Entropy of Proteins Using Multiscale Cell Correlation

DOI:
10.1021/acs.jcim.0c00611

Document Version
Accepted author manuscript

Entropy of Proteins Using Multiscale Cell Correlation

Arghya Chakravorty,[†] Jonathan Higham,[‡] and Richard H. Henchman[*,¶,§]

†*Department of Chemistry, University of Michigan, Ann Arbor, Michigan, 48109, USA*

‡*MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine, The University of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh, EH4 2XU, United Kingdom*

¶*Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, United Kingdom*

§*Department of Chemistry, The University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom*

E-mail: henchman@manchester.ac.uk

Abstract

A new multiscale method is presented to calculate the entropy of proteins from molecular dynamics simulations. Termed Multiscale Cell Correlation (MCC), the method decomposes the protein into sets of rigid-body units based on their covalent-bond connectivity at three levels of hierarchy: molecule, residue and united atom. It evaluates the vibrational and topographical entropy from forces, torques and dihedrals at each level, taking into account correlations between sets of constituent units that together make up a larger unit at the coarser length scale. MCC gives entropies in close agreement with normal mode analysis and smaller than those using quasiharmonic analysis as well as providing much faster convergence. Moreover, MCC provides an insightful

decomposition of entropy at each length scale and for each type of amino acid according to their solvent exposure and whether they are terminal residues. While the residue entropy depends weakly on solvent exposure, there is greater variation in entropy components for larger, more polar amino acids, which have increasing conformational entropy but reduced vibrational entropy with greater solvent exposure.

Introduction

Given the inherent flexibility and marginal stability of proteins, it is essential to have methods to reliably and insightfully calculate them. Entropy is both a measure of the flexibility of a protein and is a key determinant of its stability. Its calculation requires the determination of the probability distribution of the positions and momenta of all atoms in the protein. Given the size and complexity of proteins, however, this is an especially difficult challenge.

Much work has been invested in calculating the entropy of biomolecular systems[1–9] One of the oldest methods applied to proteins is Normal Mode Analysis (NMA).[10–15] It calculates the entropy from the vibrational frequencies derived from a multidimensional Gaussian distribution using the eigenvalues of the Hessian matrix at an energy minimum. It avoids the need for a simulation, can be applied directly to experimental structures, and in the quantum formulation leads to absolute entropies. It has also been implemented using coarse-grained[16–18] or hierarchical[18–25] representations of proteins, although more with a focus on vibrational frequencies and their corresponding motions than on entropy.[17] Less advantageously, NMA is known to slightly underestimate entropy compared to experiment, at least for small molecules,[26] because it derives frequencies from a system's minimum which usually has higher curvature than configurations at the temperature of interest. Also, separate NMA calculations are required for every minimum,[27] which becomes impractical for large, multimodal systems, and it is unclear how many minima should be accounted for.[28] Frequently for proteins, only the entropy of different conformational minima of side chains treated independently is considered.[29,30]

Quasiharmonic Analysis (QHA), which is another widely used method for proteins, also approximates a system's probability distribution as a multidimensional Gaussian but derives it from the coordinate fluctuations of atoms relative to their average positions in a molecular dynamics simulation,[31–34] which lead to vibrational frequencies and absolute entropy. Its advantages are that only one matrix is needed and there is no need to explicitly account for the many different minima but this comes at the cost of significantly overestimating entropy.[35–37] Another disadvantage of QHA is its slow convergence because of noise in the off-diagonal correlations[36–39], but these can be somewhat ameliorated by the use of internal coordinates[35,40,41] or a von Mises distribution[42] for dihedral angles. To go beyond the Gaussian distribution, the probability distribution may be represented as a weighted sum of multiple Gaussians.[43]

A range of other strategies exist to calculate entropy. One approach uses probability distributions of individual dihedrals from structures generated in a simulation[44–47] or from the Protein Data Bank.[48,49] While these methods do not assume a Gaussian distribution, they cannot give absolute entropies in terms of discrete quantum states or over all degrees of freedom and in their simplest formulation do not account for correlation between coordinates. To account for correlation, a number of hybrid approaches exist that combine QHA and coordinate integration. The Boltzmann quasiharmonic variant combines the entropy from the one-dimensional distributions with correlation from the covariance matrix of coordinate fluctuations,[50,51] a process that can be extended to second-order joint distributions of eigenvectors.[52] QHA entropy with and without off-diagonal correlation can be used to calculate entropy from B-factors in X-ray crystal structures[4] although with limited generality.[53] Separate coordinate distributions may be integrated along the eigenvectors of the dihedral covariance matrix[54] or vectors derived by independent component analysis.[55] Nearest-neighbor methods[56,57] permit the more efficient calculation of higher-order probability distribution functions than fixed multi-dimensional histograms and may be extended to have anisotropic probability distributions.[58] Mutual information entropy (MIE) approaches[59,60] account for

second-order correlations but their convergence is difficult and higher-order correlations are problematic to include.[61,62] One solution that gives more reliable convergence and an upper bound to entropy is the minimum spanning tree (MIST) method which considers only a minimum set of pairwise correlations.[63] PopCOEN, a variant of MIST, trains a neural network from entropy calculated using pairwise correlations along the protein backbone.[64] Higher order correlations can be successfully accounted for in mutual information methods either by combining with nearest-neighbor methods,[65,66] with conformational discretisation, or with a cut-off in the range of correlations supplemented by NMA for each conformer.[28,67] Probability distributions over all discrete conformations are limited to small proteins[68] and still require evaluation of the vibrational term. In yet more ways to calculate protein entropy, it can be derived from the number of contacts,[69,70] the variance of the protein energy and solvent free energy without direct reference to structure,[71] or the probability distributions for successively growing a protein, monomer by monomer.[72–74]

Most methods to calculate entropy use coordinates at a single length scale, either Cartesian or internal, although it could be argued that backbone and side-chain dihedrals indicate different length scales. Multiscale approaches have the advantage of the more adaptable resolution of protein motion at small and large length scales as well as greater efficiency. Cartesian coordinates do not well capture correlations in non-linear motion, especially rotational motion, and can be slow to converge. Internal coordinates suffer from redundancy and arbitrariness of definition, mathematical complexity, and sensitivity. The disadvantages of multiscale methods are the arbitrariness and complexity of a hierarchy, ensuring the different levels do not overlap, and appropriately accounting for correlations. Multiscale approaches to calculate entropy have largely been restricted to NMA[18–25] and have therefore largely been applied to single, minimised structures rather than simulation ensembles. Multiscale Cell Correlation (MCC), which was recently extended to organic liquids,[75,76] also makes use of a multiscale coordinate framework. In MCC, each level comprises multiple sets of units covalently bonded together, with each set becoming a single larger unit at the coarser level

of structure, and so on to ever coarser levels. For each set of units, space is discretised into "cells", which correspond to local arrangements of the units in a particular energy well. The entropy over position and momentum within each cell is denoted "vibrational" while the entropy over all cell occupancy probabilities is termed "topographical", a nomenclature that avoids the ambiguous "configurational" label, which may be related to minima or to position. The vibrational entropy is evaluated from the eigenvalues of force and torque covariance for all units in the set, and the topographical entropy is evaluated from the distributions of unit contacts. MCC grew out of a single-level treatment of entropy for liquids[77–81] and solutions[82–85] which treated molecules as rigid units. It became dual-length scale when it incorporated internal molecular flexibility, first using dihedral distributions for small organic molecules in water[86] and later using force covariance of united atoms for hydrocarbons and small peptides.[37] Here we extend MCC to three length scales to calculate the entropy of a protein, compare it to the established methods of NMA and QHA, and use the resulting entropy decomposition to understand the distribution of entropy in a protein and its dependence on amino-acid type and solvent exposure.

MCC Theory

Assignment of Structural Hierarchy

To extend MCC to proteins, the protein is partitioned into units at three different length scales: the molecule (M) level, residue (R) level, and united-atom (UA) level. The entropy is evaluated at each level. A fourth level, the atomic level, characterized by the strongly quantized vibrations of hydrogen stretching and bending, is not considered here owing to the small size of its entropy. A residue, as in traditional usage, comprises an amino-acid side chain and its corresponding backbone atoms. A united atom is defined as a non-hydrogen atom together with its bonded hydrogen atoms which are represented explicitly rather than merged into the heavy atom. Every unit is treated as a rigid body with three translational

and either three, one or zero rotational degrees freedom, depending on whether the unit is non-linear, linear or a point, respectively, in terms of its constituent units.

We next define coordinate systems for the molecule, residue and united-atom levels (Figure 1). At the molecule level, the origin is the center of mass and the axes are taken as the principal axes of the molecule. These are used for both translation and rotation, as well as for translation at the residue level, while a local coordinate system is used for the rotational motion of a residue. This local coordinate system is based on the strongest interactions with neighboring units, which are the covalent bonds in the backbone with neighboring amino acids. Thus the origin for a residue is the average position of the three $NC_\alpha C$ atoms in the backbone. The x axis is the N-C vector, the y axis is orthogonal to that in the $NC_\alpha C$ plane, and z is orthogonal to both of these. We refer to this system as the $NC_\alpha C$ axes. These axes are also used for translation at the united-atom level, while for rotation, again because of the adjoining covalent bonds, the position of the heavy atom defines the origin and the average of the vectors along the covalent bonds with hydrogens defines the x axis, from which both y and z axes are derived (see Appendix for details).

The vibrational entropy of each unit relates to the effective harmonic potential of its constituent units. It is partitioned into translational and rotational components, which we label more specifically as transvibrational (transvib) and rovibrational (rovib). The topographical entropy (topo) relates to the probability distribution of energy wells corresponding to different arrangements of the units in the set. At the united-atom level, it is equivalent to the conformational entropy, excluding dihedrals that involve only hydrogen twisting, which are mostly zero due to symmetry (e.g. CH_3, NH_2) or small, being correlated with other united atoms via hydrogen bonds (e.g. OH). We do not here account for the topographical entropy at the residue level, based on the assumption that the relatively stable backbone structure of the proteins studied here makes it small. The topographical entropy at the molecule level is also not considered here but could be included by discretizing translation and rotation by the surrounding solvent molecules at the appropriate protein concentration.[87] The total

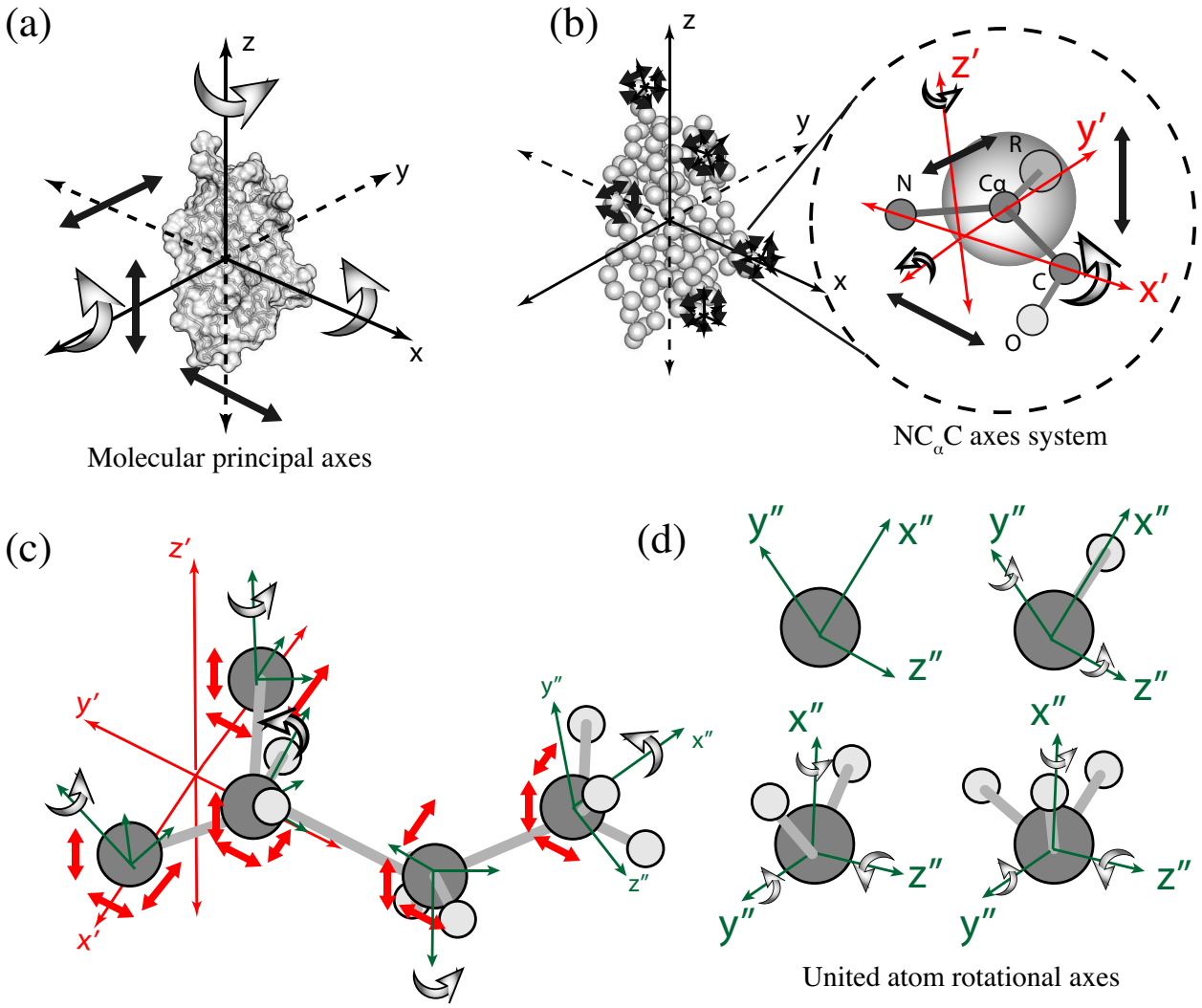

Figure 1: Axes at each level of hierarchy: (a) molecule (M) uses the principal axes of the molecule, (b) residue (R) uses the $NC_\alpha C$ axes of each residue, and (c) united atom (UA) uses the number of covalently bonded hydrogens to the central heavy atom according to (d) (see text for details of the axes).

entropy for a protein is therefore taken as the sum of seven terms

$$S_{\text{total}} = S_{\text{M}}^{\text{transvib}} + S_{\text{M}}^{\text{rovib}} + S_{\text{R}}^{\text{transvib}} + S_{\text{R}}^{\text{rovib}} + S_{\text{UA}}^{\text{transvib}} + S_{\text{UA}}^{\text{rovib}} + S_{\text{UA}}^{\text{topo}} \tag{1}$$

Vibrational Entropy

The vibrational entropy terms of a unit are calculated in the harmonic approximation using the equation for a quantum harmonic oscillator

$$S^{\text{transvib}} = k_{\text{B}} \sum_{i=1}^{N_{\text{vib}}} \left(\frac{h\nu_i / k_{\text{B}} T}{\mathrm{e}^{h\nu_i / k_{\text{B}} T} - 1} - \ln\left(1 - \mathrm{e}^{-h\nu_i / k_{\text{B}} T}\right) \right) \tag{2}$$

where k_{B} is Boltzmann's constant, h is Planck's constant, T is temperature, ν_i are the vibrational frequencies, and N_{vib} is the number of vibrations which depends on the number of constituent units. The frequencies are derived from the eigenvalues of the force and torque covariance matrices using the equation[37,75,76]

$$\nu_i = \frac{1}{2\pi} \sqrt{\frac{\lambda_i}{k_{\text{B}} T}} \tag{3}$$

The $3N \times 3N$ force covariance matrix is constructed from the mass-weighted forces of all N constituent units, where the elements of the matrix are given by $\langle F_i F_j / \sqrt{m_i m_j} \rangle$ and averaging is done over all simulation frames. The indices i and j range over the three axes x, y, z of each constituent unit. F_i is the force on a unit and m_i is its mass, both summed over all constituent units. As mentioned earlier, forces are oriented into the reference frame of the unit so that all its constituent units are in the same reference frame. If there is only unit, then that the frame of that unit is used, as for a single protein. The six smallest eigenvalues, corresponding to translation and rotation of the unit at the coarser level, are excluded to prevent double-counting at the coarser level ($N_{\text{vib}} = 3N - 6$ in Eq. 2). No eigenvalues are excluded at the highest level of hierarchy, as is the case for a single protein.

The torque covariance matrix has elements $\langle \tau_i \tau_j / \sqrt{I_i I_j} \rangle$ where τ_i and I_i are respectively the torque and moment of inertia for axis i. Values for the x axis for linear units and for all values of point units are excluded from the matrix because these have zero torque and zero moment of inertia. As mentioned earlier, torques are calculated in the reference frame of each constituent unit to effectively capture their rotational motion. Because the axes of each

constituent unit are different, correlations between torques of different constituent units are weakly correlated, and we therefore make the mean-field approximation that the pairwise forces contributing to each torque are partitioned equally between the participating atoms, such that the mean-field torque is half the calculated one.[77,78]

Topographical Conformational Entropy

The topographical term at the united-atom level, $S^{\text{topo}}_{\text{UA}}$, also called the conformational entropy, is derived from the probability distribution p_i of the N_{conf} unique sets of discrete conformations over all N_{dih} flexible dihedrals involving united atoms in the backbone and side chain of a residue using the standard equation for entropy

$$S^{\text{topo}} = -k_{\text{B}} \sum_{i=1}^{N_{\text{conf}}} p_i \ln p_i \tag{4}$$

This approach accounts for correlation within each side chain but ignores correlations between different side chains, which are assumed to be small for flexible dihedrals. An adaptive method is used to determine the different conformations of each dihedral rather than imposing fixed boundaries, similar in purpose but simpler than what has been done elsewhere.[28] Conformations are identified from peaks in the distribution of dihedral angles from the simulation, where the distribution is represented as a histogram with a 30° bin-width and a peak is defined as a bin with a higher population than both its adjacent bins. This automatically ensures that peaks are at least 60° apart. Subsequently, each dihedral is assigned to the closest peak. Figure S1 illustrates the performance of this method for dihedrals having 1–3 conformations. As mentioned earlier, residue topographical entropy is not considered here, assumed to be small for stable proteins.

Normal Mode Analysis and Quasiharmonic Analysis.

Two other methods capable of calculating the absolute entropy of proteins are tested for comparison: Normal Mode Analysis (NMA) and Quasiharmonic Analysis (QHA). NMA is performed using Gromacs (v2019)[88] from a single structure for each protein by minimising its PDB crystal structure including hydrogens with at most 5000 steps of the conjugate-gradient method and a steepest-descent step every 1000 steps, diagonalizing the mass-weighted Hessian matrix in Cartesian coordinates of the minimized structure, taking the $3N$-6 highest eigenvalues to remove whole-molecule translation and rotation, and converting to frequencies using $\nu_i = \sqrt{\lambda_i}/2\pi$ and to entropy with Eq. 2. QHA entropy is calculated using Gromacs from the same simulation trajectories as for MCC. It is evaluated using four different sets of Cartesian degrees of freedom to assess the effect of that choice:

1. All-atom. This matches NMA and closely maps to the MCC components $S_{\mathrm{UA}} + S_{\mathrm{R}}$. MCC ignores the high-frequency hydrogen-atom stretching and bending but these have negligible entropy.

2. Heavy atom. This omits the entropy of hydrogen atoms, most notably for the twisting of united atoms, and maps to the MCC components $S_{\mathrm{UA}}^{\mathrm{transvib}} + S_{\mathrm{UA}}^{\mathrm{topo}} + S_{\mathrm{R}}$.

3. C_α atom. This maps to the MCC component $S_{\mathrm{R}}^{\mathrm{transvib}}$.

4. Residue all-atom. This only involves the atoms of each amino acid. It maps to MCC components $S_{\mathrm{UA}}^{\mathrm{transvib}} + S_{\mathrm{UA}}^{\mathrm{rovib}} + S_{\mathrm{UA}}^{\mathrm{topo}}$.

For all methods, superposition is done onto the average structure using the relevant backbone C, C_α and N atoms. The mass-weighted coordinate covariance matrix is constructed and diagonalized and the lowest $3N$-6 eigenvalues are converted into frequencies ν_i using $\nu_i = \sqrt{k_{\mathrm{B}}T/\lambda_i}/2\pi$ and into entropy with Eq. 2. It should be noted that the QHA residue entropy components are not used to calculate the total entropy of each protein, unlike for

MCC. For NMA, equivalent entropy components beyond the all-atom case are non-trivial to calculate[19,89] and are not considered here.

Molecular Dynamics Simulations and Entropy Calculations

All molecular dynamics simulations are done using Gromacs (v2019).[88] A total of 74 proteins taken from the Protein Data Bank are examined. Having been used in a previous study,[90] each protein is a monomer with a crystal structure of resolution in the range 0.8–0.99 Å and number of amino acids in the range 30–200. Each protein is assigned protonation states appropriate to pH 7, solvated in its own cubic solvation box containing TIP3P[91] water with box sides placed at least 10 Å from the outermost protein atoms, and neutralised with Na^+ and Cl^- ions while maintaining an ion concentration of 150 mM. The CHARMM22/CMAP force field,[92,93] implemented as CHARMM27 in Gromacs, is used to assign force-field parameters to every component of the system and create its associated topology file.

Each system is minimized for at most 10000 steps of steepest-descent minimization followed by equilibration for 2 ns of molecular dynamics simulation under constant volume and temperature (NVT) conditions for 2 ns using the V-rescale thermostat at 300 K, during which all protein heavy atoms are harmonically restrained to their energy minimized positions using a force constant of 1000 kJ $mol^{-1}nm^{-1}$. The system is brought to the correct density with 3 ns of simulation at a pressure of 1 bar (NPT) using the Parrinello-Rahman barostat while still maintaining the harmonic restraints. Simulations are then run for a further 60 ns under the same conditions except restraints are now removed, with forces and coordinates saved every 10 ps. Only the last 50 ns are used for data collection, giving 5000 frames for the entropy calculation. Simulations use LINCS on all bonds involving hydrogen atoms, a non-bonded cutoff of 1.2 nm, periodic boundary conditions, particle-mesh Ewald summation with default parameters, and a 2 fs timestep. MCC entropy is calculated using a Python package developed in-house, which is available for download at https://github.com/arghya90/CodeEntropy.

Results

MCC Entropies versus NMA and QHA

The first comparison assesses how MCC entropy values compare with those using the established methods of NMA and QHA. These are plotted in Figure 2 using all atom, heavy atom and C_α atoms for QHA with the corresponding MCC components, noting that all methods exclude the translational and rotational entropy of the protein. It can be seen that

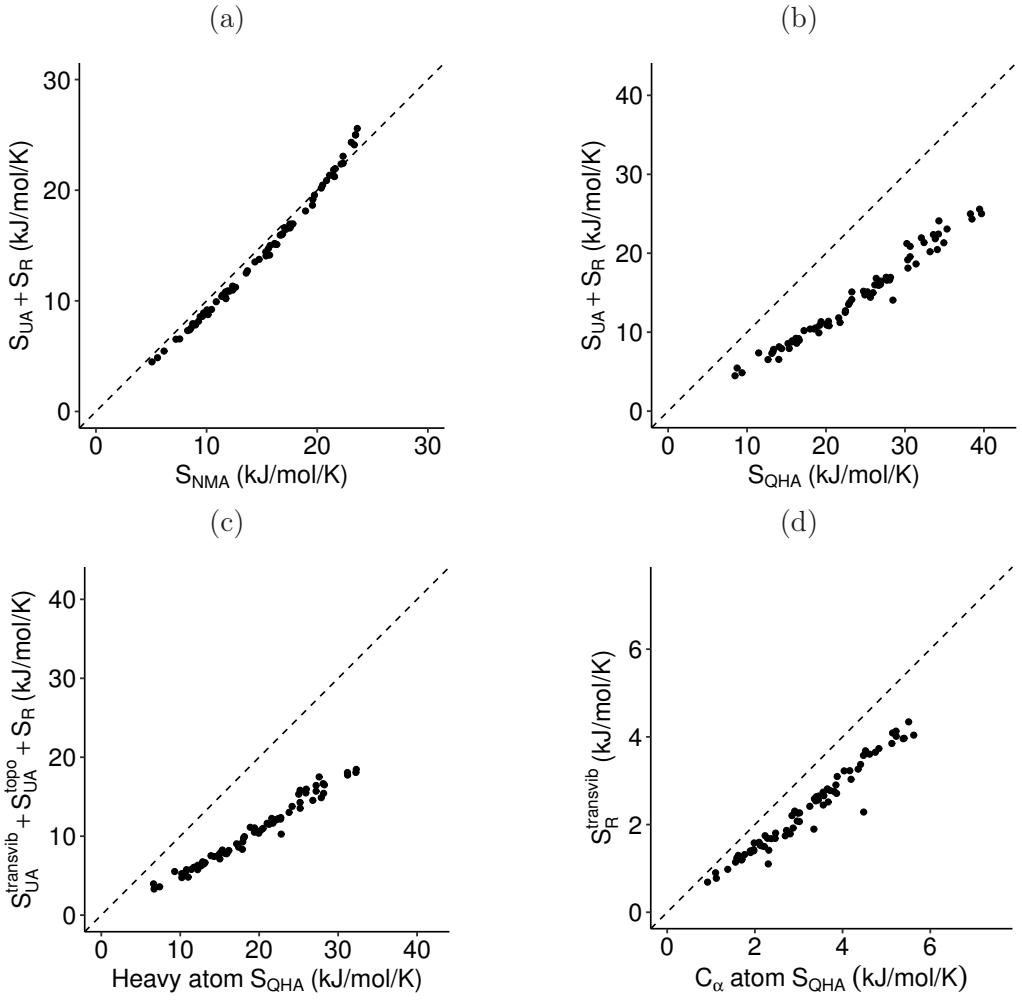

Figure 2: Total MCC protein entropies versus those by (a) NMA (a) and (b-d) QHA. The comparison with QHA involves plots of the appropriate MCC entropy components against three different sets of atoms: (b) all atoms, (c) heavy atoms, and (d) C_α atoms.

MCC values lie close to NMA values, with a linear fit yielding a slope of 1.11, y-intercept

of -2.12 kJ mol^{-1} K^{-1} and Spearman rank correlation of 0.998. QHA values, however, are about 40% larger (slope 0.70, y-intercept -2.28 kJ mol^{-1} K^{-1}, and Spearman rank correlation 0.988), which is expected based on results seen elsewhere because QHA excessively smooths over the potential energy surface.[35–37] This is reinforced by Figure S2, which illustrates for representative amino acids how the force distributionsare very closely Gaussian in striking contrast to the coordinate distributions, thereby making the harmonic approximation much more accurate for forces than for coordinates, as has been observed earlier.[37,79] For the heavy-atom entropy, which is about 2/3 of the all-atom value, there is a similar discrepancy between MCC and QHA (slope 0.62, y-intercept -1.40 kJ mol^{-1} K^{-1}, and Spearman rank correlation 0.988). At the C_α level, for which entropies are now only $\sim$16% of the all-atom values, QHA is closer and now only $\sim$30% larger than MCC (slope = 0.77, y-intercept -0.12 kJ mol^{-1} K^{-1}, and Spearman rank correlation 0.976). A clearer representation of the entropy components is given in the next section.

MCC Entropy Components at Each Level of Hierarchy

Figure 3 illustrates the sizes of the seven entropy MCC components in Eq. 1 for all proteins and how these components depend on the number of units at their respective levels of operation, namely molecule, residue and united atom. The trends make clear the extensive nature of entropy: the more units at a given level of hierarchy, the larger the entropy in a closely linear trend. However, the entropy per unit strongly depends on the type of unit. The entropy per protein molecule is $\sim$135 J K^{-1} mol^{-1}, the entropy per residue is smaller at $\sim$65 J K^{-1} mol^{-1}, while the entropy per united atom is smaller still at $\sim$9 J K^{-1} mol^{-1}. Evidently, the smaller the unit, the stronger the interactions relative to the size of the unit. The next observation is that transvibrational slightly exceeds rovibrational entropy at the molecule level, but rovibrational is larger at the other two levels. This reflects differing constraints on translation and rotation at different levels from the connecting covalent bonds, and that rotations of units, each involving different coordinate frames, are decoupled and

mean-field. The third trend is that vibrational entropy is much larger than topographical entropy, consistent with previous work for proteins.[94] This is partly because vibration occurs in three dimensions compared to conformational variation which takes place in a single dimension, and partly because the probability distribution of accessible quantum vibrational states at ambient conditions is broader than that for conformation.

An issue of further interest is the convergence of entropy with simulation time. This is plotted for the residue and united-atom MCC entropy components in Eq. 2 and total MCC and QHA entropies (Figure S3) for two representative proteins, the smallest and the largest in the dataset whose simulations have been extended from 60 to 100 ns. MCC components are seen to be well converged at 60 ns. Moreover, MCC converges much faster than QHA, decreasing by only 2 kJ K^{-1} mol^{-1} over the 80 ns window for the largest protein versus the 23 kJ K^{-1} mol^{-1} increase for QHA. This slow convergence of QHA has been noted before,[36–39] as has the opposite direction of convergence for forces compared to coordinates.[37]

Amino Acid Entropy

The entropy of each kind of amino acid averaged over all amino acids in all proteins is plotted in Figure 4 for both MCC and QHA. The MCC residue entropy comprises all three united-atom terms in Eq. 1 for all united atoms making up that residue. All methods display the same overall trends, with larger and more flexible amino acids having more entropy. However, MCC entropy values are less variable than QHA values, being larger for small amino acids like Ala, Gly and Pro and smaller for large amino acids like Arg and Lys. The larger entropy for the small amino acids primarily arises from the rovibration of united atoms, a quantity that QHA is less-well suited to resolve in Cartesian coordinates. Interestingly, Arg and Lys have comparable total entropy for MCC but for different reasons, with Arg having more transvibrational entropy owing to its greater mass and Lys having more conformational entropy because of its greater conformational sampling.[95] This contrasts with QHA which predicts a larger entropy for Lys because the coordinates now span a much larger range.

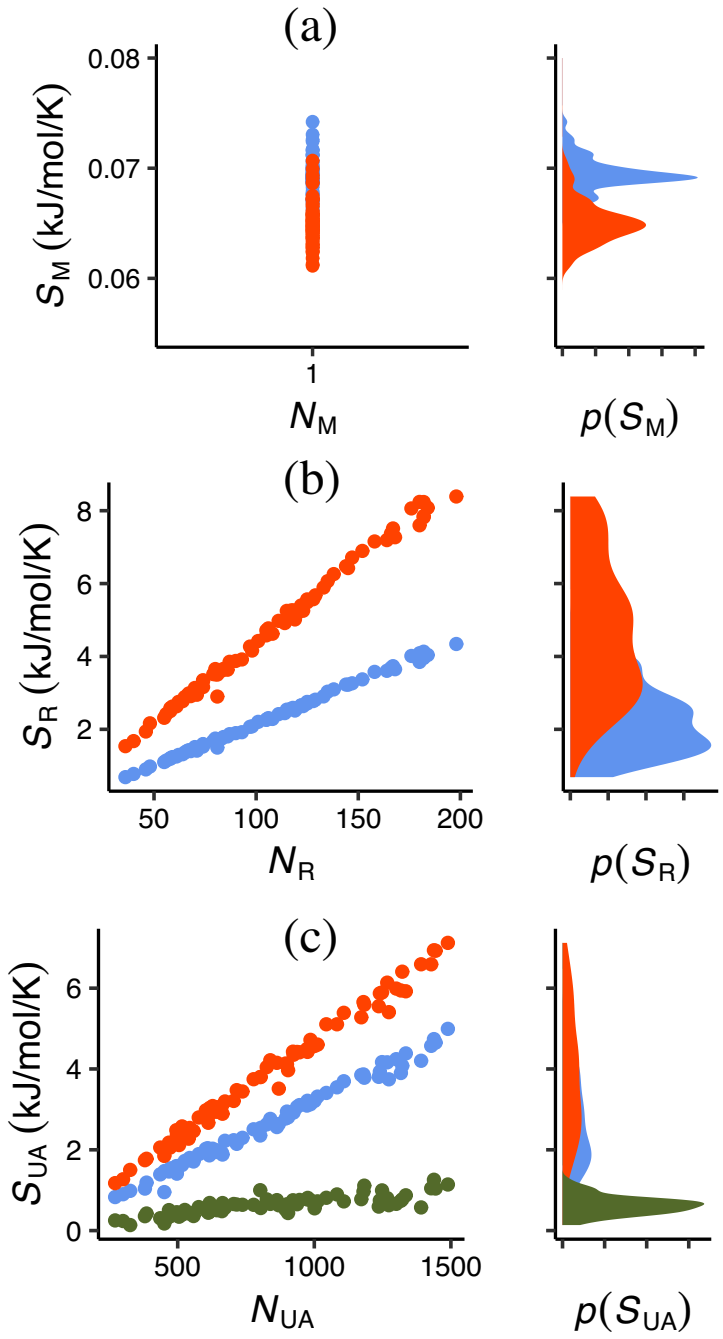

Figure 3: Left: MCC Entropy components S_i^{transvib} (blue), S_i^{rovib} (red) and S_i^{topo} (green) at the three levels of hierarchy $i = (a)$ M, (b) R or (c) UA versus the number of units at that level, N_{i}, for all proteins. Right: probability distributions $p(S_i)$ of each entropy component over all units at each level for all proteins.

Amino-Acid Entropy versus Solvent Accessible Surface Area

The entropy of amino acids naturally depends on their protein environment. The variable chosen here to explore that dependence is the solvent accessible surface area (SASA). SASA

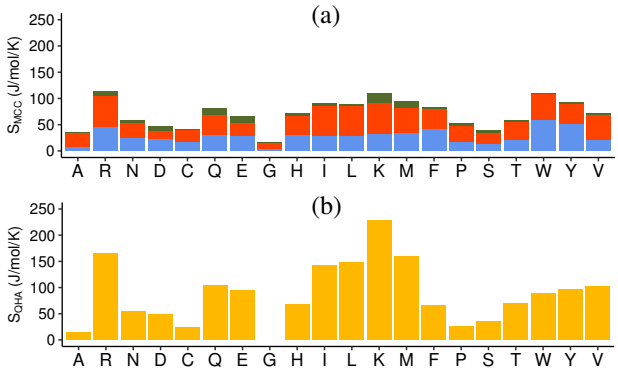

Figure 4: Average entropy of each type of amino acid for (a) MCC and its components $S_{\mathrm{UA}}^{\mathrm{transvib}}$ (blue), $S_{\mathrm{UA}}^{\mathrm{rovib}}$ (red), and $S_{\mathrm{UA}}^{\mathrm{topo}}$ (green), and (b) QHA (yellow).

is calculated using a probe radius of 1.4 Å with the Gromacs `sasa` command.[96] Figure 5 reveals a number of important trends in how the residue translational and rotational components depend on SASA. The expected outliers are the residues at the chain termini. The terminal ammonium and carboxylate groups have more transvibrational entropy when unconstrained by neighboring residues, with the freely rotating ammonium groups also having larger rovibrational entropy. For the main-chain residues, as observed earlier, rovibrational entropy is larger than transvibrational for most amino acids. The exceptions are Asp and Glu, which have fewer hydrogens, and Phe, Trp, and Tyr, which are more rigid. Rovibrational entropy for each amino acid is fairly constant and independent of environment. The only exception is Cys which has two very different rovibrational entropy values, with the entropy understandably much lower when Cys forms a disulfide bond. Transvibrational entropy is also relatively constant for the smaller and hydrophobic amino acids. However, it is more scattered and even decreases with increasing SASA for the larger, more polar amino acids, namely Arg, Asn, Asp, Gln, Glu, His, Phe, Trp, and Tyr, contrary to the expectation of greater flexibility at the surface. This indicates that these amino acids have more diverse environments and weaker interactions when inside the protein than on the surface. The curious exception is Lys, which has a relatively consistent value, suggesting it has a more homogeneous environment because of its weakly interacting hydrophobic chain and well-hydrated ammonium group.

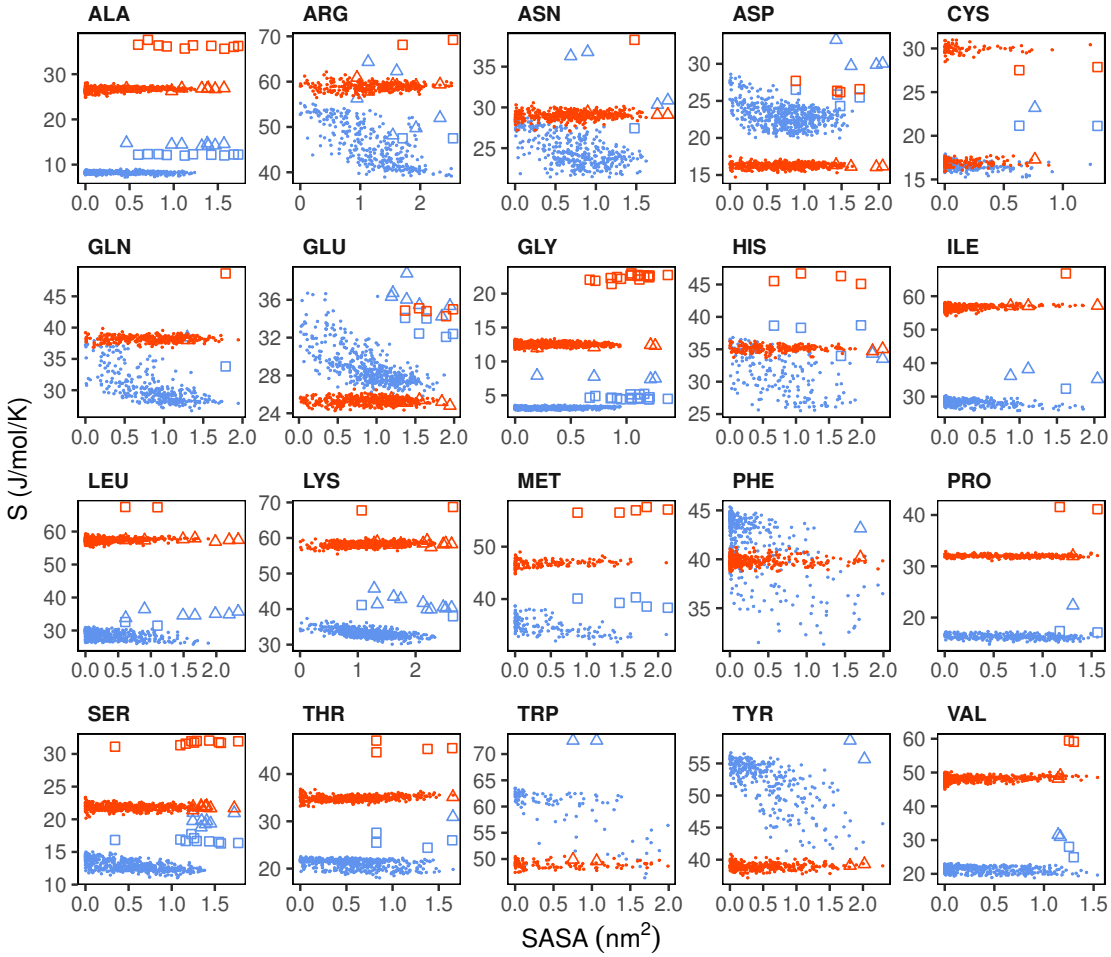

Figure 5: Transvibrational (blue) and rovibrational (red) entropy at the united atom level for all residues of a given type versus their solvent accessible surface area (SASA). N and C termini residues are indicated by open squares (□) and open triangles (△) and the remainder by solid points (•).

To reconcile the expectation of more flexible residues at the protein surface, Figure 6 illustrates how the combined vibrational entropy at the united-atom level and topographical entropy of each residue depend on SASA. Contrary to the decrease in vibrational entropy seen in Figure 5, the topographical entropy increases with SASA for most residues with flexible side-chain dihedrals, in line with the expectation that amino acids at the surface are able to access more conformations. These two trends are opposite but about the same size, even though the topographical entropy is much smaller.

A final comparison shows how residue MCC and QHA entropy depends on SASA in

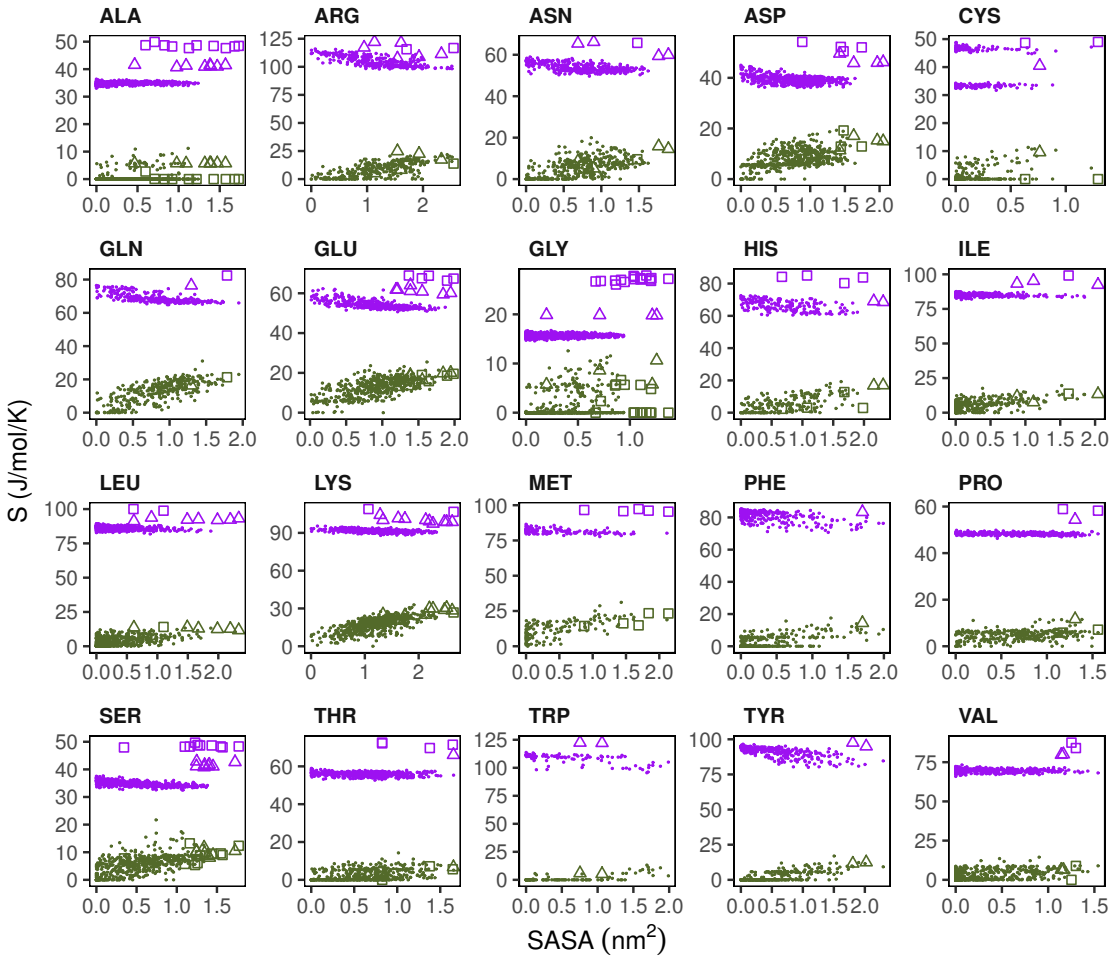

Figure 6: Vibrational (purple) and topographical entropy (green) for all residues of a given type versus their solvent accessible surface area (SASA). N and C termini residues are as in Figure 5.

Figure 7, which expands on the average values for amino acids given earlier in Figure 4. MCC values are relatively flat, a consequence of the cancellation between the vibrational and topographical entropy. QHA values, however, have a much greater spread and display a strong increase with SASA for most amino acids, varying by as much as a factor of 4, and even more for terminal amino acids.

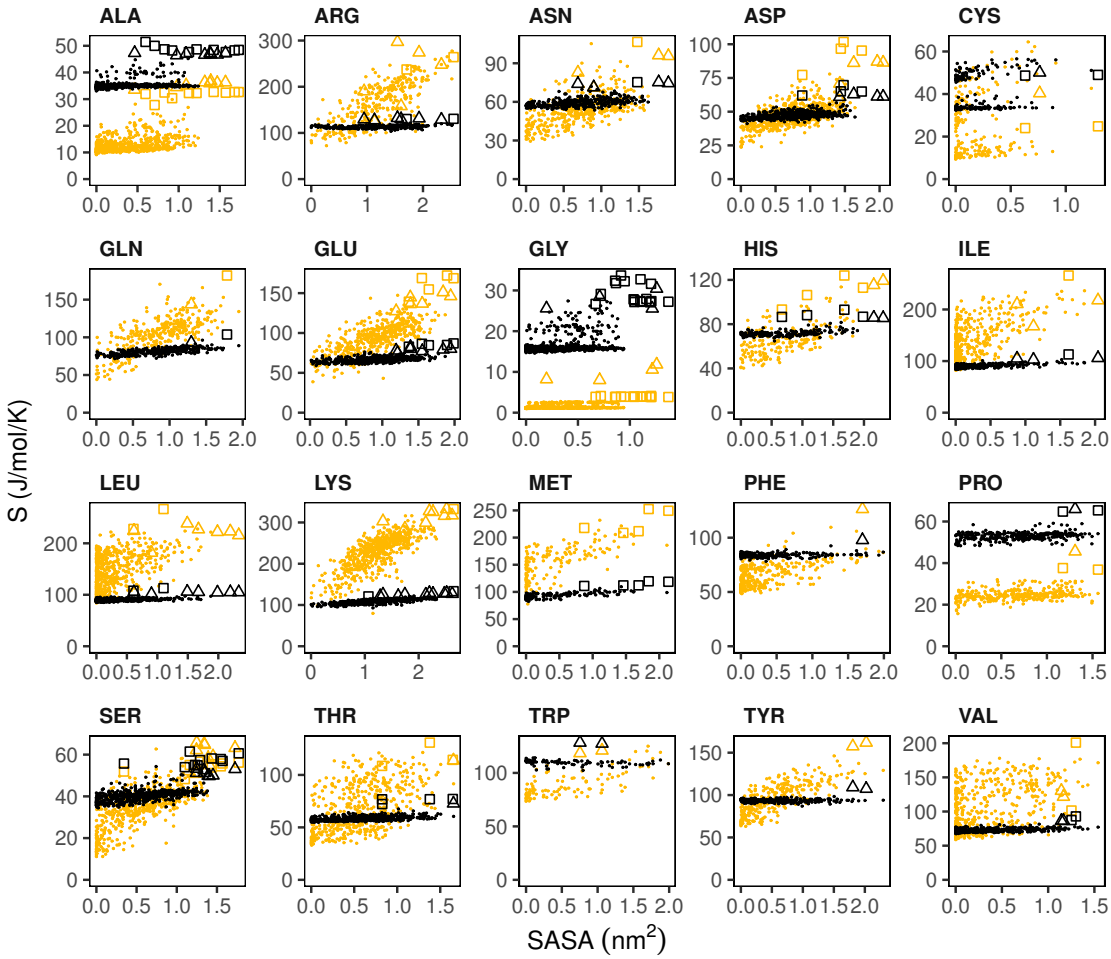

Figure 7: Entropy for all residues of a given type using MCC (black) and QHA with backbone superposition (yellow) versus solvent accessible surface area (SASA). N and C termini residues are as in Figure 5.

Discussion

Multiscale methods are widely used in simulation methodology. Here we apply the concept to the calculation of entropy where it has rarely been used. This is helpful both in understanding entropy in a more localised fashion as well as enabling its efficient calculation because it makes use of smaller matrices that converge faster than a single matrix over all atoms. We implement a three-level hierarchy using molecule, residue and united atom. Additional levels of hierarchy could be implemented at intermediate levels such as secondary-structure helices and sheets or to domains, and the same approach can be applied to liquids and therefore

to molecules in solution.[82–86] There is scope for an additional assembly level for protein complexes, and higher still if desired. Overall, this formulation shows great promise in being extended to even larger systems which are in strong need of new methods to quantify structure and stability over a wide range of length scales. Disadvantages of a hierarchy is that it is approximate, it must be assigned in an arbitrary or automated fashion,and, while many biological molecules have an inherent hierarchical structure, molecules in general do not. Nonetheless, this issue is shared by coarse-grain force fields[97–101] and can gain from developments in this area. A further issue lies in the optimal choice of reference frame for each unit. While methods based on principal axes or connecting covalent bonds are general, an automated way based on confinement by the environment would be desirable.

Concerning MCC compared to existing methods, it closely agrees with NMA. One would expect the MCC entropy to be slightly larger than NMA, given that MCC includes the topographical entropy whereas NMA does not, and that NMA tends to slightly underestimate entropy because it fits to the minimum. The slight underestimate for MCC may be because MCC averages over some level of force correlation, particularly in the flexible side chains, similar to the superposition issue of QHA.[102] It may also be because it omits force correlations between different non-covalently bonded units, or because the removal of the translational and rotational entropy for each unit to remove duplication at the higher length scale is not perfect. Finally, MCC in its current formulation omits the topographical entropy at the residue level, a term that is expected to be negligible for the stable proteins studied here but larger for unfolded and disordered proteins. QHA entropies are found to be about 40% higher than MCC and NMA, consistent with findings made elsewhere.[35–37] The large size of QHA residue entropy values, similar to what has been found elsewhere,[103] are made clearer in the amino-acid plots here as a function of solvent exposure which show large entropy spreads of up to 100–200 J K^{-1} mol^{-1}. This is in contrast to the MCC amino-acid entropy spread of $\sim$10 J K^{-1} mol^{-1} plus an additional $\sim$15 J K^{-1} mol^{-1} for terminal residues. The use by QHA of a single broad harmonic potential makes it sensitive to conformational changes, slower in

convergence, and prone to predicting a much larger entropy increase. The excessive nature of the QHA spread is highlighted by being larger than the entropy change for the vaporization of a liquid to a gas of $\sim$85 J K^{-1} mol^{-1} by Trouton's rule,[81] and yet conformational change in solution involves a much smaller gain in flexibility. Two reasons for the better performance of MCC are the use of rotational coordinates and the closely Gaussian distribution of forces to which is fitted a single harmonic potential averaged over all energy wells with a separate term to account for the multimodal dihedral distribution. Versions of QHA that reduce these effects can be implemented with multimodal potentials or using internal coordinates.[35,40,41] The harmonic approximation and the partitioning into energy wells in MCC works well for condensed phases but would break down in systems with softer potential energy surfaces such as at low density or high temperature. Methods such as MIE[59,60] and MIST[63] avoid the harmonic approximation but are limited to giving differences in entropy and require longer simulations to converge, on the order of a μs and millions of frames,[104] although fewer appear to be needed when implemented in a nearest-neighbor formulation.[105]

MCC has provided new insights into the nature of entropy in proteins. The trends for the average entropy of amino acids are informative but not unexpected, scaling with size, pre-existing notions of flexibility and whether they are chain termini. However, some trends are less obvious, such as Arg and Lys having similar overall entropies, with Arg having more vibrational entropy because of its larger mass and Lys having more conformational entropy because of its greater solvent exposure. This difference could help explain phenomena such as why hyperthermophilic proteins are better stabilized by Lys than Arg[95] because of the larger entropic stabilization by Lys in the native state relative to a misfolded confined state. This is a specific case of the more general and intriguing finding that residue entropy is largely independent of the degree of solvent exposure. For the larger, more polar amino acids, this trend arises because of a cancellation between the residue topographical and transvibrational entropy, the former increasing as expected but the latter actually decreasing with solvent exposure. This is contrary to the expectation that residue entropy increases with solvent

exposure because of the greater fluidity of the solvent, as had been found elsewhere using QHA.[103] In relation to protein folding and binding, a number of studies report an increase in vibrational entropy upon protein folding or binding,[11,106–108] some find changes in either direction[41,109,110] while others observe little variation.[30,71,94] The results here reveal that the side chains of larger, more polar residues when buried in proteins experience softer energy landscapes with fewer, flatter minima but at the surface there are more minima that are more sharply defined. This indicates there is an entropy-entropy compensation between topographical and vibrational terms because fewer, larger minima can fit in a given space. While stronger forces might be expected inside the protein due to close-packing and confinement, there is uncertainty about whether the protein interior is more or less dense than its surface[111] and there is no direct covalent confinement by buried residues compared to surface ones. One possible cause is the stronger hydrogen bonds between side chains and the surrounding water than with the rest of the protein.[112–114] Another cause may be the greater confinement of a side chain by water's flexible hydrogen-bond network. Yet another cause may be missing correlations that affect buried and surface amino acids differentially. Such relationships require further investigation.

Conclusions

We have extended the entropy method Multiscale Cell Correlation to the important case of proteins using molecule, residue and united-atom levels of hierarchy. MCC gives entropies in close agreement with normal mode analysis and smaller than quasiharmonic analysis. MCC entropy values converge noticeably faster than those for QHA. At each level of hierarchy, MCC gives an insightful decomposition of entropy into transvibrational, rovibrational and topographical entropy, and their dependence on the type of amino acid, whether they are main-chain or terminal, and their degree of solvent exposure. While a number of trends are in line with expectations, two less obvious trends have been found. First larger, more polar

amino acids have entropies that are much more variable and sensitive to their environment, namely Arg, Asn, Asp, Gln, Glu, His, Phe, Trp and Tyr but not Lys. Second, these same amino acids show a decrease in vibrational entropy with increasing solvent exposure, which cancels with their increasing topographical entropy, leading to a total residue entropy that is largely independent of solvent exposure. Overall, these findings reinforce the promise of MCC to be able to calculate entropy in molecular dynamics simulations of complex molecular systems on arbitrarily large length scales in a consistent, systematic framework.

Appendix

Construction of the Rotational Axes at the United-Atom Level

The rotational axes for a united atom are constructed in the following way. The x axis is defined along the average vector of all bond vectors between the central non-hydrogen atom and its covalently bonded hydrogen atoms. Using the formalism outlined elsewhere,[115] this vector can be represented as $\vec{r}$ with the following components in the Cartesian space

$$\vec{r} = \begin{bmatrix} r\sin\theta\cos\phi \\ r\sin\theta\sin\phi \\ r\cos\theta \end{bmatrix} \tag{5}$$

where r is the norm of $\vec{r}$, θ is the polar angle and ϕ is the azimuthal angle. From this, the unit vector along x (x'') is expressed as

$$x'' = \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix} \tag{6}$$

The y axis coincides with the $\hat{\theta}$ vector of the local curvilinear system with the components

$$y'' = \begin{bmatrix} \cos\theta\cos\phi \\ \cos\theta\sin\phi \\ -\sin\theta \end{bmatrix} \tag{7}$$

Likewise, the z axis coincides with the $\hat{\phi}$ vector with the components

$$z'' = \begin{bmatrix} -\sin\phi \\ \cos\phi \\ 0 \end{bmatrix} \tag{8}$$

The three axes with the origin at the position of the central heavy atom form an orthonormal basis in Euclidean space.

Acknowledgement

We thank Emil Alexov for his support of this work, Andrew Almond for advice on efficient coding of the software and guidance with technical details, and Jim Warwicker for helpful discussions on protein entropy.

Supporting Information Available

SI Figures:

- Figure S1: Schematic of the adaptive method to determine conformations.
- Figure S2: Force and coordinate probability distributions for three representative amino acids.
- Figure S3: Convergence of MCC and QHA entropy with time for two representative

proteins.

References

(1) Brady, G. P.; Sharp, K. A. Entropy in Protein Folding and in Protein-Protein Interactions. *Curr. Biol.* **1997**, *7*, 215–221.

(2) Meirovitch, H.; Cheluvaraja, S.; White, R. P. Methods for Calculating the Entropy and Free Energy and their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr. Prot. Peptide Sci.* **2009**, *10*, 229–243.

(3) Zhou, H. X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092–4107.

(4) Polyansky, A. A.; Zubac, R.; Zagrovic, B. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer: Berlin, Germany, 2012; Vol. 819; pp 327–353.

(5) Baron, R.; McCammon, J. A. Molecular Recognition and Ligand Association. *Ann. Rev. Phys. Chem.* **2013**, *64*, 151–175.

(6) Suárez, D.; Diaz, N. Direct Methods for Computing Single-Molecule Entropies from Molecular Simulations. *Rev. Comput. Sci.* **2015**, *5*, 1–26.

(7) Kassem, S.; Ahmed, M.; El-Sheikh, S.; Barakat, K. H. Entropy in Bimolecular Simulations: A Comprehensive Review of Atomic Fluctuations-based Methods. *J. Mol. Graph. Model.* **2015**, *62*, 105–117.

(8) Chong, S. H.; Chatterjee, P.; Ham, S. Computer Simulations of Intrinsically Disordered Proteins. *Ann. Rev. Phys. Chem.* **2017**, *68*, 117–134.

(9) Huggins, D. J.; Biggin, P. C.; Damgen, M. A.; Essex, J. W.; Harris, S. A.; Henchman, R. H.; Khalid, S.; Kuzmanic, A.; Laughton, C. A.; Michel, J.; Mulholland, A. J.;

Rosta, E.; Sansom, M. S. P.; van der Kamp, M. W. Biomolecular Simulations: From Dynamics and Mechanisms to Computational Assays of Biological Activity. *WIREs Comput. Mol. Sci.* **2019**, *9*, e1393.

(10) Go, N.; Scheraga, H. A. Analysis of Contribution of Internal Vibrations to Statistical Weights of Equilibrium Conformations of Macromolecules. *J. Chem. Phys.* **1969**, *51*, 4751–4767.

(11) Hagler, A. T.; Stern, P. S.; Sharon, R.; Becker, J. M.; Naider, F. Computer Simulation of the Conformational Properties of Oligopeptides - Comparison of Theoretical Methods and Analysis of Experimental Results. *J. Am. Chem. Soc.* **1979**, *101*, 6842–6852.

(12) Brooks, B.; Karplus, M. Harmonic Dynamics of Proteins — Normal-Modes and Fluctuations in Bovine Pancreatic Trypsin-Inhibitor. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 6571–6575.

(13) Levitt, M.; Sander, C.; Stern, P. S. Protein Normal-Mode Dynamics - Trypsin-Inhibitor, Crambin, Ribonuclease and Lysozyme. *J. Mol. Biol.* **1985**, *181*, 423–447.

(14) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

(15) Skjaerven, L.; Hollup, S. M.; Reuter, N. Normal Mode Analysis for Proteins. *J. Mol. Struc-Theochem.* **2009**, *898*, 42–48.

(16) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.

(17) Bahar, I.; Atilgan, A. R.; Demirel, M. C.; Erman, B. Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability. *Phys. Rev. Lett.* **1998**, *80*, 2733–2736.

(18) Ghysels, A.; Miller, B. T.; Pickard, F. C.; Brooks, B. R. Comparing Normal Modes across Different Models and Scales: Hessian Reduction versus Coarse-Graining. *J. Comput. Chem.* **2012**, *33*, 2250–2275.

(19) Hao, M. H.; Harvey, S. C. Analyzing the Normal Mode-dynamics of Macromolecules by the Component Synthesis Method. *Biopolymers* **1992**, *32*, 1393–1405.

(20) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. Building-block Approach for Determining Low-frequency Normal Modes of Macromolecules. *Proteins* **2000**, *41*, 1–7.

(21) Schuyler, A. D.; Chirikjian, G. S. Efficient Determination of Low-frequency Normal Modes of Large Protein Structures by Cluster-NMA. *J. Mol. Graph. Model.* **2005**, *24*, 46–58.

(22) Ghysels, A.; van Speybroeck, V.; Pauwels, E.; van Neck, D.; Brooks, B. R.; Waroquier, M. Mobile Block Hessian Approach with Adjoined Blocks: An Efficient Approach for the Calculation of Frequencies in Macromolecules. *J. Chem. Theory Comput.* **2009**, *5*, 1203–1215.

(23) Lu, M. Y.; Ming, D. M.; Ma, J. P. fSUB: Normal Mode Analysis with Flexible Substructures. *J. Phys. Chem. B* **2012**, *116*, 8636–8645.

(24) Kusumaatmaja, H.; Whittleston, C. S.; Wales, D. J. A Local Rigid Body Framework for Global Optimization of Biomolecules. *J. Chem. Theory Comput.* **2012**, *8*, 5159–5165.

(25) Mochizuki, K.; Whittleston, C. S.; Somani, S.; Kusumaatmaja, H.; Wales, D. J. A Conformational Factorisation Approach for Estimating the Binding Free Energies of Macromolecules. *Phys. Chem. Chem. Phys.* **2014**, *16*, 2842–2853.

(26) Pople, J. A.; Schlegel, H. B.; Krishnan, R.; Defrees, D. J.; Binkley, J. S.; Frisch, M. J.; Whiteside, R. A.; Hout, R. F.; Hehre, W. J. Molecular-Orbital Studies of Vibrational Frequencies. *Int. J. Quantum Chem.* **1981**, 269–278.

(27) Evans, D. A.; Wales, D. J. Free Energy Landscapes of Model Peptides and Proteins. *J. Chem. Phys.* **2003**, *118*, 3891–3897.

(28) Suárez, E.; Diaz, N.; Suárez, D. Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2011**, *7*, 2638–2653.

(29) Pickett, S. D.; Sternberg, M. J. E. Empirical Scale of Side-chain Conformational Entropy in Protein-Folding. *J. Mol. Biol.* **1993**, *231*, 825–839.

(30) Doig, A. J.; Sternberg, M. J. E. Side-chain Conformational Entropy in Protein Folding. *Protein Sci.* **1995**, *4*, 2247–2251.

(31) Karplus, M.; Kushick, J. N. Methods for Estimating the Configuration Entropy of Macromolecules. *J. Am. Chem. Soc.* **1981**, *14*, 325–332.

(32) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance-matrix. *Chem. Phys. Lett.* **1993**, *215*, 617–621.

(33) Sch afer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. Entropy Calculations on a Reversibly Folding Peptide: Changes in Solute Free Energy Cannot Explain Folding Behavior. *Proteins* **2001**, *43*, 45–56.

(34) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289–6292.

(35) Chang, C. E.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the Quasiharmonic Approximation. *J. Chem. Theory. Comput.* **2005**, *1*, 1017–1028.

(36) Genheden, S.; Ryde, U. Will Molecular Dynamics Simulations of Proteins Ever Reach Equilibrium? *Phys. Chem. Chem. Phys.* **2012**, *14*, 8662–8677.

(37) Hensen, U.; Gräter, F.; Henchman, R. H. Macromolecular Entropy Can Be Accurately Computed from Force. *J. Chem. Theory Comput.* **2014**, *10*, 4777–4781.

(38) Harris, S. A.; Gavathiotis, E.; Searle, M. S.; Orozco, M.; Laughton, C. A. Cooperativity in Drug-DNA Recognition: A Molecular Dynamics Study. *J. Am. Chem. Soc.* **2001**, *123*, 12658–12663.

(39) Gohlke, H.; Case, D. A. Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein-Protein Complex Ras-Raf. *J. Comput. Chem.* **2004**, *25*, 238–250.

(40) Hikiri, S.; Yoshidome, T.; Ikeguchi, M. Computational Methods for Configurational Entropy Using Internal and Cartesian Coordinates. *J. Chem. Theory Comput.* **2016**, *12*, 5990–6000.

(41) Goethe, M.; Fita, I.; Rubi, J. M. Testing the Mutual Information Expansion of Entropy with Multivariate Gaussian Distributions. *J. Chem. Phys.* **2017**, *147*, 224102.

(42) Li, D. W.; Bruschweiler, R. In silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides. *Phys. Rev. Lett.* **2009**, *102*, 118108.

(43) Gyimesi, G.; Zavodszky, P.; Szilagyi, A. Calculation of Configurational Entropy Differences from Conformational Ensembles Using Gaussian Mixtures. *J. Chem. Theory Comput.* **2017**, *13*, 29–41.

(44) Edholm, O.; Berendsen, H. J. C. Entropy Estimation from Simulations of Non-diffusive Systems. *Mol. Phys.* **1984**, *51*, 1011–1028.

(45) Zhang, J. F.; Liu, J. S. On Side-chain Conformational Entropy of Proteins. *PLoS Computational Biology* **2006**, *2*, e168.

(46) Bhowmick, A.; Head-Gordon, T. A Monte Carlo Method for Generating Side Chain Structural Ensembles. *Struct.* **2015**, *23*, 44–55.

(47) Towse, C. L.; Akke, M.; Daggett, V. The Dynameomics Entropy Dictionary: A Large-Scale Assessment of Conformational Entropy across Protein Fold Space. *J. Phys. Chem. B* **2017**, *121*, 3933–3945.

(48) Stites, W. E.; Pranata, J. Empirical-evaluation of the Influence of Side-chains on the Conformational Entropy of the Polypeptide Backbone. *Proteins* **1995**, *22*, 132–140.

(49) Wang, X. X.; Zhang, D.; Huang, S. Y. New Knowledge-Based Scoring Function with Inclusion of Backbone Conformational Entropies from Protein Structures. *J. Chem. Inf. Model.* **2018**, *58*, 724–732.

(50) Dinola, A.; Berendsen, H. J. C.; Edholm, O. Free-energy Determination of Polypeptide Conformations Generated by Molecular Dynamics. *Macromolecules* **1984**, *17*, 2044–2050.

(51) Harpole, K. W.; Sharp, K. A. Calculation of Configurational Entropy with a Boltzmann-Quasiharmonic Model: The Origin of High-Affinity Protein-Ligand Binding. *J. Phys. Chem. B* **2011**, *115*, 9461–9472.

(52) Baron, R.; Hunenberger, P. H.; McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theo. Comput.* **2009**, *5*, 3150–3160.

(53) Caldararu, O.; Kumar, R.; Oksanen, E.; Logan, D. T.; Ryde, U. Are Crystallographic B-factors Suitable for Calculating Protein Conformational Entropy? *Phys. Chem. Chem. Phys.* **2019**, *21*, 18149–18160.

(54) Wang, J.; Bruschweiler, R. 2D Entropy of Discrete Molecular Ensembles. *J. Chem. Theory Comput.* **2006**, *2*, 18–24.

(55) Nguyen, P. H. Estimating Configurational Cntropy of Complex Molecules: A Novel Variable Transformation Approach. *Chem. Phys. Lett.* **2009**, *468*, 90–93.

(56) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules. *J. Comput. Chem.* **2007**, *28*, 655–668.

(57) Fogolari, F.; Corazza, A.; Fortuna, S.; Soler, M. A.; VanSchouwen, B.; Brancolini, G.; Corni, S.; Melacini, G.; Esposito, G. Distance-Based Configurational Entropy of Proteins from Molecular Dynamics Simulations. *PLoS One* **2015**, *10*, e0132356.

(58) Hensen, U.; Grubmüller, H.; Lange, O. F. Adaptive Anisotropic Kernels for Nonparametric Estimation of Absolute Configurational Entropies in High-Dimensional Configuration Spaces. *Phys. Rev. E* **2009**, *80*, 011913.

(59) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y. P.; Gilson, M. K. Configurational Entropy in Protein-Peptide Binding: Computational Study of Tsg101 Ubiquitin E2 Variant Domain with an HIV-Derived PTAP Nonapeptide. *J. Mol. Biol.* **2009**, *389*, 315–335.

(60) Numata, J.; Knapp, E. W. Balanced and Bias-Corrected Computation of Conformational Entropy Differences for Molecular Trajectories. *J. Chem. Theory Comput.* **2012**, *8*, 1235–1245.

(61) Baruah, A.; Rani, P.; Biswas, P. Conformational Entropy of Intrinsically Disordered Proteins from Amino Acid Triads. *Sci. Reports* **2015**, *5*, 11740.

(62) Long, S. Y.; Wang, J. W.; Tian, P. Significance of Triple Torsional Correlations in Proteins. *RSC Advances* **2019**, *9*, 13949–13958.

(63) King, B. M.; Silver, N. W.; Tidor, B. Efficient Calculation of Molecular Configurational

Entropies Using an Information Theoretic Approximation. *J. Phys. Chem. B* **2012**, *116*, 2891–2904.

(64) Goethe, M.; Gleixner, J.; Fita, I.; Rubi, J. M. Prediction of Protein Configurational Entropy (Popcoen). *J. Chem. Theory Comput.* **2018**, *14*, 1811–1819.

(65) Hensen, U.; Lange, O. F.; Grubmüller, H. Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach. *PLoS One* **2010**, *5*, e9179.

(66) Fenley, A. T.; Killian, B. J.; Hnizdo, V.; Fedorowicz, A.; Sharp, D. S.; Gilson, M. K. Correlation as a Determinant of Configurational Entropy in Supramolecular and Protein Systems. *J. Phys. Chem. B* **2014**, *118*, 6447–6455.

(67) Suárez, E.; Suárez, D. Multibody Local Approximation: Application to Conformational Entropy Calculations on Biomolecules. *J. Chem. Phys.* **2012**, *137*, 084115.

(68) Cukier, R. I. Dihedral Angle Entropy Measures for Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2015**, *119*, 3621–3634.

(69) Gohlke, H.; Ben-Shalom, I. Y.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K. H. Rigidity Theory-Based Approximation of Vibrational Entropy Changes upon Binding to Biomolecules. *J. Chem. Theory Chem.* **2017**, *13*, 1495–1502.

(70) Sankar, K.; Jia, K. J.; Jernigan, R. L. Knowledge-based Entropies Improve the Identification of Native Protein Structures. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 2928–2933.

(71) Chong, S. H.; Ham, S. Dissecting Protein Configurational Entropy into Conformational and Vibrational Contributions. *J. Phys. Chem. B* **2015**, *119*, 12623–12631.

(72) Meirovitch, H.; Vasquez, M.; Scheraga, H. A. Stability of Polypeptide Conformational

States as Determined by Computer Simulation of the Free Energy. *Biopolymers* **1987**, *26*, 651–671.

(73) Cheluvaraja, S.; Meirovitch, H. Simulation Method for Calculating the Entropy and Free Energy of Peptides and Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9241–9246.

(74) Meirovitch, H. Methods for Calculating the Absolute Entropy and Free Energy of Biological Systems Based on Ideas from Polymer Physics. *J. Mol. Recognit.* **2010**, *23*, 153–172.

(75) Higham, J.; Chou, S. Y.; Gräter, F.; Henchman, R. H. Entropy of Flexible Liquids from Hierarchical Force-Torque Covariance and Coordination. *Mol. Phys.* **2018**, *116*, 1965–1976.

(76) Ali, H. S.; Higham, J.; Henchman, R. H. Entropy of Simulated Liquids Using Multiscale Cell Correlation. *Entropy* **2019**, *21*, 750.

(77) Henchman, R. H. Partition Function for a Simple Liquid Using Cell Theory Parametrized by Computer Simulation. *J. Chem. Phys.* **2003**, *119*, 400–406.

(78) Henchman, R. H. Free Energy of Liquid Water from a Computer Simulation via Cell Theory. *J. Chem. Phys.* **2007**, *126*, 064504.

(79) Klefas-Stennett, M.; Henchman, R. H. Classical and Quantum Gibbs Free Energies and Phase Behavior of Water Using Simulation and Cell Theory. *J. Phys. Chem. B* **2008**, *112*, 3769–3776.

(80) Henchman, R. H.; Irudayam, S. J. Topological Hydrogen-bond Definition to Characterize the Structure and Dynamics of Liquid Water. *J. Phys. Chem. B* **2010**, *114*, 16792–16810.

(81) Green, J. A.; Irudayam, S. J.; Henchman, R. H. Molecular Interpretation of Trouton's and Hildebrand's Rules for the Entropy of Vaporization of a Liquid. *J. Chem. Thermodyn.* **2011**, *43*, 868–872.

(82) Irudayam, S. J.; Henchman, R. H. Solvation Theory to Provide a Molecular Interpretation of the Hydrophobic Entropy Loss of Noble Gas Hydration. *J. Phys.: Condens. Matter* **2010**, *22*, 284108.

(83) Irudayam, S. J.; Henchman, R. H. Prediction and Interpretation of the Hydration Entropies of Monovalent Cations and Anions. *Mol. Phys.* **2011**, *109*, 37–48.

(84) Gerogiokas, G.; Calabro, G.; Henchman, R. H.; Southey, M. W. Y.; Law, R. J.; Michel, J. Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10*, 35–48.

(85) Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Heifetz, A.; Bodkin, M.; Law, R. J.; Henchman, R. H.; Michel, J. Assessment of Hydration Thermodynamics at Protein Interfaces with Grid Cell Theory. *J. Phys. Chem. B* **2016**, *120*, 10442–10452.

(86) Irudayam, S. J.; Plumb, R. D.; Henchman, R. H. Entropic Trends in Aqueous Solutions of Common Functional Groups. *Faraday Discuss.* **2010**, *145*, 467–485.

(87) Irudayam, S. J.; Henchman, R. H. Entropic Cost of Protein-Ligand Binding and its Dependence on the Entropy in Solution. *J. Phys. Chem. B* **2009**, *113*, 5871–5884.

(88) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

(89) Brooks, B. R.; Janezic, D.; Karplus, M. Harmonic-Analysis of Large Systems. 1. Methodology. *J. Comput. Chem.* **1995**, *16*, 1522–1542.

(90) Chakravorty, A.; Jia, Z.; Li, L.; Zhao, S.; Alexov, E. Reproducing the Ensemble Average Polar Solvation Energy of a Protein from a Single Structure: Gaussian-Based Smooth Dielectric Function for Macromolecular Modeling. *J. Chem. Theory Comput.* **2018**, *14*, 1020–1032.

(91) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(92) MacKerell, A. D. et al. All-atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(93) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(94) Karplus, M.; Ichiye, T.; Pettitt, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083–1085.

(95) Berezovsky, I. N.; Chen, W. W.; Choi, P. J.; Shakhnovich, E. I. Entropic Stabilization of Proteins and its Proteomic Consequences. *PLoS Comput. Biol.* **2005**, *1*, e47.

(96) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. The Double Cubic Lattice Method: Efficient Approaches to Numerical Integration of Surface Area and Volume and to Dot Surface Contouring of Molecular Assemblies. *J. Comput. Chem.* **1995**, *16*, 273–284.

(97) Tozzini, V. Coarse-Grained Models for Proteins. *Curr. Opin. Struc. Biol.* **2005**, *15*, 144–150.

(98) Riniker, S.; Allison, J. R.; van Gunsteren, W. F. On Developing Coarse-Grained Models for Biomolecular Simulation: A Review. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12423–12430.

(99) Saunders, M. G.; Voth, G. A. Coarse-Graining Methods for Computational Biology. *Annu. Rev. Biophys* **2013**, *42*, 73–93.

(100) Ingolfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The Power of Coarse Graining in Biomolecular Simulations. *WIREs Comput. Mol. Sci.* **2014**, *4*, 225–248.

(101) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and their Applications. *Chem. Rev.* **2016**, *116*, 7898–7936.

(102) Hunenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J. Mol. Biol.* **1995**, *252*, 492–503.

(103) Schäfer, H.; Smith, L. J.; Mark, A. E.; van Gunsteren, W. F. Entropy Calculations on the Molten Globule State of a Protein: Side-chain Entropies of α-lactalbumin. *Proteins* **2002**, *46*, 215–224.

(104) Fleck, M.; Polyansky, A. A.; Zagrovic, B. PARENT: a Parallel Software Suite for the Calculation of Configurational Entropy in Biomolecular Systems. *J. Chem. Theory Comput.* **2016**, *12*, 2055–2065.

(105) Fogolari, F.; Maloku, O.; Foumthuim, C. J. D.; Corazza, A.; Esposito, G. PDB2ENTROPY and PDB2TRENT: Conformational and Translational-Rotational Entropy from Molecular Ensembles. *J. Chem. Inf. Model.* **2018**, *58*, 1319–1324.

(106) Tidor, B.; Karplus, M. The Contribution of Vibrational Entropy to Molecular Association - the Dimerization of Insulin. *J. Mol. Biol.* **1994**, *238*, 405–414.

(107) Ma, B. Y.; Tsai, C. J.; Nussinov, R. A Systematic Study of the Vibrational Free Energies of Polypeptides in Folded and Random States. *Biophys. J.* **2000**, *79*, 2739–2753.

(108) Rossi, M.; Scheffler, M.; Blum, V. Impact of Vibrational Entropy on the Stability of Unsolvated Peptide Helices with Increasing Length. *J. Phys. Chem. B* **2013**, *117*, 5574–5584.

(109) Carrington, B. J.; Mancera, R. L. Comparative Estimation of Vibrational Entropy Changes in Proteins through Normal Modes Analysis. *J. Mol. Graph. Model.* **2004**, *23*, 167–174.

(110) Grünberg, R.; Nilges, M.; Leckner, J. Flexibility and Conformational Entropy in Protein-Protein Binding. *Structure* **2006**, *14*, 683–693.

(111) Richards, F. M. Areas, Volumes, Packing, and Protein Structure. *Annu. Rev. Biophys. Bio.* **1977**, *6*, 151–176.

(112) Klotz, I. M.; Franzen, J. S. Hydrogen Bonds between Model Peptide Groups in Solution. *J. Am. Chem. Soc.* **1962**, *84*, 3461–3466.

(113) Eberhardt, E. S.; Raines, R. T. Amide-Amide and Amide-Water Hydrogen Bonds - Implications for Protein-Folding and Stability. *J. Am. Chem. Soc.* **1994**, *116*, 2149–2150.

(114) Honig, B.; Yang, A. S. Free-Energy Balance in Protein Folding. *Adv. Prot. Chem.* **1995**, *46*, 27–58.

(115) Lovelock, D.; Rund, H. *Tensors, Differential Forms, and Variational Principles*; Dover Books on Mathematics Series; Dover, 1989.

Graphical TOC Entry