JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Balanced and Bias-Corrected Computation of Conformational Entropy Differences for Molecular Trajectories

Jorge Numata and Ernst-Walter Knapp*

Department of Biology, Chemistry and Pharmacy, Institute of Chemistry and Biochemistry, Fabeckstrasse 36a, 14195 Berlin, Germany

Supporting Information

ABSTRACT: The mutual information (MI) expansion is applied to two molecular systems to probe algorithms that serve to estimate conformational entropy differences more precisely. The individual terms of the MI expansion are evaluated with a histogram method. Internal coordinates are used to avoid spurious correlations, which would require higher order terms in the MI expansion. Two approaches are applied that compensate for systematic errors that occur with a histogram method: (1) Simulation data are balanced by using the same number of coordinate sets (frames) for both conformer domains considered for the entropy difference computation. Balancing puts fluctuations of the histogram bin contents on the same level for both conformer domains, allowing efficient error cancellation. (2) Bias correction compensates for systematic deviations due to a finite number of frames per bin. Applying both corrections improves the precision of entropy differences drastically. Estimates of entropy differences are compared to thermodynamic benchmarks of a simple polymer model and trialanine, where excellent agreement was found. For trialanine, the average error for the estimated conformational entropy difference is only 0.3 J/(mol K), which is 100 times smaller than without applying the two corrections. Guidelines are provided for efficiently estimating conformational entropies. The program ENTROPICAL, used for the computations, is made available, which can be used for molecular dynamics or Monte Carlo simulation data on macromolecules like oligopeptides, polymers, proteins, and ligands.

1. INTRODUCTION

A macrostate of a molecular system can be specified by appropriate thermodynamic variables. The conformational entropy of a molecular system is a measure of the missing information about the specific molecular conformation (microstate) adopted among the many available conformations of the macrostate. This interpretation follows Jaynes' work[1] and Ben-Naim's reformulation of statistical mechanics in terms of information theory.[2] The physical entropy[3] S is proportional to the dimensionless information entropy[4] $\underline{S} = -\Sigma\ p_i \ln(p_i)$ according to $S = k_B\underline{S}$, where the p_i represents the probabilities that the molecular system adopts a particular microstate i and k_B is Boltzmann's constant. The interplay of entropy S and the average internal energy $\langle H\rangle$ is described by the free energy expression $F = \langle H\rangle - TS$. The Boltzmann factor $\exp(-F/k_BT)$ involving the free energy F provides the relative probabilities of occupation for specific macrostates at a given absolute temperature T.

Using thermodynamic relations, it is possible to separate enthalpic and entropic contributions to free energy changes measured experimentally.[5] However, the separation of the total entropy change into solvent and solute components is in general not straightforward.[6,7] The reason is that conformational entropy is a measure of the microscopic variability of conformations, a level of detail which is challenging to resolve experimentally.

Macromolecules involve many degrees of freedom. Therefore, they constitute a special kind of challenge in the estimation of conformational entropy. A variety of methods have been proposed to tackle this problem.[8,9] The quasi-harmonic approximation (QHA)[10−15] is based on "principal component analysis" (PCA), which uses eigenvalue decomposition to account for linear correlations between pairs of coordinates. It fits the observed probability density for the eigenmode coordinates of an effective harmonic oscillator model for which statistical mechanical quantities like the entropy can be expressed analytically. More elaborate QHA approaches apply corrections in third order moments of the coordinates[16] or in pairwise supra-linear correlations.[17,18] A further development of pairwise supra-linear correlations is the "minimally coupled subspace" approach.[19] It combines "independent component analysis"[20] with "mutual information (MI) expansion"[21] and "adaptive kernel density estimator" approaches.[19]

Analyzing the conformational variations of the relatively rigid DNA[15] and RNA[22] duplex with molecular dynamics (MD) simulations using PCA[23] or QHA[10−15] in Cartesian coordinates, the resulting eigenmodes are approximately harmonic.[15,22] Applying these methods to a ribonucleotide duplex,[24] only small corrections for anharmonicity and pairwise supralinear correlations were needed using the nearest-neighbor method.[17] In contrast, peptides and proteins are more flexible and therefore also more anharmonic. As a consequence the internal degrees of freedom of polypeptide chains described in Cartesian coordinates are highly correlated even after removing the linear correlations with PCA or QHA.[25]

Internal Bond−Angle−Torsion (BAT) coordinates avoid spurious correlations that appear in connection with Cartesian coordiantes[26] and can also be applied in the quasi-harmonic

Received: December 20, 2011
Published: March 14, 2012

dx.doi.org/10.1021/ct200910z | J. Chem. Theory Comput. 2012, 8, 1235−1245

approximation.[27−29] Another issue where BAT coordinates are helpful is a problem occurring with high frequency vibrations of molecular systems. A proper description of those vibrations requires a quantum mechanical treatment, where in the limit of large frequencies the entropy contribution vanishes, since only the ground state is occupied. In the classical treatment, the entropy of such vibrations becomes negative for large frequencies and is proportional to the logarithm of the frequency.[17] This is an artifact of classical mechanics that assumes a continuous distribution of microstates, while the quantum mechanical treatment assumes discrete states where for high frequency vibrations mainly the ground state is populated. The advantage of the BAT coordinates is that the degrees of freedom giving rise to high frequency vibrations (mainly bond lengths and in part also bond angles) are approximately decoupled from each other (in contrast to Cartesian PCA and QHA modes where they are strongly coupled) and from the low frequency degrees of freedom (mainly torsion angles). Due to the precise decoupling, the artifactual entropy contributions from high frequency degrees of freedom nearly cancel in entropy differences that we consider in the present work.

Other approaches estimate entropy by fitting the observed distributions in torsion angle space to probability distributions given in closed form[30] like Gaussian and/or von-Mises kernel density estimators.[31−35] The latter approach is nonparametric and approximates the probability density as a sum of peaks for which an analytical expression of the entropy is available. Another unrelated method to compute entropy, inspired by polymer physics, is the rigorous but computationally demanding hypothetical scanning,[36,37] which is based on reconstructing the macromolecular chain conformer from scratch. Methods originally devised to estimate free energy differences, like thermodynamic perturbation and integration, have also been extended to estimate entropy differences.[38−40]

Developments in chaos and nonlinear dynamics theory have revealed a connection between the Kolmogorov−Sinai[41] (KS) entropy rate and thermodynamic conformational entropy.[42] Accordingly, the KS entropy rate measures the relaxation of perturbed phase space trajectories. Wissman et al.[43] propose an algorithm to compute the KS entropy rate using mutual information estimates with histogram bins. Although the KS entropy rate depends on the precise nature of the perturbation of trajectories, it may provide an interesting alternative view to obtaining thermodynamic entropies.

There are two other related approaches: Ceriotti et al.[44] reduce the dimension of a molecular conformational space by a projection algorithm. Brandman et al.[45] analyze the correlations between distant residues in a bacterial ribosome applying mutual information. In the current work, however, we are interested in the thermodynamic conformational entropy, which in principle is a function of the full-dimensional conformational space.

Knowledge of conformational entropy differences is an important ingredient to understanding binding affinities.[9,46] In the present work, we develop efficient algorithms for estimating conformational entropy differences of macro-molecular systems comprising many degrees of freedom. The model systems that we study to test these methods are a three-atom molecule in two different confined spaces and two conformer regimes of trialanine in implicit solvent. For both systems, we generate very precise benchmark entropy values to compare with. We employ the internal BAT coordinates combined with a histogram method to estimate entropy with the mutual information (MI) expansion,[21] which is capable of accounting for supralinear correlations.[47,48] Most importantly, we introduce novel techniques that expedite convergence and compensate bias in estimating conformational entropy differences.

2. THEORY

The aim of the present work is to compute conformational entropy differences of a macromolecule (solute) immersed in a solvent possessing two distinct conformer domains α and β. These are, for instance, conformational domains separated by torsional energy barrier or the native folded and denatured unfolded structures of a protein.[49] In Appendix A of the Supporting Information, we show that the entropy difference for a molecule with N atoms

$$\Delta S_{\alpha\beta} = k_{\mathrm{B}}(\underline{S}_{\alpha} - \underline{S}_{\beta}) \tag{1}$$

can be expressed as a difference of the reduced relative conformational entropy, corresponding to the Shannon differential entropy[4]

$$\begin{aligned}\underline{S}_{\delta} &= -\int_{\Omega_{\delta}} d^{(3N-6)}\mathbf{b}'\rho_{\delta}(\vec{\mathbf{b}}')\ln[\rho_{\delta}(\vec{\mathbf{b}}')] \\ &\equiv -\langle \ln(\rho_{\delta})\rangle,\ \delta = \alpha, \beta\end{aligned} \tag{2}$$

of the conformational probability density function

$$\begin{aligned}\rho_{\delta}(\vec{\mathbf{b}}') = \exp[-U_{\delta}(\vec{\mathbf{b}}')/k_{\mathrm{B}}T]/\int d^{(3N-6)}\mathbf{b}' \\ \exp[-U_{\delta}(\vec{\mathbf{b}}')/k_{\mathrm{B}}T]\end{aligned} \tag{3}$$

In expressions 2 and 3, we use local spherical polar coordinates,[50−52] also called "bond−angle−torsion" (BAT[53]) coordinates

$$\begin{aligned}&\vec{\mathbf{b}}' = (b_2, b_3, \theta_3, \mathbf{b}_4, \mathbf{b}_5, \ldots.\ \mathbf{b}_{N-1}, \mathbf{b}_N), \\ &\mathbf{b}_n = (b_n, \theta_n, \varphi_n)\end{aligned} \tag{4}$$

with bond length b_n, inclination angle θ_n, and azimuthal angle φ_n. The prime at the coordinate supervector $\vec{\mathbf{b}}'$ denotes that it describes only the conformation of a molecule and excludes the overall translational and rotational degrees of freedom. The multidimensional differential in eqs 2 and 3 is defined as (see Supporting Information Appendix B)

$$d^{(3N-6)}\mathbf{b}' = b_2^{\ 2}db_2 \times b_3^{\ 2}db_3 \times \sin\theta_3 d\theta_3 \times \prod_{n=4}^{N} d^{(3)}b_n \tag{5}$$

with

$$d^{(3)}\mathbf{b}_n = b_n^{\ 2}db_n \sin\theta_n d\theta_n d\varphi_n \tag{6}$$

If P torsions φ_n share three atoms (for a methyl group $P = 3$), geometrical correlations can be reduced furthermore by transforming $P - 1$ torsions into phase angles[54] ϕ_n. More technical details on obtaining a set of nonredundant internal BAT coordinates for a given molecular topology are given in Appendix C of the Supporting Information. These local coordinates are adapted to the molecular structure by separating degrees of freedom with high flexibility (torsion angles) from those with low flexibility (bonds and bond angles). This helps in avoiding strong but spurious correlations inherent in atomic Cartesian coordinates.[26] Since these correlations are large, they can mask the physically relevant correlations.

In case an implicit solvent model is used, the potential energy function U_δ is explicitly defined, and the configurational integral in eq 2 can, *in principle*, be evaluated directly. For an explicit solvent model, U_α depends implicitly on the thermodynamic state of the system (i.e., pressure and temperature) and involves averaging the Boltzmann factor $\exp(-U_{\mathrm{solv}}/k_{\mathrm{B}}T)$, where U_{solv} is the solute−solvent interaction, over "free" solvent configurations for each fixed solute conformation.[55] In both cases, the resulting U_α incorporates the influence of the solvent on the distribution of the molecular conformations.[56] It should be noted that a rigorous separation of the conformational entropy of a solute from the entropy of the embedding solvent[57,58] is not possible in the current scheme because of the correlations between the two molecular subsystems.

3. METHOD

3.1. Entropy Evaluation. The conformational entropy, eq 2, of a macromolecule with N atoms involves an integral in $3N - 6$ dimensions. Hence, even for a small macromolecule, solving such integrals suffers from the curse of dimensionality. It is virtually impossible to perform these integrals explicitly even for molecules of a few atoms. Alternatively, one can use sampling methods based on MD or Monte Carlo (MC) simulations, which generate molecular conformers in the frame of a canonical or quasi-canonical ensemble. In case the computation of canonical ensemble averages (free energy, enthalpy, and entropy) can be based on a single equilibrated trajectory, importance sampling with Metropolis MC or MD simulation[59] and energy averaging are straightforward techniques to evaluate them. We will use this method to obtain reliable benchmark data to compare our results with. This method is however not applicable if the problem requires the use of different trajectories from independent simulations, as is generally necessary for studying protein−ligand binding. Hence, other procedures are needed which can deal with data from different trajectories. An alternative for these cases is the method we outline below.

3.2. Mutual Information Expansion in Low Dimensional Subspaces. The convergence of the reduced conformational entropy $\underline{S}$, eq 2, suffers from the curse of dimensionality. Therefore, we approximate $\underline{S}$ by a systematic series, projecting the probability distribution function ρ from the L-dimensional space, spanned by the generalized coordinates $\vec{q}^{\,t} = (q_1, q_2, ...q_{\mathrm{L}})$ into subspaces of lower dimensions as defined below

$$\rho_{(3)i,j,k}(q_i, q_j, q_k) = -\int \rho(\vec{q})\, d^{(L-3)}q_{i,j,k}, \quad \text{with } d^{(L-3)}q_{i,j,k} = \prod_{l \neq i,j,k}^{L} dq_l \tag{7a}$$

and the analog expressions of two- and one-dimensional reduced probability distribution functions

$$\rho_{(2)i,j}(q_i, q_j) = \int \rho_{(3)i,j,k}(q_i, q_j, q_k)\, dq_k \quad \text{and } \rho_{(1)i}(q_i) = \int \rho_{(2)i,j}(q_i, q_j)\, dq_j \tag{7b}$$

Individual values of these low dimensional probability densities can be readily estimated from a finite set of simulation data. Their statistical accuracy improves the lower the dimension of the considered subspace is. With these reduced probability distributions, one can define entropy expressions in the corresponding low dimensional conformational space as, for instance

$$\underline{S}_{(3)i,j,k} = -\int \rho_{(3)i,j,k}(q_i, q_j, q_k) \ln(\rho_{(3)i,j,k}(q_i, q_j, q_k)) \prod_{l=i,j,k} J_l dq_l \tag{8a}$$

for the three-dimensional subspace and analog expressions for the two- and one-dimensional subspaces

$$\underline{S}_{(2)i,j} = -\int \rho_{(2)i,j}(q_i, q_j) \ln(\rho_{(2)i,j}(q_i, q_j)) \prod_{l=i,j} J_l dq_l \tag{8b}$$

and

$$\underline{S}_{(1)i} = -\int \rho_{(1)i}(q_i) \ln(\rho_{(1)i}(q_i)) J_i dq_i \tag{8c}$$

The factors J_n appearing in the conformational integrals of eqs 8 are from the Jacobian determinant describing the transformation of the volume element from Cartesian to generalized coordinates. In the present application, we use BAT coordinates, eq 5, where according to eq 6 the Jacobian factors J_n are

$$J_n = b_n^{\,2} \text{ for } q_n = b_n,\ J_n = \sin\theta_n \text{ for } q_n = \theta_n,\ J_n = 1 \text{ for } q_n = \varphi_n \tag{9}$$

We are now prepared to formulate the MI expansion for entropies in an L dimensional phase space[17,21,47]

$$\underline{S}_{\mathrm{MIE}} = \sum_{i=1}^{L} s_{(1)i} - \sum_{i<j=1}^{L} I_{(2)i,j} + \sum_{i<j<k=1}^{L} I_{(3)i,j,k} - ... \tag{10}$$

The terms $I_{(2)i,j}$ and $I_{(3)i,j,k}$ in eq 10 are the mutual information terms of second and third order, respectively, which are defined as

$$I_{(2)i,j} = s_{(1)i} + s_{(1)j} - s_{(2)i,j} \tag{10a}$$

and

$$I_{(3)i,j,k} = s_{(1)i} + s_{(1)j} + s_{(1)k} - (s_{(2)i,j} + s_{(2)i,k} + s_{(2)j,k}) + s_{(3)i,j,k} \tag{10b}$$

The MI expansion 10 starts with the sum of marginal entropy contributions $\underline{S}_{(1)i}$ in the individual one-dimensional subspaces, neglecting correlations between them. The next terms correct for these correlations up to a given order. The MI expansion can in principle be extended to any desired order, up to full dimensionality.[21] However, higher order terms have a notoriously difficult convergence behavior. In the present work, we will use the MI expansion up to third order. According to our experience with the model systems in this work, as well as evidence gathered from related work,[34,48] such low-order MI expansions are sufficient to evaluate entropies from peptide and protein molecular trajectories reliably when internal BAT coordinates are used.

3.3. Discretization. To evaluate the subspace entropies according to eq 8, the molecular conformer coordinates obtained from a simulation of a canonical ensemble are discretized using histogram bins. In a three-dimensional subspace spanned by the coordinates $\vec{q}_{ijk} = (q_i, q_j, q_k)$, the bins are numbered by the integer vector $\vec{m}_{ijk} = (m_i, m_j, m_k)$, where the m_l ($l = i, j, k$) run from 1 to M_l and their widths are given by Δq_l. If the total number of conformations belonging to conformer regime δ is $N^{(\delta)}$ and the number of conformations in the bin $\vec{m}_{ijk}$ belonging to the three-dimensional subspace spanned by $\vec{q}_{ijk}$ is $N^{(\delta)}_{(3)i,j,k}\vec{m}_{ijk}$, the corresponding discretized probability is

$$p^{(\delta)}_{(3)i,j,k}(\vec{m}_{ijk}) = N^{(\delta)}_{(3)i,j,k}(\vec{m}_{ijk}) \Big/ \Big(N^{(\delta)} \prod_{l=i,j,k} J^{(\delta)}_l(m_l)\Delta q_l\Big) \tag{11a}$$

such that the probability density function is normalized to unity according to

$$1 \equiv \frac{1}{N^{(\delta)}} \sum_{m_i,m_j,m_k=1}^{M_i,M_j,M_k} N^{(\delta)}_{(3)i,j,k}(\vec{m}_{ijk}) \simeq \int \rho^{(\delta)}_{(3)i,j,k}(\vec{q}_{ijk})\, dq_i dq_j dq_k$$

The $J^{(\delta)}_l(m_l)$ refers to the Jacobian factors, eq 9, for the different BAT coordinates q_l. Using analog definitions, the discretized probabilities for two- and one-dimensional subspaces (normalized the same way as in the three-dimensional subspace) are

$$p^{(\delta)}_{(2)i,j}(\vec{m}_{ij}) = N^{(\delta)}_{(2)i,j}(\vec{m}_{ij}) \Big/ \Big(N^{(\delta)} \prod_{l=i,j} J^{(\delta)}_l(m_l)\Delta q_l\Big) \tag{11b}$$

and

$$p^{(\delta)}_{(1)i}(m_i) = N^{(\delta)}_{(1)i}(m_i) / (N^{(\delta)} J^{(\delta)}_i(m_i)\Delta q_i) \tag{11c}$$

On the basis of these discretized probabilities, the entropies in the three-, two-, and one-dimensional subspaces spanned by $\vec{q}_{ijk}$, $\vec{q}_{ij}$, and q_i, respectively, can be written as

$$\underline{S}^{(\delta)}_{(3)i,j,k} \simeq -\Big(\prod_{l=i,j,k} \Delta q_l\Big) \sum_{m_i,m_j,m_k=1}^{M_i,M_j,M_k} \Big(\prod_{l=i,j,k} J^{(\delta)}_l(m_l)\Big) p^{(\delta)}_{(3)i,j,k}(\vec{m}_{ijk}) \ln(p^{(\delta)}_{(3)i,j,k}(\vec{m}_{ijk})) \tag{12a}$$

$$\underline{S}^{(\delta)}_{(2)i,j} \simeq -\Big(\prod_{l=i,j} \Delta q_l\Big) \sum_{m_i,m_j=1}^{M_i,M_j} \Big(\prod_{l=i,j} J^{(\delta)}_l(m_l)\Big) p^{(\delta)}_{(2)i,j}(\vec{m}_{ij}) \ln(p^{(\delta)}_{(2)i,j}(\vec{m}_{ij})) \tag{12b}$$

$$\underline{S}^{(\delta)}_{(1)i} \simeq -\Delta q_i \sum_{m_i=1}^{M_i} J^{(\delta)}_i(m_i) p^{(\delta)}_{(1)i}(m_i) \ln(p^{(\delta)}_{(1)i}(m_i)) \tag{12c}$$

used to evaluate the MI expansion 10. To account for the periodicity of the torsion angles, the histogram bins are placed appropriately, using an adaptive algorithm described in Supporting Information Appendix C.

3.4. Bias Correction. The entropy expressions, eq 12 are based on estimates of the probability density function using finite samples that represent the canonical ensemble. These are subject to fluctuations, which lead to deviations (bias) that systematically underestimate the true value of entropy.[60,61] A simple example to demonstrate the underestimation of entropy estimates and the bias correction is shown in Appendix D of the Supporting Information. In the limit of small probabilities to find the molecular system in one particular bin of the histogram, a simple correction (bias correction) term can be added that compensates this bias and yields bias-corrected (unbiased) entropy estimates according to

$$\begin{aligned} \hat{\underline{S}}^{(\delta)}_{(3)i,j,k} &= \underline{S}^{(\delta)}_{(3)i,j,k} + \frac{\hat{M}^{(\delta)}_{ijk} - 1}{2N^{(\delta)}}, \\ \hat{\underline{S}}^{(\delta)}_{(2)i,j} &= \underline{S}^{(\delta)}_{(2)i,j} + \frac{\hat{M}^{(\delta)}_{ij} - 1}{2N^{(\delta)}}, \\ \hat{\underline{S}}^{(\delta)}_{(1)i} &= \underline{S}^{(\delta)}_{(1)i} + \frac{\hat{M}^{(\delta)}_{i} - 1}{2N^{(\delta)}} \end{aligned} \tag{13}$$

The $\hat{M}^{(\delta)}$ counts only the occupied bins of the histograms (with $p^{(\delta)}(\vec{m}) > 0$), such that $\hat{M}^{(\delta)}_{ijk} \leq \Pi_{l=i,j,k}Ml$ and $N^{(\delta)}$ is the total number of frames (molecular conformations) in the sample. Evidently, the corrections are larger for entropy terms in higher dimensional subspaces.[62,63] The bias correction, eq 13, depends only on parameters characterizing the evaluation of the data ($\hat{M}^{(\delta)}$, $N^{(\delta)}$).

3.5. Balancing. When calculating entropies to compare with experiments, one is interested in entropy differences rather than absolute entropies. In the present application, we evaluate entropy differences between states belonging to two different conformer regimes (see eq 1). Analogous applications are possible between the bound and unbound states of a protein–ligand system. Using a finite number of frames (molecular conformations from simulation), the entropy difference may already have converged, although the entropies of individual conformer regimes have not. We have noticed that convergence of entropy differences is most efficient when the same number of effectively independent frames ($N^{(\alpha)} \approx N^{(\beta)}$) is used for both conformer regimes (α, β). When the two conformer regimes are simulated in a single trajectory, an imbalance occurs if $N^{(\alpha)} > N^{(\beta)}$. Discretizing these data in a histogram, eq 11, the systematic errors for the subspace entropies, eq 12, will differ for the two conformers. Hence, the systematic errors will not cancel in the entropy difference, eq 1, using these subspace entropies. To avoid this problem, the set of data is balanced keeping all frames of the minority conformer regime β, while reducing the number of frames of the majority conformer regime α by randomly deleting frames of the conformer α. If instead only a contiguous part of the trajectory is used to reduce the number of frames of the majority conformer, the two conformer regimes are no longer explored under the same conditions. Effects of such a nonequivalent exploration are discussed under the heading "Importance of Choosing Frames at Random in the Balancing Method" in Appendix E of the Supporting Information.

Here, we provide an explanation for the observation that with less data for the majority conformer regime, better estimates of entropy differences are obtained. Smooth probability distributions have higher entropy than rough distributions. For example, a perfectly smooth Gaussian probability distribution provides the maximum entropy for fixed variance.[64] Alternatively, a rough, multipeaked probability density of the same variance contains more information, since the multipeaked distribution "classifies" data in more detail. It is well established[63] that the statistical bias originates from statistical variations in the bin values of the histogram p, eq 11 representing the true probability density ρ, eq 7. It is evident that histograms will on average become smoother the more data are used to estimate the distribution. We conjecture that balancing works well because it produces histograms with comparable roughness in both conformer regimes. Thus, the bias from the histogram roughness cancels in the entropy difference, eq 2. This behavior is detailed in Appendix E of the Supporting Information.

Hence, we recommend applying balancing when the conformers of both regimes are taken from the same trajectory. However, balancing will also work, if different trajectories of the same molecular system are simulated under equivalent conditions. In contrast to the bias correction that applies to entropies of individual states (conformer regimes), the balancing correction applies only to entropy differences. While the bias correction term, eq 13, compensates systematic deviations (bias) connected mainly with the evaluation procedure, balancing accounts also for systematic deviations that depend on the particular system under study.

3.6. Generate Molecular Conformations in a Canonical Ensemble. Data that represent a canonical ensemble of molecular conformations can be generated by MD. We use Langevin dynamics as implemented in CHARMM35b1 as a thermostat. To avoid slowing down the dynamics as observed in ref 65, a friction constant of $\gamma_{\mathrm{Lang}} = 1\ \mathrm{ps}^{-1}$ is used.[66] However, other thermostats such as the Andersen thermostat[67] or Nosé–Hoover chains[68] may also be appropriate.[69]

Some implicit solvent models such as GBMV with standard parameters are known not to conserve energy[70] in microcanonical (NVE) MD simulations because of the complexity of the molecular surface of the solute used to approximate the Poisson–Boltzmann solvation free energy. As a consequence, these models combined with a thermostat may generate imperfect canonical ensembles. Therefore, we prefer to use the energy conserving implicit solvent model FACTS,[71] defined purely on the basis of pairwise distances between atoms.

3.7. Benchmark Entropy. The free energy difference between two molecular conformer regimes (α and β) can be calculated from a single trajectory if equilibrated simulation data reflecting Boltzmann statistics are available. If $N^{(\alpha)}$ and $N^{(\beta)}$ are the number of conformations (frames) in the conformer regime α and β, the conformational free energy difference can be calculated according to

$$\Delta F_{\alpha\beta} = F_\alpha - F_\beta = -k_{\mathrm{B}}T\ \ln(N^{(\alpha)}/N^{(\beta)}) \tag{14}$$

The entropy difference between two conformer domains α and β of a macromolecule with N atoms can be written:

$$\Delta S_{\alpha\beta,\mathrm{bench}} = (\Delta\left\langle H_{\alpha\beta}\right\rangle - \Delta F_{\alpha\beta})/T \tag{15}$$

Using eq S6, $\langle H_\delta\rangle = 3/2Nk_{\mathrm{B}}T + \langle U_\delta\rangle$, and conformers at the same temperature T, the kinetic energy cancels in the difference of the internal energy, yielding

$$\Delta\left\langle H_{\alpha\beta}\right\rangle = \Delta U_{\alpha\beta} = \left\langle U_\alpha\right\rangle - \left\langle U_\beta\right\rangle \tag{16}$$

The potential energy difference $\Delta U_{\alpha\beta}$, eq 16, can be evaluated from MD simulation data by averaging over all frames of conformer regimes α and β, respectively. Entropy differences computed from data of a single trajectory based on eq 15 converge more rapidly than an MI expansion. Therefore, they can be used as a benchmark to test the MI expansion method. The MI expansion converges more slowly because it requires estimates of multidimensional averages, whereas eq 15 works by just averaging two scalar quantities, namely internal energy and free energy, where the latter is evaluated according to eq 14. Nevertheless, the MI expansion is more generally applicable. The use of a single trajectory is limited to cases where the conformational transition kinetics is fast enough to observe such transitions often enough within the simulation to obtain good statistics. Conversely, the MI expansion can also be applied to compute entropy differences for situations where the slow transition kinetics requires the conformer regimes to be generated by independent MD simulations where relation 14 cannot be applied to compute the free energy difference. Such independent trajectories are required, for instance, to evaluate the binding affinities of ligand–receptor or protein–protein complexes.

Using MD simulation data to evaluate $\Delta F_{\alpha\beta}$ according to eq 14, $\Delta F_{\alpha\beta}$ converges more rapidly with the length of the trajectory than the evaluation of $\Delta S_{\alpha\beta}$ and $\Delta U_{\alpha\beta}$ based on eqs 15 and 16, respectively. This convergence behavior has been reported previously[38] and is discussed at length for our simulations in Appendix F. The simulation data are obtained with importance sampling based on Boltzmann statistics such that for an evaluation of $\Delta F_{\alpha\beta}$ all frames are used with equal weights, while for evaluation of $\Delta U_{\alpha\beta}$ the frames need to be reweighted using the potential energy terms $U^{(\delta)}$, $\delta = \alpha, \beta$. Hence, the effective number of frames available for the latter case is smaller, resulting in larger statistical errors. As a consequence, the convergence of the benchmark value of $\Delta S_{\alpha\beta}$ according to eq 15 is limited by the convergence behavior of $\Delta U_{\alpha\beta}$.

3.8. Monte Carlo Simulation of a Three-Atom Molecule in a Cage. The first model system that we investigate is a three-atom molecule whose conformations are generated by a continuous random walk starting at the origin with fixed step size (bond length). The first atom is considered to be fixed at the minimum of a wall opened toward the positive z axis defining a cage (see Figure 1). The wall surface obeys the relation

$$z_{\mathrm{wall}}(x, y) = \varepsilon(x^2 + y^2)^{1/2},\ \varepsilon > 0 \tag{17}$$

The second atom can change its position by varying its angle θ_2 relative to the plane rectangular to the z axis. The third atom can move by rotating around the axis formed by atoms 1 and 2 by the azimuthal angle φ and by varying the bond angle θ_3 of the three atoms. Rotations of the molecule around the z axis do not matter, since they do not change the configuration of the wall and molecule, due to the rotational symmetry of the wall surface, eq 17. The set of angular variables (θ_2, θ_3, φ) is analogous to the BAT coordinates. They are the internal coordinates of the molecule fixed with atom 1 at the wall.

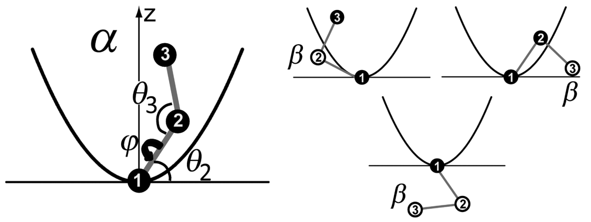

Figure 1. Three-atom molecule modeled as a continuous random walk with fixed bond lengths. Each conformation is defined by two bond angles (θ_2, θ_3) and one torsion angle φ. The conformer regime α is the ensemble of conformations where atoms 2 and 3 are both above the depicted parabolic wall, eq 17, while conformer β comprises all other conformers (right part).

The locations of atoms 2 and 3 relative to the wall surface determine the conformer regime (α or β) to which the molecule belongs. If both atoms are above the wall, the molecule belongs to conformer regime α. If one or both atoms are below the wall, the molecule is in conformer regime β. In this way, we have constructed a molecular model with an asymmetric distribution between the two conformer regimes. Choosing the parameter value $\varepsilon = 0.612$ for the wall surface, eq 17, yields an asymmetry of 1 to 10.4 in the proportion of conformations between regimes α and β. A simple MC procedure is used to generate 5×10^7 free molecular conformations. Then the wall surface, eq 17, is introduced, and the molecular conformers are assigned to one of the two conformer regimes. This procedure is described in more detail in the Supporting Information, Appendix G.

3.9. Molecular Dynamics Simulation of Trialanine.

We have chosen trialanine as a model system because it has been shown experimentally[72,73] and in simulations[74,75] to possess two conformer regimes, and because the transition kinetics between the conformer regimes is fast enough to benchmark the entropy change using the single trajectory method. Simulations of trialanine were performed with 13 different conditions, each one spanning a 1 μs trajectory. The canonical ensemble was approximated using the Langevin thermostat with coupling constant $\gamma_{\text{Lang}} = 1$ ps^{-1}. The time propagation step was 1 fs. No SHAKE constraints were used to account also for entropy contributions from hydrogen atom bond vibrations. Conformations were saved every 0.2 ps for a total of $N_{\text{frames}} = 5 \times 10^6$. The CHARMM22 force field[76] was used together with the implicit solvation model FACTS[71] with parameters $\kappa = 8$ and dielectric constant $\varepsilon = 1.0$ implemented in CHARMM35b1. In order to generate a total of 13 simulations with different entropies, we varied the hydrophobic "surface tension" term $\gamma_{\text{H}\phi}$ and scaled the attractive $1/r^6$ term of the Lennard-Jones potential by the dimensionless factor $\varepsilon_{\text{attr}}$. For vanishing surface tension ($\gamma_{\text{H}\phi} = 0.0$), we used $\varepsilon_{\text{attr}}(j) = 0.00 + 0.25j$, $j = 0, 1, 2, ...6$, and for $\gamma_{\text{H}\phi} = 0.025$ cal/(mol K Å^2) and $\gamma_{\text{H}\phi} = 0.045$ cal/(mol K Å^2), we used $\varepsilon_{\text{attr}} = 0.00$, 0.50, and 1.0.

The molecular conformations of the trajectories were postprocessed to generate two conformers by using a geometric criterion. The main anharmonic motion in trialanine is about the dihedral angle ψ_2 of the middle residue, as has also been shown experimentally,[72,73] so we have chosen ψ_2 as our "order parameter" (see Figure 2). We separate two conformers of trialanine by searching for two minima in occupation in the torsion angle ψ_2. As a result, we obtain a conformer regime α with dihedral angles similar to an α-helix and a conformer regime β with torsion angles similar to polyglycine 3_1-helix (P_{II}).

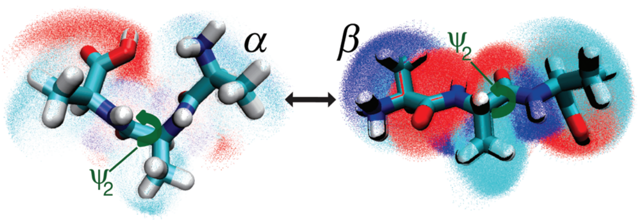

Figure 2. The compact (α) and extended (β) conformers of trialanine. The Ramachandran torsion ψ_2 is used as an order parameter to define the conformers. The compact conformer α has a lower (more favorable) potential energy $U_\alpha < U_\beta$ but also lower (more unfavorable) entropy $\underline{S}_\alpha < \underline{S}_\beta$, eq 2, than the extended conformer. By how much U and $\underline{S}$ differ is a function of the surface tension ($\gamma_{\text{H}\phi}$) and the scaled $1/r^6$ attractive term of the Lennard-Jones potential ($\varepsilon_{\text{attr}}$), which were varied in each of the 13 simulations.

Our trialanine model consists of $N = 34$ atoms. Its geometry can be described with $N - 1 = 33$ bonds, $N - 2 = 32$ angles, and $N - 3 = 31$ torsions, yielding a total of 96 BAT coordinates. Furthermore, the torsion angles can be divided into 13 main torsions and 18 associated phase angles.

4. RESULTS

4.1. Entropy Estimation for the Three-Atom Molecule in a Cage.

A two-step continuous unconstrained random walk starting at the coordinate origin can be considered as a three-atom model where the first atom is fixed at the origin. The Cartesian as well as the three internal coordinates (θ_2, θ_3, φ) (for a definition see Figure 1) of such a molecular model are by construction uncorrelated. By introducing a wall to divide the ensemble of conformers into two regimes, correlations between the coordinates are introduced. All three internal coordinates are supralinearly[77] correlated, as evidenced by nonvanishing pairwise $I_{(2)i,j}$ and third order $I_{(3)1,2,3}$ MI terms.

For the chosen value of the curvature $\varepsilon = 0.612$ of the quadratic wall, eq 17, we obtain for 5×10^7 random walks (frames), $N^{(\alpha)} = 4.38 \times 10^6$ conformers of type α and $N^{(\beta)} = 45.6 \times 10^6$ conformers of type β (Figure 1). Since β is the majority conformer, applying balancing means to randomly select N_α frames of conformer regime β for the entropy difference computation. The benchmark entropy, eq 15, may be used as a standard. It converges quickly with the number of frames (solid line in Figure 3) to the value $\Delta S_{\alpha\beta} = -\Delta F_{\alpha\beta}/T = k_{\text{B}} \ln(N^{(\alpha)}/N^{(\beta)}) = -19.5$ J/(mol K), since $\Delta U_{\alpha\beta} = 0$.

Among the estimators, the slowest convergence is observed (Figure 3) when neither balancing nor bias correction is applied, equivalent to the original method by Gilson et al.[47,48] The fastest convergence is achieved by applying both balancing and bias correction. When separating the effects, the balancing method alone provides a stronger improvement for a small number of frames (N_{frames}), while bias correction provides a stronger improvement for a larger number of frames N_{frames}. In practical applications, it is advisible to apply both balancing and bias correction, as they work synergistically to accelerate convergence.

The abscissa in Figures 3 and 4 is the total number of frames in the simulation. Because in the balancing method we actually use only a subset of those frames, the CPU requirements of the entropy evaluation are reduced by 1 order of magnitude while at the same time improving the convergence. Nevertheless, we

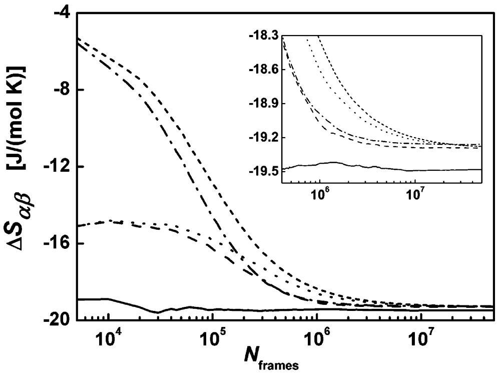

Figure 3. Entropy difference $\Delta S_{\alpha\beta}$ for the three-atom molecule as described in section 3.8, probing the two correction methods, i.e., balancing and bias correction. The computations are based on a total of 5×10^7 conformers. The solid line (—) is the benchmark entropy difference, eq 15. All entropy estimators use the third-order MI expansion with $M = 35$ histogram bins, eqs 10−12. Dashed line (– –), balanced and bias-corrected; dotted line (····), balanced and biased; dash-dotted line (– · –), unbalanced and bias-corrected (reflected to be above the benchmark); short dashed line (----), unbalanced and biased (reflected to be above the benchmark). The latter two curves have been reflected about their asymptotic values $\Delta S_{\alpha\beta}(\infty)$ according to $\Delta S'_{\alpha\beta} = \Delta S_{\alpha\beta}(\infty) - \Delta S_{\alpha\beta}$ for ease of comparison. The inlay zooms into the last phase of convergence. The fastest convergence among estimators is achieved by applying both methods: bias correction and balancing.

use the same abscissa to allow comparison between the methods.

The number of histogram bins M chosen is the resolution at which the conformational space and the correlations between the different variables will be sampled. The dependence on M is plotted in Figure 4 for the uncorrected (a) and for the balanced and bias-corrected (b) methods. If we choose M to be too large, there will not be enough data to fill the bins, and convergence will be slower and incomplete for the given amount of data, which is 5×10^7 conformers. If we choose too small an M, the resolution will not be high enough to capture the correlations. The values of M between 20 and 35 are most suitable for the third order MI expansion using 5×10^7 conformers. There is, however, a dependence of $\Delta S_{\alpha\beta}$ on M, which is reduced by using balancing and bias correction but not completely eliminated. In summary, most of the entropy estimates have reached their asymptotic value when using balancing and bias correction (Figure 4b). Conversely, the corresponding results are far from being converged, if the MI expansion method is used without corrections, i.e., unbalanced and biased (Figure 4a).

4.2. Entropy Estimation for Trialanine.

For the MD simulations of trialanine, larger values of $\varepsilon_{\text{attr}}$ enhance the attractive wing of the Lennard-Jones potential. This leads to more compact conformations ($N^{(\alpha)} > N^{(\beta)}$) and a larger entropy difference $\Delta S_{\beta\alpha}$. Larger surface tension ($\gamma_{\text{H}\phi}$) up to a value of 0.045 cal/(mol K Å^2) had a smaller and opposite effect on $\Delta S_{\beta\alpha}$. By varying $\varepsilon_{\text{attr}}$ and $\gamma_{\text{H}\phi}$, different simulation conditions are created, which are then used to test the entropy estimator on the basis of the MI expansion. The order parameter ψ_2 serves to cluster the conformers α and β (see Figure 2 and Figure S1 of the Supporting Information).

Figure 5 shows the deviation of the entropy estimates from the benchmark values. The simulation conditions are labeled as 1−13, ordered by increasing $\Delta F_{\beta\alpha}$. The simulations with conditions 1 and 2 have vanishing $\Delta F_{\beta\alpha}$, so that $K_{\text{eq}} = N^{(\beta)}/N^{(\alpha)} = \exp(-\Delta F_{\beta\alpha}/k_{\text{B}}T) \approx 1$; i.e., the numbers of frames are equal. In other words, the molecular system is naturally balanced. In contrast, the simulation with condition 13 is very unbalanced with $\Delta F_{\beta\alpha} \gg 0$ and $K_{\text{eq}} \approx 0.07$, such that there is room for improvement using the balancing method. See the caption of Figure 5 for values of the parameters and thermodynamic variables for the 13 simulation conditions.

The deviations in Figure 5 are plotted for the second-order MI expansion including all 96 BAT coordinates. Using the balancing method (green and blue symbols) results in the smallest deviations of $\Delta S_{\beta\alpha}$ from the benchmark values. In particular, combining balancing with bias correction (green) results in an average absolute deviation of less than 0.3 J/(mol K). Using the balancing method without bias correction (blue) results in an average deviation of 0.7 J/(mol K), about twice as large. The unbalanced $\Delta S_{\beta\alpha}$ values (black and red) have generally large, negative deviations and a systematic, spurious dependency on $\Delta F_{\beta\alpha}$. When only bias correction is applied (black), but no balancing, the absolute deviation becomes 7.5 J/(mol K). The red symbols in Figure 5 represent the estimates

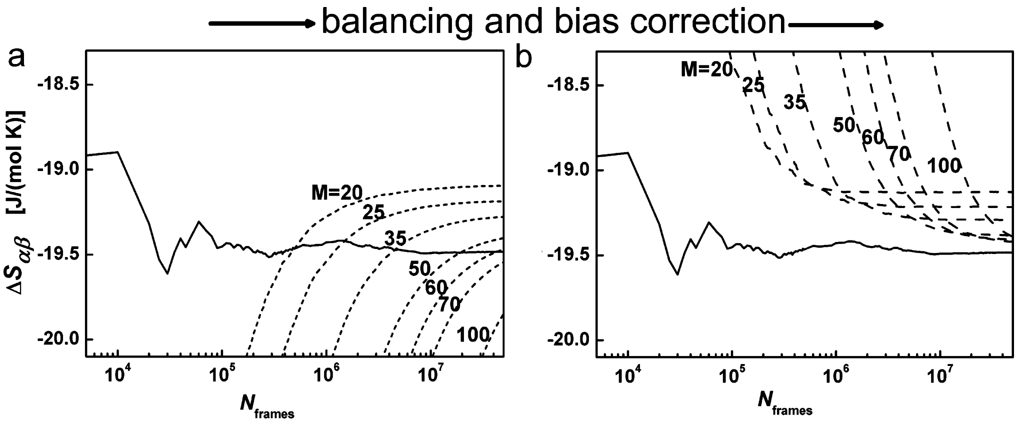

Figure 4. Entropy difference $\Delta S_{\alpha\beta}$ for the three-atom molecule probing different histogram sizes using (a) uncorrected estimates (unbalanced and biased) and (b) corrected entropy estimates (balanced and bias-corrected) with improved asymptotic convergence. The computations are based on a total of 5×10^7 conformations. Solid line is the benchmark entropy according to eq 15. All entropy estimators (dashed lines) use the third-order MI expansion varying the number of histogram bins M. Note that the curves with $M = 35$ are identical to Figure 3 except for the reflection.

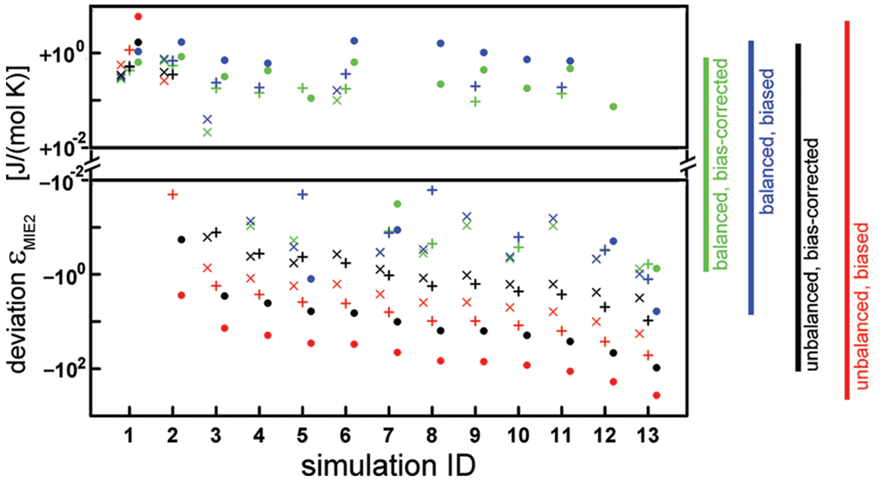

Figure 5. Deviation of the estimated conformational entropy difference $\Delta S_{\beta\alpha}$, eqs 10−12, using the second-order MI expansion (MIE2) with all 96 BAT coordinates from the benchmark value, eq 15. Based on MD simulations of 1 μs with 5 × 10^6 frames (coordinate sets) for trialanine. Smaller deviations are for symbols near the center of the discontinuous logarithmic ordinate. The MD simulations with 13 different conditions are ordered by increasing $\Delta F_{\beta\alpha}$. The color labels the two correction methods used (see bars on the right). The symbols label the number of bins used in histograms: ×, M = 20; +, M = 35; •, M = 100. It is apparent that the deviation of the estimated $\Delta S_{\beta\alpha}$ is smallest when the estimates are both balanced and bias-corrected (green). Details of MD simulations are given in section 3.9. Correspondence between simulation condition ID and parameters is as follows: ID 1 (ε_{attr} = 0.0, $\gamma_{H\phi}$ = 0.045 kcal/(mol Å^2)), 2 (0.00, 0.025), 3 (0.00, 0.000), 4 (0.50, 0.045), 5 (0.25, 0.000), 6 (0.50, 0.025), 7 (0.50, 0.000), 8 (1.00, 0.045), 9 (0.75, 0.000), 10 (1.00, 0.025), 11 (1.00, 0.000), 12 (1.25, 0.000), 13 (1.50, 0.000).

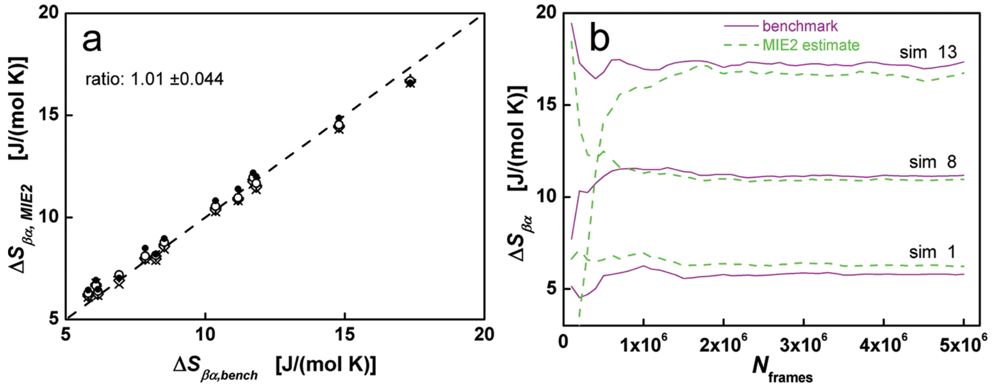

Figure 6. Entropy difference $\Delta S_{\beta\alpha}$ for the trialanine model system using all 96 BAT coordinates (bonds, bond angles, torsion angles, and phase angles) and considering 5 × 10^6 frames, which are in time order. (a) Influence of the number of histogram bins, M, on the estimated entropy difference of the second-order MI expansion (MIE2), plotted versus the corresponding benchmark values for the 13 different MD simulation conditions applying both corrections: balancing and bias correction. The number of histogram bins M was varied: ×, M = 20; ◇, M = 25; +, M = 35; ○, M = 50; •, M = 100. The dashed diagonal line corresponds to perfect agreement between the benchmark and estimate of entropy difference. Also given is the average and standard deviation of the ratio $\Delta S_{\beta\alpha,MIE}/\Delta S_{\beta\alpha,bench}$ over all five M values and data from all 13 simulation conditions (optimal result is 1.0 ± 0.0). (b) Convergence of the benchmark and the MIE2 estimate as a function of the number of frames N_{frames} using M = 35 histogram bins. For the sake of clarity, only three representative simulations (1, 8, 13) are shown. See caption of Figure 5 for correspondence between simulation (sim) condition ID and their parameters.

of $\Delta S_{\beta\alpha}$ where no corrections are applied, corresponding to the original method by Gilson et al.[47,48] In this case, the entropy difference has not converged using all available $N^{(\alpha)} + N^{(\beta)}$ = 5 × 10^6 frames. The average absolute deviation of the estimated entropy difference from the benchmark value in absence of any corrections is 32 J/(mol K), which is about 100 times larger than the corresponding results obtained applying both correction methods.

To obtain close agreement with the benchmark values for the trialanine model system (see Figure 6a), it was necessary to include all 96 BAT coordinates and pair correlations between them, as implemented in the second order MI expansion of the entropy differences. The estimate-to-benchmark ratio $\Delta S_{\beta\alpha,MIE2}/\Delta S_{\beta\alpha,bench}$ (see Figure 6a) was found to be 1.01 ± 0.044 when averaged over all five histogram schemes M (M = 20, 25, 35, 50, 100) and all 13 simulation conditions. In Figure 6b, we see that both benchmark (solid line) and estimated (dashed line) entropy differences are asymptotically converged, with the benchmark converging more quickly. This is shown for three examples in Figure 6b (and for all examples in Figures S2 and S3). The first-order MI expansion converges much more quickly than the second order, but the entropies $\Delta S_{\beta\alpha}$ obtained with the first order MI expansion have an estimate-to-benchmark ratio of 0.82 ± 0.051 (corresponding to 1 − 0.82 = 18% average underestimation; see Table 1). The third-order MI expansion does not converge for the available 5 × 10^6 frames and would likely require at least 1 order of magnitude more frames. See Appendix H for complete data for the first-, second-, and third-order MI expansions.

Table 1. Averages and Standard Deviations for the Estimate-to-Benchmark Ratio $\Delta S_{\beta\alpha,\mathrm{MIE}}/\Delta S_{\beta\alpha,\mathrm{bench}}$ over All 13 Simulation Conditions Using Histograms with M = 35 Bins[a]

	order of MI expansion	
coordinate set	MIE1	MIE2
13 main torsion angles	0.71 ± 0.089	0.82 ± 0.051
13 torsion and 18 phase angles	0.81 ± 0.071	0.97 ± 0.024
all 96 BAT coordinates	0.82 ± 0.091	1.01 ± 0.037

[a] The optimal result is 1.0 ± 0.0. The entropy estimates were computed using the first- and second-order MI expansions (MIE1 and MIE2) applying both methods (balancing and bias correction). The estimate-to-benchmark ratios vary for the different coordinate sets and orders of the MI expansion used. The best results are obtained with MIE2 using all 96 BAT coordinates, and the second best results are for the 31 "soft degrees of freedom" (13 torsions and 18 phase angles).

In the recent work of Brüschweiler and Li,[33] it was suggested to employ the main torsion angles ("soft degrees of freedom") only and to neglect the "hard degrees of freedom". They suggested furthermore[34] to use only the first-order MI expansion, neglecting correlations between the torsion angles. For dipeptides, they obtained estimate-to-benchmark ratios between 0.87 and 0.96 (see Table 1 in ref 33, last column). Neglecting 33 bonds, 32 bond angles, and 18 phase angles, 13 main torsion angles remain for the trialanine model involving 34 atoms. Applying the first-order MI expansion with M = 35, the estimate-to-benchmark ratio averaged over all 13 simulation conditions is 0.71 ± 0.089. Including also pairwise correlations by using the second-order MI expansion raises the average ratio to 0.82 ± 0.051. If we now redefine the "soft degrees of freedom" to include not only the main 13 torsions but also the 18 phase angles, we obtain ratios of 0.81 ± 0.071 (for MIE1) and 0.97 ± 0.024 (for MIE2), which are closer to unity. More details can be found in Appendix H of the Supporting Information. These data are summarized in Table 1.

5. DISCUSSION

Internal BAT coordinates allow a compact representation of the available conformational volume of a molecule. An alternative and complementary view of entropy to the "missing information" is "a measure of the phase space volume" occupied by a certain state (see section 27.3 in ref 78) in the canonical ensemble. The mobility of hydrogen atoms of a methyl group in internal BAT coordinates is characterized mainly by a single dihedral angle, while the remaining two phase angles are less important, since they belong to the stiffer degrees of freedom. All other degrees of freedom of the methyl group describe small amplitude vibrations in three bond angles and three bond lengths. Alternatively, the Cartesian representation of the methyl group hydrogens requires nine geometrically highly correlated coordinates, all of which involve large amplitude motions. Even after applying PCA or QHA,[11,13] such correlations persist[25] for polypeptide chains. Internal BAT coordinates avoid such spurious correlations inherent to Cartesian coordinates and are therefore more suitable to describe the relevant correlations of motion in a molecule, thus yielding improved entropy estimates even if the MI expansion is used in low order only. A further advantage of BAT coordinates is that they efficiently decouple the high frequency degrees of freedom from each other and from the low frequency degrees of freedom. This provides efficient cancellation in entropy differences of high frequency vibrations in the classical treatment, which otherwise would provide artifactual contributions. This can be observed from nearly vanishing contributions to entropy differences originating from bond lengths (see Table S1 and Figure S4 in the Supporting Information).

The signal processing community has designed a wealth of approaches to estimate entropy from samples of time series. They include histogram methods,[61,79] kernel density estimators,[80] and the k-nearest-neighbor approach.[17,77,81−84] For a finite number of samples, all entropy estimators suffer from statistical and systematic biases.[60,61] The systematic bias can be understood intuitively because entropy is a sensitive measure of the "variability" of a probability density, and a finite sample will tend to underestimate this variability. A major focus in signal processing[85] is to estimate entropy with mutual information (MI) estimators for a small number of variables (around 10) and a small number of samples (about 10^3). Entropy estimation for molecular simulation data presents a different type of challenge, since we compute entropy differences considering molecular systems involving 10^2 or more atomic coordinates, where one needs samples of 10^5 or more independent coordinate frames. In this study, we provide evidence that adequately bias-corrected and balanced histogram-based entropy estimators work best for data from molecular simulation. At the same time, its simplicity makes this method computationally more efficient than others like the k-nearest-neighbor approach.

6. CONCLUSION

In this work, trialanine, a small test model molecule, was used to prove that the second-order MI expansion, in conjunction with both balancing and bias correction, allows for proper convergence of the entropy difference $\Delta S_{\beta\alpha}$. This is the case even though the individual conformational entropies $\underline{S}_{\alpha}$ and $\underline{S}_{\beta}$, eq 2, are not converged (see Appendix E of the Supporting Information). Notwithstanding, the estimated values of $\Delta S_{\beta\alpha}$ are in excellent agreement with the corresponding benchmark values.

The use of local spherical polar coordinates,[50−52] the so-called BAT coordinates,[53] enables a clear-cut separation of global translation and rotation from the internal degrees of freedom. In the quasi-harmonic approximation[12,13] and other approaches,[35] the rigid rotor approximation[86] is often used to remove the translational and rotational degrees of freedom. Unfortunately, the rigid rotor introduces spurious mass dependencies[7,14] and correlations between external and internal degrees of freedom, which can be avoided by using BAT coordinates. The BAT coordinates are also adept at describing internal motions of polypeptides for numerical computations of entropy, since this coordinate system minimizes spurious geometric correlations between molecular coordinates.

Without balancing and bias correction, the method has been used before.[47,48] Here, we demonstrated that the uncorrected estimate converges when using a much larger sample size (see Figure 3). However, both the balancing and bias correction methods accelerate convergence in a synergistic fashion and enable a more efficient use of the available simulated frames. The balancing method allows for a more efficient systematic cancellation of sampling errors in entropy differences, which works well even if the individual entropy contributions are poorly converged. Applied simultaneously with balancing, the bias correction method compensates systematic bias due to a limited sample size.

However, just paying attention to the convergence of an entropy estimator does not guarantee that the algorithm works properly and that the results are reliable. In the test phase of an algorithm to compute entropies, a careful comparison with benchmark values is necessary before one can consider applying it to larger macromolecules, where benchmark values are not easily available. This is the purpose of the present study. Such comparisons have been done before,[29,33,47] proceeding then to calculate entropy for large molecular systems. We show here that the converged entropy differences obtained with the balanced and bias-corrected histogram method agree well with thermodynamic benchmarks. For the trialanine model system, the conformational entropy estimates agree with benchmarks to an average deviation of 0.3 J/(mol K), or alternatively an estimate-to-benchmark ratio of 1.01 ± 0.037 (see Table 1). A small standard deviation and an average estimate-to-benchmark ratio close to unity together indicate converged estimates and a thermodynamically relevant result.

We tested the suggestions of Brüschweiler et al.[33,34] of just using the main torsion angles (excluding phase angles, bond angles, and bonds), as well as using the first-order MI expansion only. For trialanine, this resulted in an estimate-to-benchmark ratio of only 0.71 ± 0.089. However, following this line of thought and using only the main torsions and phase angles in the second-order MI expansion yielded results almost as good as those of the full BAT coordinate set, with an estimate-to-benchmark ratio of 0.97 ± 0.024. This reduced set of coordinates is only about 1/3 of the full set of BAT coordinates. As a consequence, the computational cost for the second-order MI expansion is reduced to about 1/9.

The present work provides proof of principle of entropy computation for a small oligopeptide involving 96 degrees of freedom. The computational procedure of this study uses state of the art MI expansion combined with bias-correction and a novel balancing method, which yields converged entropy differences in agreement with benchmark results. There is encouraging evidence from the application of related methods[34,48] that the current approach will also be successful for much larger systems such as proteins and nucleotides.

When estimating conformational entropy differences with the methods presented here, the following guidelines are important: (i) The molecular dynamics trajectories need to be long enough to provide a Boltzmann distribution of equilibrated microstates representative for the considered conformer macrostates. (ii) The full set of BAT coordinates, or alternatively only the torsion and phase angles (plus any external bonds and angles analogous to the ones we defined in Figure 1), need to be considered. (iii) To perform a second-order MI expansion, necessary to achieve a sufficient accuracy of a few percent, a trajectory with a large number of independent frames is needed (at least 10^5 frames with random frame selection for balancing). (iv) Avoid using more frames for the dominant molecular conformer, since the best bias cancellation in entropy differences occurs when the number of frames is balanced. (v) When balancing requires considering only a subset of the total number of frames, select the frames randomly from the whole trajectory to utilize the conformational space explored by the simulation as completely as possible.

Since entropy estimators are generally biased, the balancing method presented here is likely also applicable for algorithms estimating entropy differences by methods other than histogram binning. The complete method, including first- to third-order MI expansion, balancing, and bias correction can be performed with the program ENTROPICAL. It can be obtained from the authors upon request and used with CHARMM and NAMD topologies and trajectories.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figures S1−S4, Appendices A−H, Tables S1 and S2, and complete ref 76 are provided. This information is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: knapp@chemie.fu-berlin.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by the collaborative research center SFB 765 project C1 from the Deutsche Forschungsgemeinschaft (DFG). We cordially thank Dr. Dennis J. Diestler for useful discussions.

■ ABBREVIATIONS

BAT, Bond−Angle−Torsion; GBMV, Generalized Born using Molecular Volume; KS, Kolmogorov−Sinai; MD, Molecular Dynamics; MC, Monte Carlo; MI, Mutual Information expansion; MIEn, expansion *n*th order MI expansion; PCA, Principal Component Analysis; QHA, Quasi-Harmonic Approximation; RW, Random Walk

■ REFERENCES

(1) Jaynes, E. T. In *The Maximum Entropy Formalism*; Levine, R., Tribus, M., Eds.; MIT Press: Cambridge, MA, 1979.
(2) Ben-Naim, A. *A Farewell To Entropy: Statistical thermodynamics based on information*; World Scientific Publishing Company: Singapore, 2008.
(3) Clausius, R. *Ann. Phys.* **1865**, *201*, 353−400.
(4) Shannon, C. E.; Weaver, W. *Bell Syst. Tech. J.* **1948**, *27*, 379−423.
(5) Salwiczek, M.; Samsonov, S.; Vagt, T.; Nyakatura, E.; Fleige, E.; Numata, J.; Cölfen, H.; Pisabarro, M. T.; Koksch, B. *Chem.—Eur. J.* **2009**, *15*, 7628−7636.
(6) Makhatadze, G. I.; Privalov, P. L. *Biophys. Chem.* **1994**, *51*, 291−309.
(7) Carlsson, J.; Åqvist, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5385−5395.
(8) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181−186.
(9) Polyansky, A. A.; Zubac, R.; Zagrovic, B. *Methods Mol. Biol.* **2012**, *819*, 327−353.
(10) Stern, O. *Ann. Phys.* **1916**, *356*, 237−260.
(11) Schlitter, J. *Chem. Phys. Lett.* **1993**, *215*, 617−621.
(12) Schäfer, H.; Mark, A. E.; van_Gunsteren, W. F. *J. Chem. Phys.* **2000**, *113*, 7809−7817.
(13) Andricioaei, I.; Karplus, M. *J. Chem. Phys.* **2001**, *115*, 6289−6292.
(14) Carlsson, J.; Åqvist, J. *J. Phys. Chem. B* **2005**, *109*, 6448−6456.
(15) Harris, S. A.; Laughton, C. A. *J. Phys.: Condens. Matter* **2007**, *19*, 076103.
(16) Rojas, O. L.; Levy, R. M.; Szabo, A. *J. Chem. Phys.* **1986**, *85*, 1037−1043.
(17) Numata, J.; Wan, M.; Knapp, E. W. *Genome Inform.* **2007**, *18*, 192.
(18) Baron, R.; Hünenberger, P. H.; McCammon, J. A. *J. Chem. Theory Comput.* **2009**, *5*, 3150−3160.

(19) Hensen, U.; Lange, O. F.; Grubmüller, H. *PLoS One* **2010**, *5*, e9179.

(20) Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; Wiley-Interscience: New York, 2001.

(21) Matsuda, H. *Phys. Rev. E* **2000**, *3*, 3096−3102.

(22) Noy, A.; Pérez, A.; Lankas, F.; Luque, F. J.; Orozco, M. *J. Mol. Biol.* **2004**, *343*, 627−638.

(23) Amadei, A.; Linssen, A.; Berendsen, H. *Proteins* **1993**, *17*, 412−425.

(24) Mukherjee, A. *J. Phys. Chem. Lett.* **2011**, *2*, 3021−3026.

(25) Chang, C.-E.; Chen, W.; Gilson, M. K. *J. Chem. Theory Comput.* **2005**, *1*, 1017−1028.

(26) Mendez, R.; Bastolla, U. *Phys. Rev. Lett.* **2010**, *104*, 228103.

(27) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325−332.

(28) Nola, A. D.; Berendsen, H. J. C.; Edholm, O. *Macromolecules* **1984**, *17*, 2044−2050.

(29) Harpole, K. W.; Sharp, K. A. *J. Phys. Chem. B* **2011**, *115*, 9461−9472.

(30) Darian, E.; Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. *J. Comput. Chem.* **2005**, *26*, 651−660.

(31) Wang, J.; Brüschweiler, R. *J. Chem. Theory Comput.* **2006**, *2*, 18−24.

(32) Li, D.-W.; Khanlarzadeh, M.; Wang, J.; Huo, S.; Brüschweiler, R. *J. Phys. Chem. B* **2007**, *111*, 13807−13813.

(33) Li, D. W.; Brüschweiler, R. *Phys. Rev. Lett.* **2009**, *102*, 118108.

(34) Li, D.-W.; Showalter, S. A.; Bruschweiler, R. *J. Phys. Chem. B* **2010**, *114*, 16036−16044.

(35) Suárez, E.; Díaz, N.; Suárez, D. *J. Chem. Theory Comput.* **2011**, *7*, 2638−2653.

(36) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2006**, *125*, 024905.

(37) Meirovitch, H. *J. Mol. Recognit.* **2010**, *23*, 153−172.

(38) Pearlman, D. A.; Rao, B. G. In *Encyclopedia of Computational Chemistry*; Schleyer, P v. R., Ed.; 1998; Vol. 2, pp 1036−1061.

(39) Peter, C.; Oostenbrink, C.; Dorp, A. v.; van Gunsteren, W. F. *J. Chem. Phys.* **2004**, *120*, 2652−2661.

(40) Chipot, C.; Pohorille, A. In *Free energy calculations: Theory and applications in chemistry and biology*; Springer: Berlin, 2007.

(41) Sinai, Y. G. *Dokl. Akad. Nauk S.S.S.R.* **1959**, *124*, 768−771.

(42) Dzugutov, M.; Aurell, E.; Vulpiani, A. *Phys. Rev. Lett.* **1998**, *81*, 1762−1765.

(43) Wissman, B. D.; McKay-Jones, L. C.; Binder, P. M. *Phys. Rev. E* **2011**, *84*, 046204.

(44) Ceriotti, M.; Tribello, G. A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A* **2011**, *108*, 12969−12970.

(45) Brandman, R.; Brandman, Y.; Pande, V. S. *PLoS One* **2012**, *7*, e29377.

(46) Watanabe, H.; Tanaka, S.; Okimoto, N.; Hasegawa, A.; Taiji, M.; Tanida, Y.; Mitsui, T.; Katsuyama, M.; Fujitani, H. *Chem—Biol. Inf. J.* **2010**, *10*, 32−45.

(47) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107.

(48) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. *J. Mol. Biol.* **2009**, *389*, 315−335.

(49) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47−L49.

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

(51) Herschbach, D. R.; Johnston, H. S.; Rapp, D. *J. Chem. Phys.* **1959**, *31*, 1652−1661.

(52) Gō, N.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 535−542.

(53) Chang, C.-E.; Potter, M. J.; Gilson, M. K. *J. Phys. Chem. B* **2003**, *107*, 1048−1055.

(54) Abagyan, R.; Totrov, M.; Kuznetsov, D. *J. Comput. Chem.* **1994**, *15*, 488−506.

(55) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300−313.

(56) Lazaridis, T.; Karplus, M. *Biophys. Chem.* **2003**, *100*, 367−395.

(57) Zhou, H.-X.; Gilson, M. K. *Chem. Rev.* **2009**, *109*, 4092−4107.

(58) Reinhard, F.; Grubmüller, H. *J. Chem. Phys.* **2007**, *126*, 014102.

(59) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(60) Paninski, L. *Neural Comput.* **2003**, *15*, 1191−1253.

(61) Schürmann, T. *J. Phys. A* **2004**, *37*, L295−L301.

(62) Steuer, R.; Kurths, J.; Daub, C. O.; Weise, J.; Selbig, J. *Bioinformatics* **2002**, *18* (Suppl. 2), S231−S240.

(63) Herzel, H.; Schmitt, A. O.; Ebeling, W. *Chaos Solitons Fractals* **1994**, *4*, 97−113.

(64) Jaynes, E. T.; Bretthorst, G. L. *Probability Theory: The Logic of Science*; 2003.

(65) Juneja, A.; Numata, J.; Nilsson, L.; Knapp, E. W. *J. Chem. Theory Comput.* **2010**, *6*, 1871−1883.

(66) Bussi, G.; Parrinello, M. *Phys. Rev. E* **2007**, *75*, 056707.

(67) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384−2393.

(68) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635−2643.

(69) Rosta, E.; Buchete, N.-V.; Hummer, G. *J. Chem. Theory Comput.* **2009**, *5*, 1393−1399.

(70) Chocholouová, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719−729.

(71) Haberthür, U.; Caflisch, A. *J. Comput. Chem.* **2007**, *29*, 701−715.

(72) Hamm, S. W. a. P. *J. Phys. Chem. B* **2000**, *104*, 11316−11320.

(73) Schweitzer-Stenner, R.; Eker, F.; Huang, Q.; Griebenow, K. *J. Am. Chem. Soc.* **2001**, *123*, 9628−9633.

(74) Mu, Y.; Kosov, D. S.; Stock, G. *J. Phys. Chem. B* **2003**, *107*, 5064−5073.

(75) Mu, Y.; Stock, G. *J. Phys. Chem. B* **2002**, *106*, 5294−5301.

(76) MacKerell, A. Jr. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(77) Numata, J.; Ebenhöh, O.; Knapp, E. W. *Genome Inform.* **2008**, *20*, 112−122.

(78) Penrose, R. *The Road to Reality: A Complete Guide to the Laws of the Universe*; Vintage Books: New York, 2005.

(79) Moddemeijer, R. *Signal Process.* **1989**, *16*, 233−248.

(80) Beirlant, J.; Dudewicz, E. J.; Györfi, L.; Meulen, E. C. v. d. *Int. J. Math. Stat. Sci.* **1997**, *6*, 17−39.

(81) Hnizdo, V.; Darian, E.; Federowicz, A.; Demchuk, E.; Li, S.; Singh, H. *J. Comput. Chem.* **2007**, *28*, 655−668.

(82) Kraskov, A.; Stögbauer, H.; Grassberger, P. *Phys. Rev. E* **2004**, *69*, 066138.

(83) Hnizdo, V.; Singh, H.; Misra, N.; Fedorowicz, A.; Demchuk, E. *Amer. J. Math. Manage. Sci.* **2003**, *23*, 301−321.

(84) Nilsson, M. *IEEE T. Inform. Theory* **2007**, *53*, 2330−2341.

(85) Bercher, J. F.; Vignat, C. *IEEE T. Signal Process.* **2000**, *48*, 1687−1694.

(86) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; Dover: Mineola, NY, 1955.