**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] (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] (2) and FASTA [3] , as well as motif searches [4] (4), contemporary approaches have explored an extensive array of methodologies, including sequence-based techniques [5] (5), omics-data integration [6](6), phylogenetic profiling [7] (7) and three-dimensional protein structures [8](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] (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](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](11, 12), which leverage information extracted from known protein structures to estimate various energies, including distance-dependent interactions [13](13), dihedral angles, and accessible surface energies [14] (14). In this study, we employ the Sippl potential function [11](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.

To assess the effectiveness of this representation in protein classification, we conducted a series of comprehensive experiments. Initially, we demonstrated the remarkable efficiency and accuracy of our approach by swiftly and accurately classifying two CATH superfamilies. This classification surpassed the performance of alternative distance metrics such as RMSD [15](15) and TM-score [16](16). Subsequently, in a more extensive investigation, we unveiled the capability of energetic profiles to precisely differentiate between All-alpha and All-beta protein classes, as well as to distinguish five substantial homologous remote superfamilies from SCOP, achieving impressive accuracies of 85% and 97%, respectively. Our findings also revealed that the energetic profiles of homologous protein structures exhibit closer similarities compared to those of non-homologous protein structures. 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 functions.

**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 pairwise 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.

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

Figure 2 displays the distribution of total energy in protein domains within the ASTRAL40 and ASTRAL95 datasets across four structural scope classes: all-alpha, all-beta, alpha+beta, and alpha/beta. The total energies are normalized by the protein length. The illustration reveals significant variations in the total energy of domains across different structural classes. Similar outcomes were noted for energies estimated from sequences.

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

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 among various classes, 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** |
|  |  |
|  |  |
|  |  |

To delve deeper into the variations 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: Boxplot of Pearson correlations of energetic profiles within three protein-fold families. All domains (≤100 residues) of three SCOP families contained in the ASTRAL 2.08 40% representative’s database were extracted: (A) homeodomain (a.4.1.1, 20 members), (B) SH3-domains (b.34.2.1, 31 members), (C) glutathione S-transferase, N-terminal domain (c.47.1.5, 9 members). A randomly chosen set of nonhomologous domains equal in members for each family. (D) All homologous pairs (16474 pairs, median = 0.81), all non-homologous pairs (4712650 pairs, median = 0.60).*

**3.1 Exploring Energy Profile Correlations in Remotely Homologous Proteins.**

**3.1 Assessing Structural Similarities Using Energy Profiles**

In this section, we delve into the investigation of the efficacy of energy profiles in quantifying structural similarities among protein structures. To achieve this, we employed the Manhattan distance to measure the dissimilarity between two energy profiles. Our analysis was conducted on a comprehensive benchmark dataset, comprising a total of 260 protein domains downloaded from the CATH v4.2.0 database. The protein domains in this dataset varied in length, spanning from 44 to 854 residues on average [19]. This diverse dataset encompassed two distinct superfamilies: the C-terminal domain of the DNA helicase RuvA subunit (CATH Code: 1.10.8.10), falling under the Alpha class, characterized by an Orthogonal Bundle Architecture, Helicase, and Ruva Protein fold; and the Homing endonucleases (CATH Code: 3.10.28.10), situated within the Alpha and Beta class, featuring the Roll Architecture, and associated with the Endonuclease I-creI fold. Notably, within the CATH protein database, these two superfamilies comprised 161 and 97 structures at the first level (Mainly-alpha class) and the third level (Alpha-Beta class), respectively.

Our results revealed that energy profiles effectively discriminate between these two superfamilies, as vividly depicted in the UMAP plot presented in Figure 1. Additionally, in our comparative analysis, we not only relied on the Manhattan distance and energy profiles but also compared our findings with those obtained using the Yau-Hausdorff distance [19], GR-Align, RMSD, and TM-score. The Yau-Hausdorff distance offers a method for assessing structural similarities between three-dimensional objects through coordinated rotations of their respective coordinate matrices. Our exploration encompassed the calculation of pairwise distances among the 260 protein domains, employing the (3D) Yau-Hausdorff method with two distinct rotation parameters: 10 and 2500, subsequently generating a distance matrix. This matrix served as the basis for classifying proteins using the k-nearest neighbor (k-NN) method, with k set to 1.

To comprehensively evaluate the performance of the methods employed, we computed accuracy metrics and conducted comparative assessments on a system with specifications of 2.4GHz and 8GB RAM. Particularly, we calculated the 1-NN accuracy rate, which involves counting the number of proteins classified in the same class as their nearest neighbors in the reference classification. Notably, both the energy profile calculation and the 1-NN algorithm were implemented efficiently within approximately 10 minutes on a system with a 2.4 GHz processor and 4GB RAM. Impressively, our methodology achieved a classification accuracy of 98% for distinguishing between the two superfamilies. Tables 1 and 2 provide a detailed breakdown of the accuracy results and the corresponding confusion matrix, respectively.

Chart, scatter chart

Description automatically generated

*Figure 2: Scatter plot showing regression line of RMSD of each pair of structures against the distance of each pair of energy profiles for Ct and Ho families. Spearman's Rank correlation coefficient is shown in the plot.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | GR-Align | RMSD | TM-Score | YH (10Rotation) | YH (2500Rotation) | STEP | SQEP |
| Accuracy  Time | 62.3%  2 min | 59.2%  1 h | 61.5%  9h 20 min | 70.8%  10 min | 81.5%  4h 10 min | 98%  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** | **2 , 6** | **0.012 , 0.037** |
| **Homing** | **3 , 4** | **94 , 93** | **0.030 , 0.041** |

*Table 2: Confusion matrix by 1-NN for energetic profile 250-D and energetic profile 210-D (SQEP).*

**3.3 Classification of SCOP Superfamilies Using Energetic Profiles**

In this section, we extend our analysis to classify proteins into five distinct SCOP superfamilies: (why we used these superfamilies and add more numbers related to the number of proteins and protein length,…) winged helix, PH domain-like, NTF-like, Ubiquitin-like, and Immunoglobulins. Our classification approach leverages energetic profiles as a critical feature, with the Support Vector Machine (SVM) and Random Forest (RF) classifiers serving as our classification models. To ensure the reliability and generalization capability of our models, we subject them to rigorous 10-fold cross-validation. The results of this cross-validation are summarized in Table 4, which presents both accuracy and F1-measure (the harmonic mean of precision and recall).

Our findings from the 10-fold cross-validation demonstrate the robust performance of the SVM classifier, achieving an impressive accuracy of 96.8% across the five superfamilies. This high accuracy indicates the SVM algorithm's efficacy in accurately classifying remotely homologous proteins, affirming the validity and reliability of our classification model.

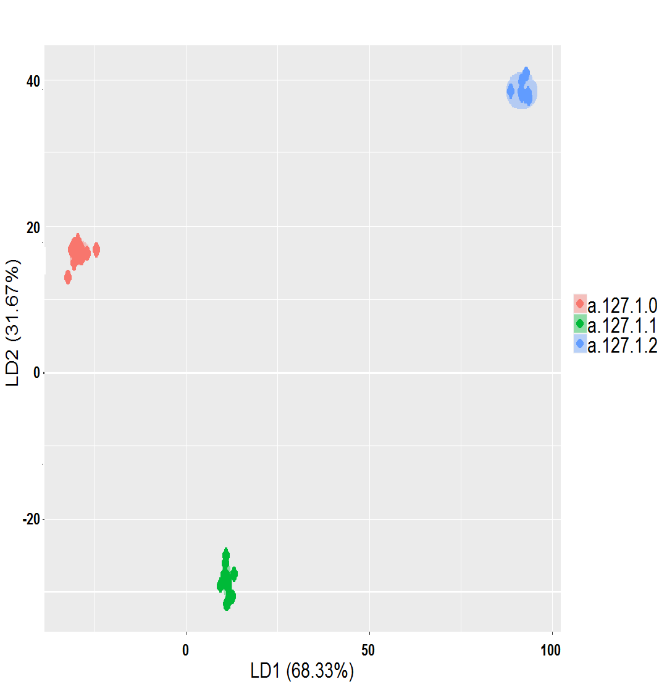
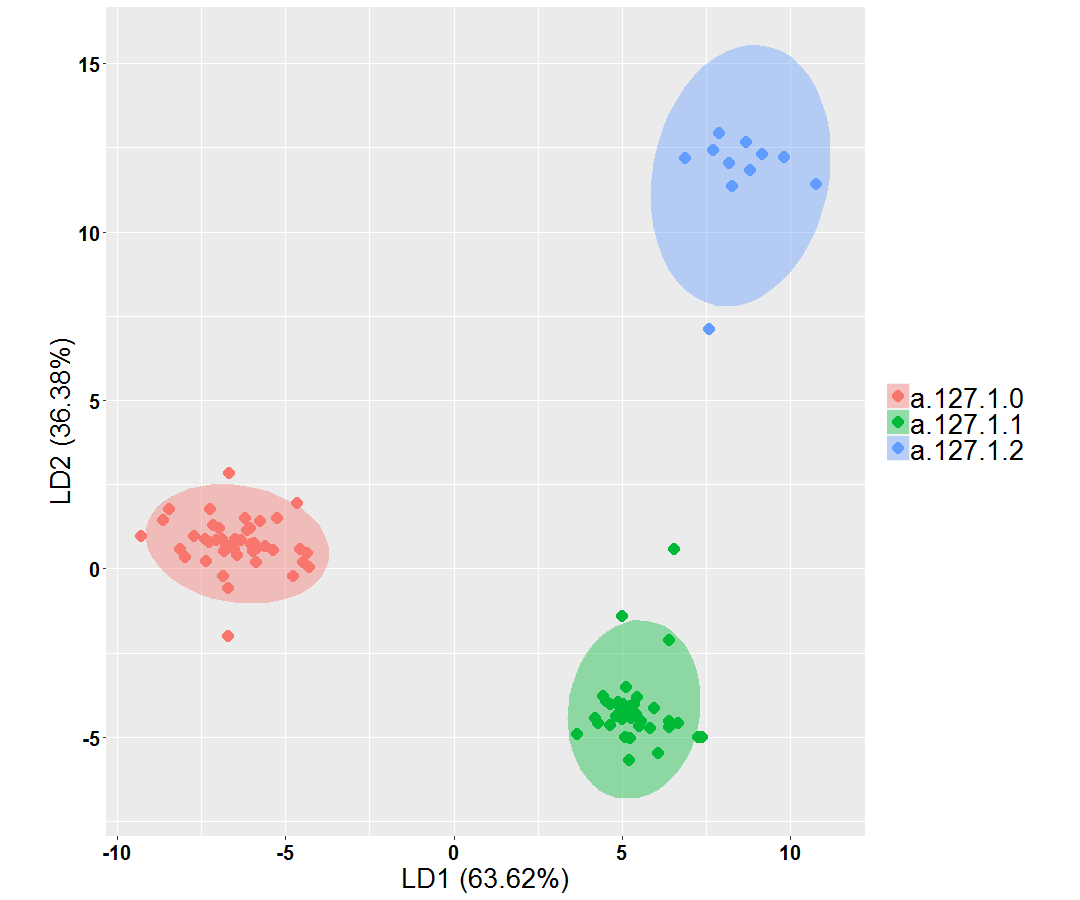
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Total-accuracy | F1\_wigend\_helix | F1\_PH.domain-like | F1\_NTF-like | F1\_Ubiquitin-like | F1\_Immunoglobulins |
| Fold01 | 0.977 | 0.963 | 0.85 | 0.92 | 0.969 | 0.992 |
| Fold02 | 0.956 | 0.935 | 0.647 | 0.963 | 0.921 | 0.981 |
| 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.*

**3.4 Classification of SCOP Superfamily: Aspartase Superfamily**

In this section, we delve into the classification of proteins within the Aspartase superfamily, which belongs to the all-alpha class of proteins in the SCOP database. A superfamily in the SCOP database encompasses a group of protein families that share a common evolutionary origin, characterized by related functions and sequence or structural similarities. The accurate classification of protein families within a superfamily is of paramount importance. For our analysis, we focused on the Aspartase superfamily, which is known for its role in catalyzing the reaction between aspartic acid and fumaric acid, producing fumaric acid and ammonium ions through a specialized mechanism. Notably, aspartases from various organisms exhibit high sequence homology and are functionally related enzymes [20]. This superfamily consists of three families with the IDs a.127.1.0, a.127.1.1, and a.127.1.2, which have 48, 40, and 11 proteins, respectively. Figure 3 visually presents a UMAP representation of the energy profiles of the proteins within this dataset. This visualization unequivocally demonstrates the effective separation of proteins belonging to the three families. The evident distinctiveness of energy profiles highlights the invaluable utility of such profiles in the precise classification of proteins at family level. Furthermore, it underscores the broader potential of energy profiles in superfamily classification across diverse protein families.

A B



*Figure 3: Two-dimensional representation of Three family in aspartase superfamily (SCOP) by LDA method with energetic profile 210-dimensional PCEP (A) and 250-dimensional SQEP (B).*

**3.6 Phylogeny Inference Using Energy Profiles**

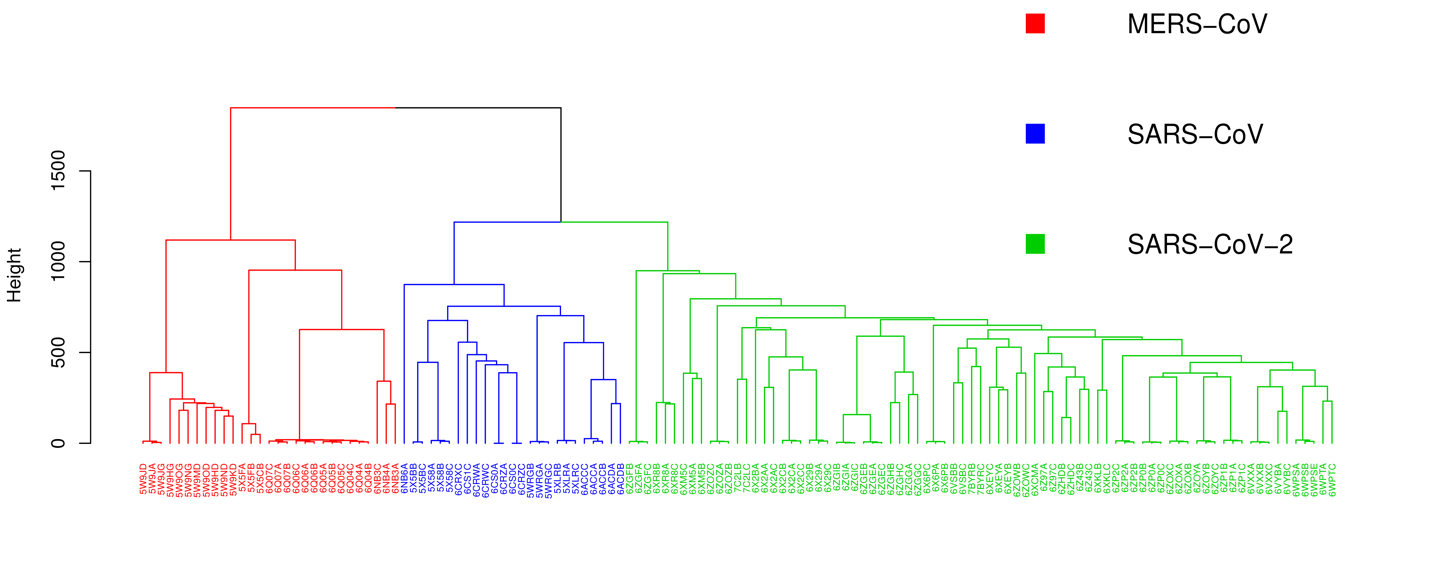
In the realm of structural biology and evolutionary analysis, three-dimensional protein structure classification and the alignment of multiple sequences stand as formidable tools for uncovering structural similarities and deducing phylogenetic relationships. A phylogeny, often visualized as a tree, serves 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 introduce our novel approach leveraging 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**

The release of SARS-CoV-2 protein structures in the Protein Data Bank (PDB) has been instrumental in advancing our understanding of this novel coronavirus. These structures encompass an array of vital viral components, including 28 spike glycoprotein structures, over 150 main protease structures, and over 60 structures of other critical SARS-CoV-2 proteins, marking a valuable resource. These high-resolution protein structures play a pivotal role in unraveling the intricacies of viral assembly and serve as indispensable tools for rational vaccine and therapeutic development. Among the arsenal of viral proteins, the spike glycoprotein, a transmembrane glycoprotein, takes center stage. This protein orchestrates viral infection by facilitating host receptor binding and serves as the primary target for neutralizing antibodies and vaccine design efforts. To comprehensively explore the structural landscape of these spike glycoproteins and gain insights into their evolutionary relationships, we harnessed the resources of the CoV3D database (https://cov3d.ibbr.umd.edu), a comprehensive repository housing a wealth of coronavirus protein structures and their intricate complexes with antibodies, receptors, and small molecules. From the CoV3D database, we extracted a dataset comprising 133 spike glycoprotein structures distinguished by the presence of the closed receptor binding domain (RBD) within their structure. This dataset encompasses 79 chains from SARS-CoV-2, 25 chains from SARS-CoV, and 29 chains from MERS-CoV. To investigate the structural divergence and relationships among these spike glycoproteins, we employed a 210-dimensional Profile of Cumulative Energy Profiles (PCEP) analysis. Calculating Manhattan distances between all pairs of energetic profiles, we effectively segregated 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, providing a visually informative representation of the structural relationships within this protein family (see Figure 7).

In a parallel study, we conducted hierarchical clustering of spike protein sequences from the SARS-CoV, MERS-CoV, and SARS-CoV-2 viruses. This clustering was grounded in the distances between their energetic profile similarities, characterized by a 210-dimensional Structural Quality Energy Profile (SQEP). The resulting dendrogram visualization unveiled the intricate evolutionary relationships among spike proteins within these three viruses, offering complementary insights into the dynamic evolution of this crucial viral component, as elucidated in previous research endeavors.



*Figure 7: Clustering of coronavirus spike glycoprotein structures. The dendrogram of was generated based on pairwise the energetic profile 210-dimensional PCEP similarities between 133 spike glycoprotein chain structures in R (*[*www.r-project.org*](http://www.r-project.org)*). leaves labels are the pdb-IDs of the chains.*



*Figure 8: Clustering of coronavirus spike glycoprotein structures. The dendrogram of was generated based on pairwise the energetic profile 210-dimensional EQEP similarities between 9 spike protein sequences in.*

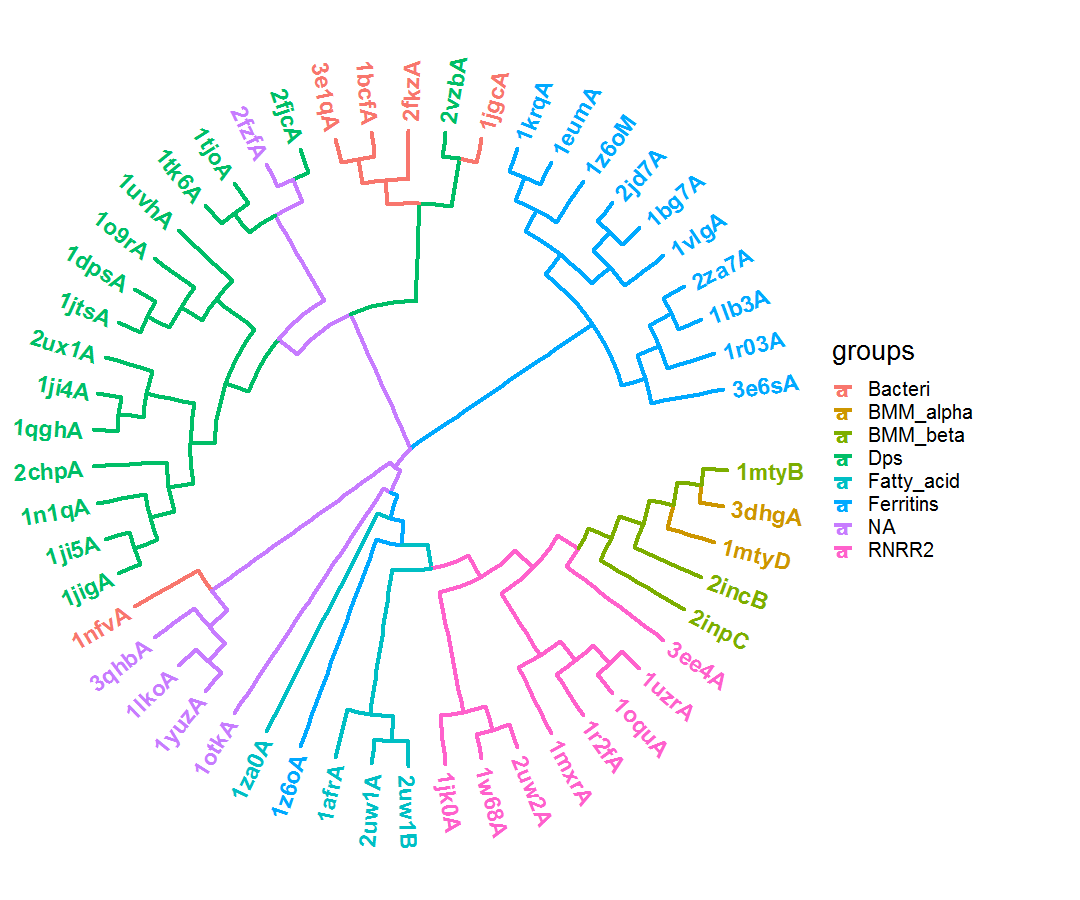
*Accession id; 2019-nCoV[Wuhan-Hu-1]: QHD43416.1, 2019-nCoV[HKU-SZ-005b]: QHN73810.1, 2019-nCoV[HKU-SZ-002a]: QHN73795.1, SARS-CoV[GZ02]: AAS00003.1, SARS-CoV[GD01]: AAP51227.1, SARS-CoV[Sino1-11]: AAR23586.1, SARS-CoV[GZ02]: AAS00003.1, SARS-CoV[NS-1]: AAR91586.1, MERS-CoV[ChinaGD01]: AKJ80137.2, MERS-CoV[2014KSA\_683]: AID55073.1.*

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

Our methodology was further applied to the expansive ferritin-like superfamily, a superfamily whose structural phylogeny has been previously reported (20). This superfamily is of particular interest due to its striking characteristics—despite exhibiting minimal sequence similarity and substantial differences in quaternary structure and function across its constituent members, ferritin-like proteins maintain a conserved structural core (20). Notably, this superfamily encompasses a diverse array 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. Lundin et al. previously delved into the phylogeny of this superfamily, constructing a phylogenetic network using the distance-based NeighborNet network method, grounded in distance calculated through structure-based alignment methods [21](20). Moreover, Malik et al. [22] (21) conducted a study in 2020 to assess whether phylogenies derived from pairwise structural comparisons are influenced by protein length and shape disparities. Their findings indicated that structural phylogenetics thrives when structures exhibit remarkably similar lengths, while our method remains robust, impervious to protein length variations.

To explore the phylogenetic relationships within this superfamily, we employed the neighbor-joining method [23](22) to reconstruct the phylogenetic tree, harnessing both Profile of Cumulative Energy Profiles (PCEP) and Structural Quality Energy Profile (SQEP) data (Figure 9). We calculated Manhattan distances between all pairs of energetic profiles encompassing 250-dimensional PCEP, ASEP, and DAEP for the superfamily proteins, subsequently utilizing this distance matrix to inform our tree-building process.

Our qualitative analysis provides compelling support for the separation of two SCOP families within this superfamily, namely ferritin (a.25.1.1) and RNR-like (a.25.1.2), as well as their subdivision into distinct groups within our phylogenetic tree. These groupings closely align with prior studies conducted by Lundin et al. [21] and Malik et al.[22] in 2020, further affirming the robustness of our approach (see Figure 8). Figure 8 visually illustrates that the overarching classification delineated in SCOP, CATH, and Pfam (http://pfam.xfam.org/) is faithfully recapitulated through our energy-based phylogenetic analysis.

****

*Figure 9: Energy-based phylogenetic of the ferritin-like superfamily. The two large SCOP families, ferritins (a.25.1.1; Bacteri, Ferritins, Dps and NA) and ribonucleotide reductase-like (a.25.1.2; BMM\_alpha, BMM\_beta, Fatty\_acid and RNRR2) separated in the tree with smaller groups.*

**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.