Energetic Profile-Based Protein Comparison: A New and Fast Approach for Structural and Evolutionary analysis

Peyman Choopanian[1, 2], Jaan-Olle Andressoo[1, 2, 3*], and Mehdi Mirzaie[1, 2*]

[1]Translational Neuroscience, Department of Pharmacology, Faculty of Medicine and Helsinki Institute of Life Science, 00014 University of Helsinki, Finland

[2]Department of Pharmacology, Faculty of Medicine, 00014 University of Helsinki, Finland

[3]Division of Neurogeriatrics, Department of Neurobiology, Care Sciences and Society (NVS), 17177 Karolinska Institutet, Sweden

Abstract

In structural bioinformatics, the efficiency of predicting protein similarity, function, and evolutionary relationships is crucial. Our innovative approach leverages protein energy profiles derived from a knowledge-based potential, deviating from traditional methods relying on structural alignment or atomic distances. This method assigns unique energy profiles to individual proteins, facilitating rapid comparative analysis for both structural similarities and evolutionary relationships across various hierarchical levels. Our study demonstrates that energy profiles contain substantial information about protein structure at class, fold, superfamily, and family levels. Notably, these profiles accurately distinguish proteins across species, illustrated by the classification of coronavirus spike glycoproteins and bacteriocin proteins. Introducing a novel separation measure based on energy profile similarity, our method shows significant correlation with a network-based approach, emphasizing the potential of energy profiles as efficient predictors for drug combinations with faster computational requirements. Our key insight is that the sequence-based energy profile strongly correlates with structure-derived energy, enabling rapid and efficient protein comparisons based solely on sequences.

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

Introduction

A thorough understanding of protein function holds paramount importance within the domains of biology, medicine, and pharmacy. While experimental methods exhibit high accuracy in protein function associations, their inherent limitations, such as being time-intensive and expensive, have instigated the exploration of computational alternatives. The evaluation of protein similarity by comparing two proteins has consistently emerged as a key methodology. This assessment plays a pivotal role in uncovering insights into the functions and evolutionary relationships of proteins. Advances in high-throughput technologies have led to the establishment of extensive repositories containing protein sequences, a substantial proportion of which, however, lack annotations [1]. The significant advancements in omics data and the evolution of machine learning techniques have propelled progress in protein research, transitioning from traditional methods like PSI-BLAST [2] to more sophisticated approaches [3]. In the realm of machine learning research, a crucial step is encoding data as input. Although there is no universal approach, encoding amino acid sequences or structural features has been widely adopted for various protein function predictions, including drug-protein interactions [4], anti-hypertensive peptides [5], and RNA-protein interactions[6]. Despite the diversity in methodologies, the underlying commonality revolves around determining protein similarity either through sequence alignment or structural comparison.

Protein structure is a fundamental feature affecting function, and activity. The energy of a protein structure plays a key role in determining its structure. Knowledge-based potentials, categorized as statistical energy functions, are derived from databases of known protein structures that empirically capture the most probable state of a protein and describe microstates of interactions within protein structure [7, 8, 9]. It is generally assuming that a native protein structure is confined to a state with the minimum total energy, and the more similar a structure is to the native state, the closer its total energy is to the native state. However, we take a step beyond this assumption and suggest a hypothesis that two similar proteins possess analogous energy profiles. To evaluate this hypothesis, we assigned an energetic feature vector to each protein, with each entry representing the summation of energies for a specific pair of amino acids. With 20 amino acids in proteins, this resulted in 210 pairwise interaction types. This is the first study to assign an energetic feature vector to each protein for comparative analysis. This 210-dimensional vector represents the intricate energy landscape inherent to the structural diversity of proteins. This vector of energies serves as the cornerstone of our analytical approach and provides a robust foundation for further investigative pursuits. Given the current issue of experimentally determining the three-dimensional structures of proteins, estimating energy based on sequence emerges as a crucial consideration. Dostari et al. [10] introduced a method to estimate energy based on amino acid composition. In our study, we drew inspiration from their approach to extract the energy profile based on protein sequence.

The stratification of proteins into distinct folds, superfamilies, and families, guided by evolutionary consanguinity or shared structural and functional attributes, is crucial for precise function prediction. Databases like CATH (Class, Architecture, Topology, Homologous superfamily) [11] and SCOP (Structural Classification Of Proteins) [12] categorize proteins into hierarchical groups based on structural feature, from

broad classifications like folds and classes to finer details such as superfamilies and families. To assess profile of energies at various levels, we utilized the ASTRAL40 (95) datasets from SCOPe as a benchmark dataset [13]. Comparing energy and distance between profiles estimated from both sequence and structure revealed a high correlation on protein domains from both ASTRAL40 and ASTRAL95 datasets. UMAP projections provided additional evidence that the profile of energy encapsulates structural information at fold, superfamily, and family levels, as observed through random selections. Our method demonstrated superior performance in terms of both accuracy and computational efficiency compared to currently available tools.

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. We also evaluated our method to reconstruct evolutionary relationships among proteins from the ferritin-like superfamily that are beyond the "twilight zone" [14] —a sequence similarity range (typically 20-35% identity) that complicates the differentiation between true homologs and random matches due to insufficient sequence conservation. 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. To assess the discriminatory capacity of energy profiles in discerning proteins across various species, we chose the spike glycoproteins from three coronavirus species [15]. Our findings indicate that both sequence-level and structural-level energy profiles successfully cluster proteins from distinct species. In a separate analysis, we computed the sequence-based energy profile for a diverse set of bacterial protein families known as bacteriocins. The identification and understanding of these peptides are crucial due to their ecological significance, but their diverse sequences and structures present challenges for conventional identification methods. The BAGEL data set includes 689 proteins[16], each with a length greater than 30 amino acids, providing a comprehensive and challenging benchmark for evaluating peptide identification techniques. Our findings highlight that the energy profile can categorize these proteins based on BAGEL annotation, demonstrating the effectiveness of our method in handling the complexity and diversity inherent in bacteriocin sequences.

The identification of effective drug combinations, essential for treating complex diseases, face challenges due to the combinatorial explosion of potential drug pairs. Cheng et al. introduced a network-based methodology leveraging the human protein-protein interactome to discover clinically effective drug combinations, demonstrating that topological relationships between drug-target modules, as indicated by a separation measure, reflect both biological and pharmacological relationships [17]. In our study, we introduce a separation measure based on the similarity between profile of energies of protein targets, revealing a significant correlation with the separation measure derived from the protein-protein interaction network. This suggests that the profile of energy holds promise as a reliable predictor for drug combinations, requiring only protein sequences and offering quicker computation compared to network-based approaches. Therefore, this study offers a means to characterize and compare proteins using profile of energies, enabling predictions of their structural and functional properties.

Results

Knowledge-based potentials are derived from databases of known protein structures. Various potential functions, such as distance-dependent, dihedral angles, and accessible surface energies leverage information from known protein structures to estimate energies of pairwise interactions [8]. In our investigation, we employed the potential function, where atom contacts were identified using the tessellation method, as outlined in the method section [18]. For each protein structure, contacts between atoms were determined through the tessellation method, and the energy for each pair of amino acid types was estimated using equation (2) in the Method section. With 20 amino acids present in proteins, this process resulted in 210 pairwise interaction types. This 210-dimensional vector represents the energy landscape inherent in the structural diversity of proteins. For each pair of proteins, the Manhattan distance between the profiles of energies is considered a measure of dissimilarity between them. Given the current issue of experimentally determining the three-dimensional structures of proteins, estimating energy based on sequence pertains as a crucial subject. We utilized equation (5) to construct the profile of energy based on sequence.

Correlation between Energy estimated based on structure and Sequence.

To examine the profile of energy at various levels of SCOP, we employed the ASTRAL40 (95) database (version 2.08) from SCOPe as a benchmark dataset, comprising domains with no more than 40% (95%) sequence similarity, as determined by BLAST identity, and filtered for E-value similarity scores [13]. This dataset offers a comprehensive description of structural and evolutionary relationships among proteins from the Protein Data Bank. At first, we calculated energies for protein domains in the ASTRAL40 and ASTRAL95 datasets using both structure- and sequence-based methods. Fig. 1A depicts the relationship between the total energy derived from the structure (on the y-axis) and from the sequence (on the x-axis), with the ASTRAL40 on the left and ASTRAL95 on the right side of the figure. The observed high correlation coefficient suggests that sequence-based energy estimation serves as a reliable approximation and can be effectively used in scenarios where the protein structure is unidentified.

For every pair of domains within the ASTRAL40 (ASTRAL95) datasets, the distances between their profile of energy were computed utilizing both structural and sequence-based energy estimation. In Fig. 1B, the x-axis denotes the distance between Compositional Profile of Energies (CPE), while the y-axis represents the distance between Structural Profile of Energies (SPE)(for more details see the method section). The figure reveals a strong correlation between the distances estimated through structural and sequence-based approaches. Hence, the energy estimation based on sequence data is deemed sufficiently reliable.

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, Fig. 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 (Fig. 2B).

3.2 Unveiling the Energy Patterns Across SCOP Hierarchy

We visualized energy profiles derived from sequence and structure for domains within the all-alpha and all-beta classes. As shown in Fig. 3, UMAP embeddings effectively capture structural characteristics distinguishing all-alpha and all-beta domains. This visualization reveals distinct energy patterns between these classes, a consistency also found in sequence-based analyses. To explore structural information at lower hierarchical levels of SCOP, two folds (a.100 and a.104) from the all-alpha class, two superfamilies (a.29.2 and a.29.3) from fold a.29, and two families (a.25.1.0 and a.25.1.2) from superfamily a.25.1 were randomly selected. Fig. 4 displays two figures per panel, with the left figure illustrating CPE profiles and the right figure showcasing SPE profiles. UMAP plots in Fig. 4 demonstrate that protein domains within the same fold, superfamily, or family share similar energy patterns and cluster together.

To delve deeper into differences in distances among protein domains within the same class, we calculated pairwise distances for domains within the all-alpha class from the ASTRAL95 dataset. Subsequently, these distances were compared with distances from domains across different classes. As shown in Fig. 5A-B, intraclass distances in purple are significantly lower than interclass distances in yellow. Similar results were obtained when calculating pairwise distances from domains within fold a.29 and comparing them with distances from domains in different folds within the all-alpha class. Likewise, distances between energy patterns of domains within the same superfamily a.29.3 are significantly less than distances between energy patterns of domains within fold a.29 that belong to different superfamilies (Fig. 5C-D). Consequently, it can be inferred that energy patterns of domains belonging to the same superfamily/fold/class exhibit higher similarity than those from different superfamilies/folds/classes.

It is commonly assumed that proteins sharing similar structures also exhibit similar functions. Several measurements have been developed to assess protein structure similarity, each offering unique insights. Root Mean Square Deviation (RMSD)[19] quantifies average spatial variance between corresponding atoms or components within superimposed proteins, providing a fundamental measure of structural deviation. The TM-score (Template Modeling score) [20] evaluates similarity by considering both residue-level alignment and overall topology, offering a nuanced assessment of structural resemblance. TM-Vec [21], a recent advancement, employs deep learning techniques trained on diverse protein structures to enhance accuracy and efficiency in similarity assessment. On the alignment front, GR-align[22] stands out, likely utilizing geometric reasoning for accurate structural alignment. Its robustness to structural variations makes it invaluable in structural bioinformatics. Additionally, the Hausdorff distance[23] provides a measure of dissimilarity between sets of points, offering further insight into structural comparisons. Here, we employed a benchmark dataset sourced from the CATH v4.2.0 database, comprising 251 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 used the 1-NN classification method to categorize proteins based on GR-Align, RMSD, TM-score, Yau-Hausdorff distance, TM-Vec, and the distance between energy profiles as a measure of protein dissimilarity. As shown in Table 1 and Fig. 6, our method achieves a computation time faster than TM-Vec. Moreover, our method significantly outperforms GR-Align, RMSD, TM-Score, and YH in terms of accuracy and efficiency, highlighting a substantial advantage. Our method eliminates the need for superimposing protein structures or conducting structural alignments; instead, we calculate energy profiles and measure the distance between them. Table 1 details result and processing times, demonstrating the efficient implementation of CPE calculation and the 1-NN algorithm, completed in about 3 minutes on a system with a 2.4 GHz processor and 4GB RAM. Our methodology achieved a remarkable classification accuracy of 97% in distinguishing between two protein families.

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 Immunoglobulins (b.1.1) [24]. Our classification strategy incorporated energetic profiles CPE as features, employing 1-nearest neighbor (1-NN) and Random Forest (RF) classifiers as our models. To ensure the robustness and generalization of our models, we subjected RF to rigorous 10-fold cross-validation. The results, summarized in Table 2, include metrics for accuracy and F1-score, demonstrating the effectiveness of our model. Both classifiers show performance levels close to 100%, as illustrated in Table 2. We compared the CPE method with TM-Vec. As depicted in Table 2, our results (CPE) are not only comparable to TM-Vec in terms of accuracy but also demonstrate a faster performance.

3.4 Phylogeny Inference of the Ferritin-Like Superfamily

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 long evolutionary distances. Lundin 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. Lundin et al. [14] 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 Ribonucleotide reductase-like (RNR R2), rubrerythrins, bacterioferritins, Dps (DNA binding protein from starved cells that protects against oxidative DNA damage),

and Dps-like proteins. As discussed by Lundin et al.[14] at the superfamily level, the classification of the "ferritin-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. Malik et al. [25], and Puente-Lelievre et al. [26] delved into the evolutionary relationships of this superfamily by creating a phylogenetic network.

They employed the distance-based NeighborNet network method [27], utilizing distances calculated through structure-based alignment methods. . Fig. 7A depicts the schematic tree built by Malik et al.[25], and Lelievre et al. [26]. We employed the same protein structures within this superfamily as utilized by Malik et al.[25] and Lelievre et al. [26] to reconstruct the phylogeny of based on profile of energies. 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 Fatty acids [14]. Following this, we computed SPE for each protein and calculated all pairwise distances between SPEs. The phylogenetic tree, constructed using the phangorn package [28], was visualized through the SplitTree software [29] and is presented in Fig. 7B.

Our results suggest that the energetic phylogenies within the ferritin-like superfamily unveil significant relationships among its members, aligning with known evolutionary relationships 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. Thus, our methodology accurately bifurcates this superfamily into two families. Delving into specifics, the family a.25.1.1 (depicted by orange color triangles) further divides into four subgroups: "ferritins", "Dps", "Rubrerythrin", and "Bacterioferritins" indicated by distinct colors in Fig. 7B. On the other hand, the second branch related to the a.25.1.2 family (dark blue triangles), 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. The protein groupings presented by Lelievre et al. in Fig. 7A are color-coded, corresponding to the colors used in Fig. 7B, C. Our approach successfully reconstructed the phylogenetic tree using the energy profile. However, as shown in Fig.7C, the phylogenetic tree generated by the TM-Vec representation and cosine similarity could delineate two distinct branches corresponding to two protein families but failed to predict the evolutionary relationships within each protein family as proposed by Lelievre et al. The dashed line in Fig. 7B-C

demonstrates that both the energy model and the vector model effectively distinguish between the Ferritin and Ribonucleotide reductase-like families.

In Fig. 7B, the energy profile model accurately orders the divergence of proteins within the Ferritin family, following Lelievre et al.'s order of rubrerythrins, Ferritins, and then Dps and bacterioferritins. Conversely, for the Ribonucleotide reductase-like family, the energy profile model reconstructs the proposed evolutionary order of Fatty acids, RNR R2, and then BMM. In contrast, the phylogenetic tree reconstructed using the TM-Vec model, while capable of differentiating the two families, does not align with the evolutionary order suggested by Lelievre et al. For example, within the Ferritin family, Lelievre et al.'s model posits that Dps diverged later, whereas the TM-Vec model indicates it as the first group to separate. Similarly, for the Ribonucleotide reductase-like family, the TM-Vec model places the BMM proteins as the earliest branch in the phylogenetic tree, whereas Lelievre et al.'s model suggests they were among the last to diverge.

3.5 Clustering of the SARS-CoV-2, SARS-CoV and 2012 MERS-CoV 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, 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 [15].

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 and evolutionary relationships within this protein family. As depicted in Fig. 8A-B, the lineage of SARS-CoV and SARS-CoV-2 is clearly distinguished from that of MERS-CoV, which belongs to a distinct subgenus within the family of coronaviruses.

As shown in Fig. 8A-D, our methods, CPE and SPE, demonstrate superior results in clustering proteins from different viruses. Based on the evolutionary history of these viruses, it is known that MERS-CoV

emerged first, followed by SARS-CoV, and then SARS-CoV-2. CPE and SPE accurately reveal this evolutionary pattern. However, Fig. 8C and 8D reveal that RMSD and TM-Vec are not effective in accurately reconstructing this pattern for certain proteins. As highlighted by the orange circles in Fig. 8C, some SARS-CoV-2 proteins are positioned far from their respective groups. Similarly, SARS-CoV proteins in the tree generated by the TM-Vec method are also mis grouped, as indicated by the orange circle. Additionally, this method failed to correctly group proteins from MERS-CoV. In addition to clustering performance metrics, our methods, CPE and SPE, are significantly more efficient in time of computation, with CPE taking only 0.9 seconds and SPE requiring just 3 minutes. In comparison, TM-Vec takes 89 seconds, RMSD takes 70 minutes, and TM-score requires a substantial 9.7 hours. This underscores the computational advantage of our methods, making them both effective and efficient for clustering analyses. To compare the clustering results of our methods with those obtained from RMSD, TM-score and TM-Vec, we used the Adjusted Rand Index (ARI). The ARI measures the similarity between two clusterings, accounting for chance agreement, with values ranging from -1 to 1, where 1 indicates perfect agreement. Fig. 8E and Table S1 show that with a cut tree of 4, our sequence-based methods achieved the highest clustering performance, with an ARI of 0.95. TM-Vec's optimal result was at a cut tree of 5, achieving an ARI of 0.87. For structure-based methods, our SPE method achieved a perfect ARI of 1 with a cut tree of 3. RMSD's best performance was at a cut tree of 6 with an ARI of 0.73, while TM-score performed best at a cut tree of 4, with an ARI of 0.56. These results, summarized in Table 3, underscore the effectiveness of our proposed methods in both sequence and structure-based clustering analyses. Fig. 8F highlights that our methods are significantly faster than others. The list of PDBIDs of the spike proteins used in this analysis is summarized in the table S2 and the result of TM-score is presented in **Fig. S1**.

3.6. Clustering of Bacteriocins

Bacteriocins are peptides produced by bacteria that act as strong antibacterial agents against other, typically closely related microbial species. We analyzed the bacteriocins family available in the BAGEL database, those with a length larger than 30 amino acids, including a total of 689 proteins [16]. 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 Fig. 9A, our analysis revealed that profile of energy (CPE) can clearly partition bacteriocins according to BAGEL annotation. Hamamsy et al. [30] leveraged the deep protein language models to develop the TM-Vec model, which is trained on pairs of protein sequences and their TM-scores. We compared CPE distances to the TM-scores of protein structures predicted by AlphaFold2 [31], OmegaFold [32], and ESMFold [33], as well as the TM-Vec predicted by the TM-Vec model. As demonstrated in Fig. 9B, the TM-score of proteins predicted by AlphaFold2, OmegaFold, and ESMFold from the same

class is similar to proteins from different classes. TM-Vec is effective at distinguishing between bacteriocins from the same class and proteins from different classes. Although there is some overlap between TM-Vec values from proteins from the same class and other classes. Our method also effectively distinguishes between proteins from the same class and those from other classes in bacteriocin dataset.

3.7. Effective Drug Combination suggestion using Energetic Signatures

The identification and validation of effective drug combinations are crucial in the treatment of various complex diseases, aiming to enhance therapeutic efficacy while minimizing toxicity. However, this task is hindered by a combinatorial explosion resulting from the multitude of potential drug pairs. Cheng et al. introduced a network-based methodology to pinpoint clinically effective drug combinations tailored to specific diseases [17]. This approach involved assessing the network-based relationships among drug targets and disease proteins within the human protein-protein interactome. By quantifying these relationships, they identified clusters of drugs that exhibited correlations with therapeutic effects. The drugs within these clusters targeted the same disease module but belonged to separate neighborhoods. This innovative network methodology presented by Cheng et al. provides a generic and powerful means to discover effective combination therapies during drug development. Disease proteins were observed to form localized neighborhoods, referred to as disease modules, rather than being randomly distributed throughout the interactome. To characterize the mutual relationship between two drugs and a disease module, they employed the following network-based proximity measure:

$$s_{AB} = < d_{AB} > - \frac{< d_{AA} > + < d_{BB} >}{2} \qquad (5)$$

This measure assessed the network proximity of drug-target modules A and B by comparing the mean shortest distance within the interactome between the targets of each drug ($< d_{AA} >$ and $< d_{BB} >$) to the mean shortest distance between A-B target pairs $< d_{AB} >$. When $s_{AB} < 0$, the targets of the two drugs are in the same network neighborhood; when $s_{AB} > 0$, the two targets are topologically separated.

The authors demonstrated that the topological relationship between two drug-target modules, as indicated by s_{AB}, reflects both biological and pharmacological relationships. They also showed that the network proximity (s_{AB}) of drug-drug pairs in the human interactome correlates with chemical, biological, functional, and clinical similarities. This led them to conclude that each drug-target module possesses a well-defined network-based footprint. If the footprints of two drug-target modules are topologically separated, the drugs are considered pharmacologically distinct. Conversely, if the footprints overlap, the magnitude of the overlap indicates the strength of their pharmacological relationship. A closer network proximity of targets in a drug pair suggests higher similarities in their chemical, biological, functional, and clinical profiles.

Here, we used the following separation measure, denoted by E_{AB}, based on similarity between profiles of energies of protein targets:

$$E_{AB} = < d_{AB} > - \frac{< d_{AA} > + < d_{BB} >}{2} \qquad (6)$$

where

$$< d_{AB} > = \frac{1}{\|A\| + \|B\|}\left(\sum_{a \in A} \min_{b \in B} d(a,b) + \sum_{b \in B} \min_{a \in A} d(a,b)\right)$$

and $d(a,b)$ represents the Manhattan distance between the energy profiles of proteins a and b.

Fig. 10 depicts the correlation between s_{AB} values, as computed by Cheng et al. [17], for a set of 65 antihypertensive drugs exhibiting complementary exposure to the hypertension disease module, and the corresponding E_{AB}. The results demonstrate a strong correlation between s_{AB} and E_{AB}, suggesting that the energy profile holds promise for predicting drug combinations. It is important that our approach only requires protein sequences and is significantly faster than computing the shortest path in a protein-protein interaction network.

Discussion

The continuous growth of protein databases highlights the importance of understanding their functional characteristics. It's widely recognized that proteins with similar structures often perform similar functions. Additionally, there's a common belief that proteins with similar structures also share similar energy levels. Therefore, our study aims to pioneer a new approach by directly linking protein energy landscapes to their functional attributes. By investigating this relationship, we seek to uncover new insights into how protein structure, energetics, and biological activity are interconnected. Knowledge-based potentials are energy functions derived from known protein structures. In our study, we used the DBNI potential function[18] to calculate the energy between pairs of amino acids, generating energy profiles based on both sequence and three-dimensional structure. A significant achievement of our study is the high correlation observed between energy estimates derived from sequence and those from structural data, allowing for the derivation of energy profiles based solely on sequence information, which enables fast and accurate computational analysis. However, it's worth noting that the reliance on knowledge-based potentials is dependent on known protein structures, potentially limiting the generalizability of results to proteins with varied structural characteristics or those are underrepresented in existing databases. Furthermore, despite the promising correlation between energy estimates derived from sequence and structural data, it is possible that there are complexities in accurately capturing the entirety of protein energetics solely from sequence information, which could affect the reliability of the resulting energy profiles. To address these issues, one possible option is to adjust the energy profile, such as through reweighting, to specific applications, such as protein remote homology detection or drug-target affinity prediction.

We employed Uniform Manifold Approximation and Projection (UMAP) to visualize energy profiles at both sequence and structural levels derived from protein domains within the ASTRAL database, revealing their capacity to distinguish proteins across various hierarchical levels, including class, fold, superfamily, and family. Notably, the Manhattan distance between energy profiles serves as a measure of dissimilarity, eliminating the necessity for structural or sequence alignment in protein comparison and resulting in significantly faster computational analyses, as demonstrated in Table 2. The comparison table highlights

notable differences in both accuracy and computational efficiency among the methods evaluated. The profile of energy (CPE) method demonstrates a remarkable accuracy of 97%, significantly surpassing other methods such as GR-Align, RMSD, and TM-Score, which range from 59.2% to 81.5%. This indicates that the CPE method excels in accurately distinguishing between protein structures at different superfamilies, showcasing its superiority in capturing structural dissimilarities effectively. In terms of computational efficiency, the CPE method stands out as the most time-efficient, requiring a mere 3 minutes for processing. In contrast, traditional methods like RMSD and TM-Score demand significantly longer computational times, ranging from 1 hour to over 9 hours. For instance, the CPE method is approximately 20 times faster than RMSD and 180 times faster than TM-Score. This stark difference underscores the efficiency of the CPE method, particularly in time-sensitive scenarios or large-scale protein structure comparison tasks.

Our method's efficacy was further assessed by comparing its results with structural dissimilarity metrics such as RMSD, TM-Score, and GR-align in classifying proteins across five distinct SCOP superfamilies, showcasing its superior accuracy and computational efficiency. Particularly challenging is elucidating evolutionary relationships among superfamilies beyond the "twilight zone," where sequence similarity alone proves inadequate for meaningful analysis. To address this, we examined energy profiles to reconstruct a phylogenetic network of the Ferritin-like superfamily, incorporating proteins from the twilight zone. Our analysis, consistent with previous studies by Lundin et al[14] and Malik et al.[25], unveiled substantial and valuable evolutionary signal preserved within energy profiles, indicating their potential as representative indicators of protein structure. Moreover, we examined the structural attributes of spike glycoproteins among three coronaviruses—SARS-CoV, MERS-CoV, and SARS-CoV-2—using a 210-dimensional energy profile combined with Manhattan distances. This study successfully grouped these proteins into specific clusters corresponding to each virus, offering insights into their structural and evolutionary relationships. Additionally, our inquiry extended to 689 proteins within the bacteriocins family, encompassing various sizes and stability levels sourced from the BAGEL database. By employing the energy profile (CPE), we effectively distinguished bacteriocins according to BAGEL classifications, showcasing the usefulness of this method in protein classification, particularly in scenarios where proteins exhibit differing stabilities. Comparative analysis involving TM-scores from a range of prediction models emphasized the effectiveness of our approach in differentiating proteins within and across classes, thereby providing valuable insights into bacteriocins. In summary, our findings underscore the valuable insights offered by energy profiles across structural, functional, and evolutionary scales.

One of the significant applications of assessing protein similarity lies in quantifying the proximity between two drugs based on their protein targets. When the protein targets of two drugs exhibit similarity, it is reasonable to anticipate similarities in the drugs themselves. Our method, capable of quantifying the dissimilarity between two proteins, potentially encodes functional information that can be leveraged to gauge the similarity between two drugs according to their protein targets. Comparative analysis with a study conducted by Cheng et al.[17] demonstrates a notable correlation between our results, derived solely from protein sequence data, and theirs, obtained using protein-protein interaction data. It is worth reiterating that

our method boasts remarkable speed compared to conventional approaches. By providing a rapid yet effective means of assessing protein similarity, our method offers promising implications for drug discovery and development, facilitating the identification of potential drug candidates with similar protein targets. This underscores the significance of leveraging computational methods to expedite drug discovery processes while maintaining robustness and accuracy. In conclusion, our research introduces the energy profile as an innovative feature set containing significant functional insights that can be utilized to represent proteins within machine learning methodologies for predicting protein function, drug-target interactions, and drug combination outcomes.

In our investigation, we examined the energy profile surrounding protein drug targets and demonstrated a strong correlation between our scoring system and that derived from protein-protein interaction networks. It's important to acknowledge that while a more sophisticated computational approach and experimental validation are crucial in drug combination study, these aspects fall beyond the purview of our manuscript. Moreover, while our method bears significant implications for drug discovery and development, its efficacy might be limited by the availability and quality of protein sequence and structural data, as well as the inherent complexity of drug-target interactions. Therefore, it is imperative for independent research endeavors to address this crucial aspect and offer comprehensive insights into the practical application of our approach in real-world therapeutic contexts.

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 [34].

Pairwise distance-dependent knowledge-based potential.

Knowledge-based potentials are derived from databases of known protein structures. Various potential functions, such as distance-dependent, dihedral angles, and accessible surface energies leverage information from known protein structures to estimate energies of pairwise interactions [8]. In our study, we employed the potential function in which the contact between atoms identified using the tessellation method [18]. To obtain nearest neighbors in the protein structure, all amino acids in a protein chain were represented by heavy atoms and a Delaunay tessellation of the resulting point set was computed using Qhull [35]. Two atoms were defined to be in contact if they are two vertices of an edge in a simplex; and therefore, they are not shielded from contact by other atoms. The distance between any two atoms was divided into 30 distance shells, starting at 0.75 Å, with a distance shell equal to 0.5 Å in width to extract the knowledge-based potential. As a result, atoms separated by less than the 6 angstroms interact. There is no direct interaction between two atoms when there is a third atom between two close atoms. All pairwise occurrences outside of this range were excluded. All atom pairs in a structure were considered except those that belonged to the same residue. The study considered 167 atom types by treating all nonhydrogen atoms as having different atom types when they are in different amino acid residues.

The energy between the two atoms i and j at distance d, is calculated as follows:

$$\Delta E^{ij}(d) = RT\left[ln\left(1 + M_{ij}\sigma\right) - ln(1 + M_{ij}\sigma(\frac{f_{ij}(d)}{f_{xx}(d)}))\right] \quad (1)$$

where RT is constant and equal to 0.582 kcal/mole. M_{ij} is the number of observations for atomic pair i and j, $f_{ij}(d)$ is the relative frequency of occurrence for i and j in distance class d, $f_{xx}(d)$ is the relative frequency of occurrence for all atomic pairs in distance shell d, and σ is the weight given to each observation. As discussed by Sippl [7], it was assumed that $\sigma = 0.02$.

The potential energy associated with the interaction of residues A and B denoted by $\Delta E(A,B)$ is estimated by summing the pairwise potentials between the atoms of each of these residues as follows:

$$\Delta E(A,B) = \sum_{i\in A, j\in B} \Delta E^{ij}(d) \quad (2)$$

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

Given that there are 210 unique amino acid-amino acid interaction types among these 20 amino acids, the total number of unique $\Delta E(A,B)$ 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 specific pairs of amino acid types. We call this 210-dimensional vector as the **Structural Profile of Energy (SPE)** of a protein structure.

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. Interactions between pairs of atoms are identified using Delaunay tessellation. For each protein S in the training set, e_i^S denotes the energy of interactions between all amino acid residues of type i and all other amino acids in the protein structure S. We estimated the value of $\widehat{e_i^S}$ using the following expression:

$$\widehat{e_i^S} = N_i^S \sum_{j=1}^{20} P_{ij} n_j^S \quad (3)$$

where, N_i^S represents the frequency of amino acid type i in the structure S, and $n_j^S = \frac{N_i^S}{L}$, is calculated as the ratio of N_i^S to the total number of amino acids in S, denoted by L, and P is the energy predictor matrix, delineating the dependence of amino acid i's energy on the jth element within the amino acid composition. The parameters of the corresponding row of the matrix P are obtained by minimizing the function

$$Z_i = \sum_S (e_i^S - \widehat{e_i^S})^2 \quad (4)$$

Letting $\frac{\partial Z_i}{\partial P_{ij}} = 0$ for all P_{ij} leads to a set of linear equations which are solved for each amino acid type by using Symbolic Math Toolbox in MATLAB.

For each pair of amino acid types i and j, we used the following equation to estimate the energy E_{ij} based on amino acid sequence composition:

$$E_{ij} = n_i \sum_{j}^{20} P_{ij}\, n_j \quad (5)$$

where P is the energy predictor matrix estimated using equation 4. 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. The profile of energies is normalized based on protein length.

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 software was used to read and analyze PDB files [36]. The “geometry” package was used to implement the Quickhull algorithm to find direct contacts and nearest neighbors of atoms in pdb files using the Delaunay tessellation method (https://cran.r-project.org/web/packages/geometry/index.html). The kNN, and RF classification algorithms were implemented using the “random Forest”, and the “caret" package [37, 38]. Figures were generated using the ggplot2 package [39]. TM-Vec representations were generated by configuring `tm_vec_model_cpnt` to `tm_vec_cath_model`. The functions can be accessed at (https://github.com/tymor22/tm-vec/tree/master). TM-scores were calculated using `tm_align` from the tmtools 0.1.1 module.

Data and code availability. The data that support the findings of this study and all code for data analysis are openly available at:

https://github.com/mirzaie-mehdi/ProteinEnergyProfileSimilarity

Funding: P.Ch. and M.M. were supported from the grants of J.-O.A - Academy of Finland (grants no. 297727 and 350678), Sigrid Juselius Foundation, ERA-NET NEURON grant nr 352077, Helsinki Institute of Life Science Research Fellow, and by European Research Council (ERC, grant no. 724922).

References

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

2. Altschul SF*, et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).

3. Kilinc M, Jia K, Jernigan RL. Improved global protein homolog detection with major gains in function identification. *Proceedings of the National Academy of Sciences* **120**, e2211823120 (2023).

4. Quan Y, Xiong Z-K, Zhang K-X, Zhang Q-Y, Zhang W, Zhang H-Y. Evolution-strengthened knowledge graph enables predicting the targetability and druggability of genes. *PNAS nexus* **2**, pgad147 (2023).

5. Du Z, Ding X, Hsu W, Munir A, Xu Y, Li Y. pLM4ACE: A protein language model based predictor for antihypertensive peptide screening. *Food Chemistry* **431**, 137162 (2024).

6. Wang Y*, et al.* RNAincoder: a deep learning-based encoder for RNA and RNA-associated interaction. *Nucleic Acids Research*, gkad404 (2023).

7. Sippl MJ. 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* **7**, 473--501 (1993).

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

9. Mirzaie M, Eslahchi C, Pezeshk H, Sadeghi M. A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys. *Proteins: Structure, Function, and Bioinformatics* **77**, 454-463 (2009).

10. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology* **347**, 827--839 (2005).

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

12. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. *Nucleic acids research* **28**, 257-259 (2000).

13. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research* **42**, D304-D309 (2014).

14. Lundin D, Poole AM, Sjberg B-M, Hgbom M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *Journal of Biological Chemistry* **287**, 20565--20575 (2012).

15. Gowthaman R, Guest JD, Yin R, Adolf-Bryfogle J, Schief WR, Pierce BG. CoV3D: a database of high resolution coronavirus protein structures. *Nucleic acids research* **49**, D282-D287 (2021).

16. van Heel AJ, de Jong A, Montalbán-López M, Kok J, Kuipers OP. BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Research* **41**, W448-W453 (2013).

17. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. *Nature communications* **10**, 1197 (2019).

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

19. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of molecular biology* **235**, 625--634 (1994).

20. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302-2309 (2005).

21. Hamamsy T*, et al.* Protein remote homology detection and structural alignment using deep learning. *Nat Biotechnol*, (2023).

22. Malod-Dognin N, Pržulj N. GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity. *Bioinformatics* **30**, 1259-1265 (2014).

23. Tian K, Zhao X, Zhang Y, Yau S. Comparing protein structures and inferring functions with a novel three-dimensional Yau–Hausdorff method. *Journal of Biomolecular Structure and Dynamics*, (2018).

24. Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of αα-turn motifs in proteins. *Journal of molecular biology* **255**, 235-253 (1996).

25. Malik AJ, Poole AM, Allison JR. Structural phylogenetics with confidence. *Molecular Biology and Evolution* **37**, 2711--2726 (2020).

26. Puente-Lelievre C*, et al.* Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *bioRxiv*, 2023.2012. 2012.571181 (2023).

27. Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution* **21**, 255-265 (2004).

28. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593 (2011).

29. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* **23**, 254-267 (2005).

30. Hamamsy T*, et al.* Protein remote homology detection and structural alignment using deep learning. *Nature biotechnology*, 1-11 (2023).

31. Jumper J*, et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

32. Wu R*, et al.* High-resolution de novo structure prediction from primary sequence. bioRxiv 2022. *Google Scholar*.

33. Lin Z*, et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).

34. Wang G, Dunbrack Jr RL. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591 (2003).

35. Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)* **22**, 469-483 (1996).

36. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**, 2695-2696 (2006).

37. Breiman L. Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA* **1**, 3-42 (2002).

38. Kuhn M*, et al.* Package 'caret'. *The R Journal* **223**, (2020).

39. Wickham H. ggplot2. *Wiley interdisciplinary reviews: computational statistics* **3**, 180-185 (2011).

Acknowledgments: The authors would like to thank Vilma Iivanainen, Elina Nagaeva, and Sakari Hietanen for reading the manuscript and providing valuable feedback.

Author Contributions: M.M. designed and supervised the research; M.M. and P.Ch. analyzed data; M.M., P.Ch., and J.-O.A. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Competing Interest Statement: The authors declare no conflict of interest.

Figure Legends and Tables

Fig. 1 | Sequence-Structure relationship. A) The correlation between total energy estimates derived from protein structure and sequence for protein domains within ASTRAL40(left) and ASTRAL95(right) data sets. B) The correlation between the distances of profile of energy estimated from sequence (CPE) and structure (SPE) for all pairs of domains in ASTRAL40(left) and ASTRAL95(right).

Fig. 2 | Energy Distribution in Protein Domain Structural Classes. The distribution of normalized total energy in protein domains from ASTRAL40 and ASTRAL95 datasets based on protein structure (A) and sequence (B) 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.

Fig. 3 | UMAP Visualization of Energy Profiles in All-Alpha and All-Beta Domains from ASTRAL40 and ASTRAL95 Datasets. UMAP projection of SPE and CPE shows the separation of the all-alpha (green point) and all-beta (pink point) proteins selected from the ASTRAL40 and ASTRAL95 datasets. A) CPE of ASTRAL40, B) SPE of ASTRAL40, C) CPE of ASTRAL95, and D) SPE of ASTRAL95. Dots represent two dimensional UMAP projection of SPE(CPE) for individual sequences. UMAP plots were generated by parameters n_neighbors = 20 and min_dist = 0.1.

Fig. 4 | UMAP Visualization of Energy Profiles. The UMAP projection of Structural Energy Profiles (SPE) and Compositional Energy Profiles (CPE) of protein domains from ASTRAL40 and ASTRAL95 represents the structural information embedded in energy profiles across hierarchical levels of SCOP; each panel includes two figures, one generated by CPE (left panel) and the other by SPE (right panel), revealing that protein domains sharing the same A) fold, B) superfamily, and C) 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 were generated using parameters n_neighbors = 20 and min_dist = 0.1.

Fig. 5 | Comparative Boxplots of Pairwise Distances among Energy Profiles in ASTRAL40 and ASTRAL95. Comparative Boxplots of Pairwise Distances among Energy Profiles in ASTRAL40 and ASTRAL95, depicting A-B) intraclass distances within the all-alpha class (in purple) versus interclass distances (in yellow), C-D) 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 E-F) 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 Compositional Energy Profiles (CPE, left panel) and the other using Structural Energy Profiles (SPE, right panel).

Fig. 6 | Performance and Computational Efficiency of Protein Dissimilarity Measures. A) Time versus accuracy for the 1-NN classifier using GR-Align, RMSD, TM-score, Yau-Hausdorff distance, TM-Vec, and the distance between energy profiles (CPE) as measures of protein dissimilarity. B) Running times of the evaluated methods, scaled to 12 hours, with an inset zooming in on the region indicated by the dashed circle. The entire circle represents 130 seconds. Each method is represented by different colors as indicated in the figure legend.

Fig. 7 | Phylogenetic network reconstruction for the ferritin-like superfamily. A) Schematic view of the relationships between the major ferritin-like protein families. B) The network recontacted by SPE demonstrates the distinct separation (red dotted line) of two SCOP families: ferritins (SCOP id a.25.1.1), which includes Bacteri, ferritins, Dps, and rubrerythrin subgroups, and the Ribonucleotide Reductase-like family (SCOP id a.25.1.2), which includes BMM_alpha, BMM_beta, Fatty_acid, and RNRR2 subgroups. Smaller groups are clearly distinguished. C) The network recontacted by TM-Vec representation and cosine similarity.

Fig. 8 | Clustering Analysis of Spike Glycoprotein Structures from SARS-CoV, SARS-CoV-2, and MERS-CoV. 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, B) protein sequence C) RMSD D) TM-Vec. Each leaf on the dendrogram is labeled with the PDB-IDs of the corresponding chains, and the leaves are color-coded to represent the host virus of the spike glycoprotein structure. E) The ARI values of CPE, SPE, TM-Vec, RMSD, and TM-Score scores, and F) The running time.

Fig. 9 | UMAP Visualization and Comparison of Embeddings for Bacteriocins. Visualization of profile of energies embeddings using UMAP for 689 peptides across three classes of bacteriocins. A) The UMAP projection of Compositional Energy Profiles (CPE) on bacteriocins at different classes. B) Comparison of CPE distances (CPE_dis) with the TM-scores produced by running TM-align on structures predicted by AlphaFold2, OmegaFold and ESMFold, and TM-Vec for 238,000 pairs of bacteriocins. CPE_dis is normalized by min-max normalization

Fig. 10 | Correlation between Protein-Protein Interaction Network Distances and Profile of Energies Distances. The correlation between separation distances estimated by protein-protein interaction network (X-axis) and the distance between profiles of energies (Y-axis).

Table 1 | The accuracy and computation time for 1-NN classifier based on GR-Align, RMSD, TM-score, Yau-Hausdorff distance, TM-Vec, and the distance between profiles of energy (CPE) as a measure of protein dissimilarity.

Method	Accuracy	Time
GR-Align	62.3%	2 min
RMSD	59.2%	1 h
TM-Score	61.5%	9 h 20 min
YH (10 Rotation)	70.8%	10 min
YH (2500 Rotation)	81.5%	4h 10 min
TM-Vec	100%	67 sec
CPE	100%	1 sec

Table 2 | Total accuracy and F1 measure for each of the five superfamilies by 1-NN and the results of 10-Fold cross validation with random forest (RF).

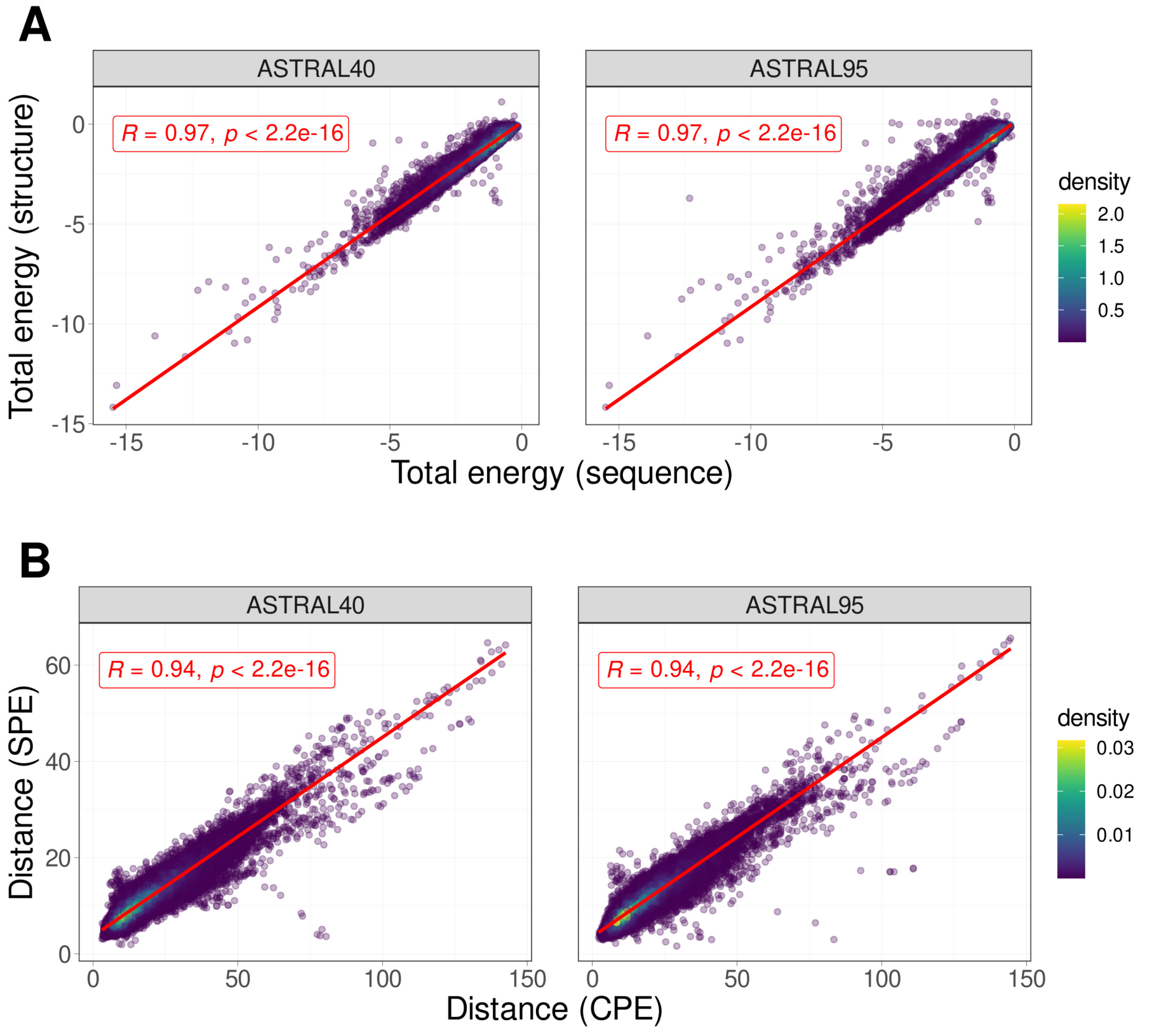
Method	Time	Accuracy	F1 Measure				
			wigend_helix	PH.domain-like	NTF-like	Ubiquitin-like	Immunoglobulins
CPE (1NN)	103 Sec	0.98	0.98	0.96	0.99	0.99	0.99
CPE (RF)	103 Sec	0.99	0.97	0.97	0.99	0.99	0.99
TM_Vec (1NN)	955 Sec	0.99	0.99	1	1	0.99	0.99

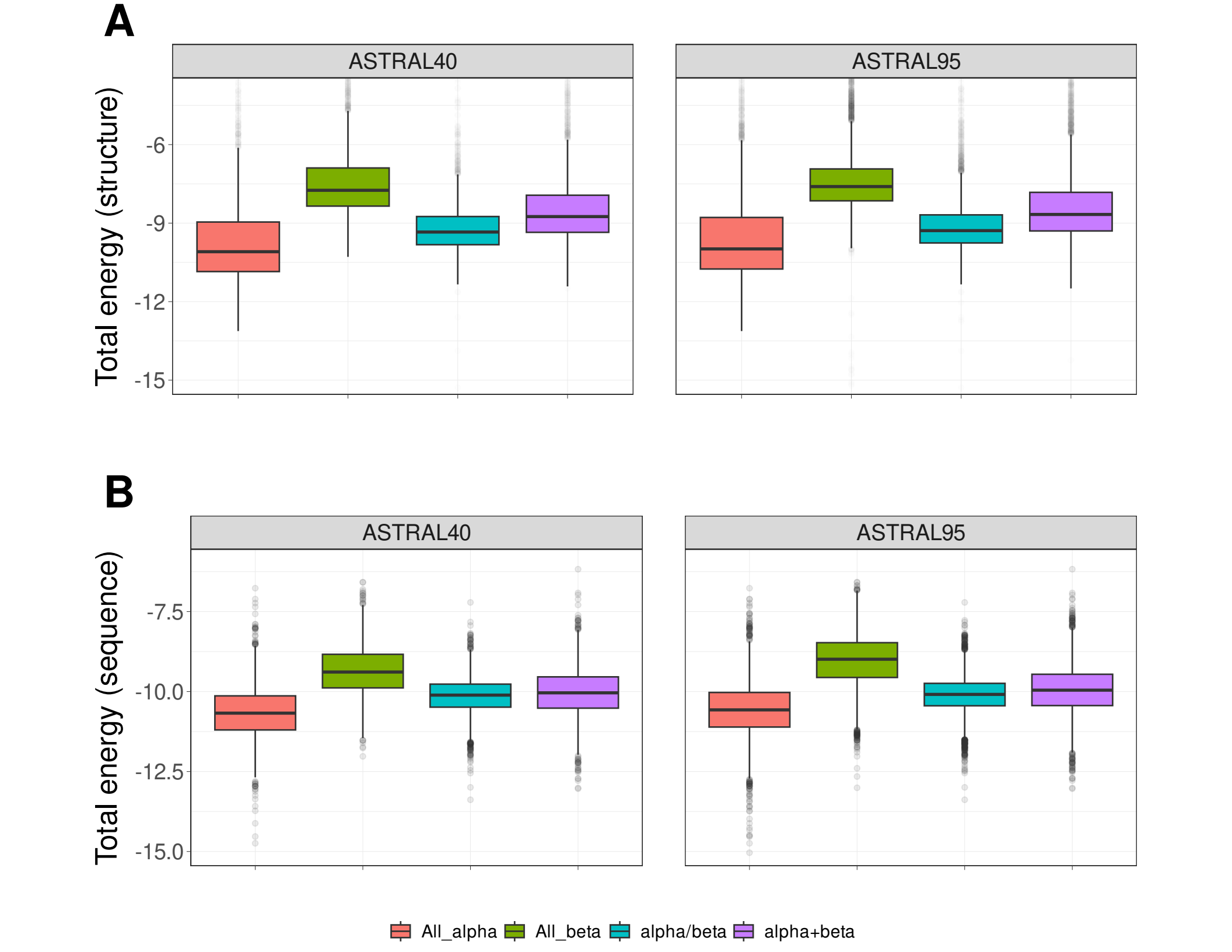
878 **Table 3 |** Comparison of clustering results using Adjusted Rand Index (ARI)

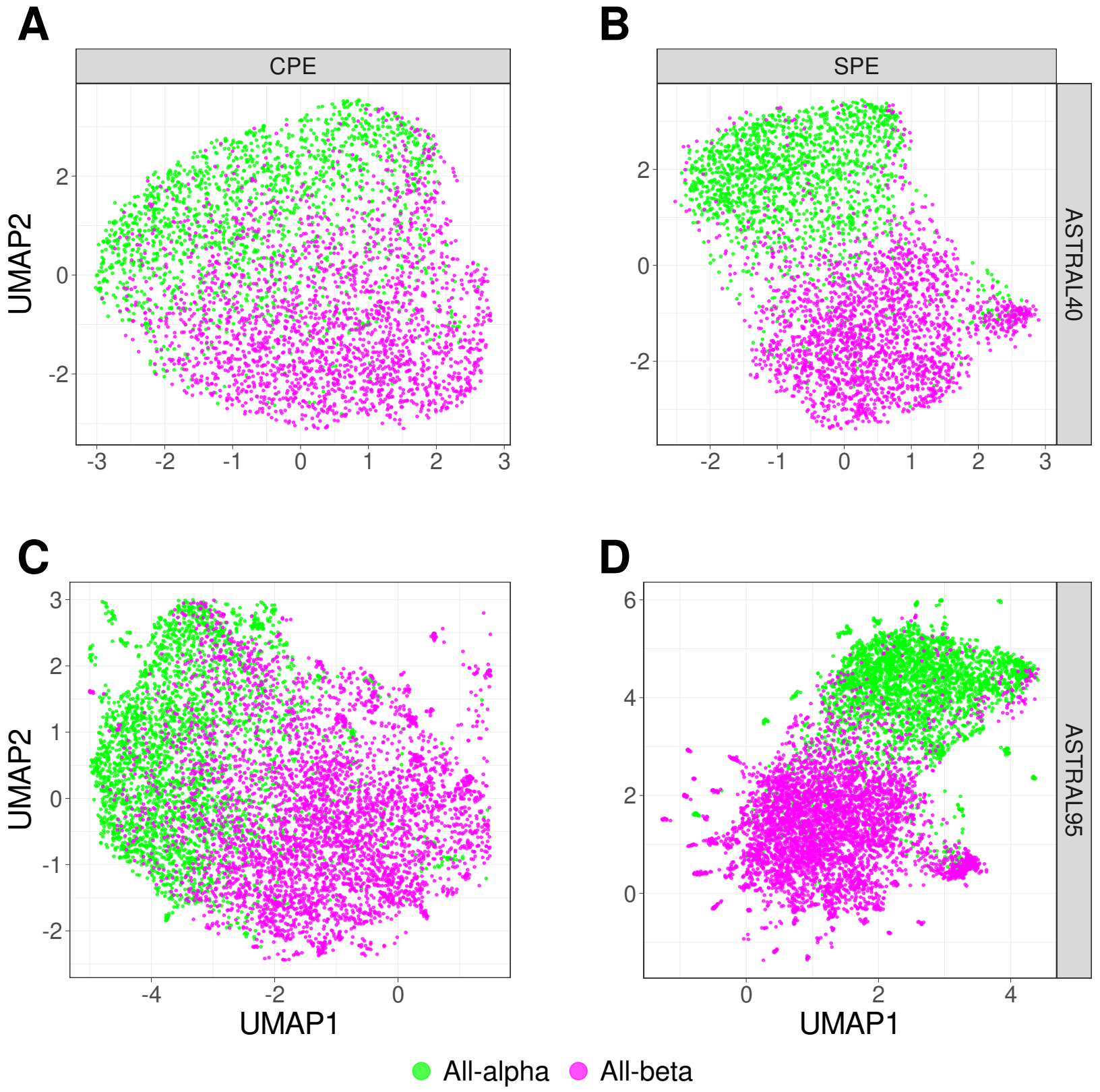
Method	Type	Cut Tree	ARI
CPE	Sequence-Based	4	0.95
TM-Vec	Sequence-Based	5	0.87
SPE	Structure-Based	3	1.00
RMSD	Structure-Based	6	0.73
TM-Score	Structure-Based	4	0.56

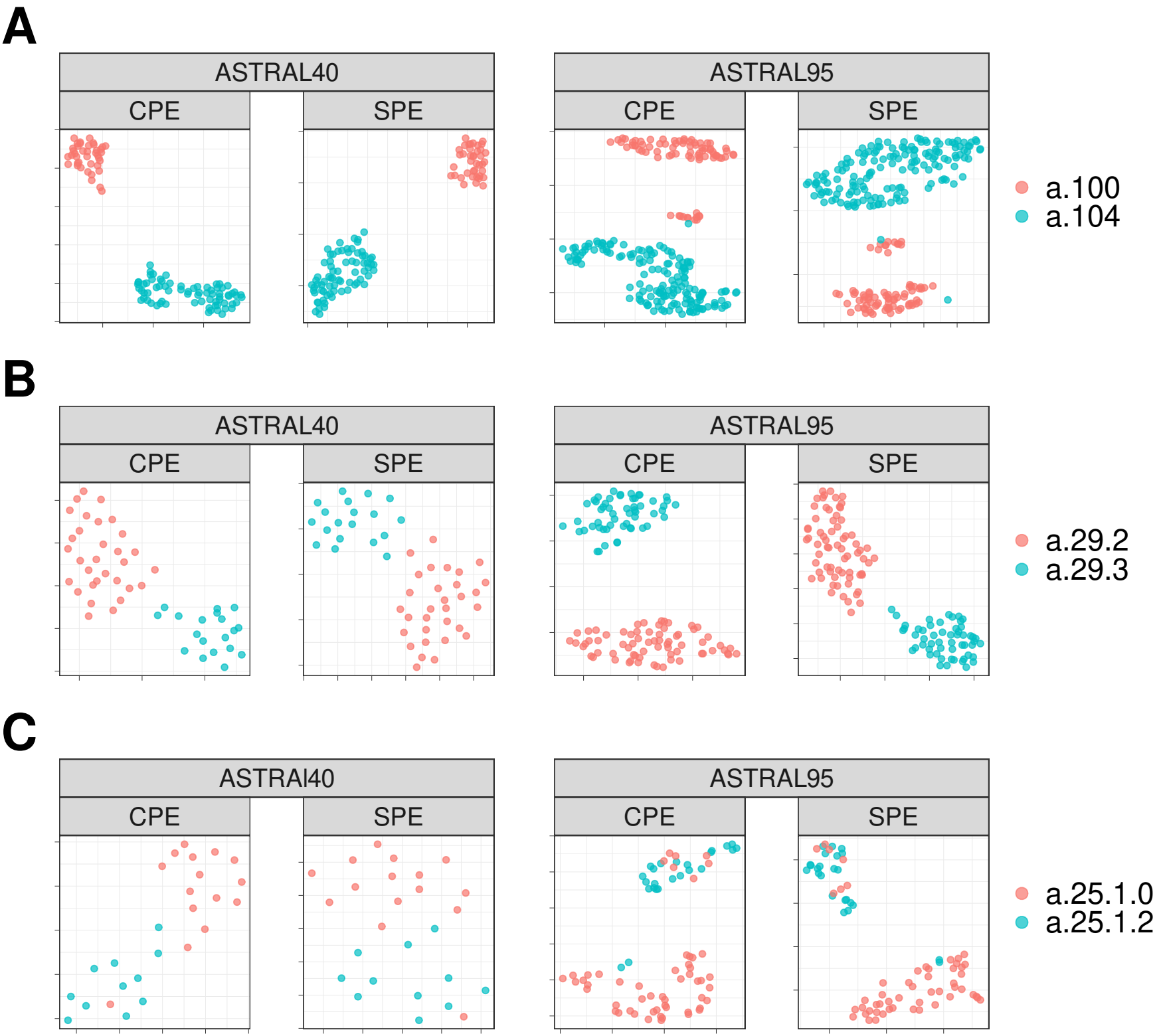
879

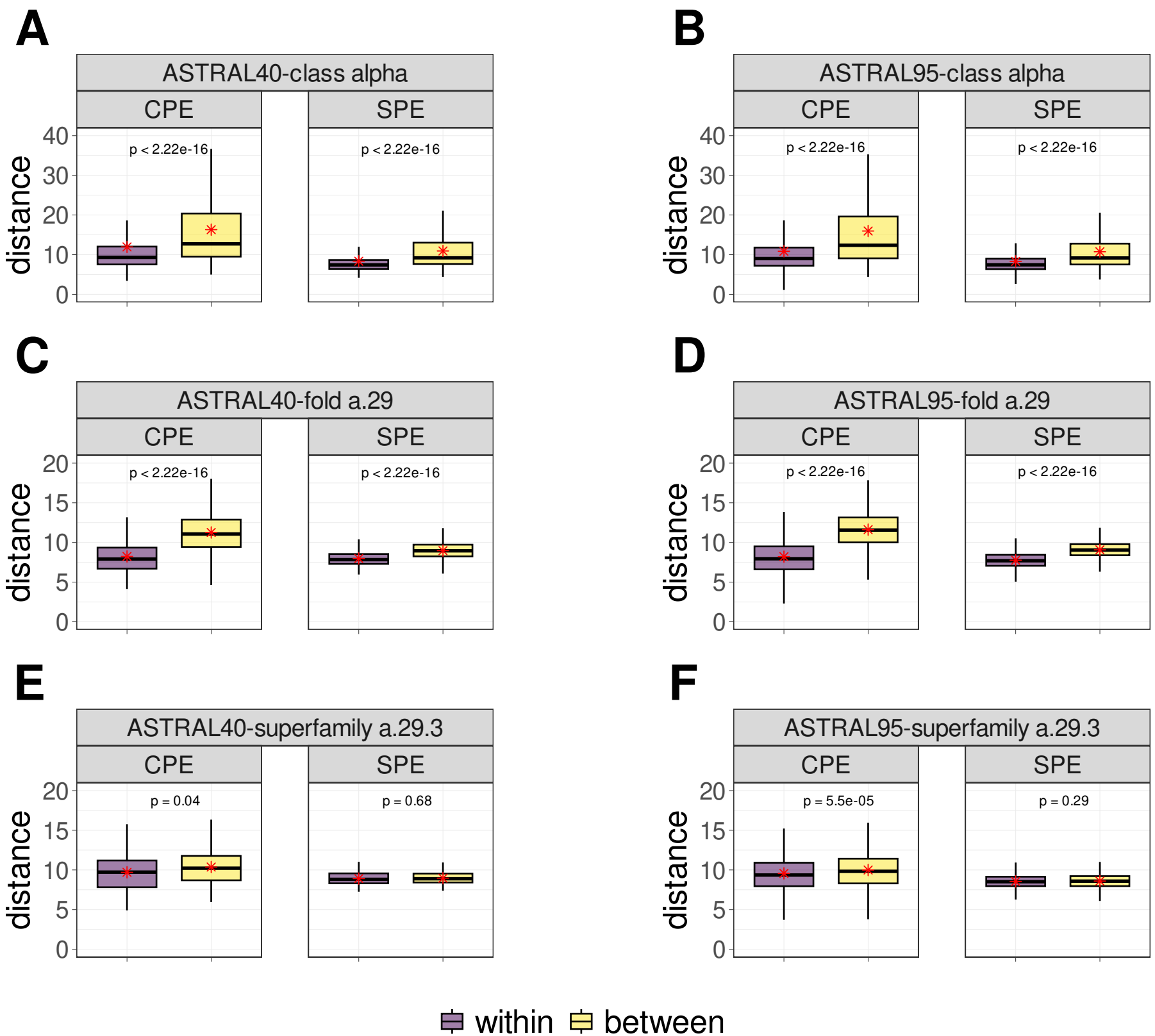
880

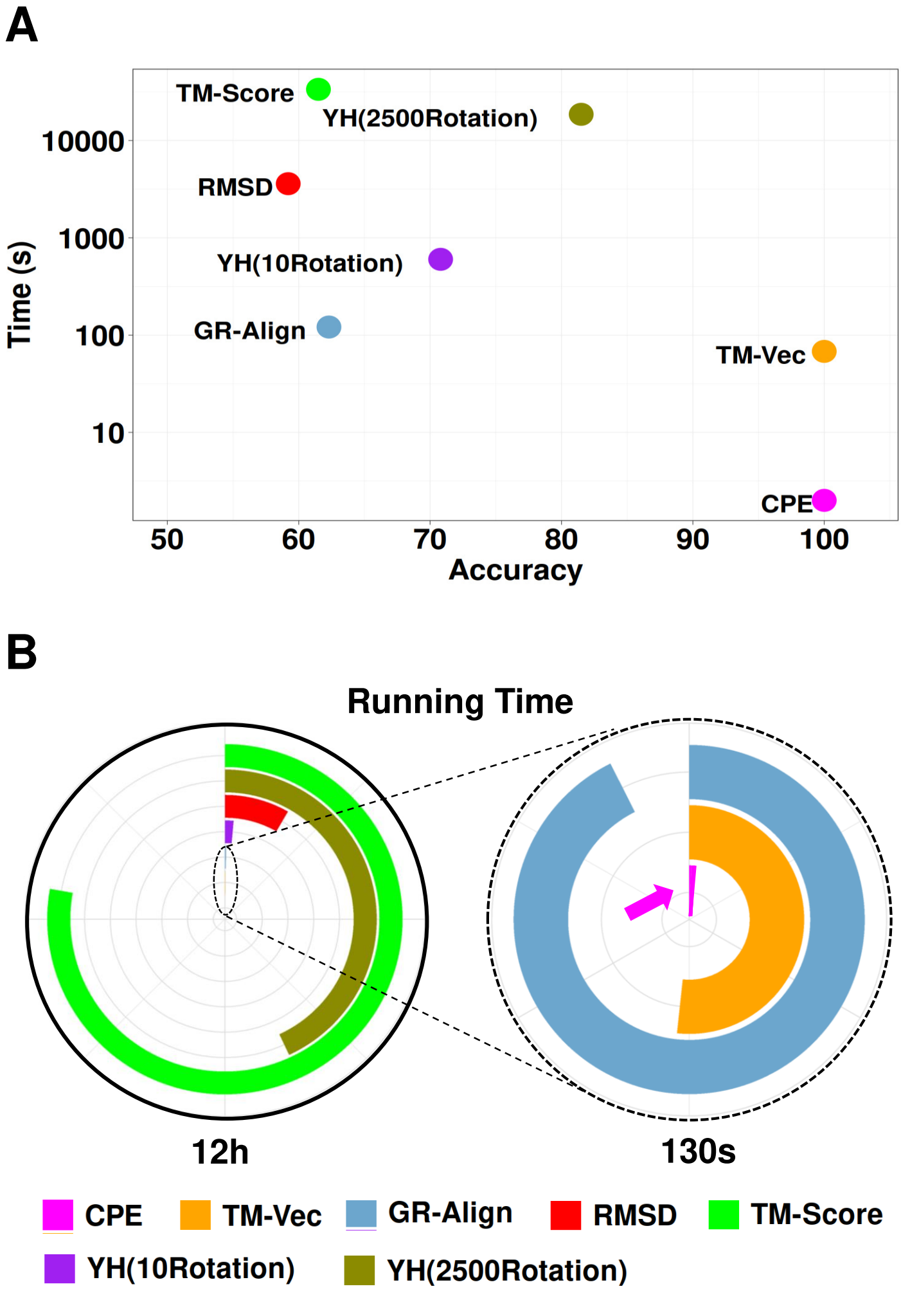

A
ASTRAL40
ASTRAL95
R = 0.97, p < 2.2e-16
R = 0.97, p < 2.2e-16
Total energy (structure)
Total energy (sequence)
density
2.0
1.5
1.0
0.5
0
-5
-10
-15
B
ASTRAL40
ASTRAL95
R = 0.94, p < 2.2e-16
R = 0.94, p < 2.2e-16
Distance (SPE)
Distance (CPE)
density
0.03
0.02
0.01
60
40
20
0
50
100
150


A
ASTRAL40
ASTRAL95
Total energy (structure)
-6
-9
-12
-15
B
ASTRAL40
ASTRAL95
Total energy (sequence)
-7.5
-10.0
-12.5
-15.0
All_alpha
All_beta
alpha/beta
alpha+beta


A
B
C
D
CPE
SPE
ASTRAL40
ASTRAL95
UMAP1
UMAP2
All-alpha
All-beta


A
ASTRAL40
CPE
SPE
ASTRAL95
CPE
SPE
a.100
a.104
B
ASTRAL40
CPE
SPE
ASTRAL95
CPE
SPE
a.29.2
a.29.3
C
ASTRAI40
CPE
SPE
ASTRAL95
CPE
SPE
a.25.1.0
a.25.1.2

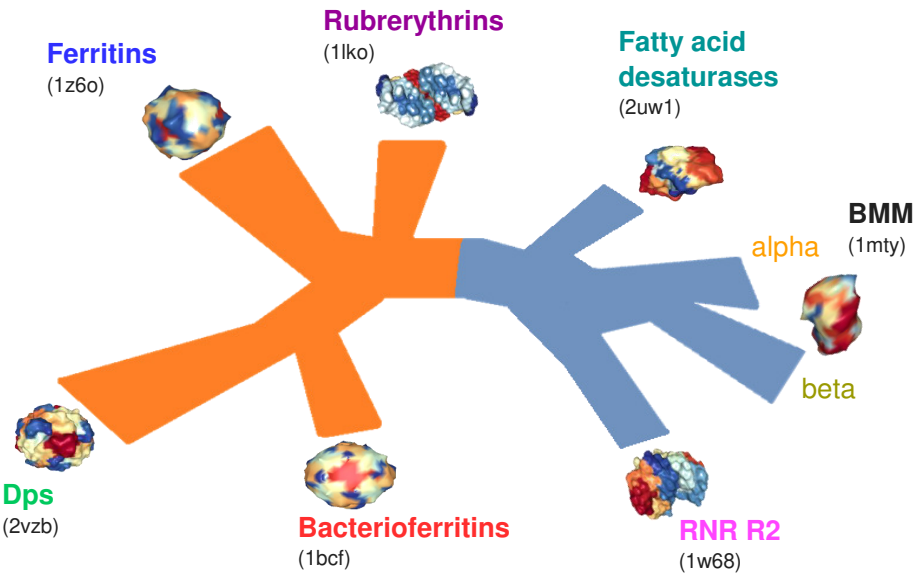

A
ASTRAL40-class alpha
CPE
SPE
distance
p < 2.22e-16
p < 2.22e-16
B
ASTRAL95-class alpha
CPE
SPE
distance
p < 2.22e-16
p < 2.22e-16
C
ASTRAL40-fold a.29
CPE
SPE
distance
p < 2.22e-16
p < 2.22e-16
D
ASTRAL95-fold a.29
CPE
SPE
distance
p < 2.22e-16
p < 2.22e-16
E
ASTRAL40-superfamily a.29.3
CPE
SPE
distance
p = 0.04
p = 0.68
F
ASTRAL95-superfamily a.29.3
CPE
SPE
distance
p = 5.5e-05
p = 0.29
within
between


A
TM-Score
YH(2500Rotation)
RMSD
YH(10Rotation)
GR-Align
TM-Vec
CPE
10000
1000
100
10
Time (s)
50
60
70
80
90
100
Accuracy
B
Running Time
12h
130s
CPE
TM-Vec
GR-Align
RMSD
TM-Score
YH(10Rotation)
YH(2500Rotation)

A

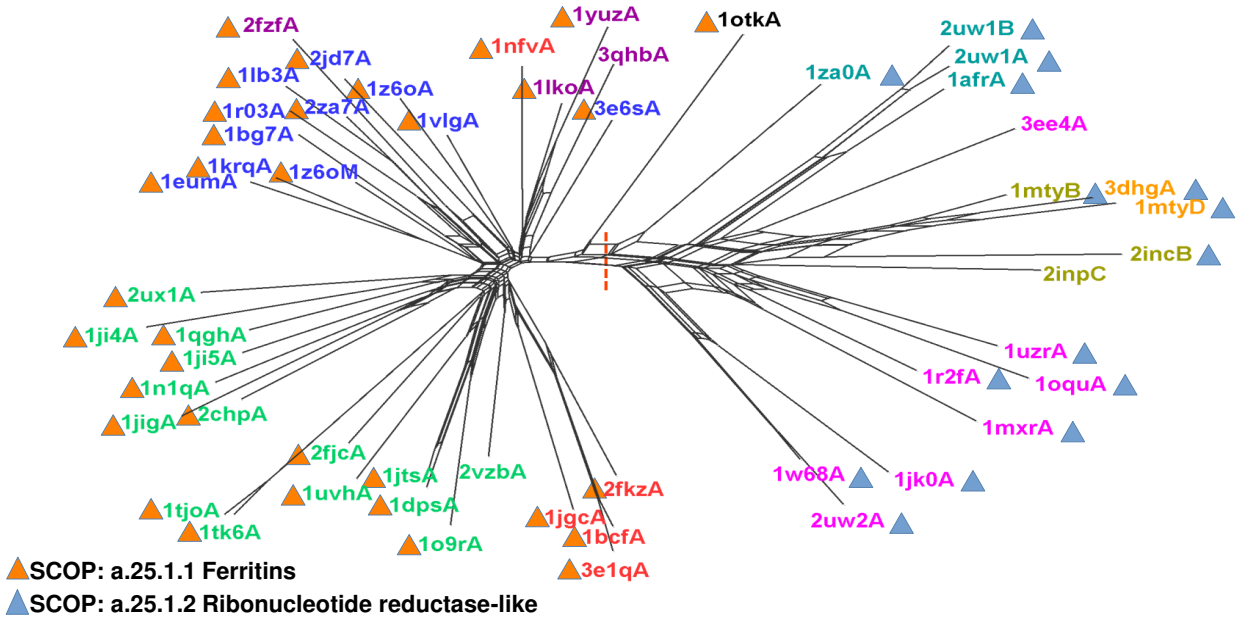
Manual Annotation

Subgroup	Pfam
Bacterioferirtin	00210 Ferritin
Dps	
Ferritin	
BMM alpha	02332 Phenol_hydrox
BMM beta	
Fatty acid desaturase	03405 FA_desaturase
RNR R2	00268 Ribonuc_red_sm
Rubrerythrins	02915 Rubrerythrins
1tokA	05138 PaaA_PaaC



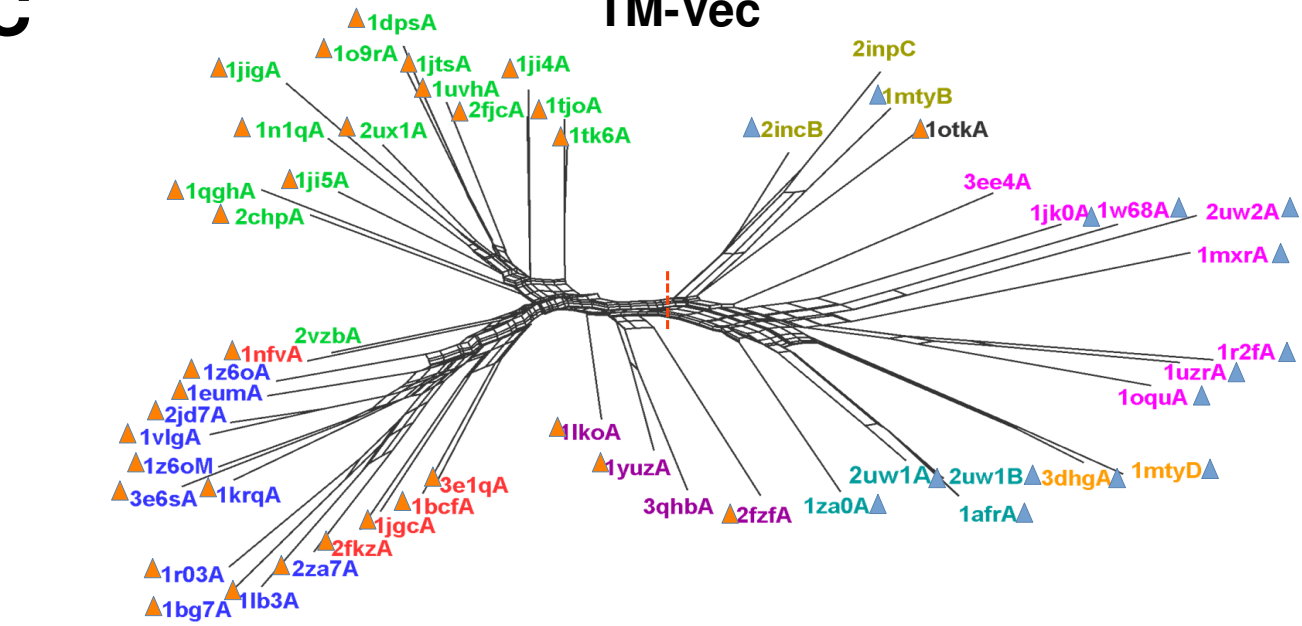
B

SPE



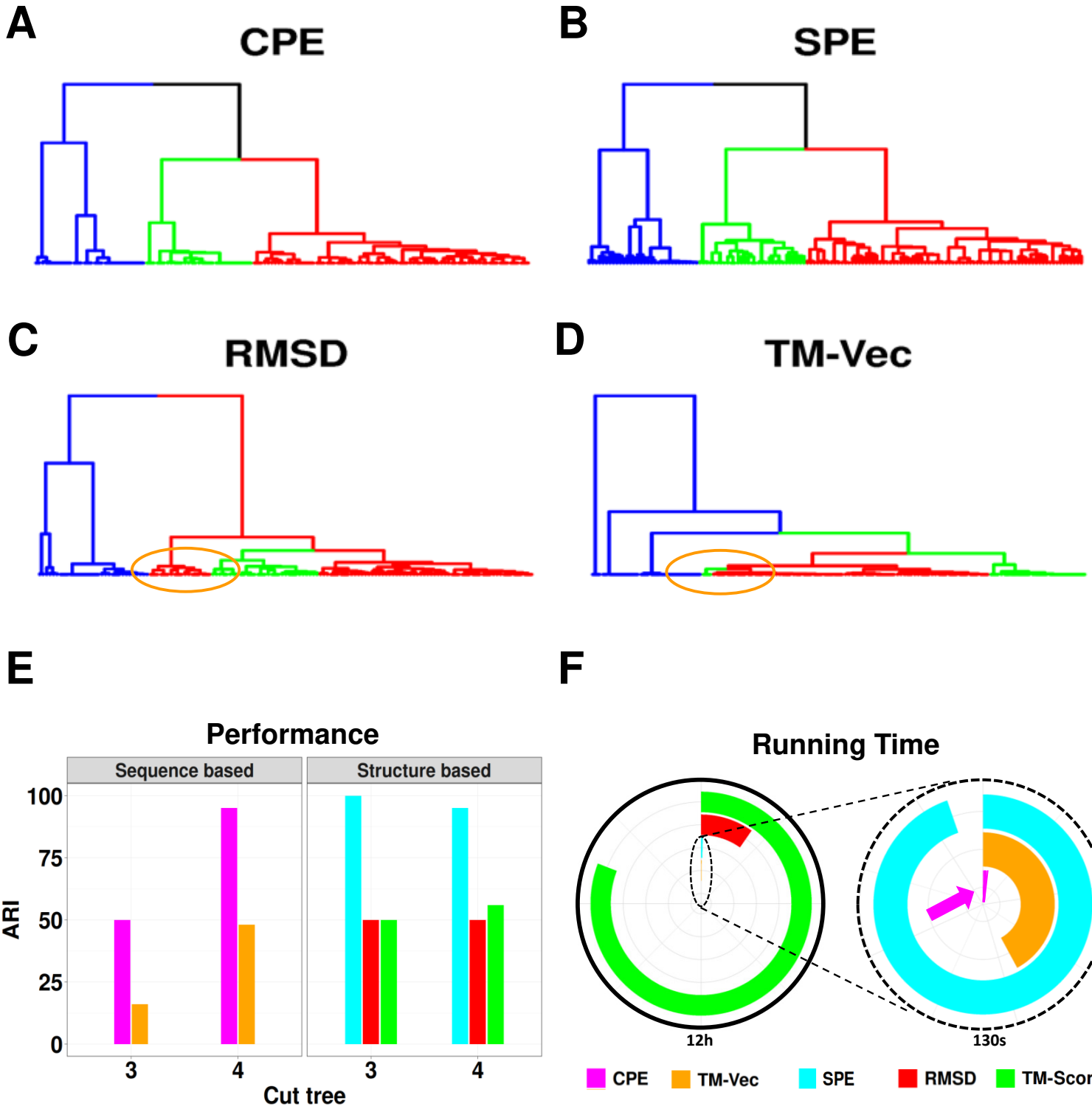
C

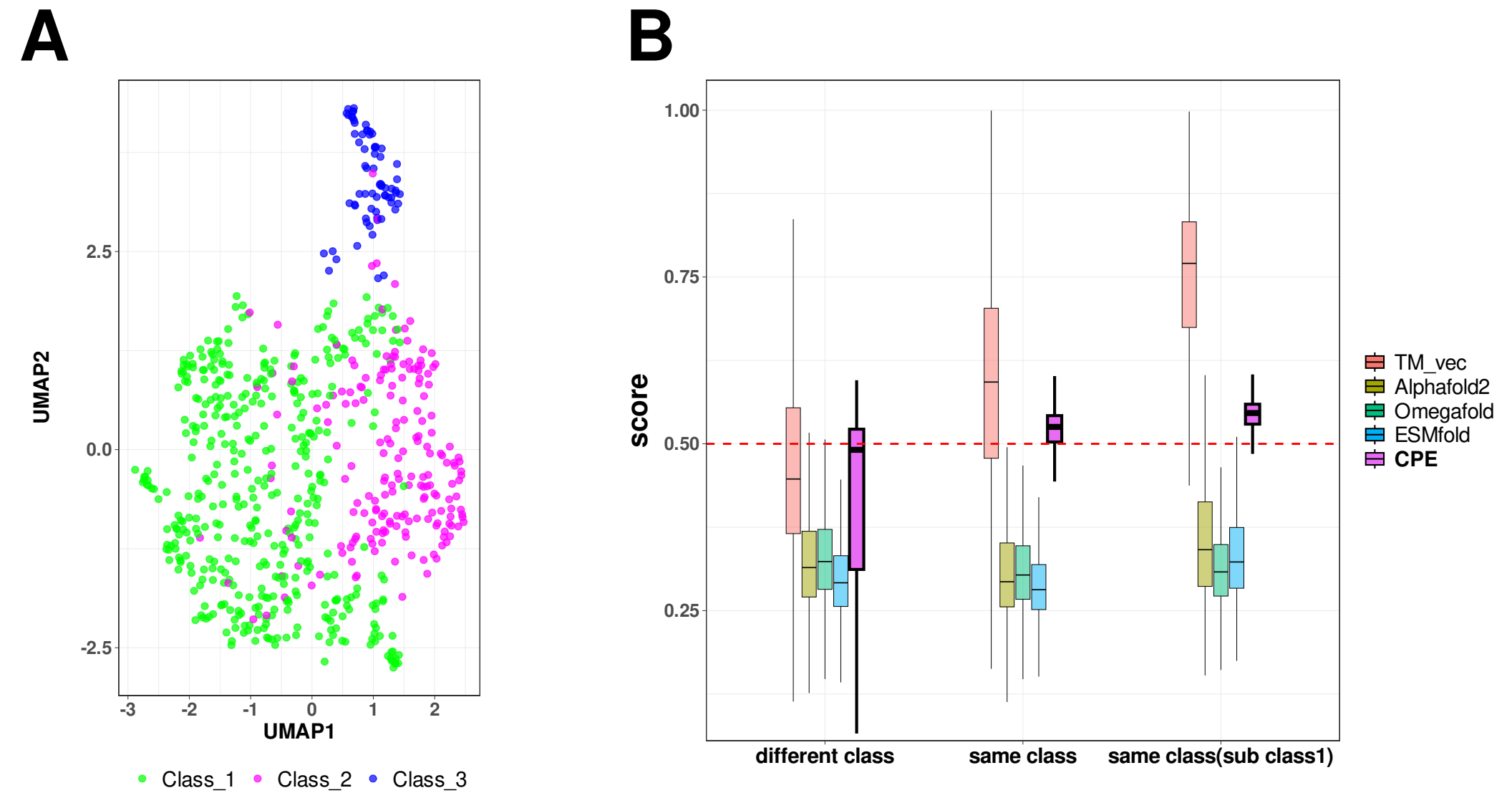
TM-Vec

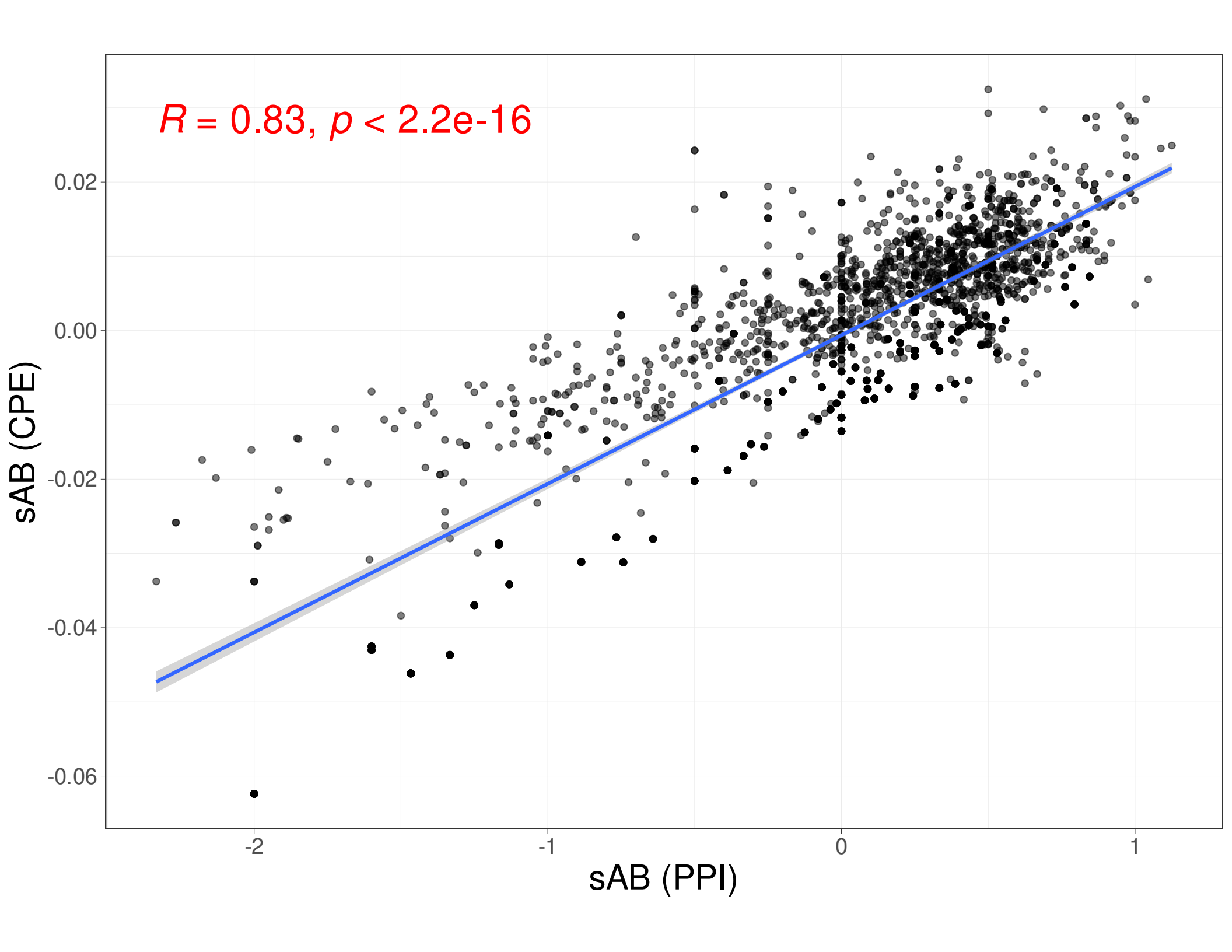


Evolution of Spike protein

MERS_Cov → SARS_Cov → SARS_Cov2




A
B
UMAP2
UMAP1
2.5
0.0
-2.5
-3
-2
-1
0
1
2
Class_1
Class_2
Class_3
score
1.00
0.75
0.50
0.25
different class
same class
same class(sub class1)
TM_vec
Alphafold2
Omegafold
ESMfold
CPE


R = 0.83, p < 2.2e-16
0.02
0.00
-0.02
-0.04
-0.06
sAB (CPE)
-2
-1
0
1
sAB (PPI)