SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

*Molecular & Cellular Biomechanics* 2024, 21(4), 202.
https://doi.org/10.62617/mcb202

Article

# Identifying sequential differences between protein structural classes using network and statistical approaches

**Xiaogeng Wan[1,*], Xinying Tan[2]**

[1] Department of Mathematics, College of Mathematics and Physics, Beijing University of Chemical Technology, Beijing 100029, China
[2] The Fourth Medical Center, PLA General Hospital, Beijing 100037, China
*** Corresponding author:** Xiaogeng Wan, wxgbj88@sina.com

**CITATION**

Wan X, Tan X. Identifying sequential differences between protein structural classes using network and statistical approaches. Molecular & Cellular Biomechanics. 2024; 21(4): 202. https://doi.org/10.62617/mcb202

**ARTICLE INFO**

Received: 19 July 2024
Accepted: 24 September 2024
Available online: 16 December 2024



**Abstract:** Protein sequence information is believed to embed the hint of their structures. To uncover the nature between protein sequence and their structures, this study motivates to inspect the dynamic interactions between various protein sequence features, and identify the sequential differences between the different protein structures. Protein sequence data from all structural classes in CATH and SCOP, and the structural disordered proteins from DisProt, as well as the structural motifs in PROSITE, are analyzed in this study. Betweenness and closeness centrality measures are employed to capture the topology of the networks constructed from amino acid feature interactions, while statistical tests are further implemented to compare the feature series distributions. Key findings suggest that in all structural classes, the features for Ala and α-helix and bend preference property, Ala and side-chain size, Ala and Gly, as well as Met and Leu attain significant interactions between each other, and the feature for Leu, Val, and Asn are acted as the critical sources of feature interactions, whereas Cys, His, Trp, and Met exhibit weak intra-type interactions with other features. These implicate that these feature interactions may have little impact in coding the structural differences. For the $\alpha$ structures, Glu, Pro and side-chain size, hydrophobicity properties exhibit high importance in feature interactions, whereas Gly, Thr and physical properties such as $\alpha$-helix and bend preference, extended structural preference, pK-C value and surrounding hydrophobicity for $\beta$ structures, show special high importance in $\beta$ structures. Both $\alpha$ and $\beta$ types of structures show Ser as the common sources of feature interactions, while the mixed $\alpha$ and $\beta$ structures not only show common characters with the $\alpha$ and $\beta$ types of structures, but also preferred interactions between Met, Lys and double-bend preference property, and between the sequence arrangements of Cys, His, Met, Tyr and amino acid composition features. The intrinsically disordered proteins (IDPs) present high frequency for the repetition patterns of certain amino acids, while the different structural motifs also show special characters. More sequential differences between the structures can also be identified from *K*-mers statistics and feature series distributions. The new discoveries reveal the nature of amino acid feature interaction mechanics, and show great importance of these interactions in coding the different types of protein structures. The results can not only contribute to future molecular design for protein-based vaccine or drug, but also enlighten the development for new protein structural classifiers.

**Keywords:** protein sequence feature; structural types; relationship measures; network; statistics

## 1. Introduction

Classifying or predicting protein 3D structures using amino acid sequence homology is hot research topics in bioinformatics, where protein sequence information shows great influences on their structures [1–5]. As technology develops, many artificial intelligence techniques have been proposed for protein structural

classifications. Alpha Fold [6] and its improved version Alpha Fold2 [7] developed by DeepMind attain the overall best accuracy in protein structural predictions [7]. Wu et al. have developed a deep learning-based protein structural model refinement method ATOM Refine [8]. Hong et al. have proposed a novel protein 3D structural modeling method A-Prot by implementing the protein language model MSA Transformer [9]. Pearce et al. have innovated an open-source protein structural prediction algorithm Deep Fold by implementing multi-task deep residual neural-networks [10]. Kruglov et al. have extended the evolutionary algorithm USPEX into a novel protein structural prediction method using global optimization [11]. Stapor et al. have invented a multi contact-based folding method Multi C Fold [12]. Stapor et al. also innovated a new lightweight deep network ProteinUnet2 with U-Net convolutional architecture [13]. Kim et al. have proposed a new accurate prediction algorithm AttSec using transformer architecture [14]. Zhang et al. [15] have proposed a new protein structural optimization algorithm based on deep learning technology. Yasin et al. [16] have proposed a novel deep learning model based on graph convolutional network. Zhang et al. Al [17] have invented a new atomic-level protein structural model by means of Cryo-EM density maps. Liu et al. [18] have invented a new protein fold identifiers using deep learning and support vector machines. Wan et al. [19] have analyzed the symmetry of intra-type feature relations, and extract monotonic centrality characteristics for protein structures. Other protein structural studies may use spatial classifiers such as Minimum-Square-Error hyper-planes [20,21], convex hulls [22] or other clustering algorithms to classify protein features.

On of the bottleneck in protein structural classification or prediction tasks is the feature extraction from protein sequence. Typical protein feature methods map the compositions, arrangements, physical properties of amino acids as well as alignment scores into real vectors or matrices [3,20,23]. These methods are for instances, the natural vector (NV) [3], averaged property factors (APF) [20], PSSM [23], PseAAC [24], Pse-in-One [25]. Since Kmer methods show good advantage in faster construction of phylogenetic trees [26], which can significantly minimize the memory requirements for their employment, many K-mers methods are developed for feature extraction [27]. Liu et al. [28] have developed a computational method based on auto-cross covariance transformation with K-mers composition. Wen has proposed a K-mers sparse matrix account for K-mer appearances in genetic sequences [29]. Recent methods such as FEGS [30] consider to use graphical tools and amino acid pairs to present better protein sequence features.

Traditional protein structural classification analysis only takes uses of protein sequences to classify or predict their structures, but never give further inspect on how these amino acid sequences encoding their structures. Therefore, how the amino acid combination and their physical properties influence the structures, and which critical sequence factors that own crucial impacts on the formation of different protein structures are still unknown. To uncover the dynamical nature between protein sequence and their structures and identifying the critical factors that influence the formation of different types of protein structures, this research aims to use complex network approaches to model the protein sequence feature interactions, and utilize statistical methods to examine the K-mers and feature series distributions. In this study, protein structural data from not only macro level of top structural classes and

folds but also micro level of structural motifs are analyzed. The outcomes of the research elucidate the mechanics of amino acid feature interaction, the key findings regarding can further be used for protein molecular design or developing new protein structural classifiers.

## 2. Methods

In this section, details of the network and statistical methods used in the study are introduced, where the flow charts are presented in **Figure 1** and **Figure 2**.

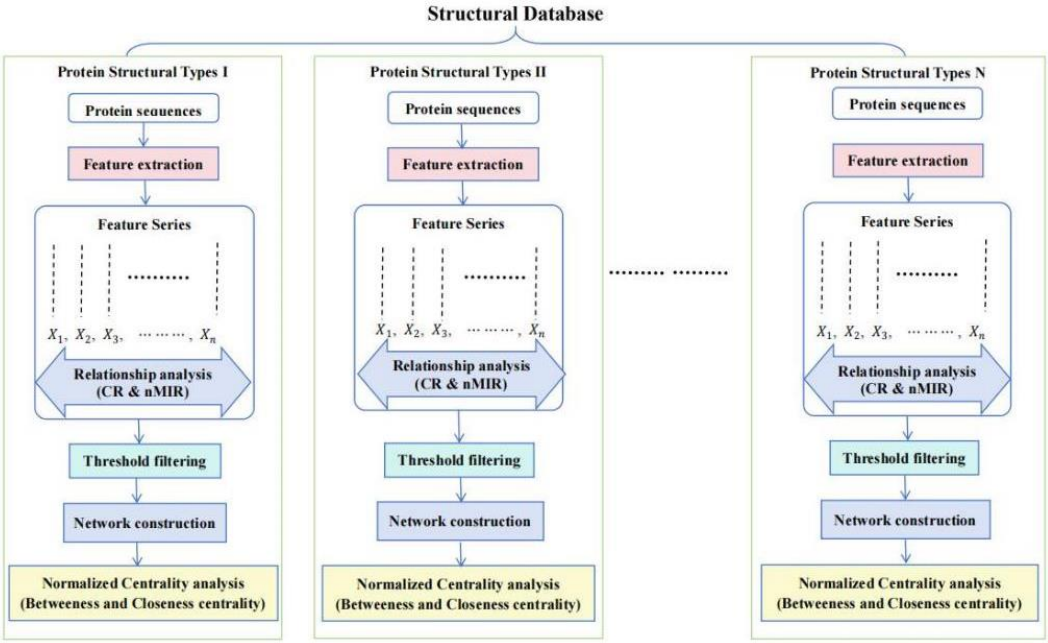


**Figure 1.** Diagram for the process of network analysis.

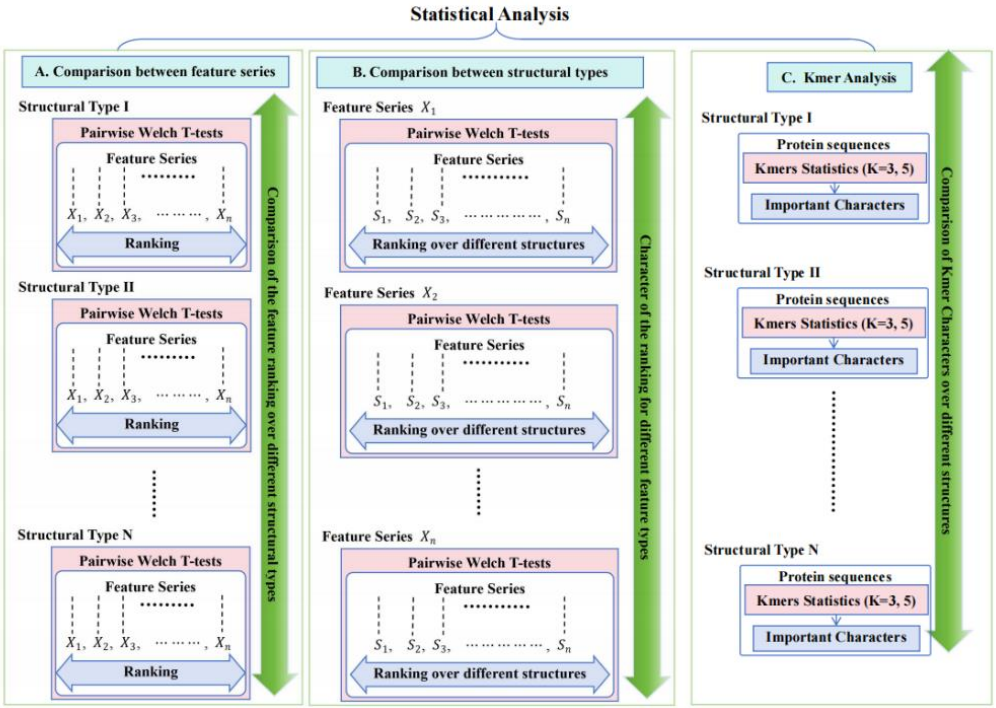


**Figure 2.** Diagram for the process of statistical analysis.

**Figure 1** presents the diagram for the process of network analysis. The network analysis is made up of feature extraction, the formation of feature series and relationship analysis, threshold filtering and network construction, as well as normalized centrality analysis.

**Figure 2** presents the diagram for the process of statistical analysis. The statistical analysis includes the comparison between feature series, the comparison between structural types, and the K-mers analysis.

## 2.1. Amino acid feature extractions

Let first recall some typical methods for amino acid sequence feature extractions.

### 2.1.1. Natural vector feature

The natural vector (NV) feature is a 60-dimensional real vector containing three parts [3], namely the amino acid composition numbers (abbreviated as the *N* features) $v_N = (n_A, n_R, \dots, n_V)$, the geometric mean distances (abbreviated as the $\mu$ features) $v_\mu = (\mu_A, \mu_R, \dots, \mu_V)$, and the second-order central moments (abbreviated as the D features) $v_D = (D_A^2, D_R^2, \dots, D_V^2)$, where $\mu_k = \frac{T_k}{n_k}$ denotes the mean distances from the amino acid k to the origin (initial amino acid), $D_k^2 = \sum_{i=1}^{n_k} \frac{(s[k][i]-\mu_k)^2}{n_k \cdot n}$ is the second order central moments of amino acid *k*, $T_k = \sum_{i=1}^{n_k} s[k][i]$, $s[k][i]$ denotes the distance between the first and the *i*-th *k*-type amino acid in the given sequence, k indicates one of the twenty types of amino acids [3], and the symbols $A, R$, ..., $V$ represent the twenty kinds of amino acids.

### 2.1.2. Averaged property factors

The averaged property factors (APF) is a 10-dimensional real vector (abbreviated as the P features) $v_P = (\langle F^{(1)} \rangle, \langle F^{(2)} \rangle, \dots, \langle F^{(10)} \rangle)$ describing the ten physical properties of amino acids [20], where $\langle F^{(i)} \rangle = \frac{1}{N} \sum_{k=1}^{N} f_k^{(i)}$(i = 1, 2, … , 10) stands for the mean value of the *i*-th factor [20,31], $f_k^{(i)}$ is the *i*-th factor of amino acid *k* and *N* is the residue number. The ten properties include the α-helix and bend preference ($P_1$, amino acid preferences for alpha-helix or bend-structures) [31], Side-chain size ($P_2$) [31], Extended structural preference ( $P_3$ , $\beta$ -structural preferences) [31], Hydrophobicity ( $P_4$ , hydrophobicity values for amino acids) [31], Double-bend preference ($P_5$, normalized proportions of double-bend specified by opposite signs of two continues virtual-bond torsion angles) [32], amino acid composition ($P_6$, mean values of amino acid compositions) [33], Flat extended preference ($P_7$, mean relative proportions of occurrence in extended structural regions $E_0$) [34], occurrence in α region ($P_8$, normalized frequency of α-regions identified by backbone torsion angles) [35], pK-*C* value ($P_9$, polarity parameter for solutes with certain dissociation degree in aqueous solution) [36], and the surrounding hydrophobicity ($P_{10}$, obtained by a group of hydrophobic indices for β-structures) [37]. Details of these properties are listed in Table V of [31].

### 2.1.3. Occurrence frequencies

Due to varying lengths of the sequences, simple amino acid numbers are not sufficient to characterize the composition features of amino acids, the 20-dimensional

occurrence frequency vector (abbreviated as the $F$ features) $v_F = (F_A, F_R, \ldots, F_V)$, corresponds to the trivial case of PseAAC when $\lambda = 0$ [24], is defined to better present the proportions of amino acids in the given protein sequences, where $F_k = n_k/N$ denotes the frequency of the *k*-type amino acid in a given sequence of *N* residues.

#### 2.1.4. Features for structural motifs

In PROSITE database, the sequence of structural motifs is recorded in patterns. To analyze the fixed amino acid combination in these patterns, the occurrence frequency feature (*F* feature): $V_F = (f_A, f_R, \ldots, f_V)$ and the averaged property factor feature (APF feature): $v_P = (\langle F^{(1)} \rangle, \langle F^{(2)} \rangle, \ldots, \langle F^{(10)} \rangle)$, are computed from the pattern sequence to extract the composition and physical property characters, where $f_k$ stands for the pseudo occurrence frequency for the *k*-type amino acid in the fixed pattern, $\langle F^{(i)} \rangle = \sum_{k=1}^{20} f_k \cdot f_k^{(i)}$, $f_k$ is the pseudo frequency for the *k*-type amino acid and $f_k^{(i)}$ is the property value for the *i*-th factor of the k-type amino acid, *k* = *A*, *R*, *N*, *D*, ...,*V*, *i* =1, 2, ..., 10.

## 2.2. Network analysis

#### 2.2.1. Relationship analysis

For a class of *N* protein sequences, when pile up all features, we can get an *N* × 90 matrix whose rows are feature vectors $v = (v_N, v_\mu, v_D, v_{APF}, v_F)$ and columns are feature series. In this matrix, all feature series are aligned in the same protein order (elements in the same positions correspond to the features extracted from the same protein). Use $X_1, X_2, \ldots, X_{90}$ to denote the 90-channel features, we compute the absolute correlation (CR):

$$R(i,j) = \left| \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{var}(X_i)\mathrm{var}(X_j)}} \right| \tag{1}$$

and the normalized mutual information rates (nMIR) [22]:

$$I(i,j) = \begin{cases} I(X_i; X_j)/H_{\max}, & \text{if } i \neq j \\ H(X_i)/H_{\max}, & \text{if } i = j \end{cases} \tag{2}$$

between the feature series $X_i$ and $X_j$ ($i, j$ = 1, 2, …, 90), here $H_{\max} = \max_i H(X_i)$ is the largest Shannon entropy for all series. The relations are all symmetric and scaled in [0,1], where CR measures the linear relations, while nMIR detects the mutual relations regardless of whether it is linear or not [22,38].

Since the order of the proteins is in consistent with the order of rows in the feature matrix, we can eliminate the protein order effect in the relationship analysis by performing random shuffles on feature series. We can shuffle the rows of the feature matrix, and compute CR and nMIR relations on every shuffled feature matrix. Since larger shuffle numbers may exhibit more robust results at the expense of longer computation times, we perform a standard deviation test on shuffle numbers as shown in Supplementary **Table S1**, where we use:

$$\overline{\sigma_R} = \frac{1}{90^2}\sum_{i,j=1}^{90}\sigma_R(i,j) \tag{3}$$

and:

$$\overline{\sigma_I} = \frac{1}{90^2}\sum_{i,j=1}^{90}\sigma_I(i,j) \tag{4}$$

to measure the average standard deviations for CR and nMIR, and $\sigma_R(i,j) = \sigma(R(i,j))$ and $\sigma_I(i,j) = \sigma(I(i,j))$ are standard deviations for CR and nMIR between $X_i$ and $X_j$. Results show that all shuffle numbers present tiny standard deviations compared to the overall relationship magnitudes, therefore we choose a moderate and balanced choice of 100 shuffles to perform our analysis.

### 2.2.2. Thresholds and significant relations

Since the relationship value ranges may vary between feature types, we define the adjusted scalar thresholds $\theta * M$ ($\theta \in [0,1]$) to filter the significant relations, where $M = \begin{pmatrix} M_{11} & \cdots & M_{15} \\ \vdots & \ddots & \vdots \\ M_{51} & \cdots & M_{55} \end{pmatrix}$ is the $5 \times 5$ block partition for the $90 \times 90$ matrix, where each block $M_{ij}$ is a constant matrix with all elements identical to the local maximum relationship value between feature types i and j, subscripts *i*, *j*=1, 2, ... , 5 denote the five feature types: *N*, *μ*, *D*, *F*, *P*, respectively. In the threshold filtering, relations above or equal to the threshold are set to 1, others are set to 0. This results in a binary adjacency matrix for unweighted networks.

The variation of the networks against the varying threshold can be observed from the spectral radius $\rho(A) = \max_i|\lambda_i|$ (leading eigenvalue) for the adjacency matrix *A*, in which $\lambda_i$ is the diagonal entry in the Jordan normal form *J* (a matrix uniquely characterizing each network) [39]. The dependency between the $\theta$ values and the number of significant relations is shown in **Figure 3**. In this figure, larger $\theta$ values may allow less significant relations, while smaller $\theta$ values may generate significant relations in massive amounts.

**Figure 3** depicts the numbers of significant feature relations and network spectral radius against different θ values in $[0.90, 0.99]$ for the different structural classes in CATH, SCOP and IDPs. The horizontal axis stands for θ values (scalar multiplicity of the thresholds), while the vertical axis represents the numbers of significant relations and the spectral radii.

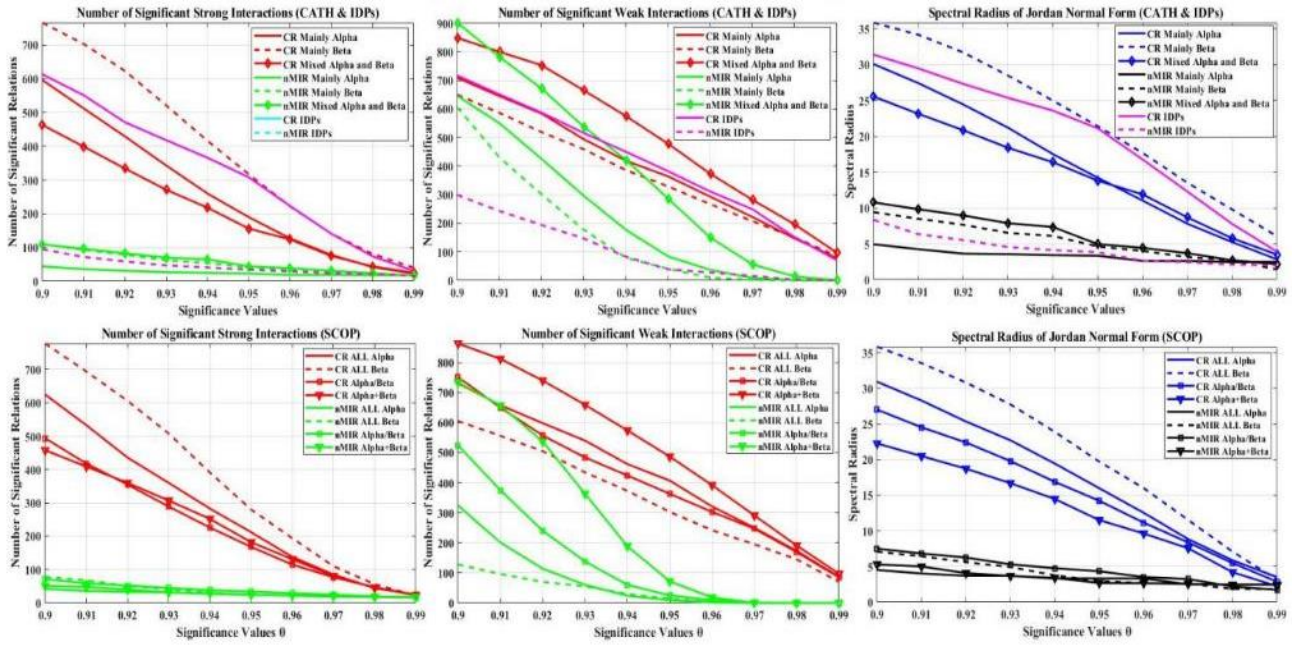


**Figure 3.** Threshold analysis of significant relations and spectral radius.

In testing the impact of thresholds on the relationship filtering, a medium value $\theta$ = 0.95 is found to be a proper choice for the filtering for the structural class data. As to the structural motifs in PROSITE, due to the sensitivity of the data on the thresholds, a lower value $\theta$ = 0.8 is used as the most appropriate choice for all nine types of the structural motifs. Worth noting, in all the filtering, the weak relations below $(1-\theta)\times M$ are massive in amount, which are less identifiable in characterizing the structures. Therefore, only significant relations above $\theta \times M$ are used in our analysis.

### 2.2.3. Construction of unweighted networks

The binary adjacency matrices $A = (a_{ij})_{90\times 90}$ obtained by the threshold filtering $(\theta \times M, \theta = 0.95)$ are used to construct unweighted feature networks. To avoid self-edges in the networks, the diagonal elements $a_{ii}(i = 1,2,\ldots,90)$ of the binary matrices are set to 0. The 90 vertices of the networks correspond to the 90 channel of feature series, the edges joining these vertices indicate the significant relations identified above the threshold $\theta \times M$. The networks globally contain both intra-type and cross-type relations. The networks are constructed for all 100 random shuffles of the feature series.

### 2.2.4. Centrality analysis

Centrality is a metric that associates every vertex a non-negative value estimating the importance of the vertices, where higher centrality values indicate higher level of importance of the vertices, which implicate more intensive feature interactions of the corresponding features. Here, we use the betweenness and closeness centrality, which address path connections, to analyse the importance of the vertices.

In a network of N vertices, the betweenness centrality is calculated as the ratio between the number of geodesic paths via $i$ and the number of all geodesic paths as follows [39]:

$$c_i = \frac{1}{N^2}\sum_{p,q}\frac{n_{pq}^i}{g_{pq}} \tag{5}$$

where $n_{pq}^i$ is the number of geodesic paths between vertices $p$ and $q$ via the vertex $i$, $g_{pq}$ is the total geodesic path number from $p$ to $q$, and the geodesic paths can be computed by specifying the powers of the binary adjacency matrices. The summation runs over all vertices $p$ and $q$, and the centrality value is normalized by the $N$ square to scale between [0,1].

The **c**loseness centrality [39], also relying on geodesic paths, accounts for the mean inverse distances from each vertex to all the other vertices and is defined as follows [39]:

$$c_i = \frac{1}{N-1}\sum_{j(\neq i)}\frac{1}{d_{ij}} \tag{6}$$

where $d_{ij}$ is the geodesic distance between vertex $i$ and $j$. The summation rules out the term when $j = i$, in that $d_{ii} = 0$ may obtain an infinite value for the centrality, which is unreasonable. This definition has a natural meaning that it assigns closer vertices with higher weights but farther vertices with lower weights, which also rules out the term $\frac{1}{d_{ij}}$ for vertices $i$ and $j$ in different connected components [39].

## 2.3. Statistical comparison between feature series

To compare feature value distributions of different structures, pairwise *T* tests are performed among the feature series. Since different types of features may have different distributions, which may lead to the non-homogeneity of the variance for the feature series. Welch *T* tests is kind of commonly used statistical tests that are widely used in biological data analysis [40], with the advantage that it is free from the homogeneity of variance for the datasets. Therefore, we choose to use the pairwise Welch *T* test in the feature comparison analysis.

### 2.3.1. Ranking of feature series

For each structural class, the 90 channel feature series correspond to the 90-dimensional features. The lengths of all series are consistent with the number of protein sequences in this class. Since different feature types may attain different value ranges, we first perform a standard Levene *F* test [41] between the feature series to check the homogeneity of the variances. All significance levels in $\Delta = \{0.25, 0.1, 0.05, 0.025, 0.01, 0.005\}$ suggest that the variances of the feature series are non-homogeneous. This may be due to the existence of both significant and non-significant feature relations. This does not disturb our study, we choose to perform pairwise Welch *T* tests [41], which do not rely on the homogeneity of variance for the feature series, to compare the feature series within each feature type.

For a specific structural class and feature type, let *K* denotes the number of feature series, the following *T* statistic is defined as:

$$T = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{S_i^2}{n_i} + \frac{S_j^2}{n_j}}} \sim T(\nu) \tag{7}$$

with $\nu = \frac{\left(S_i^2/n_i + S_j^2/n_j\right)^2}{\frac{\left(S_i^2/n_i\right)^2}{n_i - 1} + \frac{\left(S_j^2/n_j\right)^2}{n_j - 1}}$ degrees of freedom, where $n_i = n_j$ are sample sizes (the length of feature series), $\bar{X}_i$, $\bar{X}_j$ and $S_{i,k}^2$, $S_{j,k}^2$ are respectively the sample mean and sample variances for feature series $X_i$ and $X_j$ $(i \neq j, i, j = 1,2,\ldots,K)$. The following two sets of unilateral hypotheses are defined to compare the feature series:

(I)

$$H_0: \mu_i \leq \mu_j, H_1: \mu_i > \mu_j, \tag{8}$$

(II)

$$H_0': \mu_i \geq \mu_j, H_2: \mu_i < \mu_j \tag{9}$$

For a certain significance level $\delta$, if $T \geq T_\delta(\nu)$, then $\mathrm{H}_1: \mu_i > \mu_j$ holds, the feature values of $X_i$ are significantly larger than those of $\mathrm{X_j}$; otherwise, $\mathrm{H}_0: \mu_i \leq \mu_j$ holds, we need to further check (II) to judge the magnitudes. When checking hypothesis (II), if $T \leq -T_\delta(\nu)$, then $\mathrm{H}_2: \mu_i < \mu_j$ establishes, the values of $X_j$ are deemed significantly larger than $X_i$; otherwise, $\mathrm{H}_0': \mu_i \geq \mu_j$ holds and both hypotheses $\mathrm{H}_0: \mu_{i,k} \leq \mu_{j,k}$ and $\mathrm{H}_0': \mu_{i,k} \geq \mu_{j,k}$ establish, which implicates that no significant differences are found between $X_i$ and $X_j$.

As $\delta$ alters in $\Delta = \{0.25, 0.1, 0.05, 0.025, 0.01, 0.005\}$, similar results are obtained. When $\delta$ decreases, the rejection regions become narrower, fewer differences are identified; larger $\delta$ values e.g., 0.25 and 0.1, may generate wider rejection regions, thus more differences can be identified. Since we put more focus on the sequential differences rather than their equality, wider rejection regions with larger $\delta$ such as $\delta = 0.25$ are required. The pairwise *T* tests are performed for each type of the features and the structural classes.

For a feature type of *K* series, $C_K^2$ pairwise *T* tests are performed (hypotheses (I) and (II) together account for one test), and we use a ranking score $\mathrm{RK} = (s_1, s_2, \ldots, s_K)$ to visualize the comparison results. The rank score is initiated with a zero vector. If a feature series $X_i$ is deemed to have significantly larger values in one round of the *T* test, a score of 1 is added to $s_i$, and zero sores are added to the other features. If both features are equivalent, then both scores are unchanged. After $C_K^2$ rounds of pairwise *T* tests, the final score $\mathrm{RK} = (s_1, s_2, \ldots, s_K)$, denoting the total number of tests that each feature is judged with significantly higher values, defines a rank for the feature values. Higher RK scores imply larger magnitudes of the features. If two features get identical scores, the two features have deemed to attain equivalent magnitudes.

### 2.3.2. Feature comparison between different structural classes

Fixing a specific kind of feature $X_k$ (one of the 90 feature series), we compare

the values of $X_k$ between different structural classes. For two arbitrary structural classes $i$ and $j$, containing $n_i$ and $n_j$ number of protein sequences, the following two unilateral hypotheses:

(I)

$$\mathrm{H}_0: \mu_{i,k} \leq \mu_{j,k}, H_1: \mu_{i,k} > \mu_{j,k} \tag{10}$$

(II)

$$\mathrm{H}_0': \mu_{\mathrm{i,k}} \geq \mu_{\mathrm{j,k}}, \mathrm{H}_2: \mu_{\mathrm{i,k}} < \mu_{\mathrm{j,k}} \tag{11}$$

along with the $T$ statistics:

$$T = \frac{\overline{X_{\iota,k}} - \overline{X_{J,k}}}{\sqrt{\frac{S_{i,k}^2}{n_{i,k}} + \frac{S_{j,k}^2}{n_{j,k}}}} \sim T(\varepsilon) \tag{12}$$

are defined to compare the values of $X_k$ between structural classes $i$ and $j$. Here, $\varepsilon = \frac{\left(S_{i,k}^2/n_{i,k} + S_{j,k}^2/n_{j,k}\right)^2}{\frac{\left(S_{i,k}^2/n_{i,k}\right)^2}{n_{i,k}-1} + \frac{\left(S_{j,k}^2/n_{j,k}\right)^2}{n_{j,k}-1}}$) is the degree of freedom for $T$, $k$ = 1, 2, ..., 90 is the feature index, $i \neq j$ are indices for structural classes, and $\overline{X_{\iota,k}}, \overline{X_{J,k}}$ and $S_{i,k}^2$,$S_{j,k}^2$ are the sample mean and sample variances for $X_k$ in structural classes $i$ and $j$. The sample sizes $n_i, n_j$ denote the numbers of the protein sequences in classes $i$ and $j$. The pairwise Welch $T$ tests are performed for $X_k$ ($k$ = 1, 2, ..., 90) between the different structural classes. A ranking score $\mathrm{RK_{CATH}} = (s_\alpha, s_\beta, s_m, s_{\mathrm{IDPs}})$ (m represents the mixed $\alpha$ and $\beta$ class) is defined for CATH and IDPs, and $\mathrm{RK_{SCOP}} = (s_\alpha, s_\beta, s_{\alpha/\beta}, s_{\alpha+\beta}, s_{\mathrm{IDPs}})$ is defined for SCOP and IDPs, to account for the total number of tests that each structural class is judged to attain significantly larger feature values in the $C_N^2$ rounds of pairwise $T$ tests ($N$ = 4 for CATH and IDPs, and $N$ = 5 for SCOP and IDPs). Similar results with slight variations are obtained as $\delta$ varies in $\Delta = \{0.25, 0.1, 0.05, 0.025, 0.01, 0.005\}$. Here, the largest value $\delta = 0.25$ is implemented by our analysis to obtain more dissection results.

## 2.4. Kmer analysis

For the 20 types of amino acids, although there may exit $20^K$ number of possible Kmer combinations in total, however not all of them may occur in reality. The combination number $20^K$ grows exponentially as $K$ increases. To balance for both Kmer combinations and the frequency of appearance, smaller values such as $K$ = 3 and 5 are fund as the proper choices for the Kmer analysis. Here, we count the number of appearances for the K-mers, and examine the most frequent K-mers appeared in each of the different structural types.

# 3. Results

In this study, network and statistical methods are employed to analyze the feature interaction and feature value distribution for different protein structural classes and motifs.

## 3.1. Analysis of protein feature networks

### 3.1.1. Network analysis for protein structural classes

All protein sequence data (no greater than 30% similarity) in the CATH and SCOP databases as well as intrinsically structural disordered proteins (IDP) (no less than 80% content of disorder) in DisProt database are analyzed [16]. All CATH and SCOP data can be downloaded from Protein Data Bank (PDB, https://www.rcsb.org) using the PDB IDs listed in Supplementary Dataset S1. The intrinsically disordered proteins (IDPs) can be downloaded from DisProt database (https://www.disprot.org) by using the accession numbers provided in Supplementary Dataset S2. The CATH data contains three structural classes, i.e., the mainly $\alpha$ (1673 protein sequences), mainly $\beta$ (1772 protein sequences) and the mixed $\alpha$ and $\beta$ (4876 protein sequences) classes. The SCOP data contains 4 structural classes, i.e., the all-$\alpha$ (960 protein sequences), all-$\beta$ (1030 protein sequences), $\alpha/\beta$ (1490 protein sequences) and $\alpha + \beta$ (1356 protein sequences) classes. The IDPs data contains 3625 non-redundant structurally disordered region sequences. We extract the natural vector, averaged property factors and occurrence frequency features for these sequences, where each structural class corresponds to a 90-channel feature matrix, in which the row entries are the feature vectors and the column entries are the feature series aligned in the same protein order. The feature matrix results are provided in Supplementary Dataset S3.

One hundred random shuffles are performed on the feature series to eliminate protein order effects, where the CR and nMIR relations are computed between the shuffled feature series. Standard deviations for the relationship value are averaged to verify the robustness of the relations, the data are presented in **Table 1**. In this table, the average standard deviations are significantly lower than the mean relationship results, which indicates the relationship results are reliable and robust in our analysis.

**Table 1** presents the averaged standard deviations $\overline{\sigma_R}$ and $\overline{\sigma_I}$ for the CR and nMIR relations, where $\overline{\sigma_R} = \frac{1}{90^2}\sum_{i,j=1}^{90}\sigma_R(i,j)$ and $\overline{\sigma_I} = \frac{1}{90^2}\sum_{i,j=1}^{90}\sigma_I(i,j)$, $\sigma_R(i,j) = \sigma(R(i,j))$ and $\sigma_I(i,j) = \sigma(I(i,j))$ are the standard deviations of the CR and nMIR relations between features $X_i$ and $X_j$, $i, j$ = 1,2,...,90. The last row shows the mean relationship results.

**Table 1.** The average standard deviations of the relationship results.

| **Standard deviations** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **CATH** | | | **SCOP** | | | **IDPs** | |
| **Classes** | $\overline{\sigma_R}$ | $\overline{\sigma_I}$ | **Classes** | $\overline{\sigma_R}$ | $\overline{\sigma_I}$ | $\overline{\sigma_R}$ | $\overline{\sigma_I}$ |
| Mainly $\alpha$ | $5.558\times 10^{-16}$ | $7.691\times 10^{-17}$ | All-α | $5.025\times 10^{-16}$ | $9.288\times 10^{-17}$ | | |
| Mainly $\beta$ | $5.817\times 10^{-16}$ | $8.740\times 10^{-17}$ | All-β | $5.005\times 10^{-16}$ | $8.423\times 10^{-17}$ | | |
| Mixed $\alpha$ and $\beta$ | $7.929\times 10^{-16}$ | $6.000\times 10^{-17}$ | α/β | $5.103\times 10^{-16}$ | $7.463\times 10^{-17}$ | $1.837\times 10^{-15}$ | $1.37\times 10^{-16}$ |
| | | | α + β | $4.967\times 10^{-16}$ | $6.491\times 10^{-17}$ | | |
| Average $\sigma$ | $6.435\times 10^{-16}$ | $7.477\times 10^{-17}$ | Average σ | $5.025\times 10^{-16}$ | $7.916\times 10^{-17}$ | | |
| Mean relations | 0.366 | 0.0729 | Mean relations | 0.3590 | 0.0765 | 0.3544 | 0.1156 |

Take CATH data as an example, the intra-type feature relations for the three structural classes are shown in **Figures 4–6** (the heatmaps for the other datasets are

shown in Supplementary **Figures S1–S25**). In these figures, the composition and arrangement features are represented by amino acid abbreviations, the APF features are represented by their property indices. The colors indicate the magnitudes of the relations, and the blue bars indicate weak relations between the features.

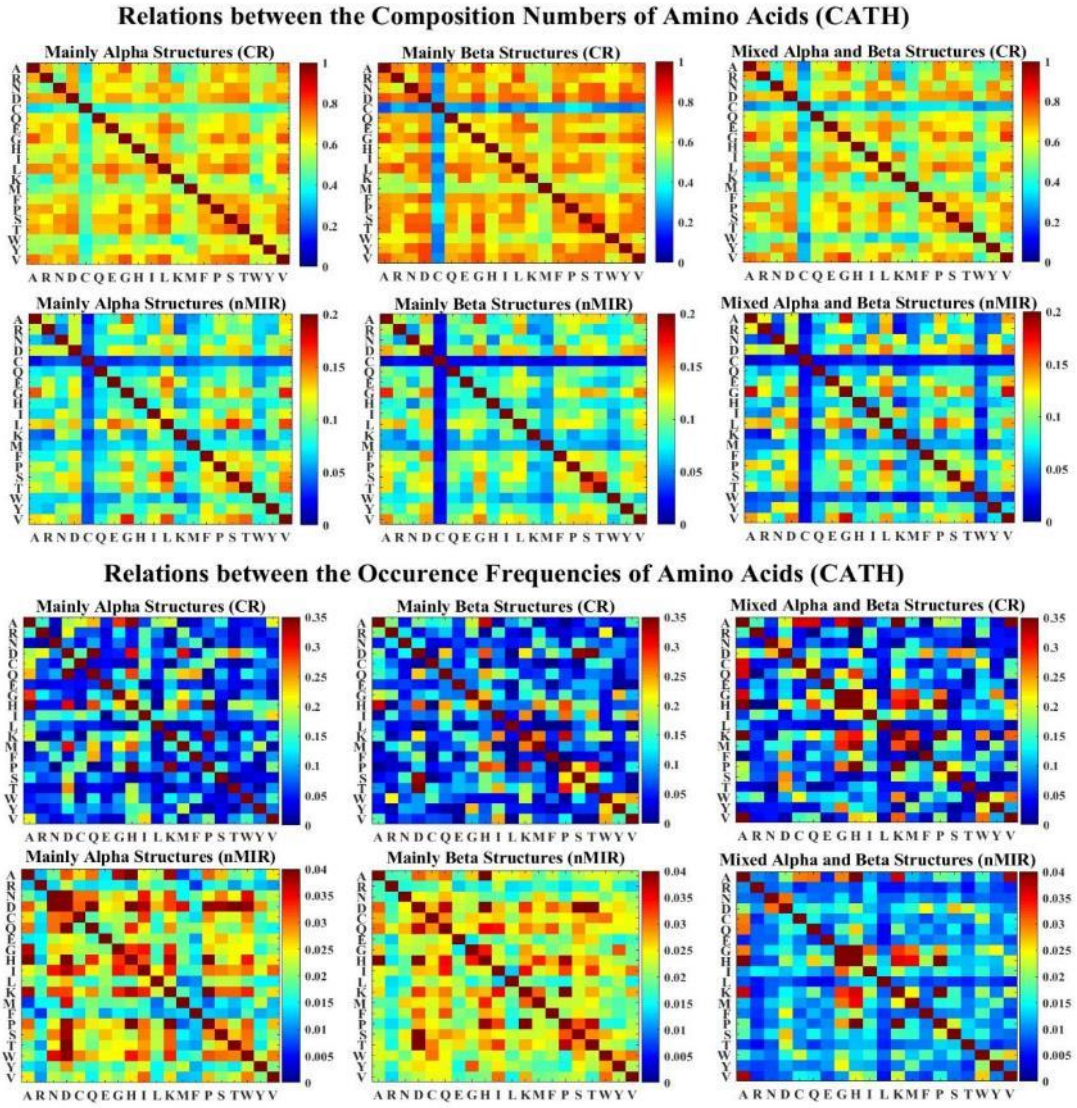


**Figure 4.** Heatmaps for the intra-type relations between composition features (CATH).

**Figure 4** shows the mean (intra-type) relationship matrices between the composition features of CATH. The colors represent the magnitudes of the relations indicated in the color bar. In the color bars, the colder the colors indicate the lower the relationship values, while the warmer the colors interpret the higher the relationship values. In each subplot, the dark blue color represents the lowest relationship value, while the dark red color represents the highest relationship value. To better display the feature interactions, the colors are upper-truncated by the maximum off-diagonal elements of the matrices. The row and column labels represent the composition features.

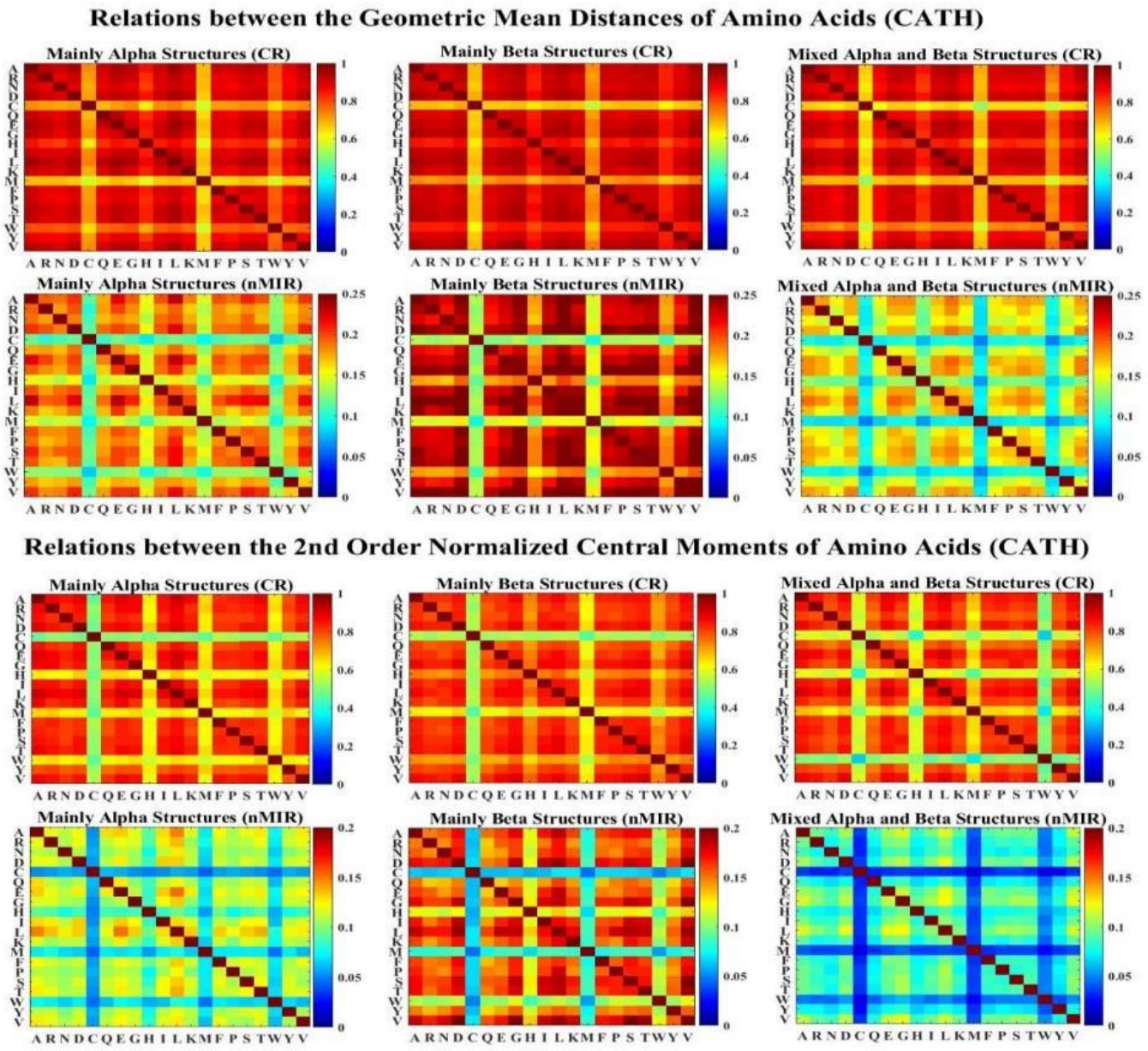


**Figure 5.** Heatmaps for the intra-type relations between arrangement features (CATH).

**Figure 5** depicts the mean (intra-type) relationship matrices between amino acid arrangement features of CATH. The colors represent the magnitudes of the relations indicated in the color bar. In the color bars, the colder the colors indicate the lower the relationship values, while the warmer the colors interpret the higher the relationship values. In each subplot, the dark blue color represents the lowest relationship value, while the dark red color represents the highest relationship value. The colors are upper-truncated by the maximum off-diagonal elements in the matrices, and the row and column indices represent the arrangement features.

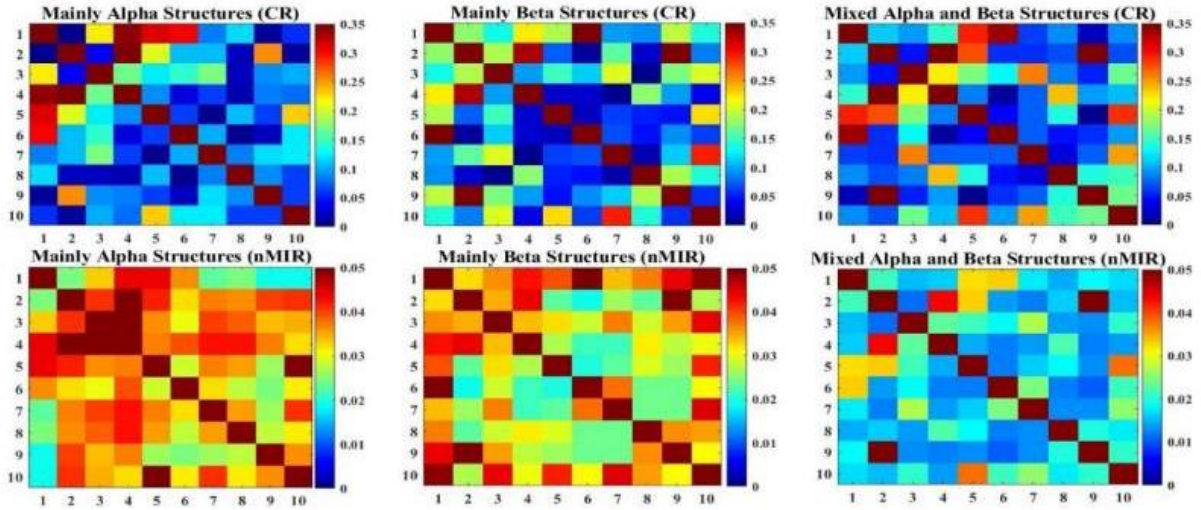


**Figure 6.** Heatmaps for the intra-type relations between physical property features (CATH).

**Figure 6** depicts the mean (intra-type) relationship matrices between physical property features of CATH. The colors represent the magnitudes of the relations indicated in the color bar. In the color bars, the colder the colors indicate the lower the relationship values, while the warmer the colors interpret the higher the relationship values. In each subplot, the dark blue color represents the lowest relationship value, while the dark red color represents the highest relationship value. The colors are upper-truncated by the maximum off-diagonal elements in the matrices. The integer channel indices denote the ten physical properties of the APF features: $P_1$ (the α-helix and bend preference, $P_2$ (side-chain size), $P_3$ (extended structural preference), $P_4$ (hydrophobicity), $P_5$ (double-bend preference), $P_6$ (amino acid composition), $P_7$ (flat extended preference), $P_8$ (occurrence in $\alpha$ region), $P_9$ (pK-$C$ value), $P_{10}$ (the surrounding hydrophobicity indices for $\beta$-structures).

In both CATH and SCOP datasets, common characters are identified between the different structural classes, e.g., the Asp, Leu, and Val present as common sources of significant intra-type feature relations, whereas His, Cys, Met, and Trp present weak intra-type relations with other features.

### 3.1.2. Significant feature relations by threshold filtering

We use the mean relationship matrices to perform the threshold analysis. An example of the significant intra-type CR relations for CATH is shown in **Figure 7**. In this figure, the relations below the thresholds $\theta \times M$ ($\theta = 0.95$) are set to 0 (background blue), where the yellow lattices indicate the significant CR relations above the thresholds. The plots of other datasets are shown in Supplementary **Figures S26–S38**, where typical significant relations are summarized in Supplementary **Tables S2–S3**. An example of the network structure constructed from the significant CR relations for the all α class of CATH is shown in **Figure 8**.

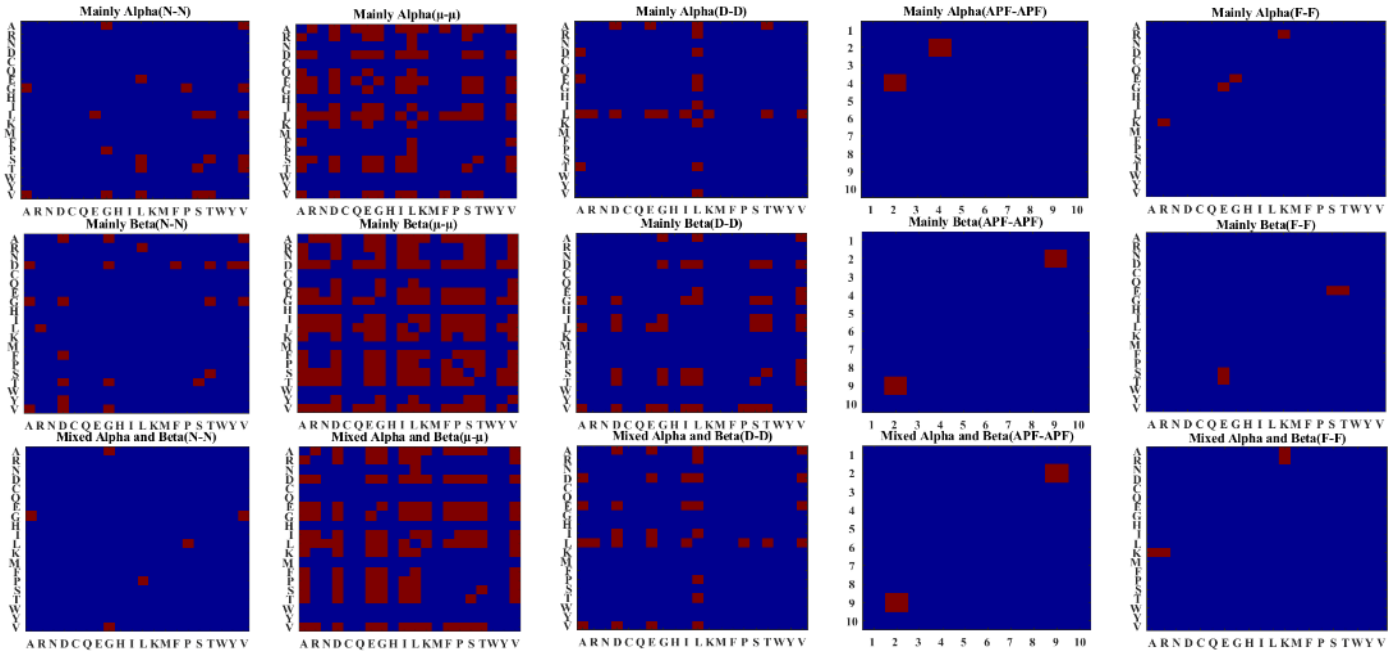


**Figure 7.** Significant intra-type CR relations between protein sequence features (CATH).

**Figure 7** plots the significant intra-type CR relations above the thresholds $\theta \times M$ ($\theta = 0.95$) for CATH. The five column panels indicate the significant feature relations for different types of features, while the three row panels indicate the significant relations for different structural classes. The yellow color indicates significant feature relations, while the background blue indicates non-significant relations. The diagonal elements of the self-relations are set off in our analysis.

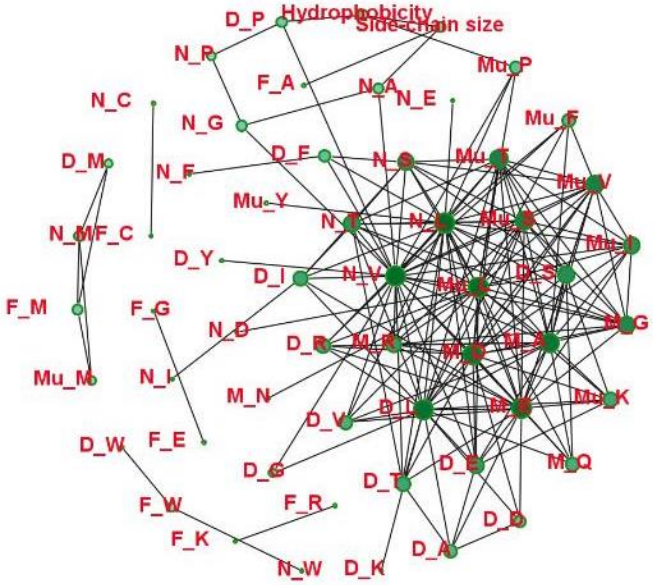


**Figure 8.** A network example of the CR relations for the all α class of CATH.

**Figure 8** shows an example of the unweighted undirected networks constructed by the significant CR relations for the all α class of CATH. The red vertex labels denote the different features, while the purple edges represent the interactions (significant relations) between the vertices. The fill color and the size of the vertices

reflects the degrees of the vertices, the darker green colors and the larger point radius indicate higher degrees, while the lighter green colors and the smaller point radius indicate smaller degrees.

Results show that the Asp, Leu, and Val act as the main sources of strong feature relations, while Cys, His, Trp, Met played as the sources of weak intra-type relations. Significant feature interactions are also identified between the compositions of Ala and side-chain size, between the numbers of Gly and Ala, Val, as well as between the composition features and the α-helix and bend preference property, and the sequence arrangements of Ala also acted as a key source of the feature interactions.

In the α structural analysis, the Glu features act as the main sources of feature interactions, while significant interactions are found between Pro and hydrophobicity, and between the frequencies of Arg and Lys. The mainly α class (CATH) exhibits significant connections between the arrangements of Leu and Ala and between the compositions of Ser and Leu, while all α class (SCOP) also shows significant relations between the arrangements of Trp and hydrophobicity, and between the arrangements of Met and double-bend preference.

However, in $\beta$ structural analysis, the Gly features are found as the main sources of features interactions, where interactions are also identified between side-chain size and pK-*C* values and between the arrangements of Thr with other amino acids. The $\beta$ structures also present intensive interactions for Thr, Phe, Tyr, side-chain size, pK-*C* value, extended structural preference, and surrounding hydrophobicity for $\beta$ structures properties. The mainly $\beta$ class structures (CATH) also show strong connections between the extended structural preference, surrounding hydrophobicity for $\beta$ structures and sequence arrangement features, and between the arrangements of Ser, Thr, Ile with other amino acids. The all $\beta$ class structures (SCOP) also show interactions between the composition of Lys and surrounding hydrophobicity for $\beta$ structures property, between the compositions of Ser and Thr, and between the composition of Gly and pK-*C* value, as well as between $\alpha$-helix and bend preference and the arrangements of Tyr.

Moreover, both the $\alpha$ and $\beta$ structures are found to admit intensive interactions between the arrangements of Ser and other amino acids. The mainly α and mainly $\beta$ classes also present intensive interactions between the compositions of Cys and Arg, while the all $\alpha$ and all $\beta$ classes also present strong connections between hydrophobicity and amino acid arrangement features.

The mixed structures share common characteristics with the $\alpha$ and $\beta$ structures. For instance, the mix type of structures show similarity with the α structures in terms of the intensive interaction for Glu, between hydrophobicity and amino acid arrangements, between the compositions of Lys and Arg, and between the composition of Ala and side-chain size. The mixed structures also show similarity with $\beta$ structures regarding the significant feature interactions for Gly, between side-chain size and pK-*C* value, between the arrangements of Trp, Tyr and $\alpha$-helix and bend preference, and between the arrangements of Cys and amino acid composition features.

The mixed structures also contain special characters in terms of the interactions between Met, Lys and double-bend preference, between the arrangements of Cys, His, Met, Tyr and amino acid composition features, and between the arrangements of Gly

and Glu, as well as between the arrangements of Ala, Gly, Glu, Ile with other amino acids. The mixed structures also show strong connections between double-bend preference, flat extended preference and surrounding hydrophobicity for $\beta$-structure properties and between the arrangement of Cys and the composition of Arg.

The IDPs show different characteristics apart from typical structural classes. It presents intensive feature interactions among the composition numbers of Glu, Val, Lys, among the arrangement features of Ala, Arg, Asp, Gln, Glu, Gly, Ile, Leu, and Val, between the numbers of Leu and Ile as well as Ser and Thr, and between the composition number of Glu and sequence arrangements of Gly.

### 3.1.3. Centrality analysis

Global feature networks are constructed by the significant relations, where betweenness and closeness centrality is computed for all vertices (features) in the networks. The centrality values are normalized in [0,1] by the maximum centrality in the same network. Higher centrality values implicate higher importance of the vertices. The normalized centrality values are plotted in **Figures 9** and **10** (CATH and IDPs) and **Figures 11** and **12** (SCOP and IDPs). In these figures, the centrality results are presented by colors, where the rows represent the feature, while the columns stand for the structural classes. Both centrality measures present similar results, where the (model-free) nMIR networks show more dissection results than the (linear) CR networks, therefore we mainly discuss the nMIR results.

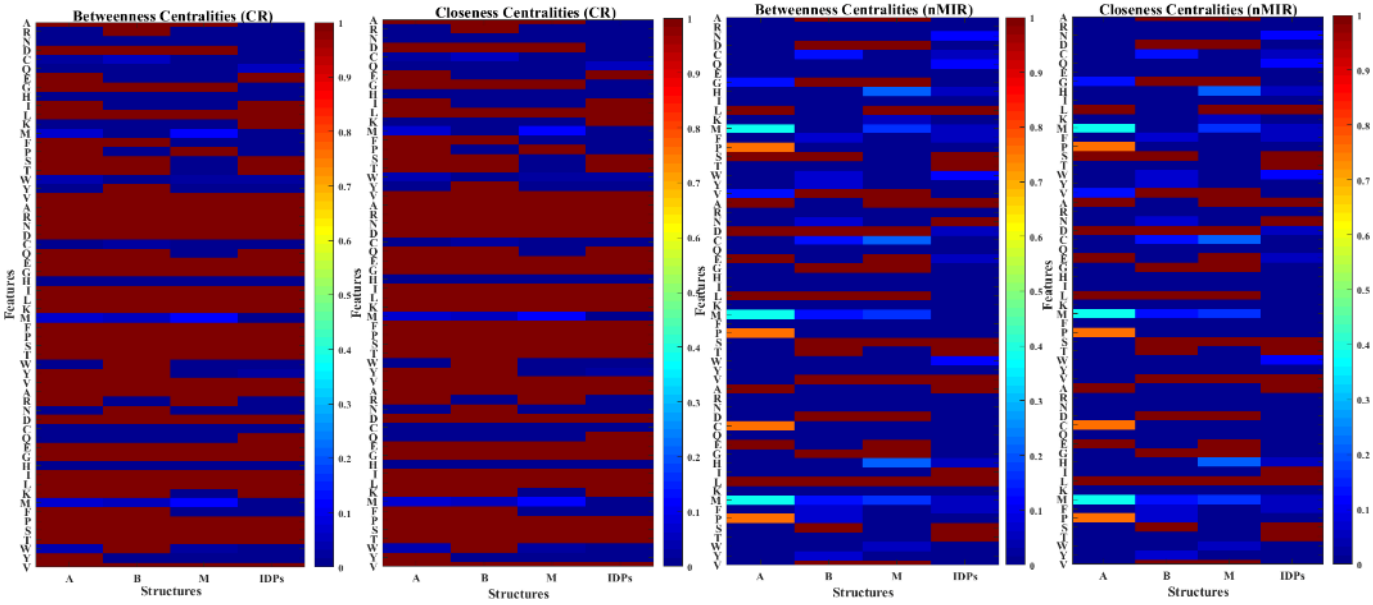


**Figure 9.** Centrality plots for the natural vector features (CATH and IDPs).

**Figure 9** shows the normalized betweenness and closeness centrality outcomes for the natural vector features of the CATH and IDPs data. Among the four panel plots, the left two panels are respectively the normalized centrality results for the CR networks, while the right two panels present the normalized centrality results for the nMIR networks. The rows are the natural vector features labeled by amino acid abbreviations (the top 20 rows are for N features, the middle 20 rows are for $\mu$ features, and the bottom 20 rows are for D features), and the columns represent the different structural classes. The colors indicate the magnitudes of the normalized centrality

values as presented in the color bar. In the color bars, the colder the colors indicate the lower the normalized centrality values, while the warmer the colors interpret the higher the normalized centrality values. In each subplot, the dark blue color represents the lowest normalized centrality value, while the dark red color represents the highest normalized centrality value.

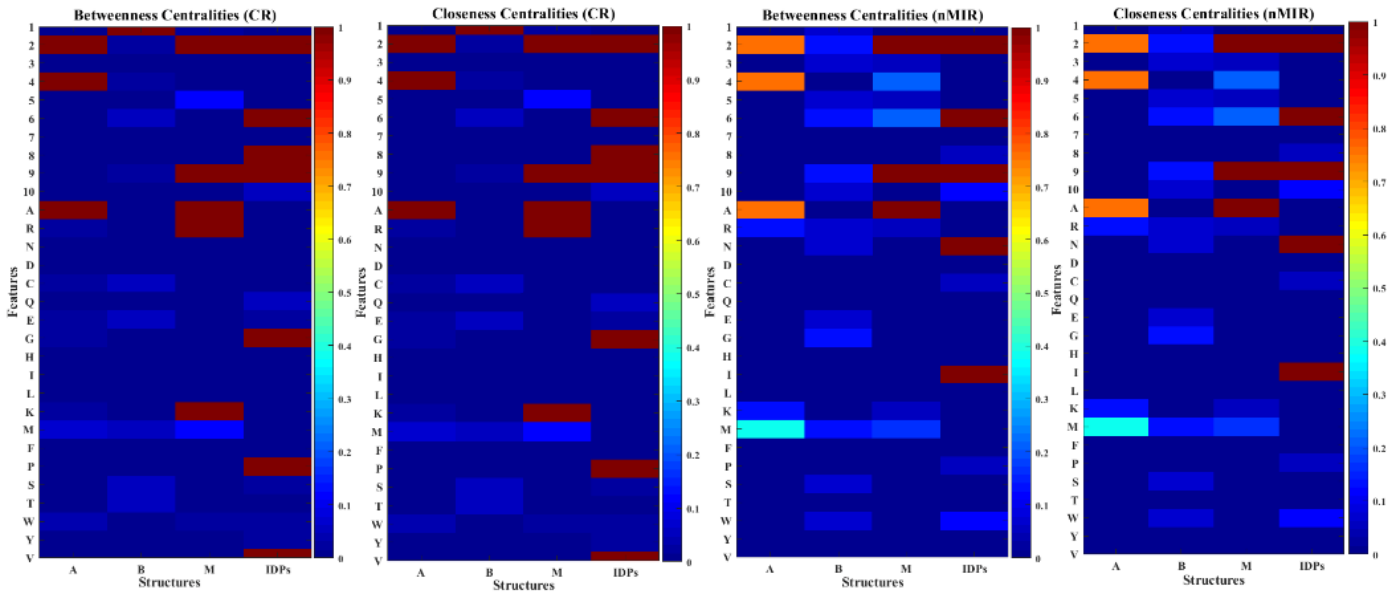


**Figure 10.** Centrality plots for the APF and F features (CATH and IDPs).

**Figure 10** shows the normalized betweenness and closeness centrality outcomes for the APF and *F* features of the CATH and IDPs data. Among the four panel plots, the left two panels are respectively the normalized centrality results for the CR networks, while the right two panels present the normalized centrality results for the nMIR networks. The rows stand for the ten physical properties and amino acid frequency features, while the columns represent the different structural classes. The colors of the lattices indicate the normalized centrality values as presented in the color bar. In the color bars, the colder the colors indicate the lower the normalized centrality values, while the warmer the colors interpret the higher the normalized centrality values. In each subplot, the dark blue color represents the lowest normalized centrality value, while the dark red color represents the highest normalized centrality value.

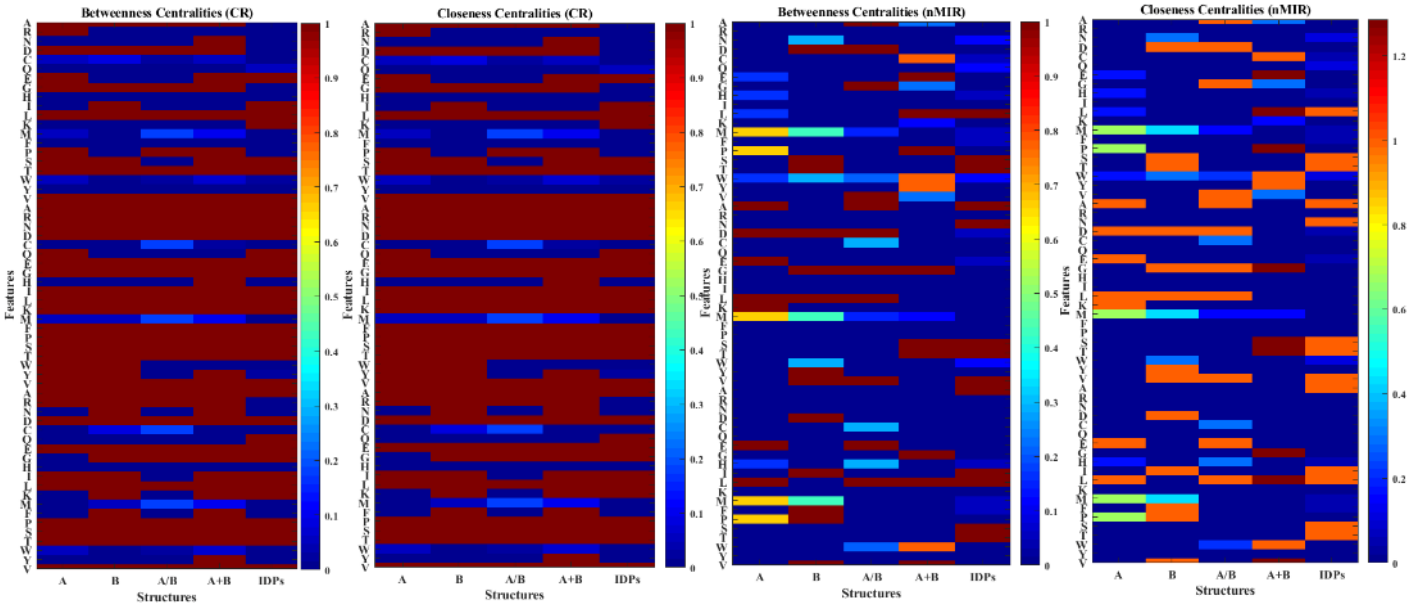


**Figure 11.** Centrality plots for the natural vector features (SCOP and IDPs).

**Figure 11** shows the normalized betweenness and closeness centrality outcomes for the natural vector features of the SCOP and IDPs data. Among the four panel plots, the left two panels are respectively the normalized centrality results for the CR networks, while the right two panels present the normalized centrality results for the nMIR networks. The rows are the natural vector features labeled by amino acid abbreviations; the columns are for the different structural classes. The colors of the lattices indicate the normalized centrality values as presented in the color bar. In the color bars, the colder the colors indicate the lower the normalized centrality values, while the warmer the colors interpret the higher the normalized centrality values. In each subplot, the dark blue color represents the lowest normalized centrality value, while the dark red color represents the highest normalized centrality value.

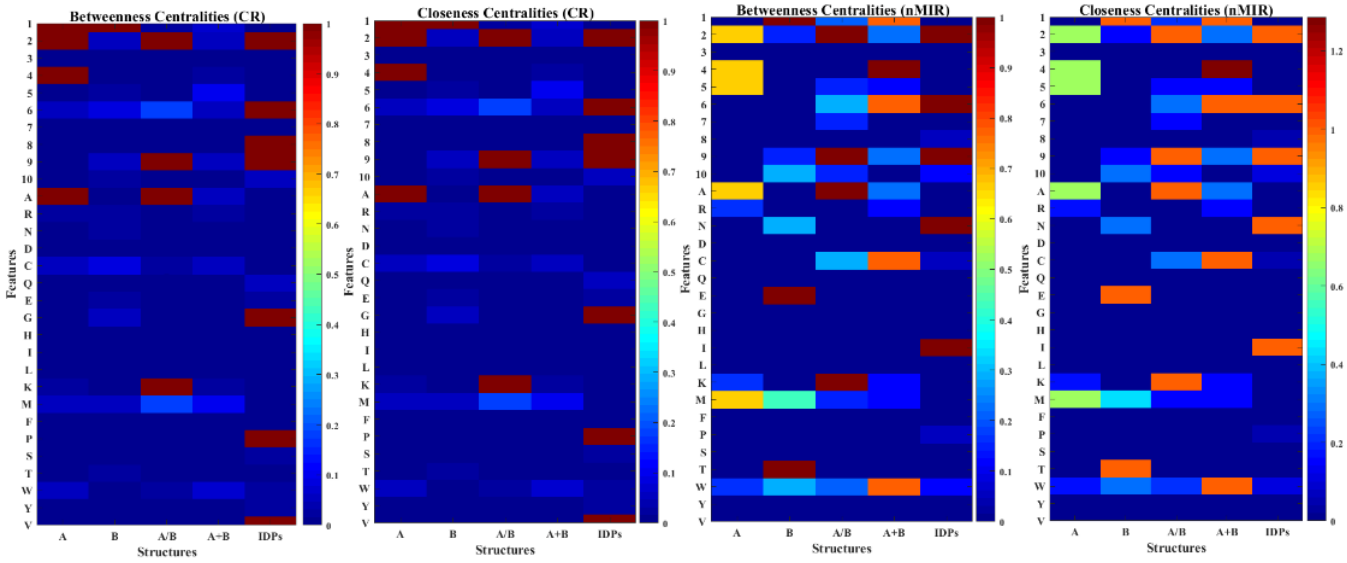


**Figure 12.** Centrality plots for the APF and *F* features (SCOP and IDPs).

**Figure 12** shows the normalized betweenness and closeness centrality outcomes

for the APF and F features of the SCOP and IDPs data. Among the four panel plots, the left two panels are respectively the normalized centrality results for the CR networks, while the right two panels present the normalized centrality results for the nMIR networks. The rows stand for the ten physical properties and amino acid frequency features, the columns are for the different structural classes. The colors of the lattices indicate the normalized centrality values as presented in the color bar. In the color bars, the colder the colors indicate the lower the normalized centrality values, while the warmer the colors interpret the higher the normalized centrality values. In each subplot, the dark blue color represents the lowest normalized centrality value, while the dark red color represents the highest normalized centrality value.

In **Figures 9** and **10**, all structural classes show high centralities for the arrangements of Leu. The mainly $\alpha$ class presents low centrality results for Gly and Val, but strong centrality results for Glu, Pro, Met, and for the compositions of Ala, Ser, side-chain size and hydrophobicity properties. The mainly $\beta$ class shows low centrality results for Glu but strong centrality results for Gly, Thr, Ser, Asp, Tyr, Val, and for the $\alpha$-helix and bend preference property. The mixed $\alpha$ and $\beta$ structures attain high centrality results for Asp, Val, Glu, Gly, and for the compositions of Ala, Arg, Lys, side-chain size and pK-*C* values. The mixed structural class shows similar trends with the mainly $\alpha$ class in terms of the strong centrality for Glu, side-chain size and the composition of Ala, while the mixed structural class also shows similarities with the mainly $\beta$ class structures in terms of the strong centrality for Gly.

The SCOP results are similar to the CATH results with slight differences. The all $\alpha$ class also obtains high centrality for double-bend preference, while the all $\beta$ class structures also attain strong centrality for the arrangements of Phe, Pro, Ile. The $\alpha/\beta$ class shows high centrality values for Ala and Lys, while the $\alpha+\beta$ class structures show strong centrality results for Trp and for the compositions of Trp, Glu, Leu, Pro, and Cys, as well as $\alpha$-helix and bend preference, hydrophobicity, and amino acid composition properties. The features of Leu are important in nearly all structural classes, while Glu, Pro, and hydrophobicity are important for α structures, and Gly, Ser, Thr, $\alpha$-helix and bend preference are important for $\beta$ structures. The mixed structures contain similarities with the $\alpha$ and $\beta$ structures but also peculiar characteristics in terms of the composition of His, side-chain size, pK-*C* values, and amino acid composition.

The IDPs present similar centrality distributions with the mixed structures in terms of the high centrality values for side-chain size and pK-*C*. It also shows special characteristics such as high centrality for the composition and arrangements of Leu, Ser, Thr, Val, Ile, and Asn; for the composition of Gly and Pro; for the arrangements of Ala; and for physical properties such as side-chain size, amino acid composition, occurrence in the α region and pK-*C* value.

### 3.1.4. Network Analysis for structural motifs

In this section, we analyse the networks for the typical structural motifs in PROSITE database (https://prosite.expasy.org/). In the PROSITE database, nine typical types of structural motifs are found with sufficient pattern data to perform our analysis, they are namely, the $\alpha$ type motifs, $\beta$ type motifs, $\alpha$ and $\beta$ type motifs, $\beta$ harping rings, EF-hands, βαβ motifs, αβα motifs, motifs with only loop structures,

and αββ motifs. The accession numbers of these data can be found in Supplementary Dataset S7.

In this analysis, we consider the composition and physical property characters for the fixed patterns in the PROSITE database. The arrangement features of amino acids are later analyzed by K-mers. We extract the pseudo compositional features (F features) and the average property factor (APF features) for the fixed patterns and compute the CR and nMIR relations between these features. Heatmaps for the relationship matrices and significant feature relations are shown in **Figures 13–16**. In these figures, the different structural motifs obtain different distributions for their significant feature relations.

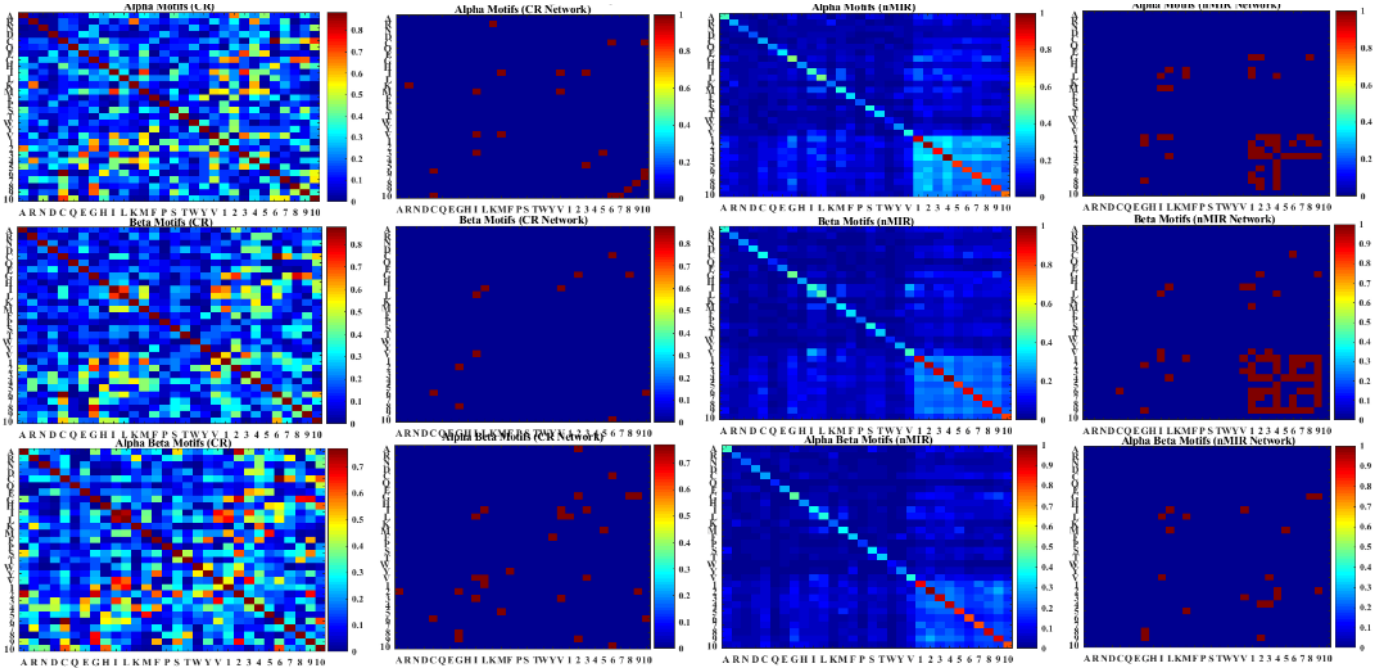


**Figure 13.** Heatmaps for the feature relations of $\alpha$, $\beta$ and $\alpha\beta$ motifs.

**Figure 13** presents the heatmaps for the mean relationship matrices and significant feature relations (filtered by thresholds) for $\alpha$, $\beta$ and $\alpha\beta$ motifs. The colors indicate the magnitudes of the relations, as shown in the color bar. The left two panels stand for CR relations, while the right two panels stand for nMIR relations. The different row panels represent the different structural motifs. The diagonal elements are self-relations, which are not concerned in our analysis, therefore, the colors of the relationship matrices are upper truncated by the maximum off-diagonal elements for better visualization. In the plots of the significant relationship matrices (the 2nd and 4th column panels), the diagonal elements are set to 0 in the process of threshold filtering to avoid self-loops in networks.

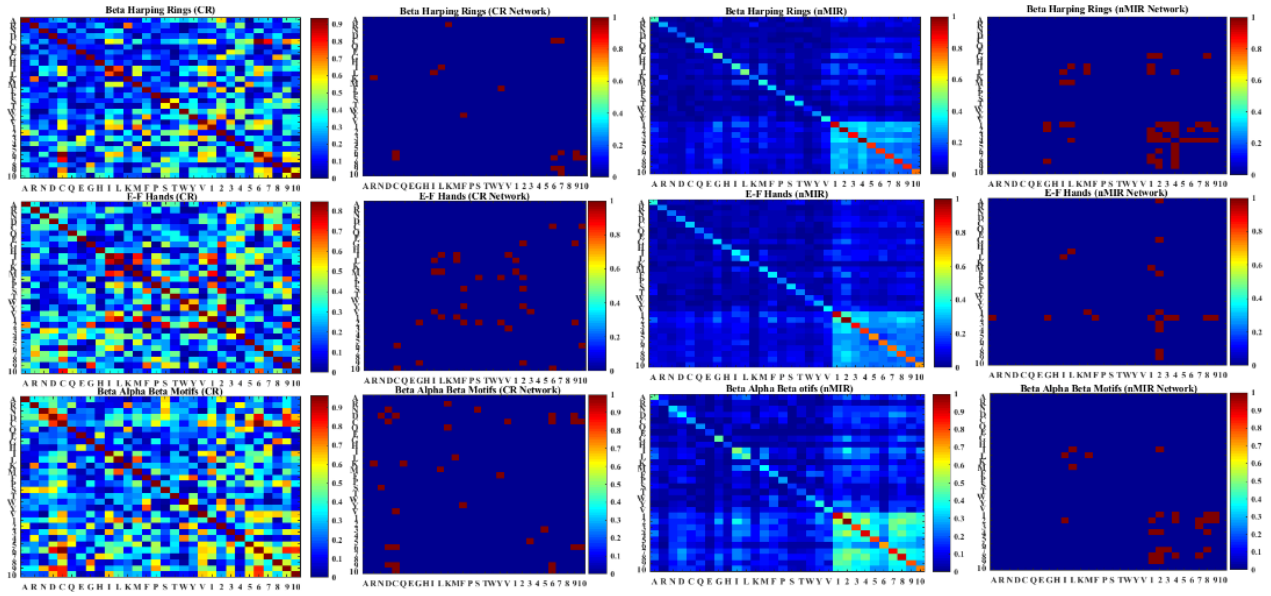


**Figure 14.** Heatmaps for the feature relations of $\beta$ harping rings, EF-hands, and $\beta\alpha\beta$ motifs.

**Figure 14** presents the heatmaps for the mean relationship matrices and significant feature relations (filtered by thresholds) for $\beta$ harping rings, EF-hands, and $\beta\alpha\beta$ motifs. The colors indicate the magnitudes of the relations, as shown in the color bar. The left two panels stand for CR relations, while the right two panels stand for nMIR relations. The different row panels represent the different structural motifs. The colors of the relationship matrices are upper truncated by the maximum off-diagonal elements, and in the plots of the significant relationship matrices (the 2nd and 4th column panels), the diagonal elements are set to 0 after threshold filtering.

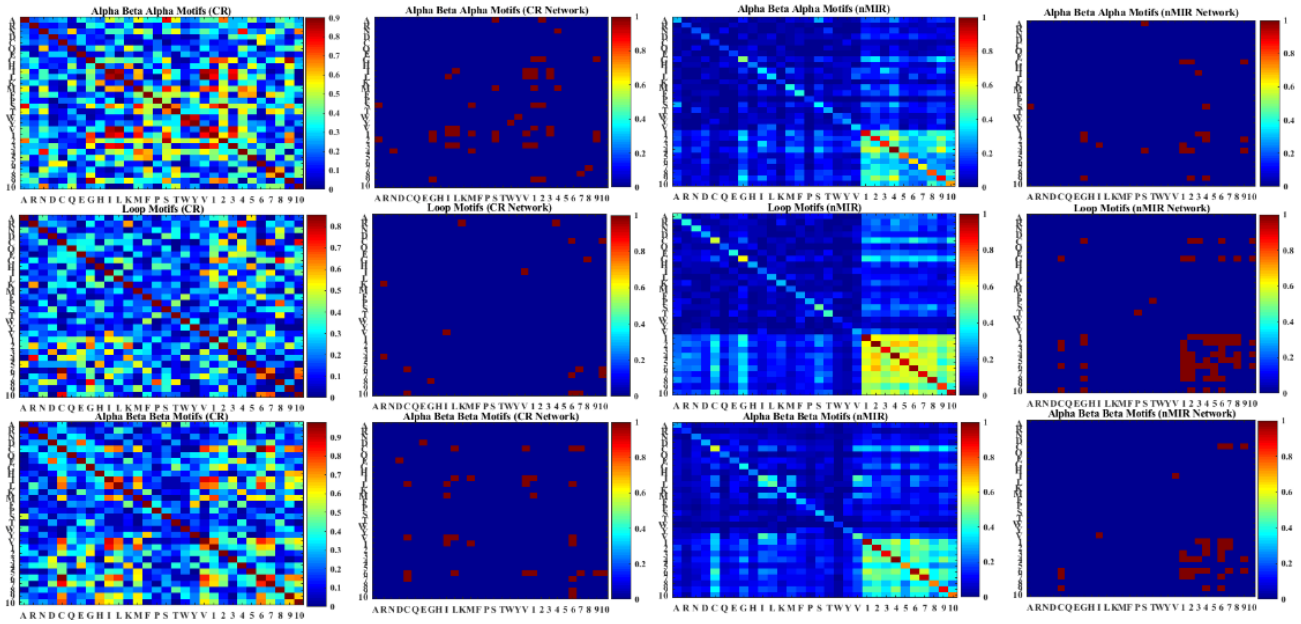


**Figure 15.** Heatmaps for the feature relations of $\alpha\beta\alpha$ motifs, motifs with mainly loop structures, and $\alpha\beta\beta$ motifs.

**Figure 15** presents the heatmaps for the mean relationship matrices and significant feature relations (filtered by thresholds) for $\alpha\beta\alpha$ motifs, motifs with mainly loop structures, and $\alpha\beta\beta$ motifs. The colors indicate the magnitudes of the

relations, as shown in the color bar. The left two panels stand for CR relations, while the right two panels stand for nMIR relations. The different row panels represent different structural motifs. The colors of the relationship matrices are upper truncated by the maximum off-diagonal elements, and in the plots of the significant relationship matrices (the 2nd and 4th column panels), the diagonal elements are set to 0 after threshold filtering.

The unweighted networks are constructed from significant relationship matrices, where the betweenness and closeness centralities are also computed for the different structural motifs. Since the networks are sparse, a lower threshold $\theta = 0.8$ is used to filter the significant relations. The betweenness and closeness centrality results are presented in **Figure 16**. In this figure, the amino acids Cys, Gly, Ile, Leu, and Val attain high centrality in the networks.

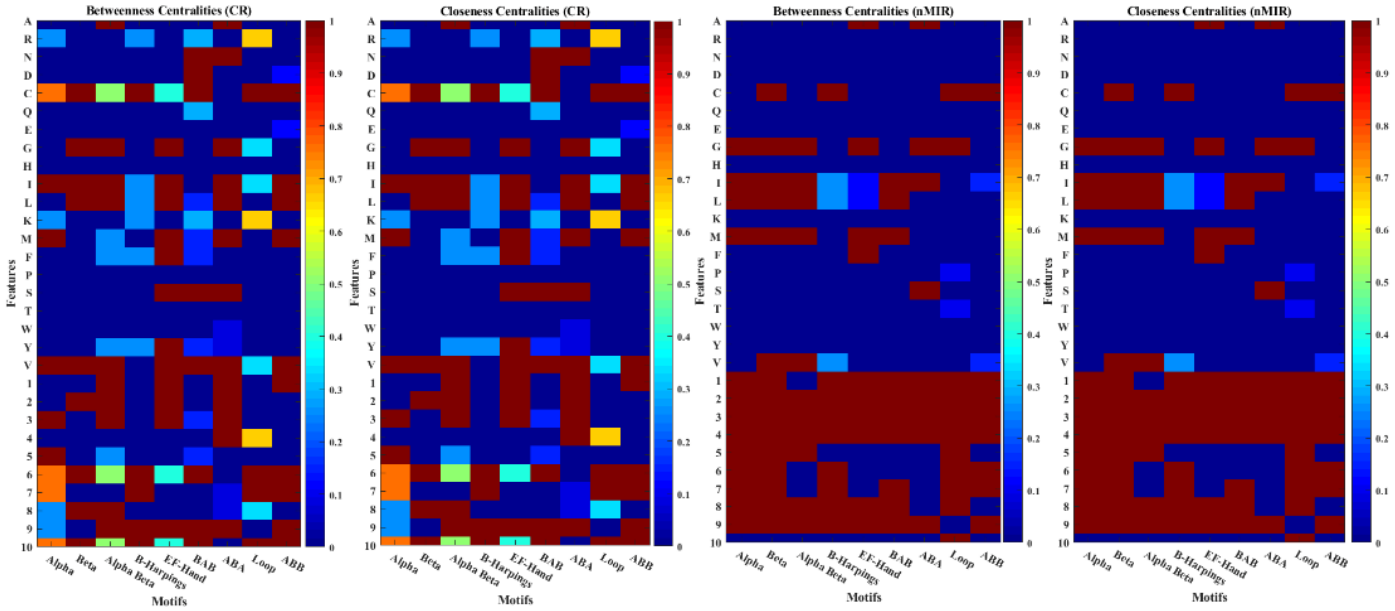


**Figure 16.** Centrality plots for the different structural motifs.

**Figure 16** presents the heatmaps for the betweenness and closeness centrality results for the different structural motifs. Among the four panel plots, the left two panels respectively show the normalized centrality results for the CR networks, while the right two panels present the normalized centrality results for the nMIR networks. The rows stand for the amino acid features, while the columns represent the different structural motifs. The colors of the lattices indicate the magnitudes of the normalized centralities as represents by the color bar. In the color bars, the colder the colors indicate the lower the normalized centrality values, while the warmer the colors interpret the higher the normalized centrality values. In each subplot, the dark blue color represents the lowest normalized centrality value, while the dark red color represents the highest normalized centrality value.

The $\alpha$ type motifs show high centrality (strong amino acid interactions) for Arg and Met, while the $\beta$ type motifs show high centralities for Cys and Gly. The $\beta$ harping rings, $\beta\alpha\beta$ motifs, loop structures and $\alpha\beta\beta$ motifs also attain high centrality for Cys. The $\beta$ harping rings and loop structures show lower centralities for Ile and Leu than other structures. The EF-hands tend to present high centralities for Met and

Phe, while the $\alpha\beta\alpha$ motifs show high centralities for Ala and Ser. The loop structures also attain high centrality for the frequency of Ala. The $\beta$ harping rings, EF-hands, $\beta\alpha\beta$ motifs, $\alpha\beta\alpha$ motifs, and $\alpha\beta\beta$ motifs together show low centralities for double-bend preference, while the $\alpha$ and $\beta$ motifs, EF-hands, and $\alpha\beta\alpha$ motifs present low centrality values for the amino acid composition and flat extended preference properties. The $\alpha\beta\beta$ motifs show low centralities for the double-bend preference and the occurrence in the α region properties.

## 3.2. Statistical analysis of feature series

### 3.2.1. Comparison between different features

We first compare the values of different features using pairwise $T$ tests ($\delta = 0.25$) within each structural class. The feature value ranks are shown in column-wise in **Figure 17**. In this figure, the colors indicate the magnitudes of the ranks indicated by the color bar. We find the features of Glu, Leu, the compositions of Ala, Asp, Val, and physical properties such as hydrophobicity, surrounding hydrophobicity for $\beta$-structure properties, achieve higher ranks than other features, whereas Cys, His, Met, Trp, and double-bend preference, amino acid composition properties show lower ranks. The values of the physical property features are determined by the amino acid compositions. Large compositions of amino acids with high values in this property may lead to a high feature value of this property. The value ranks for the twenty amino acids in the ten physical properties are presented in Supplementary **Table S4**. The Cys, His, Met, and Trp attain high values in amino acid composition properties, while His, Met, and Lys obtain high values in double-bend preference properties. However, Ala, Val, and Leu have low values for amino acid composition, and Ala, Asp, and Leu also show low values for double-bend preference. Therefore, the small compositions of Cys, His, Met, and Trp and the large compositions of Ala, Asp, Val, and Leu together lead to the low values for these two properties. Similarly, the large compositions of Glu and Asp may lead to high values of hydrophobicity, and the large compositions of Leu, Asp, Val and low composition of Trp together contribute to the high values of the surrounding hydrophobicity for $\beta$-structures.

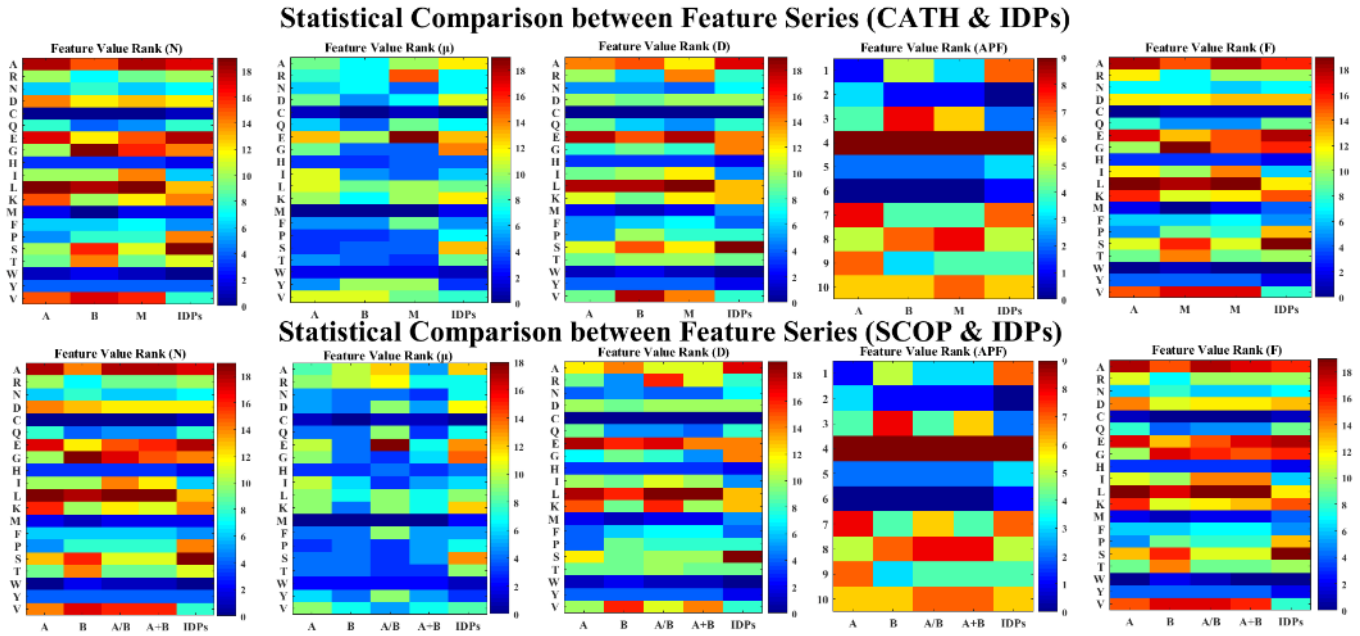


**Figure 17.** Statistical comparison between feature series.

**Figure 17** shows the feature value ranks for the different structural classes of CATH (top panel) and SCOP (bottom panel). These ranks are obtained by comparing different features using pairwise $T$ tests ($\delta = 0.25$). The colors indicate the magnitudes of the ranks as presented in the color bar. In the color bars, the colder the colors indicate the lower the ranking values, while the warmer the colors interpret the higher the ranking values. The ranking values range from 0 to 19 (for the twenty types of amino acids), the dark blue color indicates zero ranking value, while the dark red color indicates the highest-ranking value. The labels A, B, and M separately denote the $\alpha$, $\beta$ and mixed structural classes.

The compositions of Ala, Glu, and Leu show higher ranks in $\alpha$ than in $\beta$ structures. The $\alpha$ structures attain large compositions for Ala, Asp, Glu, Leu, Lys, Val, and high value of arrangement features for Glu, Leu, as well as high property values for hydrophobicity, flat extended preference, pK-*C* values, but low values for $\alpha$-helix and bend preference property. Since Ala, Glu, and Leu attain low values in $\alpha$-helix and bend preference, while Leu attains certain high values in flat extended preference and pK-*C*, the large compositions of Ala, Glu, and Leu together contribute to the low values of α-helix and bend preference and high values of flat extended preference and pK-*C*.

The compositions of Gly, Ser, Thr, and Val and the arrangements of Val and Ser, as well as the extended structural preference property, show higher ranks in $\beta$ structures than in other structures. The mainly β class also attains high values for $\alpha$-helix and bend preference, extended structural preference, and hydrophobicity. The mixed structures attain high rankings for the features of Glu, the compositions of Ala, Gly, Ile, and Val, the arrangements of Arg, and the extended structural preference, hydrophobicity, occurrence in the $\alpha$ region, and surrounding hydrophobicity for $\beta$-structure properties.

The IDPs show high rankings for the composition and arrangement of Ala, Glu,

Gly, and Ser and for the physical properties: $\alpha$ -helix and bend preference, hydrophobicity, and flat extended preference. These types of structures not only show similarity with the other structural classes in terms of the high rankings for the composition of Glu and Gly and hydrophobicity, but present special characteristics in terms of the high rankings for the arrangements of Ala and Ser and the α-helix and bend preference property.

### 3.2.2. Comparison between structural classes

The comparison results for each type of feature across different structural classes are shown in **Figure 18** and **Figure 19**. In these figures, the ranks are plotted row-wise for each structural class, along with plots for the sample mean and standard deviations. The colors indicate the values of the ranks as shown in the color bar. The sequential differences can be observed by comparing the rows of ranks. We can see apparent differences between the different structural classes. The mixed structural classes show overall higher ranks for all amino acid arrangements than other structural classes, which is followed by $\beta$ and $\alpha$ structures, and the IDPs structures show the lowest ranks for the arrangement features. The mixed structures, containing both $\alpha$, $\beta$ and loop structures, may be more complicated and require more amino acid arrangements to encode the structures. The $\beta$ structures rank the second highest in the arrangement features, which may imply that the bend and junction of parallel and anti-parallel $\beta$ sheets may need more comprehensive amino acid arrangements than α structures. The low ranking for the arrangement features of IDPs may be caused by the repetition of certain types of amino acid, as found in the Kmer analysis. The red bars for Ala, and Leu, hydrophobicity in the sample mean plots indicate larger values of these features. The dark blue bars for Cys, His, Met, and Trp indicate small values of these features.

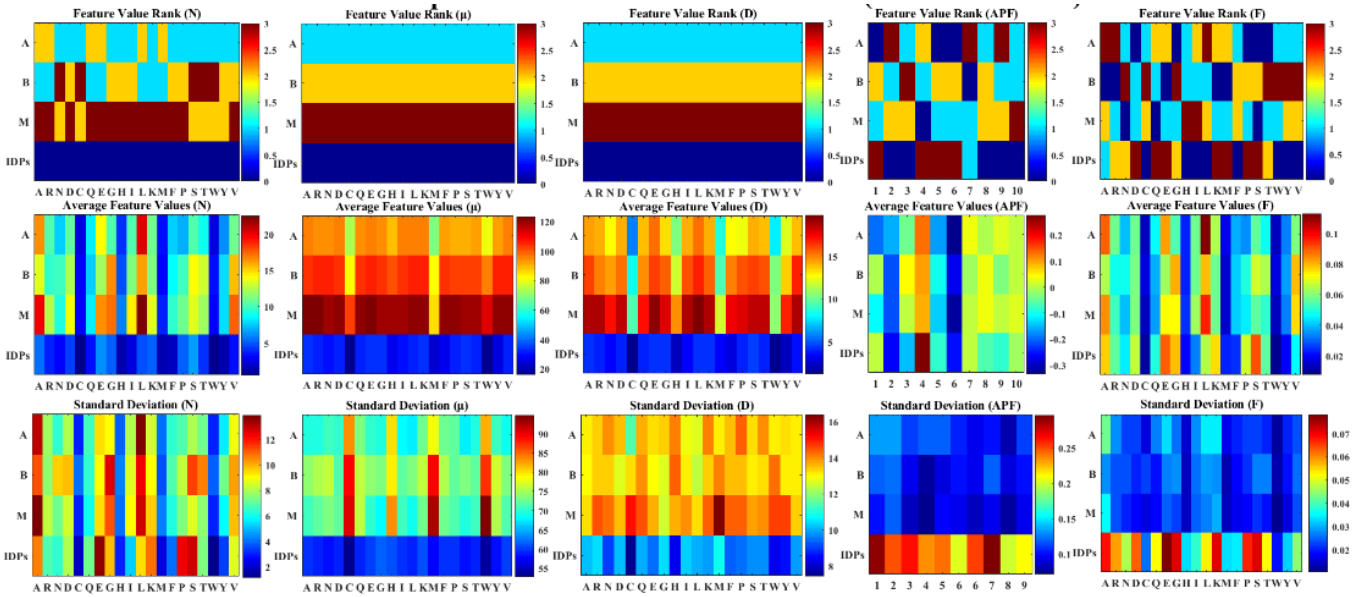


**Figure 18.** Feature comparison between structural classes (CATH and IDPs).

**Figure 18** shows the feature value ranks (top panel), sample means (middle panel) and standard deviations (bottom panel) for CATH and IDPs. The feature ranks are obtained by comparing the same feature between different structural classes using pairwise $T$ tests ($\delta = 0.25$). The sample means and standard deviations are the

common sense mean and standard deviations for the feature series. The colors of the lattices indicate the values of the ranks as presented in the color bar. In the color bars, the colder the colors indicate the lower the ranking values, while the warmer the colors interpret the higher the ranking values. In each subplot, the dark blue color means the lowest ranking value, while the dark red color means the highest-ranking value. The row label 'Mixed" stands for the mixed $\alpha$ and $\beta$ class.

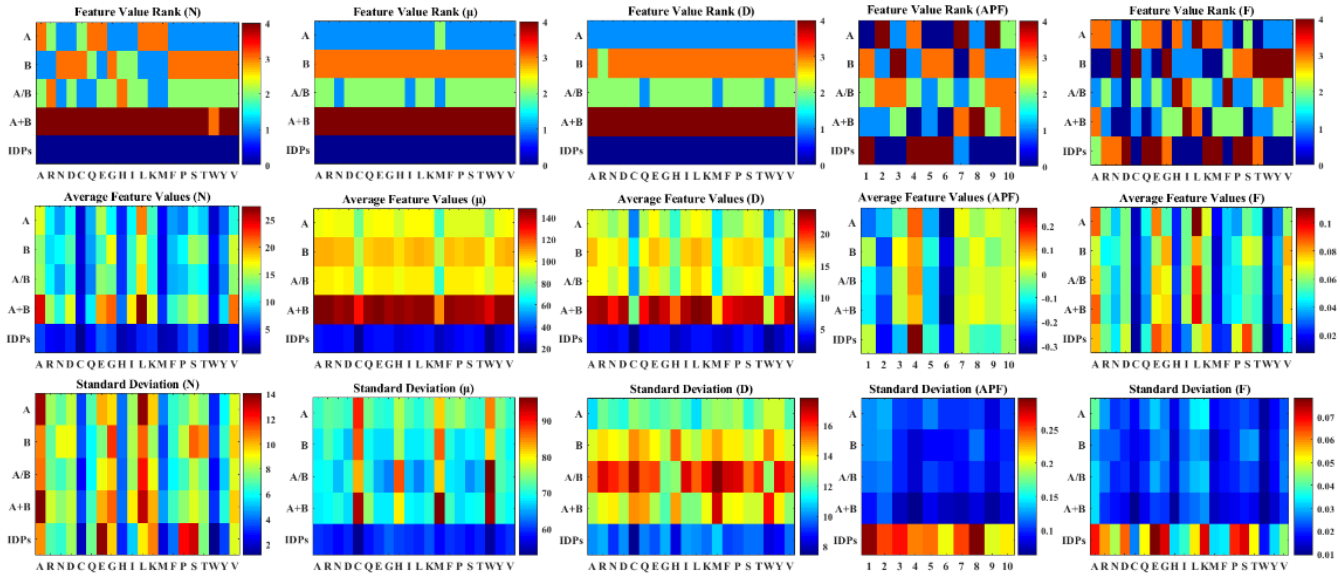


**Figure 19.** Feature comparison between structural classes (SCOP and IDPs).

**Figure 19** shows the feature value ranks (top panel), sample means (middle panel) and standard deviations (bottom panel) for SCOP and IDPs. The feature ranks are obtained by comparing the same feature between different structural classes using pairwise $T$ tests ($\delta = 0.25$). The sample means and standard deviations are the common sense mean and standard deviations for the feature series. The colors of the lattices indicate the values of the ranks, as shown in the color bar. In the color bars, the colder the colors indicate the lower the ranking values, while the warmer the colors interpret the higher the ranking values. In each subplot, the dark blue color means the lowest ranking value, while the dark red color means the highest-ranking value. The row labels A, B denote the $\alpha$ and $\beta$ classes, respectively.

We can see sharp discrepancy between the $\alpha$ and $\beta$ structures. In the $\alpha$ structures, there sequences contain large compositions for Ala, Arg, Gln, Glu, Leu, and Met and high values for side-chain size, flat extended preference, and pK-$C$ values, while the $\beta$ structures show large compositions for Asn, Cys, Gly, Ser, Thr, Trp, Tyr, and Val and high values for extended structural preference properties. The large compositions of Glu, Lys, and Gln together lead to the high values of side-chain size and flat extended preference in α structure properties, while the large compositions of Cys, Ser, Thr, and Tyr result in the high values of extended structural preference. The mixed structures show high values for surrounding hydrophobicity for $\beta$-structures, and large compositions of Ala, Arg, His, Ile, Leu, etc. The IDPs show large compositions of Asp, Gln, Glu, Lys, Met, Pro, and Ser and high values for the α-helix and bend preference properties.

### 3.2.3. Feature comparison for structural motifs

We compute the feature value rankings for each type of the motifs, which are presented in **Figure 20**. In this figure, the composition of Gly, Ile, Leu, and Val attain general higher rankings than that of other amino acids. In addition, the extended structure preference, occurrence in the α region and surrounding hydrophobicity attain higher ranking than those of the other properties, whereas the side-chain size attains low ranking in nearly all structural motifs. Except for the general high rankings of Ile, Leu and Val, the α motifs also show high rankings for Ala, Lys, while the $\beta$ motifs show high rankings for Gly, the $\beta$ harping rings show high ranking for Cys, while the motifs of loop structures show high rankings for Cys, Gly, Ser, and the $\alpha\beta\beta$ motifs also present high ranking for Cys.

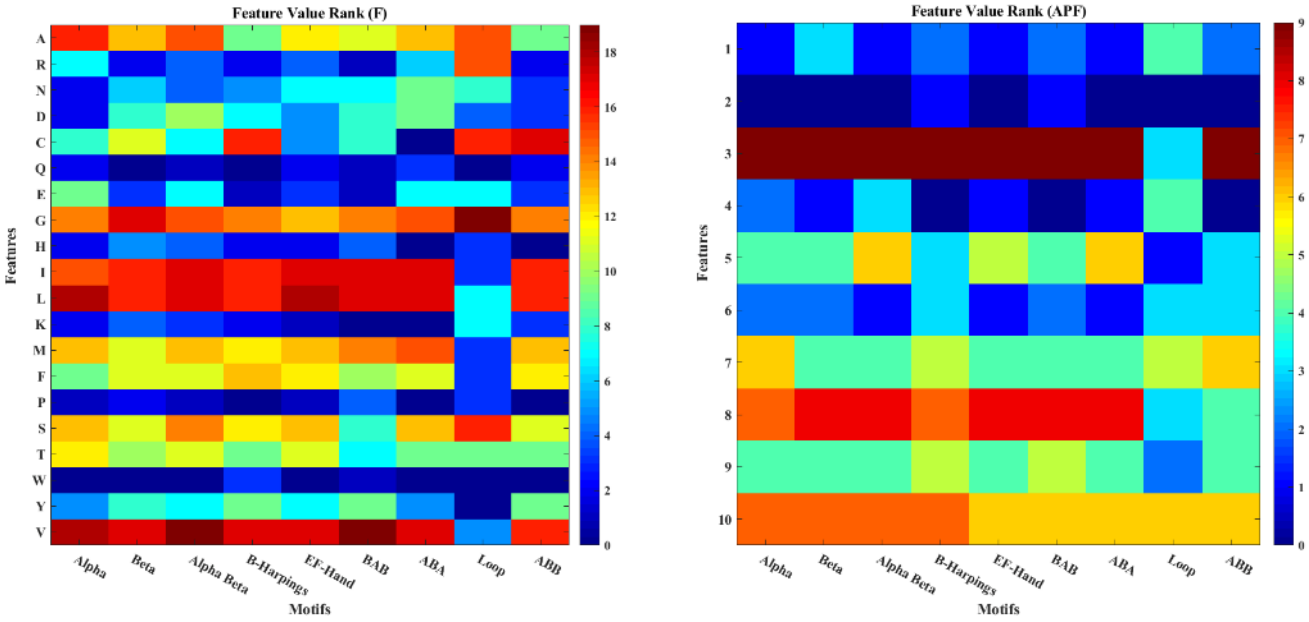


**Figure 20.** Statistical comparison between different feature series of the structural motifs.

**Figure 20** shows the feature value ranks for the different structural motifs. The ranks are obtained by comparing the same feature between different structures using pairwise $T$ tests ($\delta = 0.25$). The colors indicate the magnitudes of the ranks as indicated in the color bar. In the color bars, the colder the colors indicate the lower the ranking values, while the warmer the colors interpret the higher the ranking values. In each subplot, the dark blue color means the lowest ranking value, while the dark red color means the highest-ranking value.

The same features among different structural motifs are compared, and the results are shown in **Figure 21**. The α motifs attain a high ranking for Ala, Glu, and Thr, while the $\beta$ motifs attain a high ranking for His. The $\beta$ harping rings present high ranking for Phe and Trp, the $\beta\alpha\beta$ motifs show a high ranking for Trp and Val, and the $\alpha\beta\alpha$ motifs attain high ranking for Ile, Leu, and Met. The loop structures attain high ranking for Arg, Asn, Glu, Gly, Lys, Pro, Ser and physical properties such as $\alpha$-helix and bend preference, hydrophobicity, amino acid composition, flat extended preference, and surrounding hydrophobicity. In the second panel of **Figure 19**, the sample mean plots show generally high values for the composition of Ile, Leu, Met,

and Val and for physical properties such as extended structure preference, occurrence in the $\alpha$ region, and surrounding hydrophobicity.

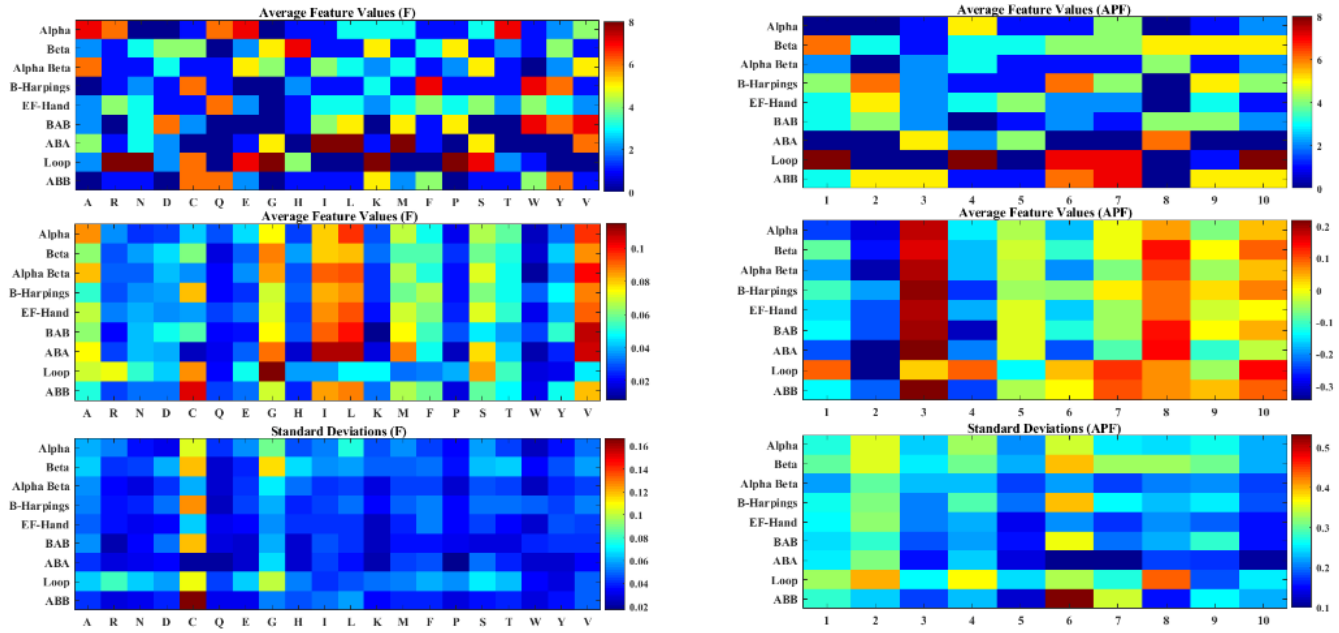


**Figure 21.** Feature comparison between the different structural motifs.

**Figure 21** shows the feature value ranks (top panel), sample mean (middle panel) and standard deviations (bottom panel) for the different structural motifs. The feature ranks are obtained by comparing the same features between the different structural motifs using pairwise $T$ tests ($\delta = 0.25$). The sample mean and standard deviations are the common sense means and standard deviations for the feature series. The colors indicate the values of the ranks as presented in the color bar. In the color bars, the colder the colors indicate the lower the ranking values, while the warmer the colors interpret the higher the ranking values. In each subplot, the dark blue color means the lowest ranking value, while the dark red color means the highest-ranking value.

### 3.3. K-mer analysis

K-mers are the strings of consequent amino acids in protein sequences, which are basic units that reflect the local arrangement of amino acids in given protein sequences. We count the statistics of K-mers for the CATH, SCOP and IDPs datasets with $K$ = 3 and 5. The most frequent 3-mers for the different structural classes are listed in descending order of their occurrence as presented in **Tables 2** and **3**. In the CATH data, the mainly $\alpha$ class admits 236,405 different 5-mers and 7954 different 3-mers, majority of these 5-mers appear only once, in which the most frequent 5-mer appears 18 times in the mainly $\alpha$ class, and the most frequent 3-mer appears 389 times. Among these K-mers, the most frequently appeared amino acids in $\alpha$-helix structures are such as Ala, Leu, Glu, Asp, Val, Tyr.

**Table 2** lists the most frequent 3-mers for the different structural classes in descending order of their appearance.

**Table 2.** The most frequent 3-mers for the different structural types.

| Classes | The most frequent 3-mers in descending order of their appearance |
|---|---|
| Mainly $\alpha$ | ALA, ELL, LLE, AAA, LAA, AAL, LLA, EAL, ALL, EEL, LLL, LAE, LAL, ELA, LEE, LLD, LLK, LAR, LEA, LLS, KLL, AEL, AEA, LKE, LLR, ALE, LEL, KEL, DLL. |
| Mainly $\beta$ | DGT, DGK, LTV, GTV, SGS, TVT, VTL, DGS, GSG, VTV, SGG, GGV, GKV, LTL, VSG |
| Mixed $\alpha$ and $\beta$ | ALA, LAA, AAA, AAL, EAL, LLA, ALL, LLE, LAE, ELL, AAG, AGL, EEL, LAL, LKE, AVA, ALE, LAG, ELA, VAA, LEE |
| All-$\alpha$ | LLE, ALA, AAA, EAL, EEL, AAL, ALL, ELL, LAE, LEE, LLA, LLD, LAA, ELA, LLL, LLK, LEA, LKE, KEL, LAL, KAL, AEL, LEK, EAA, LEL |
| All-$\beta$ | GSG, SGS, DGS, VTV, DGT, TVT, GTV, VTL, LTL, DGK, SGT, TLT, GSV, TGT, VGG, VKV, SGG, PDG, AGT, GGV, GKL, GVL, LLA, LTV, GSS, ADG, ATG |
| $\alpha/\beta$ | ALA, AAA, LAA, AAL, EAL, LLA, LLE, LAE, ALL, LAL, LAG, EEL, ALE, ALK, AVA, LKE, ELL, ELA, ALR, AGL, VLA |
| $\alpha + \beta$ | ALA, ELL, LAA, EAL, LLE, LLA, ALL, LGL, AAL, LKE, LAL, LAE, EEL, LEE, LLG, EKL, ALG, LLK, LEK, LLS, LEA, AAA, DLL, KAL, AEL, EVL, LLL |
| IDPs | EEE, SSS, AAA, GGG, PPP, EED, DDD, SPS, KKK, DEE, EDE, EEK, SDS, APA, SGS, EKK, TSP, GSS, PAA, DED, AEE, SSG, QQQ, KEK, PSS, SPT, PAP, PTS, EDD, EQE, AAP, ELE, LEE, DDE, EEA, KEE, GGS, GSG, SSE, SAS, SGG, SEE, STS, EEL, PSP, SSP, DSD, EAE, PEE, AAS, KKE, EAA, ASS |

The mainly $\beta$ class contains 286,063 different 5-mers and 7986 different 3-mers. The most frequent 5-mer appears 15 times, while the most frequent 3-mer appears 18 times. The frequent appearance of Gly, Val, Leu, Asp, Thr, Lys in K-mers is a special characteristic for the $\beta$ structures.

The mixed $\alpha$ and $\beta$ class includes 722,430 different 5-mers and 7997 different 3-mers. The most frequent 5-mer appears 40 times, while the most frequent 3-mer appears 1183 times. The frequent appearance of Ala, Val, Leu, Gly, Glu, Lys is a critical characteristic for the mixed α and $\beta$ structures.

**Table 3** lists the most frequent 5-mers for the different structural classes in descending order of their appearance.

**Table 3.** The most frequent 5-mers for the different structural types.

| Classes | The most frequent 5-mers in descending order of their appearance |
|---|---|
| Mainly $\alpha$ | ALELD, LELDP, VKLLL, NLGNA, AAALA, DEAIE, LEAGA, LGNAY, LLEAG, TPLHL, YDEAI, QKALE, EYYQK, IEYYQ, GNAYY, LHLAA, LLLEA, KALEL, PLHLA, VVKLL |
| Mainly $\beta$ | GDSGG, DSGGP, PDGTV, STDGG, GDVVL, SDPYV, SGGPL, RIAQL, RSGLA, RSTDG, DAGIY, DGALT, DGSSS, DPYVK, GGVPI, NDGKL, KLGLT, KLKLG, FTYTG, SPDGK, TLTVT |
| Mixed $\alpha$ and $\beta$ | AAAAA, AALAA, AAVAA, EELKK, NLYFQ, ALAAL, GSGKS, ALAAA, ENLYF, ALLAA, ALLEA, GSGKT, LAAAA, AARAL, LAALA, VAALA |
| All-$\alpha$ | AALAA, KKLEE, AAALA, EELEE, LAALL, LLDEL, LKALG, LTEKG, KRLEA, SEDYG, VSELL, AALLE, ARRLL, AERLL, AILAT, ALAAA, REALE, REAVE, RLEEA, RVMES, QLTEE, EAALK, EAELA, EALLA, EELLR, EEVKK, ELEEL, ELLAR, EVEAL, IRRLE, LEKAL, LLERL, LTEEE, LVKVL, KELGT, KLIEK, KKLLE, MILNS |

**Table 3.** (*Continued*).

| Classes | The most frequent 5-mers in descending order of their appearance |
|---|---|
| All-$\beta$ | DGKLK, GDSGG, STDGG, WYVDG, GTYRC, NDGKL, KLGLT, FFTYT, FTYTG, TGIVS, YVDGV, ADLSG, ANPLY, ATANI, RKDHS, DGRVF, DGKRP, DGSSS, DLRGA, DSGGP, DWVKY, QDGKR, EYDWV, HTATA, IGDVV, IVSSF, GANPL, GGALV, GGSGG, GKLKL, GFDAS, GSYNG, GTVAS, GYSNG, NQDGK, NPLYA, NTGIV, LDTGA, LETGA, LKKGD, LTSSA, KLKLG, PAHGT, SGTAL, SNGGV, SNGNL, SVNLL, TIRVT, TNKYT, TLVGH, TPGKI, YTGPA, VDAAF, VDGVL, VKYTS, VTLTC |
| $\alpha/\beta$ | AAAAA, GSGKS, ALAAL, GKTTL, GSGKT, SGKST, EELKK, TGSGK, AALAA, ALEGA, LAAAA, LLAEA, AAAAL, AAEAL, ADVVL, AGLLA, ALREA, EELRK, ELAKR, GVDVV, LEALG, VIGGG |
| $\alpha + \beta$ | DFGLA, AVLRA, RDLKP, EALEA, EEIKK, AAEKL, AKEIA, DILKE, DLIKK, EELLK, EKEKE, EKEFL, GLRRL, LAELL, LKEKL, KALGL, KEKLL, FAEAF, TPDGR |
| IDPs | EEEEE, SPTSP, YSPTS, GGGGG, PSYSP, SPSYS, SDSDS, PTSPS, SYSPT, TSPSY, DDDDD, AAAAA, DSDSD, PPPPP, SSSSS, QQQQQ, EEQEQ, QELEE, EQEQE, EQELE, DEEEE, EEDEE, EDEEE, EEDDE, QEQEL, EEEED, ELEEQ, LEEQE, RDRDR, EDDEE, EEEDE, DDEED, DEEDD, APAPA, EEEEG, EDDDD,'DDDDE, DEEED, KKKKK, PAPAP, AFSFG, DEDDD, DEEDE, EDEDE, EDEED, SSGSS, AAAKA, RSRSR, DRDRD, QQDEQ, QEGQL, EKLPG, GGGWG, PAFSF, TPTP |

In the SCOP data, the all $\alpha$ class includes 149,854 different 5-mers and 7884 different 3-mers. The most frequent 5-mer appears 9 times, while the most frequent 3-mer appears 213 times. The frequent appearance of Ala, Leu, Glu, Arg, Lys is a key characteristic for the all $\alpha$ class.

The all $\beta$ class includes 331,815 different 5-mers and 7924 different 3-mers. The most frequent 5-mer appears 9 times, while the most frequent 3-mer appears 136 times. The frequent appearance of Gly, Phe, Tyr, Val, Leu, Thr is a key characteristic for the all $\beta$ class.

The $\alpha/\beta$ class includes 236,405 different 5-mers and 7970 different 3-mers. The most frequent 5-mers appears 23 times, while the most frequent 3-mer appears 458 times. The frequent appearance of Ala, Leu, Glu, Gly, Ser, Lys is a main characteristic for the $\alpha/\beta$ class.

The $\alpha + \beta$ class includes 215,314 different 5-mers and 7942 different 3-mers. The most frequent 5-mer appears 8 times, while the most frequent 3-mer appears 223 times. The frequent appearance of Ala, Leu, Glu, Asp, Arg, Lys is a key characteristic for the $\alpha + \beta$ class.

The IDPs include 184,749 different 5-mers and 7785 different 3-mers. The most frequent 5-mer appears 137 times, while the most frequent 3-mer appears 801 times. The large repetition of Glu, Ser, Pro, Gly, Gln, Ala, Asp is a key characteristic for the IDPs, which may be one of the reasons for the structural disorder of these proteins.

For the different types of structural motifs in PROSITE, the K-mers results are summarized in Supplementary Dataset S8. The α type motifs contain frequent Kmer patterns involve Leu, Ile, Val, and Met, and the $\beta$ type motifs admit frequent patterns, including Gly, Cys, Phe, and Tyr. The $\alpha$ and $\beta$ type motifs get frequent patterns include Val, His, Tyr, Ile, Phe, Ser, Glu, Ala, and Leu, while the $\beta$ harping rings contain frequent patterns involve Cys, Arg, Met, Leu, Ile, and Val. The EF-hands show frequent patterns involve Leu, Phe, Tyr, His, Ile, and Val. The βαβ motifs show

frequent patterns involve Val, Leu, Ala, and Ile. The αβα motifs attain frequent patterns involve Leu, Met, and Val. The motifs of the loop structures admit frequent patterns involve Ala, Asp, Met, Cys, Asn, and Gly. The αββ motifs contain frequent patterns involve Cys, His, Ile, Leu, and Gly.

Since K-mers represent local units of the sequence, it is a more delicate units of the protein sequence, which decides the local structures of the proteins and acts a crucial role in the sequential analysis of protein structures. The high frequency K-mers patterns identified in our analysis reveal the sequential nature for the different protein structures. From the above analysis, we can see that the different types of structures present special K-mer combinations and amino acid compositions. Particularly for the intrinsically structural disordered proteins, the highly repeated occurrence for a same of kind of amino acid may lead to the structural disorders of these proteins. The statistical analysis of K-mers reveals critical impact of such local amino combination and arrangement patterns on the formation of the different types of protein structures. The K-mers results can be implemented to develop new features or classifiers to improve the efficiency in future protein structural classifications. Since K-mers characterize the local situation of proteins, and the 3-mers and 5-mers are already enough to characterize such local structures, the amino acid combination preferences in high frequency K-mers can also be employed for future protein molecular design especially protein-based drug or vaccine developments.

## 4. Discussion

Traditional protein structural studies take uses of the sequence homology to develop new sequence features or classifiers for the structural classifications or predictions [1–18]. However, all these studies ignore the rich information hidden behind the amino acid combinations and the feature interactions. In complexity and network analysis, the behaviour of a system can usually be modeled by the interactions o its components [39]. Eng lightened by this complex network analysis; the behaviour of how amino acid sequence encode their structures can be modeled by the interactions between the various protein sequence features. To make up the deficiency of previous studies and also uncover the dynamic nature between the interaction features, we use network and statistical tools [19] to identify the sequential differences between the different types of structures at both structural class level and motifs level. In this research, we extract the standard protein sequence features, include the amino acid composition, arrangement [3,24] and their physical properties [20], and model the sequential influences to the structures using networks. We implement centrality metrics [39] and statistical tests [40,41] to identify the sequential discrepancy between the different types of structures, where interesting results are found in respect of amino acid feature interactions and feature value distributions. The analysis covers not only macro levels of the top structural classes of CATH, SCOP and IDPs (in the DisProt database), but also micro levels of the structural motifs (in the PROSITE database) and K-mers, where both common and special characteristics are identified for the different protein structural types.

For the sequential similarities between different protein structures, we find that the all-structural classes show strong connections between Asp, Leu, and Val with

other amino acids, but weak intra-type connections for Cys, His, Trp, and Met. These results are robust and are in consistent with early findings in single type feature analysis [19], which implicate that the Asp, Leu, and Val are very important and are actively involved in coding of the different types of structures. When taking the cross-type features into account, the all-structural classes also attain significant interactions between the compositions of Gly and Ala, between the composition of Ala and side-chain size, between side-chain size and hydrophobicity properties, and between $\alpha$-helix and bend preference and amino acid composition properties. These significant interactions imply that the features of Gly and Ala have intensive coupling in forming the sequences and structures, while the Ala is largely related with the side-chain size of the peptides, and these significant feature connections are general sequential characters for all types of structures. These outcomes are consistent with early findings of amino acid physical properties [31]. For instance, the occurrence of significant interaction between the side-chain size and hydrophobicity properties are sensible, because early statistical findings convince that the bulkiness of side chains (containing polar groups) may have a significant effect on the hydrophobicity property [31], these perhaps lead to the significant connections between these two properties. The significant connection of other feature pairs may also due to the polarity and the charged groups of the amino acid residues [2]. These common characters for all structural types may be caused by the integration of the comprehensive interactions between the amino acid composition, arrangement and physical properties.

In the statistical analysis of feature series, pairwise Welch *T* tests are used for feature comparisons. The Welch T test is free from the homogeneity of the data variance [40]; hence it is an ideal test in real-world data analysis, particularly for biological data analysis. In our analysis, the all-structural classes admit large compositions of Glu, Leu, Ala, Asp, and Val and high values for hydrophobicity property, but small compositions of Cys, His, Met, and Trp. These implicate that the Glu, Leu, Ala, Asp, and Val are intensively involved in the formation and coding of the protein structures, while Cys, His, Met, and Trp are less involved in the structural coding.

When focusing on specific types of structures, the $\alpha$ and $\beta$ structures exhibit both similarities and differences in their feature interactions. The α structures show special preference and high importance (intensive feature interactions) for Glu, along with significant connections between the arrangements of Pro and hydrophobicity, and between the compositions of Arg and Lys. The α structures are largely composed of Ala, Arg, Gln, Glu, Leu, and Met in their sequence, which also attain high property values in the side-chain size and pK-C properties. These imply that the negatively charged Glu [2] are important in encoding the α structures, this outcome agrees with early findings in single type feature analysis [19]. We also find the hydrophobicity property attains intensive relations with the positively charged Arg and Lys [2], this may suggest that the positive charge of these amino acids may affect the exposure of the residues to the protein molecule surface [31] in $\alpha$ structures.

The $\beta$ structures show preferences for Gly and high importance for Thr, Phe, Tyr and pK-*C* values, extended structural preference, and surrounding hydrophobicity for $\beta$ structures. The $\beta$ structures exhibit larger arrangement features than those of $\alpha$

structures, large compositions of Asn, Cys, Gly, Ser, Thr, Trp, Tyr, and Val, and high values for the extended structural preference property. These imply that the uncharged polar amino acid Gly [2] is important in encoding of the $\beta$ type structures, and other uncharged polar amino acids such as Thr, Tyr Asn, Cys, Ser [2] along with physical properties such as extended structural preference [31], pK-*C* values [36] and surrounding hydrophobicity for $\beta$ structures [37] are also important in coding the $\beta$ structures.

The $\alpha$ and $\beta$ structures also present similarities in terms of feature interactions. For instance, both the $\alpha$ and $\beta$ structures present strong connections between the arrangements of Ser and other amino acids, between the compositions of Cys and Arg, and between hydrophobicity and amino acid arrangements. These imply that the amino acids arrangement property has great influence to hydrophobicity of amino acids [31], and the uncharged polar amino acid Ser attains strong connections with other amino acids in terms of their sequence arrangements in both the $\alpha$ and $\beta$ types of structures.

The mixed structures exhibit larger composition and arrangement features than those of the other structures. The mixed structures include large compositions of Ala, Gly, Arg, His, Ile, Leu, Phe, and high values for hydrophobicity, occurrence in $\alpha$ region, surrounding hydrophobicity for $\beta$-structures properties, and high importance for Ala, Asp, Val, Leu, Glu, Gly, and side-chain size, pK-*C* value properties. The mixed structures not only contain similarities with both $\alpha$ and $\beta$ structures but also special characters in terms of the significant connections between Met, Arg, Lys and double-bend preference, between the arrangements of Cys, His, Met, Tyr and amino acid compositions, and between the arrangements of Gly and Glu, and between the arrangements of Ala, Gly, Glu, Ile with other amino acids. These imply that except for the common characters, the non-polar amino acids Ala, Ile and Phe, the negatively charged amino acid Asp, the positively charged amino acids Arg and His [2], along with the side-chain size [31] and pK-*C* value [36] properties are crucial in forming the mixed type of structures. Moreover, the features of the non-polar amino acid Met and the positively charged amino acids Arg, Lys admit strong relations with the double-bend structures [32], and the Cys, His, Met, Tyr are strong related with the general amino acid compositions [31] property.

In the intrinsically structural disordered proteins (IDPs), there found intensive interactions for the compositions of Glu, Ser, Gln, Leu, Val, Ile, and the arrangements of Ala, Arg, Asp, Gln, Glu, Gly, Ile, Leu, Val, and between the side-chain and pK-*C* properties. The IDPs show high importance for Thr, Asn, Gly, Pro, and amino acid composition, occurrence in the α region, and pK-*C* value properties, which contain large similarity with the mixed structures. The sequences of IDPs contain large compositions of Ala, Glu, Gly, and Ser, which also admit high values for the α-helix and bend preference, hydrophobicity, and flat extended preference properties. These further convince that the above similar characters identified in both the intrinsically disordered structures and the mixed structures, such as Ile, the side-chain and pK-*C* properties, may influence the coding of the loops or junction structures, which are less ordering in spatial structures.

When specifying the local sequential characters in K-mers, special characters are found in terms of the K-mer combinations and high frequency K-mers. According to

the data statistics, longer K-mers may receive low frequency of appearances, early studies on K-mers found that the 5-mers are long enough to characterize the local protein structures [42], therefore we choose the analyze the 3-mers and 5-mers to guarantee sufficient statistics of the data. Analysis shows that the different types of structures exhibit different high frequency K-mers. In α structures, the high frequency K-mers contain large combinations of Ala, Leu, Glu, Asp, Val, Tyr, Arg, and Lys, whereas the Gly, Thr, Val, Leu, Asp, Lys, Phe, and Tyr are frequently appeared in the high frequency K-mers in $\beta$ structures. The mixed structures possess large combinations of Ala, Val, Leu, Gly, Glu, Lys, Ser, Asp, Arg in their high frequency K-mers. The IDPs show large amount of repetition strings for Glu, Ser, Pro, Gly, Gln, Ala, and Asp, from which we can suggests that if the regular structures are encoded by special K-mers combinations containing different kinds of amino acid, then the highly repetition pattern for a same type of amino acids may break this code and hence lead to the intrinsically disorder of the structures. The K-mers analysis convince the critical impact of special amino combination in K-mers on the formation of different types of protein structures. The K-mers results can be further implemented in new feature development for protein structural classifications.

For the structural motifs analysis, the results suggest that the Cys, Gly, Ile, Leu, and Val are important in nearly all typical types of structural motifs, and the compositions of Gly, Ile, Leu, and Val are comparatively large than other amino acids. The α type motifs show high importance (intensive feature interactions) for Arg and Met, large compositions of Ala, Glu, Thr, and Lys, and high frequency K-mers involving Leu, Ile, Val, and Met. The $\beta$ motifs show high importance for Cys and Gly, a large composition of Gly and His, along with high frequency K-mers containing Gly, Cys, Phe, Tyr, and His. The α and β motifs contain high frequency K-mers involving Val, His, Tyr, Ile, Phe, Ser, Glu, Ala, and Leu, which are analogous as found in the mixed $\alpha$ and $\beta$ structures. The $\beta$ harping rings show high importance for Cys, and large compositions of Cys, Phe, and Trp, as well as high frequency K-mers involving Cys, Arg, Met, Leu, Ile, and Val. The EF-hands show high importance for Met, Phe, and high frequency K-mers involving Leu, Phe, Tyr, His, Ile, and Val. The $\beta\alpha\beta$ motifs show high importance for Cys, large compositions of Trp, and Val, and high frequency K-mers containing Val, Leu, Ala, and Ile. The αβα motifs show high importance for Ala and Ser, large compositions of Ile, Leu, Met, along with high frequency K-mers involving Leu, Met, and Val. The αββ motifs show both high importance and large compositions of Cys and high frequency K-mers involving Cys, His, Ile, Leu, and Gly. The loop structures show high importance for Al, Cys, and large compositions of Arg, Asn, Glu, Cys, Gly, Lys, Pro, Ser, along with high frequency K-mers involving Ala, Asp, Met, Cys, Asn, Gly.

In this paper, we find special amino acid feature interactions and feature value distributions that characterize the different types of protein structures, where both common and special characteristics are found between the structures at both the class level and motifs level. The outcomes regarding the significant features interactions, feature value distributions and K-mer patterns help illuminate the dynamic nature between the various amino acid features. The results can be future used for developing new feature, or enhancing protein molecular design specially in protein-based drug

and vaccine development. Since the networks constructed from the mutual relations are undirected, and feature methods and network tools are standard, future research can be improved by employing causal measures e.g. deep learning causal frames to measure the directed influences between the features, and also use more advanced complex network tools such as information-based complexity measures e.g., group entropy to analyze the local and global structures of the networks, from which more evidences can be dug out from the system behaviour. Additionally, we can also delve further into deeper level structural categories such as folds and super families, and capture the geometrical features of the protein structures by performing simplicial complex modeling and persistent homology analysis, from which more interesting dynamics can be revealed from the sequence and structural feature interactions.

## 5. Conclusions

In this network and statistical analysis of protein sequence features, both common and special sequential characters are specified for the various types of structures. The significant interactions between the features of Ala and α-helix and bend preference property, between Ala and side-chain size, Ala and Gly, and between Met and Leu, as well as the weak intra-type interactions between features of Cys, His, Trp, and Met, are the common characters for all protein structural types, where the feature for Leu, Val, and Asn are acted as the critical sources of the feature interactions. For the $\alpha$ structures, it presents high importance for features of Glu, Pro and side-chain size, hydrophobicity properties, whereas the $\beta$ structures present high importance for the features of Gly, Thr and physical properties such as α-helix and bend preference, extended structural preference, pK-*C* value and surrounding hydrophobicity for $\beta$ structures. Except for these special preferences, the $\alpha$ and $\beta$ type structures also show common characters that Ser is served as the common sources of feature interactions. The mixed α and β structures not only show common characters with the $\alpha$ and $\beta$ structures, but also preferred interactions between Met, Lys and double-bend preference property, and between the sequence arrangements of Cys, His, Met, Tyr and amino acid composition features, which are suggested to have strong relations with loops and junction structures. Owning certain similarity with the mixed types of structures, the intrinsically disordered proteins (IDPs) also admit high repetitive patterns for certain kinds of amino acids in their local K-mer strings, which are suggested to the causation for the structural disorders. In the micro level, the different structural motifs not only show common characters in terms of the high importance for Cys, Gly, Ile, Leu, and Val, but also special characters. Further sequential differences can be discovered by K-mers and feature series analysis. From the network and statistical analysis, strong couplings are found between certain amino acids and physical properties. The outcomes of this study reveal the dynamic nature of amino acid feature interactions, which can be future used for protein molecular design or new feature development for protein structural classifications.

**Supplementary materials:** All data of this article is fully available in Supplementary Files. All CATH and SCOP sequence data in this analysis are downloaded from Protein Data Bank (PDB) database (https://www.rcsb.org) by using the PDB IDs in

the Supplementary Dataset S1. The IDPs and structural motifs data are respectively downloaded from DisProt (https://www.disprot.org/) and PROSITE (https://prosite.expasy.org/) database. The accession numbers of the IDPs are stored in Supplementary Dataset S2, and the accession numbers for pattern data of the structural motifs are stored in Supplementary Dataset S6. The feature matrices, adjacency matrices for unweighted networks, centrality results, and ranks obtained by statistical tests, as well as Kmer statistics are all provided in Supplementary Datasets S3–S5 and S7–S8.

**Author contributions:** Conceptualization, XW; methodology, XW; software, XW; validation, XW and XT; formal analysis, XW; investigation, XW and XT; resources, XW and XT; data curation, XW and XT; writing—original draft preparation, XW; writing—review and editing, XW and XT; visualization, XW and XT; supervision, XW; project administration, XW; funding acquisition, XW and XT. All authors have read and agreed to the published version of the manuscript.

**Acknowledgments:** Acknowledgement goes to the Beijing University of Chemical Technology for library resources.

**Ethical approval**: Not applicable.

**Conflict of interest:** The authors declare no conflict of interest.

## References

1. Levitt M. Nature of the protein universe. P. Natl. Acad. Sci. 2009; 106(27): 11079–11084.
2. Wang J, Wang, Z. & Tian, X. Bioinformatics: Fundementals and applications. Beijing: Tsinghua University Press; 2014 (In Chinese).
3. Yu, C, Deng M, Cheng SY, Yau SC, He RL, Yau ST. Protein space: A natural method for realizing the nature of protein universe. J. Theor. Biol. 2013; 318: 197–204.
4. Zhao B, He RL, Yau ST. A new distribution vector and its application in genome clustering. Mol. Phylogenet. Evol. 2011; 59: 438–443.
5. Zhao X, Wan X, He RL, Yau ST. A new method for studying the evolutionary origin of the SAR11 clade marine bacteria. Mol. Phylogenet. Evol. 2016; 98: 271–279.
6. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Zidek A, Nelson A, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Improved protein structure prediction using potentials from deep learning. Nature. 2020; 577(7792): 706-710.
7. Cramer P. AlphaFold2 and the future of structural biology. Nat. Struct. Mol. Biol. 2021; 28: 704–705.
8. Wu T, Guo Z, Cheng J. Atomic protein structure refinement using all-atom graph representations and SE(3)-equivariant graph transformer. Bioinformatics. 2023; 39(5): btad298.
9. Hong Y, Lee J, Ko J. A-Prot: protein structure modeling using MSA transformer. BMC Bioinformatics. 2022; 23: 93.
10. Pearce R, Li Y, Omenn GS, Zhan Y. Fast and accurate Ab Initio Protein structure prediction using deep learning potentials. PLoS Comput. Biol. 2022; 18(9): e1010539.
11. Rachitskii P, Kruglov I, Finkelstein AV, Oganov AR. Protein structure prediction using the evolutionary algorithm USPEX. Proteins. 2023; 91: 933–943.
12. Hou M, Peng C, Zhou, Zhang B, Zhang G. Multi contact-based folding method for de novo protein structure prediction. Brief. Bioinform. 2022; 23(1): bbab463.
13. Stapor K, Kotowski K, Smolarczyk T, Roterman I. Lightweight ProteinUnet2 network for protein secondary structure prediction: a step towards proper evaluation. BMC Bioinformatics. 2022; 23(1): 1-16.
14. Kim Y, Kim J. AttSec: protein secondary structure prediction by capturing local patterns from attention map. BMC Bioinformatics. 2023; 24(1): 183.

15. Zhang B, Liu D, Zhang Y, Shen H, Zhan G. Accurate flexible refinement for atomic-level protein structure using cryo-EM density maps and deep learning. Brief. Bioinform. 2022; 23(2): bbac026.
16. Gormez Yasin, Sabzekar M, Aydın Z. IGPRED: Combination of Convolutional Neural and Graph Convolutional Networks for Protein Secondary Structure Prediction. Proteins. 2022; 90(8): 1613.
17. Zhang B, Zhang X, Pearce R, Shen HB, Zhang Y. A New Protocol for Atomic Level Protein Structure Modeling and Refinement Using Low-to-Medium Resolution Cryo-EM Density Maps. J. Mol. Biol. 2020; 432: 5365-5377.
18. Liu B, Li CC, Yan K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Brief. Bioinform. 2019; 21 (5): 1733-1741.
19. Wan X, Tan X. A protein structural study based on the centrality analysis of protein sequence feature networks. PLoS ONE. 2021; 16(3): e0248861.
20. Rackovsky S. Sequence physical properties encode the global organization of protein structure space. P. Natl. Acad. Sci. 2009; 106(34): 14345–14348.
21. Duda RO. Pattern classification (second edition). New York: John Wiley & Sons, Inc; 2001.
22. Tian K, Xin Z, Yau S. Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. J. Theor. Biol. 2018; 456: 34–40.
23. Jeong JC, Lin X, Chen X. On position-specific scoring matrix for protein function prediction. IEEE/ACM Trans. Comput. Biol. Bioinform. 2011; 8(2): 308–315.
24. Shen H, Chou K. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 2008; 373: 386-388.
25. Liu B, Liu F, Wang X, Chen J, Fang L, Chou K. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015; W1: W65-W71.
26. Zhang Y, Wen J, Yau SS-T. Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. Genomics. 2019; 111: 1298–1305.
27. Yu C, He RL, Yau SS-T. Protein sequence comparison based on K-string dictionary. Gene. 2013; 529(2): 250-256.
28. Liu B, Wang S, Dong Q, Li S, Liu X. Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning. IEEE T. on Nanobiosci. 2016; 15(4): 328-334.
29. Wen J, Zhang Y, Yau SS-T. K-mer Sparse matrix model for genetic sequence and its applications in sequence comparison. J. Theor. Biol. 2014; 363: 145-150.
30. Mu Z, Yu T, Liu X, Zheng H, Wei L, Liu J. FEGS: a novel feature extraction model for protein sequences and its applications. BMC Bioinformatics. 2021; 22: 297.
31. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J. Protein Chem. 1985; 4(1): 23-54.
32. Isogai Y, Nemethy G, Rackovsky S, Leach SJ, Scheraga HA. Characterization of multiple bends in proteins. Biopolymers. 1980; 19: 1183-1210.
33. Jukes TH, Holmquist R, Moise H. Amino acid composition of proteins: Selection against the genetic code. Science. 1975; 189: 50-51.
34. Rackovsky S, Scheraga HA. Differential geometry and polymer confirmation. 4. Conformational and neucleation properties of individual amino acids. Macromolecules. 1982; 15: 1240-1346.
35. Maxfield FR, Scheraga HA. Status of empirical methods for the prediction of protein backbone topography. Biochemistry. 1976; 15: 5138-5153.
36. Fasman GD. Handbook of Biochemistry and Molecular Biology (3rd ed). Boca Raton: CRC Press; 1976.
37. Ponnuswamy P, Prabhakaran M, Manavalan P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim. Biophys. Acta. 1980; 623: 301-316.
38. Wan X, Tan X. A Simple protein evolutionary classification method based on the mutual relations between protein sequences. Curr. Bioinform. 2020; 15(10): 1113-1129.
39. Newman MEJ. Networks: An Introduction. New York: Oxford University Press; 2010.
40. Fang J. Statistical methods for biomedical research (2nd Edition). Beijing: Higher Education Press; 2019.
41. Joan FB. Guinness, gosset, fisher, and small samples. Stat. Sci. 1987; 2 (1), 45–52.
42. Morikawa N. Discrete differential geometry of n-simplices and protein structure analysis. Applied Mathematics. 2014; 5(16), 2458-2463.