\mathbf{x}) = \frac{1}{Z_A} e^{-E_A(\mathbf{x})/kT} \xrightarrow{\mathcal{M}^*} \eta(\mathbf{y}) = \frac{1}{Z_B} e^{-E_B(\mathbf{y})/kT}. \tag{15}$$

By Eq. (4), this implies

$$F_B - E_B(\mathbf{y}) = F_A - E_A(\mathbf{x}) - kT \ln J(\mathbf{x}), \tag{16}$$

in other words $\Phi(\mathbf{x}) = \Delta F$ *for all* $\mathbf{x}$. Hence, we have a maximally narrow distribution of values of ϕ

$$p(\phi|\mathcal{M}^*) = \delta(\phi - \Delta F). \tag{17}$$

Thus, *the convergence of Eq. (12) is immediate if the transformation is perfect*: $\Phi(\mathbf{x}) = \Delta F$ for every sampled $\mathbf{x}$.

Unfortunately, constructing a perfect transformation is likely to be much more difficult than the original problem of computing ΔF. However, it stands to reason that if $p(\phi|\mathcal{M})$ is a δ function when $A' = B$, then it will remain narrow when $A' \approx B$. Equation (17) thus gives credence to our earlier intuition [Eq. (14)]: we ought indeed to look for a transformation under which A' enjoys good overlap with B. A close resemblance between A' and B implies a narrow distribution of ϕ's, which in turn implies rapid convergence of our estimate of ΔF.

Let us summarize what has been stated here before. Equation (10) suggests a method of estimating ΔF: microstates $\mathbf{x}_n$ are sampled from the canonical ensemble A; the value $\phi_n = \Phi(\mathbf{x}_n)$ is computed for each sampled microstate; and the estimator $X_N \equiv (1/N)\Sigma_n \exp(-\phi_n/kT)$ converges to $\exp(-\Delta F/kT)$ as $N\to\infty$. Two ingredients of this scheme are (1) an invertible mapping $\mathcal{M}$, and (2) the image A' of the canonical ensemble A under $\mathcal{M}$. If A' coincides with B, then the method converges immediately. Hence, if we choose a transformation $\mathcal{M}$, which significantly improves the overlap with B, without necessarily being "perfect," then X_N ought to converge more rapidly with N than the traditional free energy perturbation (FEP) estimator, $X_N^{\mathrm{FEP}} \equiv (1/N)\Sigma_n \exp(-\Delta E_n/kT)$. We will refer to this method as *targeted free energy perturbation*, since its successful implementation requires finding a transformation $\mathcal{M}$ for which the secondary ensemble A' comes reasonably close to ‘‘hitting’’ the *target ensemble B*.

Several extensions of targeted free energy perturbation, to be developed more fully elsewhere, are potentially useful. First, for a parameter-dependent energy function $E(\mathbf{x},\lambda)$, the application of Eq. (10), to states A and B defined by infinitesimally different values of λ, leads to the identity

$$\frac{\partial F}{\partial \lambda}=\left\langle \frac{\partial E}{\partial \lambda}+\mathbf{u}\cdot\boldsymbol{\nabla}E-kT\boldsymbol{\nabla}\cdot\mathbf{u}\right\rangle_{\lambda}, \tag{18}$$

where $\mathbf{u}(\mathbf{x})$ is an arbitrary differentiable vector field of bounded magnitude [10]. While this result reduces to the widely used thermodynamic integration identity [11] for $\mathbf{u}=\mathbf{0}$, other choices of $\mathbf{u}$ might accelerate convergence of the average. It is also straightforward to incorporate Eq. (10) into the umbrella sampling [12] and overlapping distributions [13] methods. Finally, a *nonlinear* metric scaling scheme—with particular potential for enhancing the efficiency of free energy calculations based on steered molecular dynamics [14]—might result from combining the approach of the present paper with that of Ref. [9].

The utility of targeted free energy perturbation depends critically on our ability to construct a mapping $\mathcal{M}$ appropriate to the problem at hand. While intuition will in some cases reliably suggest a candidate, in others it may be very difficult or computationally expensive to devise a mapping that improves the overlap with ensemble B. In the case of *nondiffusive* systems, however, a promising and quite general strategy exists [15]. For a quasirigid system, such as a large molecule, the canonical ensemble occupies a strongly localized region of configuration space (assuming that the translational and rotational degrees of freedom of the entire molecule have been integrated out, or else pinned down by a constraining potential). Given two such molecules A and B—alchemically different, hence represented by different energy functions [16]—we can roughly approximate the associated canonical ensembles by Gaussian distributions in the many-dimensional configuration space [17]. A reasonable candidate for $\mathcal{M}$ is then the linear transformation that converts one of these Gaussians into the other. Even if the Gaussian approximation is quite crude, the mapping thus constructed is likely to result in a significantly improved overlap between A' and B (relative to that between A and B).

We conclude this paper with numerical results illustrating the targeted free energy perturbation method. The setting is the expansion of a cavity in a fluid. While the aim here is simply a comparison between methods, it bears mention that recent years have seen renewed theoretical interest in the problem of cavity formation in fluids [18], both as a fundamental problem in physical chemistry, and because of the role played by hydrophobicity in determining and stabilizing protein structure.

Consider n_p point molecules confined within a cubic container of volume L^3, but excluded from a spherical cavity of radius R located at the center ($\mathbf{r}=\mathbf{0}$) of the container. Assume periodic boundary conditions and a pairwise interaction $V_{\mathrm{int}}(\mathbf{r}_i,\mathbf{r}_j)$ between molecules. We can write the energy function for such a fluid as

$$E(\mathbf{x};R)=\Theta(\mathbf{x};R)+\sum_{i<j} V_{\mathrm{int}}(\mathbf{r}_i,\mathbf{r}_j). \tag{19}$$

Here, $\mathbf{x}=(\mathbf{r}_1,\ldots,\mathbf{r}_{n_p})$ specifies the microstate, and $\Theta(\mathbf{x};R)$ is either 0 (if all $r_i>R$) or $+\infty$ (otherwise), thus enforcing the exclusion of molecules from the spherical cavity. Treating the cavity radius R as an external parameter, let us choose two values, R_A and R_B, satisfying $0<R_A<R_B<L/2$, and let A and B denote the corresponding equilibrium states (canonical ensembles), at a given temperature T. We want to compute the associated free energy difference $\Delta F=F_B-F_A$. Physically, this is the reversible, isothermal work required to expand the cavity radius from R_A to R_B.

The quantity $\exp(-\Delta F/kT)$ is equal to the probability P that, given a microstate $\mathbf{x}$ sampled from ensemble A, the region $R_A<r\leq R_B$ will be devoid of molecules. This can be viewed as a consequence of Eq. (1), noting that ΔE is equal to either 0 or $+\infty$, depending on whether or not this region is vacant. The application of the traditional perturbation method amounts to evaluating this probability by straight sampling.

Let us now try to construct a transformation $\mathcal{M}$ that improves the efficiency of estimating $P=\exp(-\Delta F/kT)$. With the traditional method, poor convergence arises if $P\ll 1$, i.e., if for nearly every microstate $\mathbf{x}\in A$, there will be molecules located in the region $R_A<r\leq R_B$. Therefore, let us choose $\mathcal{M}$ so as to vacate this region. A candidate transformation [19], acting on each particle independently, is

$$\mathbf{r}_i\rightarrow g(r_i)\mathbf{r}_i, \quad i=1,\ldots,n_p, \tag{20}$$

where

$$g(r)=\left[1+\frac{(R_B^3-R_A^3)(L^3-8r^3)}{(L^3-8R_A^3)r^3}\right]^{1/3} \tag{21}$$

if $R_A<r\leq L/2$, and $g(r)=1$ otherwise. Under this transformation, the region of space defined by $R_A<r\leq L/2$ gets uniformly compressed into the region $R_B<r\leq L/2$. For any $\mathbf{x}$ sampled from A, the quantity $E_B(\mathbf{y})-E_A(\mathbf{x})$ is just the change in the total interaction energy (ΣV_{int}) resulting from this compression, and

$$J(\mathbf{x})=[(L^3-8R_B^3)/(L^3-8R_A^3)]^{\nu(\mathbf{x})}, \tag{22}$$

where $\nu(\mathbf{x})$ is the number of molecules in the region $R_A<r\leq L/2$.

We simulated 125 molecules inside a container of sides $L=22.28$ Å, at $T=300$ K. A Lennard-Jones interaction between molecules was used, with parameters corresponding to argon [20] ($\sigma=3.542$ Å, $\epsilon=0.1854$ kcal/mol). The values of R_A and R_B were taken to be 9.209 Å and 9.386 Å, respectively.

Sampling from ensemble A was achieved with the Metropolis algorithm. Three independent runs were carried out,

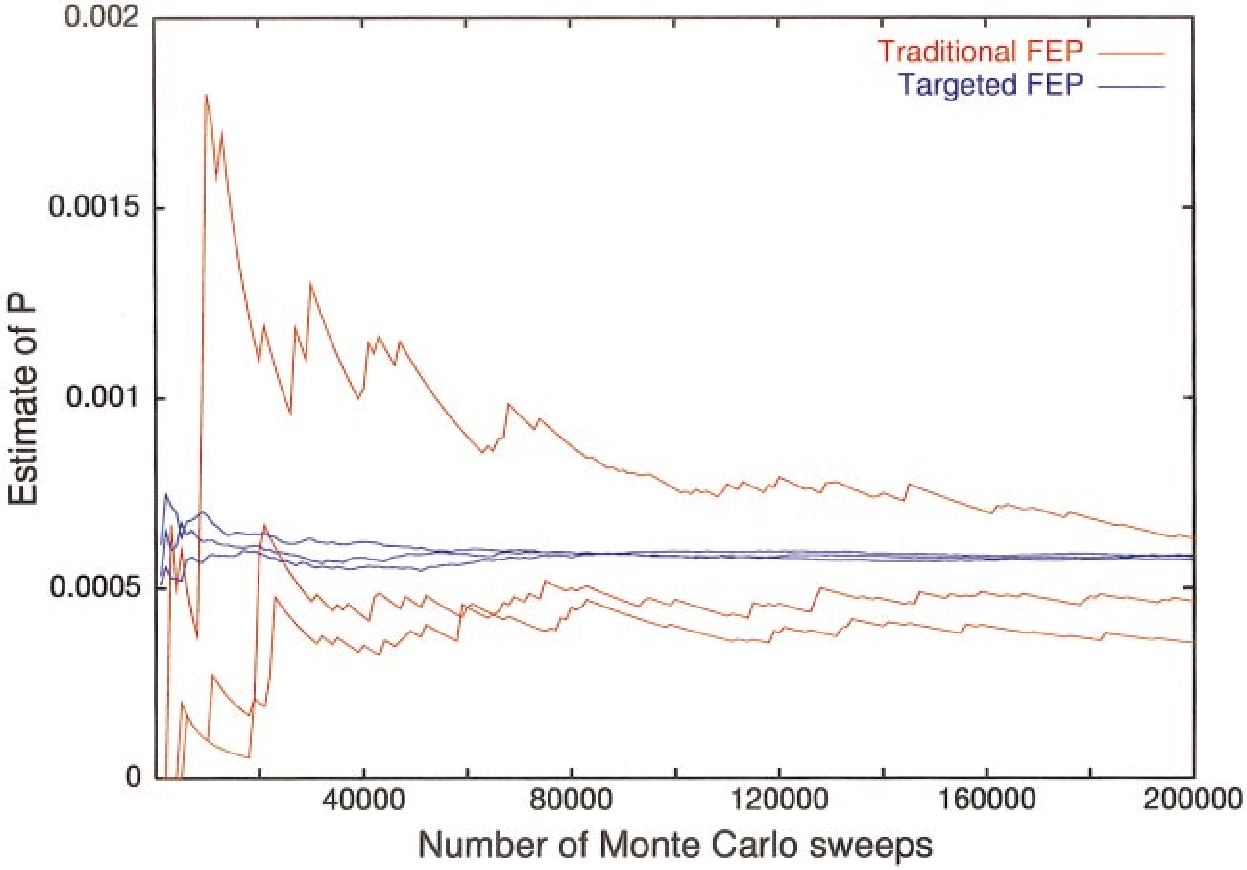

FIG. 1. (Color) Traditional and targeted free energy perturbation estimates of $P=\exp(-\Delta F/kT)$, as a function of the number of MC sweeps.

each consisting of 500 initial relaxation sweeps followed by 2×10^5 production sweeps. These runs were used to estimate $P=\exp(-\Delta F/kT)$, using both the traditional perturbation approach (i.e., observing the frequency with which the region $R_A<r\leq R_B$ is spontaneously vacant) and targeted perturbation [Eq. (12)]. In Fig. 1, each red curve shows the traditional perturbation estimate of P for a single run, accumulating as a function of number of production sweeps, N (plotted in increments of $\Delta N=1000$). The blue curves show the targeted perturbation estimates for the same runs. It is evident that the latter converge much faster than the former. Combining the data from all three runs, the two methods yield the estimates $P_{\mathrm{trad}}^{\mathrm{est}}=(4.83\pm0.49)\times10^{-4}$ and $P_{\mathrm{targ}}^{\mathrm{est}}=(5.81\pm0.05)\times10^{-4}$. The error bars are 1σ, and their ratio suggests that efficiency in this case is improved by about two orders of magnitude by using targeted (rather than traditional) free energy perturbation.

Note added. Since the original submission of this paper, it has come to my attention that an equivalent method has been developed for the estimation of ratios of normalizing constants (e.g., likelihood ratios) of probability models [21].

It is a pleasure to acknowledge stimulating discussions and correspondence with Graeme Ackland, Alastair Bruce, Angel Garcia, Gabriel Istrate, Lawrence Pratt, and Nigel Wilding. This research is supported by the Department of Energy, under Contract No. W-7405-ENG-36, and by the Polish-American Maria Skłodowska-Curie Joint Fund II, under project PAA/DOE-98-343.

[1] See, e.g., D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, New York, 1987); D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, San Diego, 1996).

[2] *Free Energy Calculations in Rational Drug Design*, edited by M.R. Reddy and M.D. Erion (Kluwer, Dordrecht, 2001).

[3] G.J. Ackland (unpublished).

[4] T. Schafer and E.V. Shuryak, Rev. Mod. Phys. **70**, 323 (1998).

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

[6] Ensembles A and B refer to the two canonical distributions representing the thermodynamic states A and B. By contrast, ensemble A', defined by Eq. (13), does not generally correspond to a physically interesting canonical distribution.

[7] A.F. Voter, J. Chem. Phys. **82**, 1890 (1985).

[8] A.D. Bruce, N.B. Wilding, and G.J. Ackland, Phys. Rev. Lett. **79**, 3002 (1997); A.D. Bruce, A.N. Jackson, G.J. Ackland, and N.B. Wilding, Phys. Rev. E **61**, 906 (2000).

[9] M.A. Miller and W.P. Reinhardt, J. Chem. Phys. **113**, 7035 (2000).

[10] To derive Eq. (18), take $\mathcal{M}$: $\mathbf{x} \rightarrow \mathbf{x} + \mathbf{u} d\lambda$, where $d\lambda$ is the difference between the parameter values defining states A and B, and expand Eq. (10) to first order in $d\lambda$.

[11] J.G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).

[12] G.M. Torrie and J.P. Valleau, Chem. Phys. Lett. **28**, 578 (1974).

[13] C.H. Bennett, J. Comput. Phys. **22**, 245 (1976).

[14] J.R. Gullingsrud, R. Braun, and K. Schulten, J. Comput. Phys. **151**, 190 (1999).

[15] This scheme has been suggested independently in the somewhat different context of lattice-switch Monte Carlo; A. Acharya, A.D. Bruce, and G.J. Ackland (unpublished); and also see Ref. [3].

[16] The problem of estimating ΔF between alchemically different molecules or molecular complexes is a central problem of rational drug design; see Ref. [2].

[17] Extensive research along these lines has been carried out for protein molecules; see, e.g., the review by A. Kitao and N. Go, Curr. Opin. Struct. Biol. **9**, 164 (1999).

[18] See, for instance, K. Lum, D. Chandler, and J.D. Weeks, J. Phys. Chem. B **103**, 4570 (1999); G. Hummer, S. Garde, A.E. Garcia, and L.R. Pratt, Chem. Phys. **258**, 349 (2000), and related references.

[19] For complicated cavity geometries, analogous mappings could be constructed numerically at relatively low cost.

[20] R. Reid, J. Prausnitz, and T. Sherwood, *The Properties of Gases and Liquids*, 3rd ed. (McGraw-Hill, New York, 1977), Appendixes A and C.

[21] X.-L. Meng and S. Schilling, J. Computat. Graph Statis. (to be published).