**Energetic Profile-Based Protein Comparison: A Novel Approach for Discovering Evolutionary Relationships**

**Abstract**

Protein stability is a fundamental attribute enabling proteins to execute specific functions. Within this context, the energy profile of a protein emerges as a pivotal determinant of its stability, a profile that can be directly inferred from its sequence or structure. Consequently, augmenting protein annotations with energy-based features becomes essential, complementing the traditional focus on amino acid composition and structural information. We propose a knowledge-based approach to estimate distance-dependent interactions in proteins using known protein structures, resulting in a multidimensional energy vector comprising 210 pairwise energy terms that comprehensively represent a protein's conformational state. This energy vector not only characterizes individual protein structures but also offers a versatile means of quantifying structural dissimilarity between proteins. Notably, our methodology extends to protein sequences, enabling comparisons independent of protein length and obviating the need for alignment or superimposition, in contrast to traditional approaches requiring structural alignment—a challenging task, especially when comparing highly dissimilar structures. We harness these capabilities on diverse datasets to classify proteins at class, fold, family, and superfamily levels, explore functional similarities, and conduct phylogenetic analyses. Our results demonstrate the precision, speed, and efficacy of our method in comparison to established structure comparison techniques, particularly in unraveling evolutionary relationships between proteins. Notably, we successfully reconstruct the evolutionary lineage of spike glycoproteins across different species of coronaviruses, exemplifying the broad applicability and power of our approach.

**Keywords**

Energy-based annotation, Structural dissimilarity, Evolutionary relationships, Profile of energy, Knowledge-based potential.

**Introduction**

A thorough understanding of protein function is essential in the realms of biology, medicine, and pharmacy. However, traditional experimental approaches for deciphering the biological functions of proteins face considerable hurdles, including significant costs and time constraints. Proteins, as molecular workhorses, provide a wide range of essential functions, including structural support, contractility, transportation, enzyme catalysis, storage, hormone signaling, and defensive roles. The growth of high throughput methods in genomics has generated an extensive repository of protein sequences across multiple databases, a large amount of which is still unannotated [1]. The stratification of proteins into distinct folds/superfamily/families, predicated upon evolutionary consanguinity or shared structural and functional attributes, emerges as an indispensable strategy for precise function prediction. Such classification augments our holistic understanding of an organism's physiological landscape. While conventional protein function prediction strategies have relied on sequence similarity search tools like BLAST [2] and FASTA [3], as well as motif searches [4], contemporary approaches have explored an extensive array of methodologies, including sequence-based techniques [5], omics-data integration [6], phylogenetic profiling [7] and three-dimensional protein structures [8], all striving to enhance the precision and depth of our insights into protein functionality. In the realm of protein structure and function classification, databases such as CATH (Class, Architecture, Topology, Homologous superfamily) [9] provide a comprehensive framework, categorizing proteins into hierarchical groups based on their structural features, including classes, architectures, topologies, and homologous superfamilies. In comparison to SCOP (Structural Classification of Proteins) [10], CATH offers a distinct perspective by emphasizing the relationships between protein structures at various levels, facilitating a more nuanced understanding of structural diversity and evolution within the protein world. In the SCOP hierarchy, superfamilies sharing the same fold exhibit analogous overall secondary structures, orientation, and connectivity, thereby manifesting a common core structure. Folds characterized by congruent core structures are grouped into the same structural class. Despite extensive studies on protein evolution, encompassing protein sequence, secondary structure, and three-dimensional structural attributes, the intrinsic energy of protein structures, which fundamentally influences macromolecular and organismal evolution, remains an underexplored dimension within this domain.

One prominent approach involves the development and application of knowledge-based potential functions [11, 12], which leverage information extracted from known protein structures to estimate various energies, including distance-dependent interactions [13], dihedral angles, and accessible surface energies [14]. In this study, we employ the Sippl potential function [11] to extract distance-dependent knowledge-based potential function. Given the existence of 20 distinct amino acids that lead to a total of 210 unique pairwise interactions, we systematically encapsulated these interactions within a multidimensional 210-element vector. Within this vector, each element represents the cumulative energy contributions attributed to a specific pair of amino acids. To illustrate, the i'th element encompasses the combined energy arising from all interactions involving ALA-GLY residues within the protein structure. Consequently, any given protein structure could be succinctly summarized as a 210-dimensional vector, serving as a comprehensive representation of the intricate energy landscape intrinsic to the protein's structure. We refer to this vector of energies as the "energetic profile," and it serves as the cornerstone of our analytical approach, providing a robust foundation for further investigative pursuits. We consider the Manhattan distance between the energetic profiles of two proteins as a measure of dissimilarity between them.

Given the prevailing challenge of experimentally ascertaining the three-dimensional structures of many proteins, the evaluation of energy based on sequence emerges as a crucial consideration. Dostari et al. devised a method rooted in amino acid composition to estimate energy, thereby facilitating predictive insights based on sequence. Employing this methodology, we computed the energy profile based on sequence in our investigation. In this study, we initially compute energy profiles for protein domains within Astral 40 and Astral 95, at both sequence and structure levels. Our findings demonstrate a highly significant correlation between the total energy estimated from sequence and structure, as well as a pronounced correlation between the distances among energy profile estimated from sequence and structure. Utilizing UMAP projections, we observe a clear separation of proteins belonging to the all-alpha and all-beta classes, indicating that the energy profile encapsulates structural information. Our findings also revealed that the energetic profiles of homologous protein structures exhibit closer similarities compared to those of non-homologous protein structures. To delve deeper, we randomly selected two folds (a.100 and a.104), two superfamilies (a.29.2 and a.29.3), and two families (a.25.1.0 and a.25.1.2) from SCOP database, calculating energy profiles at both sequence and structure levels. The resulting UMAP plot underscores the informative nature of energy profiles in clustering proteins across different folds, superfamilies, and families. Subsequently, the application of 1-NN classification on two families with CATH IDs 1.10.28.10 and 3.10.28.10 reveals that our energy profile surpasses classical methods such as TM-Score[16], RMSD [15], GR-Align, and Yau-Hausdroff distance in terms of accuracy and computational efficiency.

In conjunction with the organizational frameworks provided by SCOP, CATH, and PFAM for the protein universe, it is important to note their limitations, as they may present conflicting classifications and lack the ability to elucidate evolutionary relationships between individual superfamilies across extensive evolutionary distances. Ludin et al. conducted a comprehensive analysis of protein structures within the functionally diverse ferritin-like superfamily. They employed an evolutionary network construction approach to unveil relationships among proteins beyond the "twilight zone," where sequence similarity alone fails to facilitate meaningful evolutionary analysis. Building on this context, our study leverages profiles of energies to reconstruct a phylogenetic network. Our findings strongly suggest that a substantial and valuable evolutionary signal is preserved within the profile of energy, serving as a representative indicator of protein structure.

In a final exploration, we delved into the evolutionary relationships of spike proteins within SARS-CoV, MERS-CoV, and SARS-CoV-2 coronaviruses. This investigation was undertaken through the application of a hierarchical clustering method, shedding light on the intricate evolutionary dynamics within these viral proteins.

This study offers a means to characterize and compare proteins using energy profiles, enabling predictions of their structural and functional properties. Furthermore, this computational framework not only facilitates our understanding of individual protein behavior but also contributes to the broader exploration of evolutionary relationships, functional annotations, and drug discovery in the intricate world of proteins.

**2 Materials and Methods**

A non-redundant structural dataset of 6944 protein chains was culled by PISCES from PDB with pairwise sequence identity < 50%, resolution < 1.6 Å, R-factor < 0.25, protein length > 40 and < 1000 residues. These proteins were applied to train and calculate the knowledge-based potential function.

**2.1 Pairwise distance-dependent knowledge-based potential.**

It is generally assumed that two atoms are in contact if their spatial distance is less than a certain threshold. As a result, atoms separated by less than a constant number interact. There is no direct interaction between two atoms when there is a third atom between two close atoms. Based on Delaunay tessellation, a representation that eliminates contact between two atoms when a third atom is between them, a physical contact can be accurately modeled. In 2014, Mirzaie et al. introduced an all-atom knowledge-based potential based on Delaunay Tessellation [13]. In this study we apply this potential function to calculate the potential of pairwise contact. The energy between the two atoms i and j at distance d, is calculated as follows:

(1)

where the temperature *T* was set to 293 K, corresponding to *RT* of 0.582 kcal/mole, where *R* is the gas constant.  is the number of observations for atomic pair  and , is the relative frequency of occurrence for  and  in distance class , is the relative frequency of occurrence for all atomic pairs in distance shell , and is the weight given to each observation. As discussed by Sippl, it was assumed that. The potential energy associated with the interaction of residues A and B denoted by is estimated by summing the pairwise potentials between the atoms of each of these residues as follows:

(2)

which the sum is on all pairs of atoms in contact with the Delaunay triangulation method.

Consider the set T, which encompasses all 20 types of amino acids in proteins. Given that there are 210 unique amino acid-amino acid interaction types among these 20 amino acids, the total count of unique values are 210. As a result, we create a 210-dimensional vector to represent distance-dependent energy interactions between residues, with each dimension representing the energy interaction between distinct pairs of amino acids. We call this 210-dimensional vector as the **Structural Profile of Energy (SPE)** of a protein structure.

**2.2 The pairwise energy content estimated from amino acid composition.**

The knowledge-based potential function discussed in the previous section relies on having the three-dimensional structure of a protein. Nevertheless, it's worth noting that the three-dimensional structures of numerous proteins have not yet been determined experimentally. Dosztányi et al. proposed a method to estimate the energy from a protein sequence [17]. They approximated , the total energy per amino acid, based on the protein’s amino acid composition. Let be the length of the sequence, be the number of amino acid residues of type in the sequence and be its relative frequency. The energy per amino acid, as approximated by Dosztányi et al, is as follows [17]:

where is the energy predictor matrix estimated using protein structures from the training dataset as detailed by Dosztányi et al [17]. For each pair of amino acid types and , we used the following equation to estimate the energy based on amino acid sequence composition:

As a result, we create a 210-dimensional vector to represent energy between amino acid types using amino acid composition. We call this 210-dimensional vector as the **Compositional Profile of Energy (CPE)** of a protein sequence.

**3 Results**

**Correlation between Energy estimated based on structure and Sequence**

Initially, we calculated energies for protein domains in the ASTRAL40 and ASTRAL95 datasets using both structure- and sequence-based methods. The ASTRAL 40 (95) dataset consists of protein domains with sequence similarities of less than 40% (95%) [18]. This dataset offers a comprehensive description of structural and evolutionary relationships among proteins from the Protein Data Bank. As mentioned in Method section, we can calculate the total energy and the profiles of energy form protein structure or predicted from sequences. Figure 1A-B illustrates the correlation between the total energy estimated from structure on y-axis and from sequence on x-axis. The observed high correlation coefficient suggests that sequence-based energy estimation serves as a reliable approximation, particularly in scenarios where the protein structure is unidentified. Similarly, Figure 1C-D depicts the connection between energy profile distances for all pairs of domains in the ASTRAL40 and ASTRAL95 datasets, as estimated from both structure and sequence. Likewise, a strong correlation exists between the distance of pairs of energy profiles estimated from sequence and structure.

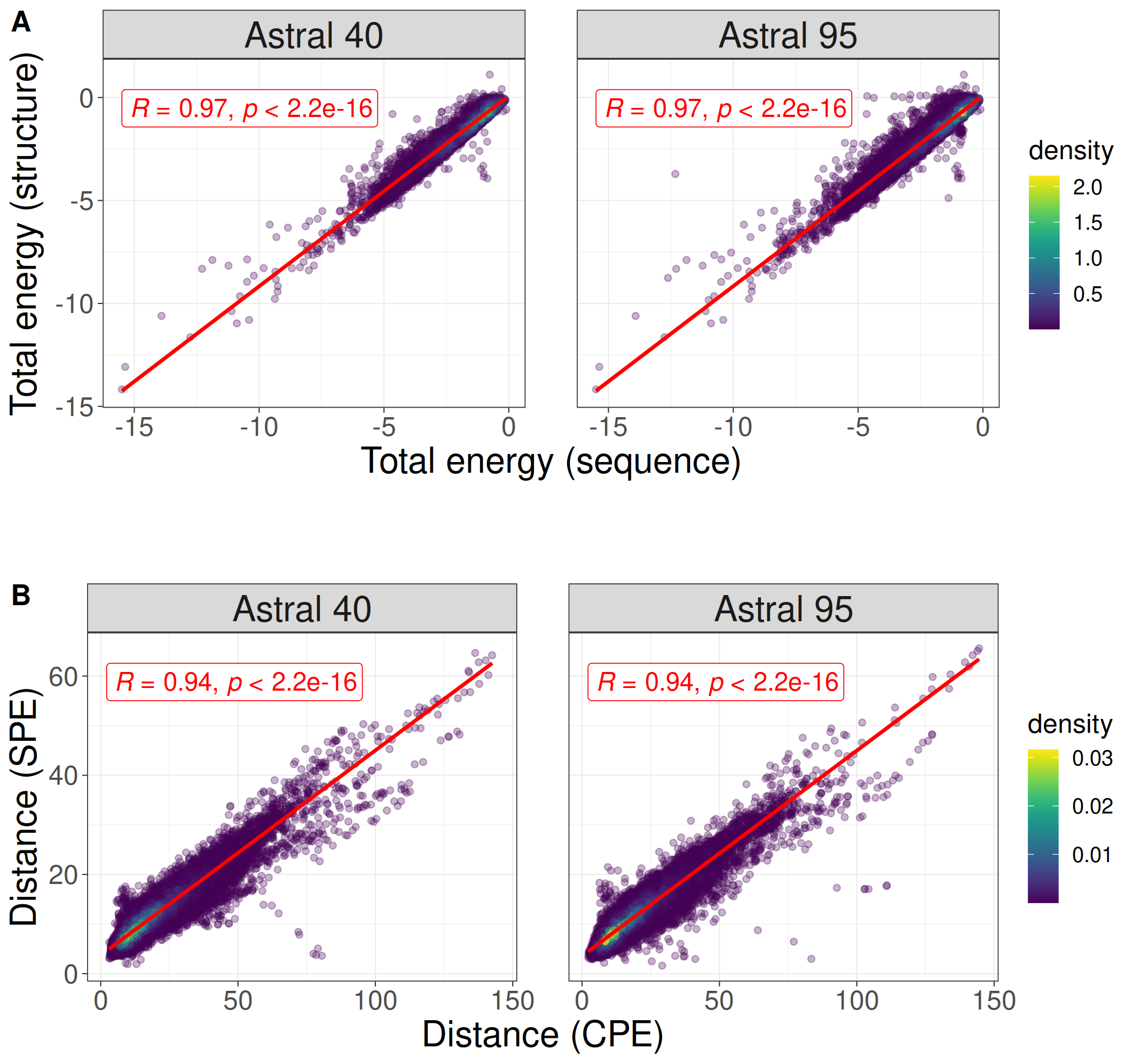


Figure 1. The correlation between total energy estimates derived from protein structure and sequence for protein domains within Astral 40 (depicted in A) and Astral 95 (depicted in B). Additionally, the figure presents the correlation between the distance of energy profile pairs for domains in Astral 40 (shown in C) and Astral 95 (shown in D)

The stability, mutational robustness, and design adaptability of α-helices relative to β-strands in natural proteins have been widely acknowledged in scientific literature. To investigate this phenomenon, Figure 2 presents the distribution of total energy within protein domains from the ASTRAL40 and ASTRAL95 datasets, categorized into four structural scope classes: all-alpha, all-beta, alpha+beta, and alpha/beta. Total energies, normalized by protein length, are analyzed to discern patterns across these structural classes. The figure highlights significant differences in total energy among domains with different structural compositions, suggesting diverse energetic landscapes associated with distinct protein structures. This observation is consistent with similar trends observed in energy estimations derived from sequence information, reinforcing the importance of structural motifs in influencing the energetic properties of proteins. (add numbers of domains at different classes)

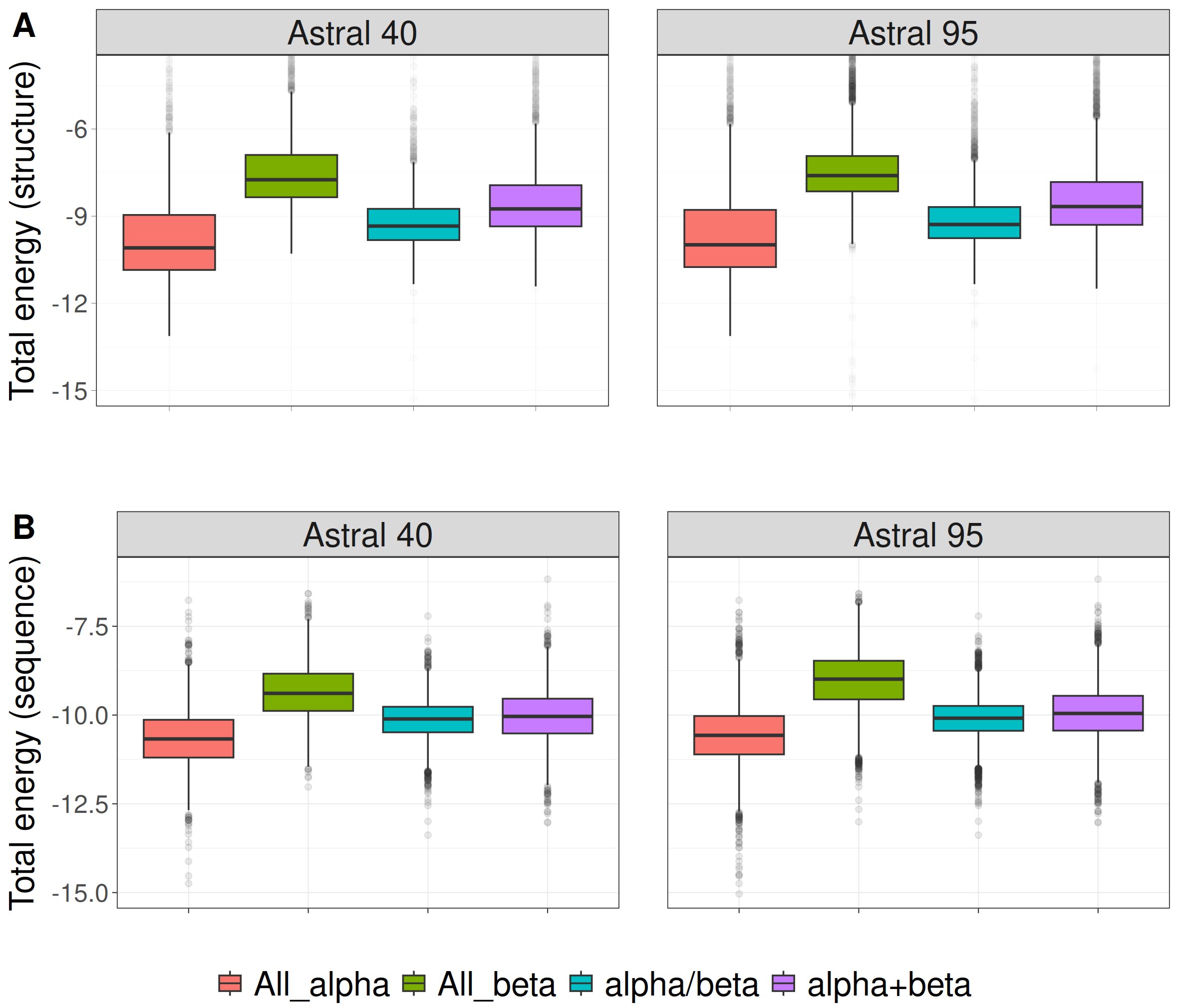


Figure 2. The distribution of normalized total energy in protein domains from ASTRAL40 and ASTRAL95 datasets based on protein structure (A, B) and sequence (C, D) across various structural scope classes. In the ASTRAL40 dataset, there are 2644, 3059, 4463, and 3653 protein domains in the all-alpha, all-beta, alpha+beta, and alpha/beta classes, respectively. Similarly, in the ASTRAL95 dataset, there are 3443, 10164, 9344, and 7474 protein domains in the all-alpha, all-beta, alpha+beta, and alpha/beta classes, respectively.

We visualized the energy profiles generated by sequence and structure for domains belonging to the all-alpha and all-beta classes. As depicted in Figure 3, the UMAP embeddings of energetic profiles effectively capture the structural characteristics that differentiate between all-alpha and all-beta domains. This visualization highlights the presence of distinct energy profiles between all-alpha and all-beta domains, a consistency that is also evident in sequence-based energy profiles.

A group of images of different colored dots

Description automatically generated

Figure 3: UMAP projection of SPE and CPE shows the separation of the all-alpha (green point) and all-beta (pink point) proteins selected from the ASTRAL 40 and 95 dataset. A) SPE of ASTRAL40, B) CPE of ASTRAL 40, C) SPE of ASTRAL 95, and D) CPE of ASTRAL 95. Dots represent two dimensional UMAP projection of SPE for individual sequences. UMAP plots were generated by parameters n\_neighbors = 20 and min\_dist = 0.1.

To explore the structural information of energy profiles at lower hierarchical levels of SCOP, we randomly selected two folds (a.100 and a.104) from the all-alpha class, two superfamilies (a.29.2 and a.29.3) from the fold a.29, and two families (a.25.1.0 and a.25.1.2) from the superfamily a.25.1. In Figure 4, each panel showcases two figures, with the left figure generated by SEP profiles and the right figure by CEP profiles. The UMAP plot in Figure 4 underscores that protein domains within the same fold, superfamily, or family share similar energy profile patterns.

|  |  |
| --- | --- |
| **Astral 40** | **Astral 95** |
|  |  |
|  |  |
|  |  |

Figure 4. The UMAP projection of Structural Energy Profiles (SEP) and Compositional Energy Profiles (CEP) on protein domains from Astral40 and Astral 95 represents the structural information embedded in energy profiles across hierarchical levels of SCOP; each panel includes two figures, one generated by SEP (left panel) and the other by CEP (right panel), revealing that protein domains sharing the same fold, superfamily, or family exhibit comparable energy profile patterns. The folds a.100 and a.104, superfamilies a.29.2 and a.29.3, as well as families a.25.1.0 and a.25.1.2, are randomly selected for analysis, and the UMAP plots are generated using parameters n\_neighbors = 20 and min\_dist = 0.1.

To delve deeper into the differences in distances among protein domains within the same class, we calculated pairwise distances between the energy profiles of protein domains within the all-alpha class from the Astral 95 dataset. Subsequently, we compared these distances with the distances of energy profiles from protein domains across different classes. In Figure 5, each panel displays two figures, with the left figure generated by SEP profiles and the right figure by CEP profiles. As depicted in Figure 5A, interclass distances are notably lower than intraclass distances. Similar results were obtained when we calculated pairwise distances from protein domains within fold a.29 and compared them with pairwise distances from protein domains in different folds within the all-alpha class (Figure 5B). Likewise, the distances between energy profile patterns of protein domains within the same superfamily a.29.3 are significantly less than the distances between energy profiles of protein domains within fold a.29 that belong to different superfamilies (Figure 5C). Consequently, it can be inferred that energy profiles of domains belonging to the same superfamily/fold/class exhibit greater similarity than those from different superfamilies/folds/classes.

|  |  |
| --- | --- |
| **Astral40** | **Astral 95** |
| A |  |
| B |  |
| C |  |

Figure 5: Comparative Boxplots of Pairwise Distances among Energy Profiles in Astral 40 and Astral 95, depicting A) intraclass distances within the all-alpha class (in purple) versus interclass distances (in yellow), B) intraclass distances within the a.29 fold (in purple) versus distances from protein domains in different folds within the all-alpha class (in yellow), and C) intraclass distances within the a.29.3 superfamily (in purple) versus distances from protein domains in different superfamilies within the fold a.29 (in yellow). Each panel presents two figures, one generated using Structural Energy Profiles (SEP, left panel) and the other using Sequence-Coupled Energy Profiles (CEP, right panel).

To assess the profile of energy in protein superfamily classification, we investigated five distinct SCOP superfamilies: winged helix (a.4.5), PH domain-like (b.55.1), NTF-like (d.17.4), Ubiquitin-like (d.15.1), and Macroglobulin (b.1.29). Our classification strategy incorporated energetic profiles as features, employing Support Vector Machine (SVM) and Random Forest (RF) classifiers as our models. To ensure the robustness and generalization of our models, we subjected them to rigorous 10-fold cross-validation. The outcomes, presented in Table 4, encompass both accuracy and F1-measure, revealing the performance of our models. The 10-fold cross-validation results highlight the SVM classifier's robust performance, achieving an impressive 96.8% accuracy across the five superfamilies. This high accuracy underscores the efficacy of the SVM algorithm in precisely classifying remotely homologous proteins, substantiating the credibility and reliability of our classification model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Total-accuracy | F1\_wigend\_helix | F1\_PH.domain-like | F1\_NTF-like | F1\_Ubiquitin-like | F1\_Immunoglobulins |
| Fold01 | 0.977 | 0.963 | 0.85 | 0.92 | 0.969 | 0.992 |
| Fold02 | 0.956 | 0.935 | 0.647 | 0.963 | 0.921 | 0.981 |
| Fold03 | 0.968 | 0.948 | 0.78 | 0.981 | 0.951 | 0.985 |
| Fold04 | 0.972 | 0.94 | 0.821 | 0.964 | 0.938 | 0.993 |
| Fold05 | 0.977 | 0.963 | 0.857 | 0.962 | 0.957 | 0.992 |
| Fold06 | 0.963 | 0.955 | 0.718 | 0.963 | 0.926 | 0.985 |
| Fold07 | 0.973 | 0.978 | 0.913 | 0.963 | 0.942 | 0.984 |
| Fold08 | 0.963 | 0.922 | 0.8 | 0.863 | 0.944 | 0.991 |
| Fold09 | 0.97 | 0.923 | 0.884 | 0.898 | 0.937 | 0.996 |
| Fold10 | 0.959 | 0.937 | 0.722 | 0.898 | 0.908 | 0.99 |
| Average | 0.968 | 0.946 | 0.799 | 0.937 | 0.939 | 0.989 |

*Table 4: Classification performance statistics are given for each of the five superfamilies by 10-Fold cross validation with SVM. Statistics presented in the table include: accuracy and the F1-measure for each fold.*

It is commonly assumed that proteins sharing similar structures also exhibit similar functions. Various measurements, such as RMSD, TM-score, RG-align, and Yau-Hausdorff distance, have been developed to assess protein structure similarity. Here, we employed a benchmark dataset sourced from the CATH v4.2.0 database, comprising 260 protein domains from two distinct protein families: the C-terminal domain in the DNA helicase RuvA subunit (representing the Alpha class, characterized by Orthogonal Bundle Architecture, Helicase, and Ruva Protein fold, with CATH Code: 1.10.8.10), and the Homing endonucleases (belonging to the Alpha and Beta class, featuring Roll Architecture, and Endonuclease I-creI fold, with CATH Code: 3.10.28.10). The protein domains varied in the number of residues, ranging from 44 to 854, with an average of 211.

We employed the 1-nearest neighbor (1-NN) classification method to categorize proteins based on GR-Align, RMSD, TM-score, Yau-Hausdorff distance, and the distance between profiles of energy as a measure of protein dissimilarity. As outlined in Table 1, our approach exhibited superior accuracy and faster processing times compared to other methods. It is noteworthy that our method obviates the need for superimposing two protein structures or conducting structural alignments; instead, we calculate the profiles of energies and determine the distance between these profiles. The computations were executed on a PC with a configuration of 2.40 GHz and 8 GB RAM. Table 1 provides a detailed breakdown of results and processing times, highlighting the efficient implementation of energy profile calculation and the 1-NN algorithm, accomplished in approximately 10 minutes on a system with a 2.4 GHz processor and 4GB RAM. Impressively, our methodology achieved a remarkable classification accuracy of 99% in distinguishing between the two protein families.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | GR-Align | RMSD | TM-Score | YH (10Rotation) | YH (2500Rotation) | SEP | CEP |
| Accuracy  Time | 62.3%  2 min | 59.2%  1 h | 61.5%  9h 20 min | 70.8%  10 min | 81.5%  4h 10 min | 99%  10 min | 96%  3 min |

*Table 1: Summary of the classification results with our proposed energy model, common methods and YH distance.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **C-Terminal** | **Homing** | **Error** |
| **C-Terminal** | **159 , 155** | **0 , 6** | **0.012 , 0.037** |
| **Homing** | **3 , 4** | **94 , 93** | **0.030 , 0.041** |

*Table 2: Confusion matrix by 1-NN for energetic profile 210-D*

**3.6 Phylogeny Inference Using Energy Profiles**

In the realm of structural biology and evolutionary analysis, three-dimensional protein structure classification and the alignment of multiple sequences stand as formidable tools for uncovering structural similarities and deducing phylogenetic relationships. A phylogeny, often visualized as a tree, serves as a narrative of evolutionary processes, elucidating the intricate relationships that exist among various entities, be they genes, populations, species, or other biological units.

In this section, we leverage energy profiles for the inference and reconstruction of phylogenetic trees. Specifically, we apply this method to unveil the phylogenetic relationships within two distinct biological contexts: the coronavirus spike glycoprotein structures and the ferritin superfamily. Through this analysis, we aim to shed light on the evolutionary histories and interconnections that underlie these essential biological entities, ultimately enriching our understanding of their structural evolution and functional relationships.

**3.6.1 Exploring the Phylogeny of the SARS-CoV-2 proteins**

Over the past two decades, Coronaviruses (CoVs) have been associated with various outbreaks, including the 2002–2003 SARS-CoV outbreak, the 2012 MERS-CoV incident (1), and the recent COVID-19 pandemic initiated by SARS-CoV-2 in late 2019. Since February 2020, a considerable number of SARS-CoV-2 protein structures have been recorded in the Protein Data Bank (PDB). One pivotal viral protein, the spike glycoprotein, has garnered significant attention. As a transmembrane glycoprotein, it plays a central role in viral infection by facilitating host receptor binding and stands as the primary target for neutralizing antibodies and vaccine design. To thoroughly investigate the structural landscape of these spike glycoproteins and gain insights into their evolutionary connections, we utilized the CoV3D database (https://cov3d.ibbr.umd.edu), a comprehensive repository containing diverse coronavirus protein structures and their complex interactions with antibodies, receptors, and small molecules.

From the CoV3D database, we curated a dataset comprising 143 spike glycoprotein structures distinguished by the presence of the closed receptor binding domain (RBD) within their structure. This dataset encompasses 80 chains from SARS-CoV-2, 31 chains from SARS-CoV, and 32 chains from MERS-CoV. To scrutinize the structural variations and relationships among these spike glycoproteins, we generated a 210-dimensional profile of energies at both sequence and structure levels. By calculating Manhattan distances between all pairs of energetic profiles, we successfully categorized the spike glycoprotein structures into three distinct clusters through unsupervised clustering based on these distances. These clusters correspond to the SARS-CoV, MERS-CoV, and SARS-CoV-2 viruses, offering a visually informative representation of the structural relationships within this protein family ( Figure 7).

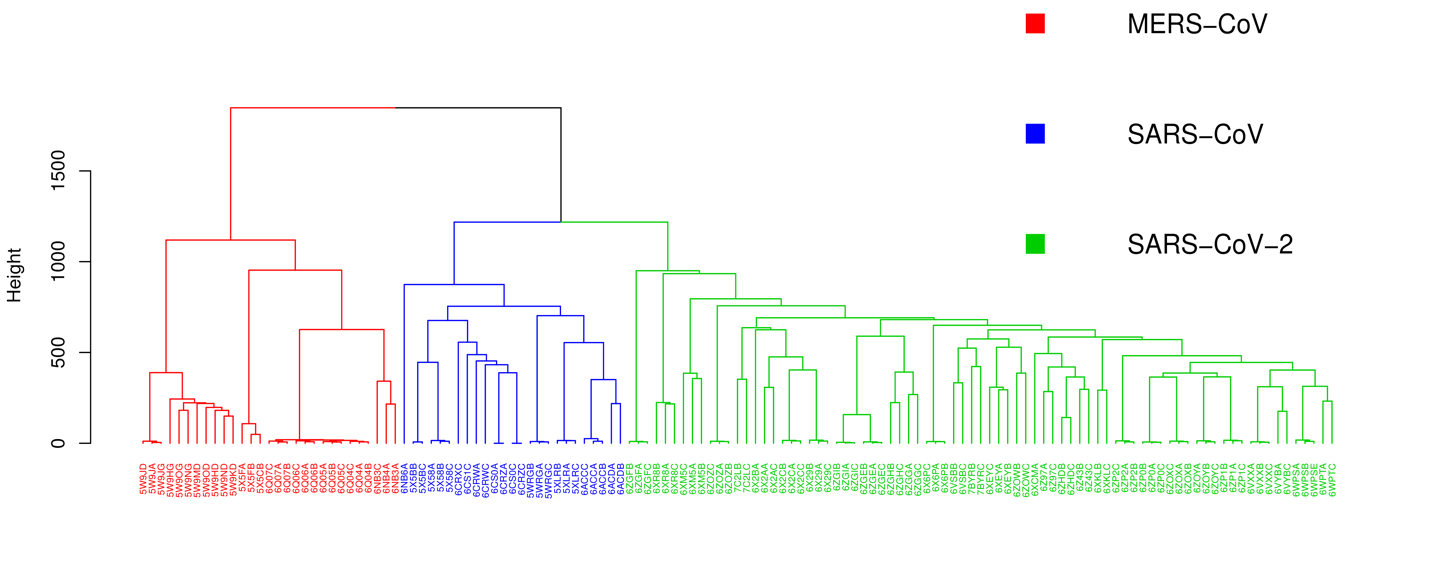


Figure 7: The dendrograms illustrate the clustering of spike glycoprotein structures from three viruses SARS-CoV, SARS-CoV-2, and MERS-CoV. The clustering is based on pairwise distances of energy profiles derived from A) protein structure and B) protein sequence. Each leaf on the dendrogram is labeled with the PDB-IDs of the corresponding chains, and the leaves are color-coded to represent the virus of the spike glycoprotein structure).

**3.6.2 Exploring the Phylogeny of the Ferritin-Like Superfamily**

Several significant databases aim to organize the protein universe at a high level, such as Pfam relying on sequence information, and both SCOP and CATH utilizing protein structural data. These databases categorize proteins into families or superfamilies based on measures of either sequence or structural similarity. While these databases are essential for outlining broad structural relationship, they often present conflicting classifications, and lacking information on evolutionary relationships among individual superfamily components[21]. SCOP superfamilies contain protein families that are assumed to be evolutionary related based on sequence and structural similarity and functional commonalities. Ludin et.al [21] investigated how ferritin-like proteins are classified across Pfam, SCOP, and CATH. Notably, this superfamily encompasses a diverse range of proteins, including iron-storing ferritins, methane monooxygenases, the small subunit of RNR R2, rubrerythrins, bacterioferritins, Dps (DNA binding protein from starved cells that protects against oxidative DNA damage), and Dps-like proteins. As discussed by Ludin et.al at the superfamily level, the classification of the “ferritine-like” superfamily appears consistent across these databases but does differ in the amount of information provided regarding the relationships and functions of superfamily constituents. So although the classification in all three databases is hierarchical, they do not encompass all level of functional and evolutionary information. The low sequence similarities across this superfamily make it feasible to construct sequence-based phylogenies only for specific subsets. Consequently, addressing this challenge requires efforts to integrate structural information with sequence-based phylogenies. Lundin et al. [21] and Malik et al. [22] delved into the evolutionary relationships of this superfamily by creating a phylogenetic network. They employed the distance-based NeighborNet network method, utilizing distances calculated through structure-based alignment methods. To reconstruct the previously published structural phylogeny of the ferritin-like superfamily, we utilized the same protein structures within this superfamily as Lundin et al. and Malik et al. The dataset specifically focuses on the SCOP superfamily, Ferritin-like (a.25.1) encompassing two manually curated protein families: Ferritin (a.25.1.1) and RiboNucleotide Reductase-like [RNR] (a.25.1.2). The “Ferritin” family contains ferritins, bacterioferritins, and Dodecameric ferritin homolog (Dps) proteins and the “RiboNucleotide Reductase-like” family contains the activating subunit of class I ribonucleotide reductase (RNR R2), BMM, and Fads. Following this, we computed the Structure Profile Energetics (SPE) for each protein and determined the distances between SPEs. The reconstruction of the phylogenetic tree was achieved using the neighbor-joining method and showed in figure 8[23].

Our results suggest that the energetic phylogenies within the ferritin-like superfamily unveil significant relationships among its members, aligning with known evolutionary connections and functional roles. In line with prior investigations, a key observation is that the resulting phylogenetic tree exhibits two primary branches, corresponding to two families a.25.1.1 and a.25.1.2. Consequently, our methodology accurately bifurcates this superfamily into two families. Additionally, our approach faithfully replicates the grouping presented in the figure by Lundin et al. in 2012. Delving into specifics, the family a.25.1.1 further divides into four subgroups: "ferritins," "Dps and related," "Rubrerythrin," and "Bacterioferritins." On the other hand, the second branch related to the a.25.1.2 family, despite SCOP and CATH assigning these proteins to a unified RNR-like family, reveals three distinct families according to Pfam—Phenol\_Hydrox (PF02332), Ribonuc\_red\_sm (PF00268), and Fatty acid desaturase (PF03405). Our results consistently support this more detailed sequence-based classification, as well as the further subdivision of the BMMs into BMMa and BMMb.

There are some proteins in our networks fall outside of the major groupings in our networks, all of which are classified as Ferritins by CATH, with most also classified as ferritins by SCOP. For instance, in our network, 1otkA is closer to the RNRs rather than the ferritins. It is noteworthy that Pfam classifies this protein into PaaA\_PaaC, with 1otkA the only member of PaaA\_PaaC. Another protein is 3ee4A which Pfam classifies as Ribonuc\_red\_sm, possibly because of its sequence similarity to RNR R2 proteins (30). In our network this structure clearly occupies an outgroup position relative to the RNR R2 structures. This is functionally consistent with its ligand-binding pocket, which indicates that it is a substrate oxidizing enzyme, and its lack of competence as an RNR R2 (30, 31).

In the Fatty-acid group, there is a distinct cluster of plant Fads (2uw1A-B, 1afrA), whereas the Mycobacterium tuberculosis protein (1za0A) appears more distantly related. As discussed by Lundin et al. [21] unfortunately, this is the only solved structure of a bacterial Fad. It is also one of a paralogous pair and not the one considered functional. The structure of the functional Fad has not yet been possible to solve (28). With such a skewed data set, it is difficult to judge how well our energetic-based network identifies evolutionary relationships within the Fad group.

Proteins 1mtyB-D, 2incB, 2inpC, and 3dhgA are members of the PF02332 family. This protein family includes several components of multicomponent enzyme systems predominantly found in Proteobacteria and Actinobacteria, including subunits alpha and beta of a methane monooxygenase and an alkene monooxygenase system, small and large subunits of propane 2-monooxygenase system and P1 protein of phenol hydroxylase. Through sequence-based phylogenies (22, 27), it has been proposed that bacterial multicomponent monooxygenases (BMMs) evolved through duplication and divergence, resulting in distinct catalytic (α) and non-metal binding (β) subunits. While BMMs generally exhibit low substrate specificity, the discrimination between α and β subunits into two clans—one with proteins annotated as metal iron-bindings (1mhyD B, 3dhgA) and another subgroup comprising non-metal bindings (2inpC, 2incB)—is evident. This suggests that our energetic analysis can uncover both recent and more distant evolutionary relationships.

Concerning the 2fzfA protein, SCOP and CATH classify it as Ferritin, and Pfam categorizes it within the Rubrerythin group. In the study by Ludin et al., this protein is not placed in the Ferritin group, but in our study, it is classified as belonging to Ferritins. It's worth noting that RCSB classifies it as UNKNOWN FUNCTION.

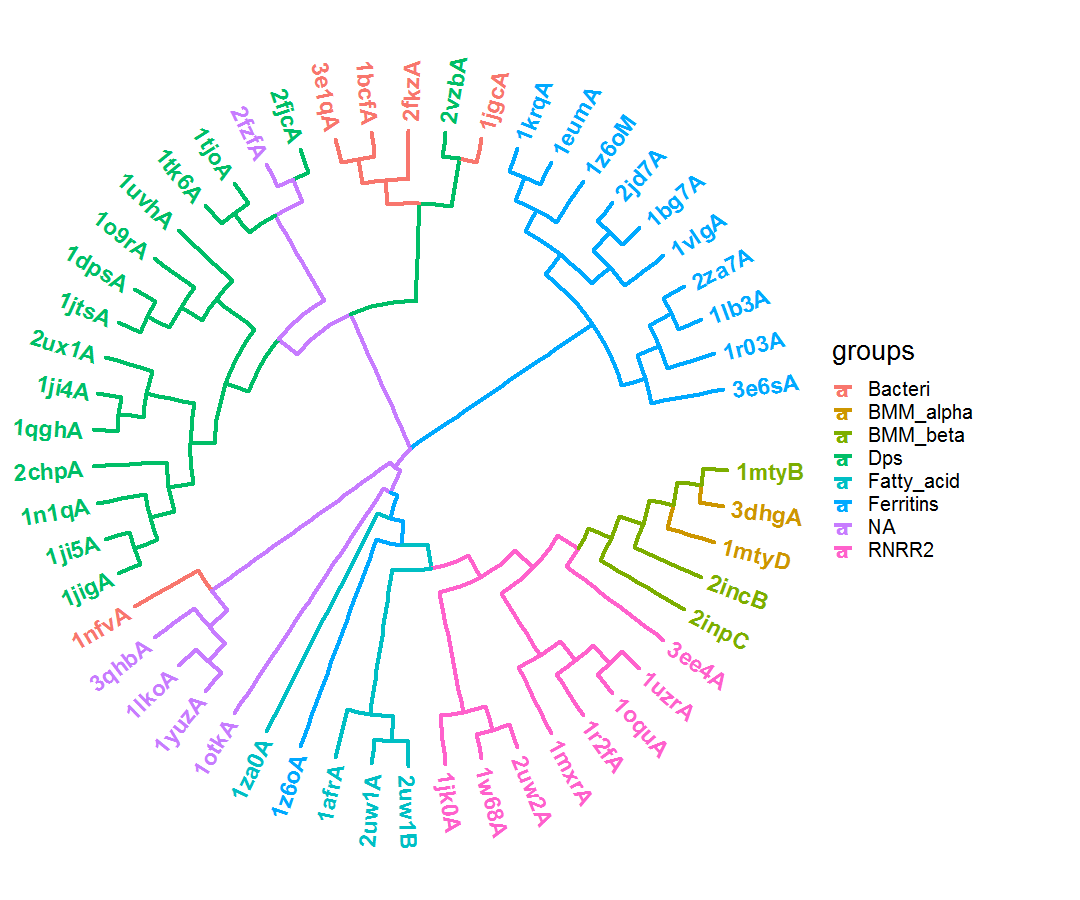
****

Figure 8: Energy-based phylogenetic network of the ferritin-like superfamily. The two large SCOP families, ferritins (a.25.1.1; Bacteri, Ferritins, Dps and Rubrerythrin) and ribonucleotide reductase-like (a.25.1.2; BMM\_alpha, BMM\_beta, Fatty\_acid and RNRR2) separated in the tree with smaller groups.

In this section, we analysed a diverse set of bacterial families called bacteriocins, utilizing the BAGEL database for analysis. Bacteriocins are small peptides produced by bacteria, functioning as potent antimicrobial agents that target other microbial species. Detecting and understanding these peptides is crucial due to their ecological importance, but their diverse sequences and structures make them challenging to identify using traditional methods. To address this issue, the BAGEL tool was developed in 2006, specifically designed for identifying Ribosomally synthesized and post-translationally modified peptides (RiPP) and bacteriocin biosynthetic gene clusters (BGCs). BAGEL categorizes bacteriocins based on size and stability into RiPPs (also defined as class I bacteriocins by BAGEL), class II bacteriocins (small heat stable proteins < 10 kDa) and class III bacteriocins (large heat-labile proteins > 10 kDa). As shown in figure 9, our analysis revealed that profile of energy can clearly partition bacteriocins according to BAGEL annotation.

**Analysis Tools and Packages**

All computational analyses were conducted using the versatile R programming language (www.r-project.org), with the utilization of various specialized packages tailored for specific tasks. Below is an overview of the packages and tools employed throughout our analysis:

The BIO3D package was used to read PDB files and analyze them. The Quickhull algorithm in the geometry package was used to find direct contacts and nearest neighbors of atoms using the Delaunay tessellation method. Class, random Forest, and e1071 packages were used for kNN, RF, and SVM classification methods, respectively, and cross-validation was performed using the caret package. To visualize the results, the ggplot2 package was used. The Ape and ggtree packages were used to implement and visualize hierarchical clustering and the NJ method.

**Discussion**

In this study, we have presented a comprehensive analysis of the utility of energy profiles in various aspects of structural biology and bioinformatics. Our investigation has spanned the classification of proteins based on structural similarities, the classification of proteins into distinct structural classes, the classification of proteins into SCOP superfamilies, and the inference of phylogenetic relationships among proteins and protein families.

Our initial exploration focused on discerning potential correlations within the energy profiles of remotely homologous proteins. We demonstrated that even proteins with relatively low sequence identity and structural similarity, such as the TetR family repressor M. tuberculosis EthR and the putative transcriptional regulator ycdc, can exhibit significant correlations in their energy profiles. This finding suggests that energy profiles can capture meaningful structural and functional information, even in cases where traditional sequence or structural alignment methods may fail. Furthermore, our extensive analysis using the ASTRAL SCOPe 2.08 dataset confirmed that energy profiles of proteins within the same family exhibit stronger correlations compared to non-homologous proteins. The statistical significance of these differences was established through the Kruskal-Wallis test, highlighting the potential of energy profiles as a valuable tool for protein classification and functional annotation.

Our investigation into the efficacy of energy profiles in quantifying structural similarities among protein structures revealed promising results. We employed a diverse benchmark dataset representing two distinct superfamilies: the C-terminal domain of the DNA helicase RuvA subunit and the Homing endonucleases. Our analysis demonstrated that energy profiles effectively discriminate between these two superfamilies, achieving a classification accuracy of 98%. This impressive accuracy underscores the potential of energy profiles as a powerful feature for structural classification.

We also compared our energy profile-based approach with other common methods, such as GR-Align, RMSD, TM-score, and the Yau-Hausdorff distance. Our methodology consistently outperformed these methods in terms of accuracy, while remaining computationally efficient.

In the context of classifying proteins into structural classes (all-alpha and all-beta), energy profiles proved to be highly effective. Our analysis using Support Vector Machine (SVM) and Random Forest (RF) classifiers consistently achieved an average accuracy score of 0.85 in 10-fold cross-validation. This robust performance highlights the potential of energy profiles for automated protein classification based on their underlying structural characteristics.

Extending our analysis to classify proteins into SCOP superfamilies, we employed SVM and RF classifiers and subjected them to rigorous 10-fold cross-validation. Our results demonstrated a high classification accuracy of 96.8% across five distinct SCOP superfamilies. This finding suggests that energy profiles can capture subtle structural variations that are characteristic of specific superfamilies, enabling accurate classification even for remotely homologous proteins.

Within the Aspartase superfamily, our analysis further showcased the effectiveness of energy profiles in classifying proteins at the family level. Despite minimal sequence similarity, energy profiles successfully separated proteins into three distinct families within the superfamily. This result emphasizes the potential of energy profiles in superfamily classification across diverse protein families, contributing to our understanding of their structural evolution.

In the realm of phylogenetic inference, we introduced a novel approach using energy profiles to uncover the evolutionary relationships among proteins and protein families. Our analysis focused on two biological contexts: the coronavirus spike glycoprotein structures and the ferritin-like superfamily.

In the case of the coronavirus spike glycoprotein structures, our energy profile-based approach revealed distinct clusters corresponding to SARS-CoV, MERS-CoV, and SARS-CoV-2. This clustering offered valuable insights into the structural relationships among these spike glycoproteins, complementing sequence-based phylogenetic analyses.

In the ferritin-like superfamily, our phylogenetic tree reconstruction based on energy profiles successfully separated two SCOP families and further subdivided them into distinct groups, aligning with previous studies. This outcome underscores the robustness of our energy profile-based phylogenetic approach and its ability to recapitulate known structural classifications.

Overall, our findings demonstrate the versatility and effectiveness of energy profiles as a valuable tool in structural biology and bioinformatics. Energy profiles can capture subtle structural and functional information that may be missed by traditional sequence and structural alignment methods, making them a promising avenue for future research in protein classification and phylogenetics.

**References**

1. Sayers, E.W., et al., *Database resources of the national center for biotechnology information.* Nucleic acids research, 2021. **49**(D1): p. D10.

2. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.

3. Lipman, D.J. and W.R. Pearson, *Rapid and sensitive protein similarity searches.* Science, 1985. **227**(4693): p. 1435-1441.

4. Mistry, J., et al., *Pfam: The protein families database in 2021.* Nucleic acids research, 2021. **49**(D1): p. D412-D419.

5. Jain, A., et al., *Analyzing effect of quadruple multiple sequence alignments on deep learning based protein inter-residue distance prediction.* Scientific Reports, 2021. **11**(1): p. 7574.

6. Szklarczyk, D., et al., *STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic acids research, 2019. **47**(D1): p. D607-D613.

7. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proceedings of the National Academy of Sciences, 1999. **96**(8): p. 4285-4288.

8. Zhu, X., Y. Xiong, and D. Kihara, *Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2. 0.* Bioinformatics, 2015. **31**(5): p. 707-713.

9. Sillitoe, I., et al., *CATH: increased structural coverage of functional space.* Nucleic acids research, 2021. **49**(D1): p. D266-D273.

10. Lo Conte, L., et al., *SCOP: a structural classification of proteins database.* Nucleic acids research, 2000. **28**(1): p. 257-259.

11. Sippl, M.J., *Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.* Journal of computer-aided molecular design, 1993. **7**: p. 473--501.

12. Mirzaie, M. and M. Sadeghi, *Knowledge-based potentials in protein fold recognition.* Archives of Advances in Biosciences, 2010. **1**(4).

13. Mirzaie, M. and M. Sadeghi, *Delaunay‐based nonlocal interactions are sufficient and accurate in protein fold recognition.* Proteins: Structure, Function, and Bioinformatics, 2014. **82**(3): p. 415-423.

14. Melo, F., R. Snchez, and A. Sali, *Statistical potentials for fold assessment.* Protein science, 2002. **11**(2): p. 430--448.

15. Maiorov, V.N. and G.M. Crippen, *Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins.* Journal of molecular biology, 1994. **235**(2): p. 625--634.

16. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins: Structure, Function, and Bioinformatics, 2004. **57**(4): p. 702--710.

17. Dosztanyi, Z., et al., *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* Journal of molecular biology, 2005. **347**(4): p. 827--839.

18. Fox, N.K., S.E. Brenner, and J.-M. Chandonia, *SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures.* Nucleic acids research, 2014. **42**(D1): p. D304--D309.

19. Tian, K., et al., *Comparing protein structures and inferring functions with a novel three-dimensional Yau--Hausdorff method.* Journal of Biomolecular Structure and Dynamics, 2019. **37**(16): p. 4151--4160.

20. Viola, R.E., *L-aspartase: new tricks from an old enzyme.* Advances in enzymology and related areas of molecular biology, 2000. **74**: p. 295--341.

21. Lundin, D., et al., *Use of structural phylogenetic networks for classification of the ferritin-like superfamily.* Journal of Biological Chemistry, 2012. **287**(24): p. 20565--20575.

22. Malik, A.J., A.M. Poole, and J.R. Allison, *Structural phylogenetics with confidence.* Molecular Biology and Evolution, 2020. **37**(9): p. 2711--2726.

23. Gascuel, O., *BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.* Molecular biology and evolution, 1997. **14**(7): p. 685--695.