Published in final edited form as:
J Chem Phys. 2007 July 14; 127(2): 024107. doi:10.1063/1.2746329.

Extraction of Configurational Entropy from Molecular Simulations via an Expansion Approximation

Benjamin J. Killian, **Joslyn Yudenfreund Kravitz**, and **Michael K. Gilson**
Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850

Abstract

A method is presented for extracting the configurational entropy of solute molecules from molecular dynamics simulations, in which the entropy is computed as an expansion of multi-dimensional mutual information terms, which account for correlated motions amongst the various internal degrees of freedom of the molecule. The mutual information expansion is demonstrated to be equivalent to estimating the full-dimensional configurational probability density function (pdf) using the Generalized Kirkwood Superposition Approximation (GKSA). While the mutual information expansion is derived to the full dimensionality of the molecule, the current application uses a truncated form of the expansion in which all fourth- and higher-order mutual information terms are neglected. Truncation of the mutual information expansion at the *n*th-order is shown to be equivalent to approximating the full-dimensional pdf using joint pdfs with dimensionality of *n* or smaller by successive application of the GKSA. The expansion method is used to compute the absolute (classical) configurational entropy in a basis of bond-angle-torsion internal coordinates for several small molecules as well as the change in entropy upon binding for a small host-guest system. Convergence properties of the computed entropy values as a function of simulation time are investigated and comparisons are made with entropy values from the second generation Mining Minima software. These comparisons demonstrate a deviation in -*TS* of no more than about 2 kcal/mol for all cases in which convergence has been obtained.

I. INTRODUCTION

The total entropy of a molecule in solution can be separated into two parts, a solvent entropy associated with solvent motions, and a solute, or configurational, entropy associated with solute motions[1]. Changes in the latter are thought to make substantial contributions to important biochemical processes like protein folding and molecular association. As a consequence, methods of quantifying the configurational entropy could help explain how biomolecules function, and methods of controlling it could be valuable for molecular design.

In recent years, it has become possible to estimate changes in configurational entropy from NMR data[2-7]. This approach has yielded remarkable insights; for example, binding in at least two systems appears to be associated with reduced local entropy in the binding site of a protein, balanced by increased local entropy far from the binding site[8,9]. However, NMR methods of assessing entropy currently rely upon assumptions regarding the orientational distributions underlying observed order parameters, and are not yet able to account for motional correlations. It is thus of interest to seek further insight by other means.

Computer modeling can be used to study the configurational entropy of a solute molecule, and have the potential to provide a more detailed look at the nature and biological roles of configurational entropy. For example, the Mining Minima (M2) method has provided evidence that configurational entropy is an important determinant of affnities for host-guest and protein-ligand binding[10]. However, M2 is not immediately applicable to a molecule as complex as a protein. Also, Mining Minima might be expected to underestimate entropy because it focuses on local energy wells, rather than explicitly including higher energy conformations as in the case of Monte Carlo or molecular dynamics simulations. Another approach, the hypothetical scanning method, can be used to determine the entropy associated with a specified microstate of a molecular system[11,12], and has been used to compare the entropy of a helical and a hairpin conformation of a peptide. However, both hypothetical scanning and M2 require specialized software, and a method of extracting configurational entropy from conventional molecular dynamics (MD) or Monte Carlo simulations would be of considerable interest.

The quasiharmonic approximation[13] is perhaps the best known method of extracting the configurational entropy from a molecular simulation. It involves diagonalizing the covariance matrix of the atomic coordinates and approximating their probability distribution function (pdf) as a multidimensional Gaussian distribution with variances equal to its eigenvalues. The configuration integral of this pdf can then be obtained analytically. Perhaps the chief drawback of the quasiharmonic approximation is that it can substantially overestimate the entropy for systems with multimodal pdfs; i.e., with multiple occupied energy minima.[14,15]. Also, when applied to a trajectory in Cartesian coordinates, it requires that the snapshots be rotationally and translationally superimposed so that a reasonable covariance matrix may be calculated, and different results can be obtained depending upon how the superposition is done.

Other simulation-based approaches also deserve mention. For example, histogramming methods can be used to compute entropies via the density of states[16,17], and have been successfully applied to the solvation of small molecules[18]. They have also been applied to peptides[19,20], though convergence properties in these applications have not yet been reported. An intriguing approach based upon covariance analysis of torsion angles in the complex plane has recently been presented[21]. Methods of calculating entropy differences from simulations are further addressed in a helpful recent review and analysis by Peter *et al*[22].

The present paper describes a new approach to computing configurational entropy based on superposition approximations to the full-dimensional probability density in terms of one-dimensional, two-dimensional, three-dimensional, and higher-order pdfs. Such superposition approximations are well known in the theory of homogeneous fluids[23-29] with references too numerous to fully enumerate here, and have previously been used to establish entropy expansions for liquids[30-34]. More recently entropy expansion methods have been applied to signal processing[35] and more general complex systems, such as frustrated spin systems[36]. Despite the rich body of work in this area, these density-based entropy expansions have not, to our knowledge, been applied to configurational entropy of a solute molecule.

We follow in particular a formulation in which the entropy is expressed in terms of pairwise and higher-order mutual information terms[36], where an *n*th-order mutual information captures the degree to which the *n*-dimensional correlation influences the entropy. This mutual information expansion can be directly related back to a superposition approximation of the pdf, as demonstrated in the Theory section. The accuracy of the computed entropy should in principle increase as higher order terms in the series are included. However, the computations become more difficult because of the rapid increase in the number and the order of the multi-particle pdfs that need to be analyzed. Here, calculations at the pairwise level are found to yield remarkably good agreement with independent M2 calculations for model systems, and it is proposed that the pairwise level provides a favorable balance of accuracy and efficiency.

II. THEORY

Although the concepts employed here are not new, they have not previously been applied to the calculation of the configurational entropy of a molecule from a simulation, to our knowledge, and therefore are detailed here in connection with configurational entropy for the sake of clarity and specificity.

A. Partial molar configurational entropy

The standard chemical potential of a molecule in solution can be written as[37]

$$\mu^{\mathrm{o}} = -RT \quad \ln\left(\frac{8\pi^2}{C^{\mathrm{o}}}Z\right) \tag{1}$$

where R is the gas constant, T is absolute temperature, C^{o} is standard concentration, and Z is the configuration integral

$$Z \equiv \int e^{-\beta E(\mathbf{r})} d\mathbf{r}. \tag{2}$$

Here $\beta = (RT)^{-1}$ and the integration is performed over a set of Cartesian internal coordinates, $\mathbf{r}$ that define the molecular conformation, such as Anchored Cartesian coordinates[38,39]. This expression omits a Jacobian determinant contribution that is nearly constant across thermodynamically accessible conformations (in Ref. 38, see Equation (13) and associated text). For a molecule in the gas phase, $E(\mathbf{r})$ is the vacuum potential energy as a function of conformation, $U(\mathbf{r})$ [40]. For a molecule in solution, the potential energy is supplemented by the solvation energy as a function of conformation: $E(\mathbf{r}) = U(\mathbf{r})+W(\mathbf{r})$ [37]. The factor of $\frac{8\pi^2}{C^{\mathrm{o}}}$ in effect accounts for the rotational and translational freedom of the molecule in solution[37,41] at standard concentration.

The partial molar configurational entropy S^{o} is given by[1]

$$-TS^{\mathrm{o}} = \mu^{\mathrm{o}} - \langle E \rangle , \tag{3}$$

where the angle brackets are used to indicate an ensemble average. By introducing the concept of the full-dimensional pdf, $\rho(\mathbf{r})$, over the configuration space of the molecule, the partial molar configurational entropy can be rewritten in the form (see Appendix A)

$$-TS^{\mathrm{o}} = -RT\ln\frac{8\pi^2}{C^{\mathrm{o}}} + RT\int \rho(\mathbf{r}) \quad \ln \quad \rho(\mathbf{r})\, d\mathbf{r}. \tag{4}$$

The negative of the integral in Equation (4) is the Shannon entropy associated with the pdf $\rho(\mathbf{r})$[42] and is closely related to the Gibbs H-function[43,44]. This quantity, when multiplied by the gas constant, becomes a physical entropy associated with the distribution in the configuration space. We will denote this as $S = -R\int \rho(\mathbf{r}) \ln\rho(\mathbf{r})\, d\mathbf{r}$, such that one can write

$$-TS^{\mathrm{o}} = -RT \quad \ln \quad \frac{8\pi^2}{C^{\mathrm{o}}} - TS. \tag{5}$$

The quantity - ln $\rho(\mathbf{r})$ is called the information, or self-information, associated with a measurement on the distribution $\rho(\mathbf{r})$[45,46]. It is of interest to note that, in a statistical thermodynamic sense, the quantity S is itself an ensemble average of the information associated with the pdf; that is $S = R\langle -\ln \rho(\mathbf{r})\rangle$, where the gas constant is used to set the units.

B. Bond-angle-torsion coordinates

The present method uses an internal coordinate system comprising bond-lengths, bond-angles, and dihedrals, rather than Cartesian coordinates[38,39,41,47-49]. For a molecule with N atoms, a system of $3N$ - 6 bond-angle-torsion (BAT) coordinates can be defined as follows (see Figure 1). Three atoms connected by serial bonds (atoms 1, 2, 3 in Figure 1) are chosen as root atoms. Assuming no external field, the internal energy of the molecule is independent of origin. Therefore, one can define the position of atom 1 as the origin of the internal coordinate system, thus eliminating three external degrees of freedom. Additionally, the overall orientation of the molecule does not affect the energy. For this reason, one may define the bond between atoms 1 and 2 to lie along the z-axis of the internal frame, eliminating the azimuthal and polar angles of atom 2 with respect to the origin. Finally, the plane formed by atoms 1, 2, and 3 can be identified as the xz-plane of the internal coordinate system. This eliminates the polar angle that defines atom 3 relative to the z-axis. Fixing the orientation of the molecule in such a manner results in the removal of an additional three external degrees of freedom. These six external degrees of freedom will not be dealt with further here.

The internal coordinates are then constructed in a tree, starting with the root atoms, where the coordinates of atom i are given by its bond-length, bond-angle, and dihedral angle with respect to a series of three other atoms closer in the tree to the root atoms. When the dihedral angles of two atoms are defined with respect to the same three prior atoms (e.g., atoms 5, 6 in Figure 1), they can be defined independently as $\boldsymbol{\phi}_i$ and $\boldsymbol{\phi}_j$. Alternatively, one can be defined as a phase angle with respect to the other[49]: $\phi'_j = \phi_j - \phi_i$. In this way, the torsion of one methyl hydrogen will range from 0 to 2π, but the torsions of the other two will have tight probability distributions around $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$. For a receptor-ligand complex, a pseudo-bond is formed between one atom of the receptor and one atom of the ligand; this leads to incorporation of pseudo-bond length, 2 pseudo-bond angles, and 3 pseudo-bond torsions.

Transforming from Cartesian to BAT coordinates $\{\mathbf{b}, \boldsymbol{\theta}, \boldsymbol{\phi}\}$ requires multiplying the volume element of the integral by a Jacobian determinant of the form $J(b,\theta) = b_2^2 \prod_{i=3}^{N} b_i \sin\theta_i$, where b_i is the bond length between atoms i - 1 and i and θ_i is the bond angle defined by atoms i - 2, i - 1 and i[39]. Hence,

$$S = -R \int J(\mathbf{b},\theta) \quad \rho(\mathbf{b},\theta,\phi) \quad \ln \quad \rho(\mathbf{b},\theta,\phi) \quad d\mathbf{b}\, d\theta\, d\phi. \tag{6}$$

The Jacobian determinant can often be omitted when calculating a unimolecular entropy change, because it depends only upon bond-lengths and bond-angles, which are rather rigid. However, for a binding reaction, the density of the free ligand is expressed in units of *length*$^{-3}$, so a correct calculation of the change in translational entropy requires including the Jacobian terms that convert the pseudobond coordinates (previous paragraph) into units of *length*$^{-3}$. Furthermore, the present study compares the entropy expansion results with M2 calculations, which include bond-stretch and angle-bend coordinates with a full Jacobian, so the full Jacobian is included in all data reported here.

C. First order entropy approximation

To simplify notation $\{\mathbf{b}, \boldsymbol{\theta}, \boldsymbol{\phi}\}$ is now replaced by x ≡ $(x_1, \cdots, x_m)$, where $m = 3N - 6$, N being the number of atoms. Likewise, $J(x_i, x_j, x_k)$, for example, now symbolizes those parts of the Jacobian determinant associated with variables x_i, x_j, x_k. Thus, if $x_i = b_3$, $x_j = \theta_5$ and $x_k = \boldsymbol{\phi}_4$, then $J\left(x_i, x_j, x_k\right) = b_3^2 \sin\theta_5$. The total configurational entropy for this molecule is then

$$S\ (x_i,\cdots,x_m)=-R\int J\,(x_1,\cdots,x_m)\,\rho\,(x_1,\cdots,x_m)\ \ln\rho\,(x_1,\cdots,x_m)\,dx_1\cdots dx_m. \tag{7}$$

Although the *m*-dimensional pdf could in principle be determined by means of a simulation, through histogramming or via parametric or non-parametric density estimators, achieving adequate convergence is not practical for most systems of chemical or biological interest. As a result, an approximation must be employed.

Prior to establishing the first order approximation to the *m*-dimensional entropy, it is necessary to develop a consistent notation for use in the subsequent derivations. In this work, the exact form of a marginal or joint density, entropy, or mutual information of a specific dimensionality will be identified using a subscript equal to the dimensionality. For example, the three-dimensional joint probability density of x_3, x_4, and x_5 will be denoted as

$$\rho_3\,(x_3,x_4,x_5)\equiv\int J\,(x_1,x_2,x_6,\cdots,x_m)\,\rho\,(x_1,\cdots,x_m)\,dx_1dx_2dx_6\cdots dx_m. \tag{8}$$

Further, when a quantity is approximated at a particular order, this will be denoted using a superscript with the order of the approximation in parentheses. As an example, the third order approximation to the full-dimensional entropy will be expressed as $S^{(3)}$.

One may begin by considering the approximation that motions along the *m* degrees of freedom are completely uncorrelated with each other, as previously noted[50]. This assumption allows the pdf to be estimated to first order as

$$\rho^{(1)}\,(x_1,\cdots,x_m)\approx\prod_{i=1}^{m}\rho_1\,(x_i) \tag{9}$$

where $\rho_1(x_i)$ is the marginal probability density for x_i; e.g.,

$$\rho_1\,(x_1)\equiv\int J\,(x_2,\cdots,x_m)\,\rho\,(x_1,\cdots,x_m)\,dx_2\cdots dx_m. \tag{10}$$

Substituting $\rho^{(1)}(x_1,\cdots,x_m)$ for $\rho(x_1,\cdots,x_m)$ within the logarithm in Equation (7) allows one to separate the entropy into a sum of integrals of the form

$$S_1\,(x_i)\equiv-R\int J\,(x_i)\,\rho_1\,(x_i)\quad\ln\quad\rho_1\,(x_i)\,dx_i. \tag{11}$$

The integral in Equation (11) is the marginal entropy associated with x_i. From this, the first order approximation to the full-dimensional entropy becomes the sum of the marginal entropies,

$$S^{(1)}=\sum_{i=1}^{m}S_1\,(x_i) \tag{12}$$

Thus, the assumption that the degrees of freedom are uncorrelated allows simplification from an intractable *m*-dimensional pdf to *m* far more tractable one-dimensional pdfs. However, the resulting first-order approximation to the entropy is inaccurate when correlations exist. More particularly, it is an upper bound which applies in the limit of completely uncorrelated motions[36,50].

D. Approximate factorizations of the probability density

In order to account for correlations among the various degrees of freedom, it is desirable to approximate the full-dimensional density at a higher order. This is a well developed method in the field of liquid theory[23-29], and the connection between the approximated density and the expansion approximation of the entropy has been previously investigated[30-36]. The derivation of the mutual information expansion in this section follows these works.

In a series of papers, Kirkwood sought to formulate a general theory of molecular distributions in liquids[23,24], which led to the approximation of a three-point distribution in the liquid as a superposition of contributions from two-point distributions; i.e., to the Kirkwood Superposition Approximation (KSA)[25]. The Kirkwood approximation was later shown to be a specific case of more general expansions which allow an n-point distribution to be estimated using corresponding (n - 1)-point distributions[26-29,35]. This estimation is called the Generalized Kirkwood Superposition Approximation (GKSA). Reiss[28] and Singer[29] demonstrated that the KSA and the GKSA are, from a variational standpoint, the optimal approximations of an n-particle distribution for $n \geq 3$.

We employ the GKSA to approximate the full-dimensional pdf for the internal coordinates of a molecular species. We then demonstrate that the full-dimensional entropy can be approximated using a mutual information expansion obtained by substituting the GKSA-estimated pdf inside the logarithm of the entropy integral. This approximation allows one to expand the logarithm as a summation, and permits the entropy integral to be separated into a sum of lower-dimensional terms.

We begin by returning to the m-dimensional pdf, $\rho(x_1, x_2, ..., x_m)$. If the pdf spans a space of dimension larger than three, one may approximate it using the GKSA: $\rho^{(m\text{-}1)}(x_1, x_2, ..., x_m) \approx \rho(x_1, x_2, ..., x_m)$. The Kirkwood approximation for the three-dimensional pdf is[25,29,35,36]

$$\rho^{(2)}(x_1,x_2,x_3) = \frac{\rho_2(x_1,x_2)\,\rho_2(x_1,x_3)\,\rho_2(x_2,x_3)}{\rho_1(x_1)\,\rho_1(x_2)\,\rho_1(x_3)}. \tag{13}$$

This can be expressed in a shortened notation as

$$\rho^{(2)}(x_1,x_2,x_3) = \frac{\prod_{C_2^3}\rho_2}{\prod_{C_1^3}\rho_2}, \tag{14}$$

where ρ_2 denotes a two-dimensional joint pdf, and the product notation $\prod_{C_n^m}$ indicates that all of the possible $\binom{m}{n}$ unique combinations of the given subsets of degrees of freedom must be included in the product.

One can then insert the KSA-approximated density of Equation (14) into the logarithm of the full-dimensional entropy expression in Equation (7) and expand the argument of the logarithm with product and quotient rules for logarithms, into the form

$$S^{(2)}(x_1,x_2,x_3)=-R\int J(x_1,x_2,x_3)\,\rho(x_1,x_2,x_3)\left[\ln\quad\rho_2(x_1,x_2)+\ln\quad\rho_2(x_1,x_3)+\ln\quad\rho_2(x_2,x_3)-\ln\quad\rho_1(x_1)-\ln\rho_1(x_2)-\ln\rho_1(x_3)\right]dx_1dx_2 \tag{15}$$

This separates into 6 integrals that can be evaluated individually. In each integral, one can integrate out the degrees of freedom that are not included inside the logarithm, thus reducing the full three-dimensional pdf in front to a lower dimensional marginal pdf. The result is

$$\begin{aligned} S^{(2)}(x_1,x_2,x_3)= \quad & -R\int J(x_1,x_2)\,\rho_2(x_1,x_2)\ln\rho_2(x_1,x_2)\,dx_1dx_2 \\ & -R\int J(x_1,x_3)\,\rho_2(x_1,x_3)\ln\rho_2(x_1,x_3)\,dx_1dx_3 \\ & -R\int J(x_2,x_3)\,\rho(x_2,x_3)\ln\rho_2(x_2,x_3)\,dx_2dx_3 \\ & +R\int J(x_1)\,\rho_1(x_1)\ln\rho_1(x_1)\,dx_1 \\ & +R\int J(x_2)\,\rho_1(x_2)\ln\rho_1(x_2)\,dx_2 \\ & +R\int J(x_3)\,\rho_1(x_3)\ln\rho_1(x_3)\,dx_3. \end{aligned} \tag{16}$$

Each integral in Equation (16) is now just a marginal or pairwise joint entropy:

$$S^{(2)}(x_1,x_2,x_3)=S_2(x_1,x_2)+S_2(x_1,x_3)+S_2(x_2,x_3)-S_1(x_1)-S_1(x_2)-S_1(x_3). \tag{17}$$

The above equation can be simplified by introducing the pairwise joint mutual information[45]

$$I_2\left(x_i,x_j\right)\equiv S_1(x_i)+S_1\left(x_j\right)-S_2\left(x_i,x_j\right), \tag{18}$$

which is a measure of correlation between two degrees of freedom that relates the amount of information about x_j that is gained by fully characterizing x_i[51]. Using this definition, Equation (17) can be rewritten in the form

$$S^{(2)}(x_1,x_2,x_3)=S_1(x_1)+S_1(x_2)+S_1(x_3)-I_2(x_1,x_2)-I_2(x_1,x_3)-I_2(x_3,x_2). \tag{19}$$

This approximate expression equals the first order approximation of Equation (12), corrected for pairwise correlations by including the pairwise joint mutual information. This offers a conceptually intuitive method of approximating the full-dimensional entropy, by first summing the individual contributions from each degree of freedom, then correcting for correlations among higher order terms. One may generalize this mutual information expansion to higher dimensionality.

For higher dimensionality, the GKSA allows one to approximate an *m*-dimensional pdf using the form[29,35,36]

$$\rho^{(m-1)}\left(x_1,x_2,\cdots,x_m\right)=\frac{\prod_{C^m_{m-1}}\rho_{m-1}}{\frac{\prod_{C^m_{m-2}}\rho_{m-2}}{\frac{\vdots}{\frac{\prod_{C^m_2}\rho_2}{\prod_{C^m_1}\rho_1}}}}, \tag{20}$$

where, as before, the term ρ_n represents a specific n-dimensional joint pdf and the products are over all unique combinations of the indicated dimensionality.

Returning to the definition of the entropy given in Equation (7), one can expand the logarithm term by approximating the argument of the logarithm using the GKSA of Equation (20),

$$S^{(m-1)}=-R\int J\left(x_1\cdots,x_m\right)\rho\left(x_1\cdots,x_m\right)\quad \ln\quad \rho^{(m-1)}\left(x_1\cdots,x_m\right)dx_1\cdots dx_m. \tag{21}$$

Following the derivation of the three-dimensional case (generalized in Appendix B), the m - 1 order approximation to the full-dimensional entropy can be written as

$$S^{(m-1)}=\sum_{n=1}^{m-1}(-1)^{m-n+1}\sum_{C^m_n}S_n. \tag{22}$$

Using the definition of higher-order joint mutual information[36], it can be shown that Equation (22) can be written as

$$S^{(m-1)}=\sum_{i=1}^{m}S_1\left(x_i\right)-\sum_{C^m_2}I_2\left(x_i,x_j\right)+\sum_{C^m_3}I_3\left(x_i,x_j,x_k\right)-\sum_{C^m_4}I_4\left(x_i,x_j,x_k,x_l\right)+\cdots. \tag{23}$$

This is the mutual information expansion offered by Matsuda[36].

Looking at Equation (23) from a computational standpoint, it becomes clear that one has not eliminated the sampling and storage issues that plague the computation of the full-dimensional entropy: instead of having to populate one m-dimensional histogram, one must populate m different m - 1 dimensional histograms. Further, there is an insurmountable combinatorial explosion of the lower dimensional histograms to be constructed, for a system of any size. What Equation (23) does offer is a systematic method for including higher order corrections to the approximated full-dimensional entropy. This expansion can be truncated at any level desired, allowing the entropy to be computed using a maximum dimensionality of the histograms to be constructed that is suitable for a given molecular system and available computational resources. Note that it is unnecessary ever to form a representation of the full-dimensional pdf by multiplying these lower dimensionality histograms; the low-dimensionality histograms themselves are all that is needed to compute the terms of the entropy expansion. However, it is of interest to discern the relation between the order of truncation and the level of approximation of the density in the GKSA.

E. Truncating the mutual information expansion: the connection to density

The GKSA can now be used as justification for a particular level of approximation to the entropy, providing a meaningful connection between estimation of the density at a given dimensionality and the highest order correction included in the truncated mutual information expansion. Suppose that one desires to limit the dimensionality of the joint pdfs to a small number such as two or three. One may then apply the GKSA successively to obtain an approximation of the *m*-dimensional pdf using joint pdfs of the desired maximum dimensionality. The simplest example occurs when one approximates a four-dimensional pdf using one- and two-dimensional joint pdfs only. This is done by first applying the GKSA to the four-dimensional pdf[26,28,29,35],

$$\rho^{(3)}(x_1,x_2,x_3,x_4)=\frac{\rho_3(x_1,x_2,x_3)\,\rho_3(x_1,x_2,x_4)\,\rho_3(x_1,x_2,x_5)\,\rho_3(x_2,x_3,x_4)}{\frac{\rho_2(x_1,x_2),\rho_2(x_1,x_3)\rho_2(x_1,x_4)\rho_2(x_2,x_3)\rho_2(x_2,x_4)\rho_2(x_3,x_4)}{\rho_1(x_1)\rho_1(x_2)\rho_1(x_3)\rho_1(x_4)}}. \tag{24}$$

Into this expression, one may substitute the Kirkwood approximation for each of the three-dimensional joint pdfs. After some simplifying algebra, the new approximation becomes

$$\rho^{(2)}(x_1,x_2,x_3,x_4)\approx\frac{\rho_2(x_1,x_2)\,\rho_2(x_1,x_3)\,\rho_2(x_1,x_4)\,\rho_2(x_2,x_3)\,\rho_2(x_2,x_4)\,\rho_2(x_3,x_4)}{\left[\rho_1(x_1)\,\rho_1(x_2)\,\rho_1(x_3)\,\rho_1(x_4)\right]^2}. \tag{25}$$

Substitution of Equation (25) into the logarithm of Equation (7) and subsequent integration yields the expression

$$\begin{aligned}S^{(2)}(x_1,x_2,x_3,x_4)= \quad & S_2(x_1,x_2)+S_2(x_1,x_3)+S_2(x_1,x_4)+S_2(x_2,x_3)+S_2(x_2,x_4)\\ & +S_2(x_3,x_4)-2\left[S_1(x_1)+S_1(x_2)+S_1(x_3)+S_1(x_4)\right].\end{aligned} \tag{26}$$

By using the definition of the joint pairwise mutual information, this can be rewritten as

$$\begin{aligned}S^{(2)}(x_1,x_2,x_3,x_4)= \quad & S_1(x_1)+S_1(x_2)+S_1(x_3)+S_1(x_4)-I_2(x_1,x_2)-I_2(x_1,x_3)\\ & -I_2(x_1,x_4)-I_2(x_2,x_3)-I_2(x_2,x_4)-I_2(x_3,x_4).\end{aligned} \tag{27}$$

which is equivalent to applying Equation (23) to a four-dimensional pdf and truncating the expansion after the second order corrections. From this it is clear that truncation of the mutual information expansion of the entropy at the pairwise level is equivalent to approximating the density by the GKSA using only one- and two-dimensional joint pdfs.

In general, one can estimate an *m*-dimensional pdf using only one-and two-dimensional joint pdfs by successive application of the GKSA. The general approximation takes the form

$$\rho^{(2)}(x_1\cdots,x_m)\approx\frac{\prod_{C_2^m}\rho_2}{\left[\prod_{C_1^m}\rho_1\right]^{m-2}}, \tag{28}$$

and the corresponding entropy expansion takes the form

$$\begin{aligned} S^{(2)}(x_1,\cdots,x_m) &\approx \sum_{C_2^m} S_2 - (m-2)\sum_{C_1^m} S_1 \\ &= \sum_{C_1^m} S_1 - \sum_{C_2^m} I_2. \end{aligned} \tag{29}$$

This expansion is identical to the general mutual information expansion in Equation (23) but truncated at the pairwise level of correlation. This can be easily extended to include three-dimensional joint pdfs as well. The general form of the density approximation is now

$$\rho^{(3)}(x_1\cdots,x_m) \approx \frac{\prod_{C_3^m}\rho_3}{\dfrac{\left[\prod_{C_2^m}\rho_2\right]^{m-3}}{\left[\prod_{C_1^m}\rho_1\right]^{C_2^m-2}}}. \tag{30}$$

The associated expansion of the entropy becomes

$$\begin{aligned} S^{(3)}(x_1,\cdots,x_3) &\approx \sum_{C_3^m} S_3 - (m-3)\sum_{C_2^m} S_2 + C_2^{m-2}\sum_{C_1^m} S_1 \\ &= \sum_{C_1^m} S_1 - \sum_{C_m^2} I_2 + \sum_{C_m^3} I_3. \end{aligned} \tag{31}$$

Again, this is the same as the expression given in Equation (23) truncated at the level of three-fold correlations. While one can generalize this approximation to any desired level, we will not extend beyond three dimensions in this paper.

It is worth noting that, assuming an nth order approximation of the density,

$$S^{(n)} = R\left\langle -\ln\rho^{(n)}\right\rangle = -R\int\rho \quad \ln \quad \rho^{(n)}dx, \tag{32}$$

where the angle brackets imply that -ln $\rho^{(n)}$ is averaged over the true pdf ρ rather than $\rho^{(n)}$. The quantity $S^{(n)}$ provides a measure of the amount of information we gain about the true distribution ρ by analysis of the approximate distribution $\rho^{(n)}$. This measure is closely related the Kullback-Leibler (K-L) divergence, or relative entropy, which provides a measure of difference between the true and approximate pdfs[46]. It is clear from the above analysis that the information about ρ that is lost by assuming $\rho^{(n)}$ (i.e., the K-L divergence between ρ and $\rho^{(n)}$) is precisely contained in the mutual information terms of order n+1 and larger.

F. Computational Details

The dynamical trajectories were all generated using the University of Houston Brownian Dynamics computer code[52]. The stochastic dynamics module was employed, along with a distance dependent dielectric model, $D_{ij} = 4r_{ij}$, where r_{ij} is the distance in Angstroms between atoms *i* and *j*. The program Quanta[53] was used to type the atoms and generate CHARMm22[54] force parameters.

Each trajectory was run for a minimum simulated time of 50 ns with a time-step of 0.001 ps, and was recorded at a rate of 10^5 snapshots for each ns of simulated time. The output trajectories were then converted into BAT coordinate files. The mutual information expansion method was applied through a computer code called the Algorithm for Computing Configurational

ENTropy from Molecular Mechanics (ACCENT-MM), in which all pdfs were generated by histograming. Marginal and pairwise joint histograms were constructed for all combinations of bond-length, bond-angle, and torsional degrees of freedom, with 120 bins used for each dimension. Threefold joint histograms were constructed only for torsional degrees of freedom. For computing the threefold mutual information associated with the torsions, 60 bins were used for each dimension, as this was found to produce the most stable numerical results. For the host-guest complex, the BAT coordinate system included a pseudobond that connected an atom in the guest molecule to an atom in the host molecule, as noted in the Theory section.

The results obtained using the entropy expansion method were compared with secondgeneration Mining Minima (M2) calculations[10] for the same molecules and the same energy model. The M2 calculations yield not only the chemical potential of the molecule but also the average energy, allowing one to calculate the configurational entropy using Equation (3). In the entropy expansion method, the average energy obtained from the dynamical trajectories was combined with the configurational entropy computed by the ACCENT-MM software to yield the chemical potential, again using Equation (3). Comparison with M2 offers cross-validation of the entropy expansion results against a tested methodology that relies on different computation formalisms.

Corrections for internal symmetries must be made before the ACCENT-MM results can be compared with those from M2. The M2 code utilizes a symmetry filtering algorithm that prevents overcounting of the energy minima by including only a single representative of each distinct configuration[55]. This is accomplished by finding all configurations that can be made to superimpose through motions along rotational or internal coordinates and removing duplicates. This becomes an issue for methyl rotations, for which M2 reduces three configurations to one, so that the ACCENT-MM entropy values must be correspondingly reduced by ln 3 for each methyl group present in the molecule, since the dynamics simulations allow each methyl to rotate 0 - 2π radians. In the case of cyclohexane, M2 includes a single instance of the thermodynamical prodominant chair configuration, while the MD simulations include two. Therefore, the ACCENT-MM entropy is reduced by ln 2 for comparison. It should be noted that chosing to use phase angles for constructing torsional pdfs does not reduce the configurational symmetry, and therefore the same corrections apply.

III. RESULTS

A. Hydrogen Peroxide

The simplest system that was investigated here is the hydrogen peroxide molecule. With four atoms, hydrogen peroxide has six internal degrees of freedom: three bonds, two angles, and one dihedral. Figure 2 demonstrates the structures of the pdfs for the H-O-O-H dihedral at 300, 500, and 1000 K, and Figure 3 documents the convergence of the first- and second-order approximations to the total entropy, $S^{(1)}$ and $S^{(2)}$, respectively, for all three simulation temperatures. Table I compares the average potential energies, temperature weighted standard entropies, and standard chemical potentials with the corresponding M2 results. The first- and second-order entropy values are very similar and agree well with the M2 results. Furthermore, the mean energies from the simulations and from M2 are comparable. As a consequence, the chemical potentials computed using the two methods agree favorably.

B. Methanol

Methanol is the simplest system investigated that includes strongly coupled torsional degrees of freedom, those for the three methyl hydrogens. The plots in Figure 4 demonstrate the marginal pdfs for a single H-O-C-H dihedral at 300, 500, and 1000 K. The pdfs of the torsions corresponding to the remaining two methyl hydrogens are essentially identical to the first, as expected for a well-converged simulation, and are therefore not shown here. Table II compares the entropies computed using ACCENT-MM at the first-, second-, and third-order with M2 results; chemical potentials and mean energies are again included. The first- and second-order approximations to the entropy do not agree well with the M2 results, deviating by as much as 9 kcal/mol. However, including the third-order mutual information terms bring the expansion results within 0.5 kcal/mol of the M2 entropies.

We conjectured that the large magnitudes of the third-order mutual informations resulted from the strong—and foreseeable—coupling of the three methyl hydrogens. This suggested that the importance of the third-order terms could be reduced by treating the torsions associated with two of the hydrogens as phase angles relative to the first (see the BAT coordinate discussion in the Theory section). These phase angles display tight distributions centered at $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$, respectively (data not shown). Repeating the calculations with this phase angle approach leads to excellent agreement between the second-order ACCENT-MM entropies with M2, with deviations no greater than 0.2 kcal/mol (Table III). In addition, the third-order mutual information values all drop to nearly zero, as evidenced by the equality of $S^{(2)}$ and $S^{(3)}$. The present analysis shows that a suitably chosen coordinate system can markedly reduce the importance of higher-order terms in the entropy expansion. The use of phase angles in effect builds an important part of the correlation information into lower-order joint pdfs. We observed similar results for other molecular species. As a consequence, the remainder of this paper uses only the phase angle treatment of coupled torsional angles.

Figure 5 demonstrates the convergence of the temperature weighted standard entropy as a function of simulation time for the temperatures investigated. The plots on the left hand side are for the “full torsion” data in which each torsion was treated independently, while those on the right hand side are from the “phase angle” approach. Excellent convergence appears to have been attained using both approaches.

Torsional contributions to the configurational entropy of methanol have previously been investigated by Demchuk and Singh[56]. They performed Monte Carlo simulations on ensembles of methanol molecules at 300 and 1000 K using AMBER-style force field parameters, and constructed marginal pdfs using parameterized trimodal von Mises fitting functions[56]. Figures 4a and 4c offer direct comparison between the histograms constructed using ACCENT-MM with the von Mises functions generated by Demchuk and Singh. The two sets of pdfs have similar shapes, but those from the prior study are less sharply peaked, presumably due to differences in the force fields employed. The study by Demchuk and Singh yielded average values of 1.744 and 1.831 for the unitless quantity $S_1(\phi)/R$ associated with the methyl torsions at 300 K and 1000 K, respectively. For comparison, the corresponding marginal pdfs from this study yield values of 1.626 and 1.815 for the same respective temperatures. The two sets of data are quite similar, though the previous calculations resulted in slightly higher entropies, consistent with the flatter pdfs shown in Figure 4.

C. 1,2-dichloroethane

The BAT coordinates for 1,2-dichloroethane include five torsional angles, four of which can be defined as narrowly distributed phase angles with respect to the fifth. The fifth torsion, chosen to be the Cl-C-C-Cl torsion, is allowed to vary through the full range of 0 to 2π, as evident from its pdf (Figure 6). The value of $S^{(2)}$ again agrees well with the M2 results, as shown in Table IV. As in the case of methanol when phase angles were employed, the third-order mutual information terms for this molecule contribute little to the entropy. Also, the mean energies again demonstrate good agreement with the M2 values, so the chemical potentials compare favorably. Finally, Figure 7 documents good convergence for the three temperatures over the course of the 50 ns simulations.

Hnizdo and co-workers[57] studied the torsional entropy of 1,2-dichloroethane using Monte Carlo dynamics at 500 K with the OPLS force field. The marginal pdf for the Cl-C-C-Cl torsion was fit to a Fourier series expansion. As shown in Figure 6, their pdf agrees well with that obtained here, given the use of two different energy models. To compare the entropy values, we computed the Shannon entropy associated with the Fourier series expansion provided in the previous paper. The resulting value of $S_1(\phi)/R$ was found to be 1.19, compared to 1.01 in the present work.

D. Alkanes

In this section, we present data for ACCENT-MM computations on seven straight-chain and cyclic alkanes in order to evaluate convergence properties and accuracy for larger molecules. The compounds investigated are butane, pentane, hexane, heptane, octane, nonane, and cyclohexane, all simulated at 1000 K to promote conformational sampling. The results of 50 ns simulations of these alkanes are summarized and compared with M2 results in Table V. The first-order entropies deviate from the M2 results by more than 10 kcal/mol, but the second-order corrections improve the ACCENT-MM results to within about 2 kcal/mol of the M2 values for the straight-chain species. Convergence graphs in Figure 8 show that the second-order entropies are relatively well converged after 50 ns (long dash curves in Figure 8). The cyclohexane entropies converge faster than the hexane entropies (compare Figures 8c and 8g), as might be expected given the reduced volume of torsional space accessible to this cyclic molecule. This also coincides with the lower entropy displayed by the cyclohexane in comparison to hexane.

The third-order entropies for the straight-chain alkanes deviate from M2 more than the second-order entropies (Table V). However, these deviations must be interpreted in light of the convergence plots (dotted curves in Figure 8), which indicate that the third-order entropies are not well converged at 50 ns, yet are approaching the M2 results. Thus, the fully converged third-order results may well be similar to the second-order values. As for the previous systems, the average energies from M2 agree with the simulations to within 1 kcal/mol (Table V). Consequently, the chemical potentials obtained from M2 and from ACCENT-MM at the second-order agree to within about 2 kcal/mol.

Correlations contribute more strongly to entropy for cyclohexane than for the linear alkanes, as might be expected for this fairly stiff ring. This is apparent from the fact that the first-order approximation to the entropy deviates more from the M2 entropy for cyclohexane than for the straight-chain alkanes (Figure 8g and Table V). Moreover, whereas for the linear alkanes the second-order approximation converges to near the M2 results, the second-order entropy for cyclohexane deviates from M2 by more than 10 kcal/mol. This indicates that the second-order approximation still overestimates the total entropy of cyclohexane, presumably due to pronounced higher-order correlations. Accordingly, Figure 8g indicates that the third-order approximation is converging to a number close to the M2 value. It is concluded that correlations above second-order are modest but non-negligible in the case of cyclohexane.

The convergence properties were more fully examined for butane and nonane, the smallest and largest alkanes investigated here. Figure 9 shows the convergence of these two molecules for 150 ns of simulation. By 150 ns, the butane results have converged at all orders of approximation (Figure 9a). For nonane, although the second-order results have converged, the third-order approximation is still unconverged after 150 ns (Figure 9b). Despite the lack of convergence, one can make a comparison between the M2 values and an extrapolated prediction of the third-order entropy for these systems. We fit the third-order convergence data using a function of the form $S^{(3)}(t)=at^{-b}+S_{\infty}^{(3)}$, where t is the simulation time in ns and a, b, and $S_{\infty}^{(3)}$ are adjustable parameters. This functional form has been previously shown to offer good estimation of configurational entropy convergence data[15]. The benefit of using this functional form is that the parameter $S_\infty^{(3)}$ is the optimal value of the third-order entropy for an infinite simulation time and can be used as an estimation of the true third-order entropy. The parameters *a*, *b*, and $S_\infty^{(3)}$ were optimized using standard linear least squares (LLS) techniques.

The LLS optimization for the butane simulation data yielded values of *a* = 92.6702, *b* = 0.8342, and $S_\infty^{(3)}=23.0329$ kcal/mol. Thus the extrapolated third-order entropy is within about 1 kcal/mol of both the second-order approximation and the M2 value and differs from the reported third-order entropy at 150 ns by about 1.5 kcal/mol, indicating that good convergence has been achieved for butane. For the nonane data, the LLS optimized parameters were found to be *a* = 1202.0011, *b* = 0.8205, and $S_\infty^{(3)}=67.9114$ kcal/mol. The extrapolated estimation of the third-order entropy for nonane is about 20 kcal/mol less than the reported value at 150 ns. However, the estimated value the third-order entropy for nonane is in deviation from the M2 results by only about 5 kcal/mol, confirming that the third-order entropy appears to be converging to a value similar to that obtained using M2.

E. Urea Receptor

As a final example, we report the computed entropy changes upon binding for a simple host-guest complexation system, a dimethyl-substituted variant of a synthetic barbiturate receptor[58,59], which binds ethyleneurea via four hydrogen bonds[14,58] (Figure 10).

Table VI provides the thermodynamic quantities computed using both the ACCENT-MM and M2 methods. The absolute second-order entropies for the ligand and the receptor agree with M2 almost exactly, while differing by about 1.6 kcal/mol for the complex. Furthermore, as with the other systems studied here, the average energies from M2 agree closely with those from the simulations. The final result is that the change in chemical potential at second-order deviate from the M2 values by less than 1.2 kcal/mol. Figure 11 demonstrates convergence of the second-order approximations to the entropy (long dash curves). These are relatively well converged after the 50 ns simulation time. The third-order approximations do not compare as favorably with the M2 results (Table VI) but, as with the alkanes, their deviation appears to be attributable in large part to incomplete convergence of the third-order mutual information contributions (Figure 11).

IV. DISCUSSION

This study demonstrates the theoretical and practical application of a rigorous entropy expansion to flexible molecules and complexes. The first level of the expansion neglects all correlations and therefore provides an upper limit of the entropy. The second level accounts in addition for pairwise correlations and always yields a lower entropy than the first-order approximation. Successive levels correct for tertiary, quaternary and higher-order correlations and can either raise or lower the entropy, depending on the nature of the correlation[36]. At all levels, correlations are conceptualized and incorporated through pairwise and higher-order mutual information terms. It is furthermore shown that the entropy approximation at each order is a cross-entropy between the factorization approximation of the pdf at that order and the true pdf. For example, the second-order entropy approximation is based upon an approximation of the full pdf in terms of first- and second-order marginal pdfs; and higher order approximations of the entropy correspond to factorizations incorporating higher-order marginal pdfs.

The well-defined hierarchy of approximations in the present approach is appealing because it provides a clear path, though not always an easy one, to improving the quality of entropy calculations. In this sense, the expansion approach puts simulations on a similar footing to quantum chemical calculations, for which the incorporation of higher-order correlations also represents a systematic path of improvement.

¡¡MKG: edited this paragraph in light of corrected cyclohexane results.¿¿ The central issues of convergence and accuracy are addressed in practical terms here by the simulation studies. In all cases the first-order entropy appears to converge promptly, while the second-order entropy typically converges adequately within ~50 ns. Convergence to third order is considerably harder to achieve. It is thus encouraging that second-order approximations of the entropy agree well with reference results from the very different M2 method, except for the highly coupled cyclohexane ring, for which agreement appears improved by including third-order information. These observations arguably support the validity of both the M2 and the entropy expansion approaches. It is also encouraging to note that previous investigations of homogeneous liquids have demonstrated that third- and higher-order corrections are modest, even if non-trivial: it appears that, in the case of simple fluids, more than 80% of the total entropy can be accounted for by including the second-order (i.e., two-particle) densities[31, 33]. It thus seems likely that the second order approximation will, in many cases, provide a good balance of accuracy and computational speed.

It is important to use a suitable coordinate system with the entropy expansion method. For example, it is found that treating each hydrogen of a methyl group as having an independent torsion pushes substantial correlation into the third-order terms and therefore makes the calculation more complex. This particular problem can be avoided by treating two of the hydrogen torsions as phase angles relative to the other hydrogen torsion. More generally, it is likely that using Cartesian coordinates instead of BAT coordinates would make the entire methodology intractable. Conversely, there may exist some sets of collective coordinates that will make the method more tractable.

The entropy expansion method should be useful for systems of biological and chemical interest, and such an application is currently being pursued by the present authors. Although the calculations are computationally taxing because they involve large systems, it seems likely that adequate convergence can still be achieved at the second order and that this level of approximation will be instructive. In fact, the present results already are relevant because they indicate that second-order correlations make remarkably large contributions to the configurational entropy. This is evident from comparisons of the first- and second-order entropy data in the Results tables. Even for the simple host-guest system studied here, the second-order mutual information terms reduce the binding entropy penalty by several kcal/mol. This observation raises interesting questions regarding the interpretation of NMR-derived configurational entropies, which currently do not account for correlations. It is also worth noting that the present methodology can provide a structural picture of the origins of entropy changes, a capability of potential value for insight and design. It is not clear that the density of states approaches enable so detailed an analysis.

For densely packed molecules like proteins or systems containing flexible rings (such as cyclohexane above), third-order correlations may become more important and should be addressed by efforts to compute third-order mutual information terms. The nearest-neighbor entropy method[15,60-62] is likely to be particularly valuable for this purpose because it appears to make better use of the available simulation data than the histogramming approach taken here, and thus may speed convergence. Hnizdo *et al* have employed the nearest-neighbor method for computing configurational entropy of dihedral angles[15], while Kraskov and coworkers have applied the method to more general systems[61,62]. In addition, the theory presented here provides a clear basis for formulating mixed levels of approximation in the entropy expansion or, equivalently, the pdf, where highly correlated degrees of freedom would be treated at third order, say, while others would be treated only at second order. Such an approach could be of considerable value in the study of challenging molecular systems. Thus, a range of enhancements and applications of the present expansion approach can readily be envisioned.

Acknowledgments

The authors thank Drs. Vladimir Hnizdo, Robert Mnatsakanov, and Dan S. Sharp for bringing the nearest-neighbor entropy method to our attention, and for valuable comments regarding the manuscript. This publication was made possible by grant no. GM61300 from the National Institute of General Medical Sciences of the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of General Medical Sciences.

APPENDIX A

For a molecule with a potential energy $E(\mathbf{r})$ that is dependent only upon generalized position coordinates, the probability density of the spatial coordinates is given by[63]

$$\rho(\mathbf{r})=\frac{e^{-\beta E(\mathbf{r})}}{\int e^{-\beta E(\mathbf{r})}d\mathbf{r}}, \tag{A1}$$

such that normalization is imposed in the form $\int\rho(\mathbf{r})\,d\mathbf{r}=1$. This expression is obtained through the assumption of separability of the Hamiltonian into momentum and spatial parts within the Gibbs' distribution[63]. The corresponding momentum probability will not be dealt with here. In accordance with probability theory, the average value for an observable $X(\mathbf{r})$ over a continuous distribution is defined as[64]

$$\langle x(\mathbf{r})\rangle \equiv \int X(\mathbf{r})\,\rho(\mathbf{r})\,d\mathbf{r}. \tag{A2}$$

The configurational entropy is related to the standard chemical potential and average energy through the expression in Equation (3)

$$-TS^{o}=\mu^{o}-\langle E\rangle\,. \tag{A3}$$

Into this one can substitute the definition of the average energy and the definition of μ^{o} to obtain

$$-TS^{o}=-RT\frac{8\pi^{2}}{C^{o}}-\left[RT\ln\left(\int e^{-\beta E(\mathbf{r})}d\mathbf{r}\right)\right]\int\rho(\mathbf{r})\,d\mathbf{r}-\int E(\mathbf{r})\,\rho(\mathbf{r})\,d\mathbf{r}, \tag{A4}$$

where the definition of μ^{o} has been multiplied by the normalization constraint. The term in square brackets above is constant for a given temperature, and can be taken inside the normalization integral. This allows the last two terms on the right hand side to be combined as a single integral,

$$-TS^{o}=-RT\frac{8\pi^{2}}{C^{o}}-RT\int\rho(r)\left[\ln\left(\int e^{-\beta E(\mathbf{r})}d\mathbf{r}\right)+\beta E(\mathbf{r})\right]d\mathbf{r}. \tag{A5}$$

The second term in the square brackets can be written in the form of a logarithm, such that

$$-TS^{\mathrm{o}} = -RT\frac{8\pi^2}{C^{\mathrm{o}}} - RT\int \rho(r)\left[\ln\left(\int e^{-\beta E(\mathbf{r})} d\mathbf{r}\right) - \ln \quad e^{\beta E(\mathbf{r})}\right] d\mathbf{r}. \tag{A6}$$

Finally, the logarithms can be combined to yield the expression

$$-TS^{\mathrm{o}} = -RT\frac{8\pi^2}{C^{\mathrm{o}}} + RT\int \rho(\mathbf{r}) \ln\left(\frac{e^{-\beta E(\mathbf{r})}}{\int e^{-\beta E(\mathbf{r})} d\mathbf{r}}\right) d\mathbf{r}, \tag{A7}$$

which becomes

$$-TS^{\mathrm{o}} = -RT\frac{8\pi^2}{C^{\mathrm{o}}} + RT\int \rho(\mathbf{r}) \ln\rho(\mathbf{r})\, d\mathbf{r}, \tag{A8}$$

as was to be shown.

APPENDIX B

Equation (20) shows the GKSA for a general *m*-dimensional pdf. Substitution of this approximate pdf into the logarithm of the entropy integral found in Equation (7) results in Equation (21). One may use the product rule of logarithms to expand this expression (with special attention to the sign of the terms) to obtain

$$S^{(m-1)} = -R\int J(x_1, \cdots, x_m)\, \rho(x_1, \cdots, x_m) \left\{\sum_{n=1}^{m-1} (-1)^{m-n+1} \ln \quad \left[\prod_{C_n^m} \rho_n\right]\right\} dx_1 \cdots dx_m \tag{B1}$$

The summation can be taken outside of the integral and the negative sign absorbed into the exponent, yielding the expression

$$S^{(m-1)} = \sum_{n=1}^{m-1} (-1)^{m-n+2} R\int J(x_1, \cdots, x_m)\, \rho(x_1, \cdots, x_m) \ln \quad \left[\prod_{C_n^m} \rho_n\right] dx_1 \cdots dx_m. \tag{B2}$$

Each logarithm is a product of C_n^m different *n*-dimensional joint pdfs. The logarithm of each product of *n*-dimensional joint pdfs can be expanded using the product rule,

$$S^{(m-1)} = \sum_{n=1}^{m-1} (-1)^{m-n+2} R\int J(x_1, \cdots, x_m)\, \rho(x_1, x_2 \cdots, x_m) \sum_{C_n^m} [\ln\rho_n]\, dx_1 dx_2 \cdots dx_m, \tag{B3}$$

where, as before, the summation can be taken outside of the integral,

$$S^{(m-1)} = \sum_{n=1}^{m-1} (-1)^{m-n+2} \sum_{C_n^m} R\int J(x_1, \cdots, x_m)\, \rho(x_1, x_2 \cdots, x_m) \ln\rho_n dx_1 dx_2 \cdots dx_m. \tag{B4}$$

As any given *n*-dimensional joint pdf depends on exactly *n* of the *m* degrees of freedom, one may integrate over the remaining *m* - *n* dimensions independently of the logarithm terms. This integration will result in the reduction of the full *m*-dimensional pdf to the corresponding *n*-dimensional joint pdf, ρ_n. If one defines $d\tau^n$ as the differential element of volume corresponding to the *n* dimensions that remain to be integrated (including the remaining portions of the Jacobian determinant), the entropy can now be written as

$$S^{(m-1)}=\sum_{n=1}^{m-1}(-1)^{m-n+1}\sum_{C_n^m}\left\{-R\int\rho_n\ln\rho_n d\tau^n\right\}. \tag{B5}$$

The term in braces is exactly the *n*th-order joint entropy associated with a specific set of *n* degrees of freedom. Thus, one has the final form of the expansion

$$S^{(m-1)}=\sum_{n=1}^{m-1}(-1)^{m-n+1}\sum_{C_n^m}S_n. \tag{B6}$$

References

1. Chang C-EA, Chen W, Gilson MK. Proc. Natl. Acad. Sci. USA 2007;104:1534. [PubMed: 17242351]
2. Akke M, Brüschweiler R, Plamer AG III. J. Comp. Chem 1994;15:488.
3. Yang D, Kay LE. J. Mol. Biol 1996;263:369. [PubMed: 8913313]
4. Stone MJ. Acc. Chem. Res 2001;34:379. [PubMed: 11352716]
5. Prabhu NV, Lee AL, Wand AJ, Sharp KA. Biochemistry 2003;42:562. [PubMed: 12525185]
6. Homans SW. Chembiochem 2005;6:1585. [PubMed: 16038002]
7. Spyracopoulos L. Protein. Pept. Lett 2005;12:235. [PubMed: 15777271]
8. Lee AL, Kinnear SA, Wand AJ. Nat. Struct. Biol 2000;7:72. [PubMed: 10625431]
9. Arumugam S, Gao G, Patton BL, Semenchenko V, Brew K, Doren SRV. J. Mol. Biol 2003;327:719. [PubMed: 12634064]
10. Chang C-E, Gilson MK. J. Am. Chem. Soc 2004;126:13156. [PubMed: 15469315]
11. Cheluvaraja S, Meirovitch H. Proc. Natl. Acad. Sci. USA 2004;101:9241. [PubMed: 15197271]
12. Cheluvaraja S, Meirovitch H. J. Chem. Phys 2006;125:024905.
13. Karplus M, Kushick JN. Macromolecules 1981;14:325.
14. Chang C-E, Chen W, Gilson MK. J. Chem. Theory Comput 2005;1:1017.
15. Hnizdo V, Darian E, Fedorowicz A, Demchuk E, Li S, Singh H. J. Comput. Chem 2007;28:655. [PubMed: 17195154]
16. Ferrenberg AM, Swendsen RH. Phys. Rev. Lett 1989;63:1195. [PubMed: 10040500]
17. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. J. Comput. Chem 1992;13:1011.
18. Rick SW. J. Chem. Theory Comput 2006;2:939.
19. Ohkubo YZ, Brooks CL III. Proc. Natl. Acad. Sci. USA 2003;100:13916. [PubMed: 14615586]
20. Ohkubo YZ, Thorpe IF. J. Chem. Phys 2006;124:024910. [PubMed: 16422651]
21. Wang J, Brüschweiler R. J. Chem. Theory Comput 2006;2:18.
22. Peter C, Oostenbrink C, van Dorp A, van Gunsteren WF. J. Chem. Phys 2004;120:2652. [PubMed: 15268408]
23. Kirkwood JG. J. Chem. Phys 1935;3:300.
24. Kirkwood JG. J. Chem. Phys 1939;7:919.
25. Kirkwood JG, Boggs EM. J. Chem. Phys 1942;10:394.
26. Fisher IZ, Kopeliovich BL. Sov. Phys. Docl 1960;5:761.
27. Stell, G. The Equilibrium Theory of Classical Fluids. Frisch, HL.; Lebowitz, JL., editors. W. A. Benjamin, Inc.; New York: 1964.
28. Reiss H. J. Stat. Phys 1972;6:39.

29. Singer A. J. Chem. Phys 2004;121:3657. [PubMed: 15303932]
30. Raveché HJ. J. Chem. Phys 1971;55:2242.
31. Mountain RD, Raveché HJ. J. Phys. Chem 1971;55:2250.
32. Wallace DC. J. Chem. Phys 1987;87:2282.
33. Baranyai A, Evans DJ. Phys. Rev. A 1989;40:3817. [PubMed: 9902600]
34. Lazaridis T, Paulaitis ME. J. Phys. Chem 1992;96:3847.
35. Attard P, Jepps OG, Marčelja S. Phys. Rev. E 1997;56:4052.
36. Matsuda H. Phys. Rev. E 2000;62:3096.
37. Gilson MK, Given JA, Bush BL, McCammon JA. Biophys. J 1997;72:1047. [PubMed: 9138555]
38. Potter MJ, Gilson MK. J. Phys. Chem. A 2002;106:563.
39. Chang C-E, Potter MJ, Gilson MK. J. Phys. Chem. B 2003;107:1048.
40. Hill, TL. An Introduction to Statistical Thermodynamics. Addison-Wesley; Reading, MA: 1960.
41. Herschbach DR, Johnston HS, Rapp D. J. Chem. Phys 1959;31:1652.
42. Shannon CE. Bell System Tech. J 1948;27:379.
43. Tolman, RC. The Principles of Statistical Mechanics. Oxford University Press; London: 1938.
44. Jaynes ET. Am. J. Phys 1965;33:391.
45. Reza, FM. An Introduction to Information Theory. Dover Publications Inc.; New York: 1994.
46. Nalewajski, RF. Information Theory of Molecular Systems. Elsevier; Amsterdam: 2006.
47. Pitzer KS. J. Chem. Phys 1946;14:239.
48. Gō N, Scheraga HA. Macromolecules 1976;9:535.
49. Abagyan R, Totrov M, Kuznetsov D. J. Comp. Chem 1994;15:488.
50. Edholm O, Berendsen HJC. Mol. Phys 1984;51:1011.
51. Ihara, S. Information Theory for Continuous Systems. World Scientific Publishing; Singapore: 1993.
52. Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Ilin A, Antosiewicz J, Gilson MK, Bagheri B, Scott LR, et al. Comp. Phys. Comm 1995;91:57.
53. QUANTA 2005. Accelrys, Inc.; San Diego, CA: 2005.
54. MacKerell AD Jr. Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, et al. J. Phys. Chem. B 1998;102:3586.
55. Chen W, Huang J, Gilson MK. J. Chem. Inf. Comput. Sci 2004;44:1301. [PubMed: 15272838]
56. Demchuk E, Singh H. Mol. Phys 2001;99:627.
57. Hnizdo V, Fedorowicz A, Singh H, Demchuk E. J. Comput. Chem 2003;24:1172. [PubMed: 12820124]
58. Goswami S, Mukherjee R. Tetrahedron Lett 1997;38:1619.
59. Chang S-K, Hamilton AD. J. Am. Chem. Soc 1988;110:1318.
60. Singh H, Misra N, Hnizdo V, Fedorowicz A, Demchuk E. Am. J. Math. Manag. Sci 2003;23:301.
61. Kraskov A, Stögbauer H, Grassberger P. Phys. Rev. E 2004;69:066138.
62. Kraskov A, Stögbauer H, Andrezejak RG, Grassberger P. Europhysics Lett 2005;70:278.
63. Landau, L.; Lifshitz, E. Statistical Physics. Oxford University Press; London: 1938.
64. Parzen, E. Modern Probability Theory and Its Applications. John Wiley & Sons, Inc.; New York: 1960.

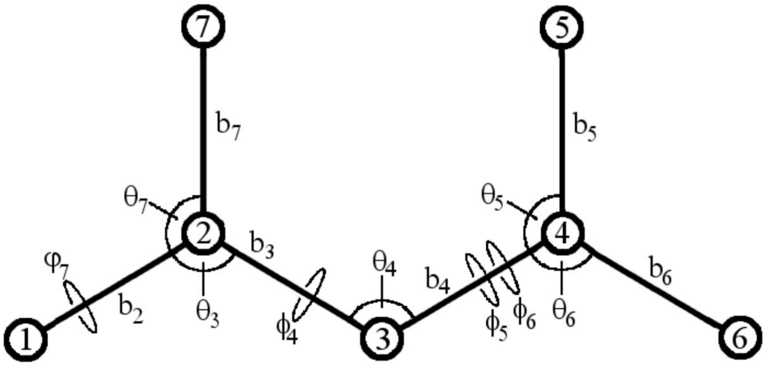

FIG. 1.
Schematic representation of the 3*N* - 6 bond-angle-torsion (BAT) coordinates for a small molecule. Atoms 1, 2, and 3 are chosen as the root atoms. The symbol ***ϕ*** denotes a proper torsion and the symbol *φ* denotes an improper torsion.

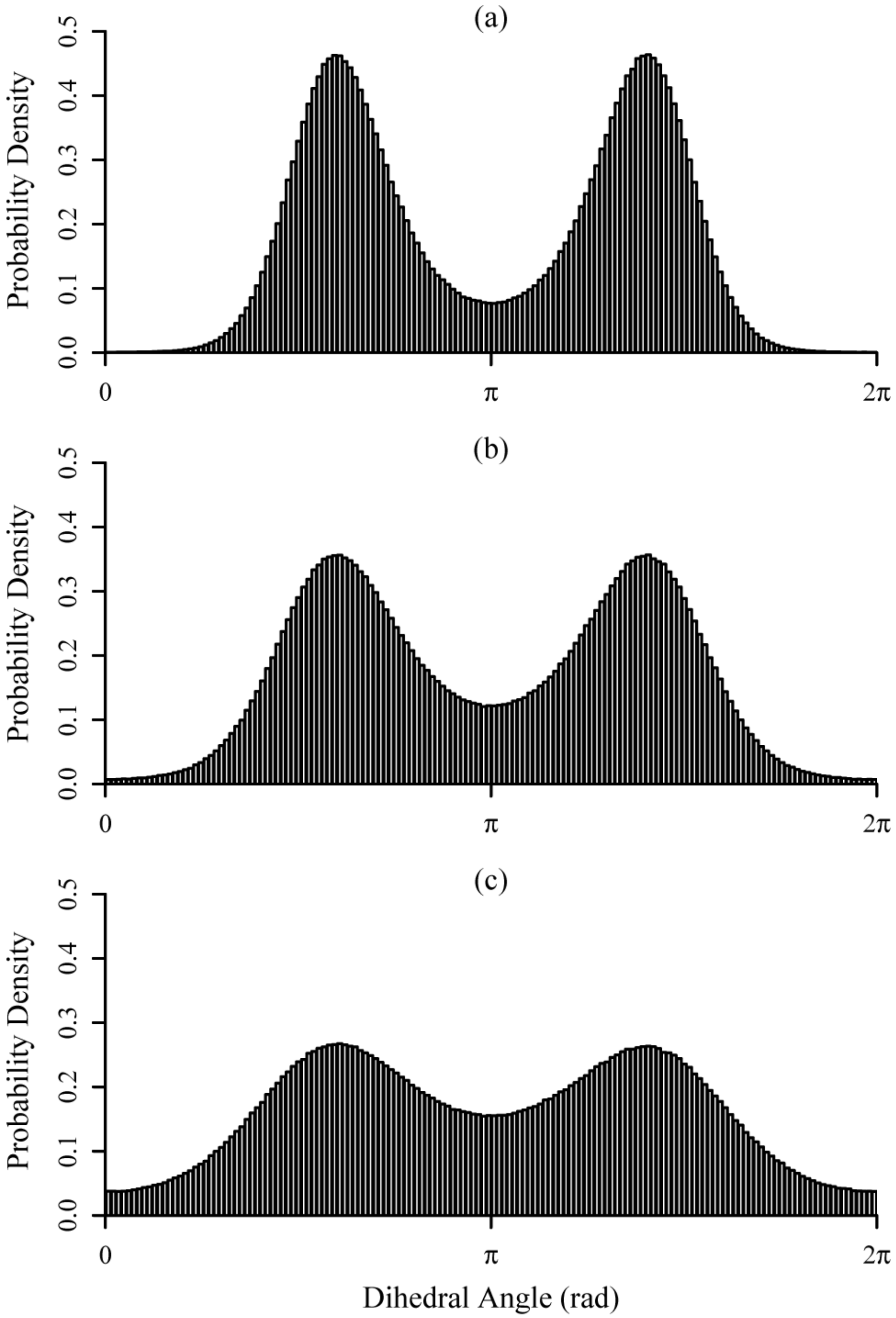

FIG. 2.

Plots of the ACCENT-MM probability density functions for the H-O-O-H dihedral angle in hydrogen peroxide for simulation temperatures of (a) 300 K, (b) 500 K, and (c) 1000 K.

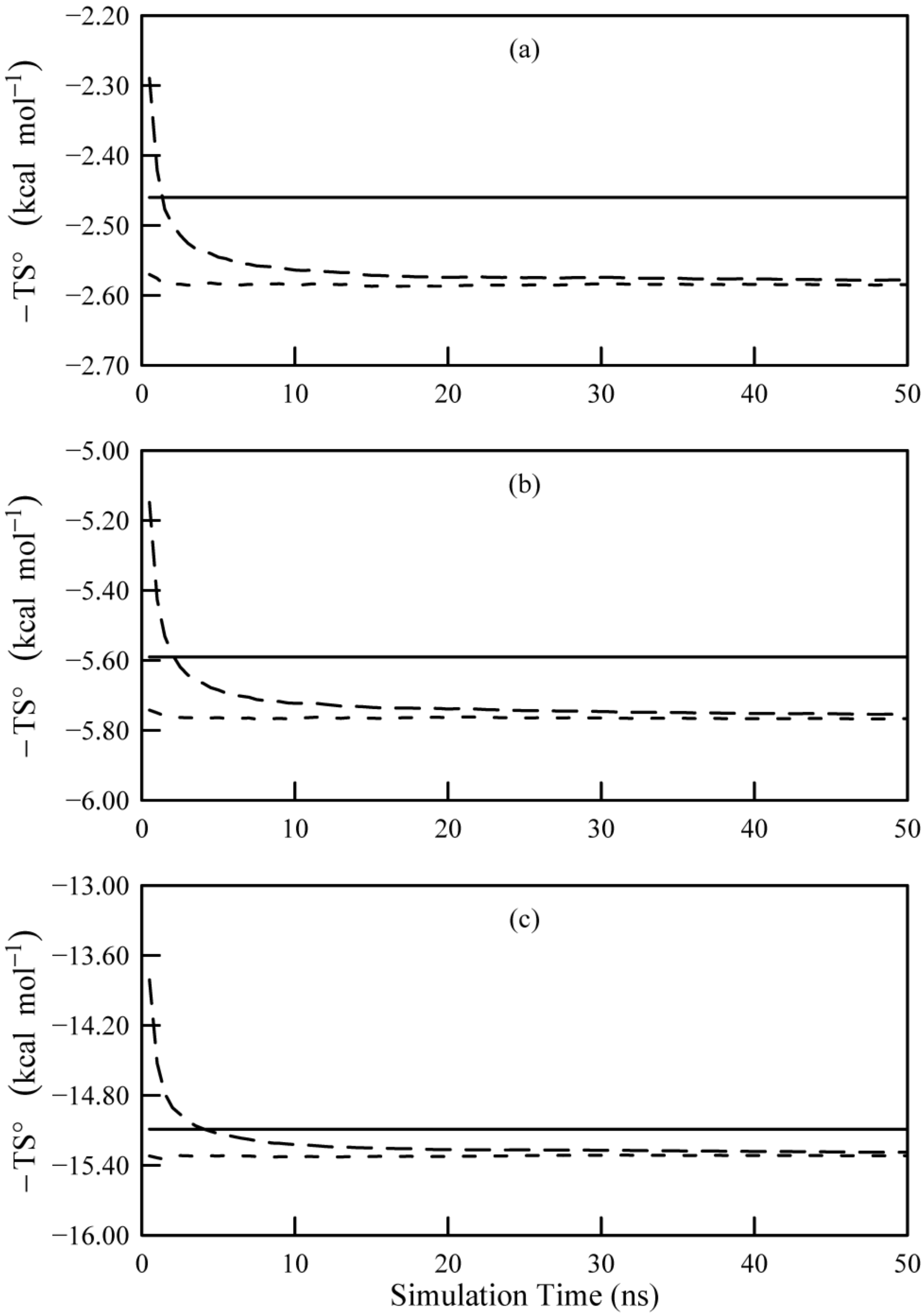

FIG. 3.

Convergence plots of the temperature weighted entropy for hydrogen peroxide as computed using ACCENT-MM for simulations at (a) 300 K, (b) 500 K, and (c) 1000 K. The data include the first-order approximation to the total entropy (short dash) and the second-order approximation to the total entropy (long dash). The solid line shows the value obtained using M2.

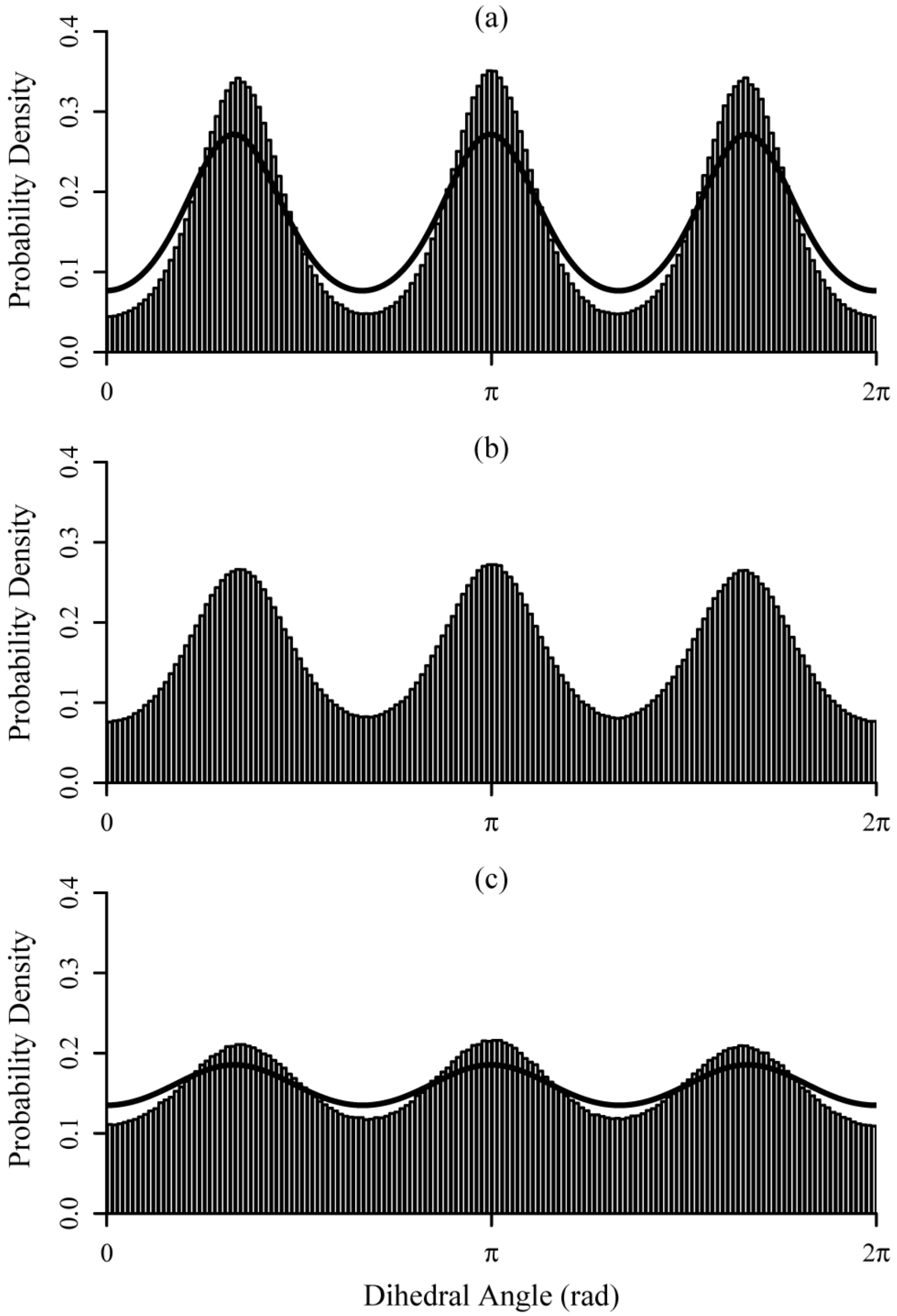

FIG. 4.

Plots of the ACCENT-MM probability density functions for a single H-O-C-H dihedral angle in methanol for simulation temperatures of (a) 300 K, (b) 500 K, and (c) 1000 K. The solid lines demonstrate pdfs generated by Demchuk and Singh[56].

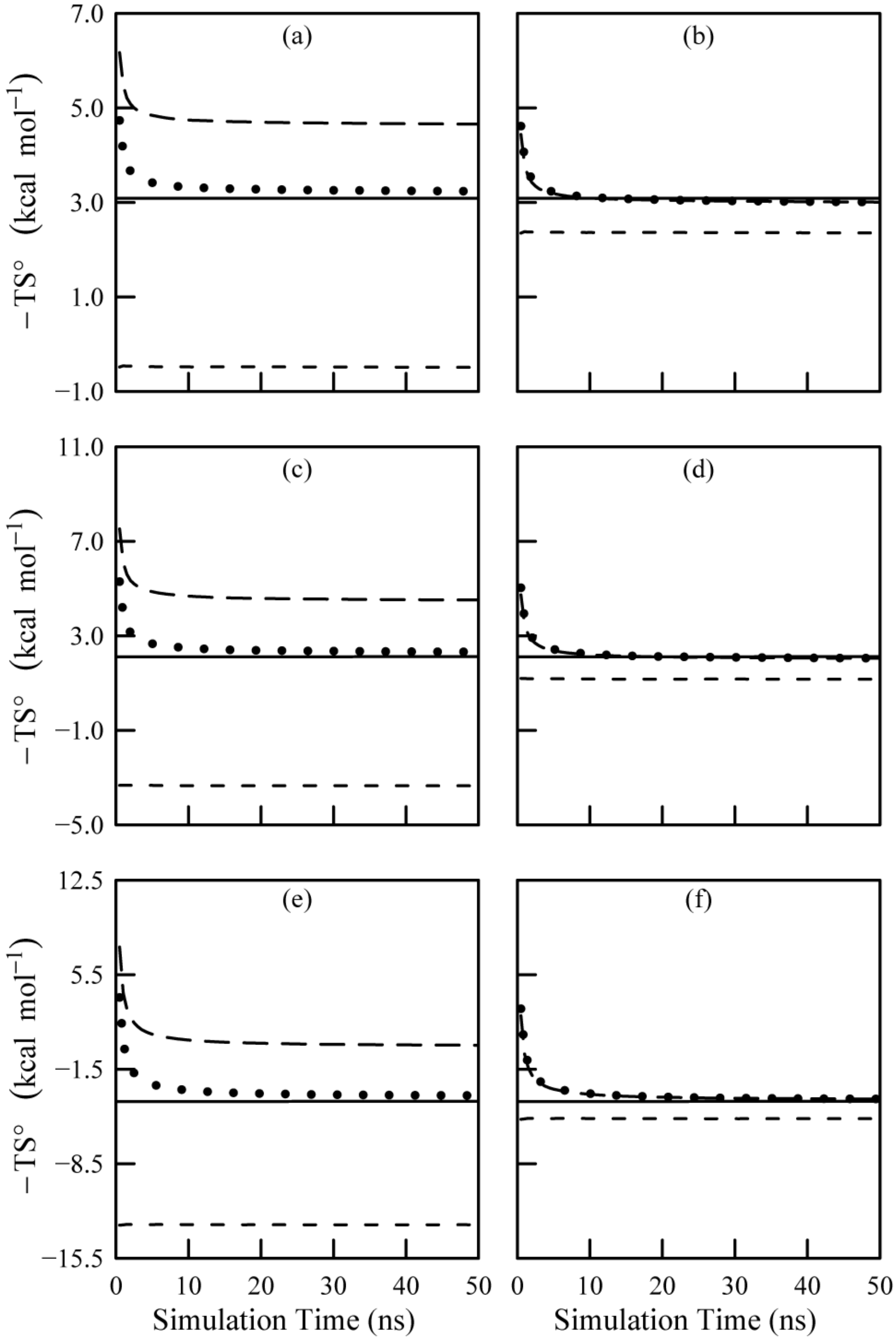

FIG. 5.

Convergence plots of the temperature weighted entropy for methanol computed with ACCENT-MM. Frames (a) and (b) are for a simulation temperature of 300 K using the “full torsion” and “phase angle” methods, respectively. Frames (c) and (d) are the same, but at 500 K. Frames (e) and (f) are the same, but at 1000 K. The data include the first-order approximation to the total entropy (short dash), the second-order approximation to the total entropy (long dash), and the third-order approximation to the total entropy (dotted). The solid line shows the value obtained using M2.

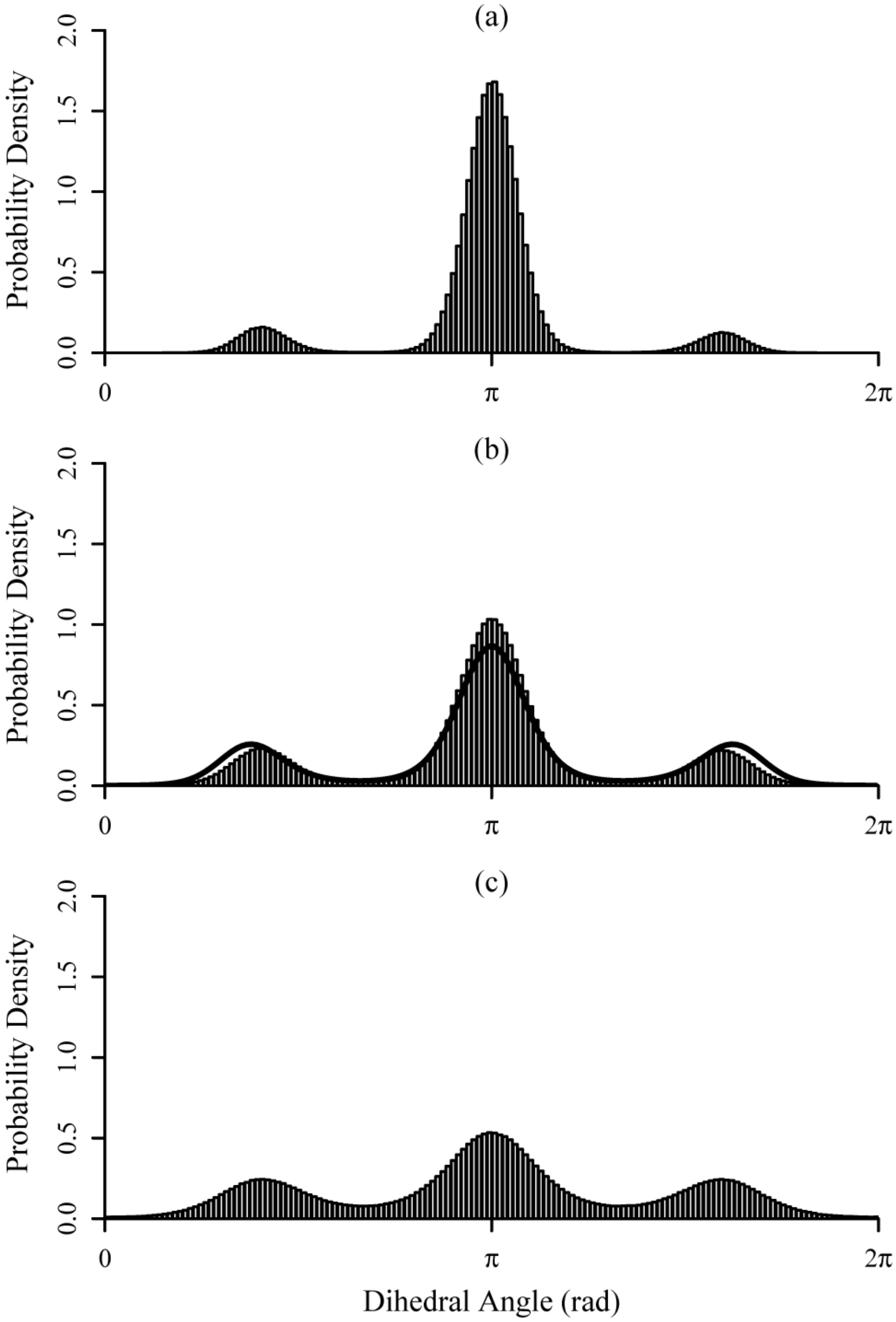

FIG. 6.
Computed probability density functions for the Cl-C-C-Cl dihedral angle in 1,2-dichloroethane, for simulation temperatures of (a) 300 K, (b) 500 K, and (c) 1000 K. The solid line shows the pdf generated by Hnizdo *et al*[57].

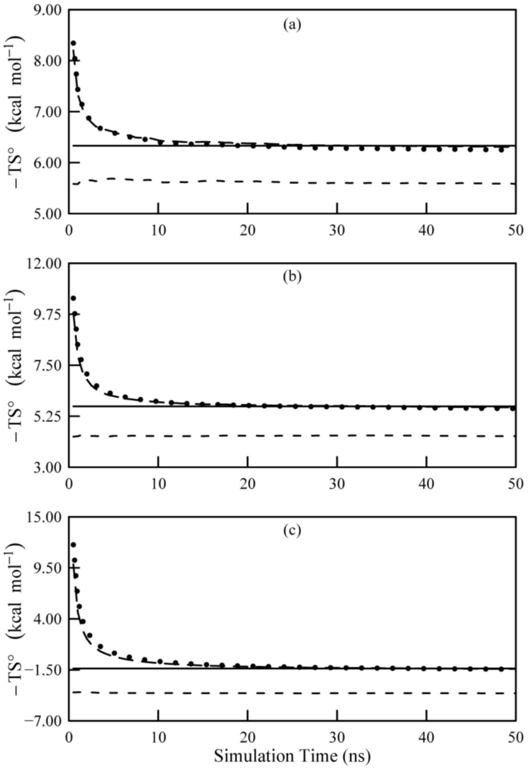

FIG. 7.

Convergence plots of the temperature-weighted entropy for 1,2-dichloroethane computed with ACCENT-MM for simulations at (a) 300 K, (b) 500 K, and (c) 1000 K. The data include the first-order approximation to the total entropy (short dash), the second-order approximation to the total entropy (long dash), and the third-order approximation to the total entropy (dotted). The solid line shows the value obtained using M2.

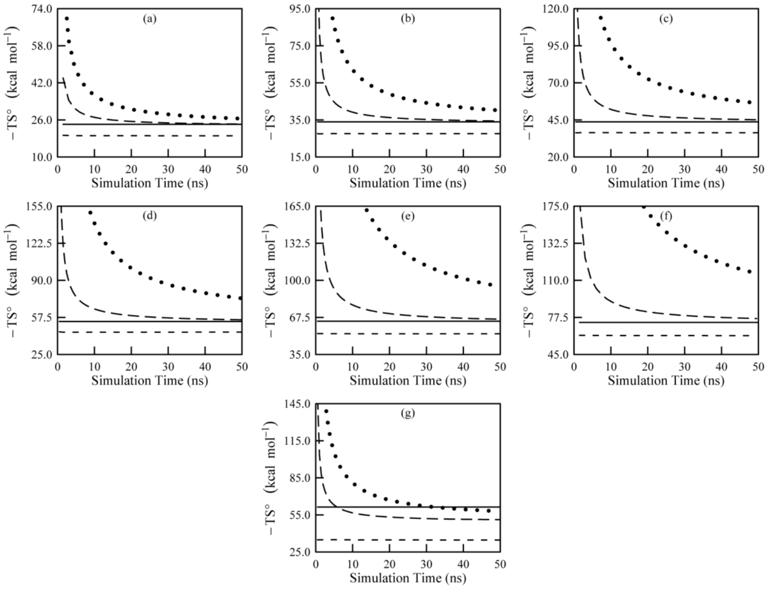

FIG. 8.

Convergence plots of the temperature weighted entropy for alkanes as computed using ACCENT-MM for a simulation temperature of 1000 K. The frames are: (a) butane; (b), pentane; (c), hexane; (d), heptane; (e), octane; (f), nonane; and (g), cyclohexane. The data include the first-order approximation to the total entropy (short dash), the second-order approximation to the total entropy (long dash), and the third-order approximation to the total entropy (dotted). The solid line shows the value obtained using Mining Minima.

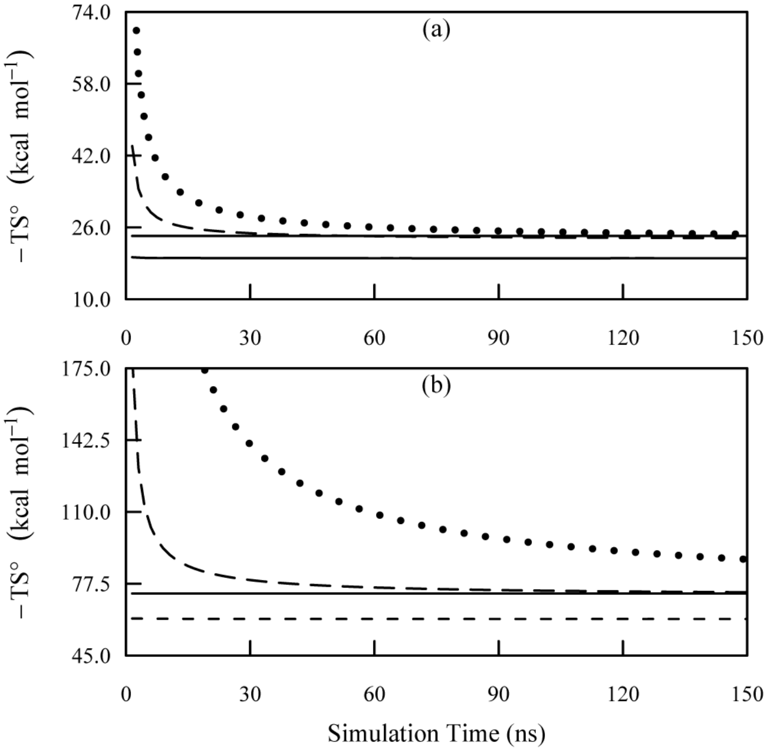

FIG. 9.

Convergence plots of the temperature weighted entropy for butane and nonane 150 ns of simulation as computed using ACCENT-MM for a simulation temperature of 1000 K. The data include the first-order approximation to the total entropy (short dash), the second-order approximation to the total entropy (long dash), and the third-order approximation to the total entropy (dotted). The solid line shows the value obtained using M2.

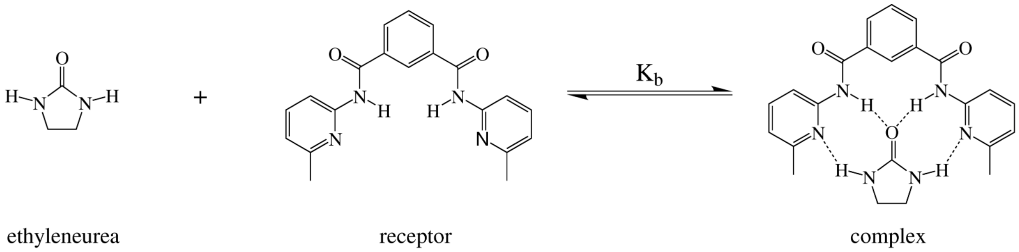

FIG. 10.

Complexation reaction for ethyleneurea with a synthetic receptor. Only those hydrogen atoms that participate in hydrogen bond interactions are shown explicitly, as dashed lines.

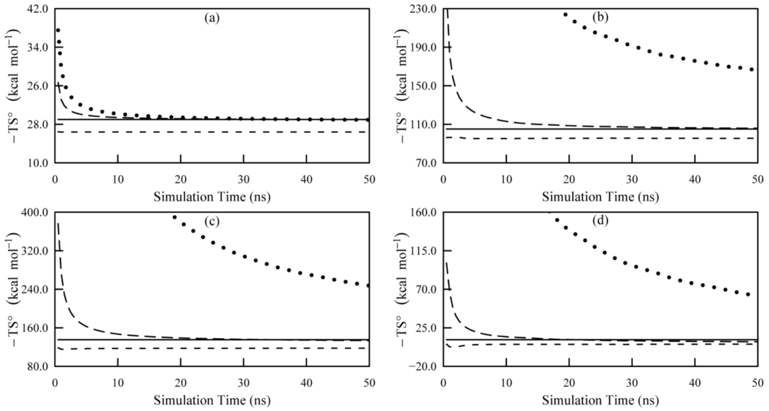

FIG. 11.

Convergence plots of the temperature weighted entropy for the complexation as computed using ACCENT-MM at a simulation temperature of 300 K. The frames are: (a) ethyleneurea; (b), receptor; (c), complex; (d), change on binding. The data include the first-order approximation to the total entropy (short dash), the second-order approximation to the total entropy (long dash), and the third-order approximation to the total entropy (dotted). The solid line shows the value obtained using Mining Minima.

TABLE I

Thermodynamic quantities for hydrogen peroxide computed with ACCENT-MM as compared with M2. All values are in kcal/mol

			ACCENT-MM	
		M2	**(1)[a]**	**(2)[b]**
300 K	$\langle E \rangle$	11.94	11.98	11.98
	$-TS^o$	-2.46	-2.58	-2.58
	μ^o	9.48	9.40	9.40
500 K	$\langle E \rangle$	13.13	13.14	13.14
	$-TS^o$	-5.59	-5.77	-5.75
	μ^o	7.54	7.37	7.39
1000 K	$\langle E \rangle$	16.11	15.89	15.89
	$-TS^o$	-15.09	-15.32	-15.29
	μ^o	1.02	0.57	0.60

[a] First-order approximation to the entropy

[b] Second-order approximation to the entropy

TABLE II

Thermodynamic quantities for methanol computed with ACCENT-MM using the “full torsion” pdfs as compared with M2. All values are in kcal/mol

			ACCENT-MM		
		M2	**(1)**[a]	**(2)**[b]	**(3)**[c]
	⟨E⟩	11.95	12.01	12.01	12.01
300 K	$-TS^{o}$	3.09	-0.49	4.66	3.24
	μ^{o}	15.04	11.52	16.67	15.25
	⟨E⟩	14.34	14.32	14.32	14.32
500 K	$-TS^{o}$	2.12	-3.34	4.52	2.32
	μ^{o}	16.46	10.98	18.84	16.64
	⟨E⟩	20.29	19.95	19.95	19.95
1000 K	$-TS^{o}$	-3.88	-13.02	0.28	-3.44
	μ^{o}	16.41	6.93	20.23	16.51

[a] First-order approximation to the entropy

[b] Second-order approximation to the entropy

[c] Third-order approximation to the entropy

TABLE III

Thermodynamic quantities for methanol computed with ACCENT-MM using the “phase angle” pdfs as compared with M2. All values are in kcal/mol

			ACCENT-MM		
		M2	**(1)[a]**	**(2)[b]**	**(3)[c]**
	⟨E⟩	11.95	12.01	12.01	12.01
300 K	-TS^o	3.09	2.36	3.01	3.01
	μ^o	15.04	14.37	15.02	15.02
	⟨E⟩	14.34	14.32	14.32	14.32
500 K	-TS^o	2.12	1.17	2.05	2.06
	μ^o	16.46	15.49	16.37	16.38
	⟨E⟩	20.29	19.96	19.96	19.96
1000 K	-TS^o	-3.88	-5.16	-3.70	-3.70
	μ^o	16.41	14.80	16.26	16.26

[a] First-order approximation to the entropy

[b] Second-order approximation to the entropy

[c] Third-order approximation to the entropy

TABLE IV

Thermodynamic quantities for 1,2-dichloroethane computed with ACCENT-MM as compared with M2. All values are in kcal/mol

			ACCENT-MM		
		M2	**(1)[a]**	**(2)[b]**	**(3)[c]**
300 K	⟨E⟩	5.72	5.77	5.77	5.77
	-*TS*^{o}	6.33	5.58	6.30	6.31
	μ^{o}	12.05	11.35	12.07	12.08
500 K	⟨E⟩	9.51	9.62	9.62	9.62
	-*TS*^{o}	5.69	4.37	5.64	5.67
	μ^{o}	15.20	13.99	15.26	15.29
1000 K	⟨E⟩	18.68	18.77	18.77	18.77
	-*TS*^{o}	-1.36	-4.03	-1.41	-1.30
	μ^{o}	17.32	14.74	17.36	17.47

[a] First-order approximation to the entropy

[b] Second-order approximation to the entropy

[c] Third-order approximation to the entropy

TABLE V

Thermodynamic quantities for various alkanes computed at 1000 K with ACCENT-MM as compared with M2. All values are in kcal/mol

			ACCENT-MM		
		M2	**(1)**[a]	**(2)**[b]	**(3)**[c]
Butane	⟨E⟩	38.90	38.92	38.92	38.92
	$-TS^o$	24.09	19.13	23.62	24.51
	μ^o	62.99	58.05	62.54	63.43
Pentane	⟨E⟩	48.60	48.33	48.33	48.33
	$-TS^o$	33.69	27.59	34.41	40.14
	μ^o	82.29	75.92	82.74	132.37
Hexane	⟨E⟩	54.95	54.71	54.71	54.71
	$-TS^o$	43.67	36.34	45.08	56.30
	μ^o	98.62	91.05	99.79	111.01
Heptane	⟨E⟩	67.84	67.35	67.35	67.35
	$-TS^o$	53.89	44.65	55.41	74.21
	μ^o	121.73	112.00	122.76	141.56
Octane	⟨E⟩	77.37	76.89	76.89	76.89
	$-TS^o$	64.22	53.11	65.98	94.62
	μ^o	141.59	130.00	142.87	171.51
Nonane	⟨E⟩	86.85	86.11	86.11	86.11
	$-TS^o$	73.08	61.55	73.64	88.53
	μ^o	159.93	147.66	159.75	174.64
Cyclohexane	⟨E⟩	61.99	61.86	61.86	61.86
	$-TS^o$	61.38	34.74	51.17	57.93
	μ^o	123.37	96.60	113.03	119.79

[a] First-order approximation to the entropy

[b] Second-order approximation to the entropy

[c] Third-order approximation to the entropy

TABLE VI

Thermodynamic quantities computed for the complexation reaction at 300 K using ACCENT-MM as compared with M2. All values are in kcal/mol

			ACCENT-MM		
		M2	**(1)[a]**	**(2)[b]**	**(3)[c]**
	$\langle E \rangle$	-9.64	-9.64	-9.64	-9.64
Ligand	$-TS^o$	19.00	16.40	18.98	18.92
	μ^o	9.36	6.76	9.34	9.28
	$\langle E \rangle$	16.15	15.57	15.57	15.57
Receptor	$-TS^o$	105.12	95.45	105.87	166.28
	μ^o	121.27	111.02	121.44	181.85
	$\langle E \rangle$	-9.96	-9.33	-9.33	-9.33
Complex	$-TS^o$	135.25	117.73	133.61	247.43
	μ^o	125.29	108.40	124.28	238.10
	$\langle E \rangle$	-16.47	-15.26	-15.26	-15.26
Net	$-T\Delta S^o$	11.13	5.88	8.76	62.23
	ΔG^o	-5.34	-9.38	-6.50	46.97

[a] First-order approximation to the entropy

[b] Second-order approximation to the entropy

[c] Third-order approximation to the entropy