

Atomic density distributions in proteins: structural and functional implications

Sotirios Touloupoulos , Nicholas M. Glykos ^{*} 

Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, 68100, Alexandroupolis, Greece

ARTICLE INFO

Keywords:

Protein structure
Hydrophobic core
Protein structure/function relationships
Packing density
Clustering structures

ABSTRACT

Atomic packing is an important metric for characterizing protein structures, as it significantly influences various features including the stability, the rate of evolution and the functional roles of proteins. Packing in protein structures is a measure of the overall proximity between the proteins' atoms and it can vary notably among different structures. However, even single domain proteins do not exhibit uniform packing throughout their structure. Protein cores in the interior tend to be more tightly packed compared to the protein surface and the presence of cavities and voids can disrupt that internal tight packing too.

Many different methods have been used to measure the quality of packing in proteins, identify factors that influence it, and their possible implications. In this work, we examine atomic density distributions derived from 21,255 non-redundant protein structures and show that statistically significant differences between those distributions are present. The biomolecular assembly unit was chosen as a representative for these structures. Addition of hydrogen atoms and solvation was also performed to emulate a faithful representation of the structures *in vitro*.

Several protein structures deviate significantly and systematically from the average packing behavior. Hierarchical clustering indicated that there are groups of structures with similar atomic density distributions. Search for common features and patterns in these clusters showed that some of them include proteins with characteristic structures such as coiled-coils and cytochromes. Certain classification families such as hydrolases and transferases have also a preference to appear more frequently in dense and loosely-packed clusters respectively.

Regarding factors influencing packing, our results support knowledge that larger structures have a smaller range in their density values, but tend to be more loosely packed, compared to smaller proteins. We also used indicators, like crystallographic water molecules abundance and B-factors as estimates of the stability of the structures to reveal its relationship with packing.

1. Introduction

Atomic packing has been an important metric for characterizing protein structures since 1974, when it was observed that the average packing density within proteins' interiors is roughly equivalent to that of small organic molecule crystals [1]. Although numerous methods had been developed to calculate the packing and interactions of amino acid residues within proteins, the use of packing density as a criterion for evaluating model protein structures was developed explicitly in 1990 [2].

To date, several approaches have been tested to measure the atomic packing in structures. The Voronoi procedure is a widely-used method, in which a unique volume is assigned to individual atoms in order to

study variations in packing of proteins [3–7]. Another well-established method for analyzing packing interactions in proteins is based on the calculation of the occluded molecular surface [8]. Other methods and approaches for analyzing protein packing have also been reported [9–15].

Packing is an important aspect of protein structures. A compact packing of amino acid residues is known to affect both the thermal stability and folding rate of proteins. [16–25]. Protein stability is of significant interest to the biotechnological, pharmaceutical, and food industries. The effects of packing on protein stability has extensively been studied to the point that modeling programs have incorporated packing as a parameter, aiming to predict protein stability after mutations [26]. Moreover, hydrogen bonds, which increase the packing

^{*} Corresponding author. <https://utopia.duth.gr/glykos/>

E-mail address: glykos@mbg.duth.gr (N.M. Glykos).

<https://doi.org/10.1016/j.jmglm.2025.109228>

Received 17 August 2025; Received in revised form 8 November 2025; Accepted 17 November 2025

Available online 19 November 2025

1093-3263/© 2025 Elsevier Inc. All rights reserved, including those for text and data mining, AI training, and similar technologies.

density in the protein interior are known for their indirect contribution to protein stability [27]. Recent studies indicated that the major determinants of protein stability include packing and van der Waals interactions [28–30].

As the structural comparison shown in Figs. 1 and 2 exemplifies, proteins do not exhibit uniform packing throughout their structure [31]. Localized packing defects appear as cavities, and their presence can compromise the stability of the protein [32]. Additionally, the distribution of these voids (cavities) is highly heterogeneous across different proteins [33]. Despite the presence of occasional cavities, the interior of spherical proteins remains tightly packed. The Voronoi volumes of surface atoms, modeled with solvent surrounding the protein, are approximately 7 % larger [34,35], indicating that packing is less dense on the protein surface.

Experimental studies have shown that mutations in protein cores, where small residues are replaced with larger ones, generally destabilize the protein. This suggests that there is minimal empty space available to accommodate additional atoms [23,36]. This can be explained by the α -helical and β -sheet secondary structures in globular proteins. These elements organize in a manner that allows non-polar side chains to interlock like jigsaw puzzle pieces, creating densely packed cores. As a result of this tight packing, van der Waals forces are considerably stronger in the interior [27]. However, due to energetically unfavorable atomic overlaps, protein cores cannot exceed some density limits [37]. The rigidity of protein cores is also shown to be strongly correlated with packing density [38,39]. Furthermore, studies have shown that the interior of proteins evolves slowly, in contrast to the surface which has more rapid evolution [40,41]. Solvent accessibility has become the *de facto* structural measurement to use in protein evolution studies. However, more recent work has called the central role of solvent accessibility into question and has identified packing as an important factor too [42]. The two packing measures most frequently employed in evolutionary studies are the **contact number** and the **weighted contact number**. For a given amino acid, the **contact number** represents the total count of other residues within its local structural neighborhood. In contrast, the **weighted contact number** considers all residues in the protein, assigning weights to them based on the square of their inverse distance

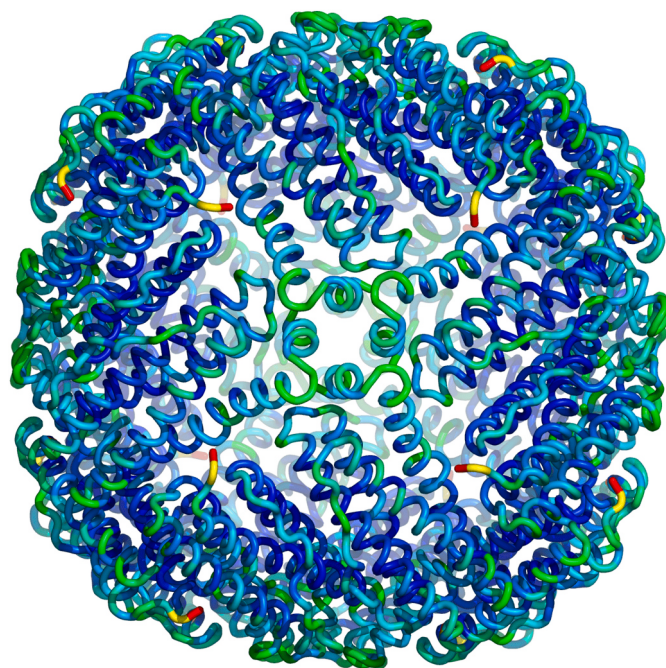


Fig. 1. Schematic diagram of a ferritin-like protein structure with a low density profile. This is the biological assembly corresponding to PDB entry 3r2k colored according to atomic temperature factors.

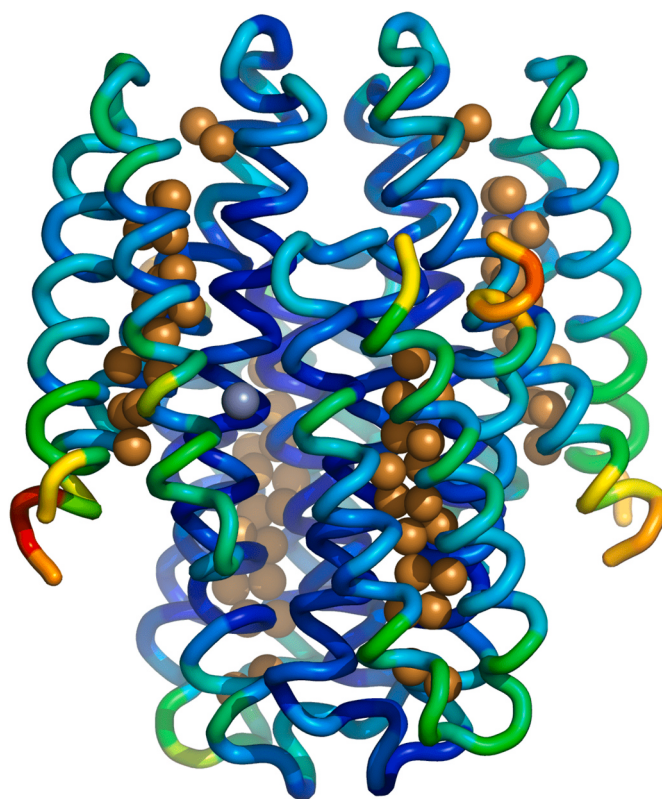


Fig. 2. Schematic diagram of the structure of a protein with a high density profile. This is a cytosolic copper storage protein (PDB entry 6zif) colored according to atomic temperature factors, with the copper atoms indicated as solid spheres.

to the amino acid under examination [43,44].

Atomic packing is affected by a combination of different factors. A statistical analysis of the radius of gyration for 3769 protein domains across four major classes (α , β , α/β , and $\alpha+\beta$) revealed that each class exhibits a characteristic radius of gyration, indicating its specific level of structural compactness. For example, α -helical proteins exhibit the highest radius of gyration across the considered protein size range, indicating a less compact packing compared to β and $(\alpha + \beta)$ proteins. In contrast, α/β proteins display the lowest radius of gyration, characteristic of the most compact packing among the classes [45].

Another study showed that for proteins with a molecular weight below 20 kDa, the average density shows a positive deviation that becomes more pronounced as molecular weight decreases, indicating that smaller proteins are more densely packed than larger ones, which tend to have a looser packing structure [46]. Additionally, an analysis of 152 non-homologous proteins demonstrated that variations in protein packing are influenced by a complex interplay of protein size, secondary structure, and amino acid composition. They showed that helices appear to be more efficiently packed compared to strands and that large proteins are expected to have increased overall packing [47].

In this communication we attempt to approach the problem of characterizing and analyzing protein density not through average statistical or structural properties, but by building and directly comparing individual density profiles which were created for an extended set of more than 21,000 proteins. The essence of our approach is the following: For each atom of each structure we calculate the density (in $\text{Da}/\text{\AA}^3$) inside a sphere centered on that atom. If, for example, a given protein structure contains 5000 atoms, then we would calculate 5000 density values (one for each atom). These density values are then used to calculate a histogram of their distribution which is characteristic of the protein structure under examination. Having collected the density

distributions, we can quantify and analyze their similarities and differences using established metrics such as the Euclidean distance (calculated between any given pair of distributions). By doing an all-to-all comparison of those distributions, we can quantitatively characterize structural and functional patterns present in these distributions, as well as evolutionary relationships between diverse families of proteins. In the following paragraphs we present details of this method, and of the structural, functional, and evolutionary results obtained from its application.

2. Methods

2.1. Calculation of density distributions

The starting set of protein structures (see previous paragraph) is an extended sample obtained from the PDB and comprising representative culled proteins (identity cutoff $\leq 50\%$), hydrogenated and hydrated to simulate *in vitro* conditions (see section §2.2 below for details of how the representative structures were obtained and how hydrogens and waters were added). Throughout our calculations we have tested three different sphere radii (5 Å, 6 Å, 7 Å) to remove the bias that this otherwise arbitrary choice would incur. Greater values (>7 Å) have also been tested but were found lacking the resolution needed by our method. In the first stage, our method parses the x,y,z coordinates of atoms and calculates the distance between every possible combination of atoms (i, j). For any given radius R (5 Å, 6 Å or 7 Å), it calculates whether the distance between the atoms i and j is smaller than R , and if so, the mass (in Dalton) of atom j is counted as lying inside the sphere of atom i . This is performed recursively for every atom, and after division with the volume of the sphere, the atomic density distribution of the protein under examination is obtained (in units of Dalton per cubic Ångström). As will be discussed later, our choice to perform the analysis in units of Dalton/Å³ is crucial for being able to differentiate functionally important groups of proteins (for example, metal-containing proteins). We should also note that this approach is based on counting only the presence or absence of atoms lying inside spheres centered on every other atom. Each atom contributes as a whole to the density calculation when it lies inside a sphere, with no correction being made for the part of the atom's volume that is outside the sphere. Finally, we note that this algorithm is only applied to protein atoms (and not, for example, to the water molecules that were added to simulate a fully hydrated protein structure or to the protein's ligands). Fig. 3 shows a graphical representation of this procedure.

2.2. Representative structures and system preparation

The PISCES server [48] was used to obtain a set of proteins with diverse structural and functional characteristics from the Protein Data Bank.

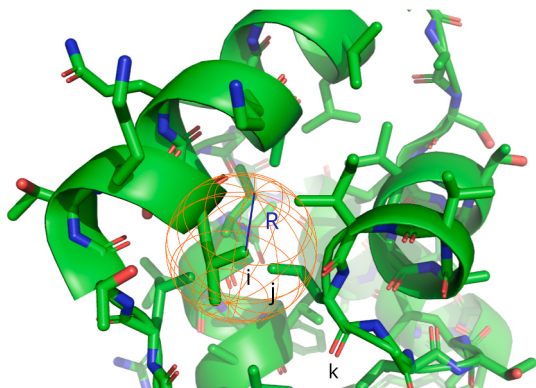


Fig. 3. Graphical representation for centering atoms on the calculation of density.

Bank. The full list of selection criteria given to PISCES are shown in Table 1, which resulted to a set of 22478 structures. For all structures, the biological assembly was used for all further calculations. An additional cutoff of a maximum of 80,000 protein atoms per structure was applied to final list of structures.

OpenBabel [49,50] was used to add missing hydrogen atoms to PDB files. The program Solvate (<https://www.mpinat.mpg.de/grubmueller/solvate>) was used to perform hydration of structures by adding and energy minimizing a sufficiently large box of pre-equilibrated water molecules around the solute, emulating a physically convincing fully hydrated protein structure *in vitro*. Care was taken to avoid adding water molecules in buried protein cavities. Fig. 4 shows a schematic illustration of this procedure for the case of a small protein.

3. Results

3.1. Probability distributions indicate the presence of significant density variability

The density distributions (one for each protein examined) were calculated as described in section §2.1. To simplify the subsequent calculation of distance metrics between different distributions, the same number of histogram bins (equal to 100) was used for all proteins. This choice for the number of bins was guided by the application of the Freedman-Diaconis rule to a randomly selected subset of protein structures. In the final step of preparing the initial data set, the individual distributions were normalized by dividing with the total number of atoms of each protein, thus converting the units to frequencies (of observing the corresponding density, see Fig. 5 below). The final data set comprises three (21255×100) matrices, corresponding to the three radii we examined (5 Å, 6 Å, 7 Å). In each of these matrices, every row corresponds to a different structure and every column to a different bin from the density distribution of the given protein (and for the given radius). Fig. 5 shows the probability density distributions (i.e., over all 21255 proteins) for each of the radii we examined.

It can be seen that several data points are significantly distant from the mean, with several of them deviating by more than 3σ from the average. This implies that there could be certain distributions that would be characterized as ‘outliers’ when compared with the average distribution. The observation that as the radius increases, the distributions become tighter about their mean is fully consistent with our expectations: as the volume of the spheres increases, the density calculated from each sphere approaches the same value (which is the average density of protein structures). Equivalently, as the radius is increasing, the ‘high resolution’ information about the variation of density inside a protein structure is diminished due to extensive averaging.

3.2. Principal component analysis allows the identification of unusual density distributions

Before proceeding with the main theme of our analysis –which is based on the clustering of the primary data through hierarchical clustering methods (discussed below)– we used PCA as a preliminary step to

Table 1

Criteria given to PISCES server for culling the Protein Data Bank.

Criteria	Value
Resolution	0.0–2.2
R-factor	0.25
Sequence length	50–10000
Sequence percentage identity	≤ 50.0
X-ray entries	Included
EM entries	Excluded
NMR entries	Excluded
Chains with chain breaks	Included
Chains with disorder	Included

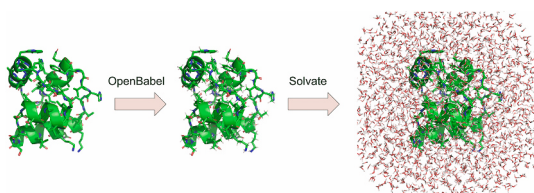


Fig. 4. Preparation of the final systems before density calculation: Addition of hydrogens using *OpenBabel*, and of a pre-equilibrated water box using *Solvate*.

visualize the distribution of data in the reduced principal component space. Fig. 6(a,b,c) show the PCA distributions obtained from the matrix calculated with the 6 Å radius and projected on the top three principal components. We observe that the data points are not harmonically distributed in these projections and that they deviate significantly from two-dimensional Gaussian distributions. There is a pronounced high density area (corresponding to an ‘average’ density distribution), but significant excursions from normality are immediately obvious (for example the tails clearly seen in Fig. 6(a)). To help visualize the amount of deviation present in the density distributions, we show in Fig. 6(d) the average density distribution (black line) versus the density distributions obtained from two outliers (structures 1JOP and 3NIO). The 1JOP structure is a cytochrome and its density distribution is markedly shifted to the right (higher density values). The 3NIO structure on the other hand is shifted to the left (lower density values) and corresponds to a guanidinobutyrase protein. This is an early indication that structures with uncommon distributions exist. Further examination of outliers reveals structures with a preference in light-harvesting (cytochromes) and copper-storage proteins. The biological significance of these deviations from a harmonic behavior are analyzed in the following sections.

3.3. Hierarchical clustering allows the identification of distinct groups of proteins

Starting from the three (21255×100) matrices corresponding to the three radii we examined (see section §3.1), we calculated the corresponding distance matrices by calculating the Euclidean distance between all possible pairs of distributions. These distance matrices are the primary data upon which hierarchical clustering methods are based. Fig. 7 shows a visualization of these symmetrical distance matrices in the form of heatmaps where distances are encoded as colors ranging for dark blue (small distances), through yellow (intermediate distances), to red (large distances). The lower three panels in Fig. 7 show the same matrices but after exclusion of the 100 most distant proteins (this was done in order to increase the dynamic range of these graphs). Please note that all matrices shown in Fig. 7 have been scaled to the same maximum

distance (they are on the same color scale).

The general appearance of the distance matrices shown in Fig. 7 is fully consistent with the results discussed in sections §3.1 and §3.2. For example, the increase of the average distances as we move from the 5 Å matrix, to 6 Å, and to 7 Å (going from mostly blue colors to mostly yellows in Fig. 7), is consistent with the tighter and higher density distributions seen in Fig. 5: Higher values of frequencies (as seen, for example, in Fig. 5(c)) lead to larger on average Euclidean distances, leading to the systematic trend observed in Fig. 7. To put this in numbers, Table 2 shows the averages, standard deviations and maximal distances recorded for the three distance matrices shown in Fig. 7.

The second important feature of these matrices concerns their internal consistency (which, however, is more difficult to discern due to their relatively low contrast of the heatmaps). Closer examination of the graphs in Fig. 7, however, does show that the patterns of small/large distances (dark/light colors) is more-or-less the same irrespectively of which matrix is being examined. To put this observation in numbers, we compared these three matrices by calculating the values of the linear correlation coefficient between all possible pairwise combinations. The comparison of the (5 Å matrix) vs. (6 Å matrix) gave a value of the linear correlation coefficient of +0.86, the 6 Å vs. 7 Å comparison gave a value of +0.93, and the 5 Å-7 Å combination a value of +0.76. The fact that the matrices are so similar and internally consistent is reassuring: it implies that the subsequent calculation of dendrograms is robust and not highly sensitive to the value of the averaging radius. The lower correlation for the 5 Å-7 Å pair, combined with the value of +0.93 for the 6 Å-7 Å pair indicates the information content of the matrices has stabilized (converged) once we reach the 6 Å radius. For this reason, all further calculations reported in this communication were based on the matrix calculated with the 6 Å averaging radius [diagrams (b) and (e) in Fig. 7].

Fig. 8(a) shows the dendrogram obtained by performing hierarchical clustering of the 6 Å distance matrix using the R package for statistical computing. Hierarchical clustering produced well-separated clusters of structures, with the individual clusters having similar atomic density distributions and being of a size suitable for further statistical analysis. Other clustering algorithms such as k-means and HDBSCAN have also been tested, but failed to produce well-separated clusters (which is not unexpected given the relatively uniform distribution of the raw data as principal component analysis clearly indicated, see Fig. 6). However, k-means and HDBSCAN did provide additional information on outliers or small groups of structures with common characteristics.

The fact that the dendrogram is well-structured does not alleviate the problem of how to select the cutoff distance (‘height’) needed for cluster definition. Using the data shown in Table 2, a cutoff of (mean+2σ) for the 6 Å matrix would have given a value for the cutoff of 0.15 units. We have elected to slightly lower this number to 0.14 units in order to differentiate between the two clusters that their lineage separated at that

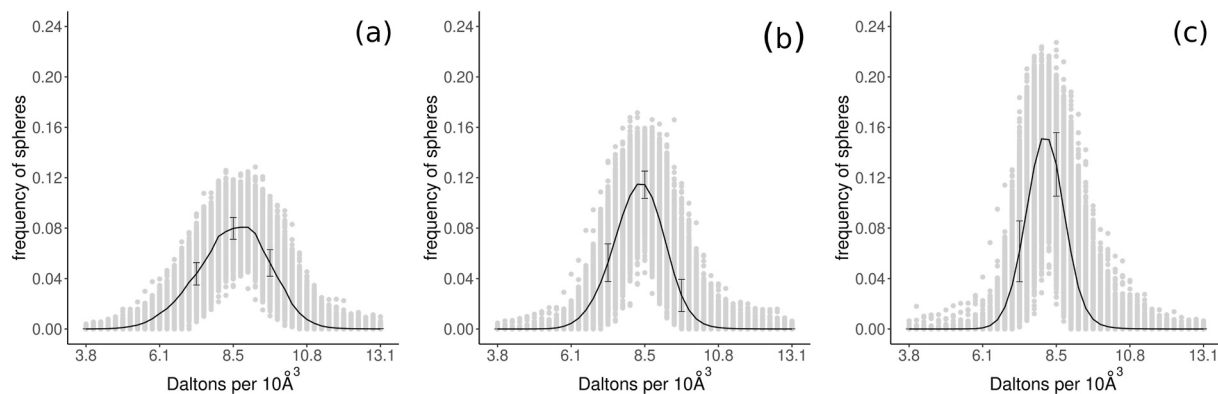


Fig. 5. Probability density distributions for each of the radii examined: (a) is for the 5 Å radius, (b) 6 Å, (c) 7 Å. In each diagram, the black line corresponds to the mean value of the respective bins. A representative sample of the corresponding standard deviations has also been added to the diagrams to help establish the amount of variance present in the individual distributions.

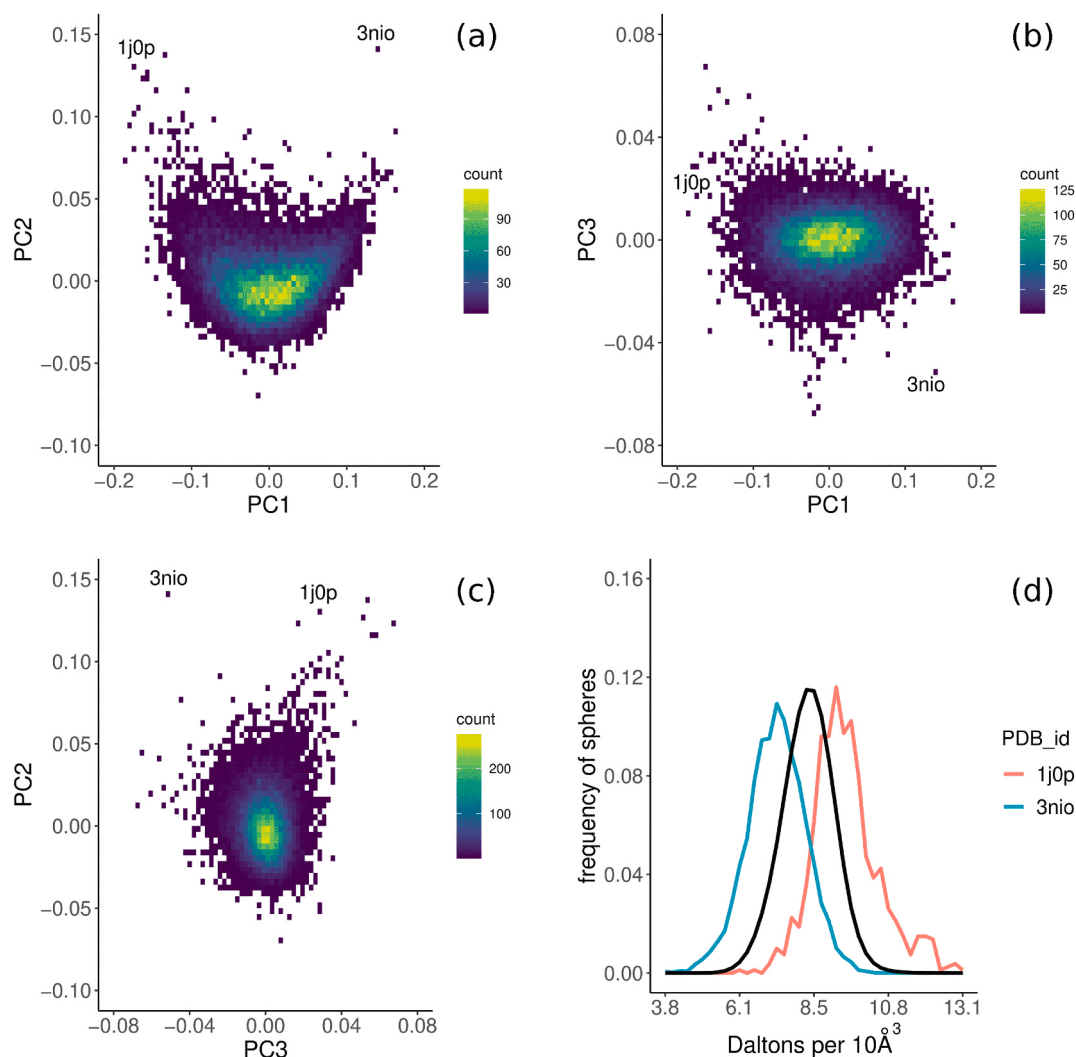


Fig. 6. Principal Component Analysis of the matrix obtained with a 6 Å radius. See text for details of this analysis.

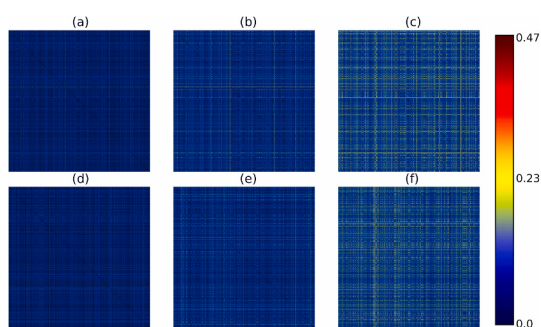


Fig. 7. Distance matrices visualized in the form of heatmaps. Panels (a), (b) and (c) correspond to the complete 5 Å, 6 Å, and 7 Å matrices. The lower three panels (d,e,f) are the same matrices but after removal of the 100 most distant structures to allow visualization of smaller distances in the matrices.

Table 2

Statistics for the three distance matrices shown in Fig. 7.

Matrix	Average distance	Standard deviation	Maximum distance
R = 5 Å	0.05	0.02	0.29
R = 6 Å	0.07	0.04	0.35
R = 7 Å	0.10	0.06	0.47

height (these are the clusters shown in purple and cyan in Fig. 8(b), the 2nd and 3rd from the left). This selection resulted to a total of 12 clusters. All further analyses discussed below will be referred to these clusters.

Examination of the dendrogram indicates that these 12 clusters vary significantly not only in size, but also on their distance separation (the ‘height’ of their last common ancestor). To be able to focus on the distant clusters (which are the most informative), we decided to quantify the distances between these 12 clusters by calculating a Z-score matrix (one Z-score for each possible pairwise cluster combination).

To perform this calculation, we started by fitting the atomic density distributions of the proteins that belong to each cluster to a Gaussian (see Fig. 9(b) for a pictorial explanation of the procedure). Once the means and standard deviations were available for each cluster, the Z-score matrix could be constructed, and used for another round of hierarchical clustering. The resulting dendrogram (this time showing relationships between *clusters* of proteins) is shown in Fig. 9(a).

The availability of this dendrogram, and after exclusion of clusters with less than 100 members, allowed us to focus on just five clusters of proteins that demonstrate significant deviations between them. Statistics for those five clusters are shown in Table 3. Fig. 9(b) and (c) show an explicit (raw data-based) comparison between the density distributions of proteins that belong to clusters 1, 4, 5 and 6. These graphs clearly indicate the presence of significant deviations at the raw-data level, with two clusters showing a shift to lower densities (clusters 1 & 4), and two

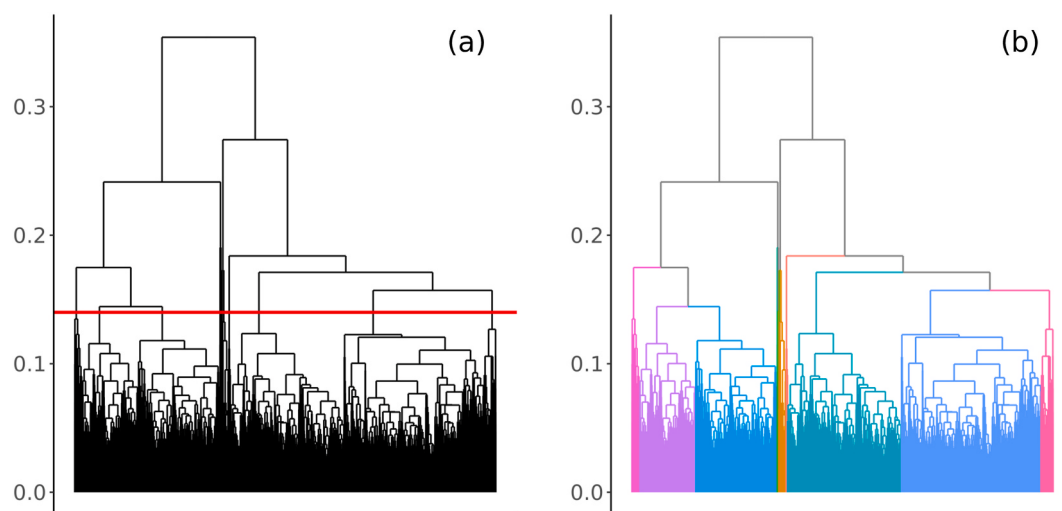


Fig. 8. Hierarchical clustering. Panel (a) shows the dendrogram produced from the 6 Å distance matrix using the “complete” linkage method. The red line corresponds to the height ($h = 0.14$) where the dendrogram was cut to produce clusters, see text for details. Panel (b) is the same dendrogram but with the individual clusters color-coded to aid identification. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

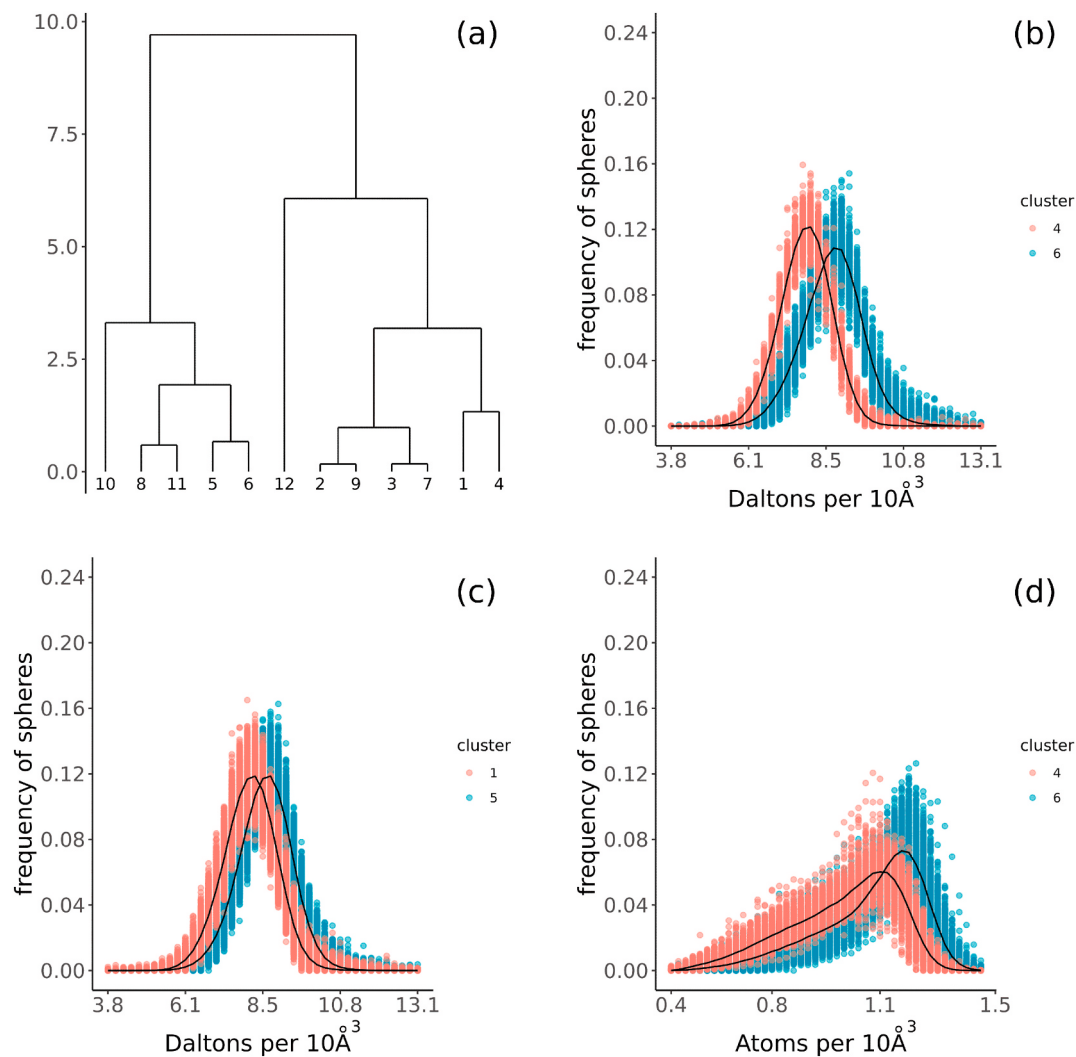


Fig. 9. Groups of structures deviate significantly from the average behavior. (a) Dendrogram from hierarchical clustering of the Z-scores matrix using the “complete” linkage method. (b) Comparison of the raw data of the clusters 4 and 6. (c) Comparison of the raw data of the clusters 1 and 5. (d) Comparisons of the raw data of the clusters 4 and 6, but in units of atoms per cubic Ångström, see text for details.

Table 3

Statistics for the five clusters of interest sorted on the basis of their median atomic density.

Cluster	Sample size	Median density (Dalton/Å ³)
4	344	0.77
1	5813	0.79
3	7045	0.81
5	2821	0.84
6	400	0.85

showing a shift to higher densities (clusters 5 & 6).

One important question at this point, is whether the differences of the distributions seen for example in Fig. 9(b) are due to the presence of ligands containing heavy atoms, or whether they reflect genuine differences in the packing of atoms. To tackle this issue, we recalculated for clusters 4 & 6 the protein density distributions, but instead of using units of “Daltons per cubic Ångström”, we calculated the distributions in units of “atoms per cubic Ångström”, thus ignoring the atomic weights of the atoms involved. The results from this calculation are shown in Fig. 9(d). Comparison between panels (b) and (d) in this figure indicates that a genuine difference of the atomic packing of the proteins that belong to these clusters appears to be present (and is not just systematic differences in the presence of ligands that lead to the observed differences). In the sections that follow we present some indications concerning the source of the observed systematic density deviations.

3.4. Structural implications of the atomic density distributions

In this section we examine the relationships between the density distributions we obtained above *versus* structural characteristics of the respective proteins such as size, secondary structure, temperature factors, and abundance of water molecules.

The first observation is that the median distribution has a wider spread (it is more scattered) and a right-skew for smaller proteins, see Fig. 10. Additionally, it appears that the number of residues is weakly anti-correlated with the median atomic density. This weak anti-correlation is quantified by the value of the spearman coefficient which is equal to -0.18 for the 6 Å data, and -0.27 for the 5 Å radius data. This finding is in good agreement with data from other studies which indicated that larger proteins tend to be packed more loosely than smaller ones [51,52]. Examination of the scatter plot in Fig. 10(b), indicates that as the number of amino acids increases, the variance of the median density values decreases. This finding may indicate the presence

of systematic structural/thermodynamic or evolutionary constraints that make a tight atomic packing of larger proteins uncommon. To put this observation in numbers, we divided the sample based on the quartiles of the residues. The median and standard deviation values of each quartile are shown in Table 4. There is a pattern in both metrics when going from samples representing small proteins to samples representing large structures, with both median and standard deviation decreasing in larger structures. Small proteins seem to have increased variability when it comes to packing, as indicated by their wider density limits. This finding may align well with the conclusion of a previous study that “Proteins are not optimized by evolution to eliminate packing voids” [52]. It is also in good agreement with an additional study, which showed a reduction in the range of density values in larger proteins. In their analysis small proteins exhibit a broad range of packing densities, varying from 0.67 to 0.87, while for large proteins densities range from 0.69 to 0.74 [53].

Furthermore, and in terms of molecular evolution, it is expected that archaic proteins would be small size and evolutionarily actively selected to be resistant to thermal denaturation and unfolding, properties that imply the presence of increased packing density and stability. This increased stability is also known to be associated with an increased resistance to proteolysis [54,55]. With the emergence of larger proteins within the progressively more complex and compartmentalized cellular environment, the evolutionary pressure to maintain dense and stable packing was reduced, and flexible and less well-packed proteins emerged [56].

Among the rest of the factors characterizing crystallographically determined protein structures, the abundance of ordered water molecules, and the mean value of the atomic temperature factors (“B-factors”) were found to be strongly correlated with atomic density. The B-factor is a measure of an atom’s displacement about their mean position, and it provides a measure of the flexibility and stability of the structure [57,58]. Ordered water molecules, on the other hand, become

Table 4

Median and standard deviation values in groups of structures with different size.

Size group	Median (Dalton/Å ³)	Standard Deviation (Dalton/Å ³)
51 < aminoacids <182	0.822	0.020
182 < aminoacids <281	0.818	0.018
281 < aminoacids <406	0.817	0.017
406 < aminoacids <664	0.814	0.017
664 < aminoacids <4682	0.812	0.016

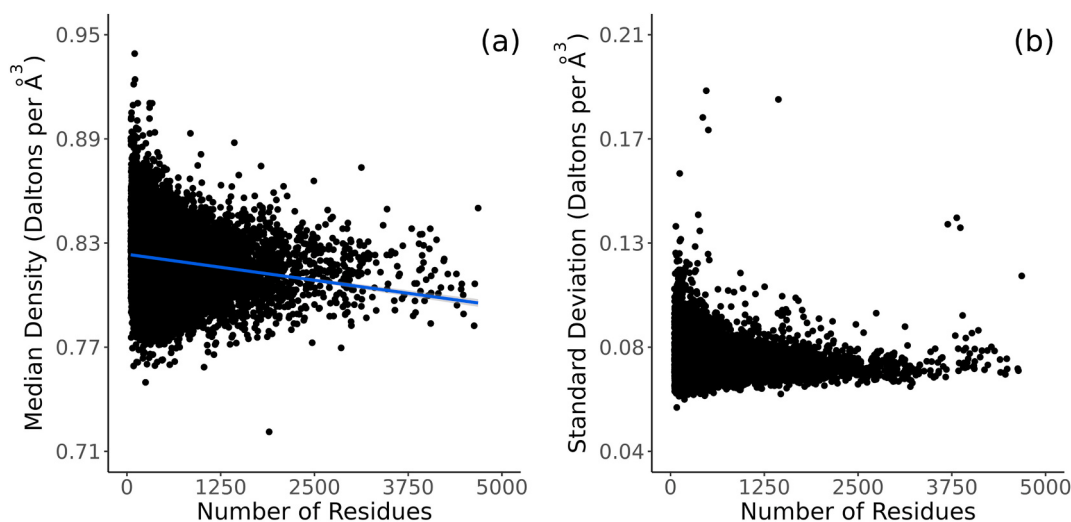


Fig. 10. Protein size *versus* atomic density. (a) Scatter plot with regression line for number of amino acid residues vs. median density. The Spearman correlation is -0.18 . (b) Scatter plot of the corresponding standard deviations. Notice the change of scale in the two graphs.

immobilized on the surface of stable structures. In these stable structures, the reduced mobility of side chains promotes favorable interactions and bond formation with water molecules. Thus, crystallographic water abundance and B-factors can be used as indicators of stability, to examine the relationship between stability and packing. The relationship between packing density and these two factors is shown in Fig. 11(c) and (d) respectively. Water abundance shows a strong positive correlation with median density values (Pearson +0.76, Spearman +0.78), indicating that the more densely packed a structure is, the more crystallographic waters are likely to be resolved. Conversely, the median of B-factors shows a pronounced anti-correlation with the median density (Pearson -0.56 , Spearman -0.57), which indicates that the flexibility and displacements of atoms in densely packed structures are reduced, compared to loosely packed structures. This shows that packing density significantly affects structures, by providing an increased stabilization (lower B-factors), and thus an increased probability of crystallographically observing ordered water molecules.

We have also examined the data for the presence of putative correlations between secondary structure composition and the atomic density profiles. The program STRIDE [59] was used to assign secondary structure state to each amino acid of the protein clusters identified in section §3.3 and shown in Table 3, and to also calculate the per residue

Solvent Accessible Surface Area (SASA). Although no outstanding pattern was found connecting the SASA with our density analysis (see Fig. S2 in Supplementary Material), this was not the case for the secondary structure: When the percentage of all secondary structure assignments was calculated across all structures of each cluster, significant correlations were detected for β -strands and α -helices, but not for other elements, such as turns and loops. The distribution of the elements' percentages is shown in Fig. 11(a) and (b). Examination of these figures indicates that as we move from loose (clusters 4 & 1) to dense (clusters 5 & 6) clusters, the percentage of β -strands is decreasing, while the percentage of α -helices is increasing. This is an indication that certain types of structures such as, for example, coiled-coils may appear only in the dense clusters. To investigate this, we collected the structure-related keywords from the corresponding HEADER records of the PDB files. The keyword "COILED" appeared in 6.5 % of the proteins of the dense cluster, but not at all in the loosely packed cluster. We have also visually inspected the coiled-coil structures that matched the keyword with molecular graphics to validate the existence of the motif. Coiled-coils are thus found to be structures with increased packing density. The coiled-coil, a slightly twisted arrangement of two or more α -helices frequently found in fibrous proteins, was proposed by Crick in 1953 [60]. These structures have a 'knobs-into-holes' type of packing in which a hydrophobic core residue from one helix is packed in a "hole" formed

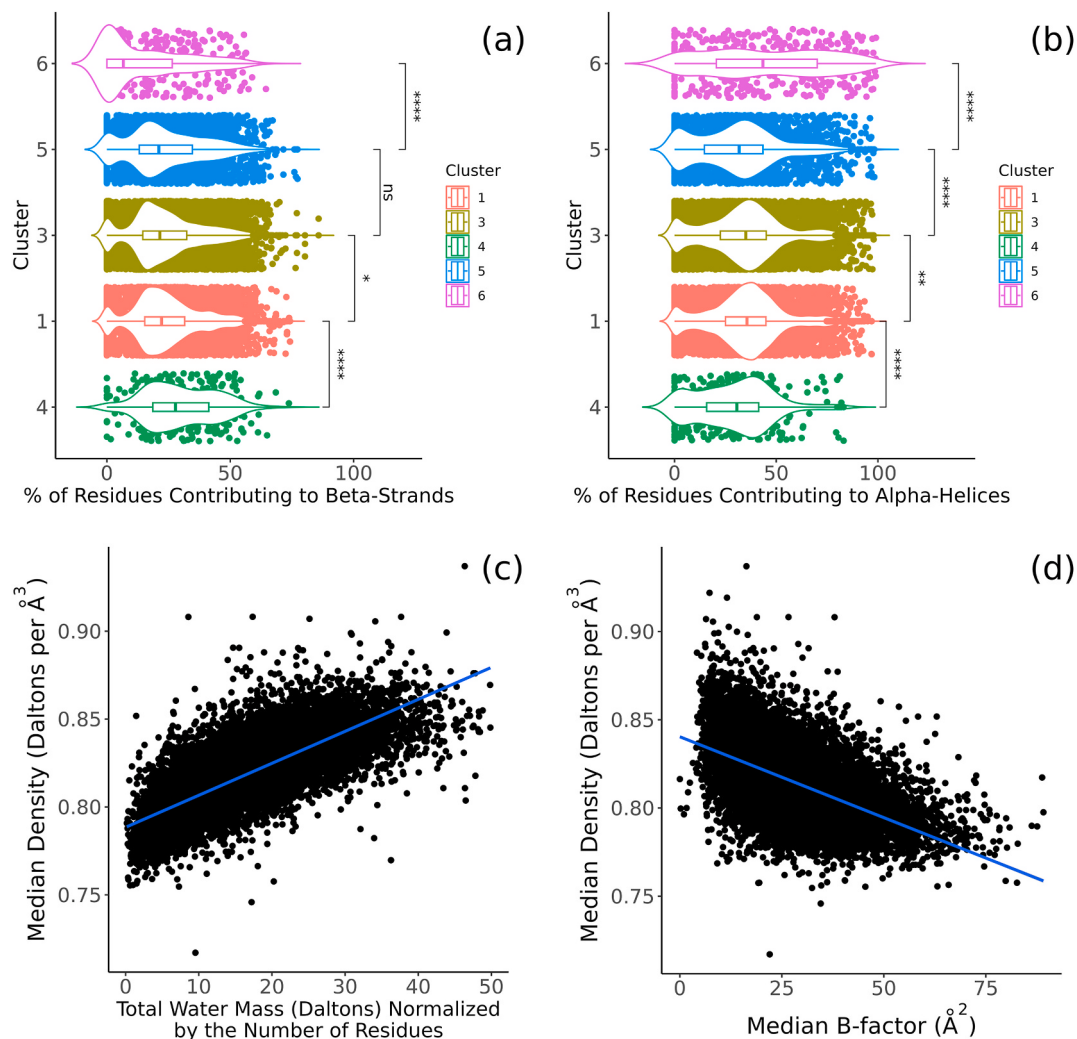


Fig. 11. Relationship between packing density and secondary structure elements, water molecules abundance and atomic temperature factors. (a) Violin plot of the percentage of residues contributing to β -strands across clusters. The order of clusters is sorted on decreasing density (dense clusters on the top, less dense towards the bottom), (b) Violin plot of the percentage of residues contributing to α -helices across clusters. (c) Correlation between median density and abundance of water molecules. (d) Correlation between median density and the atomic temperature factors (B-factors). See text for details.

by four residues of the other helix, resulting in a tight side-by-side arrangement of the hydrophobic core residues [61]. In addition, the hydrophobic core present in this motif offers stabilization in these structures [62]. Since some coiled-coils appear in fibrous proteins (e.g. keratin, myosin, kinesin) with structural and motor roles, mechanical stability is needed, in order to function properly. This stability, especially when accompanied by large internal cavities (to store substances) in polymeric coiled-coil domains, makes these motifs ideal for therapeutic applications in the form of efficient drug delivery systems. Lastly, the simplicity and structural robustness of this motif makes it ideal for many other key biological processes such as transcription and communication [63].

We extended this keyword-based procedure beyond coiled-coils, and we calculated percentages of all possible PDB-derived keywords between the five clusters shown in Table 3. We then isolated those with a differential abundance in the various clusters. The one keyword that stands-out in this procedure is the keyword “CYTOCHROME” which appeared in 5.25 % of the proteins of the dense cluster, but not at all at the less densely packed clusters. Cytochromes form a diverse group of proteins with only a few features in common. They all contain protoheme IX or one of its derivatives and function in electron transport [64]. Packing density is a crucial parameter for effective electron transport [65,66]. It has been shown that even slight reductions in the distances of through-space jumps in electron transport pathways, or enhancements in atomic packing density, can significantly accelerate the rate of transfer [67]. Our data are in good agreement with those observations as indicated by the over-representation of cytochromes in the dense clusters.

3.5. Functional implications of the atomic density distributions

In this last section of our analysis we attempt to identify and characterize possible relationships between the atomic density distributions and the functional properties of the respective proteins. We aimed to explore whether certain families of proteins show a detectable preference for a loose or dense packing in order to function properly. For that purpose, we collected the classification terms from the HEADER area of each PDB file and calculated the abundances for each of the clusters we described in section §3.3. Results are shown in Fig. 12.

It can be seen that the distribution of protein families across clusters is not uniform. Instead, certain groups appear in higher percentages in

specific density clusters. Since cluster sizes and total family abundances vary, two different normalization approaches were applied to get comparable percentages. For Fig. 12(a), the abundance of each family in a cluster was divided by the total abundance of the family across the sample of the 5 clusters (so percentages are comparable only within each cluster). We can see that Hydrolases have a higher percentage in cluster 5 (dense cluster) compared to Transferases, whereas in cluster 1 (loose cluster) the percentage of Transferases is higher than the one of Hydrolases. For Fig. 12(b), the abundance of each family in a cluster was divided by the total abundance of the 4 families within this cluster and thus, family percentages are directly comparable across clusters. We can see that ligases and transferases prefer loose clusters, while hydrolases the denser ones. For oxidoreductases there is a slight preference for the average cluster (cluster 3). Furthermore, the electron transport family which was seen increased in the dense clusters does not appear in the plot. This family includes cytochromes, that were analyzed in a previous section. Other known protein families with no significant changes across clusters do not appear in the plot.

The observed differences across protein families indicate that through protein evolution, a variation in packing density occurred, and that this variation is linked with the functional properties of the respective protein families. Multiple studies have noted that the emergence of new enzymatic specificities is often linked to a decrease in the protein's thermodynamic stability, indicating the presence of a trade-off between gaining new enzymatic functions and maintaining stability [68–71]. A fine balance between stability and activity is essential for enzymes to function optimally. However, the extent of this trade-off across different protein regions and its dependence on environmental conditions remains unclear [72]. It is also important to mention that enzymes catalyzing reactions with relatively simple mechanisms (for example, hydrolysis) were likely to be one of the earliest to evolve. This, alongside the fact that earlier proteins were probably more tightly packed (see previous section), is connected with the question of whether hydrolases were among the first protein classes to be established in terms of molecular evolution. In addition, transferases, which transfer non-water functional groups, are able to prevent the mechanistically similar process of hydrolysis in the cellular environment, where water is more abundant than any other substrate [73,74]. Some transferases, which are seen to prefer a looser packing, could share a common ancestor with hydrolases especially when considering the function/stability trade-off and their similar reaction mechanism. This discussion,

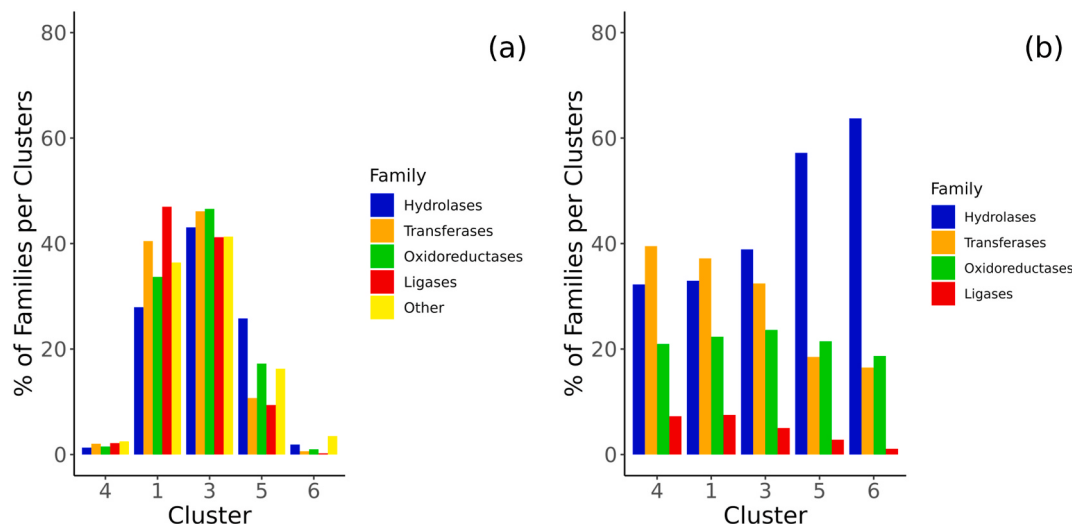


Fig. 12. Distribution of protein families across the various clusters. In both of these diagrams, low density clusters are to the left, high density clusters to the right. Panel (a) shows how the members of a given protein family are distributed across all five clusters. The scaling is such that summing all frequencies for any given family adds-up to 100 %. This diagram has not been corrected for cluster size, which explains the marked differences observed. Panel (b) shows the relative distribution of protein families on a per-cluster basis (and not across all clusters as in panel (a)).

however, remains largely speculative and should be viewed as an open-ended question, requiring further studies on protein evolution to gain a better understanding of the underlying molecular evolution mechanisms.

4. Discussion

In this work, we explored the role of packing in proteins, by analyzing a large sample of crystallographically resolved, *in silico* hydrogenated and hydrated protein structures. Our results agree with and further validate previous studies on the relationships between packing and stability on one hand, and packing and protein size on the other. Small proteins are seen to have a wider variability of density values, but in general are more compact than larger ones. The reduced density of larger proteins, may result from biological constraints, as discussed in §3.4. The correlations between median density and the abundance of crystallographic water molecules and B-factors, shown in Fig. 11, supports existing knowledge concerning the effect of packing on stability. It also indicated that our algorithm for calculating atomic density distributions provides quality metrics that can be used to classify the packing level of a protein structure.

Furthermore, search for structural patterns across clusters of interest, revealed special folding patterns such as coiled-coils and cytochromes, with a high percentage in the most dense cluster. Both cases are characteristic examples of how a protein's atomic density may be connected with the functional requirements. Mechanical stability (coiled-coils) and effective electron transport (cytochromes) actively promote a denser packing which is probably needed for them to function properly.

Regarding the functional implications of atomic density, we observe a preference for some protein families on specific density clusters. This further supports the packing-function relationship we discussed for coiled-coils and cytochromes. Except for the stability aspect, which plays a crucial role in determining a protein's function, the close proximity of atoms within a structure may also be an important factor for some enzymatic reactions to happen. This is because density of packing may affect the volume and flexibility of active sites and determine which substrates have the appropriate size to be accommodated and stabilized inside them. Following this point of view may provide further insight into how transferases minimize their hydrolytic activity given that water is the most abundant substrate in the cell environment: Transferases with their preference being for less well-packed structures, may be unable to interact with and stabilize water molecules. Conversely, hydrolases could immobilize water molecules more easily, as they are more stable themselves and the distances between side chains of their structures tend to be smaller.

Although no similar work of functionally clustering proteins based on their atomic density profiles exists, there are several studies that aim to group proteins from other features and reveal functional insights. For instance, sequence-structure-function relationships can be visualized through clustering of protein networks, providing a unified view of how proteins are organized and classified across various databases. Using methods such as minimum span clustering (MSC), thousands of enzymes have been systematically grouped to investigate discrepancies between classifications based on sequence, structure, and function [75]. More recently, structural alignment-based approaches such as Foldseek cluster have enabled the clustering of hundreds of millions of protein structures, including all entries from the AlphaFold database. This large-scale effort identified over two million non-singleton clusters, a substantial fraction of which lack existing annotations, suggesting the presence of previously undescribed or species-specific folds [76]. Beyond global clustering, recent advances in deep learning have facilitated finer-grained functional grouping. Embedding-based representations derived from protein language models such as ProtBERT and PB-Tucker allow the subdivision of protein superfamilies into more functionally consistent subfamilies [77]. At an even more granular level, microenvironment-based clustering approaches have been proposed to

rapidly group millions of local residue environments across thousands of protein chains, revealing recurrent spatial motifs associated with shared structural or functional roles [78].

To conclude, we have exhaustively analyzed and compared density profiles for an extended set of more than 21,000 proteins. Our analysis and subsequent all-to-all comparison of those distributions allowed us to quantitatively characterize structural and functional patterns present in these distributions, and to validate and further elaborate previous results in the field. Based on our analysis, we can corroborate the general view on the subject of protein density distributions: although systematic patterns of differences in atomic density are indeed present, these are *not* on a fixed one-to-one correspondence with structural and functional characteristics of the respective proteins. In a sense, the variability of atomic density that we observe in present-day proteins may be viewed as a remnant of the molecular evolution processes that led to those proteins, and not the direct result of a presently active selection process.

CRediT authorship contribution statement

Sotirios Touloupoulos: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Nicholas M. Glykos:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization.

Program and data availability

All data and programs used for the analyses reported in this communication are available via https://github.com/SotirisTouloupoulos/atomic_density.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jmglm.2025.109228>.

Data availability

Data and programs available from public repositories.

References

- [1] F.M. Richards, The interpretation of protein structures: total volume, group volume distributions and packing density, *J. Mol. Biol.* 82 (1974) 1–14.
- [2] L.M. Gregoret, F.E. Cohen, Novel method for the rapid evaluation of packing in protein structures, *J. Mol. Biol.* 211 (1990) 959–974.
- [3] Random Packings and the Structure of Simple Liquids. I. The Geometry of Random Close Packing, *Proc. R. Soc. A* 319 (1970) 479–493.
- [4] J. Pontius, J. Richelle, S.J. Wodak, Deviations from standard atomic volumes as a quality measure for protein crystal structures, *J. Mol. Biol.* 264 (1996) 121–136.
- [5] M. Gerstein, E.L.L. Sonnhammer, C. Chothia, Volume changes in protein evolution, *J. Mol. Biol.* 236 (1994) 1067–1078.
- [6] O.B. Ptitsyn, M.V. Volkenstein, Protein structures and neutral theory of evolution, *J. Biomol. Struct. Dyn.* 4 (1986) 137–156.
- [7] A.M. Lesk, C. Chothia, How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins, *J. Mol. Biol.* 136 (1980) 225–270.
- [8] N. Pattabiraman, K.B. Ward, P.J. Fleming, Occluded molecular surface: analysis of protein packing, *J. Mol. Recogn.* 8 (1995) 334–344.
- [9] S. Tanaka, H.A. Scheraga, Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins, *Macromolecules* 9 (1976) 945–950.
- [10] P.K. Warme, R.S. Morgan, A survey of atomic interactions in 21 proteins, *J. Mol. Biol.* 118 (1978) 273–287.
- [11] S.V.L. Narayana, P. Argos, Residue contacts in protein structures and implications for protein folding, *Int. J. Pept. Protein Res.* 24 (1984) 25–39.

- [12] S. Miyazawa, R.L. Jernigan, Residue – residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading, *J. Mol. Biol.* 256 (1996) 623–644.
- [13] R.A. Abagyan, M.M. Totrov, Contact area difference (CAD): a robust measure to evaluate accuracy of protein models, *J. Mol. Biol.* 268 (1997) 678–685.
- [14] J.M. Word, S.C. Lovell, T.H. LaBean, H.C. Taylor, M.E. Zalis, B.K. Presley, J. S. Richardson, D.C. Richardson, Visualizing and quantifying molecular Goodness-of-Fit: small-probe contact dots with explicit hydrogen atoms, *J. Mol. Biol.* 285 (1999) 1711–1733.
- [15] J.C. Gaines, A.H. Clark, L. Regan, C.S. O'Hern, Packing in protein cores, *J. Phys. Condens. Matter* 29 (29) (2017) 293001.
- [16] V. Erokhin, P. Facci, A. Kononenko, G. Radicchi, C. Nicolini, On the role of molecular close packing on the protein thermal stability, *Thin Solid Films* 284–285 (1996) 805–808.
- [17] J.T. Kellis Jr., K. Nyberg, D. S  il, A.R. Fersht, Contribution of hydrophobic interactions to protein stability, *Nature* 333 (1988) 784–786.
- [18] M. Mravic, J.L. Thomaston, M. Tucker, P.E. Solomon, L. Liu, W.F. DeGrado, Packing of apolar side chains enables accurate design of highly stable membrane proteins, *Science* 363 (2019) 1418–1423.
- [19] J. Chen, W.E. Stites, Packing is a key selection factor in the evolution of protein hydrophobic cores, *Biochemistry* 40 (2001) 15280–15289.
- [20] M. Vlassi, G. Cesareni, M. Kokkinidis, A correlation between the loss of hydrophobic core packing interactions and protein stability, *J. Mol. Biol.* 285 (1999) 817–827.
- [21] G. Vogt, P. Argos, Protein thermal stability: Hydrogen bonds or internal packing? *Folding Des.* 2 (1997) S40–S46.
- [22] W.S. Sandberg, T.C. Terwilliger, Influence of interior packing and hydrophobicity on the stability of a protein, *Science* 245 (1989) 54–57.
- [23] W.A. Lim, R.T. Sauer, The role of internal packing interactions in determining the structure and stability of a protein, *J. Mol. Biol.* 219 (1991) 359–376.
- [24] D.E. Makarov, A. Keller, K.W. Plaxco, H. Metiu, How the folding rate constant of simple, single-domain proteins depends on the number of native contacts, *Proc. Natl. Acad. Sci.* 99 (2002) 3535–3539.
- [25] K. Oda, M. Kinoshita, Physicochemical origin of high correlation between thermal stability of a protein and its packing efficiency: a theoretical study for staphylococcal nuclease mutants, *Biophys. Physicobiol.* 12 (2015) 1–12.
- [26] A.P. Pandurangan, B. Ochoa-Montano, D.B. Ascher, T.L.S.D.M. Blundell, A server for predicting effects of mutations on protein stability, *Nucleic Acids Res.* 45 (2017) W229–W235.
- [27] D. Schell, J. Tsai, J.M. Scholtz, C.N. Pace, Hydrogen bonding increases packing density in the protein interior, *Proteins: Struct., Funct., Bioinf.* 63 (2006) 278–282.
- [28] Y. Sugita, A. Kitao, N. Go, Computational analysis of thermal stability: effect of ile→val mutations in human lysozyme, *Folding Des.* 3 (1998) 173–181.
- [29] M. Prevost, S.J. Wodak, B. Tidor, M. Karplus, Contribution of the Hydrophobic Effect to Protein Stability: Analysis Based on Simulations of the Ile-96→Ala Mutation in Barnase, vol. 88, *Proceedings of the National Academy of Sciences*, 1991, pp. 10880–10884.
- [30] H. Kono, M. Saito, A. Sarai, Stability analysis for the cavity-filling mutations of the Myb DNA-binding domain utilizing free-energy calculations, *Proteins: Struct., Funct., Bioinf.* 38 (2000) 197–209.
- [31] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, The packing density in proteins: standard radii and volumes, *J. Mol. Biol.* 290 (1999) 253–266.
- [32] S. Sonavane, P. Chakrabarti, Cavities and atomic packing in protein structures and interfaces, *PLoS Comput. Biol.* 4 (2008) e1000188.
- [33] B. Halle, Flexibility and packing in proteins, *Proc. Natl. Acad. Sci.* 99 (2002) 1274–1279.
- [34] M. Gerstein, J. Tsai, M. Levitt, The volume of atoms on the protein surface: calculated from simulation, using voronoi polyhedra, *J. Mol. Biol.* 249 (1995) 955–966.
- [35] M. Gerstein, C. Chothia, Packing at the protein-water interface, in: *Proceedings of the National Academy of Sciences*, vol. 93, 1996, pp. 10167–10172.
- [36] W.A. Lim, R.T. Sauer, Alternative packing arrangements in the hydrophobic core of λ repressor, *Nature* 339 (1989) 31–36.
- [37] P.-S. Huang, S.E. Boyken, D. Baker, The Coming of Age of de Novo Protein Design, *Nature* 537 (2016) 320–327.
- [38] K. Gekko, Volume and compressibility of proteins, *Subcell. Biochem.* 72 (2015) 75–108.
- [39] T.V. Chalikian, V.S. Gindikin, K.J. Breslauer, Volumetric characterizations of the native, molten globule and unfolded states of cytochrome c at acidic pH, *J. Mol. Biol.* 250 (1995) 291–306.
- [40] M.F. Perutz, J.C. Kendrew, H.C. Watson, Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence, *J. Mol. Biol.* 13 (1965) 669–678.
- [41] M. Kimura, T. Ohta, On some principles governing molecular evolution, in: *Proceedings of the National Academy of Sciences*, vol. 71, 1974, pp. 2848–2852.
- [42] J. Echave, S.J. Spielman, C.O. Wilke, Causes of evolutionary rate variation among protein sites, *Nat. Rev. Genet.* 17 (2016) 109–121.
- [43] C. Shih, C. Chang, Y. Lin, W. Lo, J. Hwang, Evolutionary information hidden in a single protein structure, *Proteins: Struct., Funct., Bioinf.* 80 (2012) 1647–1657.
- [44] C. Lin, S. Huang, Y. Lai, S. Yen, C. Shih, C. Lu, C. Huang, J. Hwang, Deriving protein dynamical properties from weighted protein contact number, *Proteins: Struct., Funct., Bioinf.* 72 (2008) 929–935.
- [45] M. Yu Lobanov, N.S. Bogatyreva, O.V. Galzitskaya, Radius of gyration as an indicator of protein structure compactness, *Mol. Biol.* 42 (2008) 623–628.
- [46] H. Fischer, I. Polikarpov, A.F. Craievich, Average protein density is a molecular-weight-dependent function, *Protein Sci.* 13 (2004) 2825–2828.
- [47] P.J. Fleming, F.M. Richards, Protein packing: dependence on protein size, secondary structure and amino acid composition, *J. Mol. Biol.* 299 (2000) 487–498.
- [48] G. Wang, R.L. Dunbrack Jr., PISCES: a protein sequence culling server, *Bioinformatics* 19 (2003) 1589–1591.
- [49] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open babel: an open chemical toolbox, *J. Cheminf.* 3 (2011) 33.
- [50] J.P.G.L.M. Rodrigues, J.M.C. Teixeira, M. Trellet, A.M.J.J. Bonvin, Pdb-Tools: a Swiss Army Knife for Molecular Structures, vol. 7, *F1000Research*, 2018, p. 1961.
- [51] J. Liang, K.A. Dill, Are proteins well-packed? *Biophys. J.* 81 (2001) 751–766.
- [52] J. Zhang, R. Chen, C. Tang, J. Liang, Origin of scaling behavior of protein packing density: a sequential monte carlo study of compact long chain polymers, *J. Chem. Phys.* 118 (2003) 6102–6109.
- [53] N. Rawat, P. Shape Biswas, Flexibility and packing of proteins and nucleic acids in complexes, *Phys. Chem. Chem. Phys.* 13 (2011) 9632–9643.
- [54] S. Ahmad, V. Kumar, K.B. Ramanand, N.M. Rao, Probing protein stability and proteolytic resistance by loop scanning: a comprehensive mutational analysis, *Protein Sci.* 21 (2012) 433–446.
- [55] R.M. Daniel, D.A. Cowan, H.W. Morgan, M.P. Curran, A correlation between protein thermostability and resistance to proteolysis, *Biochem. J.* 207 (1982) 641–644.
- [56] Q.-Y. Tang, W. Ren, J. Wang, K. Kaneko, The statistical trends of protein evolution: a lesson from AlphaFold database, *Mol. Biol. Evol.* 39 (2022).
- [57] H. Frauenfelder, S.G. Sligar, P.G. Wolynes, The energy landscapes and motions of proteins, *Science* 254 (1991) 1598–1603.
- [58] Z. Sun, Q. Liu, G. Qu, Y. Feng, M.T. Reetz, Utility of B-Factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability, *Chem. Rev.* 119 (2019) 1626–1665.
- [59] M. Heinig, D. Frishman, STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins, *Nucleic Acids Res.* 32 (2004) W500–W502.
- [60] F.H.C. Crick, The packing of α -Helices: simple coiled-coils, *Acta Crystallogr.* 6 (1953) 689–697.
- [61] A.N. Lupas, M. Gruber, The structure of α -Helical coiled coils, *Adv. Protein Chem.* 70 (2005) 37–38.
- [62] Yu, Y.B. Coiled-Coils, Stability, specificity, and drug delivery potential, *Adv. Drug Deliv. Rev.* 54 (2002) 1113–1129.
- [63] A.A. McFarlane, G.L. Orriss, J. Stetefeld, The use of coiled-coil proteins in drug delivery systems, *Eur. J. Pharmacol.* 625 (2009) 101–107.
- [64] F. Scott Mathews, The structure, function and evolution of cytochromes, *Prog. Biophys. Mol. Biol.* 45 (1985) 1–56.
- [65] C. Nowak, D. Schach, J. Gebert, M. Grosserueschkamp, R.B. Gennis, S. Ferguson-Miller, W. Knoll, D. Walz, R.L.C. Naumann, Oriented immobilization and electron transfer to the cytochrome c oxidase, *J. Solid State Electrochem.* 15 (2011) 105–114.
- [66] H.-B. Shao, H.-Z. Yu, J.-W. Zhao, H.-L. Zhang, Z.-F. Liu, The effect of packing density on the electron-transfer kinetics of an azobenzene thiol monolayer on gold, *Chem. Lett.* 26 (1997) 749–750.
- [67] H.R. Williams, B.A. Dow, V.L. Davidson, Mechanisms for control of biological electron transfer reactions, *Bioorg. Chem.* 57 (2014) 213–221.
- [68] N. Tokuriki, F. Stricher, L. Serrano, D.S. Tawfik, How protein stability and new functions trade off, *PLoS Comput. Biol.* 4 (2008) e1000002.
- [69] R.A. Studer, P.-A. Christin, M.A. Williams, C.A. Orengo, Stability-activity tradeoffs constrain the adaptive evolution of RubisCO, in: *Proceedings of the National Academy of Sciences*, 111, 2014, pp. 2223–2228.
- [70] M.A. DePristo, D.M. Weinreich, D.L. Hartl, Missense meanderings in sequence space: a biophysical view of protein evolution, *Nat. Rev. Genet.* 6 (2005) 678–687.
- [71] N. Tokuriki, D.S. Tawfik, Stability effects of mutations and protein evolvability, *Curr. Opin. Struct. Biol.* 19 (2009) 596–604.
- [72] Q. Hou, M. Rooman, F. Pucci, Enzyme stability-activity trade-off: new insights from protein stability weaknesses and evolutionary conservation, *J. Chem. Theor. Comput.* 19 (2023) 3664–3671.
- [73] D.E. Koshland, Enzyme flexibility and enzyme action, *J. Cell. Comp. Physiol.* 54 (1959) 245–258.
- [74] S.H. Light, L.A. Cahoon, K.V. Mahasenan, M. Lee, B. Boggess, A.S. Halavaty, S. Mobashery, N.E. Freitag, W.F. Anderson, Transferase versus hydrolase: the role of conformational flexibility in reaction specificity, *Structure* 25 (2017) 295–304.
- [75] T.-L. Mai, G.-M. Hu, C.-M. Chen, Visualizing and clustering protein similarity networks: sequences, structures, and functions, *J. Proteome Res.* 15 (2016) 2123–2131.
- [76] I. Barrio-Hernandez, J. Yeo, J. J  nes, M. Mirdita, C.L.M. Gilchrist, T. Wein, M. Varadi, S. Velankar, P. Beltrao, M. Steinegger, Clustering predicted structures at the scale of the known protein universe, *Nature* 622 (2023) 637–645.
- [77] M. Littmann, N. Bordin, M. Heinzinger, K. Sch  tze, C. Dallago, C. Orengo, B. Rost, Clustering FunFams using sequence embeddings improves EC purity, *Bioinformatics* 37 (2021) 3449–3455.
- [78] S. Yoon, J.C. Ebert, E.-Y. Chung, G. De Micheli, R.B. Altman, Clustering protein environments for function prediction: finding PROSITE motifs in 3D, *BMC Bioinf.* 8 (S4) (2007).