Supporting Information

Balanced and bias-corrected computation of conformational entropy differences for molecular trajectories

Jorge Numata[1] and Ernst-Walter Knapp[1]*

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

Table of contents

Figures and tables mentioned explicitly in the main text:

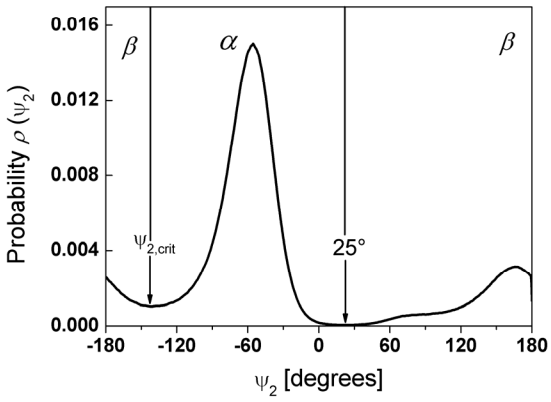

Figure S1: Probability density for the Ramachandran dihedral ψ_2 in simulation 8. The torsion angle ψ_2 (see Fig. 2 of main text) is used as order parameter dividing the conformers α and β. This circular variable delimits the conformers at two positions: $\psi_{2,crit}$ is computed as the region with minimum population near ψ_2=-140°, which varies according to the simulation conditions (See Table S2). The second cut position is fixed at ψ_2=25°, since it depends on the repulsive wing of the Lennard Jones potential and is identical for all 13 simulation conditions.

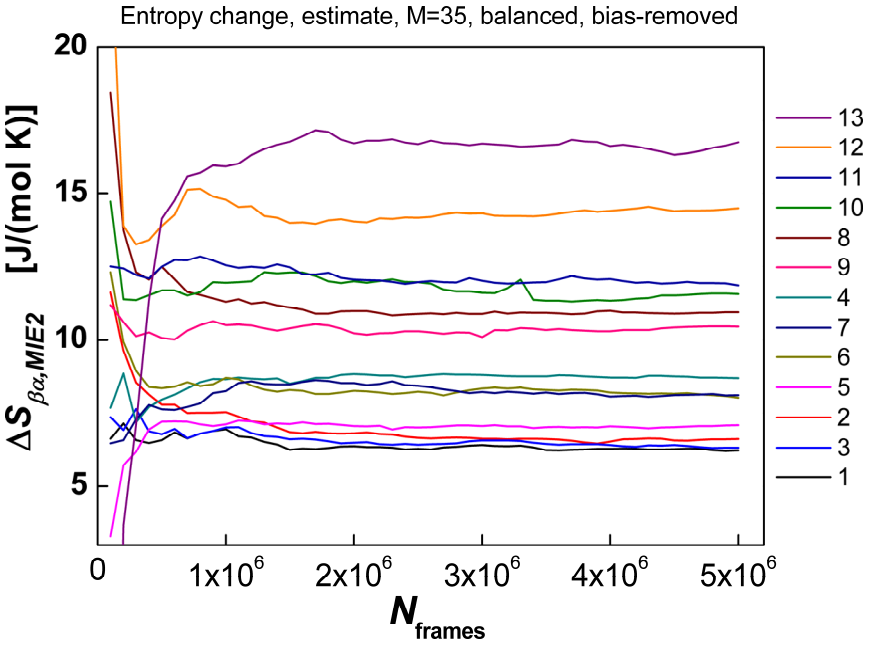

Figure S2: Convergence of the entropy estimates with the second order MI expansion, using balancing and bias correction. The frames are used in time order.

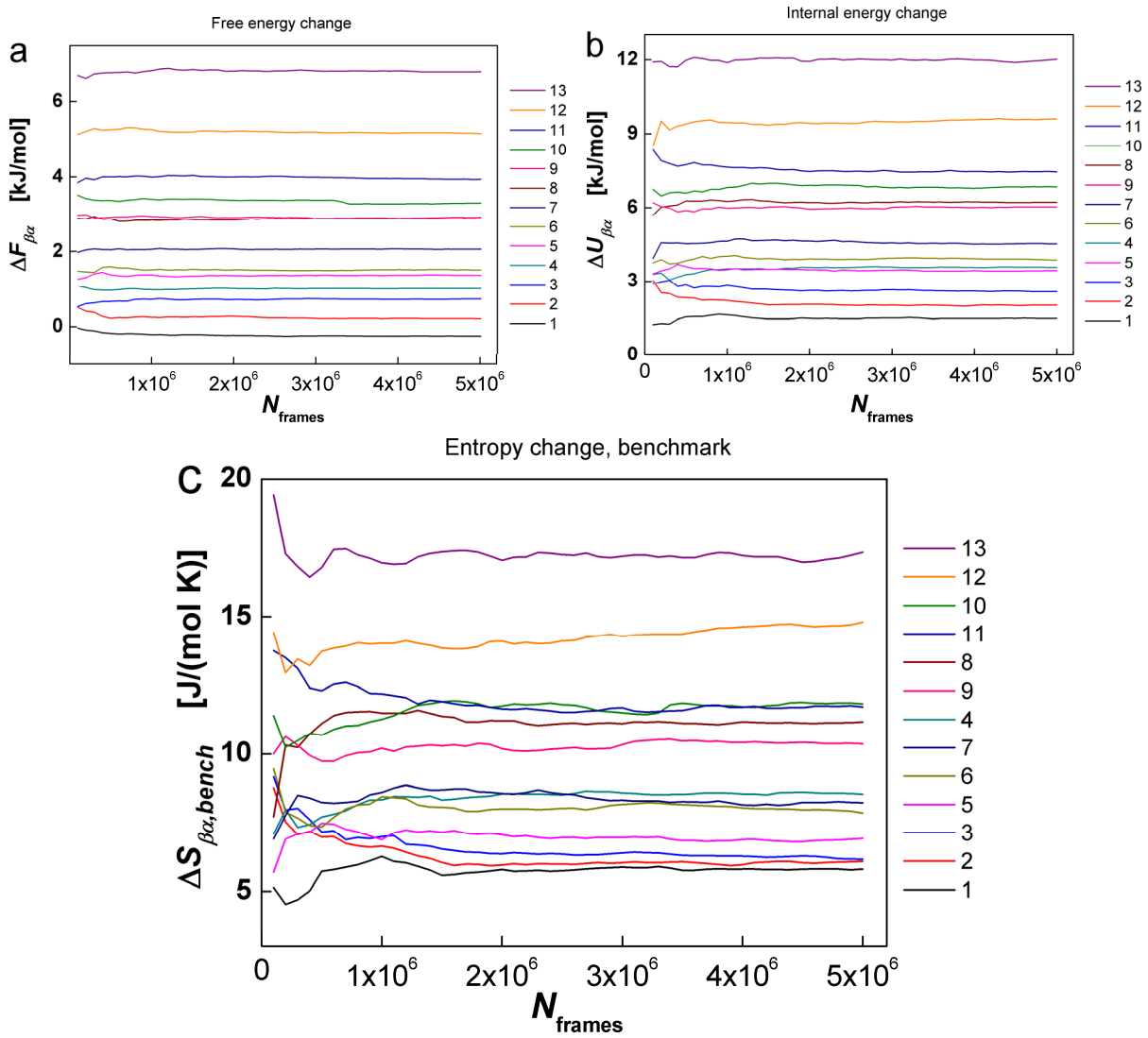

Figure S3: Convergence of the thermodynamic variables in the 1 μs trialanine simulation using $5x10^6$ frames. The frames are used in time order. **a:** Free energy change. **b:** Internal energy change. **c:** Entropy change benchmark.

Table S1: Normalized entropy contributions relative to the total conformational entropy change $\Delta s_{\beta\alpha,MIE2}$ in 2nd order MI expansion using both balancing and bias-correction. The data shown are based on MD simulations of 1 μs time for trialanine with simulation condition 8 (parameters: $\gamma_{H\phi}$ = 0.045 cal/(mol K Å^2) and ε_{attr} = 1.00). The symbol $\sum \Delta\hat{s}_{(1)} / \Delta s_{\beta\alpha,MIE2}$ denotes the normalized sum of entropy contributions from 1st order MI expansion with specified types of coordinates. The symbol $-\sum \Delta I_{(2)} / \Delta s_{\beta\alpha,MIE2}$ denotes the normalized 2nd order mutual information component for specified types of coordinate pairs. The coordinate types are denoted by B (bond lengths), A (bond angles) and T (torsions and phase angles). The contributions are ordered by absolute magnitude.

coordinates	normalized entropy contribution
T $\sum \Delta s_{(1)} / \Delta s_{\beta\alpha,MIE2}$	88.2%
T–T $-\sum \Delta I_{(2)} / \Delta s_{\beta\alpha,MIE2}$	11.5%
A $\sum \Delta s_{(1)} / \Delta s_{\beta\alpha,MIE2}$	2.32%
A–T $-\sum \Delta I_{(2)} / \Delta s_{\beta\alpha,MIE2}$	-0.84%
A–A $-\sum \Delta I_{(2)} / \Delta s_{\beta\alpha,MIE2}$	-0.82%
B–T $-\sum \Delta I_{(2)} / \Delta s_{\beta\alpha,MIE2}$	-0.12%
B $\sum \Delta s_{(1)} / \Delta s_{\beta\alpha,MIE2}$	-0.09%
B–B $-\sum \Delta I_{(2)} / \Delta s_{\beta\alpha,MIE2}$	-0.08%
B–A $-\sum \Delta I_{(2)} / \Delta s_{\beta\alpha,MIE2}$	-0.04%
Total	100%

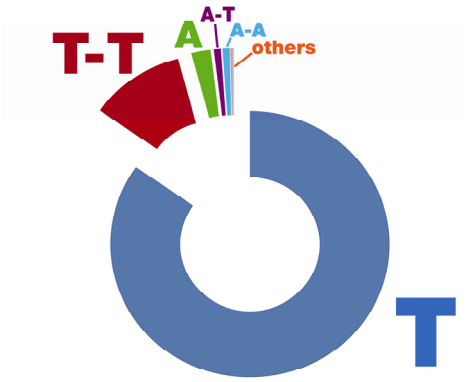

Figure S4: Graphic representation of the information in Table S1 above.

Appendix A: Configurational entropy of a macromolecule

To define entropy in the canonical ensemble of a macromolecule with N atoms, we start with the partition function of the conformer domain α

$$Q_\alpha = h^{-3N} \int d\mathbf{p}^N \int_{\Omega_\alpha} d\mathbf{r}^N \exp(-H_\alpha / k_B T). \tag{S1}$$

The Hamiltonian

$$H_\alpha = \sum_{n=1}^{N} p_n^2 / 2m_n + U_\alpha(\mathbf{r}^N). \tag{S2}$$

in eq (S1) involves kinetic and potential energy terms of the N atom macromolecule. The symbol Ω_α signifies the domain of configurations that identify the conformer α. The potential energy $U_\alpha(\mathbf{r}^N)$ is a function in the $3N$ Cartesian coordinates denoted by the $3N$-dimensional vector $\mathbf{r}^N$. The energy $U_\alpha(\mathbf{r}^N)$ is infinite outside of the domain α that defines the conformer. Integrating over the $3N$ momenta in eq (S1), we can write Q_α as

$$Q_\alpha = Z_\alpha \Big/ \prod_{n=1}^{N} \Lambda_n^3$$

in terms of the configuration integral

$$Z_\alpha = \int_{\Omega_\alpha} \exp[-U_\alpha(\mathbf{r}^N)/(k_B T)]\, d\mathbf{r}^N \tag{S3}$$

and the "momentum" contribution, expressed by the $3N$-fold product of the thermal de Broglie wavelengths

$$\Lambda_n = h / \sqrt{2\pi m_n k_B T}\,, \tag{S4}$$

where m_n is the mass of atom n, k_B is Boltzmann's constant, h is Planck's constant, and T is the absolute temperature.

The free energy F_α of the conformer in domain α is

$$F_\alpha = -k_B T \ln(Q_\alpha). \tag{S5}$$

The ensemble average of the internal energy is

$$\langle H_\alpha \rangle = k_B T^2 \left(\frac{\partial \ln Q_\alpha}{\partial T} \right)_{N,V} = \tfrac{3}{2} N k_B T + \langle U_\alpha \rangle \quad , \tag{S6}$$

where the ensemble average of the potential energy can be written as

$$\langle U_\alpha \rangle = \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) U_\alpha(\mathbf{r}^N) \tag{S7}$$

using the probability density function

$$P_\alpha(\mathbf{r}^N) = \exp[-U_\alpha(\mathbf{r}^N)/(k_B T)] \Big/ Z_\alpha\,. \tag{S8}$$

Rearranging eq (S8) and taking logarithms of both sides, we get

$$U_\alpha(\mathbf{r}^N) = -k_B T \ln[P_\alpha(\mathbf{r}^N) Z_\alpha]. \tag{S9}$$

Substitution of eq (S9) into eq (S7) gives

$$\langle U_\alpha \rangle = -k_B T \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) \ln[P_\alpha(\mathbf{r}^N) Z_\alpha] \tag{S10}$$

Now we define the configurational entropy of the conformer domain α as

$$S_\alpha = (<E_\alpha> - F_\alpha)/T. \tag{S11}$$

Using eq (S6) and (S10) we can rewrite the absolute configurational entropy, eq (S11), as

$$S_\alpha = \tfrac{3}{2} N k_B - k_B \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) \ln[P_\alpha(\mathbf{r}^N) \prod_{n=1}^{N} \Lambda_n^3], \tag{S12}$$

or

$$S_\alpha = \tfrac{3}{2} N k_B - 3 k_B \sum_{n=1}^{N} \ln \Lambda_n - k_B \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) \ln[P_\alpha(\mathbf{r}^N)]. \tag{S13}$$

In the configurational entropy difference, $\Delta S_{\alpha\beta} = S_\alpha - S_\beta$, the first two terms in eq (S13) cancel, if both entropies refer to the same temperature, yielding

$$\Delta S_{\alpha\beta} = k_B (\hat{s}_\alpha - \hat{s}_\beta), \tag{S14}$$

where the relative configurational entropy is defined by

$$\hat{s}_\delta = -\int_{\Omega_\delta} d\mathbf{r}^N P_\delta(\mathbf{r}^N) \ln[P_\delta(\mathbf{r}^N)], \; \delta = \alpha, \beta, \tag{S15}$$

which is analog to the Shannon differential entropy[1] for the probability density $P_\delta(\mathbf{r}^N)$.

Eq (S15) is the expression for a relative entropy for two reasons: (i) Its actual value varies by an additive constant term dependent on the length units (e.g. Ångström) used for the coordinates $\mathbf{r}^N$. (ii) It is a differential (continuous[2]) entropy, which may assume negative or positive values (see sec. 20 of Shannon[1] and appendix I of ref [3]). Conversely, the expression (S13) is an absolute entropy[4] because: (i) The length units used in the conformational integral cancel. (ii) Planck's constant *h* discretizes (quantizes) the phase space (cf. eq 7.12 of Landau & Lifshitz[5]). If entropy differences at different temperatures are evaluated, eq (S13) should be used. In this work we can use eq (S15), since we compute entropy differences at the same temperature.

Appendix B: Local spherical polar coordinates (BAT) coordinates.

To simplify the configurational integrals as for instance eqs (S3) or (S15) of appendix A, we introduce local spherical polar coordinates, also referred to as 'bond-angle-torsion' (BAT) coordinates[6-8]. This coordinate system is local because the frame of reference is shifted and rotated at each new bond to accommodate the molecular topology. These coordinates are defined by fixing the coordinate $\mathbf{r}_1$ of the terminal atom 1 of the macromolecule at the origin

of the coordinate system. All other coordinates refer to the bond vectors $\mathbf{b}_n$. The local spherical coordinates for the bond vector are $\mathbf{b}_n = (b_n, \theta_n, \varphi_n)$, $n = 2, 3, \ldots N$, (bond length b_n, inclination angle θ_n, azimuthal angle φ_n). We begin with bond vector $\mathbf{b}_2 = \mathbf{r}_2 - \mathbf{r}_1$ of the end atom 1, using the z- and x-axes from a lab frame as a reference for rotations θ_2 and φ_2. For the second bond vector $\mathbf{b}_3 = \mathbf{r}_3 - \mathbf{r}_2$, we use $\mathbf{b}_2$ as a reference for θ_2 but still need the x-axis from the lab frame as a reference for φ_3. For a linear molecule, the bond vectors are consecutively $\mathbf{b}_n = \mathbf{r}_n - \mathbf{r}_{n\text{-}1}$, $n = 4, 5, \ldots N$. For the local spherical coordinates of bond vector $\mathbf{b}_n$ we take atom position $\mathbf{r}_{n\text{-}1}$, as the coordinate origin, the preceding bond vector $\mathbf{b}_{n\text{-}1}$ as z-axis, and the unit vector parallel to the cross product $\mathbf{b}_{n-2}\times\mathbf{b}_{n-1}$ as x-axis. In a non-linear, branched molecule, we use for all bond vectors following a branch point (atom with more than two covalent bonds) the bond vector of the preceding two bonds as reference for z- and x-axes. Independently of the degree of branching, a molecule with N atoms and no ring structure possesses N–1 covalent bonds. Each ring introduces an additional bond. To avoid overcompleteness, one covalent bond in each ring is ignored, which automatically transforms the molecular topology back to a branched structure. Thus, together with the coordinates r_1 of the initial atom 1 a complete set of 3N BAT coordinates (Figure S5) is obtained for an N atom molecule. These BAT coordinates are collected in the 3N-dimensional supervector

$$\vec{\mathbf{b}} = (\mathbf{r}_1, \mathbf{b}_2, \mathbf{b}_3, \ldots \mathbf{b}_{N\text{-}1}, \mathbf{b}_N), \text{ with } \mathbf{b}_n = (b_n, \theta_n, \varphi_n), \; n = 2, 3, \ldots N. \qquad \text{(S16)}$$

The potential energy function U_α is independent of position and orientation of the solute in the solvent. Therefore, we can separate contributions of those degrees of freedom and perform the corresponding integrations in configurational integrals as for instance eqs (S3) or (S15) of appendix A in closed form using BAT coordinates[6,7,9,10]. The integration over $\mathbf{r}_1$ in configuration integrals like eq (S15) of the appendix A can be performed directly, yielding as a result the volume V available to solvent and solute together.

Rotating the first bond vector $\mathbf{b}_2 = \mathbf{r}_2 - \mathbf{r}_1$ together with the whole solute molecule is described by varying the polar coordinate angles (θ_2, φ_2). Similarly the whole molecule can be rotated about the bond $\mathbf{b}_2$ described by the azimuthal angle φ_3. The potential energy function U_α of the solute does not depend on the orientation of the whole solute molecule.

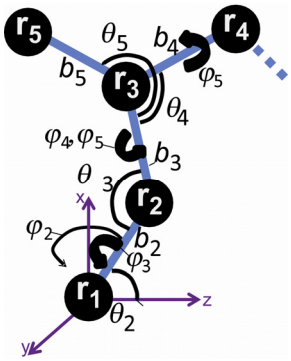

Figure S5: Local spherical polar coordinates (BAT coordinates) of a branched molecule. The lab frame (purple) is the initial reference for external rotations θ_2, φ_2 and φ_3. Further up the chain, the frame of reference is local and defined by the chemical bonds.

Hence, the integrations over θ_2, φ_2 and φ_3 in the configuration integrals like eq (S15) in appendix A can be performed directly to give the factor $8\pi^2$. As a result we have for instance for the (configurational) state sum, eq (S3)

$$Z_\alpha = V 8\pi^2 \int db_2\, b_2^2 \int db_3\, b_3^2 \int_0^{2\pi} d\theta_3 \sin\theta_3 \prod_{n=4}^{N} \int d^{(3)}\mathbf{b}_n \exp[-U_\alpha(\vec{\mathbf{b}}')/k_B T], \tag{S17}$$

with the vector differential of the local spherical polar coordinates

$$d^{(3)}\mathbf{b}_n = b_n^2 db_n \sin\theta_n\, d\theta_n\, d\varphi_n \tag{S18}$$

and the $3N$–6 BAT variables combined in the ($3N$-6)-dimensional vector

$$\vec{\mathbf{b}}' = (b_2, b_3, \theta_3, \mathbf{b}_4, \mathbf{b}_5, \ldots . \mathbf{b}_{N-1}, \mathbf{b}_N). \tag{S19}$$

In analogy to eq (S19) we also define the vector-valued differential form

$$d^{(3N-6)}\mathbf{b}' = b_2^2 db_2 \times b_3^2 db_3 \times \sin\theta_3\, d\theta_3 \times \prod_{n=4}^{N} d^{(3)}\mathbf{b}_n . \tag{S20}$$

The notation of $\vec{\mathbf{b}}'$ and $d^{(3N-6)}\mathbf{b}'$ is used in the main text. Thus, we can write now the (configurational) state sum, eq (S17) in the compact form

$$\frac{Z_\alpha}{V 8\pi^2} = \int d^{(3N-6)}\mathbf{b}' \exp[-U_\alpha(\vec{\mathbf{b}}')/k_B T] \equiv \check{z}_\alpha \ , \tag{S21}$$

where $\check{z}_\alpha$ of the second part of eq (S21) is now the conformational state sum exclusive of the position and orientations of the solute. We can now define the reduced conformational probability distribution

$$\rho_\alpha(\vec{\mathbf{b}}') = \exp[-U_\alpha(\vec{\mathbf{b}}')/k_B T]/\check{z}_\alpha , \tag{S22}$$

and the reduced relative conformational entropy, which neglects translation and orientation of the considered macromolecule, is

$$s_\delta = -\int_{\Omega_\delta} d^{(3N-6)}\mathbf{b}'\, \rho_\delta(\vec{\mathbf{b}}') \ln[\rho_\delta(\vec{\mathbf{b}}')], \ \delta = \alpha, \beta. \tag{S23}$$

Hence, the entropy differences of a molecular system can be expressed by the dimensionless configurational entropies as it is done in eq (S14) of Appendix A or alternatively by the reduced dimensionless conformational entropies, eq (S23) according to

$$\Delta S_{\alpha\beta} = k_B\,(\hat{s}_\alpha - \hat{s}_\beta) \equiv k_B\,(s_\alpha - s_\beta)\,. \qquad \text{(S24)}$$

The above expression for the conformational entropy difference will be used in the main text.

Appendix C: Automated selection of BAT coordinates

For a given molecular topology, a set of non-redundant internal BAT coordinates is constructed using the procedure described above. In practice, this translates into a tree algorithm also described by Gilson et al[8]. The PERL implementation of the BAT tree algorithm by Thomas Steinbrecher[11], which in turn uses ptraj[12], is adapted and modified to use Charmm/NAMD trajectories.

Phase angles

Azimuthal (torsion) angles, which are defined through three shared atoms, tend to show highly correlated motions. The hydrogens of a methyl group display such behavior, for which we define a master torsion angle, say φ_i, and two phase angles[13] ϕ_k. Generally, if the torsion angles φ_i and φ_j have three atoms in common, we keep φ_i and substitute φ_j by the phase angle

$$\phi_j = \varphi_j - \varphi_i. \qquad \text{(S25)}$$

This transformation has a unit Jacobian and preserves a complete geometric description of the molecule. In Figure S5, the atoms with coordinates $\mathbf{r}_4$ and $\mathbf{r}_5$ give rise to torsions φ_4 and φ_5. According to eq (S25) we substitute torsion φ_5 by the phase angle $\phi_5 = \varphi_5 - \varphi_4$. Such phase angles[13] have narrower distributions than torsion angles.

In our algorithm, main chain torsions (of the polypeptide backbone) are kept as full torsions, and the ones defined at branches (describing side chain orientations) are converted into phase angles. Both phase angles and the ability to define main chain atom types are implemented in our modified version of the BAT tree algorithm.

Continuity Maximization for Torsions

In contrast to molecular bond angles, torsion angles can vary over the whole angular regime from 0 to 2π, such that the 2π periodicity must be considered to avoid discontinuities. We apply a 'continuity maximization' algorithm to deal with this problem. For each torsion angle, its one-dimensional probability distribution is discretized with a large number of histogram bins (say 1000), many more than will finally be used for entropy computations. In this

histogram, the longest continuous stretch of empty bins is detected. The end points of the angular interval for the histogram used to evaluate the entropies are placed such that they exclude this regime. If no histogram bin is empty, the original angular distribution is kept and used for the entropy evaluation. For a torsional coordinate with values that cover the whole 2π span, the choice of the end points formally has no effect, and numerically it would only have a vanishing one. However, for a torsional coordinate that covers only part of the 2π span, this algorithm avoids considering a large number of empty (unused) histogram bins.

Appendix D: Underestimating entropy for finite samples and its correction

Entropy is underestimated (biased) when using finite samples[14,15]. To demonstrate the need for bias correction, we show here for a toy example that $s_{\mathrm{est}} \leq s_{\mathrm{bench}}$ and how eq (13) of the main text functions.

Take a system consisting of a single conformer characterized by M = 2 histogram bins. Assume that the actual probability density (source probability) of occupying either bin is $\rho_{\mathrm{i}} = ½$ (uniform distribution) yielding

$$s_{\mathrm{bench}} = -\Sigma\, \rho_{\mathrm{i}} \ln \rho_{\mathrm{i}} = \ln(2) \approx 0.6931\ . \qquad \text{(S26)}$$

The sample entropy estimate can be calculated from the sample frequency p_{i} as

$$s_{\mathrm{est}} = -\Sigma\, p_{\mathrm{i}} \ln p_{\mathrm{i}}\ . \qquad \text{(S27)}$$

Let us assume that a simulation consists of $N_{\mathrm{frames}} = 4$. If we repeat the simulation and the entropy estimation several times, we get results as shown below in Figure S6.

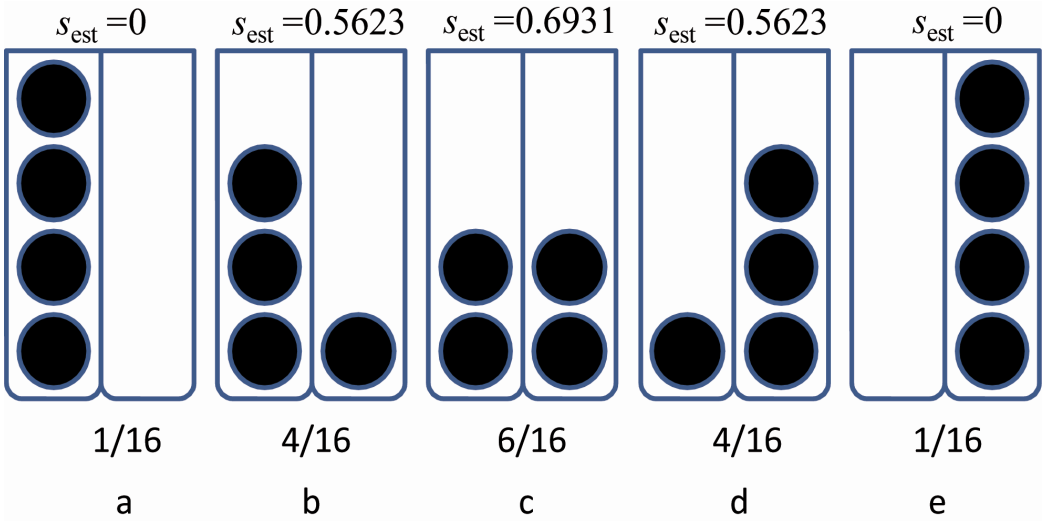

Figure S6: The 5 possible outcomes of the simulation with two discrete states (M=2 histogram bins). Above each letter is the binomial probability of obtaining this outcome from the simulation. The resulting entropy estimates are given above the histograms.

In Figure S6, we show the (binomially distributed) probabilities of obtaining the five possible distributions (a-e) of the histograms from the simulation. We see that only for 6/16=37.5% of the simulation results (case c) we get the correct value of entropy. For the remaining 10/16=62.5% (cases a,b,d,e) we observe underestimated entropy values. For cases b and d in Figure S6, the bias correction (eq (13) of the main text) essentially compensates for the underestimation yielding $\hat{s}_{est} = s_{est} + \frac{\hat{M}_i - 1}{2N_{frames}} = 0.5623 + 1/8 = 0.6873$. In realistic applications the number of frames and histogram bins (N_{frames} and M, respectively) is much larger such that the bias correction is smaller and works more precisely.

Appendix E: Convergence of the entropy estimates for trialanine

In this section, we analyze the convergence properties of entropy and entropy difference estimates. For the sake of clarity, only the final converged benchmark values of the entropy difference $\Delta S_{\beta\alpha,bench}$ (eq (15) of the main text) are shown as dashed lines in Figure S7 and Figure S8. The convergence properties of the entropy difference for the benchmark values are treated separately in the next section.

In the following discussion we will use the example of trialanine with simulation condition 8 (parameters $\gamma_{H\phi} = 0.045$ cal/(mol K Å^2) and $\varepsilon_{attr} = 1.00$) using the 2nd order MIE expansion (MIE2). The individual entropies S_δ are not fully converged, whether unbalanced (Figure S7a, b) or balanced data are used (Figure S8a, b), and independently of whether bias correction a→b is applied or not. As matter of fact, balancing will slow down the convergence of entropies of the majority conformer S_α. However, our main focus is to compute entropy differences $\Delta S_{\beta\alpha}$. There, we observe a beneficial effect of balancing. Without balancing, the entropy difference $\Delta S_{\beta\alpha}$ diverges (Figure S7c, d), while with balancing the entropy difference converges (Figure S8c, d). This is due to the fact that after balancing the individual conformer entropies ($S_{\alpha,\text{MIE2}}$ and $S_{\beta,\text{MIE2}}$) possess similar systematic errors, which cancel in the entropy difference $\Delta S_{\beta\alpha,MIE2}$. The bias correction method c→d provides an additional beneficial fine-tuning for the entropy difference.

Importance of choosing frames at random in the balancing method

In the balancing method, only a subset of the frames of the majority conformer is used. It is important to choose those frames at random[16] instead of simply taking a contiguous subset of the trajectory, since that results in a nonequivalent exploration of the phase space. While the convergence of the individual entropies $S_{\delta,MIE2}$ using time order or random order is indistinguishable to the eye due to the large magnitude of the individual entropies (Figure

S8a, b), the consequences for the convergence of the entropy difference $\Delta S_{\beta\alpha,MIE2}$ are clearly visible. In Figure S8c, d we see that the convergence of the entropy difference is accelerated by choosing the frames at random. The reason for this does not lie in the numerical properties of the bias of the histogram method, but rather in the fact that the randomly ordered conformations result in a more complete phase space exploration at a given number of frames. Choosing the frames at random is important for MD and MC simulations, where the frames are correlated with each other. The convergence behavior of $\Delta S_{\beta\alpha,MIE2}$ for all 13 simulation conditions using balancing and bias correction is presented in Figure S2.

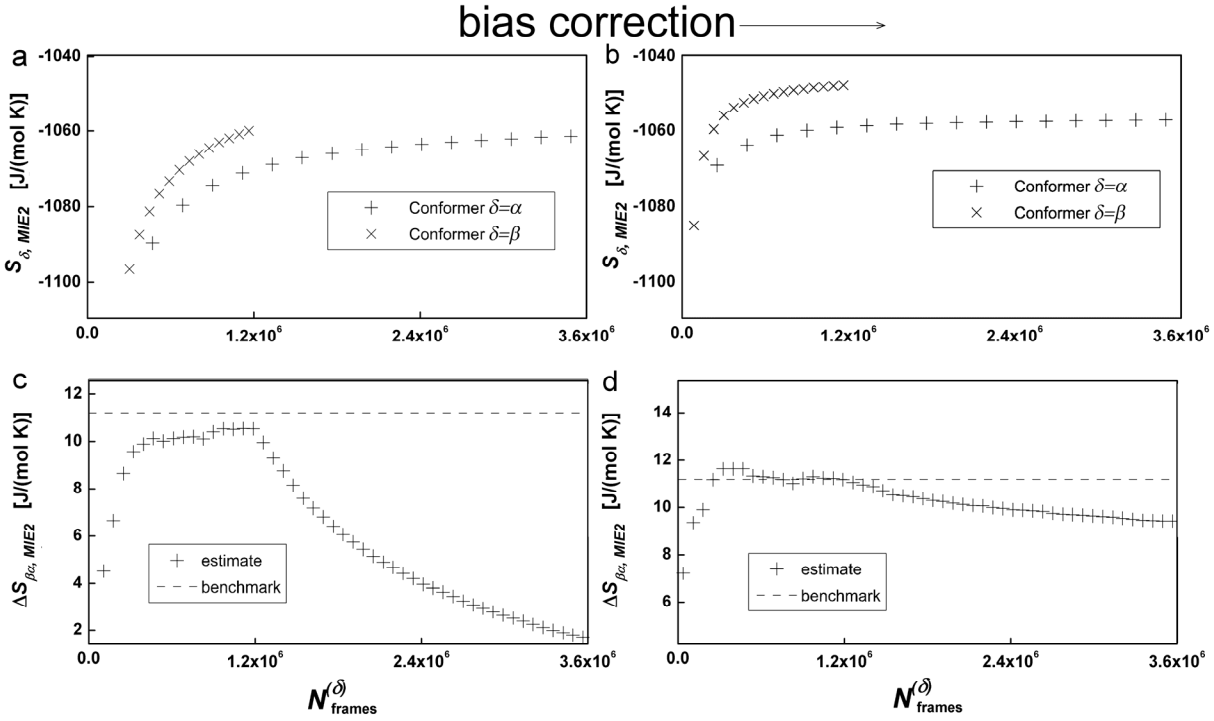

Figure S7: Unbalanced number of frames, 2nd order estimator (MIE2) used to plot individual (relative) entropies S_α, S_β and the entropy difference $\Delta S_{\alpha\beta}$. Convergence of the entropy estimates versus number of frames used for the trialanine simulation condition 8 (parameters: $\gamma_{H\phi}$ = 0.045 cal/(mol K Å^2) and ε_{attr} = 1.00). Frames are used in time order. The abscissa denotes with $N^{(\delta)}_{frames}$ the effective number of frames used for $\delta=\alpha, \beta$. This differs from N_{frames} used elsewhere, which refers to all frames of the simulation. The dashed line marks the final benchmark value. **a:** Individual conformer entropies without bias correction. **b:** Individual conformer entropies using bias correction. **c:** Entropy difference without bias correction. **d:** Entropy difference using bias correction.

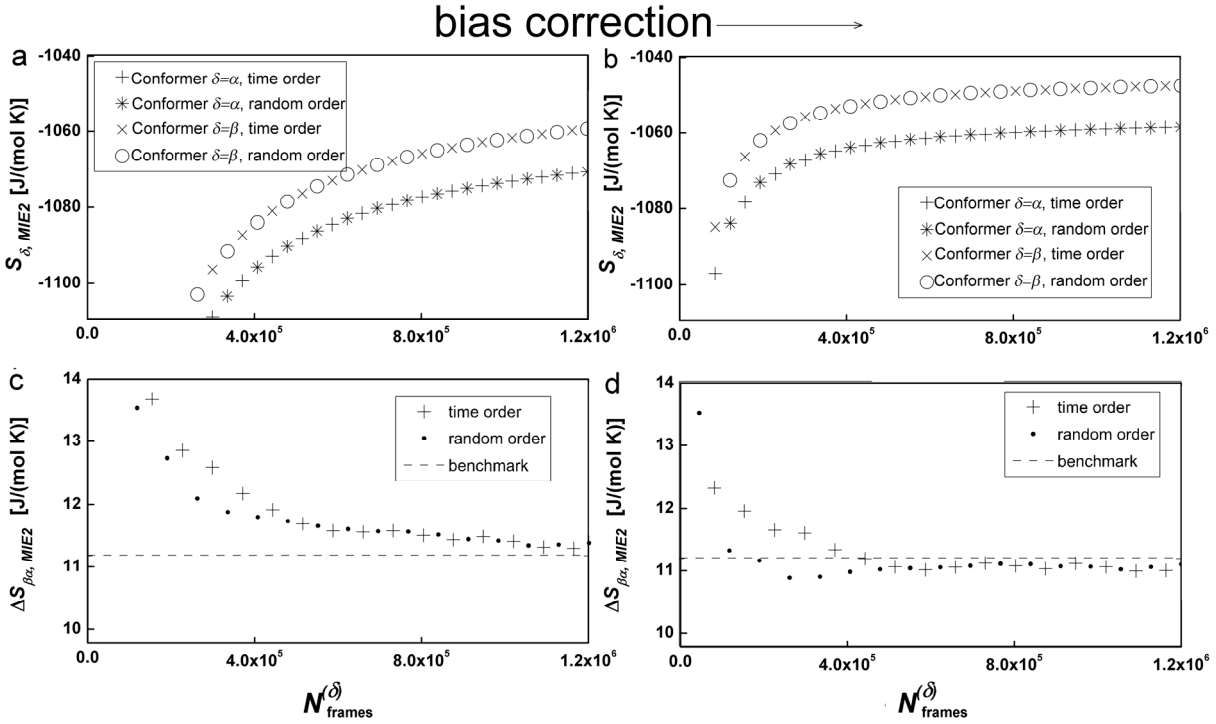

Figure S8: Balanced number of frames, 2nd order estimator (MIE2) used to plot individual (relative) entropies S_α, S_β and the entropy difference $\Delta S_{\alpha\beta}$. Convergence of the entropy estimates versus number of frames used for the trialanine simulation condition 8 (parameters: $\gamma_{H\phi} = 0.045$ cal/(mol K Å^2) and $\varepsilon_{attr} = 1.00$). Frames are used in time and random order as indicated in the figure. The abscissa denotes with $N^{(\delta)}_{frames}$ the effective number of frames used, which is identical for δ=α, β when applying the balancing method. This differs from N_{frames} used elsewhere, which refers to all frames of the simulation. The dashed line marks the final benchmark value. **a:** Individual conformer entropies without bias correction. **b:** Individual conformer entropies using bias correction. **c:** Entropy difference without bias correction. **d:** Entropy difference using bias correction.

Appendix F: Convergence of benchmarks and clustering of conformers for trialanine

In Figure S3, we observe that the free energy difference $\Delta F_{\beta\alpha}$ converges the fastest among thermodynamic variables. The energy difference $\Delta U_{\beta\alpha}$ is slower in convergence, and $\Delta S_{\beta\alpha,bench}$, being calculated as a difference, is the slowest one to converge. The simulation ID is assigned by ascending values of $\Delta F_{\beta\alpha}$. The order in the values of the energy difference $\Delta U_{\beta\alpha}$ and the entropy difference $\Delta S_{\beta\alpha,bench}$ differs somewhat with respect to the ascending $\Delta F_{\beta\alpha}$ order (see bars on the right of Figure S3). The free energy $\Delta F_{\beta\alpha}$ is the result of the interplay of energetic and entropic contributions, which are related but not identical. A given potential energy surface (which varies among simulation conditions 1 to 13) determines which microstates are accessible to each conformer at temperature T. The energetic component results from the microstates' average energy (average "funnel depth"), and the entropic component from their multiplicity (average "funnel width"), adapting the concepts of

Wolynes[17] to our system. Thus, there is no a priori reason to believe that the ascending order of the values of energy, entropy and free energy differences should be identical. See color labels in Figure S3a, b and c.

Krivov et al. simulated tetraalanine[18] with the PARAM19 force field of CHARMM[19] and the ACS[20] implicit solvent model. To evaluate entropy, the tetraalanine conformers were clustered using not a geometric, but a kinetic criterion. The simulation was done both with Langevin dynamics and with a method that confines and explores conformations in a given conformer basin. They also find that the extended β conformer has higher energy but is stabilized by entropy. The entropy difference between the helical α and extended β conformations of tetraalanine was found to be $\Delta S_{\beta\alpha} = 20.4$ J/(mol K)[18], comparable to our results for trialanine, which range from about 5.8 to 17.3 J/(mol K) depending on the simulation conditions.

Table S2: Converged values of the thermodynamic variables for trialanine simulation with 13 different conditions.

Simulation condition	ε_{attr} [dimless]	$\gamma_{H\phi}$ [cal/(mol K Å^2)]	$\Delta F_{\beta\alpha}$ [kJ/mol]	$\Delta U_{\beta\alpha}$ [kJ/mol]	$\Delta S_{\beta\alpha,bench}$ [J/(mol K)]	$\psi_{2,crit}$ [degrees]
1	0.00	0.045	-0.25	1.49	5.80	-134.5
2	0.00	0.025	0.21	2.03	6.08	-138.5
3	0.00	0.000	0.73	2.58	6.15	-140.5
4	0.50	0.045	1.01	3.57	8.54	-135.5
5	0.25	0.000	1.37	3.44	6.92	-139.5
6	0.50	0.025	1.51	3.86	7.85	-139.5
7	0.50	0.000	2.06	4.53	8.23	-141.5
8	1.00	0.045	2.87	6.22	11.17	-140.5
9	0.75	0.000	2.92	6.03	10.37	-140.5
10	1.00	0.025	3.30	6.85	11.83	-141.5
11	1.00	0.000	3.93	7.45	11.72	-141.5
12	1.25	0.000	5.16	9.60	14.79	-144.5
13	1.50	0.000	6.81	12.01	17.34	-144.5

In Table S2 the final asymptotic values for the thermodynamic variables are provided. For each simulation condition, the numerical values for the hydrophobic "surface tension" term $\gamma_{H\phi}$ and the $1/r^6$ attractive Lennard Jones potential scaling factor ε_{attr} used in each simulation can be read. Also, the critical value of ψ_2, a Ramachandran dihedral angle of the middle residue of trialanine[21,22], which we use as order parameter, is provided. $\psi_{2,crit}$ is the value of that angle at which the ensemble population is the lowest, and used to divide the conformers α and β. The second value at which the circular variable ψ_2 is cut is fixed at 25°, a

value identical for all simulations. It is the consequence of the repulsive wing of the Lennard Jones potential (identical in all 13 simulations) and physically interpretable as a steric constraint. See Figure S1 for an example of the probability distribution $\rho(\psi_2)$ corresponding to simulation ID 8.

Appendix G: Generation of conformations of the three atom molecule

A simple Monte Carlo (MC) procedure is used to generate 5×10^7 frames of the free, unrestricted 3-atom molecule by a Random Walk, (RW). To generate each frame in Cartesian coordinates, we proceed as follows:

1. Place the first atom at the origin: $\mathbf{r}_1 = (0, 0, 0)$
2. Place atom 2 at $\mathbf{r}_2 = \mathbf{r}_1 + \mathbf{b}_2$
3. Place atom 3 at $\mathbf{r}_3 = \mathbf{r}_2 + \mathbf{b}_3$.

Here, $\mathbf{b}_n$ is a vector whose tip is uniformly randomly distributed on a sphere of radius b, where b is the fixed bond length. This is accomplished through an algorithm due to Marsaglia[23], which is an optimized version of von Neumann's algorithm[24]. The independent, identically distributed pseudorandom numbers required by Marsaglia's algorithm[23] are generated by the pseudorandom number generator Taus088 due to L'Ecuyer[25].

The ensemble of free conformations is now subject to restriction by a hard wall described by, eq (17) of the main text with $\varepsilon = 0.612$. The constant ε is chosen arbitrarily to provide a positive curvature and divide the conformers unevenly. The conformer regime α comprises the frames where all atoms are above z_{wall}. The rest, where any or all atoms are below z_{wall}, is denominated β.

Appendix H: Trialanine simulations: Detailed results for 1st, 2nd and 3rd order MI expansion

Entropy estimates using all BAT coordinates

In the main text, we present the 2nd order MI expansion (MIE2) estimators for the entropy differences $\Delta S_{\beta\alpha}$ between the two the conformers (α, β) of the trialanine model. The MIE2 results are chosen, since the entropy estimates are well converged and agree best with the benchmark. Here, we present the results for the 1st and 3rd order MI expansion, and more detailed results for the 2nd order MI expansion. In Figure S9, Figure S10 and Figure S11, the four panels demonstrate the effect of using either or both correction methods. Upper left: unbalanced, biased; upper right: unbalanced, bias-corrected; lower left: balanced, biased; lower right: balanced, bias-corrected. In these figures, we consider all 96 BAT degrees of freedom of the trialanine model: bonds, angles, torsions and if necessary phase angles that

replace corresponding torsion angles. The lower right panel (d) presents the best results using both methods: balancing and bias correction. Also shown in each panel is the average and standard deviation of the estimate-to-benchmark ratio $\Delta S_{\beta\alpha,MIE1} / \Delta S_{\beta\alpha,bench}$. Average and standard deviation for this ratio are calculated over all 13 simulation conditions and all five of histogram schemes with different numbers of bins M.

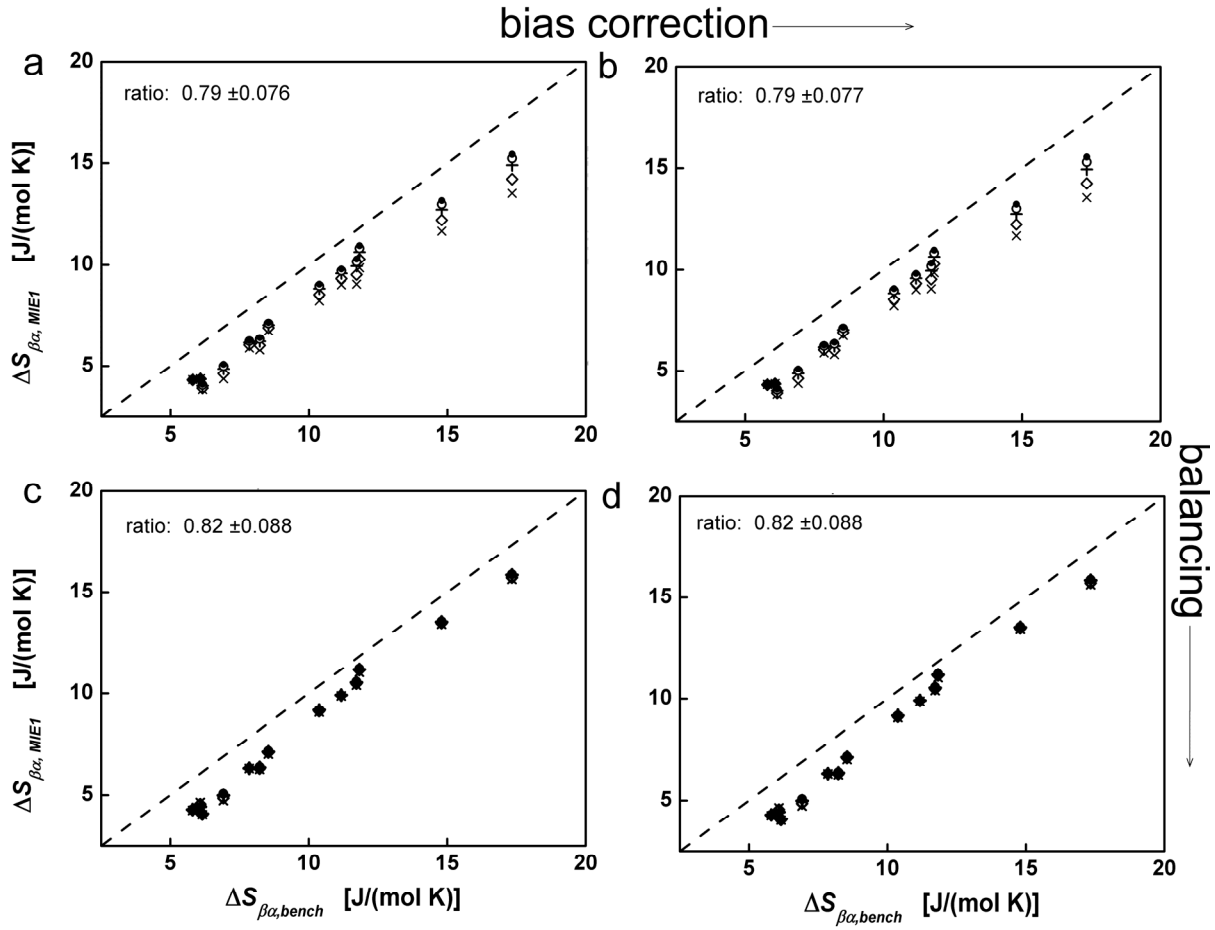

Figure S9: Results with first order MI expansion (MIE1). Entropy difference estimates $\Delta S_{\beta\alpha}$ (abscissa) between the two conformers β and α for the trialanine model are compared with benchmark entropies (ordinate). All 96 BAT degrees of freedom are used. The symbols stand for the number of histogram bins used: × M=20; ◇ M=25; + M=35; ○ M=50; ● M=100. The arrows show application of the two correction methods: none (**a**), either (**b**, **c**) or both (**d**). Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1} / \Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0.

The first order MI expansion (MIE1) in Figure S9 is well converged. Nevertheless, the converged value does not agree well with the benchmarks, as can be seen by the deviation of the computed results from the dashed diagonal line representing the perfect agreement. In MIE1, the individual entropies are estimated as the sum of the marginal entropies (first term of eq (10) of the main text). Compensating the bias according to eq (13) of the main text yields for MIE1 a small correction only, which results in no noticeable change from a→b and

c→d in Figure S9. The size of the correction is small because in the 1st order MI expansion the number of histogram bins is small such that the bins are well filled and exhibit small fluctuations. This contrasts with MIE2 and MIE3 having quadratically and cubically as many histogram bins, respectively. Thus, for MIE1 the major correction comes from balancing (a→c and b→d). The balancing method narrows the spread between the estimators for the different number of histogram bins M (different symbols), but as expected cannot correct for the lack of correlation in MIE1.

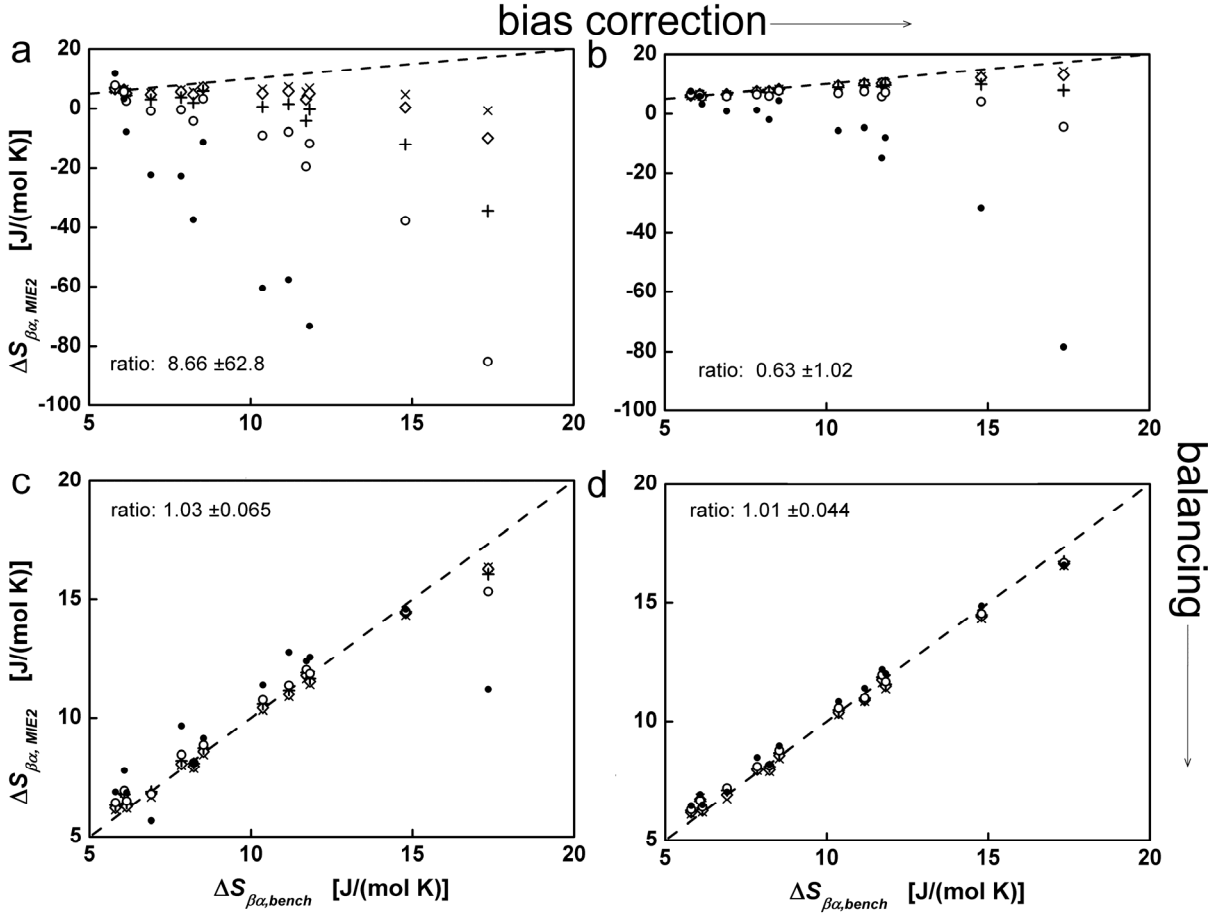

Figure S10: Results with second order MI expansion (MIE2). Entropy difference estimates $\Delta S_{\beta\alpha}$ (abscissa) between the two conformers β and α for the trialanine model are compared with benchmark entropies (ordinate). All 96 BAT degrees of freedom are used. The symbols stand for the number of histogram bins used: × M=20; ◇ M=25; + M=35; ○ M=50; ● M=100. The arrows show application of the two error correction methods: none (**a**), either (**b**, **c**) or both (**d**). Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1}$ / $\Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0.

The results for MIE2 using all 96 BAT coordinates are commented extensively in section 4.2 of the main text. In Figure S10, we see a large and beneficial effect of the balancing method (a→c and b→d). The bias correction acts to fine-tune the entropy

differences in c→d. It becomes evident that balancing and bias correction act synergistically to improve the accuracy of the entropy estimates. If we separate the contributions of the different types of coordinates [bonds (B), angles (A), torsions and phase angles (T)] in the 1st and 2nd order MI expansions, we realize that the coordinates of type T have the largest influence, 99.7% (as shown in Table S1 and Figure S4) when taking the trialanine simulation condition 8 as representative [parameters: $\gamma_{H\phi}$ = 0.045 cal/(mol K Å^2) and ε_{attr} = 1.00]. The contribution of the coordinates of types B and A, and their correlations with T approximately cancel. In particular, the influence of B in the 1st and 2nd order MI expansion has a vanishing influence of 0.17% on the final result.

The MIE3 entropy difference estimates in Figure S11 show poor agreement with the benchmarks. There is definite improvement by using bias correction and balancing, but even Figure S11d where both methods have been used is far from optimal. From this we conclude that we need more frames than the 5x10^6 frames used here to obtain well converged MIE3 estimates.

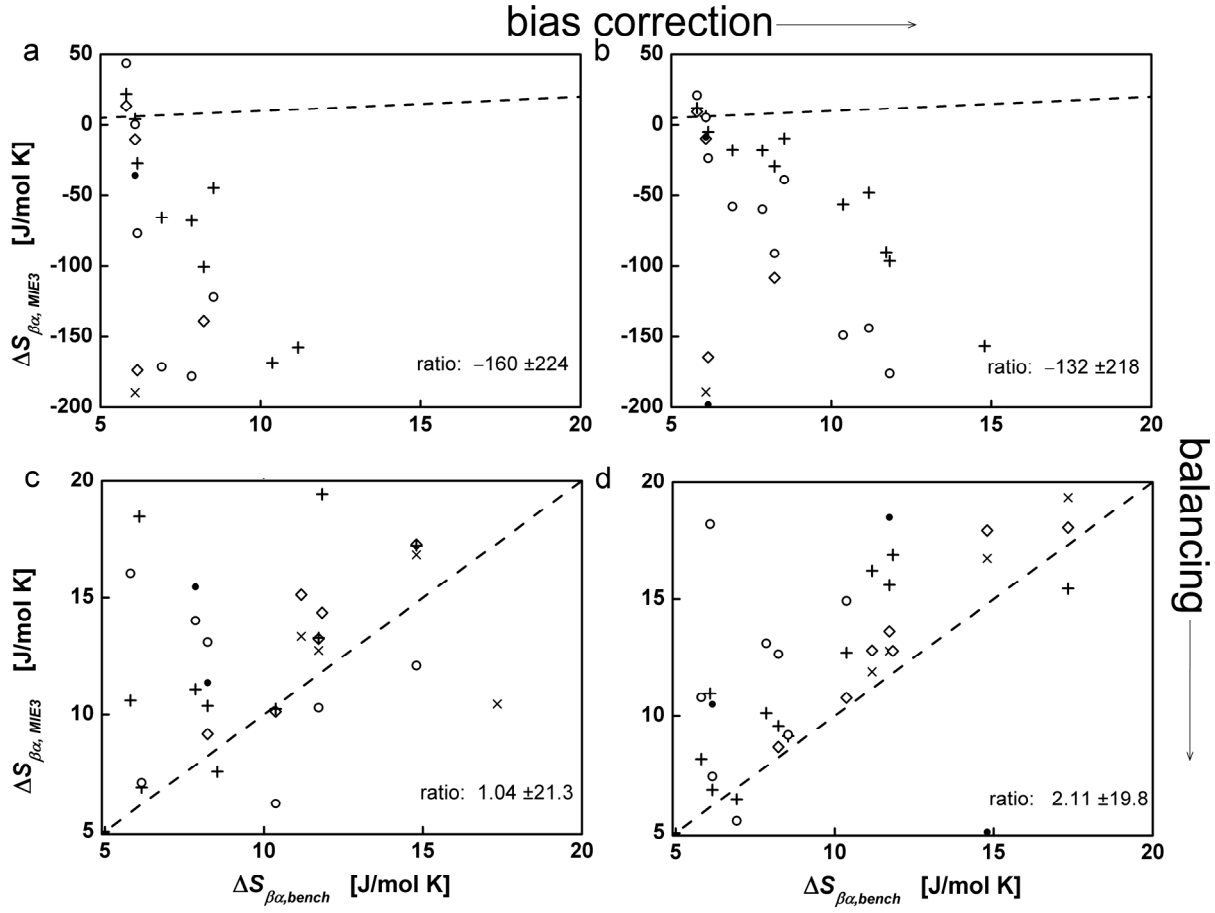

Figure S11: Results with third order MI expansion (MIE3). Entropy difference estimates $\Delta S_{\beta\alpha}$ (abscissa) between the two conformers β and α for the trialanine model are compared with benchmark entropies (ordinate). All 96 BAT degrees of freedom are used. The symbols stand for the number of histogram bins used: × M=20; ◇

M=25; + M=35; O M=50; ● M=100. The arrows show application of the corection methods: none (**a**), either (**b**, **c**) or both (**d**). Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1} / \Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0.

Entropy estimates using only soft degrees of freedom

Recent work from Brüschweiler et al.[26] suggested employing only the main torsion angles ('soft degrees of freedom') and neglecting conformational contributions from 'hard degrees of freedom', including phase angles. In their work, Brüschweiler et al. calculate only the momenta contribution (cf. second term of eq (S13)) for the hard degrees of freedom, which is required because the entropy difference is estimated for conformers at two different temperatures (T = 380 K and T = 270 K). They assume that the Jacobian determinant (which only arises from hard degrees of freedom) will be conformation-independent and thus cancel. The momenta contribution and the constant Jacobian are embodied into eq (2) of ref [26]. Using only torsions as soft degrees of freedom resulted in estimate-to-benchmark ratios between 0.87 and 0.96 when testing entropy differences of dipeptide conformers at two different temperatures (see Table I, last column, of Brüschweiler et al.[26]).

Furthermore, Brüschweiler et al. studied the conformational entropy change between the bound and unbound conformers of a protein[27]. They found that linear correlations (as obtained from the covariance matrix[28]) between torsion angles are fairly similar in the bound and unbound states. Based on this fact, Brüschweiler et al. suggested[27] to neglect correlations between the torsion angles (as estimated from mutual information, which includes non-linear correlations[29]). In defining 'soft degrees of freedom' Brüschweiler et al. considered only one main torsion angle per shared pair of bonds. This is confirmed in the statement that the alanine dipeptide "has a total of 7 soft degrees of freedom"[26]. Translated to our definition of BAT coordinates, trialanine has 13 main torsions. However, trialanine also has 18 associated phase angles, which may or may not count as 'soft degrees of freedom'. The remaining 33 bond lengths and 32 bond angles are considered stiff or 'hard degrees of freedom'. Although their entropy estimation employs different numerical methods[26,27], their results are on similar footing with ours since: (i) They employ (a subset of) BAT coordinates. (ii) Their data are naturally balanced, as their conformers belong to two independent simulations, from which they likely take the same number of frames for their analysis.

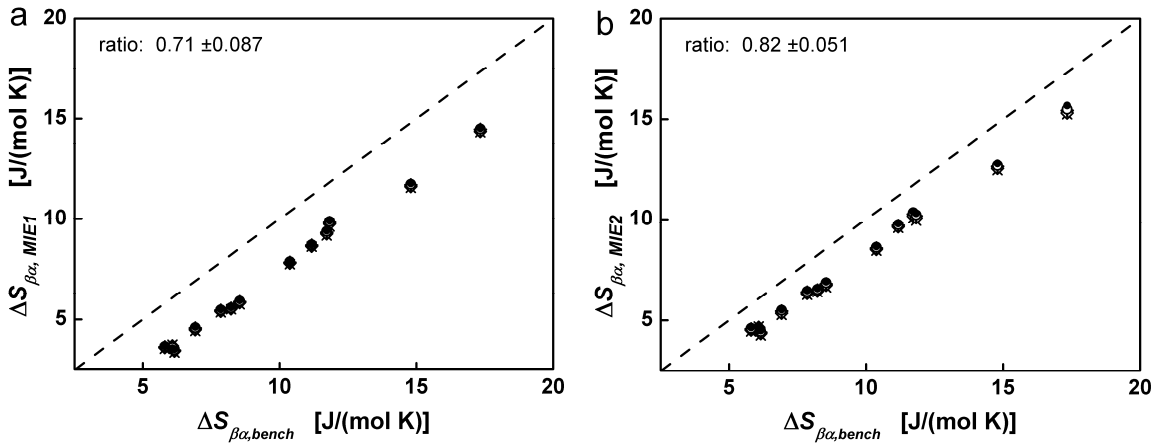

Figure S12: Entropy estimates for the trialanine model using only the main 13 torsion angles as 'soft degrees of freedom', and neglecting the conformational variations of phase angles, bond angles and bond lengths. Both methods (balancing and bias corrections) are used, as they yield the best results. Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1} / \Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0. **a:** First order MI expansion (MIE1); **b:** Second order MI expansion (MIE2).

We applied their suggestions to our model. In Figure S12a, we follow both suggestions. Using only the main 13 torsions with the 1st order MI expansion (MIE1) yields a low value of the estimate-to-benchmark ratio of 0.71 ±0.087. In Figure S12b, we switch to the 2nd order (MIE2), obtaining a larger estimate-to-benchmark ratio of 0.82 ±0.051. If we now alter the definition of soft degrees of freedom to include all 31 torsion and phase angles, we obtain a ratio of 0.81 ±0.069 for MIE1 and a ratio of 0.97 ±0.027 for MIE2 (Figure S13).

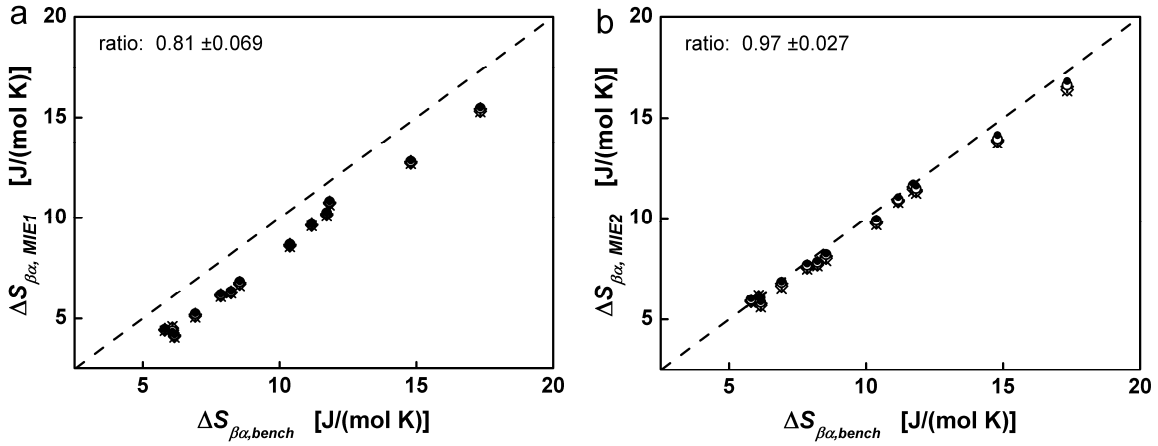

Figure S13: Entropy estimates for the trialanine model using 31 'soft degrees of freedom' (13 torsions and 18 phase angles), and neglecting the conformational variations of angles and bonds. Both methods (balancing and bias corrections) are used, as they yield the best results. Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1} / \Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0. **a:** First order MI expansion (MIE1); **b:** Second order MI expansion (MIE2).

In summary, the best estimates for trialanine are obtained when applying both correction methods: balancing and bias correction in the 2nd order MI expansion. Furthermore, using all 96 BAT coordinates with M = 35 bins histogram (Figure S10d) leads to the best estimate-to-benchmark ratio of 1.01 ±0.037. The second best results are obtained using only the 'soft degrees of freedom' defined as the torsion and phase angles (Figure S13b). Note that most data points in Figure S13b are below the identity line (ratios below 1.0), pointing to a slight systematic underestimation of the entropy differences due to small contributions from the hard degrees of freedom.

References

Complete ref (76) of main text: MacKerell Jr., A.; Bashford, D.; Bellott, M.; Dunbrack_Jr, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.; Prodhom, B.; Reiher_III, W.; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. J. Phys. Chem. B 1998, 102, 3586-3616.

(1) Shannon, C. E.; Weaver, W. *Bell Syst. Tech. J* **1948**, *27*, 379-423

(2) Ihara, S. *Information Theory for Continuous Systems*; World Scientific Publishing, 1993.

(3) Ben-Naim, A. *A Farewell To Entropy: Statistical thermodynamics based on information*; World Scientific Publishing Company: Singapore, 2008.

(4) Planck, M. *Ann. der Physik* **1922**, *371*, 365-372.

(5) Landau, L. D.; Lifshitz, E. M. *Statistical Physics Part 1-Vol. 5*; 3rd ed. ed.; Oxford: Pergamon Press, 1980.

(6) Potter, M. J.; Gilson, M. K. *J. Phys. Chem. A* **2002**, *106*, 563-566.

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

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

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

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

(11) Steinbrecher, T., *amber2accent v0.4*, 2007. http://ambermd.org/amber2accent/

(12) Case, D. A.; III, T. E. C.; Darden, T.; Gohlke, H.; Luo, R.; Merz_Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J Comp. Chem.* **2005**, *26*, 1668-1688.

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

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

(15) Schürmann, T. *J. Phys. A* **2004**, *37*, L295-L301.

(16) Matsumoto, M.; Nishimura, T. *ACM T. Model. Comput. S.* **1998**, *8*, 3-30.

(17) Wolynes, P. G. *Phil. Trans. R. Soc. A* **2005**, *363*, 453-467.

(18) Krivov, S.; Chekmarev, S. F.; Karplus, M. *Phys. Rev. Lett.* **2002**, *88*, 038101.

(19) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187-217.

(20) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835-848.

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

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

(23) Marsaglia, G. *Ann. Math. Statist.* **1972**, *43*, 645-646.

(24) Neumann, J. v. *NBS Appl. Math. Ser.* **1951**, *12*, 36-38.

(25) L'Ecuyer, P. *Mathematics of computation* **1996**, *65*, 203-213.

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

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

(28) Numata, J.; Wan, M.; Knapp, E. W. *Genome Inform.* **2007**, *18*, 192.

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