ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Entropies Derived from the Packing Geometries within a Single Protein Structure

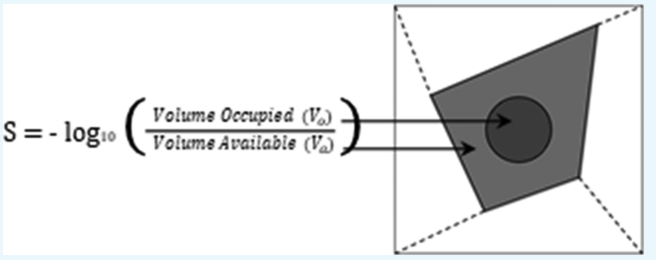
Pranav M. Khade and Robert L. Jernigan*

Cite This: *ACS Omega* 2022, 7, 20719−20730

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A fast, simple, yet robust method to calculate protein entropy from a single protein structure is presented here. The focus is on the atomic packing details, which are calculated by combining Voronoi diagrams and Delaunay tessellations. Even though the method is simple, the entropies computed exhibit an extremely high correlation with the entropies previously derived by other methods based on quasi-harmonic motions, quantum mechanics, and molecular dynamics simulations. These packing-based entropies account directly for the local freedom and provide entropy for any individual protein structure that could be used to compute free energies directly during simulations for the generation of more reliable trajectories and also for better evaluations of modeled protein structures. Physico-chemical properties of amino acids are compared with these packing entropies to uncover the relationships with the entropies of different residue types. A public packing entropy web server is provided at packing-entropy.bb.iastate.edu, and the application programing interface is available within the PACKMAN (https://github.com/Pranavkhade/PACKMAN) package.



INTRODUCTION

Entropy plays a crucial role in identifying a protein's lowest free energy conformation(s) in the course of structure prediction or modeling. Apart from these applications, entropy can also be key in providing insights into protein mechanisms, particularly for the role of disordered binding regions or intrinsically disordered proteins[1] and many other aspects of protein dynamics, mechanism, and interactions. Entropies are particularly important for assessing the effects of protein binding for both small ligands and to other proteins, where there are usually changes in entropy upon binding.[2−5] Unfortunately, almost all the current simulations or sampling of protein conformations ignore entropy, relying solely on energetics.

There have been many proposed methods to calculate entropies of biomolecules[6−11] that are based on molecular dynamics (MD) as well as discussions of experimental ways to measure entropies.[4,12,13] Many of these approaches calculate entropy by using MD or other simulations specifically to compute the fluctuations of atoms around their mean positions.[14−18] There is also quasiharmonic analysis (QHA) that approximates a probability distribution as a multi-dimensional Gaussian distribution derived from the coordinate fluctuations of atoms around their average positions in an MD simulation, yielding both vibrational frequencies and entropies[19−23] or methods that use multiple conformations from NMR ensembles or MD trajectories to calculate entropies.[24] The statistical distributions of torsion angles from MD simulations have also been used to calculate entropies.[25−27] One multiscale entropy calculation[28] even used MD simulations to obtain seven different contributing terms for the molecule, residue, and atomic levels to treat entropy as a sum of the different types of entropies. However, these methods all rely on relatively computationally expensive processes such as MD or other simulations. Another backbone entropy calculation method is Popcoen,[29] which does not directly rely on MD for entropy calculations. Instead, it is a neural network model trained on 961 protein MD simulations. Problems existing in MD, such as potential functions, parameter bias, butterfly effect, and possibly too short simulations, are also limitations for Popcoen. However, ignoring the entropies during the simulations is not the proper thing to do. Entropies should be explicitly included as a term that affects the trajectory pathway, and collecting data during a simulation for subsequent analysis does not accomplish this. Our entropy computation could enable the rapid evaluation of entropies for each conformation along a trajectory, so it would directly influence the trajectory and could be built into MD as a major improvement in MD methods.

All these previous methods rely on complex approaches and statistics to calculate a molecular property that has not been

Received: February 18, 2022
Accepted: May 17, 2022
Published: June 9, 2022

readily calculated for proteins. Experimental validation is not usually possible, so this cannot be used to determine which methods are more accurate than the others. Thus, there is a need for a simpler yet statistically robust method that can compute protein entropies in a far simpler way without relying on MD or more complex considerations *ex post facto*.

It may be important for some cases to consider entropies in terms of energy landscapes. If the sampling for entropy evaluation is performed locally on this surface, then one would be considering only conformations within a local basin. Sampling over the entire landscape means considering the collective entropies over all such basins. Clearly, using the present method only considers a highly localized set of conformations. The overarching question is whether this yields a sufficient sample or not. If the jumps between basins have high intervening barriers and are thus rare, this should be a valid approach; otherwise, a broader sample of conformations would be required. The sampling from MD should provide a test of whether this local sampling is sufficient. Indeed, as will be seen, there is excellent agreement between our results and entropies derived from MD simulations, so this permits us to conclude that local sampling should usually be sufficient. This suggests that performing this type of entropy evaluation on fly during the simulation could be an important way to improve MD simulations.

We know that there are limited numbers of favorable protein folds[30] mainly because of the stability of their hydrophobic interactions.[31−33] Because of these limited folds, the patterns in the protein structures have been used to analyze protein stability,[34] motion,[35] evolution, and function. Protein packing is another such concept that reflects the patterns within protein structures. We have seen that disturbances in protein packing because of mutations of different sizes and characteristics can lead to disrupted functions.[36] Therefore, we believe that protein packing can yield insights into protein stability as characterized by entropies. Over the past several years, we have investigated protein packing[37] as an important consideration for dynamics,[38,39] where we have demonstrated that hinges in proteins can be identified immediately from the static structure to be localized within the least densely packed regions. The present study aims to examine protein packing as the basis for protein entropy. Similar to the use of Delaunay tessellations to model protein packing, Voronoi diagrams[40] have also commonly been used[41−44] to characterize protein geometry. In this study, we employ a technique similar to packing fraction[45] or packing density[46] that has been used in polymer and liquid entropy characterizations to calculate protein entropies directly, as pioneered by Henry Eyring and extensively utilized by Flory.[47] Unlike all other entropy calculation methods, this method does not rely on any statistical considerations and instead is simply a measure of how much space is available within a static protein structure to move locally. Therefore, these data are free of most types of bias since they are based on single protein structures and their geometries only, and even though a complete sampling of these free volumes cannot occur completely independently, nonetheless, the underlying assumption of independence is an excellent one, as we will see.

In this study, we introduce a concept of "packing entropy" that is based on the packing fraction, which is defined as the volume occupied by the amino acid divided by the total volume available to the amino acid in a static structure. As shown below, the packing-based entropy has a high Pearson correlation coefficient with other completely different methods that use quasi-harmonic motion, quantum mechanics, simulations, and statistical distributions of backbone torsional angles along with other variables. We have, in addition, carried out one MD simulation to thermally denature a protein and analyzed the entropies over time and temperature. The present study demonstrates that the packing entropy of a single protein structure is usually sufficient to provide an excellent value of conformational entropy. There is an underlying assumption that a given structure does not undergo large-scale rearrangements. This approach is general and could be applied to any structure, including RNA or DNA and not just proteins.

MATERIALS AND METHODS

Example Proteins. We have used three different datasets in this study. Dataset 1 is a set of high-quality, diverse structures used to collect various statistics related to protein packing entropies, whereas Datasets 2 and 3 are obtained from different publications so that a direct comparison with the already published methods can be made.

Dataset 1. This dataset of proteins has been obtained from the Pisces server,[48] with the PDB IDs that are used as listed in Table S1. A total of 2079 unique protein-chain combinations are included in this dataset. This data is mainly used for deriving various statistics about protein packing entropy. The parameters for this collection are listed in Table 1

Table 1. Pisces Web Server[48] Culling Parameters

parameter	Value
percent sequence identity	≤15
resolution	0.0−1.5 Å
R-factor	<0.3
sequence length	40−10,000 residues
non-X-ray entries	Excluded
CA-only entries	Excluded
cull PDB by chain	True

Dataset 2. This dataset of proteins is obtained from our previous study on entropies,[49] with the PDB IDs listed in Table S2. This dataset can be found in the Supporting Information of the noted publication. The original authors generated this data to compare their method with other existing and well-established entropy calculation methods. This dataset is comprised of 30 examples that are used to compare packing entropies against those from Schlitter,[16] Andricioaei[18] method, and FoldX.[29]

Dataset 3. This dataset of proteins is obtained from the Chakravorty et al. study,[28] and the PDB IDs for the proteins are listed in Table S3. The authors have shown this data (the PDB IDs and corresponding entropy values) in the form of graphs in the referenced publication. The numerical entropy data shown was shared by the original authors. The original data contains 14 different terms (7 translational + 7 rotational) of entropy that we summed, as was done by those authors, for the comparison with the present packing entropies. This dataset includes 73 examples that are used to compare against the packing entropies from the QHA and MCC[28] methods.

All datasets are downloaded from the Protein Data Bank (PDB).[50]

Residue Entropy Calculation. We calculate the entropy by using the packing fraction[45] of the residues. The packing fraction is the ratio of the volume occupied by the atoms of amino acid *j*,

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

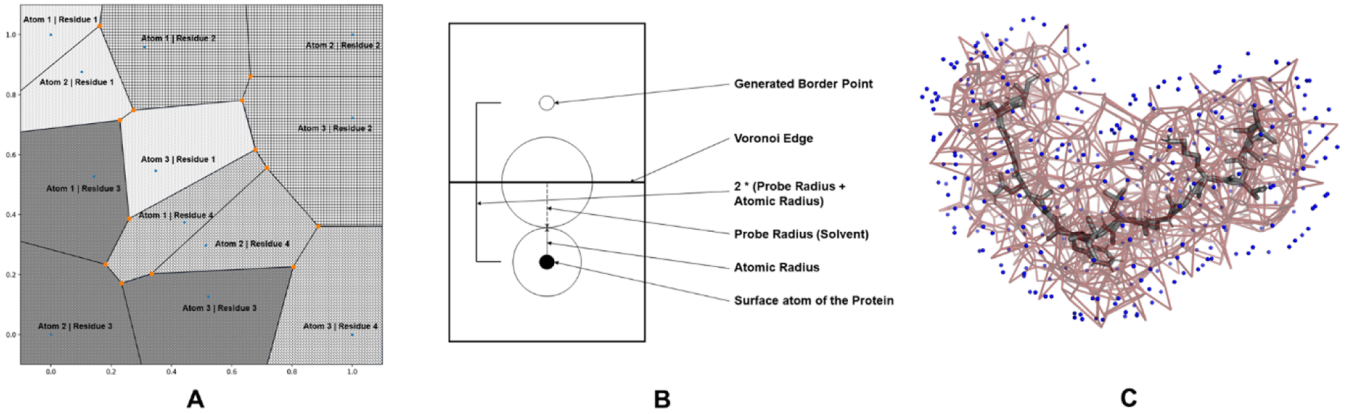


Figure 1. Voronoi diagram is generated for all atoms and after that atoms belonging to the same residue are combined (A) Voronoi diagram of a hypothetical example of an all-atom 2D structure; each cell is around an atom; lines separating (Voronoi edges) any two hypothetical atoms are equidistant from both atoms. Each pattern represents a Voronoi cell group (atoms belonging to the same residue). (B) Logic behind the Voronoi border determination in 2D space. The border points to limit the Voronoi diagrams should be generated at twice the distance of addition of the probe radius (usually solvent) and van der Waals radius. (C) Example of 3D Voronoi diagram; a Voronoi cell in 3D is derived identical to 2D; however, except that Voronoi edges are replaced by Voronoi planes. The gray-colored structure is an amino acid chain, and the blue points represent the border points described in the Methods section to trim the Voronoi diagram. The red lines represent the Voronoi plane boundaries.

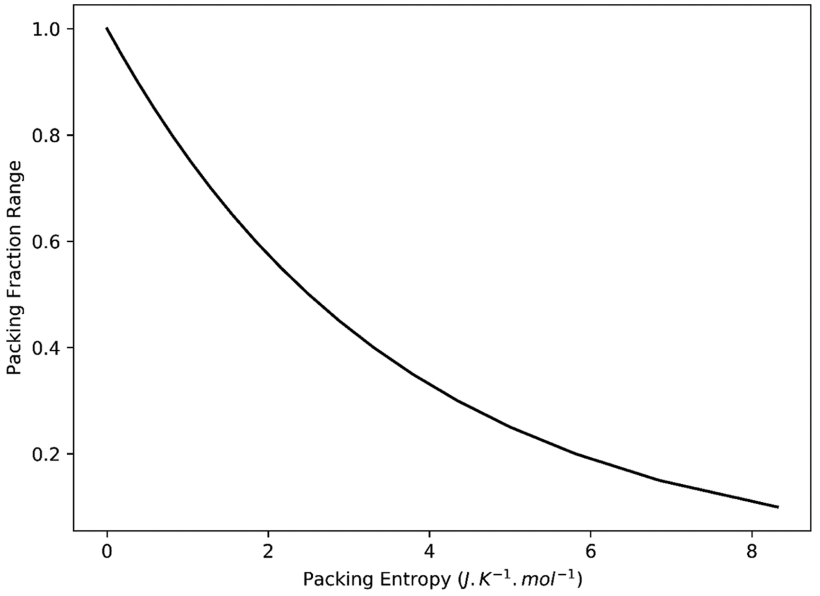


Figure 2. Packing entropy values (X-axis) are plotted against the packing fraction interval (Y-axis). This shows a non-linear relationship between entropy and packing fraction, as shown in eq 3.

$V_o(R_j)$ to the volume available, in the actual protein structure, to the amino acid j, $V_a(R_j)$, as identified in the Voronoi diagram.

Voronoi Tessellation. Voronoi tessellations or diagrams/polygons in 2D are also known as Dirichlet tessellations or Thiessen polygons.[51] They occur in nature in bubbles, dried-up wetlands, skin color patterns on a giraffe, and so forth and have been studied and applied extensively. In 2D space, given a set of distinct 2D points (p_i), a Voronoi diagram is a set of vertices and edges such that each Voronoi vertex is always equidistant to all of the three closest neighboring points (Figure 1A). The point that lies inside the "Voronoi cell" formed by the Voronoi vertices and edges around a particular atom is always closer to that atom than to any other atom in the Voronoi diagram. Also, the Voronoi edges are equidistant to both points that are being separated by it. We are interested in the Voronoi vertices obtained because they are used to calculate the volumes $V_a(R_j)$ using convex hulls.

Convex Hull. Given points p in 2D space, a convex hull is a polygon that contains all the points in p on the surface or inside the polygon and all the angles of the polygons are less than 180°. This is extended into 3D space here, with the polygon replaced by a polyhedron and 3D points replacing 2D points. The Convex hull of the Voronoi points is calculated to obtain the $V_a(R_j)$, and the convex hull of all the points in the particular residue is calculated to obtain $V_o(R_j)$.

Creating Boundaries for Voronoi Diagrams. Voronoi diagrams have boundaries stretching to infinity for most exterior points. This can cause a problem for the surface atoms of the proteins. To deal with this problem, we generate an arbitrary point cloud of 30 points (can be changed as a user parameter) around the surface protein atoms with a radial distance of 2 × (1.4 Å + van der Waal's atomic radius[52]) in such a way that no other atom in the protein is closer to any point in the generated cloud of points. This creates the Voronoi boundaries for the

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

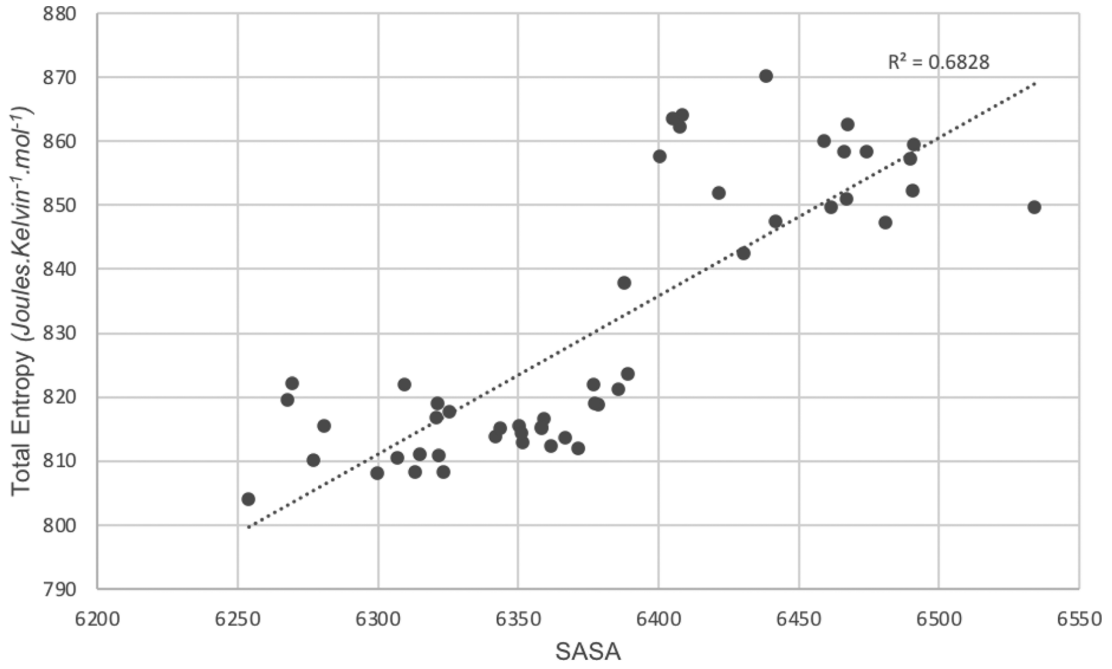


Figure 3. Total Entropy vs SASA. The R^2 score of the linear regression model (shown as a dotted line) is 0.6828. Pearson's correlation coefficient is 0.826 for the same data. The values are provided in Table S4.

atoms on the surface of the residue, as shown in Figure 1B. Doing this creates a boundary around the Voronoi diagrams, and volume-based analysis becomes more manageable. The choice of 30 points around each atom has been chosen so that there will be a sufficiently fine boundary around the protein that limits the outside of the 3D Voronoi diagram. The same concept explained in Figure 1B is easily extended to 3D structures, as shown in Figure 1C. Please read the PACKMAN documentation about generating these point clouds around each surface atom and adjusting the corresponding parameter.

Volume and Entropy Calculation. After this boundary is fixed, the Voronoi diagram[40] for all atoms in the protein along is calculated. In the Voronoi diagram, each atom will have a Voronoi tessellation/cell with multiple faces (planes), with each face being a boundary with a neighboring tessellation/cell surrounding itself so that the shared boundary between the neighbor atoms' tessellation will be equidistant from both atoms. To calculate the volume available to the residue $V_a(R_j)$ in protein i, we sum the volume of the convex hulls[53] built from the Voronoi boundary points (Voronoi vertices) of the atoms belonging to the same residue (K_j) as shown in eq 1. Since each atom's Voronoi boundaries will be equidistant from its neighbors, it will provide a good estimate of the volume available for each atom to move.

$$V_a(R_j) = \sum_{k=1}^{K_j} V(k) \tag{1}$$

The occupied volume of the residue j (in protein i) $V_o(R_j)$ is the volume of the Convex Hull[53] formed by all-atoms in the residue j. After obtaining the $V_a(R_j)$ and $V_o(R_j)$, we can calculate the packing fraction for the residue j $\mathrm{PF}(R_j)$ as

$$\mathrm{PF}(R_j) = \frac{V_o(R_j)}{V_a(R_j)} \tag{2}$$

The packing fraction of the residue j will measure how tightly it is packed in the protein structure. If the nearer $\mathrm{PF}(R_j)$ is 1, it is packed denser, and a value close to 0 indicates that the residue is sparsely packed and has much more room to move around. The packing fraction is then used to calculate the packing entropy for the residue j $S(R_j)$, as shown in eq 3, where R is the gas constant.

$$S(R_j) = -R \log_{10}(\mathrm{PF}(R_j)) \tag{3}$$

It is important to note that this relationship between the packing fraction and packing entropy has the non-linear form shown in Figure 2.

The total packing entropy of the protein can be calculated by summing all the entropies of the individual residues of the protein as shown in eq 4, where n is the number of residues in the protein.

$$\text{total packing entropy } (S) = \sum_{j=1}^{n} S(R_j) \tag{4}$$

Entropy Calculation for Multiple Conformations. This calculation can also be applied to any or all of the multiple structures from NMR structures or MD trajectories. If we consider F separate structures taken from an NMR structure of an MD trajectory, we can calculate an average as shown in eq 5, where R_{fj} is the jth residue in the fth structure.

$$\mathrm{PF}(R_j) = \frac{1}{F}\sum_{f=1}^{F} \frac{V_o(R_{fj})}{V_a(R_{fj})} \tag{5}$$

After this, the residue and protein entropies can be calculated with eq 4. It is important to note that the packing entropy for all the residues $S(R_{fj})$ and the protein f (S_f) can be calculated by treating each frame (time step) in the simulation as a separate structure; this can provide valuable insights into the motion and the mechanism of a protein as shown in Supporting Information Movie 1.

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

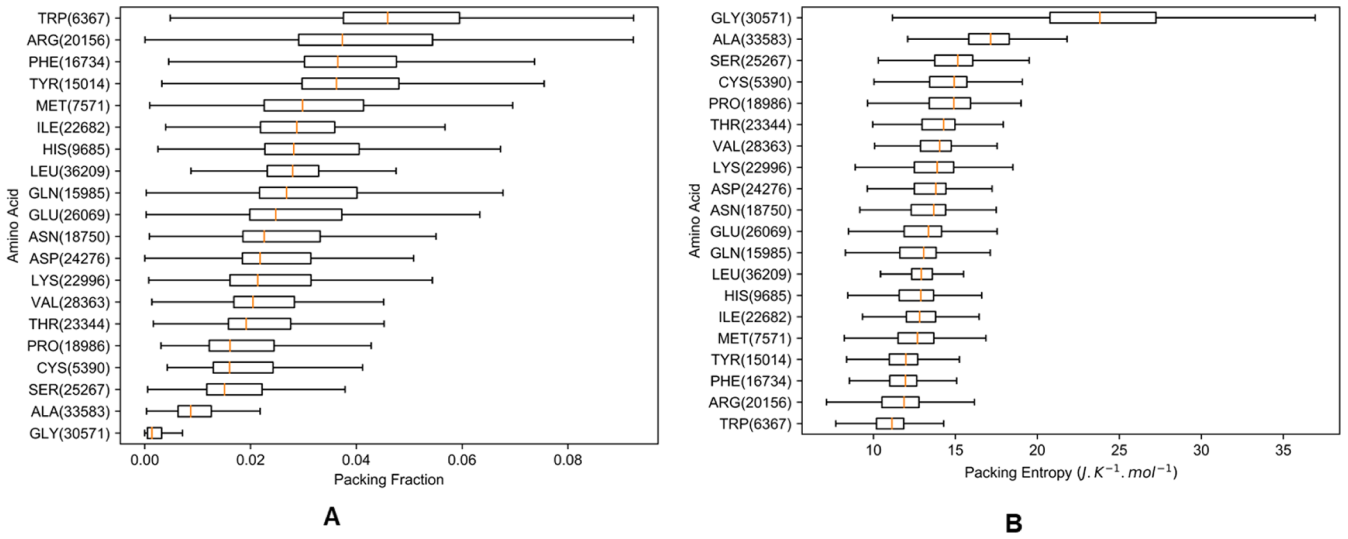


Figure 4. Box plots of residue type packing fractions and the corresponding packing entropies. The number of occurrences of the different types of amino acids in Dataset 1 for each residue is noted in parentheses following each amino acid three-letter code on *Y*-axis in both A and B. (A) Highest median value of packing fraction is around 0.05, which means that even for one of the largest amino acids, there is usually a significant volume available to move around and the relative difference in the volume occupied by the amino acids is reflected in the graph, together with overall higher variances for the larger amino acids. (B) There is a general trend for smaller amino acids to have higher entropies as well, with the high value for glycine being particularly notable. This means that they pack less tightly than the larger amino acids; these extreme values for glycine may reflect its frequent appearance on the surface and in turns, but also its lack of any side chain degrees of freedom, which suggests that side-chain flexibility is important to achieve the higher packing densities. Also, the means for most amino acid types lie between 10 and 30 J ° K^{-1} mol^{-1}. Entropy for each type of amino acid.

Entropy Units. The packing entropy for each residue is multiplied by gas constant (R) and takes on its units. The packing entropy units here are taken as Joules degrees Kelvin^{-1} mol^{-1}. We have compared the results with several other methods such as Schlitter, Andricioaei, Multiscale Cell Correlation (MCC), and QHA entropies that also used these same units.

Only Entropy Differences Are Meaningful. It is important to note is that when entropy is considered, only the entropy differences are meaningful, such as those between the total entropies of two different conformers of the same protein. We can compare different frames of an MD simulation or members of an NMR ensemble or even pairs of conformers generated from elastic network models to assess the entropy difference; more importantly, multiple conformers can also be analyzed in the presence of a ligand to obtain more information about entropy change in the process.

Protein L Denaturation Simulation. We carried out denaturation simulations of Protein L to compare with the results from Rocco et al.[54] for denaturation. We made slight changes to the simulation parameters (not the simulation box dimensions). Instead of 30 nanoseconds, we simulated the protein for 100 nanoseconds with thermal annealing from 300 to 550 K over 100 nanoseconds. Apart from the duration of the simulation, the rate of change in temperature, and the total time for the simulation, there is no change from the Rocco et al. procedure.[54] The ultimate goal is to follow the entropy during the denaturation of the protein.

RESULTS AND DISCUSSION

Protein L Denaturation Analysis. The denaturation of the protein was carried out to learn how well the Voronoi packing entropy can capture the increases in entropy of the protein during denaturation. Movie 1 was produced by taking 50 frames from our simulation. We can see from Movie 1 that as the temperature and time increase for protein L, the entropy increases and we observe the peak in entropy at model 38 (indexed from 0). The data for each frame is also available in Table S4. We also calculate Pearson's correlation coefficient for the solvent accessible surface area (SASA) and packing entropy that is equal to 0.826, and the linear regression model R^2 value shown in Figure 3 is 0.683. This means that the more exposed the surface area, the higher is its entropy, as would be expected as per the "hydrophobic effect".[5,31,55]

We have collected the values across all proteins in Dataset 1. It is clear from Figure 4A,B that certain amino acids tend to have specific packing fraction and entropy values depending on the amino acid type. We can see from the bar plots that the highest packing fraction and lowest entropy value is for tryptophan, which means, on average, that tryptophan tends to pack more densely than the other kinds of amino acids. As expected, glycine and alanine have the highest packing entropy values. Another notable entropic observation is for leucine; perhaps, because it has ∼36 k observations, its range of entropy values is relatively limited compared to the ranges for the other amino acids. Cysteine is an interesting case because it often forms a covalent bond with another cysteine. However, its packing entropy still has an extremely wide range of values, suggesting that chemically bonded cysteine pairs introduce rigidity that is not always so well accommodated within a given protein structure or that unlinked cysteine together with the covalently linked ones combine to have higher variability in entropy.

The packing fraction range lies between 0 to 0.1 for all of the amino acid types. This means that the occupied volume is less than 10% of the available volume. This, in some sense, is an artifact of the model's characteristics; the occupied volume is generated based on the fraction of the convex hull of all atoms of the residues, which because of its convex character, systematically overestimates to some extent the actual volumes, and the available volume is calculated using the Voronoi diagram that extends to 2*(solvent radius + van der Waal's radius). The

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

Table 2. Comparison of Entropies Calculated by Different Methods[a]

compared methods	Pearson's correlation coefficient (without normalization)	Spearman rank-order correlation coefficient (without normalization)	Pearson's correlation coefficient (with normalization)	Spearman rank-order correlation coefficient (with normalization)
Schlitter entropies–Andricioaei entropies (Dataset 2)	1	1	0.999	0.998
Schlitter entropies–Foldx entropies (Dataset 2)	0.991	0.981	0.873	0.575
Andricioaei entropies–Foldx entropies (Dataset 2)	0.99	0.981	0.863	0.565
Schlitter entropies–Packing entropies (Dataset 2)	0.979	0.932	0.939	0.856
Andricioaei entropies–Packing entropies (Dataset 2)	0.978	0.932	0.931	0.843
Foldx entropies–Packing entropies (Dataset 2)	0.983	0.947	0.959	0.734
QHA entropies–MCC entropies (Dataset 3)	0.985	0.988	0.954	0.685
QHA entropies–Packing entropies (Dataset 3)	0.86	0.861	0.899	0.723
MCC entropies–Packing entropies (Dataset 3)	0.885	0.881	0.935	0.672
Schlitter entropies–NRES (Dataset 2)	0.996	0.989	0.933	0.824
Andricioaei entropies–NRES (Dataset 2)	0.996	0.989	0.925	0.811
Foldx entropies–NRES (Dataset 2)	0.997	0.992	0.965	0.773
Packing entropies–NRES (Dataset 2)	0.984	0.95	0.998	0.988
MCC entropies–NRES (Dataset 3)	0.989	0.987	0.965	0.69
QHA entropies–NRES (Dataset 3)	0.967	0.973	0.925	0.713
Packing entropies–NRES (Dataset 3)	0.906	0.915	0.993	0.99

[a]The dataset used is noted in the bracket. NRES means the number of residues in the proteins for which the total entropy of the protein structure is calculated (length of the proteins). Normalized entropies are obtained by dividing the total entropy of the protein by $3N - 6$, where N is the number of heavy atoms in that protein.

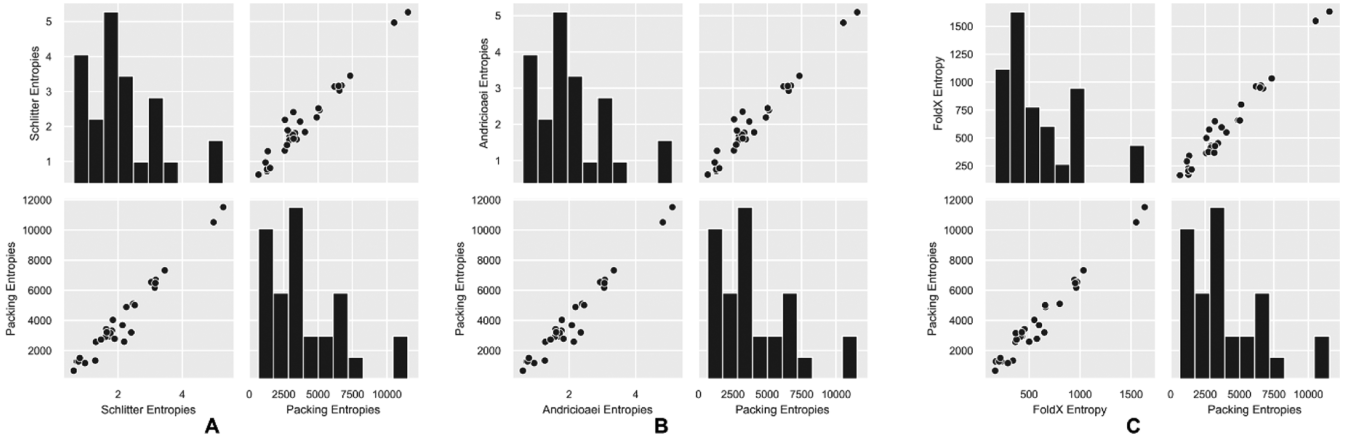


Figure 5. Comparison of entropies for the set of 30 proteins from Dataset 2. All entropies are in J ° K^{-1} mol^{-1}. (A) Packing entropy compared with Schlitter QM entropies.[16] (B) Packing entropy compared with Andricioaei MD entropies.[18] (C) Packing entropies compared with FoldX Entropies.[58] High correlations are observed for all cases as shown in Table 2.

critical information here is that the range of packing fractions are different for different amino acids, and many of them are distinct, although overlapping, in their ranges, as seen in Figure 4.

Comparison with Other Methods. We compared our results with well-established NMA and Quasi-harmonic methods, and the results are shown in Table S2. We obtained a 0.979 Pearson correlation coefficient with Schlitter entropies and 0.978 Pearson correlation coefficient with Andricioaei entropies (calculated using the appropriate columns in Table S2). Comparison of the methods with one another is listed in Table 2 along with the dataset used. It is important to note that

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

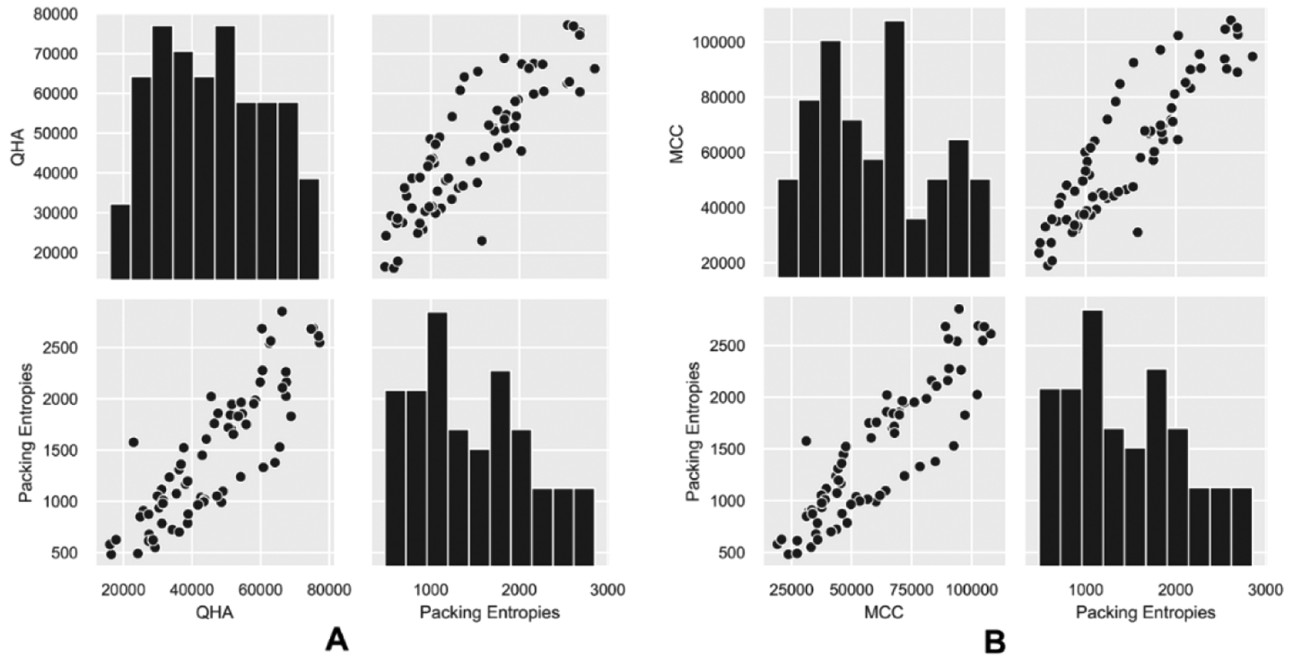


Figure 6. Comparison of entropies for the set of 73 proteins from Dataset 3 from the same study. All entropies are in J K^{-1} mol^{-1}. (A) Packing entropies compared with QHA.[28] (B) Packing entropies compared with MCC.[28]

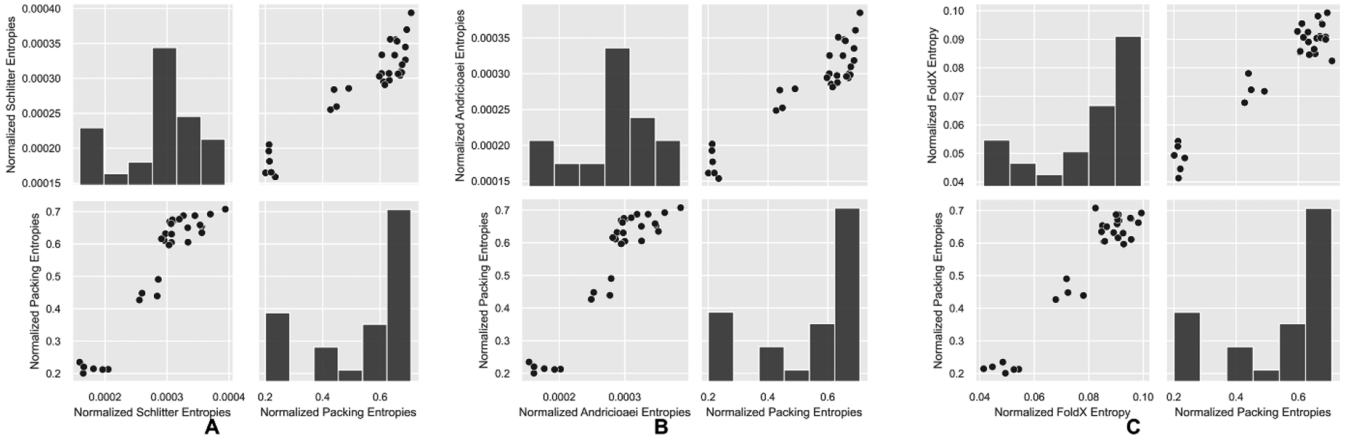


Figure 7. Comparison of entropies for the set of 30 proteins from Dataset 2. All entropies are in J ° K^{-1} mol^{-1}. Normalized entropies are obtained by dividing the total entropy of the protein by $3N - 6$, where N is the number of heavy atoms in that protein. (A) Normalized packing entropy compared with Normalized Schlitter QM entropies.[16] (B) Normalized packing entropy compared with Normalized Andricioaei MD entropies.[18] (C) Normalized packing entropies compared with normalized FoldX entropies.[58] High correlations are observed for all cases as shown in Table 2.

the Schlitter entropies are calculated using computer simulations and a quantum-mechanical approach, the Andricioaei entropies are calculated using the Quasi-harmonic method based on the covariance matrix obtained from MD, and our packing entropies are based only on a single conformation without carrying out any simulations. We also compared the results with the FoldX[58] entropies (Sidechain + Backbone) and obtain a Pearson correlation coefficient of 0.978. The linear correlation between the different entropies with the packing entropies can be seen clearly in Figures 5 and 6.

It is important to note that the total entropy correlation between different methods may not be the best way to check the performance of the method because we demonstrate in Table 2 that all of the methods' total entropies are highly correlated with the length of the protein with the help of Pearson's correlation coefficient. However, we can use the total entropy to compare the proteins of the same lengths/different conformations of the same proteins.

We also compared our packing entropies with the MCC entropies, where we obtained a Pearson correlation coefficient of 0.939, and QHA entropies, where we obtained a Pearson correlation of 0.924 (calculated using corresponding entropy columns from Table S3). Both data are the sum of all the entropy terms from the Chakravorty et al. study.[28] The comparison is tabulated in Table S3.

Normalized Entropies. We normalized the total entropies for dataset 2 and dataset 3 with $3N - 6$ where N is the number of heavy atoms. The results after the normalization are shown in Table 2, Figures 7, and 8. After the normalization, unlike before (Figures 5 and 6), we see that the entropies still are linear but form clusters instead of being evenly distributed along the diagonal. This might be because of the removed length bias with the normalization. We investigated both the Dataset 2 and 3 to

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

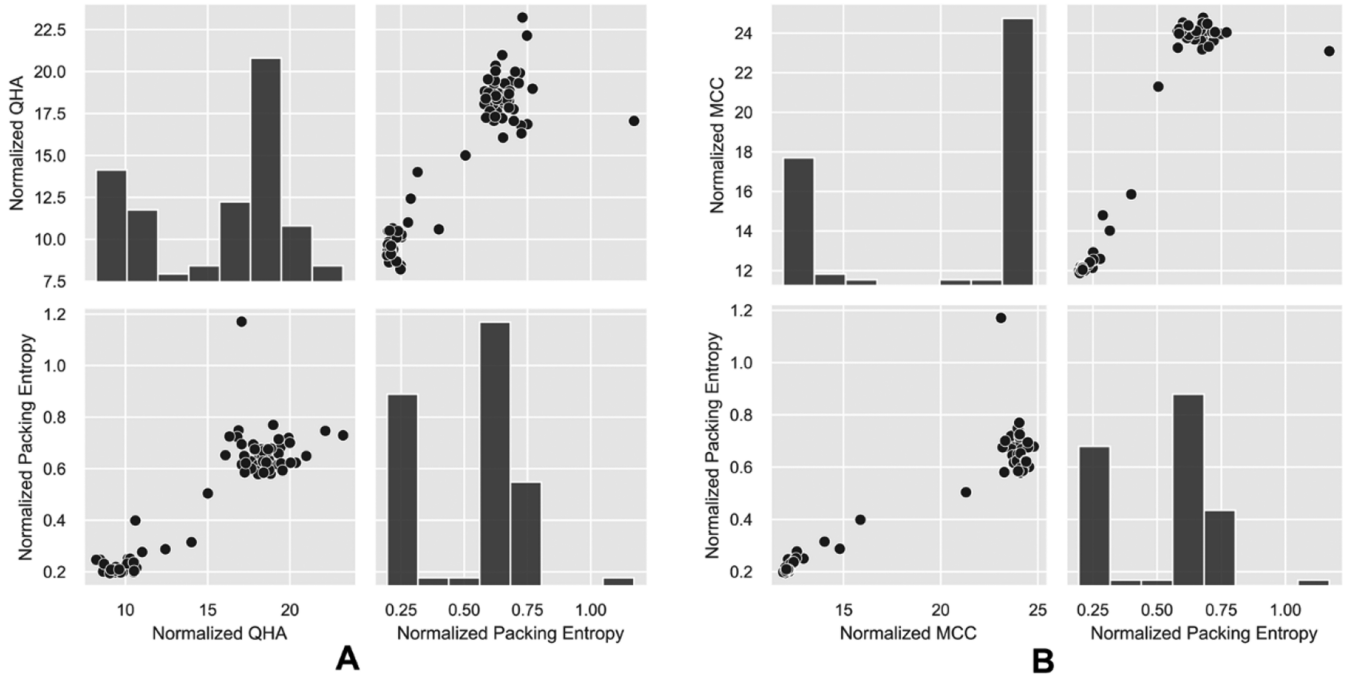


Figure 8. Comparison of entropies for the set of 73 proteins from Dataset 3 from the same study. All entropies are in J K^{-1} mol^{-1}. Normalized entropies are obtained by dividing the total entropy of the protein by $3N - 6$, where N is the number of heavy atoms in that protein. (A) Normalized packing entropies compared with normalized QHA.[28] (B) Normalized packing entropies compared with normalized MCC.[28]

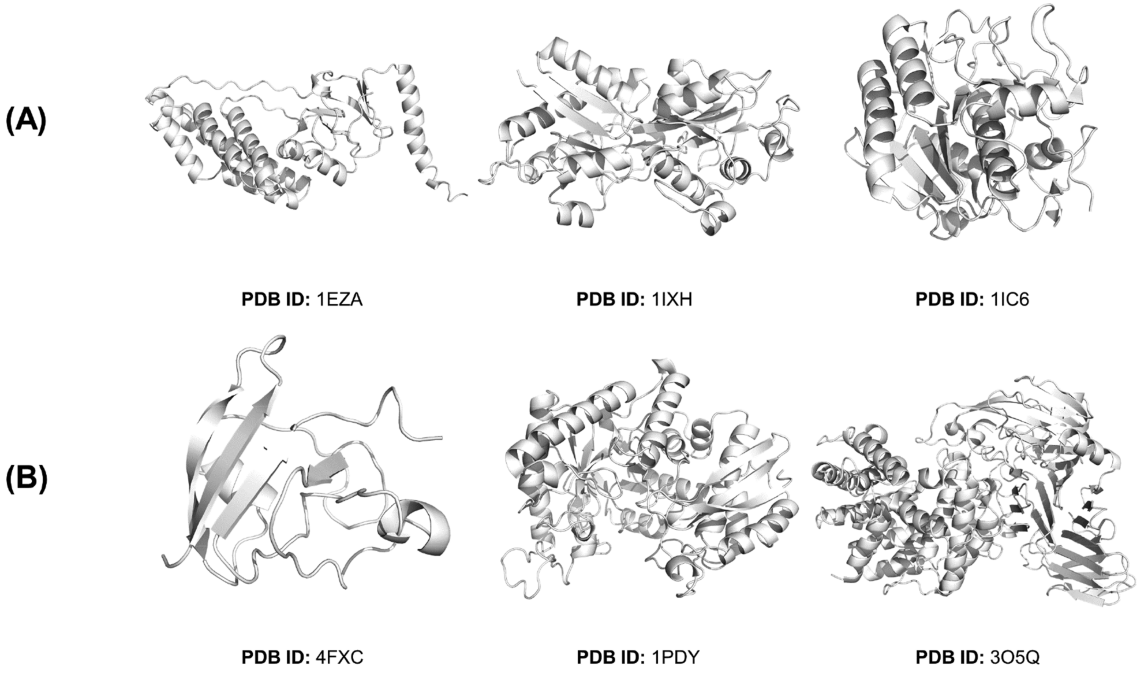


Figure 9. Subset of examples in Dataset 2 that are selected from individual clusters as seen in Figure 7C by sorting with normalized FoldX and normalized packing entropy values. (A) Structures sorted from low to high with normalized entropy appear as data points on the bottom left side of Figure 7C (B). Structures sorted from high to low with normalized entropy appear as data points on the bottom left side of Figure 7C.

find the subfigure with distinct cluster. We observed that normalized FoldX entropy (Dataset 2, presented in Figure 7) had visually distinct clusters forming when compared to the normalized packing entropy. We selected three examples each from both the clusters by sorting, and the examples can be seen in Figure 9A,B. We can see that there is no significant difference between the A and B figures, except we see more beta sheets in B and alpha helices in A. However, it is not a significant difference to notice. This might be because of the small number of examples in Dataset 2. Therefore, we decided to investigate Dataset 3 with 73 examples, where the MCC method had two distinct clusters. We followed the same procedure as described above. The Dataset 3 normalized entropy comparison can be seen in Figure 10, where we clearly see the low entropy cluster consisting of alpha helices and high entropy clusters being dominated by beta sheets. This could be because of the fact that

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

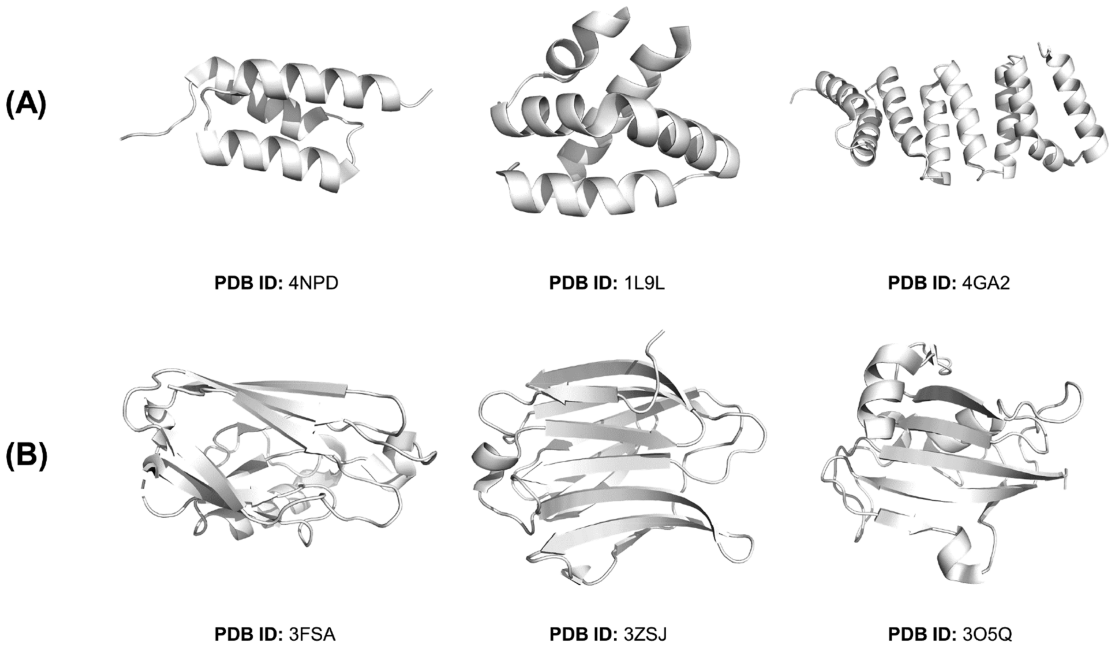


Figure 10. Subset of examples in Dataset 3 that are selected from individual clusters as seen in Figure 8 by sorting with normalized MCC and normalized packing entropy values. (A) Structures sorted from low to high with normalized entropy appear as data points on the bottom left side of Figure 8C (B). Structures sorted from high to low with normalized entropy appear as data points on the bottom left side of Figure 8C.

alpha helices have more interhelical contacts and therefore usually used by densely packed proteins such as hemoglobin.

Hydrophobicity and Kidera Factor Entropies. We also compare the packing entropies with the amino acid hydrophobicities on the Kyte−Doolittle scale[56] for Dataset 1, as shown in Figure 11. We observe regions with high density in various parts of the plot. This is significant and might be useful in the future if an empirical entropy model was developed for the packing entropies; such a model could be made continuous with respect to the hydrophobicity/other amino acid factors strongly. Also, this plot could have the same utility as detecting outliers on the Ramachandran plot to assess protein quality.

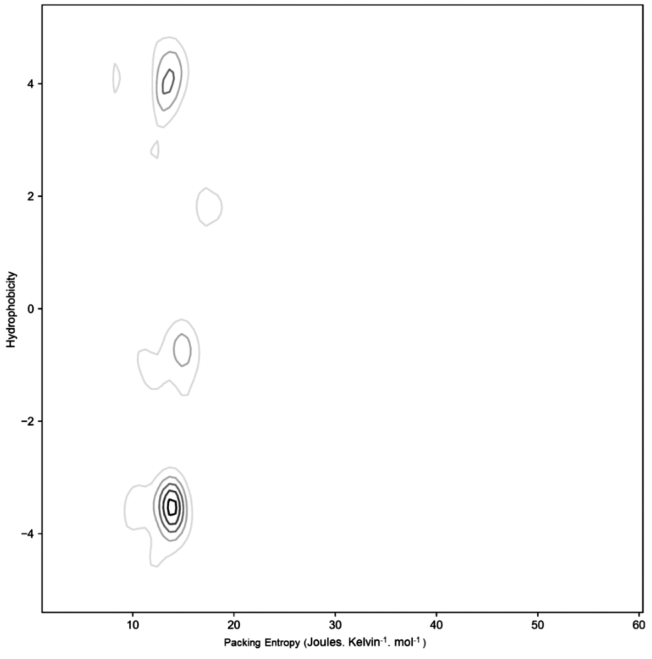


Figure 11. Contour plot of the packing entropy vs hydrophobicity. The data displayed is for 396,783 residues, and consequently, the contours are fairly smooth. Several noticeable peaks are observed on the different spectrum of hydrophobicity.

We have further investigated the packing entropy with 10 best Kidera factors[57] covering 86% variance as explained in the Kidera factors paper for the same purpose as the hydrophobicity comparison. Like hydrophobicity, Kidera factors are also sequence-based physical properties that can provide an informative landscape like the Ramachandran plot. The results are shown in Figure 12.

Comparison with *B*-Factors. We have compared the packing entropy for each residue in Dataset 1 using the Pearson correlation coefficient. We have compared the *B*-factors in four different ways: (1) by adding the *B*-factors for all the atoms in each residue, (2) by taking the median of *B*-factors of all atoms in the residue, (3) by taking the mean of *B*-factors for all atoms in a residue, and (4) by comparing Cα atom *B*-factors with the packing entropies. The detailed results from these comparisons are in Table S5. We observe correlations as high as 0.76 (higher *B*-factors are usually associated with higher entropies) for the Cα atom *B*-factors and above 0.70 for many other types of comparisons. We see a trend where the sum of *B*-factors is anti-correlated (−0.78) with the entropy for collagens and other fibrous proteins. We see similar values for the mean and median *B*-factor similarities with the Cα *B*-factors. However, it is important to note that *B*-factors may also depend on intermolecular interactions in the crystal.

Utility of Packing Entropy. The packing entropy's most significant attribute is that they do not depend on any statistical distribution/model or simulation. This creates an opportunity for energy-entropy simulations/conformer generation software to apply them without relying on any biased methods. Also, it is important to note that protein packing information could be as important as the torsional angle distributions or other parameters used for the protein dynamics.

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

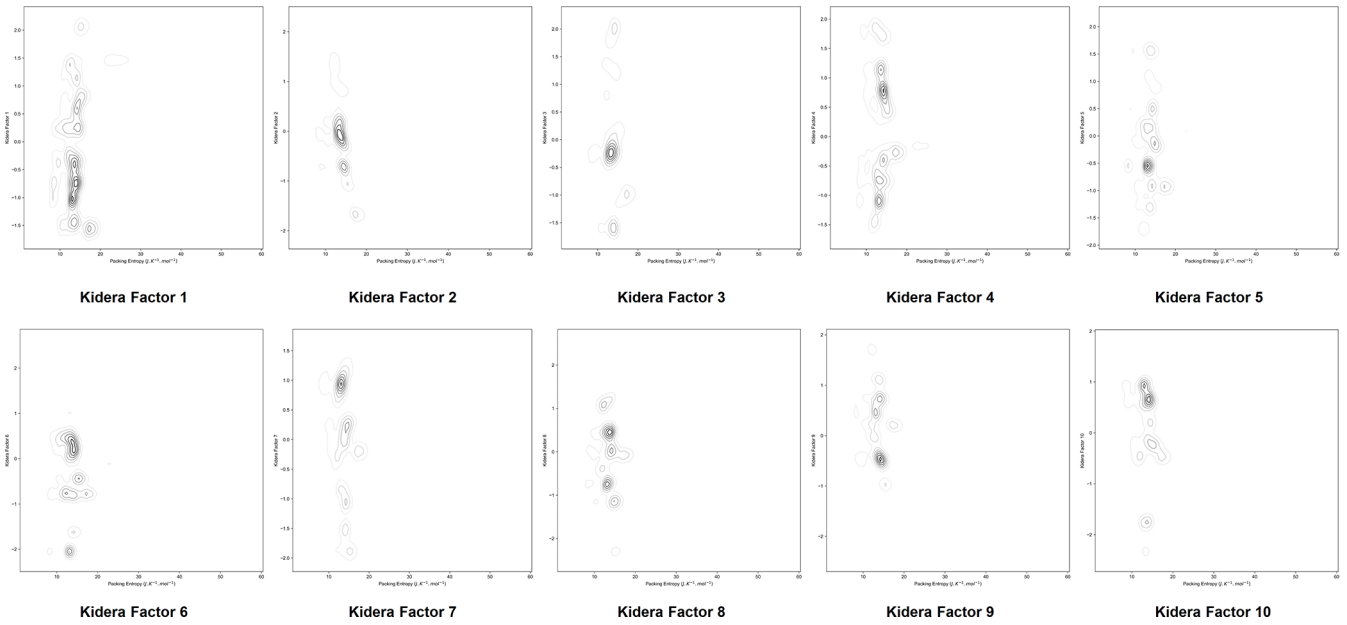


Figure 12. Packing entropy (*X*-axis) compared with various Kidera factors (*Y*-axis).

Although this method's strength is mainly distribution-independent, it is also possible to derive a distribution of protein packing entropies and use them as a contribution to the additive entropies of a system term in a hybrid approach. In the future, we will incorporate packing entropy into the entropy model built by Chakravorty et al.[28] as an additional term to the seven different entropy terms to explore this approach. Their method calculates all entropy terms and treats all the individual entropy terms as an additive. This means that packing entropy could easily be one more term that could be added to the others to obtain the total entropy of a protein.

In considering various critical aspects of this study, it is possible that certain Voronoi tessellation(s) have extreme acute angles that would lead to overestimates of volumes. However, these overestimates might not be a serious error since acute angles mean that a slight change in the neighboring atom's location could drastically affect the calculated Voronoi tessellation volume.

The packing entropy could also be useful to study mutations from the free volume perspective in terms of fitting additional atoms into the structure. If a bulky amino acid such as tryptophan was replaced by a much smaller amino acid, this would disturb the cohesion/hydrophobic core and would likely change the dynamics. There have been large-scale studies on such mutations.[59] The relative change in the entropy after mutation could provide an estimate of the disturbance to protein cohesion. The same observation could be made for replacing smaller with other larger amino acids and their effect on function.[60] It can also be helpful for enzyme design techniques where entropies are used.[61]

We are trying to explore whether packing entropy can be used to investigate the disordered regions in proteins. This is a particularly difficult challenge because of the absence of structural data for the disordered region. However, as a hypothesis, regions of the proteins with consistently high packing entropies could serve to identify the disordered regions.

Schlitter et al.[16] concluded that the time taken by simulations is a major drawback to the entropy calculation process. This difficulty is presently a problem for all simulation-based methods. However, the present packing entropies show that the entropy calculation does not have to rely on values derived from the simulations themselves. Simulation databases, such as the MoDEL database,[62] have trajectories of duration of 10 ns, and the μMoDEL database has results for 30 proteins extending from 0.1 up to 1 μs for the nMoDEL subset. This indicates that any simulation-based entropies will be limited to a specific timescale. Therefore, obtaining the entropies that are simulation-independent will be difficult to achieve.

The present method relies only on a single structure, which is its major advantage; it means that this consideration is independent of the method used to evaluate the entropies, which could be carried out by methods other than the ones used here. Also, reliance on eq 3 could likewise be changed.

This method is freely available as a web server at packing-entropy.bb.iastate.edu, and the application programing interface is also available in the PACKMAN package (https://github.com/Pranavkhade/PACKMAN).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c00999.

PDB IDs of Dataset 1; entropy values of Schlitter method, Andricioaei method, and Voronoi protein packing method; entropy values of QHA, MCC, and Voronoi protein packing method; total packing entropy (J K^{-1} mol^{-1}) and SASA calculated for the frames taken from the denaturation simulation of protein L; *B*-factor and packing entropy Pearson correlation coefficients for Dataset 1; residue specific comparisons for Cα atoms vs tip atoms, and a few specific references (PDF)

Motion and the mechanism entropy calculation for various progressing frames of the denaturation simulation of Protein L (MOV)

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730

AUTHOR INFORMATION

Corresponding Author

Robert L. Jernigan − *Bioinformatics and Computational Biology Program and Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011, United States;* orcid.org/0000-0003-0996-8360; Email: jernigan@iastate.edu

Author

Pranav M. Khade − *Bioinformatics and Computational Biology Program and Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011, United States;* orcid.org/0000-0002-0756-9828

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c00999

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank ResearchIT@iastate for help with the web server and for other support. We also thank Dr. Arghya Chakravorty and Dr. Richard H. Henchman for sharing Dataset 3. We gratefully acknowledge support from the NIH grant R01-GM127701 and the NSF grant DBI-1661391.

REFERENCES

(1) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C. H.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; Obradovic, Z. Intrinsically Disordered Protein. *J. Mol. Graphics Modell.* **2001**, *19*, 26.

(2) Sun, Z.; Yan, Y. N.; Yang, M.; Zhang, J. Z. H. Interaction Entropy for Protein-Protein Binding. *J. Chem. Phys.* **2017**, *146*, 124124.

(3) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21−42.

(4) Wand, A. J.; Sharp, K. A. Measuring Entropy in Molecular Recognition by Proteins. *Annu. Rev. Biophys.* **2018**, *47*, 41−61.

(5) Caro, J. A.; Harpole, K. W.; Kasinath, V.; Lim, J.; Granja, J.; Valentine, K. G.; Sharp, K. A.; Wand, A. J. Entropy in Molecular Recognition by Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 6563−6568.

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

(7) Meirovitch, H.; Cheluvaraja, S.; White, R. Methods for Calculating the Entropy and Free Energy and Their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr. Protein Pept. Sci.* **2009**, *10*, 229−243.

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

(9) Suárez, D.; Díaz, N. Direct Methods for Computing Single-Molecule Entropies from Molecular Simulations. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 1−26.

(10) Kassem, S.; Ahmed, M.; El-Sheikh, S.; Barakat, K. H. Entropy in Bimolecular Simulations: A Comprehensive Review of Atomic Fluctuations-Based Methods. *J. Mol. Graphics Modell.* **2015**, *62*, 105−117.

(11) Meirovitch, H. Recent Developments in Methodologies for Calculating the Entropy and Free Energy of Biological Systems by Computer Simulation. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181−186.

(12) 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.

(13) Fleck, M.; Polyansky, A. A.; Zagrovic, B. Self-Consistent Framework Connecting Experimental Proxies of Protein Dynamics with Configurational Entropy. *J. Chem. Theory Comput.* **2018**, *14*, 3796−3810.

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

(15) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14*, 325−332.

(16) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance Matrix. *Chem. Phys. Lett.* **1993**, *215*, 617−621.

(17) Schäfer, 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: Struct., Funct., Genet.* **2001**, *43*, 45−56.

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

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

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

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

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

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

(24) Gyimesi, G.; Závodszky, P.; Szilágyi, A. Calculation of Configurational Entropy Differences from Conformational Ensembles Using Gaussian Mixtures. *J. Chem. Theory Comput.* **2017**, *13*, 29−41.

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

(26) Zhang, J.; Liu, J. S. On Side-Chain Conformational Entropy of Proteins. *PLoS Comput. Biol.* **2006**, *2*, No. e168.

(27) 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.

(28) Chakravorty, A.; Higham, J.; Henchman, R. H. Entropy of Proteins Using Multiscale Cell Correlation. *J. Chem. Inf. Model.* **2020**, *60*, 5540−5551.

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

(30) Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225*, 563−564.

(31) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **2002**, *29*, 7133−7155.

(32) Chothia, C. Principles That Determine the Structure of Proteins. *Annu. Rev. Biochem.* **1984**, *53*, 537−572.

(33) Kuntz, I. D.; Kauzmann, W. Hydration of Proteins and Polypeptides. *Adv. Protein Chem.* **1974**, *28*, 239−345.

(34) Harpaz, Y.; Gerstein, M.; Chothia, C. Volume Changes on Protein Folding. *Structure* **1994**, *2*, 641−649.

(35) Jernigan, R. L.; Kloczkowski, A. Packing Regularities in Biological Structures Relate to Their Dynamics. *Methods Mol. Biol.* **2007**, *350*, 251−276.

(36) Eriksson, A. E.; Baase, W. A.; Zhang, X. J.; Heinz, D. W.; Blaber, M.; Baldwin, E. P.; Matthews, B. W. Response of a Protein Structure to Cavity-Creating Mutations and Its Relation to the Hydrophobic Effect. *Science* **1992**, *255*, 178−183.

(37) Liang, J.; Dill, K. A. Are Proteins Well-Packed? *Biophys. J.* **2001**, *81*, 751.
(38) Khade, P. M.; Kumar, A.; Jernigan, R. L. Characterizing and Predicting Protein Hinges for Mechanistic Insight. *J. Mol. Biol.* **2020**, *432*, 508.
(39) Scaramozzino, D.; Khade, P. M.; Jernigan, R. L.; Lacidogna, G.; Carpinteri, A. Structural compliance: A new metric for protein flexibility. *Proteins* **2020**, *88*, 1482.
(40) Aurenhammer, F. Voronoi diagrams-a survey of a fundamental geometric data structure. *ACM Comput. Surv.* **1991**, *23*, 345.
(41) Poupon, A. Voronoi and Voronoi-Related Tessellations in Studies of Protein Structure and Interaction. *Curr. Opin. Struct. Biol.* **2004**, *14*, 233−241.
(42) Cazals, F. Revisiting the Voronoi Description of Protein-Protein Interfaces. *Protein Sci.* **2006**, *15*, 2082.
(43) Andronov, L.; Orlov, I.; Lutz, Y.; Vonesch, J.-L.; Klaholz, B. P. ClusterViSu, a Method for Clustering of Protein Complexes by Voronoi Tessellation in Super-Resolution Microscopy. *Sci. Rep.* **2016**, *6*, 24084.
(44) Gerstein, M.; Tsai, J.; Levitt, M. The Volume of Atoms on the Protein Surface: Calculated from Simulation, Using Voronoi Polyhedra. *J. Mol. Biol.* **1995**, *249*, 955.
(45) Kumar, V. S.; Kumaran, V. Voronoi Cell Volume Distribution and Configurational Entropy of Hard-Spheres. *J. Chem. Phys.* **2005**, *123*, 114501.
(46) Richards, F. M. The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density. *J. Mol. Biol.* **1974**, *82*, 1−14.
(47) Flory, P. J. *Principles of Polymer Chemistry*; Cornell University Press, 1953.
(48) Wang, G.; Dunbrack, R. L. PISCES: A Protein Sequence Culling Server. *Bioinformatics* **2003**, *19*, 1589−1591.
(49) Sankar, K.; Jia, K.; Jernigan, R. L. Knowledge-Based Entropies Improve the Identification of Native Protein Structures. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 2928−2933.
(50) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(51) THIESSEN, A. H. PRECIPITATION AVERAGES FOR LARGE AREAS. *Mon. Weather Rev.* **1911**, *39*, 1082−1089.
(52) Bondi, A. Van Der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.
(53) Delaunay, B. Sur La Sphere Vide. *Izv. Akad. Nauk SSSR, Otd. Mat. Estestv. Nauk* **1934**, *7*, 793−800.
(54) Rocco, A. G.; Mollica, L.; Ricchiuto, P.; Baptista, A. M.; Gianazza, E.; Eberini, I. Characterization of the Protein Unfolding Processes Induced by Urea and Temperature. *Biophys. J.* **2008**, *94*, 2241−2251.
(55) Sturtevant, J. M. Heat Capacity and Entropy Changes in Processes Involving Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 2236−2240.
(56) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157*, 105−132.
(57) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A. Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. *J. Protein Chem.* **1985**, *4*, 23−55.
(58) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res* **2005**, *33*, W382.
(59) Gray, V. E.; Hause, R. J.; Fowler, D. M. Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics* **2017**, *207*, 53.
(60) Loo, T. W.; Clarke, D. M. Functional Consequences of Glycine Mutations in the Predicted Cytoplasmic Loops of P-Glycoprotein. *J. Biol. Chem.* **1994**, *269*, 7243−7248.
(61) Xie, W. J.; Asadi, M.; Warshel, A. Enhancing Computational Enzyme Design by a Maximum Entropy Strategy. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119*, No. e2122355119.
(62) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Pérez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpí, J. L.; Orozco, M. MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure* **2010**, *18*, 1399−1409.

https://doi.org/10.1021/acsomega.2c00999
ACS Omega 2022, 7, 20719−20730