Accelerated convergence of molecular free energy via superposition approximation-based reference states

Sandeep Somani¹ and Michael K. Gilson^{2,a)}

¹*Institute for Bioscience and Biotechnology Research, 9600 Gudelsky Drive, Rockville, Maryland 20850, USA*

²*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093-0736, USA*

(Received 16 November 2010; accepted 9 March 2011; published online 6 April 2011)

The free energy of a molecular system can, at least in principle, be computed by thermodynamic perturbation from a reference system whose free energy is known. The convergence of such a calculation depends critically on the conformational overlap between the reference and the physical systems. One approach to defining a suitable reference system is to construct it from the one-dimensional marginal probability distribution functions (PDFs) of internal coordinates observed in a molecular simulation. However, the conformational overlap of this reference system tends to decline steeply with increasing dimensionality, due to the neglect of correlations among the coordinates. Here, we test a reference system that can account for pairwise correlations among the internal coordinates, as captured by their two-dimensional marginal PDFs derived from a molecular simulation. Incorporating pairwise correlations in the reference system is found to dramatically improve the convergence of the free energy estimates relative to the first-order reference system, due to increased conformational overlap with the physical distribution. © 2011 American Institute of Physics. [doi:10.1063/1.3571441]

I. INTRODUCTION

The calculation of the free energy (or, more formally, the chemical potential) of a molecule in solution is a central problem in computational chemistry and biophysics, with applications to conformational stability,¹ solvation,² and molecular recognition.^{3,4} Systems of interest in this regard include relatively small (<100 atoms) druglike molecules, moderately sized supramolecular systems (100–1000 atoms), and proteins with thousands of atoms. The challenge of calculating molecular free energies stems chiefly from the complexity and high dimensionality of the energy surfaces involved.⁵ In one approach, the problem is recast as a calculation of the free energy difference between the system of interest and a reference system whose free energy is known. The effectiveness of this reference system approach increases as the conformational probability distribution of the reference system more closely approximates that of the physical system of interest. Contrariwise, if the reference distribution has little overlap with the physical system, then it will be very difficult to obtain a reliable, converged result. As a consequence, the choice of reference system is of critical importance.

In some calculations of this type, the free energy of the reference system can be computed analytically, as in the harmonic or “Einstein solid”^{6,7} reference system for solids or the ideal gas reference system for liquids.^{8,9} A reference system approach to computing the absolute free energy of a molecule was apparently first employed by Stoessel and Nowak,¹⁰ whose reference system was a collection of independent harmonic oscillators centered at atomic coordinates.

More recently, simulation data have begun to be used to construct numerical, rather than analytic, reference systems¹¹ for free energy calculations. Such approaches can be advantageous, because the resulting reference systems can better capture the flexibility and inhomogeneity of biomolecules, thereby increasing conformational overlap with the physical system. They also avoid the need for *ad hoc* tuning of adjustable parameters, such as spring constants. The fidelity of such data-based references can be maximized by using enhanced sampling methods¹² and high performance computing technologies.^{13,14}

The present work is inspired by the reference system method of Zuckerman and Ytreberg,¹¹ which uses molecular dynamics (MD) simulation data to construct one-dimensional (1D) probability density functions (PDFs) of internal bond, angle, and torsion (BAT) coordinates. Their reference state, whose conformational PDF is simply the product of these 1D PDFs, is relatively simple to construct, yet it still captures a great deal of information about the conformational distribution of the physical system. However, it does not capture correlations among the internal coordinates, and this neglect of correlations in the reference distribution reduces its overlap with the physical distribution. In particular, as the system size (dimensionality) increases, there is a rapid rise in the fraction of conformations sampled from the reference system that have high energies in the physical system, mainly due to steric clashes among the atoms. This can lead to poor convergence of the free energy estimate, restricting the applicability of the method to small molecules with weak correlations, such as linear chains with weak nonbonded interactions. Zuckerman and co-workers have recently addressed this issue by an innovative fragment-based sampling method.¹⁵ Here, we describe a different, potentially complementary, approach.

^{a)} Author to whom correspondence should be addressed. Electronic mail: mgilson@ucsd.edu. Tel.: 858-822-0622. Fax: 858-822-7726.

We have previously introduced a method for constructing approximations to the high-dimensional Boltzmann distribution of a molecule, which accounts for specified correlations among the internal coordinates.¹⁶ These approximations build upon the generalized Kirkwood superposition approximation (SA), which is a probability closure equation that approximates an N -dimensional PDF in terms of PDFs of order $N-1$ and below. Further generalization allows approximations in terms of PDFs of order up to an arbitrarily chosen level l , where l is typically much less than N . We also presented an algorithm, based on these l -level superposition approximations (SA- l), to sample molecular conformations from a PDF which approximates the full physical PDF (the Boltzmann distribution) using only the low-order PDFs. Importantly for the present application, the sampled distribution is normalized by construction, allowing us to set up a reference system with known free energy. As a consequence, the free energy of the physical system can be computed from the known free energy of the reference system plus the free energy difference between the physical and reference systems. This free energy difference is computable here, as done previously,¹¹ via thermodynamic perturbation with samples drawn from the reference distribution. We find that using a reference that incorporates pairwise correlations among the internal coordinates can dramatically increase the overlap with the physical distribution and thus reduce the variance and speed the convergence of the free energy estimates.

Section II presents the theory of SA- l based conformational sampling and the reference system method. The subsequent section describes the implementation of free energy calculations for molecular systems. The method is validated with tests on propane with a simplified energy function that allows analytic calculation of the free energy. It is then applied to calculate the free energy of polypeptides with one, three, and five alanine residues, which were recently used as test cases for the fragment-based free energy calculation method mentioned above.¹⁵ The Discussion section then assesses the significance of the results and addresses possible future directions.

II. THEORY

The configuration integral, Z_{phys} , of an M -atom molecule with $N = 3M-6$ generalized internal coordinates $\vec{X} = (X_1, \dots, X_N)$ and physical potential energy function $U_{\text{phys}}(\vec{X})$ is

$$Z_{\text{phys}} = \int \exp(-\beta U_{\text{phys}}(\vec{X})) J(\vec{X}) d\vec{X}, \quad (1)$$

where $\beta = 1/kT$, k being the Boltzmann constant and T the absolute temperature; $d\vec{X} = dX_1 \times \dots \times dX_N$ is the volume element of the configurational space; and $J(\vec{X})$ is the Jacobian factor of transforming the generalized internal coordinates to Cartesian coordinates.^{17,18} The integral is over the region corresponding to the conformation state of interest, e.g., the helical state of a peptide in solution, or the bound state of a protein-small molecule system. The absolute free energy, F_{phys} , of the molecule is given by

$$\beta F_{\text{phys}} = -\ln Z_{\text{phys}}. \quad (2)$$

Note that Z_{phys} is treated as a dimensionless number and the units of F_{phys} are set by kT .

Because N is large for typical molecules of interest, direct calculation of the configurational integral, and therefore the free energy, is computationally infeasible. However, the free energy difference with respect to a reference system (denoted by “ref”) can be computed, if there is sufficient conformational overlap between the two systems. If the free energy of the reference system, F_{ref} , is zero, as for the reference systems to be used here, then the free energy difference

$$\Delta F = F_{\text{phys}} - F_{\text{ref}}, \quad (3)$$

gives the desired physical free energy, F_{phys} . In the present work, a molecular force-field is used as the physical potential energy U_{phys} , and the reference potential energy, U_{ref} , is defined in terms of SA- l based sampling distributions.¹⁶ We use two reference systems, the $l = 1$, or singlet reference system, which ignores all correlations; and the $l = 2$, or doublet reference system, which accounts for all pairwise correlations among the internal coordinates.

A. Internal coordinate systems and low-order distributions

Based on our prior studies,¹⁶ the BAT internal coordinate system is used to set up the reference distributions. The conformation of an M -atom molecule in BAT coordinates is denoted by $\vec{X} = (\vec{b}, \vec{a}, \vec{t}) \in \mathbb{R}^N$, where $N = 3M-6$ and $\vec{b}$, $\vec{a}$, and $\vec{t}$ are vectors denoting $M-1$ bond lengths, $M-2$ bond angles, and $M-3$ torsions, respectively. Another internal coordinate system, which enables calculation of force-field energy using standard software, is the anchored Cartesian (XYZ) internal coordinate system.^{19,20} The two coordinate systems and the labeling scheme used here are illustrated in Fig. 1 using propane as an example for branched molecules considered in this study. The Jacobian of the transformation between XYZ and BAT coordinates, which is independent of the torsion angles, is given by^{18,20,21}

$$J(\vec{b}, \vec{a}, \vec{t}) = b_3 \prod_{i=4}^M (b_i^2 \sin a_i). \quad (4)$$

To set up the reference systems, the N -dimensional BAT conformational space is discretized by discretizing each coordinate into B equally spaced bins. The discretization is based on MD simulation data, with the bin width for coordinate X_i set to $\delta_i = \Delta_i/B$, where Δ_i is the range of values observed in the simulation. A microstate or conformation in the discrete space is denoted by $\vec{\bar{X}} = (\bar{X}_1, \dots, \bar{X}_N) \in \mathbb{Z}^N$ with $\bar{X}_i \in \{1, \dots, B\}$ denoting a bin number. (The bar above the symbol denotes a discrete-space variable.) A discrete-space conformation $\vec{\bar{X}}$ maps to the continuous space conformation $\vec{X}(\vec{\bar{X}})$, where the mapping for each individual coordinate is given by

$$X_i(\bar{X}_i) = X_{i,\min} + \left(\bar{X}_i - \frac{1}{2} \right) \delta_i. \quad (5)$$

$X_{i,\min}$ being the minimum value observed for the i th coordinate in the simulation data. As previously detailed,¹⁶ we

FIG. 1. Internal coordinate systems illustrated using propane. The anchored Cartesian system is defined in terms of three root atoms such that atom 1 is at the origin, atom 2 is on the x axis, and atom 3 is in the x - y plane (shaded in blue). For the peptides in Fig. 2, a C-terminal hydrogen is labeled as atom 1 and two subsequent carbon atoms in the chain are labeled, in order, as atoms 2 and 3. The bond, angle, and torsion coordinates in the BAT system are labeled. In all, propane has $11 \times 3 - 6 = 27$ internal coordinates specified by ten bond lengths (top left), nine bond angles (bottom left), and eight torsion angles (bottom right). Torsion angles for atoms 10 and 11 are improper torsions (top right), and the others are proper torsions. Shaded circles: carbon; unshaded circles: hydrogen.

compute normalized histograms of internal coordinates, based upon conformations sampled from the physical distribution via MD simulation. Individual binning of each coordinate gives the N singlet (or 1D) PDFs, $p_1(\bar{X}_{i_1})$, and joint binning of all pairs of coordinates gives the $C_2^N = N(N-1)/2$ doublet (or 2D) PDFs, $p_2(\bar{X}_{i_1}, \bar{X}_{i_2})$, where $i_1, i_2 \in \{1, \dots, N\}$. These low-order PDFs, which represent marginal PDFs of the physical distribution function, are used to set up the singlet and doublet reference systems studied here. Since the histograms are normalized, we have

$$\sum_{\bar{X}_{i_1}} p_1(\bar{X}_{i_1}) = 1, \quad \sum_{\bar{X}_{i_1}, \bar{X}_{i_2}} p_2(\bar{X}_{i_1}, \bar{X}_{i_2}) = 1. \quad (6)$$

In general, the joint distribution of a subset of k ($< N$) variables will be denoted by $p_k(\bar{X}_{i_1}, \dots, \bar{X}_{i_k})$, where $i_1, \dots, i_k \in \{1, \dots, N\}$ and subscript “ k ” denotes the dimensionality of the PDF.

B. Superposition approximation at level l

The superposition approximation at level l (SA- l) approximates the k -dimensional marginal of the physical PDF in terms of PDFs at the highest order l , where $N > k > l$. The lowest order superposition approximation in this sense, SA-1, assumes that all variables are independent. This singlet level approximation, $p_k^{(1)}$, is given by the product of the 1D marginal

PDFs:

$$p_k^{(1)}(\bar{X}_{i_1}, \dots, \bar{X}_{i_k}) \equiv p_1(\bar{X}_{i_1}) \times \dots \times p_1(\bar{X}_{i_k}), \quad (7)$$

where $i_1, \dots, i_k \in \{1, \dots, N\}$ denotes a subset of k variables. Higher level superposition approximations (SA- l , with $l > 1$) are obtained by recursive application of the generalized Kirkwood SA.^{22–26} In particular, the doublet level approximation, $p_k^{(2)}$, in addition to the 1D PDFs, incorporates all pairwise correlations among the k variables:

$$p_k^{(2)}(\bar{X}_{i_1}, \dots, \bar{X}_{i_k}) \equiv \left(\prod_{1 \leq \alpha \leq k} p_1(\bar{X}_{i_\alpha}) \right)^{-(k-2)} \left(\prod_{1 \leq \alpha < \beta \leq k} p_2(\bar{X}_{i_\alpha}, \bar{X}_{i_\beta}) \right)^{+1}. \quad (8)$$

In this work, only the singlet ($l = 1$) and doublet ($l = 2$) level superposition approximations are used.

The SA- l based sampling algorithm¹⁶ described in Sec. II C requires the conditional PDF of a variable $\bar{X}_{i_k}$ ($i_k \leq N$) given the values of the other $k-1$ variables $\bar{X}_{i_1}, \dots, \bar{X}_{i_{k-1}}$. The product rule of probability theory²⁷ gives the conditional PDF as

$$p_1(\bar{X}_{i_k} | \bar{X}_{i_1}, \dots, \bar{X}_{i_{k-1}}) = \frac{p_k(\bar{X}_{i_1}, \dots, \bar{X}_{i_{k-1}}, \bar{X}_{i_k})}{p_{k-1}(\bar{X}_{i_1}, \dots, \bar{X}_{i_{k-1}})}. \quad (9)$$

At the singlet level, since all variables are assumed to be independent, the conditional PDF of any variable is given

simply by its 1D PDF:

$$p_1^{(1)}(\bar{X}_{i_k} | \bar{X}_{i_1}, \dots, \bar{X}_{i_{k-1}}) = p_1(\bar{X}_{i_k}). \quad (10)$$

At the doublet level, we approximate the conditional PDF by substituting $p_k^{(2)}$ from Eq. (8) into Eq. (9), yielding

$$p_1^{(2)}(\bar{X}_{i_k} | \bar{X}_{i_1}, \dots, \bar{X}_{i_{k-1}}) = \frac{\prod_{1 \leq \alpha \leq k-1} p_2(\bar{X}_{i_\alpha}, \bar{X}_{i_k})}{(p_1(\bar{X}_{i_k}))^{k-2}} \frac{1}{v_k(\bar{X}_{i_1}, \dots, \bar{X}_{i_{k-1}})}, \quad (11)$$

where the normalization factor, v_k , is given by

$$v_k(\bar{X}_1, \dots, \bar{X}_{k-1}) = \sum_{\bar{X}_{i_k}} \frac{\prod_{1 \leq \alpha \leq k-1} p_2(\bar{X}_{i_\alpha}, \bar{X}_{i_k})}{(p_1(\bar{X}_{i_k}))^{k-2}}. \quad (12)$$

Equation (11) is equivalent to Eq. (20) of Ref. 16, after the extra factors are absorbed into the normalization factor. Note that, if there are no correlations among the variables, so that $p_2(\bar{X}_{i_\alpha}, \bar{X}_{i_\beta}) = p_1(\bar{X}_{i_\alpha})p_1(\bar{X}_{i_\beta})$, the doublet approximation in Eq. (12) reduces to Eq. (10).

C. SA-*l* based conformational sampling

The reference distributions constructed from the singlet and doublet PDFs are sampled as follows. At the singlet level, which uses the SA-1 approximation, correlations among variables are ignored, so each variable is sampled independently from its 1D physical PDF, and the probability of sampling a conformation $\bar{X} = (\bar{X}_1, \dots, \bar{X}_N)$ is given by the SA-1 probability:

$$\bar{p}_N^{(1)}(\bar{X}) = p_1(\bar{X}_1) \times \dots \times p_1(\bar{X}_N) = p_N^{(1)}(\bar{X}). \quad (13)$$

Since the 1D PDFs are normalized, the sampling probability $\bar{p}_N^{(1)}$ is also normalized. This singlet level approximation was used as reference by Zuckerman and co-workers.^{11,15} At the doublet level, which uses the SA-2 approximation, sampling is more complicated. Here, a conformation is generated via the ancestral sampling framework,²⁸ as follows:¹⁶

$$\bar{x}_1 \sim p_1(\bar{X}_1)$$

$$\bar{x}_2 \sim \frac{p_2(\bar{x}_1, \bar{X}_2)}{p_1(\bar{x}_1)}$$

$$\bar{x}_3 \sim \frac{p_2(\bar{x}_1, \bar{X}_3)p_2(\bar{x}_2, \bar{X}_3)}{p_1(\bar{X}_3)} \frac{1}{v_3(\bar{x}_1, \bar{x}_2)}$$

FOR $k = 4$ to N

$$\bar{x}_k \sim p_1^{(2)}(\bar{X}_k | \bar{x}_1, \dots, \bar{x}_{k-1}) \text{ given by Eq. (11)}$$

ENDFOR

where “ $\sim$ ” means “sampled from” and the lower case denotes variables that have already been sampled. The normalization factors of the conditional PDFs at each step are obtained numerically by summing up the conditional probability for all possible values of the variable to be sampled. In this algorithm, each variable defining a single molecular conformation is sampled from a one-dimensional distribution conditioned upon the values of all variables that have

already been sampled. The algorithm is iterated until the desired number of samples has been generated. Note that the distribution of conformations [Eq. (14)] from the doublet level sampling depends upon the order in which the variables are sampled. Based on prior work,¹⁶ we sample the torsion coordinates first, followed by the bond angles and, finally, the bond lengths. Each conformation generated by the above algorithm is associated with a known probability density in its corresponding multidimensional bin, given by

$$\bar{p}_N^{(2)}(\bar{X}) = p_2(\bar{X}_1, \bar{X}_2) \times p_1^{(2)}(\bar{X}_3 | \bar{X}_1, \bar{X}_2) \times \dots \times p_1^{(2)}(\bar{X}_N | \bar{X}_1, \dots, \bar{X}_{N-1}) \neq p_N^{(2)}(\bar{X}). \quad (14)$$

The inequality in the second line is included to clarify that our reference system sampling probability is not equal to the N -dimensional superposition approximation at level 2 (SA-2) but is instead an approximation to it, which is easier to sample from. The normalization of the conditional PDFs at each step of the sampling leads to the normalization of the final sampling probability $\bar{p}_N^{(2)}$, even though the SA-2 [Eq. (8)] distribution itself is not normalized (see Appendix of Ref. 16).

This sequential sampling algorithm is analogous to chain building approaches in polymer physics^{29,30} with the conditional PDF here similar to the transition probability used in extending the chain at each stage. Note that, if the low-order PDFs used for doublet sampling have cells with zero probability, then the conditional PDF of a variable during a sampling iteration can become identically zero, resulting in failure to sample the variable and hence an incomplete conformation. Here, such samples are discarded, and a new conformation is begun from the beginning. (The conditions that can result in such incomplete samples are further described in Ref. 16). Based on our tests for multiple molecular systems, only a small fraction (<1%) of doublet level sampling iterations result in incomplete samples.

Each coordinate X_i of the conformations sampled by the present approach necessarily lies in the range Δ_i established by the original simulation data. This restricts the conformational space accessible to the reference distribution to a subset of the full conformational space. Further restrictions are imposed by the PDFs used to construct the reference distribution. Since any conformation with nonzero sampling probability $\bar{p}_N^{(l)}$ necessarily has a nonzero SA-*l* probability, the product form of $p_N^{(l)}$ ensures that all coordinates of a sampled conformation lie in the nonzero regions of *all* the marginal PDFs used to construct the reference system PDF. Therefore, in general, as higher order correlations are incorporated, the conformational region accessible to sampling shrinks and approaches the region sampled by the original simulation data. Thus, it is important that the original simulation data sample the relevant regions of the conformational space. On the other hand, since higher order PDFs are ignored in both singlet and doublet level sampling, the above sampling can generate conformations not observed in the original simulation, and the region of the conformational space accessible to the SA-*l* based sampling is expected to be greater in this sense than that covered by the original MD simulation. Another important property of the present sampling algorithm is that, unlike successive steps in Gibbs sampling or in a molecular

dynamics or Monte Carlo (MC) simulation, here successively sampled conformations are uncorrelated with each other. This should facilitate thorough sampling of the accessible regions of the conformational space and also makes the sampling algorithm trivial to parallelize.

D. Reference system and physical free energy estimation

The partition function of the physical system, which is given by the N -dimensional integral in Eq. (1), can be approximated as a sum over the discrete states on which the reference distribution is defined:

$$Z_{\text{phys}} \approx \bar{Z}_{\text{phys}} \equiv \Delta V \sum_{\vec{X} \in \mathbf{S}} \exp(-\beta U_{\text{phys}}(\vec{X}(\vec{X}))) J(\vec{X}(\vec{X})), \quad (15)$$

where $\mathbf{S}$ denotes the set of all possible microstates in the discrete space, $\Delta V \equiv \delta_1 \dots \delta_N$ is the volume of a cell in the discretized BAT space, and the Jacobian is given by Eq. (4). Defining the effective physical energy in the discretized space as

$$\bar{U}_{\text{phys}}(\vec{X}) \equiv U_{\text{phys}}(\vec{X}(\vec{X})) - \frac{1}{\beta} \ln J(\vec{X}(\vec{X})) - \frac{1}{\beta} \ln \Delta V, \quad (16)$$

gives

$$Z_{\text{phys}} \approx \bar{Z}_{\text{phys}} = \sum_{\vec{X} \in \mathbf{S}} \exp(-\beta \bar{U}_{\text{phys}}(\vec{X})). \quad (17)$$

The approximation in Eq. (17) gives the continuous integral of Eq. (1) in the limit of infinitely fine discretization, $\delta_i \rightarrow 0$ and if the ranges of the coordinates, Δ_i , cover the desired region of the conformational space. The effective energy function for the reference system is based on the SA- l based sampling distributions [Eqs. (13) and (14)]:

$$\bar{U}_{\text{ref}}^{(l)}(\vec{X}) \equiv \begin{cases} -\frac{1}{\beta} \ln \bar{p}_N^{(l)}(\vec{X}), & \text{if } \bar{p}_N^{(l)}(\vec{X}) \neq 0 \\ \infty, & \text{if } \bar{p}_N^{(l)}(\vec{X}) = 0. \end{cases} \quad (18)$$

Thus, conformations drawn from the sampling distributions $\bar{p}_N^{(l)}$ are effectively sampled from a canonical distribution corresponding to the reference energy function. Since the reference energy function is defined in the discrete space, the partition function for the reference system is given by the discrete sum:

$$\bar{Z}_{\text{ref}}^{(l)} = \sum_{\vec{X} \in \mathbf{S}} \exp(-\beta \bar{U}_{\text{ref}}^{(l)}(\vec{X})) = \sum_{\vec{X} \in \mathbf{S}} \bar{p}_N^{(l)}(\vec{X}) = 1. \quad (19)$$

The last equality derives from the normalization of the sampling distribution. The free energy of the reference systems $F_{\text{ref}} = -k_B T \ln \bar{Z}_{\text{ref}}^{(l)}$ is therefore identically zero for both singlet and doublet references. The discrete sum approximation of the physical free energy can therefore be written as

$$\beta \bar{F}_{\text{phys}} = -\ln \frac{\bar{Z}_{\text{phys}}}{\bar{Z}_{\text{ref}}^{(l)}}. \quad (20)$$

Defining the energy difference for state $\vec{X}$ as

$$\Delta \bar{U}^{(l)}(\vec{X}) \equiv \bar{U}_{\text{phys}}(\vec{X}) - \bar{U}_{\text{ref}}^{(l)}(\vec{X}), \quad (21)$$

we can write the ratio of the partition functions in Eq. (20) in the form of a thermodynamic perturbation³¹

$$\beta \bar{F}_{\text{phys}}^{(l)} = -\ln \langle \exp(-\beta \Delta \bar{U}^{(l)}) \rangle_{\text{ref}^{(l)}}. \quad (22)$$

This estimate is computed using samples drawn from the singlet or doublet reference canonical distributions. Thus,

$$\begin{aligned} \beta F_{\text{phys}} &\approx \beta \bar{F}_{\text{phys}} \approx \beta \bar{F}_{\text{phys}}^{(l)} \\ &\doteq -\ln \left(\frac{1}{N_{\text{ref}}} \sum_{n=1}^{N_{\text{ref}}} \exp(-\beta \Delta \bar{U}^{(l)}(\vec{X}_n)) \right), \end{aligned} \quad (23)$$

where the first approximation is associated with the discretization of the coordinate space, the second approximation depends on the reference system used, and the final expression is the numerical implementation of the perturbation estimate using N_{ref} samples. The superscripts again denote use of the singlet ($l=1$) or doublet ($l=2$) reference system, respectively. Note that any conformation with zero probability in the reference distribution, and hence infinite reference energy, is never sampled from the reference distributions.

E. Bias and convergence of the thermodynamic perturbation estimate

Bias and convergence of the free energy estimate [Eq. (23)] can be understood by analyzing the asymptotic or infinite sampling limit, $N_{\text{ref}} \rightarrow \infty$. Consider

$$\begin{aligned} \lim_{N_{\text{ref}} \rightarrow \infty} \exp(-\beta \bar{F}_{\text{phys}}^{(l)}) &= \lim_{N_{\text{ref}} \rightarrow \infty} \frac{1}{N_{\text{ref}}} \sum_{n=1}^{N_{\text{ref}}} \exp(-\beta \Delta \bar{U}^{(l)}(\vec{X}_n)) \\ &= \sum_{\vec{X}_i \in \mathbf{S}} \bar{p}_N^{(l)}(\vec{X}_i) \exp(-\beta \Delta \bar{U}^{(l)}(\vec{X}_i)), \end{aligned} \quad (24)$$

where the second equality uses the fact that, in the limit of infinite sampling, the samples are distributed according to $\bar{p}_N^{(l)}$. Note that if there are zero probability cells or “holes” in the PDFs used to set up the reference distributions, some conformations in $\mathbf{S}$ will have $\bar{p}_N^{(l)} = 0$ and, therefore, can be dropped from the summation. Substituting $\Delta \bar{U}^{(l)}$ from Eq. (21) and using Eq. (18) gives

$$\begin{aligned} \lim_{N_{\text{ref}} \rightarrow \infty} \exp(-\beta \bar{F}_{\text{phys}}^{(l)}) &= \sum_{\vec{X}_i \in \mathbf{S}; \bar{p}_N^{(l)} \neq 0} \bar{p}_N^{(l)}(\vec{X}_i) \exp(+\beta \bar{U}_{\text{ref}}^{(l)}(\vec{X}_i)) \times \exp(-\beta \bar{U}_{\text{phys}}(\vec{X}_i)) \\ &= \sum_{\vec{X}_i \in \mathbf{S}; \bar{p}_N^{(l)} \neq 0} \exp(-\beta \bar{U}_{\text{phys}}(\vec{X}_i)). \end{aligned} \quad (25)$$

Comparing Eq. (25) with Eq. (17), since the asymptotic limit of the perturbation estimate does not include contributions from any microstates which have zero probability in the reference distribution, the perturbation estimate in Eq. (23) may be asymptotically biased. Also, since this contribution is strictly positive, the asymptotic free energy becomes more positive as the set of accessible states shrinks.

Due to the exponential in Eq. (25), the free energy is dominated by conformations with low physical energy. Therefore, if the reference system is able to access these dominant conformations, the asymptotic bias will be low. Note that, due to their larger Boltzmann factor, the dominant conformations are more likely to be sampled from the physical distribution. Therefore, if the conformational overlap between samples from the reference and physical distributions is high, then the bias is likely to be low and convergence faster.

The above analysis has implications for the use of SA-*l* based references. Compared with the singlet reference, the doublet reference system is expected to have a smaller accessible conformational region because the 2D PDFs are harder to populate and, therefore, more likely to have holes. On the other hand, due to incorporation of pair correlations, the overlap of the doublet reference is likely to be greater than that of the singlet reference. Therefore, the doublet free energy estimate is expected to converge more rapidly than singlet reference, though the asymptotic bias may be somewhat higher.

III. METHODS

We use the same overall procedure to compute the free energy of each molecular system with the singlet and doublet reference systems:

1. Run a constant temperature MD simulation of the molecule of interest, saving N_{phys} snapshots.
2. For each saved snapshot of the MD trajectory, compute the corresponding BAT coordinates and use the observed coordinate ranges to set up the bin ranges, Δ_i , and widths, δ_i , for discretization.
3. Use the N_{phys} binned BAT coordinates to generate normalized histograms representing the first- and second-order PDFs of all coordinates and pairs of coordinates, and use these to set up the singlet and doublet SA-based reference distributions.
4. Sample N_{ref} molecular conformations and their associated sampling probabilities $\tilde{p}_N^{(i)}$ from the reference distribution, via the ancestral sampling approach described in Sec. III C, using Eq. (13) for singlet or Eq. (14) for doublet versions.
5. For each reference sample, compute the reference energy $\bar{U}_{\text{ref}}^{(i)}$ from Eq. (18), construct the corresponding Cartesian molecular coordinates of the molecule, compute the corresponding physical (force-field) energy U_{phys} , and evaluate the energy difference $\Delta\bar{U}^{(i)}$ from Eqs. (16) and (21).
6. Finally, compute the thermodynamic perturbation estimate of the free energy ($\bar{F}_{\text{phys}}^{(1)}$ and $\bar{F}_{\text{phys}}^{(2)}$) by applying Eq. (23) to the sets of energy differences corresponding to the two reference systems.

A. Molecular systems

We first validate the theory and implementation with tests on a simplified representation of all-atom propane (11 atoms), for which the analytic free energy can be computed. Starting with a standard force-field representation, we delete all nonbonded (Lennard-Jones and Coulombic) energy terms, as

FIG. 2. Chemical structures of the molecules.

well as all bonded terms that do not correspond to the BAT coordinates used to specify the conformation. These simplifications decouple all internal coordinates from each other, so that there are no correlations in the thermal fluctuations of the molecule. This allows the partition function to be computed analytically (see Appendix). However, this is still a high-dimensional system, with 27 internal coordinates, and thus a useful test case. We then test the methodologies for full force-field representations of three peptides previously studied with a closely related method by Zhang *et al.*¹⁵: alanine dipeptide (Ace-Ala-Nme, 22 atoms, 60 BAT coordinates), dialanine (Ace-(Ala)₂-Nme, 32 atoms, 90 BAT coordinates) and tetra-alanine (Ace-(Ala)₄-Nme, 52 atoms, 150 BAT coordinates) where Ace is acetyl (CH₃-CO) and Nme is *N*-methylamide (NH-CH₃) (Fig. 2). GAFF (Ref. 32) force-field parameters were used in all cases. Force-field parameter files were generated with Amber AnteChamber³³ and converted to GROMACS format using amb2gmx from ffAMBER tools.^{34,35} Note that the force-field and temperature used in this study are different from those in Ref. 15, so that a quantitative comparison is not possible.

For all molecules except simplified propane, 10 sets of 50 ns vacuum MD simulations were done at 1000 K using GROMACS 4.0.5 (Ref. 36) with a time step of 1 fs; for simplified propane, a single 100 ns run was carried out. Conformations were saved every picosecond to generate a total of 5×10^6 conformations for each peptide and 1×10^6 conformations for simplified propane. Free energies are reported in units of kT ($= 8.36$ kJ/mol at 1000 K).

Each BAT coordinate was discretized into $B = 30$ bins equally spaced between the minimum and maximum values found in the MD snapshots, and the coordinate snapshots were used to populate the 1D and 2D PDFs which were used to generate 10^6 reference samples for simplified propane and 5×10^6 reference samples for each peptide.

B. Assessment of free energy estimates

We monitor convergence of the two free energy estimates, $\bar{F}_{\text{phys}}^{(1)}$ and $\bar{F}_{\text{phys}}^{(2)}$, as a function of the number of

TABLE I. Mean and standard deviation (in parenthesis) of absolute free energy (in kT) of molecules from 100 bootstrap resampled datasets.

	Number of Atoms	Analytic	Singlet	Doublet
Simplified propane	11	74.6920	74.77 (0.0002)	74.77 (0.0004)
Alanine dipeptide	22	...	207.4 (0.8)	207.5 (0.005)
Dialanine	32	...	331.5 (1.3)	311.43 (0.01)
Tetra-alanine	52	...	631.7 (17.3)	519.3 (0.15)

reference samples N_{ref} . Error analysis is done using the bootstrap method^{37,38} in which the original set of samples from the reference system are resampled with replacement to generate 100 new sets of samples, and the perturbation formula [Eq. (23)] is applied to each dataset. The mean and standard deviation of these 100 estimates are reported as the final free energy estimate and its uncertainty, respectively. We furthermore compare the distributions of the physical (force-field) energies of the original MD sampled and reference sampled conformations as a measure of the conformational overlap.

IV. RESULTS

A. Validation with simplified propane

Both free energy estimates for simplified propane converge to $74.77 kT$, within $0.08 kT$ of the analytic free energy of $74.692 kT$, as shown in Table I, and appear to be well converged, as shown in Fig. 3 (note the small range of the vertical axis). Nonetheless, there is evidently a small bias in the estimates, which is likely due to the discretization and restriction of the conformational space, since the analytic partition function is computed from continuous integrals with full coordinate ranges instead of those observed in the simulation (see Appendix). The positive sign of the small bias is consistent with the analysis of asymptotic bias in Sec. II E. Figure 4 plots the physical (force-field) energy distributions

FIG. 3. Convergence of free energy estimates for simplified propane. The solid line indicates the mean of the estimate using 100 bootstrap samples (error bars are not shown because the standard deviation is smaller than the thickness of the line).

of conformations sampled from MD and the two references. The three energy distributions are virtually identical, indicating high similarity between the conformations sampled from the reference distributions and the physical Boltzmann distribution. Overall, the accuracy of the results for this simplified propane test, for which the free energy is available in a reliable analytic form, validates the theory and implementation of the sampling and free energy calculations.

B. Peptides with full force-field representations

Analytic free energy values are not available for the three peptides (Fig. 2), because the internal coordinates are coupled by force-field energy terms so the configuration integrals cannot be factorized. We, therefore, assess the reliability of the four free energy estimates obtained for each molecule (Table I) by examining convergence plots (Fig. 5) and the overlaps of the reference and physical energy (Fig. 6).

A central result of this study is that the doublet-level reference systems lead to dramatically faster free energy convergence than the singlet level reference systems, as seen in Figs. 5(a)–5(c) and the bootstrap standard deviations in Table I.

The excellent convergence of the doublet estimate is consistent with the strong overlap between the force-field energy distribution (Fig. 6) of doublet reference samples with the MD samples from the physical Boltzmann distribution. The singlet reference systems yield much worse overlap with the physical distributions; indeed, the energies of over 99% of the singlet samples are greater than the maximum energy on the x axis in

FIG. 4. Normalized histogram (100 equally spaced bins) of potential energy of simplified propane conformations sampled by MD, singlet level sampling and doublet level sampling.

FIG. 5. Convergence of free energy estimates using singlet (thick line) and doublet (thin line) reference system for (a) alanine dipeptide, (b) dialanine, and (c) tetra-alanine. (Inset figures show the convergence of the doublet estimate.) The solid line indicates the mean of the estimate using 100 bootstrap resampled datasets, and the bars indicate the standard deviations.

Fig. 6, so the singlet distributions are not graphed. These high energies result mainly from steric clashes. It is also worth remarking that, even for the doublet reference state, the fraction of high-energy samples increases with the size of the molecules, and the energy distribution correspondingly shifts toward higher energies.

Some of the convergence graphs display relatively long plateaus followed by sudden drops. This is particularly evident in the singlet results for tetra-alanine [Fig. 5(c)]. Such plateaus risk generating the deceptive appearance of a converged result. Thus, if a free energy estimate appears to be converged, but the energy overlaps are poor, then the apparent convergence may be illusory. On the other hand, if the overlap is extensive, then the convergence is more credible.

V. DISCUSSION

We have described a reference system method for computing the absolute free energy of a molecule, which achieves convergence with a relatively small number of conformational samples. The present approach builds on important prior work,¹¹ in which simulations were used to build singlet PDFs of internal molecular coordinates and these PDFs in turn were used to define a reference system with a known free energy. The singlet level reference state was then used as a basis for the calculation of molecular free energies. The

innovation in the present study grows from our prior development of a conformational sampling method that incorporates correlations of any given order, through the use of superposition approximations.^{22,26} The low-order PDFs needed to construct the superposition approximations are, as in the method of Ytreberg and Zuckerman,¹¹ derived from molecular simulations. Here, we show that this sampling approach can be used to construct a correlated reference system for use in calculating molecular free energies. A central result is that incorporating pair correlations in the reference system dramatically improves the convergence of the free energy estimates, in comparison to the singlet level reference.

It is worth elaborating on the two sources of bias in the free energy method presented here. In order to set up the reference distributions, the conformational space is discretized. This leads to a bias since the method effectively computes the discrete-space approximation [Eq. (17)] of the configurational integral [Eq. (1)]. However, the results for propane, where the deviation of the computed discrete-space approximation from the continuous-space analytic free energy is $<0.1\%$, suggest that the bias due to the discretization is small. Based on these propane results, we expect the discretization bias for peptides to be small as well, since the force-field and discretization are similar. The second source of bias stems from the fact that the reference distributions are constructed from a finite set of MD data. The MD simulation may not cover the desired

FIG. 6. Probability distribution of potential energy of conformations sampled by MD (thick line) and by doublet level sampling (thin dotted line) for (a) alanine dipeptide: energies of 99.9% singlet samples and 1.6% doublet samples was greater than 55 kT, (b) dialanine: energies of 99.9% singlet samples and 2.4% doublet samples was greater than 80 kT, and (c) tetra-alanine: energies of all singlet samples and 11.5% doublet samples was greater than 130 kT.

region of the configurational space, in which case the region accessible in the discrete space is restricted. Also, the limited data may lead to holes, or zero probability cells, in the PDFs used to set up the reference system, instead of small nonzero probabilities if the MD sampling had been exhaustive. These zero probability cells further restrict the conformational space accessible to the SA- l based ancestral sampling used to sample the reference distributions. However, although difficult to assess *a priori*, the bias due to the restrictions on the conformational space is expected to be low if the accessible region contains the low physical energy conformations that dominate the configurational integral.

As noted below, the present SA framework can be generalized to construct reference systems based upon various sets of joint PDFs, including second and potentially higher orders. It is, thus, worth noting that there can be a trade-off between the bias and convergence of the free energy estimate. As more PDFs are included, the conformational overlap increases, thereby speeding convergence, but the bias may increase because it is increasingly difficult to populate higher order PDFs so the reference distribution may have more holes. The optimal SA-based reference will be one that uses the least number of higher order PDFs while maintaining sufficient overlap to achieve convergence with reasonable number of reference samples. It might be feasible to reduce the number of PDFs used substantially by dropping those corresponding to weak correlations in the molecules. For instance, in a long-chain molecule, PDFs between torsions that are far apart in sequence and three-dimensional space might not be critical to ensuring good overlap. To identify the stronger correlations, simple rules such as close proximity in the bonded topology or in the three-dimensional space could be used, though these rules will likely be molecule dependent. A more sophisticated and automatic approach could be to use low mutual information^{39,40} as an indicator of weakly correlated sets of coordinates. Once the least important PDFs have been identified by this criterion, a “mixed-SA” reference state can be set up by approximating the dropped PDFs in terms of their lower order marginals. It would be of interest to learn how well such an approach can capture the correlations of more complex molecular systems while controlling computational costs.

The computational requirements of the current method depend on the dimensionality of the system, N , and the level of the reference system, l . The memory required to store all l -order PDFs scales as $O(N^l B^l)$, where B is the number of bins used to discretize each coordinate, and the cost of the sampling from the l -level reference distribution scales as $O(N^l)$. At the doublet level ($l = 2$), the largest molecular system in this work, tetra-alanine ($N = 150$), required roughly 124 MB memory and 1.26 s to sample a single conformation based on a MATLAB 7.5[®] implementation of the sampling algorithm on a 2.27 GHz personal computer. Note that the method will likely need to be used in conjunction with an implicit solvent model, in order to limit the number of degrees of freedom to a computationally manageable number. Indeed, even with an implicit solvent model, the computational requirements will be significant for most systems large enough to be of practical interest. The above strategy of using a subset of higher

order PDFs can mitigate the computational requirement substantially, primarily memory required to store the PDFs. Also, in order to scale the method to larger systems, advantage can be taken of the fact that sampling of conformations from the reference distribution using the ancestral sampling approach is trivially parallelizable over a distributed computing environment, since successively sampled conformations are uncorrelated.

In prior work, the use of a singlet level reference system was found to limit the applicability of the reference system method to small molecules,¹¹ because the overlap of the reference distribution with the physical ensemble falls with increasing system size. This limitation motivated the development of an innovative approach where the molecule is divided into fragments and samples are drawn separately for each fragment.¹⁵ The fragments are then assembled to obtain the free energy of the full molecule. The present study shows that incorporating correlation into the reference system yields good convergence for larger molecules without recourse to the fragment-based method. Ultimately, a combined approach may be of value for still larger molecules, especially when correlations above second order are expected to be important.

We envision using the free energy approach described here in practical applications, such as the calculation of host-guest or protein-small molecule binding affinities. This would be an “end-point” method, in which the absolute free energies of the bound and free states are subtracted to provide a binding free energy. Other end-point methods include Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) (Refs. 41 and 42), LIE (Ref. 43), and M2 (Ref. 44). These may be contrasted with pathway-based approaches, in which a free energy change is computed as the sum of free energy differences for a series of steps along a pathway joining the initial and final states. Examples include double decoupling⁴⁵ and equilibrium⁴ or nonequilibrium⁴⁶ “pulling” approaches. The strength of the current method is that, unlike MD or MC based sampling, the SA-based ancestral sampling is not limited by the energy barriers in the physical energy surface since the samples are uncorrelated. This feature should enable more exhaustive sampling of the physical energy surface leading to more accurate results.

ACKNOWLEDGMENTS

The authors thank Dr. Chris Jarzynski and his students Suryanarayanan Vaikuntanathan and Andy Ballard for helpful discussions. S.S. acknowledges funding from Ann G. Wylie Dissertation Fellowship. This publication was made possible by Grant No. GM61300 from the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health (NIH).

APPENDIX: DERIVATION OF EXACT FREE ENERGY FOR SIMPLIFIED PROPANE

We derive the partition function of propane, with $M = 11$ atoms, for a simplified energy function lacking nonbonded energy terms and possessing a single energy term corresponding

to each of the $N = 3 \times M - 6 = 27$ BAT coordinates (Fig. 1) used to specify the conformation. Thus, the energy function is given by

$$U_{\text{phys}}(\vec{b}, \vec{a}, \vec{t}) = \sum_{i=2}^M U_B(b_i) + \sum_{i=3}^M U_A(a_i) + \sum_{i=4}^M U_T(t_i). \quad (\text{A1})$$

Harmonic functions are used for bond-stretch and angle-bend energy terms:

$$\begin{aligned} U_B(x) &= \frac{1}{2} k_B (x - b_{\text{eq}})^2, \\ U_A(x) &= \frac{1}{2} k_A (x - a_{\text{eq}})^2, \end{aligned} \quad (\text{A2})$$

and the Ryckaert–Bellemans⁴⁷ potential is used for torsions:

$$U_T(x) = \sum_{i=0}^5 C_i (\cos(x - \pi))^i, \quad (\text{A3})$$

where k_B , b_{eq} , a_{eq} , k_A , and C_i are the force-field parameters. The configurational integral is then given by

$$\begin{aligned} Z_{\text{phys}} &= \int \exp(-\beta U_{\text{phys}}(\vec{b}, \vec{a}, \vec{t})) dV \\ &= \int \exp\left(-\beta \sum_{i=2}^M U_B(b_i)\right) \exp\left(-\beta \sum_{i=3}^M U_A(a_i)\right) \\ &\quad \times \exp\left(-\beta \sum_{i=4}^M U_T(t_i)\right) \\ &\quad \times b_3 \prod_{i=4}^M (b_i^2 \sin a_i) db_2 db_3 da_3 \left(\prod_{i=4}^M db_i da_i dt_i \right). \end{aligned} \quad (\text{A4})$$

Grouping bond, angle, and torsion variables gives

$$\begin{aligned} Z_{\text{phys}} &= \int \exp\left(-\beta \sum_{i=2}^M U_B(b_i)\right) b_3 b_4^2 \dots b_M^2 db_2 \dots db_M \\ &\quad \times \int \exp\left(-\beta \sum_{i=3}^M U_A(a_i)\right) \sin a_4 \dots \sin a_M da_3 \dots da_M \\ &\quad \times \int \exp\left(-\beta \sum_{i=4}^M U_T(t_i)\right) dt_4 \dots dt_M. \end{aligned} \quad (\text{A5})$$

By separating variables further, we can write the partition function as a product of one-dimensional integrals of the form:

$$\begin{aligned} \text{Bonds : } &\int_0^\infty \exp(-\beta U_B(b)) db; \int_0^\infty b \exp(-\beta U_B(b)) db; \\ &\int_0^\infty b^2 \exp(-\beta U_B(b)) db, \\ \text{Angles : } &\int_0^{2\pi} \exp(-\beta U_A(a)) da; \\ &\int_0^{2\pi} \sin(a) \exp(-\beta U_A(a)) da, \\ \text{Torsions : } &\int_0^{2\pi} \exp(-\beta U_T(t)) dt, \end{aligned} \quad (\text{A6})$$

all of which, except the integral over torsion coordinates can be computed analytically. The force-field parameters for propane are as follows. Parameters for the two bond types are

$$\begin{aligned} \text{C-H : } k_B &= 2.8225 \times 10^5, \quad b_{\text{eq}} = 0.1092, \\ \text{C-C : } k_B &= 2.5363 \times 10^5, \quad b_{\text{eq}} = 0.1535. \end{aligned} \quad (\text{A7})$$

The three angle types have parameters

$$\begin{aligned} \text{H-C-C or C-C-H : } k_A &= 3.8828 \times 10^2, \quad a_{\text{eq}} = 1.9207, \\ \text{C-C-C : } k_A &= 3.8828 \times 10^2, \quad a_{\text{eq}} = 1.9309, \\ \text{H-C-H : } k_A &= 3.8828 \times 10^2, \quad a_{\text{eq}} = 1.8911. \end{aligned} \quad (\text{A8})$$

The units of spring constants, k_B and k_A , are in kJ/mol/nm² and kJ/mol/rad², respectively, equilibrium bond lengths, b_{eq} are in nm and equilibrium angles, a_{eq} , are in radians. Finally parameters for the two torsion types, in kJ/mol, are

$$\begin{aligned} \text{H-C-C-C or C-C-C-H : } C_0 &= 0.66944, C_1 = 2.00832, \\ &C_3 = -2.67776 \\ \text{H-C-C-H : } C_0 &= 0.62760, C_1 = 1.88280, \\ &C_3 = -2.51040 \\ &(C_2 = C_4 = C_5 = 0). \end{aligned} \quad (\text{A9})$$

Substituting these parameters with $kT = 8.36$ kJ/mol corresponding to 1000 K gives the final free energy (in kT units) as

$$\begin{aligned} F_{\text{phys}} &= -\ln Z_{\text{phys}} \\ &= 74.6920. \end{aligned} \quad (\text{A10})$$

¹F. M. Ytreberg and D. M. Zuckerman, *J. Phys. Chem. B* **109**, 9096 (2005).

²M. R. Shirts and V. S. Pande, *J. Chem. Phys.* **122**, 134508 (2005).

³M. K. Gilson and H. Zhou, *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21 (2007).

⁴H. Woo and B. Roux, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6825 (2005).

⁵T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G. Rylance, R. Johnston, and D. Wales, *J. Chem. Phys.* **122**, 084714 (2005).

⁶W. G. Hoover, *J. Chem. Phys.* **55**, 1128 (1971).

⁷D. Frenkel and A. J. C. Ladd, *J. Chem. Phys.* **81**, 3188 (1984).

⁸W. G. Hoover, *J. Chem. Phys.* **47**, 4873 (1967).

⁹L. M. Amon and W. P. Reinhardt, *J. Chem. Phys.* **113**, 3573 (2000).

¹⁰J. P. Stoessel and P. Nowak, *Macromolecules* **23**, 1961 (1990).

¹¹F. M. Ytreberg and D. M. Zuckerman, *J. Chem. Phys.* **124**, 104105 (2006).

¹²S. W. Rick, *J. Chem. Theory Comput.* **2**, 939 (2006).

¹³V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, *J. Am. Chem. Soc.* **132**, 1526 (2010).

¹⁴D. E. Shaw, M. M. Deneroff, R. O. Dror, J. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, R. C. Ho, D. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. Mcleavy, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang, *SIGARCH Comput. Archit. News* **35**, 1 (2007).

¹⁵X. Zhang, A. B. Mamonov, and D. M. Zuckerman, *J. Comput. Chem.* **30**, 1680 (2009).

¹⁶S. Somani, B. J. Killian, and M. K. Gilson, *J. Chem. Phys.* **130**, 134102 (2009).

¹⁷V. Hnizdo and M. K. Gilson, *Entropy* **12**, 578 (2010).

¹⁸N. Gö and H. A. Scheraga, *Macromolecules* **9**, 535 (1976).

¹⁹K. S. Pitzer, *J. Chem. Phys.* **14**, 239 (1946).

²⁰C. Chang, M. J. Potter, and M. K. Gilson, *J. Phys. Chem. B* **107**, 1048 (2003).

²¹D. R. Herschbach, H. S. Johnston, and D. Rapp, *J. Chem. Phys.* **31**, 1652 (1959).

- ²²P. Attard, O. G. Jepps, and S. Marčelja, *Phys. Rev. E* **56**, 4052 (1997).
- ²³A. Singer, *J. Chem. Phys.* **121**, 3657 (2004).
- ²⁴G. Stell, in *The Equilibrium Theory of Classical Fluids*, edited by H. L. Frisch and J. L. Lebowitz, 1st ed. (Benjamin, New York, 1964).
- ²⁵H. Reiss, *J. Stat. Phys.* **6**, 39 (1972).
- ²⁶J. Kirkwood and E. Boggs, *J. Chem. Phys.* **10**, 394 (1942).
- ²⁷E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, England, 2003).
- ²⁸C. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., Information Science and Statistics (Springer, New York, 2007).
- ²⁹M. N. Rosenbluth and A. W. Rosenbluth, *J. Chem. Phys.* **23**, 356 (1955).
- ³⁰H. Meirovitch, *J. Mol. Recognit.* **23**, 153 (2010).
- ³¹R. Zwanzig, *J. Chem. Phys.* **22**, 1420 (1954).
- ³²J. Wang, R. Wolf, J. Caldwell, P. Kollman, and D. Case, *J. Comp. Chem.* **25**, 1157 (2004).
- ³³J. Wang, W. Wang, P. A. Kollman, and D. A. Case, *J. Mol. Graphics Modell.* **25**, 247 (2006).
- ³⁴E. Sorin and V. Pande, *Biophys. J.* **88**, 2472 (2005).
- ³⁵A. DePaul, E. Thompson, S. Patel, K. Haldeman, and E. Sorin, *Nucleic Acids Res.* **38**, 4856 (2010).
- ³⁶B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, *J. Chem. Theory Comput.* **4**, 435 (2008).
- ³⁷A. Zoubir and B. Boashash, *IEEE Signal Process. Mag.* **15**, 56 (1998).
- ³⁸A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*, 1st ed. (Cambridge University Press, Cambridge, England, 1997).
- ³⁹B. J. Killian, J. Y. Kravitz, S. Somani, P. Dasgupta, Y. Pang, and M. K. Gilson, *J. Mol. Biol.* **389**, 315 (2009).
- ⁴⁰C. McClendon, G. Friedland, D. Mobley, H. Amirkhani, and M. Jacobson, *J. Chem. Theory Comput.* **5**, 2486 (2009).
- ⁴¹J. Srinivasan, T. E. Cheatham III, P. Cieplak, P. A. Kollman, and D. A. Case, *J. Am. Chem. Soc.* **120**, 9401 (1998).
- ⁴²H. Gouda, I. D. Kuntz, D. A. Case, and P. A. Kollman, *Biopolymers* **68**, 16 (2003).
- ⁴³J. Aqvist, C. Medina, and J. Samuelsson, *Protein Eng.* **7**, 385 (1994).
- ⁴⁴C. Chang and M. K. Gilson, *J. Am. Chem. Soc.* **126**, 13156 (2004).
- ⁴⁵M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, *Biophys. J.* **72**, 1047 (1997).
- ⁴⁶F. M. Ytreberg, *J. Chem. Phys.* **130**, 164906 (2009).
- ⁴⁷J. Ryckaert and A. Bellemans, *Faraday Discuss. Chem. Soc.* **66**, 95 (1978).