Sampling conformations in high dimensions using low-dimensional distribution functions

Sandeep Somani, Benjamin J. Killian, and Michael K. Gilson[a)]
Center for Advanced Research in Biotechnology, UMBI, 9600 Gudelsky Drive, Rockville, Maryland 20850, USA

(Received 5 September 2008; accepted 4 February 2009; published online 2 April 2009)

We present an approximation to a molecule's N-dimensional conformational probability density function (pdf) in terms of marginal pdfs of highest order l, where l is much less than N. The approximation is constructed as a product of conditional pdfs derived by recursive application of the generalized Kirkwood superposition approximation. Furthermore, an algorithm is presented to sample conformations from the approximate full-dimensional pdf based upon all input marginal pdfs. The sampling algorithm is tested for three small molecule systems by using the algorithm to sample conformations at levels $l=1$, 2, or 3 and comparing the distributions of sampled conformations with those from the molecular dynamics (MD) simulations. The distributions of conformations sampled at third ($l=3$) order resemble the MD distributions rather well and significantly better than those sampled at second ($l=2$) or first ($l=1$) order. In addition to highlighting the importance of correlations among internal degrees of freedom, these results suggest that low-order correlations suffice to describe most of the conformational fluctuations of molecules in a thermal environment. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3088434]

I. INTRODUCTION

Molecules are flexible structures in a thermal environment, so their conformations fluctuate continuously. Characterization of molecular fluctuations is critical to understanding and ultimately controlling molecular properties and functions, such as binding and catalysis. Fluctuations also are of central importance in statistical thermodynamics, which links molecular motions to thermodynamic observables such as free energy, entropy, and enthalpy. A molecule's configurational entropy, in particular, depends upon not only the magnitude of its conformational fluctuations along each degree of freedom but also the degree to which these fluctuations are mutually correlated.[1] As a consequence, calculations of configurational entropy can shed light on correlations.

The relationship between correlation and entropy is especially clear in the mutual information expansion (MIE), a series expansion of the Gibbs entropy in terms of mutual information sums of increasing order.[1–3] The mutual information at a given order goes to zero if there are no correlations at that order, and truncating the MIE series at a given order yields a value of the entropy under the approximation that higher-order correlations are absent. One may therefore assess the order of correlations in a given system by comparing the entropy from corresponding truncations of the MIE with the entropy from a second computational method that accounts for all physically relevant correlations. The mining minima (M2) method[4] can serve as this second method for molecular systems of modest size (e.g., small molecules and host-guest systems). M2 obtains the Helmholtz free energy and average energy of a molecule by computing its configurational integral as a sum over local energy minima on an energy surface defined by an empirical force field and an implicit solvent model.[4] The configurational entropy may then be computed as essentially the difference between the free energy and the mean energy. Since the M2 method uses the full configurational integral (subject to the approximation of summing over local energy wells) it effectively accounts for all physically relevant correlations at a given temperature.

It has been observed[1] that molecular entropies estimated with MIE at the doublet and triplet levels—i.e., neglecting correlations above second and third orders, respectively—can agree well with independent M2 calculations for small molecules. Because M2 includes essentially all correlations, this observation suggests that most of the physically relevant fluctuations of these molecules involve only low-order correlations. Thus, it leads to the hypothesis that conformational probability distributions can be well described with a tractable set of low-order distribution functions. The present study describes a novel approach to test this hypothesis and provides results for molecular test systems.

II. THEORY

The conformational fluctuations of a molecule with M atoms at equilibrium are fully described by a Boltzmann probability distribution function (pdf) over its $N=3M-6$ internal degrees of freedom:

$$p_N(X_1, \ldots, X_N) = \frac{\exp(-\beta E(X_1, \ldots, X_N))}{\int \exp(-\beta E(X_1, \ldots, X_N)) dX_1 \cdots dX_N}, \quad (1)$$

a)Author to whom correspondence should be addressed. Electronic mail: gilson@umbi.umd.edu. Tel.: 240-314-6217. FAX: 240-314-6255.

where $\beta=(kT)^{-1}$, k being the Boltzmann constant and T the temperature, and $E(X_1,\ldots,X_N)$ is the energy as a function of internal coordinates $\vec{X}=(X_1,\ldots,X_N)$. For a molecule in solution, the energy function comprises the molecule's internal energy and an additional contribution from the solvent.[5] The pdf, Eq. (1), can account for correlations of all orders up to N, but its high dimensionality makes it computationally unworkable for most molecules of interest.

The present study poses the following question: how accurately can the conformational fluctuations of a molecule in a thermal environment be described without accounting for high-order correlations? This is addressed by using a molecule's low-order marginal pdfs, populated with well-converged simulation data, as a basis for sampling full molecular conformations. The resulting conformational distributions, which, by construction, account only for low-order correlations, are then evaluated by comparing them with those obtained from the simulations. The sampling is done with a novel sampling approximation in which the ancestral sampling algorithm[6] is combined with a further generalization[1] of the generalized Kirkwood superposition approximation[3,7] (GKSA) that allows the required high-order conditional pdfs to be estimated in terms of the available low-order marginals.

The following sections begin by introducing a notation enabling superposition approximations to be expressed compactly. The l-level superposition approximation (SA-l) is then described, followed by derivation of SA-l based conditional distributions in terms of given marginal pdfs. Finally, algorithms to sample conformations using the conditional distributions are presented.

A. Notation

Let $\vec{X}=(X_1,\ldots,X_N)$ be N random variables representing, for example, the N internal coordinates of a molecule. Specific numerical values of the random variables are denoted by lowercase letters, so that $\vec{x}=(x_1,\ldots,x_N)$ represents a particular conformation of the N-dimensional system. To make subsequent equations more readable, we introduce a compact notation for the pdf over a subset of k coordinates $(X_{i_1},\ldots,X_{i_k})$, with $i_1,\ldots,i_k\in[1,N]$, in which the variables are denoted by just their index i_k:

$$[i_1,\ldots,i_k]\equiv p_k(X_{i_1},\ldots,X_{i_k}). \tag{2}$$

The order of the pdf is the total number of variable labels in the square bracket. For example, $[3]\equiv p_1(X_3)$ denotes the one-dimensional (1D) marginal pdf, or singlet distribution, in X_3; $[3,5]\equiv p_2(X_3,X_5)$ denotes the two-dimensional (2D) marginal pdf, or doublet distribution, in X_3 and X_5; and $[3,5,6]\equiv p_3(X_3,X_5,X_6)$ denotes the three-dimensional (3D) marginal pdf, or triplet distribution, in variables X_3, X_5, and X_6. Moreover, a specific value of the probability density $p_k(x_{i_1},\ldots,x_{i_k})$ at $X_{i_1}=x_{i_1},X_{i_2}=x_{i_2},\ldots$ is denoted by $[x_{i_1},\ldots,x_{i_k}]$, i.e.,

$$[x_{i_1},\ldots,x_{i_k}]\equiv p_x(X_{i_1}=x_{i_1},\ldots,X_{i_k}=x_{i_k}). \tag{3}$$

Specifying the values of a subset of variables effectively lowers the order of a distribution. For example, if variable X_5 has a value of x_5, the triplet distribution $p_3(X_3,X_5,X_6)\equiv[3,5,6]$ becomes a 2D distribution $p_2(X_3,X_5=x_5,X_6)$, denoted by $[3,x_5,6]$.

The conditional distribution of variable X_l, given the condition that k other variables $X_{i_i},\ldots,X_{i_k}$ have values $x_{i_1},\ldots,x_{i_k}$, is denoted by

$$[l|x_{i_1},\ldots,x_{i_k}]\equiv p_1(X_l|x_{i_1},\ldots,x_{i_k}), \tag{4}$$

where $l,i_1,\ldots,i_k\in[1,N]$ and $l\neq i_1\neq\cdots\neq i_k$. Note that $[l|x_{i_1},\ldots,x_{i_k}]$ is a 1D distribution of variable X_l, and that the product rule[8] allows this distribution to be expressed as

$$[l|x_{i_1},\ldots,x_{i_k}]=\frac{[l,x_{i_1},\ldots,x_{i_k}]}{[x_{i_1},\ldots,x_{i_k}]}. \tag{5}$$

Since $x_{i_1},\ldots,x_{i_k}$ are specific values of the variables, the denominator of Eq. (5) is just a number while the numerator is a 1D pdf.

We will be presenting an approximation of a k-dimensional pdf in terms of products of its lower-order marginal pdfs. The first-order (singlet), second-order (doublet), third-order (triplet), and fourth-order (quadruplet) marginals of a k-order distribution are given by

$$\begin{aligned}
[l]&=\int[i_1,\ldots,i_k]\prod_{j\neq l}dX_j,\\
[l,m]&=\int[i_1,\ldots,i_k]\prod_{j\neq l\neq m}dX_j,\\
[l,m,n]&=\int[i_1,\ldots,i_k]\prod_{j\neq l\neq m\neq n}dX_j,\\
[l,m,n,o]&=\int[i_1,\ldots,i_k]\prod_{j\neq l\neq m\neq n\neq o}dX_j.
\end{aligned} \tag{6}$$

When the distributions are discretized, as in the present work, the integrals are replaced by summations. We furthermore denote the products of all unique singlet, doublet, triplet, and quadruplet distributions as

$$\begin{aligned}
S_k&\equiv\prod_{1\le i_1\le k}[i_1],\\
D_k&\equiv\prod_{1\le i_1<i_2\le k}[i_1,i_2],\\
T_k&\equiv\prod_{1\le i_1<i_2<i_3\le k}[i_1,i_2,i_3],\\
Q_k&\equiv\prod_{1\le i_1<i_2<i_3<i_4\le k}[i_1,i_2,i_3,i_4].
\end{aligned} \tag{7}$$

For N random variables, there are C_1^N, C_2^N, C_3^N, and C_4^N singlet, doublet, triplet, and quadruplet marginal distributions, respectively.

B. Superposition approximations

A superposition approximation approximates a higher-order pdf in terms of lower-order distributions. The original Kirkwood superposition approximation (KSA),[9] which approximates a triplet correlation function in terms of singlet and doublet distribution functions,

$$[1,2,3] \approx \frac{[1,2][1,3][2,3]}{[1][2][3]} = (S_3)^{-1}(D_3)^{+1}, \tag{8}$$

was introduced in the distribution function theory of liquids as a closure equation to truncate the Bogoliubov-Born-Green-Kirkwood-Yvon (BBGYK) (Ref. 10) hierarchy of distribution functions and thereby enable calculation of the pair correlation function, which is equivalent to the doublet distribution here. The KSA was derived by approximating the three-particle potential of mean force as the sum of three two-particle potentials of mean force, one for each pair of the three particles. The Fisher–Kopelovich superposition approximation[11] is an extension of the KSA, also introduced in theory of liquids, which approximates a four-particle distribution function in terms of three-particle marginals:

$$[1,2,3,4] \approx \frac{[1,2,3][1,2,4][1,3,4][2,3,4]}{\underline{[1,2][1,3][1,4][2,3][2,4][3,4]}} \\ [1][2][3][4]$$

$$= (S_4)^{+1}(D_4)^{-1}(T_4)^{+1}. \tag{9}$$

The GKSA,[3,7] introduced as a probabilistic closure equation, extends the KSA to a general order, approximating a k-order distribution function in terms of all lower-order (up to $k-1$) marginals:

$$[1,\ldots,k] \approx \prod_{n=1}^{k-1} \left[\prod_{1\le i_1<i_2\cdots<i_n\le k} [i_1,\ldots,i_n] \right]^{(-1)^{k-1-n}}. \tag{10}$$

Note that the KSA and its subsequent generalizations are not normalized pdfs.[10]

By recursive application of the GKSA, the k-order pdf can be approximated in terms of marginal pdfs of order l and lower to yield an l-level superposition approximation.[1] For example, approximating the triplet distributions in the numerator of Eq. (9) by their KSAs according to Eq. (8) yields a second-level ($l=2$) approximation to a four-dimension ($k=4$) pdf:

$$[1,2,3,4] \approx \frac{[1,2][1,3][1,4][2,3][2,4][3,4]}{([1][2][3][4])^2}. \tag{11}$$

Furthermore, approximating the doublets in the numerator of Eq. (11) as products of singlets yields the first-level ($l=1$) approximation to the four-dimension ($k=4$) pdf:

$$[1,2,3,4] \approx [1][2][3][4]. \tag{12}$$

In general, approximations of a k-dimensional pdf at the singlet (SA-1), doublet (SA-2), and triplet (SA-3) levels are given by

$$[1,\ldots,k] \approx S_k, \quad \text{SA-1}, \tag{13}$$

$$[1,\ldots,k] \approx (S_k)^{-(k-2)}(D_k)^{+1}, \quad \text{SA-2}, \tag{14}$$

and

$$[1,\ldots,k] \approx (S_k)^{+(k-3)(k-2)/2}(D_k)^{-(k-3)}(T_k)^{+1}, \quad \text{SA-3}, \tag{15}$$

where the symbols on the right-hand side are defined in Eq. (7). The level of approximation l may be much less than the dimensionality k of the full distribution. In this work, singlet ($l=1$), doublet ($l=2$), and triplet ($l=3$) level superposition approximations are used. Clearly a mixed level of approximation could also be written; for example, one could approximate only a subset of the triplets in Eq. (9) by their KSAs. As noted above, the SA-l pdfs are not normalized at any level except for the trivial singlet-level approximation.

C. Superposition approximations to conditional distributions

The SA-l approximation to the k-dimensional joint pdf can be used to derive approximate conditional distributions of a variable given values of all other variables. These approximate conditionals are obtained simply by substituting the SA-l approximations for the joint pdfs in Eq. (5). For example, using the SA-2 expression for $[x_1,x_2,3]$ yields

$$[3|x_1,x_2] \approx \frac{[x_1,3][x_2,3]}{[x_1][x_2][3]} \frac{1}{N_3(x_1,x_2)} \tag{16}$$

and using the SA-2 expression for $[x_1,x_2,x_3,4]$ and $[x_1,x_2,x_3]$ gives

$$[4|x_1,x_2,x_3] \approx \frac{[x_1,4][x_2,4][x_3,4]}{[x_1][x_2][x_3][4]^2} \frac{1}{N_4(x_1,x_2,x_3)}, \tag{17}$$

where N_3 and N_4 denote the normalization factor as a function of the fixed variables. In general, the conditional distributions are not normalized due to the use of superposition approximations for the pdfs in Eq. (5). The normalization factors may be obtained by using the standard procedure of summing over all values of the conditioned variable as follows:

$$N_3(x_1,x_2) = \sum_{x_3} \frac{[x_1,x_3][x_2,x_3]}{[x_1][x_2][x_3]} \tag{18}$$

and

$$N_4(x_1,x_2,x_3) = \sum_{x_4} \frac{[x_1,x_4][x_2,x_4][x_3,x_4]}{[x_1][x_2][x_3][x_4]^2}. \tag{19}$$

The normalization factors can be computed efficiently as the summation is over values of a single variable.

More generally, the normalized conditional probability distribution of variable X_k ($k>2$), given values of the other $k-1$ variables, is approximated at the doublet level by

$$[k|x_1,\ldots,x_{k-1}] \approx \frac{(\Pi_{1\le i\le k-1}[x_i,k])}{([k])^{k-2}(\Pi_{1\le i\le k-1}[x_i])} \frac{1}{N_k(x_1,\ldots,x_{k-1})}. \tag{20}$$

Similarly, Eq. (5) may be approximated at the triplet level by applying the SA-3, yielding the following conditional distribution:

$$[k|x_1,\ldots,x_{k-1}] \approx \frac{(\Pi_{1\le i\le k-1}[x_i])^{(k-3)}([k])^{(k-3)(k-2)/2}(\Pi_{1\le i<j\le k-1}[x_i,x_j,k])}{(\Pi_{1\le i<j\le k-1}[x_i,x_j])(\Pi_{1\le i\le k-1}[x_i,k])^{(k-3)}} \frac{1}{N_k(x_1,\ldots,x_{k-1})}. \tag{21}$$

For example, the conditional distributions at the triplet level for $k=4$ and $k=5$ are, respectively,

$$[4|x_1,x_2,x_3] \approx \frac{[x_1][x_2][x_3][4][x_1,x_2,4][x_1,x_3,4][x_2,x_3,4]}{[x_1,x_2][x_1,x_3][x_2,x_3][x_1,4][x_2,4][x_3,4]} \frac{1}{N_4(x_1,x_2,x_3)} \tag{22}$$

and

$$[5|x_1,x_2,x_3,x_4] \approx \frac{([x_1][x_2][x_3][x_4])^2[5]^3[x_1,x_2,5][x_1,x_3,5][x_1,x_4,5][x_2,x_3,5][x_2,x_4,5][x_3,x_4,5]}{[x_1,x_2][x_1,x_3][x_1,x_4][x_2,x_3][x_2,x_4][x_3,x_4]([x_1,5][x_2,5][x_3,5][x_4,5])^2} \frac{1}{N_5(x_1,..,x_4)}. \tag{23}$$

D. Algorithms for sampling based on low-order marginals

We wish to sample molecular conformations while accounting for the correlations represented in marginal pdfs of order l and below. This is easily done at the singlet level ($l=1$), at which all variables are assumed to be independent: one simply samples each variable from its 1D marginal distribution without reference to the other variables. Accounting for correlations by including higher-order marginals is more complicated. We now describe an approach that combines the ancestral sampling algorithm[6] with superposition approximations of the required conditional pdfs.

1. Ancestral sampling algorithm

Any N-dimensional joint pdf can be factorized into a product of conditional distributions by repeated application of the product rule of probability theory:

$$[1,2,\ldots,k] = [N|1,2,\ldots,N-1][N-1|1,2,\ldots,N-2]\cdots[2|1][1]. \tag{24}$$

Equation (24) can be represented graphically by the directed, acyclic graph shown in Fig. 1, and the ancestral sampling algorithm[6] can be used to sample from any pdf once it is represented by a directed acyclic graph. The following pseudocode explains how the ancestral algorithm samples a conformation $\vec{x}=(x_1,\ldots,x_N)$ from the N-dimensional pdf $[1,2,\ldots,N]$ expressed by the graph in Fig. 1:

Algorithm 1: Ancestral sampling algorithm.

$x_1 \sim [1]$,

$x_2 \sim [2|x_1]$,

$x_3 \sim [3|x_1,x_2]$,

FOR $k=4$ to N

$x_k \sim [k|x_1,\ldots,x_{k-1}]$.

ENDFOR

where "$\sim$" means "sampled from." The first variable is sampled from its 1D singlet distribution, and each subsequent variable is sampled from its 1D distribution conditioned upon the values of all the variables that have been sampled so far. Since the full N-dimensional pdf is known, all conditional distributions can be computed exactly and are normalized. The ancestral algorithm rigorously samples conformations from the full N-dimensional joint pdf. However, in many real-world applications, such as the present molecular studies, the required conditional distributions are not accessible because of the high dimensionality of the systems. This problem is addressed in the following subsection.

2. Ancestral sampling with the superposition approximation

The present study uses an approximation of the standard ancestral sampling method wherein the exact conditional pdfs are replaced with superposition approximations of the conditional pdfs. Thus a doublet-level sampling algorithm is obtained by using conditional pdfs computed using only singlet and doublet distributions [Eq. (20)]:

Algorithm 2: Doublet-level sampling algorithm.

$x_1 \sim [1]$,

$x_2 \sim [2|x_1] = \frac{[x_1,2]}{[x_1]}$,

$x_3 \sim [3|x_1,x_2] = \frac{[x_1,3][x_2,3]}{[x_1][x_2][3]} \frac{1}{N_3(x_1,x_2)}$,

FOR $k=4$ to N

$x_k \sim [k|x_1,\ldots,x_{k-1}] = \frac{\Pi_{1\le i\le k-1}[x_i,k]}{[k]^{k-2}\Pi_{1\le i\le k-1}[x_i]} \frac{1}{N_k(x_1,\ldots,x_{k-1})}$.

ENDFOR

The triplet-level sampling algorithm is obtained simply by approximating the conditional distributions with the triplet-level superposition approximation [Eq. (21)]:

Algorithm 3: Triplet-level sampling algorithm.

$$x_1 \sim [1],$$

$$x_2 \sim [2|x_1] = \frac{[x_1,2]}{[x_1]},$$

$$x_3 \sim [3|x_1,x_2] = \frac{[x_1,x_2,3]}{[x_1,x_2]},$$

$$x_4 \sim [4|x_1,x_2,x_3] = \frac{[x_1][x_2][x_3][4][x_1,x_2,4][x_1,x_3,4][x_2,x_3,4]}{[x_1,x_2][x_1,x_3][x_2,x_3][x_1,4][x_2,4][x_3,4]} \times \frac{1}{N_4(x_1,x_2,x_3)}.$$

FOR $k=5$ to N

$$x_k \sim [k|x_1,\ldots,x_{k-1}] = \frac{(\Pi_{1\le i\le k-1}[x_i])^{(k-3)}([k])^{(k-3)(k-2)/2}(\Pi_{1\le i<j\le k-1}[x_i,x_j,k])}{(\Pi_{1\le i<j\le k-1}[x_i,x_j])(\Pi_{1\le i\le k-1}[x_i,k])^{(k-3)}} \times \frac{1}{N_k(x_1,\ldots,x_{k-1})}.$$

ENDFOR

These algorithms may be iterated to generate any desired number of sampled conformations, where the separate iterations are, of course, independent of each other and hence amenable to parallelization. These two specific algorithms also generalize to higher levels of approximation—i.e., to larger values of l—and the standard ancestral algorithm is recovered when $l=N$. Indeed, the l-level sampling algorithm is identical to ancestral sampling for the first l variables, because the conditional distributions become approximate only at the $(l+1)$th variable in the graph.

These approximate sampling algorithms require as input all marginals of the full pdf up to the level of the approximation, l; these input marginals are termed *reference marginals*. Thus, the doublet-level algorithm requires access to all singlet- and doublet-level reference marginals, and the triplet-level algorithm requires the triplet-level reference marginals as well. Importantly, these algorithms do not require computation or storage of the full N-level pdf or its N-dimensional superposition approximation. For the molecular systems studied here, the reference marginal pdfs may be computed from molecular simulations. For example, the doublet distribution of two torsion angles can be computed and stored simply as their normalized 2D histogram. When the distributions are represented as normalized histograms, as done in this work, then a conformational sample refers to a vector of bin numbers, one for each variable.

The distribution of sampled conformations obtained from these algorithms, unlike the standard ancestral algorithm, depends on the order in which the variables are sampled. This can be seen by recognizing that the first l variables are sampled directly from their reference marginals, whereas subsequent variables are sampled from approximations to their original, reference distributions. Note, too,

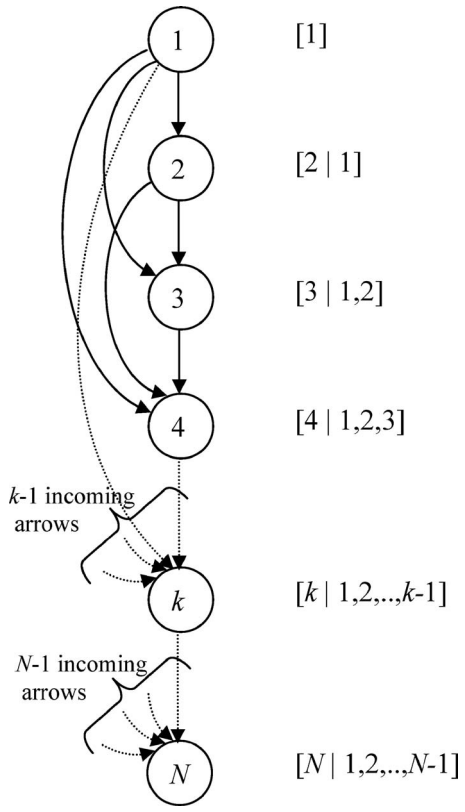

FIG. 1. Representation of an N-dimensional distribution function $[1,2,\ldots,N]$ as a directed graph. Variables are represented by circles containing the label of the respective variables. Solid arrows indicate the conditional dependencies of each variable as per the 1D conditional distribution indicated on the right of each node. The ancestral sampling uses exact conditional distributions for each variable. The l-level sampling algorithms presented in use approximations to conditional pdfs for $(l+1)$th and following variables. Doublet-level algorithm uses Eq. (20) while triplet-level algorithm uses Eq. (21). Dashed lines schematize elided portions of the graph.

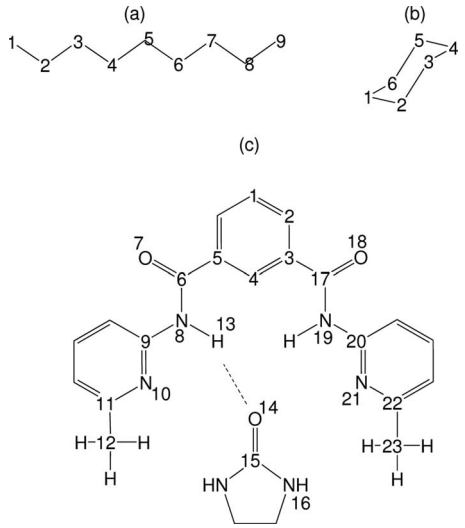

FIG. 2. Molecular systems used for testing the superposition approximation-based sampling algorithms: (a) nonane, (b) cyclohexane, and (c) host-guest complex (Ref. 13). Atoms included in the sampled structure are numbered. The internal coordinate system is set up such that atom 1 is on the origin, the bond between atoms 1 and 2 is along the x axis, and atom 3 is in the x-y plane. For nonane and cyclohexane, only the carbon chain is sampled, although the MD simulations included all hydrogens. For the host-guest complex, a dotted line represents the pseudobond used in defining the position and orientation of the guest relative to the host.

TABLE I. List of BAT degrees of freedom corresponding to the 23-atom skeleton of the host-guest complex. All 22 bonds, 21 angles, and 20 torsions are listed using the atom numbers shown in Fig. 2. Active variables are indicated by A. For the inactive variables, the equilibrium values based on the force field are listed.

Bond	Bond equilibrium value (Å)	Angle	Angle equilibrium value (rad)	Torsion	Torsion equilibrium value (rad)
1-2	1.383	1-2-3	2.0944	1-2-3-4	0
2-3	1.383	2-3-4	2.0944	1-2-3-17	A
3-4	1.383	2-3-17	A	2-3-4-5	0
3-17	1.46	3-4-5	2.0944	2-3-17-18	Phase of 2-3-17-19
4-5	1.383	3-17-19	A	2-3-17-19	A
5-6	1.46	3-17-18	A	3-17-19-20	A
6-8	1.345	4-5-6	A	3-4-5-6	A
6-7	1.225	5-6-8	A	4-5-6-7	Phase of 4-5-6-8
8-9	1.355	5-6-7	A	4-5-6-8	A
8-13	1.0	6-8-9	A	5-6-8-13	A
9-10	1.327	6-8-13	A	5-6-8-9	Phase of 5-6-8-13
10-11	1.327	8-9-10	A	6-8-9-10	A
11-12	1.5	8-13-14	A	6-8-13-14	A
13-14	A	9-10-11	2.0159	8-9-10-11	A
14-15	1.225	10-11-12	A	8-13-14-15	A
15-16	1.345	13-14-15	A	9-10-11-12	A
17-19	1.345	14-15-16	A	13-14-15-16	A
17-18	1.225	17-19-20	A	17-19-20-21	A
19-20	1.355	19-20-21	A	19-20-21-22	A
20-21	1.327	20-21-22	2.0159	20-21-22-23	A
21-22	1.327	21-22-23	A	⋯	⋯
22-23	1.5	⋯	⋯	⋯	⋯

that if the reference marginals have regions of zero probability, the present algorithms can fail to yield a complete set of variables for a conformation. This may be understood by considering the example of an SA-3 based conditional distribution:

$$[4|x_1,x_2,x_3] \approx \frac{[x_1][x_2][x_3][4][x_1,x_2,4][x_1,x_3,4][x_2,x_3,4]}{[x_1,x_2][x_1,x_3][x_2,x_3][x_1,4][x_2,4][x_3,4]} \times \frac{1}{N_4(x_1,x_2,x_3)}. \quad (25)$$

After the first three variables have been sampled, the factors containing the next variable, X_4, are essentially 1D histograms. When some bins contain values of zero, the product of two such histograms can have zero in all bins. This "null" condition makes it impossible to sample the fourth variable according to the algorithm. A related characteristic of the present approach is that the last variable is sampled from a conditional distribution constructed from the N-dimensional SA-l, which includes all reference marginal pdfs up to and including level l. Consequently, the populated regions of the marginal pdfs computed from the sampled conformations (termed the *sampled marginals*) will be contained within the populated regions of the corresponding reference marginal pdfs. Section IV D examines these issues by comparing sampled marginal pdfs with the corresponding reference marginal pdfs.

The probability of sampling each conformation with the present SA-l based sampling algorithms, $\tilde{p}_l$, is readily obtained during the sampling procedure by substituting the conditional pdf used to sample each variable into the chain rule given by Eq. (24). The sampling probability thus obtained is automatically normalized since all conditional pdfs used in the chain rule are normalized. Note that, due to the normalization of the approximate conditional distributions, the sampling probability of a conformation with the present algorithm is different from the N-dimensional SA-l probability of that conformation, $p_{\text{SA-}l}$. The Appendix provides an analytic expression for the probability of sampling a conformation in three dimensions using a doublet-level sampling algorithm, $\tilde{p}_2(x_1,x_2,x_3)$, and shows that it is normalized. The sampling distribution $\tilde{p}_2(x_1,x_2,x_3)$ also is compared with the second-level superposition approximation, $p_{\text{SA-2}}(x_1,x_2,x_3)$.

III. METHODS

A. Overview

We wish to determine whether the molecular conformations sampled with the algorithms described above are distributed similarly to those sampled by the molecular dynamics (MD) calculations from which the reference marginals used in the algorithms were obtained. The overall strategy is to

(1) run a MD simulation of a molecule;
(2) use the resulting trajectory to compute first-, second-, and third-order reference marginal pdfs—essentially normalized histograms—for internal coordinates of the molecule;
(3) use these reference marginals in the sampling algorithms to sample molecular conformations in the absence of explicit high-order correlations;

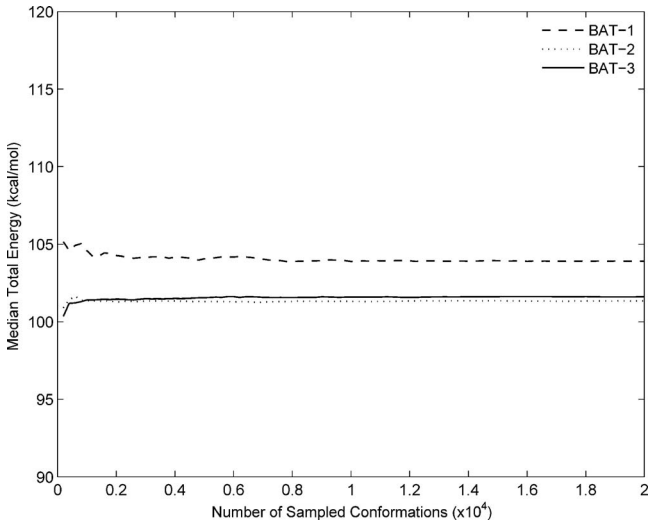

FIG. 3. Convergence of median total energy as a function of the number of conformations sampled in the BAT coordinate system at the singlet (BAT-1), doublet (BAT-2), and triplet (BAT-3) levels for the host-guest complex. The means and standard deviations of the energy converge similarly, and similar results are obtained for the other molecular systems in this study.

(4) reconstruct the 3D conformations of the molecule dictated by the sampled internal coordinates;
(5) compare the distributions of internal coordinates, conformations, and energies obtained by sampling to those from the original MD run.

Two different systems of internal coordinates are considered, bond-angle-torsion (BAT) and anchored Cartesian (XYZ).[12] For simplicity, distributions are not computed for some internal coordinates with very narrow distributions, such as dihedrals within a phenyl ring; instead, these coordinates are held fixed at their equilibrium values established by the force field. The coordinates which are sampled are termed "active."

B. Molecular systems

Three molecular systems were studied: linear nonane, cyclohexane, and a small host-guest complex.[13] Figure 2 diagrams the chemical structures of these molecules and the subsets of atoms used for testing the sampling algorithms. For each system, 5×10^6 MD snapshots spanning 50 ns of a single MD simulation were processed. The MD trajectories are those used in a previous study of configurational entropy[1] from our group. The MD simulations use an all-hydrogen energy model and approximate the effects of solvent with a simple distance-dependent dielectric model.

The analyses of nonane and cyclohexane were simplified by neglecting the coordinates of all-hydrogen atoms. This leaves 12 internal coordinates for the six carbons of cyclohexane (for BAT coordinates, these comprise 5 bond lengths, 4 bond angles, and 3 torsions) and 21 internal coordinates for the nine carbons of nonane (for BAT coordinates, these comprise 8 bond lengths, 7 bond angles, and 6 torsions). Conformations were sampled at the singlet level by independently sampling from the singlet marginal pdfs and at the doublet and triplet levels by using the algorithms described in Sec. II D 2 in both BAT and Cartesian coordinates.

The analysis of the host-guest system was simplified by limiting attention to a skeleton of 23 atoms (see Fig. 2) out of the simulated 56-atom complex. The retained atoms correspond to $63(3\times23-6)$ internal degrees of freedom. Conformations were sampled at the singlet, doublet, and triplet levels in the BAT coordinate system. Out of the 63 BAT coordinates associated with the 23 atoms, 32 coordinates were treated as active: all bond-angle and torsional degrees of freedom except the ones in the rings, along with one pseudobond, two pseudoangles, and three pseudodihedrals[4] that together specify the position and orientation of the guest with respect to the host. Table I lists the active degrees of freedom as well as the equilibrium values of the inactive ones. Three torsion angles (see Table I) that might be viewed as flexible are treated as phase angles[14] of flexible torsions that share the same rotatable bond; these phase angles have narrow distributions and are therefore treated as inactive.

C. Computing marginal and conditional distributions

The reference marginal pdfs required for the sampling algorithms were computed from MD trajectories by binning the active coordinates to create histograms and then normalizing the histograms. For example, the second-order marginal distribution of two torsion angles was obtained through a 2D histogram with square bins. For each variable, $N_B=30$ equally spaced bins were set up between the minimum and maximum values of the data to be binned, except for the torsional coordinates of cyclohexane. To effectively capture the bimodal distribution of these variables, 15 equally spaced bins were used in the two intervals of 0°–110.2° and 248.7°–360°, for a total of 30 bins.

TABLE II. Statistics of end-to-end distances (Å) in nonane from MD and sampled conformations. See text for symbols.

	Median	Mean	Standard deviation	Minimum	Maximum
MD	8.16	8.01	1.14	2.99	10.98
BAT-1	7.91	7.60	1.64	0.09	10.89
BAT-2	8.18	7.96	1.29	0.24	10.80
BAT-3	8.18	8.01	1.18	0.67	10.85
XYZ-1	7.93	7.76	2.28	0.31	14.84
XYZ-2	8.98	8.96	0.69	3.87	11.08
XYZ-3	7.23	7.24	1.66	3.05	10.97

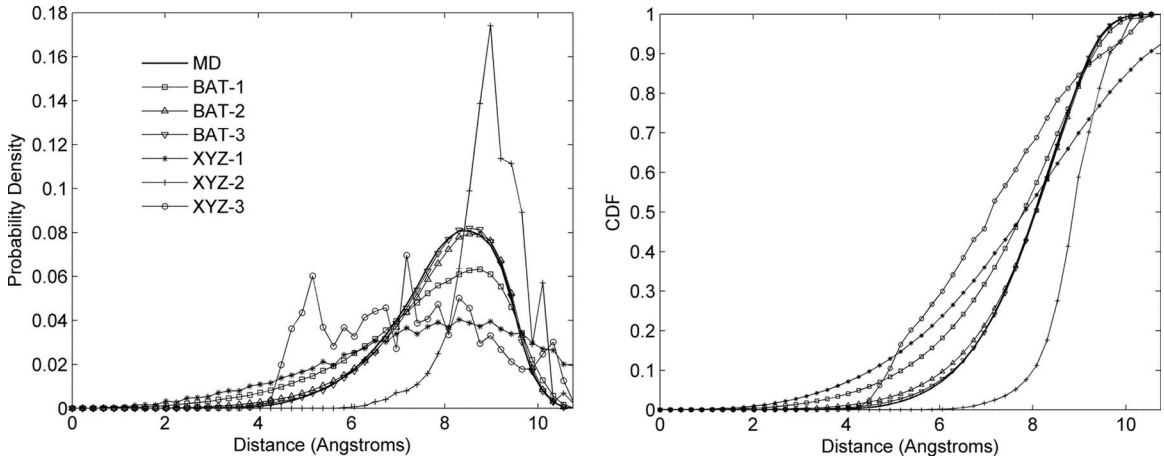

FIG. 4. Probability distributions (left) and corresponding cumulative distributions functions (right) of end-to-end distances for nonane from MD conformations and sampling algorithms at singlet, doublet, and triplet levels in BAT and XYZ coordinates.

For N active coordinates with N_B bins along each coordinate, the total numbers of elements in all matrices for the doublet (N_2) and triplet (N_3) distributions are, respectively,

$$N_2 = \frac{N(N-1)}{2} N_B^2,$$
$$N_3 = \frac{N(N-1)(N-2)}{6} N_B^3. \tag{26}$$

These numbers can become large. For example, the largest molecular system considered here has $N=32$, giving $N_2 = 446\,400$ and $N_3 = 133\,920\,000$, corresponding to about 1 Gbyte of memory at double precision (8 byte floats). However, we found that the storage requirements could be reduced about sixfold for all systems studied here by recognizing that many elements of the doublet and triplet distributions are zero and using sparse matrix storage techniques in the MATLAB® 7.5 (Ref. 15) programming environment.

The 1D conditional probability distribution constructed at each step of the SA-l based sampling algorithms consists of a product of distributions [Eqs. (20) and (21)], and the number of factors in this product becomes large as the number of variables increases. Direct multiplication of these factors can lead to underflow errors because the probabilities that are multiplied lie between 0 and 1. This problem was addressed by taking the logarithms of the factors and adding them instead of multiplying.

D. Evaluation of sampled conformations

In order to assess the contributions of successively higher-order correlations, distributions of conformations generated via sampling at the singlet, doublet, and triplet levels were compared to the distributions obtained directly from MD. Three types of comparisons were done.

First, the coordinates generated by the sampling algorithms were binned just as done for the MD snapshots, and the resulting 1D, 2D, and 3D sampled marginal pdfs were compared with the reference marginal pdfs obtained directly from MD. The difference between a sampled distribution and the corresponding reference distribution from MD is reported as the root mean square deviation (RMSD) across the bins of the stored distributions.

Second, the sampled distributions were compared with the reference MD distributions by comparing the distributions of energies and key intramolecular distances. Doing this requires reconstructing the molecular conformation as-

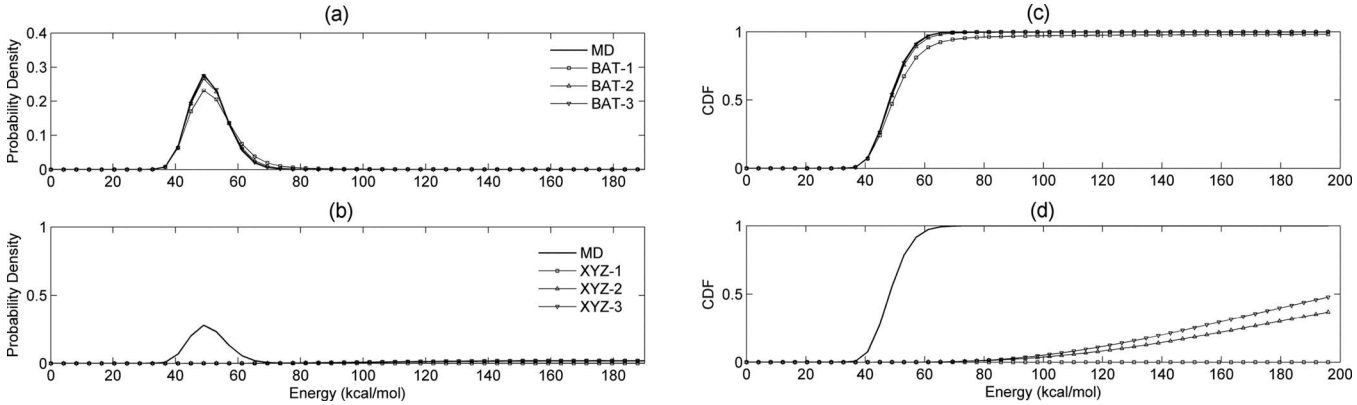

FIG. 5. Probability distributions and corresponding cumulative distributions functions (right) of total energy of nonane based upon MD conformations and conformations sampled in [(a) and (c)] BAT and [(b) and (d)] XYZ coordinate systems at singlet, doublet, and triplet levels.

TABLE III. Statistics of energy distributions (kcal/mol) from MD and sampled conformations for nonane.

	Bond	Angle	Torsion	Coulomb	vdW	Total
			Median			
MD	7.7	4.4	3.23	34.5	−0.63	50.2
BAT-1	7.8	4.5	3.21	34.9	−0.57	51.5
BAT-2	7.8	4.5	3.22	34.4	−0.62	50.6
BAT-3	7.8	4.5	3.22	34.4	−0.63	50.4
XYZ-1	1.86×10^{4}	318.4	4.91	36.0	15.33	2.2×10^{4}
XYZ-2	141.2	48.6	2.78	31.7	−0.52	232.3
XYZ-3	114.4	41.2	3.85	34.7	−0.61	202.8
			Mean			
MD	8.34	4.9	3.22	34.8	−0.6	50.7
BAT-1	8.49	5.0	3.23	36.3	3.6×10^{3}	3.6×10^{3}
BAT-2	8.48	5.0	3.23	35.0	182.6	234.3
BAT-3	8.50	5.0	3.22	34.8	11.5	63.0
XYZ-1	2.1×10^{4}	325.8	4.90	37.7	9.7×10^{4}	1.2×10^{5}
XYZ-2	171.93	55.6	2.86	32.1	179.7	442.2
XYZ-3	136.75	47.7	3.84	35.1	252.2	475.6
			Standard deviation			
MD	4.15	2.6	1.00	2.46	0.3	5.9
BAT-1	4.22	2.7	1.11	5.59	1.24×10^{5}	1.24×10^{5}
BAT-2	4.22	2.7	1.03	3.20	2.67×10^{4}	2.67×10^{4}
BAT-3	4.26	2.7	1.00	2.66	4.62×10^{3}	4.62×10^{3}
XYZ-1	1.21×10^{4}	120.0	1.14	7.94	6.51×10^{5}	6.50×10^{5}
XYZ-2	120.9	36.2	0.92	1.85	2.41×10^{4}	2.41×10^{4}
XYZ-3	93.4	31.9	1.10	3.22	3.33×10^{4}	3.33×10^{4}
			Minimum			
MD	0.04	0.02	0.01	28.1	−1.49	31.7
BAT-1	0.16	0.02	0.00	28.2	−1.57	32.2
BAT-2	0.17	0.02	0.00	28.4	−1.51	32.0
BAT-3	0.31	0.10	0.00	28.2	−1.53	32.8
XYZ-1	1.92×10^{2}	3.69	0.48	22.6	−1.35	307.1
XYZ-2	0.88	0.47	0.06	27.6	−1.01	43.1
XYZ-3	2.17	0.23	0.00	27.6	−1.42	44.0
			Maximum			
MD	45.6	34.5	7.64	49.0	5.06	100.2
BAT-1	49.1	28.1	7.88	244.4	9.98×10^{6}	9.99×10^{6}
BAT-2	45.7	27.4	7.87	130.7	9.00×10^{6}	9.00×10^{6}
BAT-3	42.0	23.7	7.22	80.5	2.54×10^{6}	2.54×10^{6}
XYZ-1	1.16×10^{5}	955.5	9.56	352.4	9.99×10^{6}	1.00×10^{7}
XYZ-2	1.55×10^{3}	452.4	7.44	84.7	9.96×10^{6}	9.97×10^{6}
XYZ-3	1.27×10^{3}	386.5	8.52	80.6	8.95×10^{6}	8.95×10^{6}

sociated with each set of internal coordinates generated by the sampling algorithm. Those internal coordinates not included in the sampling (Sec. III B), such as bond stretches in the case of the host-guest complex, were set to their equilibrium values established by the force field. Because not all internal coordinates were allowed to vary in the sampling, it was necessary to modify the MD snapshots, in which all internal coordinates fluctuate, so that they would be on a comparable footing. This was done by extracting active internal coordinates from the MD snapshots and substituting the equilibrium values for the inactive coordinates, precisely as done for the sampled conformations. Then the MD conformations were reconstructed with these idealized coordinates and compared with the sampled conformations. For the host-guest complex, distributions were compared for multiple interatomic distances that characterize the conformations. For nonane, the distance between the terminal carbons was examined. For cyclohexane, the distance between carbons 1 and 6 was examined; this distance is not part of the BAT coordinate system and depends upon the values of the internal coordinates in the same way that nonane's end-to-end distance depends upon its internal coordinates. Note that, in the XYZ coordinate system, the end-to-end distances for both nonane and cyclohexane are, in principle, functions of only the Cartesian coordinates of the last carbon. Nonetheless, since these coordinates are the last three variables to be

TABLE IV. Fraction of sampled conformations with energies greater than the maximum energy obtained in MD simulation of each test system.

	Number of conformations sampled	Number of high-energy conformations	Fraction of high-energy conformations
		Nonane	
BAT-1	500 000	14 500	2.9×10^{-2}
BAT-2	500 000	1 450	2.9×10^{-3}
BAT-3	500 000	170	3.4×10^{-4}
XYZ-1	500 000	500 000	1.00
XYZ-2	500 000	480 000	0.96
XYZ-3	500 000	475 000	0.95
		Cyclohexane	
BAT-1	500 000	401 539	0.80
BAT-2	500 000	192 204	0.38
BAT-3	500 000	48 680	0.09
XYZ-1	500 000	488 561	0.98
XYZ-2	500 000	112 324	0.22
XYZ-3	500 000	12 392	0.02
		Host-guest complex	
BAT-1	200 000	22 000	0.11
BAT-2	200 000	2 000	0.01
BAT-3	200 000	1 000	0.005

sampled, the distributions of end-to-end distances in the sampled conformations depend on all sampled XYZ coordinates.

Third, the distributions of energy for the sampled conformations, computed with the same CHARMm[16] force field model, were compared with those for the reconstructed (see above) MD conformations. Comparisons were made for both the total molecular energy and the separate terms provided by the force field. The separate terms provide additional physical insight; for example, if the sampled conformations were to yield more conformations with high Lennard-Jones energies, this would imply steric clashes.

IV. RESULTS

As detailed above, we (1) compute marginal probability distributions for the internal coordinates of nonane, cyclohexane, and a host-guest complex based upon MD simulations; (2) use these marginals to construct superposition approximations to conditional pdfs for the internal coordinates; (3) sample conformations with algorithms based upon these conditional distributions; and (4) compare the sampled conformations with the original MD conformations. For nonane and cyclohexane, 500 000 conformations were sampled, while for the host-guest complex, 200 000 conformations were sampled. Convergence was established based on the medians of energy and the sampled interatomic distances; a representative convergence plot of median total energy for the host-guest system is shown in Fig. 3.

The first three subsections here assess the accuracy of conformations sampled at the singlet, doublet, and triplet levels by studying the geometries and energies of the sampled conformations from the superposition approximations. The fourth subsection compares marginal distributions

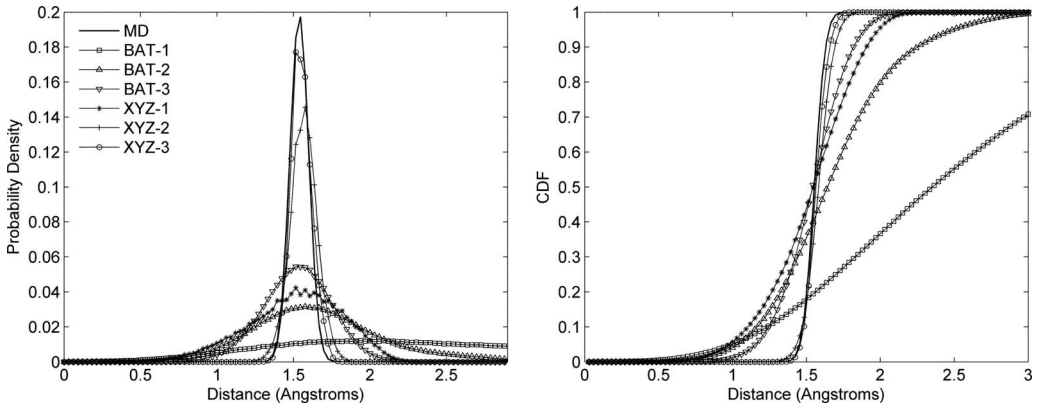

FIG. 6. Probability distributions (left) and corresponding cumulative distribution functions (right) of cyclohexane end-to-end distances computed from MD conformations and sampled conformations.

TABLE V. Statistics of end-to-end distances (Å) from MD and sampled conformations for cyclohexane.

	Median	Mean	Standard deviation	Minimum	Maximum
MD	1.54	1.54	0.06	1.24	1.83
BAT-1	2.34	2.45	1.00	0.03	5.82
BAT-2	1.63	1.66	0.43	0.09	3.74
BAT-3	1.53	1.52	0.22	0.34	2.41
XYZ-1	1.53	1.51	0.31	0.25	2.40
XYZ-2	1.57	1.57	0.09	1.08	1.93
XYZ-3	1.55	1.55	0.07	1.27	1.84

of the superposition approximations with the corresponding marginals of the original MD samples. Finally, the computational complexity of the sampling algorithms is analyzed.

Here BAT-1, BAT-2, and BAT-3 refer to conformational sampling in BAT coordinates at the singlet, doublet, and triplet levels, respectively, while XYZ-1, XYZ-2, and XYZ-3 refer to sampling in Cartesian coordinates at the corresponding levels of approximation. As noted in Sec. II D, the distributions provided by the present algorithms depend to some degree upon the sequence in which the variables are sampled. Changing the sequence of sampling, such as by sampling bond lengths first versus sampling torsions first, when BAT coordinates are used, was found to alter the results in detail but had little effect on the overall accuracy of the sampled distributions. The sampling sequence in XYZ coordinates follows the atom numbering of Fig. 2 where, for each atom, the x coordinate is sampled first, followed by the y and z coordinates. In BAT coordinates, for nonane and cyclohexane the bond coordinates of all atoms in their indexed order were sampled first, followed by angles and torsions in the same order. Sampling for the host-guest system in BAT coordinates also followed the sequence of bonds, angles, and torsions, according to the order in Table I.

As discussed in Sec. II D 2, the present algorithms can fail to yield a complete set of variables for a conformation. For the molecular systems studied here, such null samples never occurred for the doublet-level algorithm. At the triplet level, they occurred for $<0.01\%$ of the iterations for nonane and the complex and never occurred for cyclohexane.

A. Nonane

First-, second-, and third-order marginal distributions were computed for nonane in anchored Cartesian and BAT coordinates and used to sample conformations at the singlet, doublet, and triplet levels in both coordinate systems. Thus, conformations are sampled based on a total of $2\times3=6$ sampling approximations. The probability distributions of end-to-end distances (carbon 1 to carbon 9) from MD and from the six sampled sets are compared numerically in Table II, and Fig. 4 graphs the corresponding distributions.

For BAT coordinates, both the doublet- and triplet-level superposition approximations provide excellent agreement with the MD results and are markedly more accurate than singlet-level sampling. In particular, Fig. 4 shows that doublet- and triplet-level samplings produce a substantially smaller fraction of conformations with excessively short end-to-end distances than does singlet-level sampling. This is also evident from the shorter minimum distance for singlet sampling (0.09 Å) as compared to doublet (0.24 Å) and triplet samplings (0.67 Å) (Table II). The triplet level is slightly more accurate than the doublet, but the difference is less striking than that between the doublet and singlet.

For Cartesian coordinates, all three sampled cases yield distance distributions that deviate markedly from the MD distributions: although the numerical statistics in Table II look reasonable, Fig. 4 shows that the shapes of the distributions are poor. As with sampling in BAT coordinates, the population in the short-distance end of the distribution is notably higher for the singlet than for the doublet or triplet samples. Interestingly, in Cartesian coordinates the triplet-level approximation does not appear to yield greater accuracy than the doublet-level distribution, at least by the present measure.

The BAT coordinate samples of nonane yield distributions of total energy that agree very well with MD overall at the doublet and triplet levels, as shown in Fig. 5(a). Interestingly, the total energy distribution at the singlet level is only slightly inferior to those at higher levels, indicating low correlations among various BAT internal coordinates. However, the energy distributions from superposition approximations in Cartesian coordinates disagree strongly with the reference MD distribution: they are wide and shifted to much higher energies, as shown in Fig. 5(b). The singlet-level results are particularly poor, as the minimum total energy among all conformations is 307.1 kcal/mol (see Table III), which is outside the range of energies in Fig. 5(b). The fraction of such high-energy conformations is lower for doublet- and triplet-level samples. It is not surprising that the energy distributions from Cartesian sampling are poor, given the similarly poor end-to-end distance distributions described above.

The distributions of the total energy of nonane for the BAT samples show excellent agreement with the MD distributions especially at the doublet and triplet levels [in Figs. 5(a) and 5(b)]. This is consistent with the good agreement between the median total energy of the sampled conformations with the MD conformations (Table III). The larger deviations for the mean energy and other statistics result from the small fraction of high-energy conformations among the sampled conformations, as evident from Fig. 5(b) and Table IV. Closer examination of Table III indicates that the bond-stretch and bond-angle energies are well behaved, but van der Waals energies are sometimes much too large. This indicates that the high-energy conformations among the BAT samples result from excessively close atom-atom contacts,

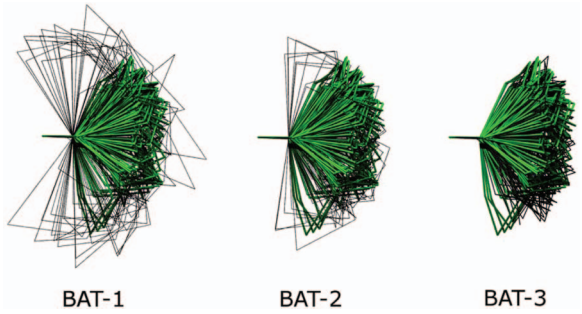

FIG. 7. (Color) Representative cyclohexane conformations from MD (in green) and from sampling (in black) in BAT coordinates at different levels. Each conformation is oriented such that atom 1 is at the origin, atom 2 is along the x axis, and atom 3 is in the x-y plane. Each set contains 100 conformations.

consistent with the small values of the minimum end-to-end distances listed in Table II. In contrast, the data in Table III show that the large errors in the energy distributions based upon XYZ coordinates result not only from van der Waals overlaps but also from severe errors in bond lengths and angles. The superior performance of the BAT coordinates, relative to Cartesians, is traceable to the fact that the BAT coordinates do an excellent job of capturing the marginal pdfs of those bond lengths and angles which are included in the coordinate set. Thus, the energetics of these stiff degrees of freedom are well reproduced.

B. Cyclohexane

As for nonane, we studied the singlet-, doublet-, and triplet-level samples in Anchored Cartesian and BAT coordinates for a total of six superposition approximations. Results are assessed geometrically in terms of the distribution of distances between carbon 1 and carbon 6, the only two successive atoms in the ring whose bond length is not part of the BAT coordinate system. As shown in Fig. 6 and further detailed in Table V, here the Cartesian coordinate system yields a more accurate distribution than does the BAT coordinate system, and the triplet-level distribution is more accurate than the doublet level, which in turn is substantially better than singlet-level distribution. The improvement upon including correlations is much more apparent in the distance distributions for samples in BAT coordinates. In BAT samples, the singlet-level distributions deviate drastically from the MD results, but the doublet- and triplet-level distributions are similar in structure to those from MD, being unimodal and centered at roughly the same bond length. Thus, for cyclohexane, higher-order correlations are required to accurately capture the geometry in both BAT and Cartesian coordinates. In absolute terms, the distribution of end-to-end distances from sampling in BAT coordinates for cyclohexane is more accurate than the distribution of end-to-end distances from sampling in Cartesian coordinates for nonane.

The MD trajectory used to compute the reference marginals includes chair, boat, and twist-boat conformations. In BAT coordinates, these conformations are established by the three internal torsions. Figure 7 compares representative conformations from MD and BAT-sampling at the three levels $l=1$, 2 and 3. The BAT-3 results resemble MD closely, and the BAT-2 results are similar, though somewhat less accurate. However, many of the BAT-1 conformations are quite distorted. The reason for the poor BAT-1 conformations has to do with the fact that singlet-level sampling completely ignores correlations, so that the effective 2D pdf linking each pair of torsions is just the product of their respective 1D pdfs. Figure 8 (left) shows that the correct 2D pdf of a pair of torsions has two maxima, corresponding to two different chair conformations. The corresponding singlet distributions (Fig. 8, middle) also have two maxima, so their product (Fig. 8, right) has four rather than two maxima. Two of the four maxima correspond to the correct maxima seen in the reference 2D pdf (Fig. 8, left); the other two are artifacts of singlet-level sampling and produce distorted conformations. Table VI summarizes the statistics of energy components computed from sampled conformations and reference MD conformations. On the whole, in both BAT and XYZ coordinate systems, triplet-level samples are closer to MD values than doublet-level samples. At the triplet level, the statistics for the bond energies from Cartesian sampling match the reference MD values better than those from BAT sampling. However, the statistics are similar for the other energy components (angle, torsion, and van der Waals). This is generally consistent with the observation that sampling in BAT coordinates leads to an inappropriately wide distribution of the atom 1–atom 6 bond length, as noted in the previous paragraph. Overall, the distribution of total energy shown in Fig. 9 indicates that triplet-level samples yield the most accurate conformational distributions, Cartesian coordinates being slightly better than BAT, and that the results become progressively worse on going to doublet and then single-level sampling. The cumulative distribution functions in Fig. 9, as well as the data in Table IV, further document marked reductions in the numbers of abnormally high-energy conformations as

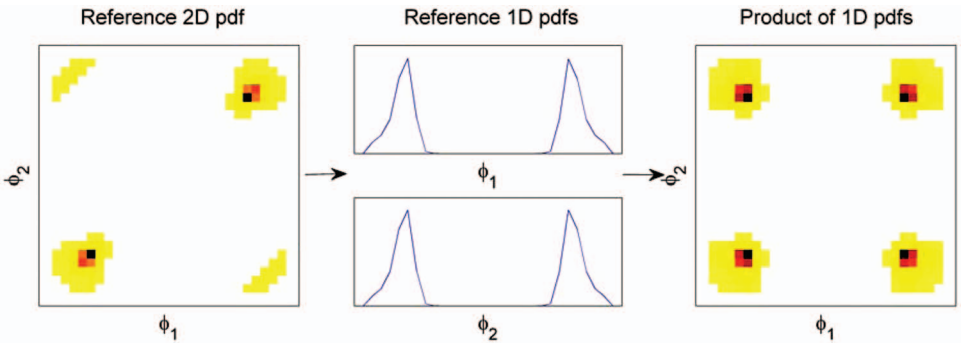

FIG. 8. (Color) The 2D pdf (right) obtained by multiplying 1D pdfs (middle) of cyclohexane has nonzero probability in regions not present in the true 2D pdf (left). Bins with zero probability in the 2D pdfs are colored white.

TABLE VI. Statistics of energy (kcal/mol) for cyclohexane computed from MD and sampled conformations. Coulombic term is not reported as it is always zero for this molecule.

	Bond	Angle	Torsion	vdW	Total
			Median		
MD	5.3	3.9	7.9	1.0	18.8
BAT-1	198.4	36.3	7.1	1.5	245.4
BAT-2	24.4	8.4	7.6	1.1	44.1
BAT-3	11.7	4.9	7.9	1.0	26.8
XYZ-1	288.8	56.8	6.9	1.8	363.4
XYZ-2	20.5	7.6	8.1	0.9	38.8
XYZ-3	10.0	5.0	7.8	1.1	25.2
			Mean		
MD	5.9	4.4	7.8	1.2	19.4
BAT-1	498.7	53.8	7.1	3.4	563.0
BAT-2	60.3	14.3	7.5	1.7	83.8
BAT-3	18.5	5.8	7.8	1.4	33.5
XYZ-1	448.8	67.6	7.0	8.1	531.5
XYZ-2	27.7	9.2	8.0	1.2	46.1
XYZ-3	12.7	6.1	7.8	1.3	27.9
			Standard deviation		
MD	3.4	2.6	0.6	0.9	4.5
BAT-1	669.9	50.4	0.8	8.4	713.0
BAT-2	95.7	15.9	0.8	2.1	107.9
BAT-3	19.9	4.1	0.7	1.2	21.5
XYZ-1	444.0	48.4	0.9	90.2	485.9
XYZ-2	24.2	6.5	0.7	0.9	27.0
XYZ-3	10.2	4.3	0.6	0.9	12.1
			Minimum		
MD	0.02	0.02	5.28	−0.21	7.87
BAT-1	0.16	0.09	4.71	−0.30	9.09
BAT-2	0.03	0.05	5.17	−0.29	8.27
BAT-3	0.01	0.04	5.32	−0.26	8.26
XYZ-1	0.32	0.15	3.99	−0.30	12.44
XYZ-2	0.08	0.07	5.37	−0.16	8.63
XYZ-3	0.08	0.04	5.66	−0.12	8.47
			Maximum		
MD	39.9	36.1	10.0	20.5	58.9
BAT-1	4.9×10^3	292.4	9.9	1.6×10^3	5.2×10^3
BAT-2	1.3×10^3	162.5	9.9	152.7	1.4×10^3
BAT-3	379.3	79.8	9.9	28.4	412.8
XYZ-1	4.3×10^3	353.3	10.0	2.8×10^4	2.8×10^4
XYZ-2	369.1	108.0	10.0	30.4	415.8
XYZ-3	180.1	76.0	10.0	18.1	245.2

more correlation is accounted for in both BAT and XYZ coordinates. Thus, the energy distributions, like the distance distributions, highlight the importance of including higher-order correlations for this constrained yet flexible chemical ring.

C. Host-guest complex

Conformations of the host-guest complex were sampled in BAT coordinates at the singlet, doublet, and triplet levels. Table VII and Fig. 10 analyze the distances between eight atom pairs: 1-11, 1-22, 12-15, 15-23, 11-22, 12-23, 8-19, and 1-15. [See Fig. 2(c) for atom numbers.] Overall, the distance distributions from doublet- and triplet-level samplings agree well with the reference MD distributions, the triplet level being somewhat more accurate than doublet. The singlet-level samples give notably poorer distributions, especially for distances between host and guest atoms, as shown in Figs. 10(c), 10(d), and 10(h).

For both BAT-2 and BAT-3 samples, the median values of all energy components are in excellent agreement with MD, as are the mean values of all energy components other than van der Waals (Table VIII). Although the tabulated statistics of BAT-1 samples are comparable to those of BAT-2 and BAT-3 samples, Fig. 11 shows that the distribution of total energies is substantially inferior with singlet-level sam-

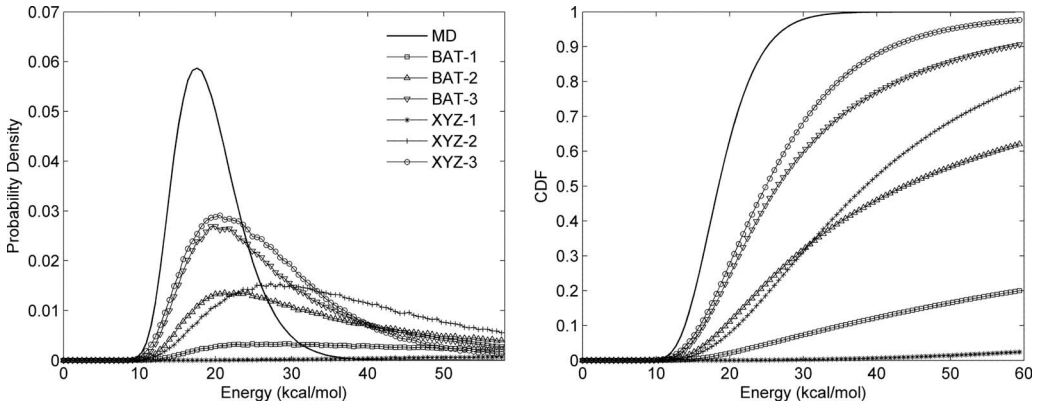

FIG. 9. Probability distributions (left) and corresponding cumulative distribution functions (right) of the total energy of cyclohexane computed from MD conformations and sampled conformations using different sampling schemes.

pling. It is also evident that the mean van der Waals energy of the BAT-3 samples is substantially better than that of BAT-1 and BAT-2 samples, though all are skewed toward higher values due to the presence of a few conformations with bad contacts. These lead to small tails of high-energy conformations for BAT-2 and BAT-3 samplings, as evident from the cumulative probability distributions of energy (Fig. 11). Table IV furthermore documents sharp reductions in the number of conformations with abnormally high energy as more correlation is accounted for in the sampling.

D. Comparing sampled and reference marginal pdfs

Another way to assess the accuracy of the sampled distributions is to compare their marginals with those of the original MD simulations. For singlet-level sampling, all 1D sampled marginal pdfs converge to the corresponding reference pdfs. For higher-level sampling, only the marginals of the first l variables converge to the reference marginals (data not shown), as expected based upon the structure of the sampling algorithm. The marginals of the subsequent ($>l$) vari-

TABLE VII. Statistics of seven key distances (Å) of host-guest conformations reconstructed from active MD and sampled BAT coordinates. Column headings give the atom pairs using atom numbers from Fig. 2(c).

	1-11	1-22	12-15	15-23	8-19	11-22	12-23	1-15
				Median				
MD	8.12	8.12	5.10	4.99	4.94	9.20	9.47	6.60
BAT-1	8.13	8.14	5.63	5.47	4.98	9.17	9.35	6.31
BAT-2	8.11	8.10	5.22	5.06	4.93	9.15	9.43	6.57
BAT-3	8.13	8.12	5.11	5.04	4.94	9.18	9.43	6.62
				Mean				
MD	8.10	8.09	5.14	5.03	4.94	9.17	9.43	6.54
BAT-1	8.11	8.12	5.61	5.71	4.99	9.18	9.43	6.18
BAT-2	8.09	8.08	5.24	5.17	4.93	9.15	9.43	6.50
BAT-3	8.11	8.09	5.12	5.09	4.94	9.17	9.42	6.58
				Standard deviation				
MD	0.11	0.12	0.43	0.45	0.13	0.46	0.79	0.45
BAT-1	0.11	0.10	1.00	1.34	0.19	0.67	1.20	0.86
BAT-2	0.11	0.13	0.64	0.88	0.14	0.52	0.88	0.54
BAT-3	0.11	0.12	0.52	0.68	0.13	0.48	0.83	0.43
				Minimum				
MD	7.06	7.07	3.42	3.12	4.27	6.31	4.95	3.22
BAT-1	7.00	7.26	1.37	1.66	4.21	6.43	5.23	1.66
BAT-2	7.19	7.16	2.20	2.13	4.30	6.58	5.22	2.20
BAT-3	7.04	7.28	2.79	2.66	4.36	7.07	6.06	3.69
				Maximum				
MD	8.37	8.38	10.34	10.33	5.92	12.17	14.50	7.92
BAT-1	8.35	8.35	9.60	12.53	5.96	12.29	14.79	10.11
BAT-2	8.37	8.36	9.00	11.58	5.68	11.45	13.56	8.41
BAT-3	8.36	8.37	8.17	10.22	5.51	11.10	12.64	8.09

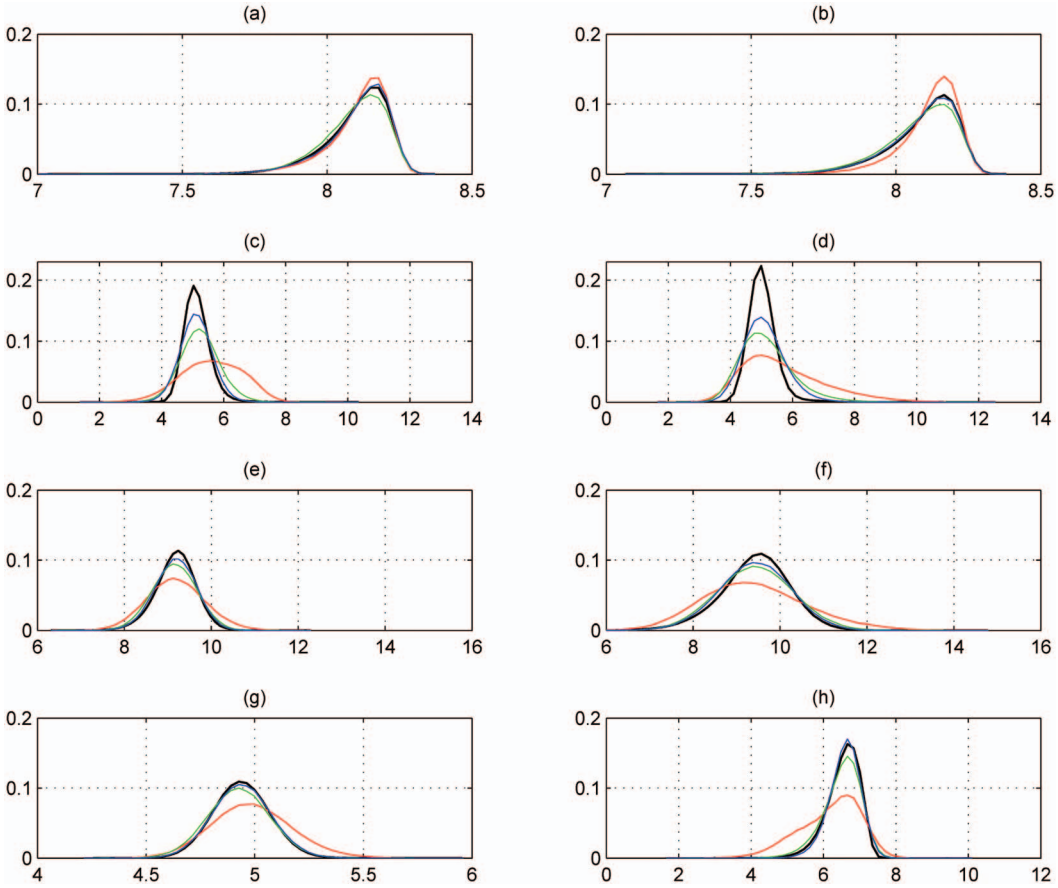

FIG. 10. (Color) Probability distributions of interatomic distances (Å) in the host-guest complex for atom pairs (a) 1-11, (b) 1-22, (c) 12-15, (d) 15-23, (e) 11-22, (f) 12-23, (g) 8-19, and (h) 1-15. Color code: MD in black, BAT-1 in red, BAT-2 in green, and BAT-3 in blue.

ables are sampled from approximate conditional distribution and therefore are expected to deviate from the reference marginals.

The deviations were quantified by computing the RMSDs between all 1D, 2D, and 3D sampled and reference marginal pdfs for each sampling case. Table IX reports the mean RMSD for the three marginal pdfs in each sampling case. As expected, for a specific molecule and coordinate system, the mean RMSD of singlet marginal pdfs is lower for singlet-level sampling than for higher-level sampling. However, the mean RMSDs of doublet and triplet marginal pdfs from doublet- or triplet-level samplings are, in general, lower than those from singlet-level sampling, indicating the presence of correlations similar to those in MD simulations. Sampling of nonane in XYZ coordinates does not follow this trend, indicating that triplet level is not sufficient to capture the correlations in this coordinate system. This observation is consistent with the analyses of distance and energy distributions above.

Marginal pdfs generated from sampled conformations are graphically compared with the corresponding reference marginals from the MD simulations in Figs. 12 and 13. Figure 12 shows the doublet marginal pdf from the XYZ study of nonane which yielded the highest RMSD values at the singlet, doublet, and triplet levels, and Fig. 13 similarly analyzes the doublet marginal which yielded the highest RMSD in BAT-1 sampling. It is worth noting that the doublet marginal pdf from BAT-1 sampling [Fig. 13(b)] is much closer to its reference MD marginal than is the XYZ-1 result for nonane [Fig. 12(b)], confirming earlier results where BAT coordinate performed better than XYZ.

E. Range of configurational space sampled

The part of the N-dimensional configuration space accessible to the present sampling algorithms is largely dependent on the reference marginal pdfs used to construct the superposition approximations in the algorithms. In one extreme case, if the reference distribution sampled only a single conformation, then all the marginals of the reference distribution would be delta functions, and only the single reference conformation would be sampled by the present algorithm. More generally, as discussed in Sec. II D 2, l-level sampling will only generate conformations that have nonzero probability in all the reference marginals. This property is apparent in Figs. 12 and 13, where the populated bins of a representative 2D marginal from doublet- and triplet-level samplings (right panel) are a subset of those sampled by the MD trajectories (left panel), unlike the marginal from singlet-level sampling. In this way, the reference marginals constrain the range of configuration space accessible to the sampling algorithm.

However, this constraint does not imply that the sampling algorithm generates only conformations identical to those used to compute the reference marginals. Rather, the omission of higher-level marginals from the superposition

TABLE VIII. Statistics of energy (kcal/mol) of host-guest complex for conformations reconstructed from active MD and sampled BAT coordinates.

	Bond	Angle	Torsion	Improper	Coulomb	vdW	Total
				Median			
MD	0.0014	7.16	3.87	0.12	65.1	24.7	101.5
BAT-1	0.0015	7.69	3.98	0.16	62.4	26.9	104.0
BAT-2	0.0015	7.20	4.14	0.17	63.6	25.1	101.4
BAT-3	0.0015	7.19	3.85	0.16	64.4	25.0	101.6
				Mean			
MD	0.0015	7.35	4.07	0.27	65.1	24.9	101.7
BAT-1	0.0015	7.94	4.21	0.27	64.0	5.41×10^{4}	5.41×10^{4}
BAT-2	0.0015	7.38	4.35	0.28	64.4	3.46×10^{3}	3.53×10^{3}
BAT-3	0.0015	7.36	4.05	0.25	65.1	959.7	1.04×10^{3}
				Standard deviation			
MD	0.0004	1.89	1.53	0.38	2.97	1.66	4.26
BAT-1	0.0004	2.19	1.65	0.39	11.3	4.82×10^{5}	4.82×10^{5}
BAT-2	0.0004	1.90	1.72	0.40	5.49	1.15×10^{5}	1.15×10^{5}
BAT-3	0.0004	1.87	1.50	0.34	4.75	6.20×10^{4}	6.20×10^{4}
				Minimum			
MD	0.0003	1.98	0.19	0.00	46.6	23.0	82.9
BAT-1	0.0003	2.37	0.48	0.00	−583.1	23.0	80.8
BAT-2	0.0003	2.11	0.49	0.00	−111.6	23.0	79.8
BAT-3	0.0004	2.29	0.61	0.00	40.6	23.2	84.8
				Maximum			
MD	0.0044	24.1	17.2	8.23	84.0	2.51×10^{3}	2.62×10^{3}
BAT-1	0.0042	23.1	16.4	7.48	1.22×10^{3}	9.99×10^{6}	9.99×10^{6}
BAT-2	0.0041	19.4	16.9	4.54	238.7	9.20×10^{6}	9.20×10^{6}
BAT-3	0.0040	19.0	12.5	2.97	262.9	8.97×10^{6}	8.97×10^{6}

approximations leaves the sampling algorithm free to generate new conformations. In fact, for all of the present test systems, over 99% of the $(2-5)\times10^{5}$ sampled conformations are new relative to the 5×10^{6} conformations in the original MD trajectory used to generate the reference marginals. (Conformations were compared after binning their coordinates.) This small degree of overlap is intuitively reasonable, given the large number of potential conformations, $\sim30^{N_D}$, where 30 is the number of bins used to discretize each coordinate and N_D is the dimensionality. Here N_D is 12, 21, and 32 for cyclohexane, nonane, and host-guest complex, respectively, so the number of conformations ranges from $\sim10^{17}$ to $\sim10^{47}$. To better judge whether the high fraction of novel conformations generated by the sampling algorithm is

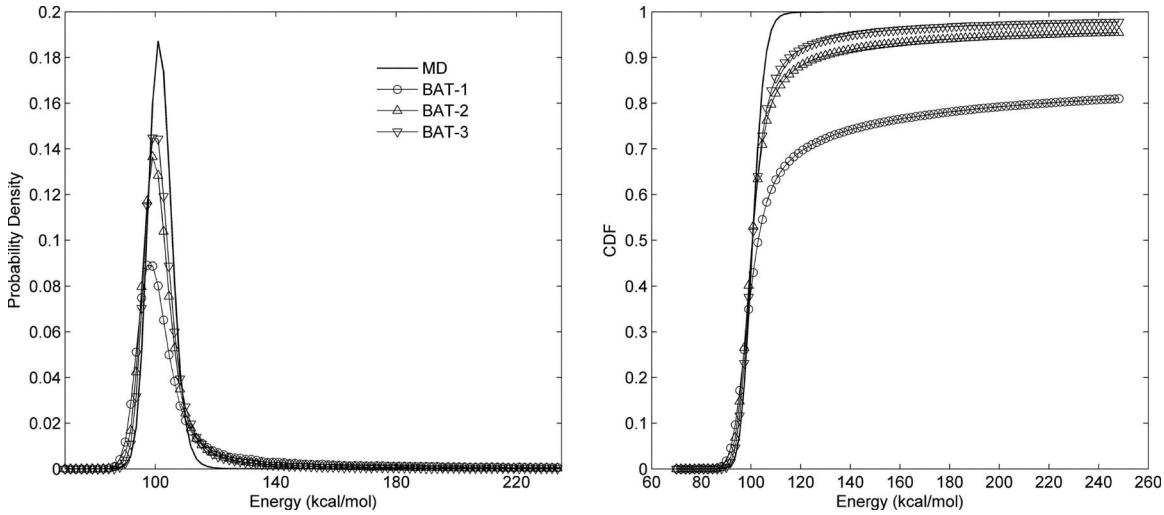

FIG. 11. Probability distributions (left) and corresponding cumulative distribution functions (right) of total energy of the host-guest complex for MD and sampled conformations.

TABLE IX. Accuracy of singlet, doublet, and triplet marginal distributions (columns) computed from conformations sampled at the singlet-, doublet-, and triplet-levels (rows) for the three molecular systems. Results are reported as mean RMSD across all marginals for a specific sampling case. Note that the table includes mean RMSDs of doublet and triplet marginal pdfs for conformations sampled at the singlet level, as well as mean RMSDs of triplet marginal pdfs for conformations sampled at the doublet level.

	Singlet	Doublet	Triplet
	Nonane		
BAT-1	0.0013	0.0023	0.0020
BAT-2	0.0013	0.0014	0.0015
BAT-3	0.0020	0.0019	0.0019
XYZ-1	0.0012	0.0267	0.0157
XYZ-2	0.0915	0.0620	0.0336
XYZ-3	0.0775	0.0530	0.0290
	Cyclohexane		
BAT-1	0.0012	0.0062	0.0045
BAT-2	0.0046	0.0037	0.0025
BAT-3	0.0028	0.0022	0.0017
XYZ-1	0.0012	0.0207	0.0129
XYZ-2	0.0238	0.0180	0.0093
XYZ-3	0.0195	0.0148	0.0083
	Host-guest complex		
BAT-1	0.0021	0.0060	0.0050
BAT-2	0.0098	0.0085	0.0055
BAT-3	0.0063	0.0052	0.0039

reasonable, we did the following test on the MD trajectories of cyclohexane and nonane. For both molecules, we counted the number of conformations in the first 5×10^5 MD that were repeated in the following 4.5×10^6 conformations. We found no repeats for nonane, while $<1\%$ of the first 5×10^5 of cyclohexane were repeated, consistent with the low overlap between the sampled with corresponding MD conformations.

F. Computational complexity of the sampling algorithms

The computational cost of generating one conformation with the sampling algorithm (Sec. II D 2) depends mainly on the level of approximation (i.e., singlet, doublet, or triplet), the number of degrees of freedom sampled, and the number of bins used to represent the distribution functions. The computational cost of sampling at the singlet level is negligible as no conditional distributions need to be computed for each variable. Because the conditional distribution functions at the triplet level [Eq. (21)] have more factors than at the doublet level [Eq. (20)], sampling at the triplet level is computationally more expensive. The time complexities of the doublet- (Algorithm 2) and triplet-level (Algorithm 3) sampling algorithms are linear and quadratic, respectively, in the number of degrees of freedom, but both algorithms are linear in the number of bins used. The present MATLAB implementation took 0.0017, 0.04, and 0.6 s on a 3.8 GHz Pentium 4 personal computer to sample a conformation of the host-guest complex (32 degrees of freedom and 30 bins) at the singlet, doublet, and triplet levels, respectively. Computer time could be substantially reduced by implementation of the algorithm in a lower-level language, such as C or Fortran, and by straightforward code optimizations, such as substitution of an improved logarithm operation, since logarithms take 35% of the central processing unit time in the current implementation of the doublet- and triplet-level sampling algorithms.

V. DISCUSSION

The present paper introduces algorithms for sampling molecular conformations in a manner that includes correlations up to a desired order by means of superposition approximations and employs this algorithm to test the importance of correlations in capturing conformational fluctuations. We find that incorporating higher-order correlations systematically improves the distributions of sampled conformations as compared with the MD conformations, and that conformations sampled via superposition approximations at the doublet or triplet levels resemble reference conformations from MD simulations rather well for one or both of the BAT or Cartesian coordinate systems.

This observation supports the hypothesis that molecular fluctuations may be described to good approximation in the absence of high-order correlations. This assessment relies on the results obtained for the three molecular systems considered here, but we expect the picture to be similar for other molecular systems of similar size and type. It would clearly be of interest to know whether the same is true for larger systems, such as proteins.

This study was motivated by evidence that configurational entropy may be approximated to good accuracy without accounting for high-order mutual information terms (see Sec. I). Comparing sampled conformations, as done here, provides a more stringent test of our hypothesis than merely comparing entropy values, because the same entropy could be obtained for two very different conformational distributions. The present study confirms that neglecting high-order correlations still allows generation of reasonably good conformations. As noted below, the present sampling algorithms can be generalized to use a selected set of higher-order reference pdfs (see also Sec. II B). The present approach therefore provides a framework for investigating the contributions of selected correlations to fluctuations in a multidimensional system.

The present sampling algorithms are approximations to the well-known ancestral sampling algorithm.[6] The ancestral sampling algorithm is a rigorous approach to sampling from a desired pdf, but it becomes difficult to apply to large, highly correlated systems because it requires access to all high-order marginals of the full pdf. This difficulty is addressed here by using approximate conditionals constructed from superposition approximations up to a selected order. This approach leads to some dependence upon sampling sequence, but this sequence dependence could readily be removed, if so desired, simply by randomizing the order in which the coordinates are sampled. The sequence dependence of the sampled conformations could also be removed by using the facts that we can readily compute the sampling probability to each conformation generated by the present

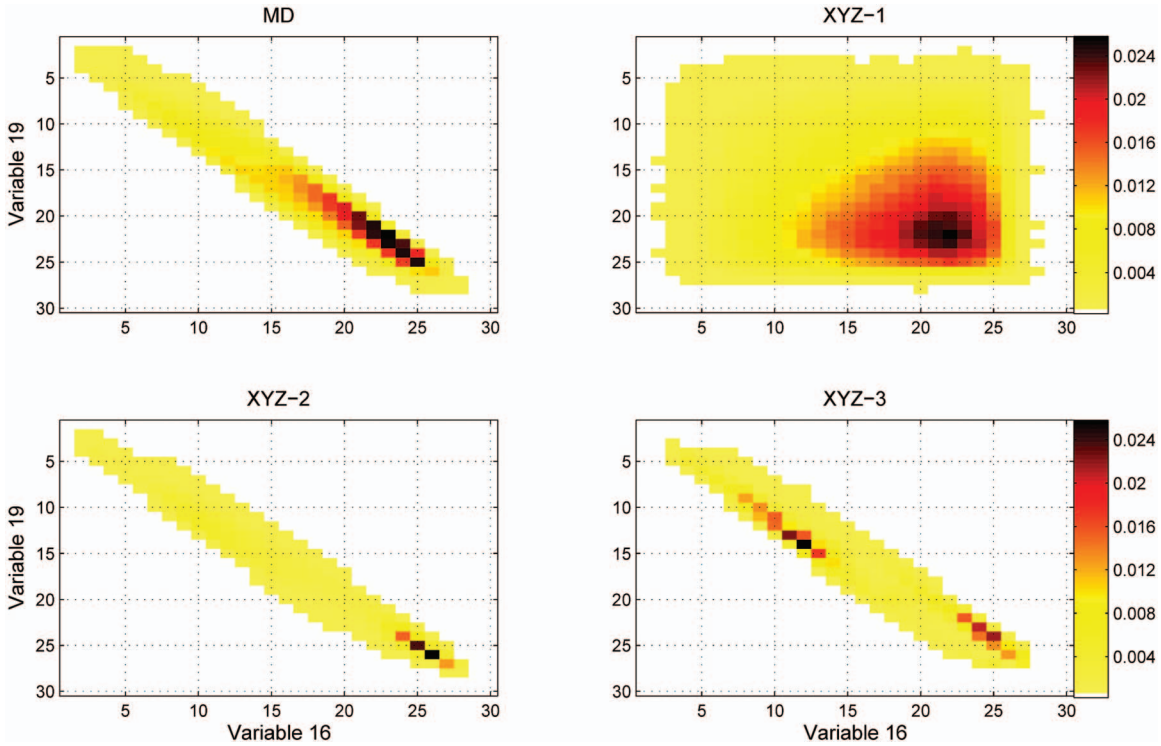

FIG. 12. (Color) Comparison of doublet marginals from MD with XYZ-1, XYZ-2, and XYZ-3 samplings of variables 16 and 19 of nonane, which correspond to the x coordinates of atoms 8 and 9 (Fig. 2). The RMSDs of the sampled doublet marginal pdfs are 0.0504, 0.0955, and 0.0692 for XYZ-1, XYZ-2, and XYZ-3, respectively. Cells are colored on a linear scale of probability, except that cells with identically zero probability are colored white.

algorithms, and that this distribution is expected to have good overlap with the SA-l approximation to the full distribution, which is symmetric in all variables. As a consequence, the present sampling algorithm could be used as a proposal distribution and corrected to the full-dimensional SA-l approximation by standard biased sampling methods.[17]

The generation of null samples (Sec. II D 2) represents a potential problem with the present approach, but this occurred very rarely in the present tests, and we have no reason to expect this to become more problematic for other molecular systems of similar size. A strength of the present approach is that using superposition approximations to approximate the conditional distributions limits the sampled conformations to regions of configuration space for which the reference marginal pdfs are populated. This helps avoid sampled conformations with grossly unrealistic conformations while still allowing construction of new conformations, i.e., ones not present in the MD samples used to construct the marginals.

Several avenues in accelerating and enhancing the present sampling algorithms can be discerned. Straightforward approaches to speeding the sampling are considered in Sec. IV F. It should also be noted that, unlike other methods of sampling from high-dimensional distributions, such as Gibbs sampling, the present approach generates successive samples that are uncorrelated with each other: in effect, the algorithm has no memory of prior conformations. As a result the sampling can be done in an embarrassingly parallel fashion on a distributed computing platform. As to improving accuracy, one obvious approach would be to construct higher-level (say, quadruplet) superposition approximations. However, this could lead to a combinatorial explosion in the number of higher-order marginals to be included. This explosion may be avoided by including only a few critically important high-level marginals in the superposition approximation. In addition, storage needs could be moderated by representing the higher-order marginals with a more sophisticated basis set, such as Gaussian or von Mises distributions, which require far fewer parameters than needed by the present histogram representation. Another approach in improving the accuracy of the sampled conformations is suggested by the dependence of the present results upon the choice of the coordinate system: BAT coordinates are clearly superior for nonane, while Cartesians appear to perform better for cyclohexane. More generally, different coordinate systems result in different degrees of correlation, or coupling, among coordinates. Accordingly, other coordinate systems, such as principal components of the MD trajectory in Cartesian or BAT (Ref. 18) coordinates, might better capture complex molecular fluctuations and other high-dimensional distributions in terms of tractable sets of low-order marginals.

The sampling algorithms introduced here may have practical uses. An immediate application would be to Ytreberg and Zuckerman's reference system approach to computing absolute free energies.[19] The approach entails computing energies of conformations sampled from an approximate reference distribution, and the convergence rate of the free energy depends critically on the degree of overlap between the reference distribution and the true physical distribution. Ytreberg and Zuckerman implemented the method with a reference distribution that neglects all correlations, as it uses only singlet marginal distributions computed from

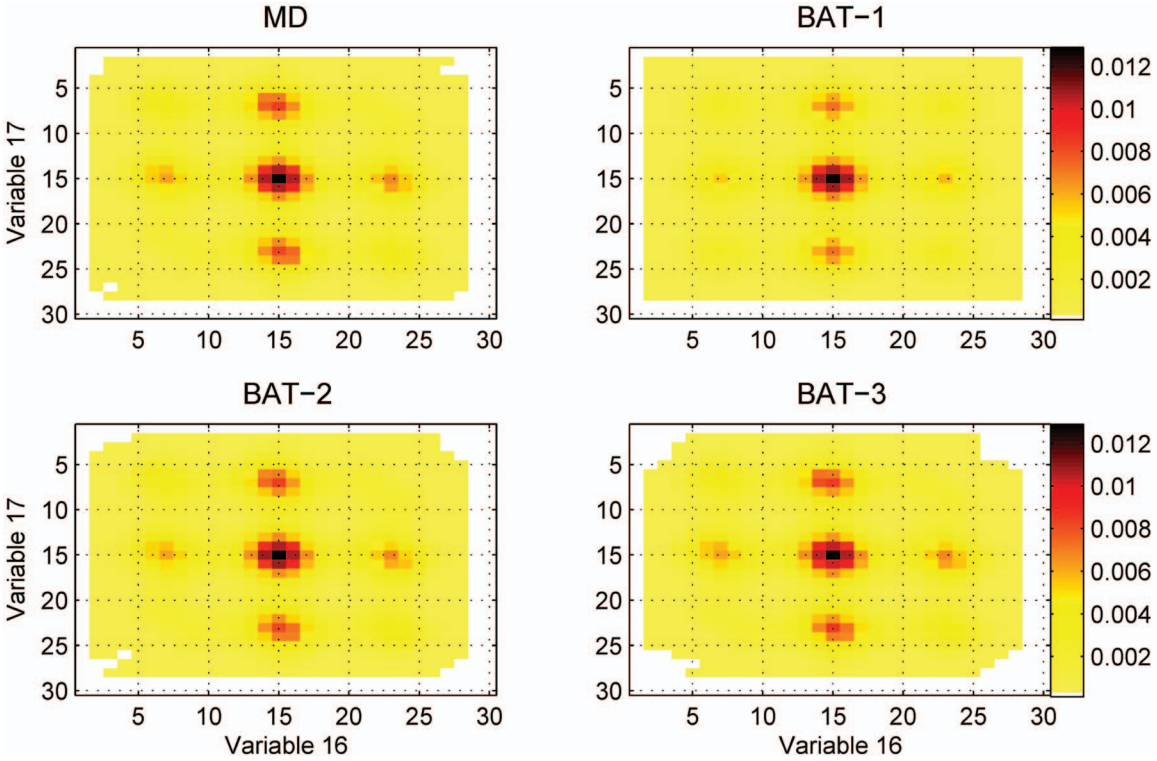

FIG. 13. (Color) Comparison of doublet marginals from MD with BAT-1, BAT-2, and BAT-3 samplings of variables 16 and 17 of nonane, corresponding to torsions 1-2-3-4 and 2-3-4-5, respectively (Fig. 2). The RMSDs of the sampled doublet marginal pdfs are 0.0106, 0.0015, and 0.0018 for BAT-1, BAT-2, and BAT-3, respectively. Cells are colored on a linear scale of probability, except that cells with identically zero probability are colored white.

MD simulation data, but the authors noted that including correlations should improve convergence. The present sampling algorithm could represent an efficient means of introducing correlation into the reference distribution and thus speeding convergence. Another application of the present approach may be to characterize conformational uncertainty—rather than thermal fluctuations—and thus qualify molecular conformations inferred from experimental data, such as crystallography, NMR, or crosslinking studies.

Approximating high-dimensional distributions in terms of low-order marginals is a common theme in the many fields of inquiry that deal with high-dimensional systems.[1–3,20] Although one often has enough data to compute low-order marginals in such cases, the available data are typically too sparse to allow a full, high-dimensional distribution to be evaluated. The present paper describes a novel approach to approximate the high-dimensional distribution in terms of tractable low-order marginals and, furthermore, to sample from this approximate distribution. This work may have applications in fields beyond statistical mechanics, including bioinformatics, structural biology, data mining, and machine learning.

ACKNOWLEDGMENTS

This publication was made possible by Grant No. GM61300 from 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 Institutes of Health.

APPENDIX: DERIVATION OF ANALYTIC EXPRESSION OF SAMPLING PROBABILITY

The l-level sampling algorithms of Sec. II D 2 sample conformations from a distribution, $\tilde{p}_l(X_1,\ldots,X_N)$, which is different from the corresponding l-level superposition approximation $p_{\text{SA-}l}(X_1,\ldots,X_N)$. In this section, we derive an analytic expression for the sampling distribution corresponding to doublet-level ($l=2$) sampling of 3D conformations, $\tilde{p}_2(X_1,X_2,X_3)$, and compare it with the corresponding superposition approximation, $p_{\text{SA-2}}(X_1,X_2,X_3)$. We show that $\tilde{p}_2$, unlike $p_{\text{SA-2}}$, is identically normalized. Additional differences between $\tilde{p}_l$ and $p_{\text{SA-}l}$ will also be highlighted.

To sample a conformation in three dimensions (x_1,x_2,x_3), the doublet-level sampling algorithm requires the following normalized singlet and doublet marginal pdfs as input: [1], [2], [3], [1,2], [1,3], and [2,3]. Let variable X_1 be sampled first, followed in order by X_2 and X_3. The variables are sampled from the following 1D pdfs;

$$f_1(X_1) = [1], \tag{A1}$$

$$f_2(X_2) = \frac{[x_1,2]}{[x_1]}, \tag{A2}$$

and

$$f_3(X_3) = \frac{[x_1,3][x_2,3]}{[x_1][x_2][3]} \frac{1}{N_3(x_1,x_2)}. \tag{A3}$$

Substituting $N_3(x_1,x_2)$ from Eq. (18) into Eq. (A3) gives

$$f_3(X_3) = \frac{[x_1,3][x_2,3]}{[x_1][x_2][3]} \frac{1}{\sum_{x_3} \frac{[x_1,x_3][x_2,x_3]}{[x_1][x_2][x_3]}}$$

$$= \frac{[x_1,3][x_2,3]}{[3]} \frac{1}{\sum_{x_3} \frac{[x_1,x_3][x_2,x_3]}{[x_3]}}. \quad \text{(A4)}$$

Using the chain rule of Eq. (24), the probability $\tilde{p}_2(x_1,x_2,x_3)$ is given by the product of the above 1D pdfs,

$$\tilde{p}_2(x_1,x_2,x_3) = f_1(x_1)f_2(x_2)f_3(x_3) \quad \text{(A5)}$$

or

$$\tilde{p}_2(x_1,x_2,x_3) = \frac{[x_1,x_2][x_1,x_3][x_2,x_3]}{[x_3]} \frac{1}{\sum_{\bar{x}_3} \frac{[x_1,\bar{x}_3][x_2,\bar{x}_3]}{[\bar{x}_3]}}, \quad \text{(A6)}$$

where $\bar{x}_3$ is used to distinguish the summation label from the value of the variable X_3. Next we show that $\tilde{p}_2(x_1,x_2,x_3)$ is normalized, i.e.,

$$\sum_{x_1}\sum_{x_2}\sum_{x_3} \tilde{p}_2(x_1,x_2,x_3) = 1. \quad \text{(A7)}$$

Substituting $\tilde{p}_2(x_1,x_2,x_3)$ from Eq. (A6) in the left hand side of Eq. (A7) and extracting the terms independent of x_3 from the summation over x_3 gives

$$\sum_{x_1}\sum_{x_2}\sum_{x_3} \tilde{p}(x_1,x_2,x_3)$$

$$= \sum_{x_1}\sum_{x_2} \left([x_1,x_2] \frac{1}{\sum_{\bar{x}_3} \frac{[x_1,\bar{x}_3][x_2,\bar{x}_3]}{[\bar{x}_3]}} \left(\sum_{x_3} \frac{[x_1,x_3][x_2,x_3]}{[x_3]} \right) \right). \quad \text{(A8)}$$

The factors involving summations over x_3 and $\bar{x}_3$ cancel, giving

$$\sum_{x_1}\sum_{x_2}\sum_{x_3} \tilde{p}(x_1,x_2,x_3) = \sum_{x_1}\sum_{x_2} [x_1,x_2] = 1. \quad \text{(A9)}$$

The last equality uses the fact that all input marginals are normalized. Thus $\tilde{p}_2(x_1,x_2,x_3)$ is identically normalized. The above procedure can be generalized to distributions in higher dimensions and to higher-level sampling.

Next we point out some general differences between the sampling distribution $\tilde{p}_l$ and $p_{\text{SA-}l}$ by comparing the two distributions for the above case. The $p_{\text{SA-2}}$ in three dimensions is given by setting $k=3$ in Eq. (14) which gives the KSA of Eq. (8). For easy comparison, we rewrite the two distributions,

$$\tilde{p}_2(x_1,x_2,x_3) = \frac{[x_1,x_2][x_1,x_3][x_2,x_3]}{[x_3]} \frac{1}{\sum_{\bar{x}_3} \frac{[x_1,\bar{x}_3][x_2,\bar{x}_3]}{[\bar{x}_3]}} \quad \text{(A10)}$$

and

$$p_{\text{SA-2}}(x_1,x_2,x_3) = \frac{[x_1,x_2][x_1,x_3][x_2,x_3]}{[x_1][x_2][x_3]}. \quad \text{(A11)}$$

The factors $[x_1]$ and $[x_2]$ in $p_{\text{SA-2}}$, which are absent from $\tilde{p}_2$, are guaranteed to be nonzero for any sampled conformation. As a result, any conformation with a nonzero $\tilde{p}_2$ probability will also have a nonzero $p_{\text{SA-2}}$ probability. In this sense, the two distributions overlap. Also, while $p_{\text{SA-2}}$ is symmetric in all three variables, $\tilde{p}_2$ is symmetric in only x_1 and x_2. More generally, the sampling probability $\tilde{p}_l$ is symmetric in only the first l sampled variables in the present sampling algorithms; the remaining asymmetry accounts for the dependence of $\tilde{p}_2$ on the sequence in which the variables are sampled.

[1] B. J. Killian, J. Y. Kravitz, and M. K. Gilson, J. Chem. Phys. **127**, 024107 (2007).
[2] H. Matsuda, Phys. Rev. E **62**, 3096 (2000).
[3] P. Attard, O. G. Jepps, and S. Marcelja, Phys. Rev. E **56**, 4052 (1997).
[4] C. E. Chang and M. K. Gilson, J. Am. Chem. Soc. **126**, 13156 (2004).
[5] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, Biophys. J. **72**, 1047 (1997).
[6] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer, New York, 2006), p. 365.
[7] H. Reiss, J. Stat. Phys. **6**, 39 (1972); A. Singer, J. Chem. Phys. **121**, 3657 (2004); G. Stell, in *The Equilibrium Theory of Classical Fluids*, edited by H. L. Frisch and J. L. Lebowitz (Benjamin, New York, 1964).
[8] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).
[9] J. G. Kirkwood and E. M. Boggs, J. Chem. Phys. **10**, 394 (1942).
[10] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 3rd ed. (Academic, New York, 2006), pp. 83–85.
[11] I. Z. Fisher and B. L. Kopeliovich, Sov. Phys. Dokl. **5**, 761 (1960).
[12] M. J. Potter and M. K. Gilson, J. Phys. Chem. A **106**, 563 (2002).
[13] S. K. Chang and A. D. Hamilton, J. Am. Chem. Soc. **110**, 1318 (1988); S. Goswami and R. Mukherjee, Tetrahedron Lett. **38**, 1619 (1997).
[14] R. Abagyan, M. Totrov, and D. Kuznetsov, J. Comput. Chem. **15**, 488 (1994).
[15] MATLAB, The MathWorks, Inc., 2007 (http://www.mathworks.com/products/matlab/).
[16] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-Mccarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[17] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, New York, 1992), pp. 316–328.
[18] A. Altis, P. Nguyen, R. Hegger, and G. Stock, J. Chem. Phys. **126**, 244111 (2007).
[19] F. M. Ytreberg and D. M. Zuckerman, J. Chem. Phys. **124**, 104105 (2006).
[20] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer, New York, 2006); C. Daub, R. Steuer, J. Selbig, and S. Kloska, BMC Bioinf. **5**, 118 (2004); A. Deshpande, M. Garofalakis, and R. Rastogi, Bell Labs Technical Report, 2001; G. J. Stephens and W. Bialek, http://arxiv.org/abs/8081.0253.