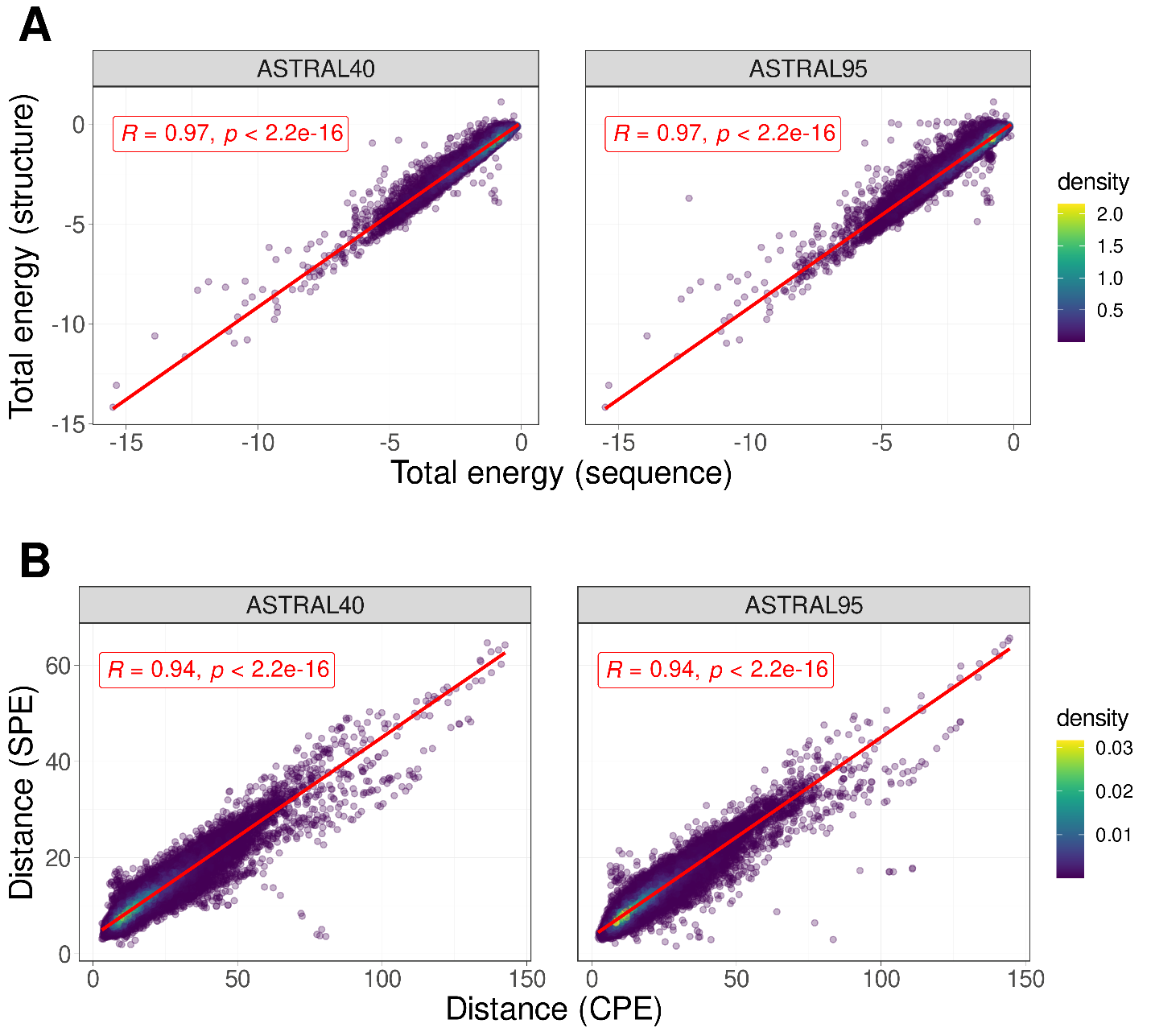
**Figures and Tables**

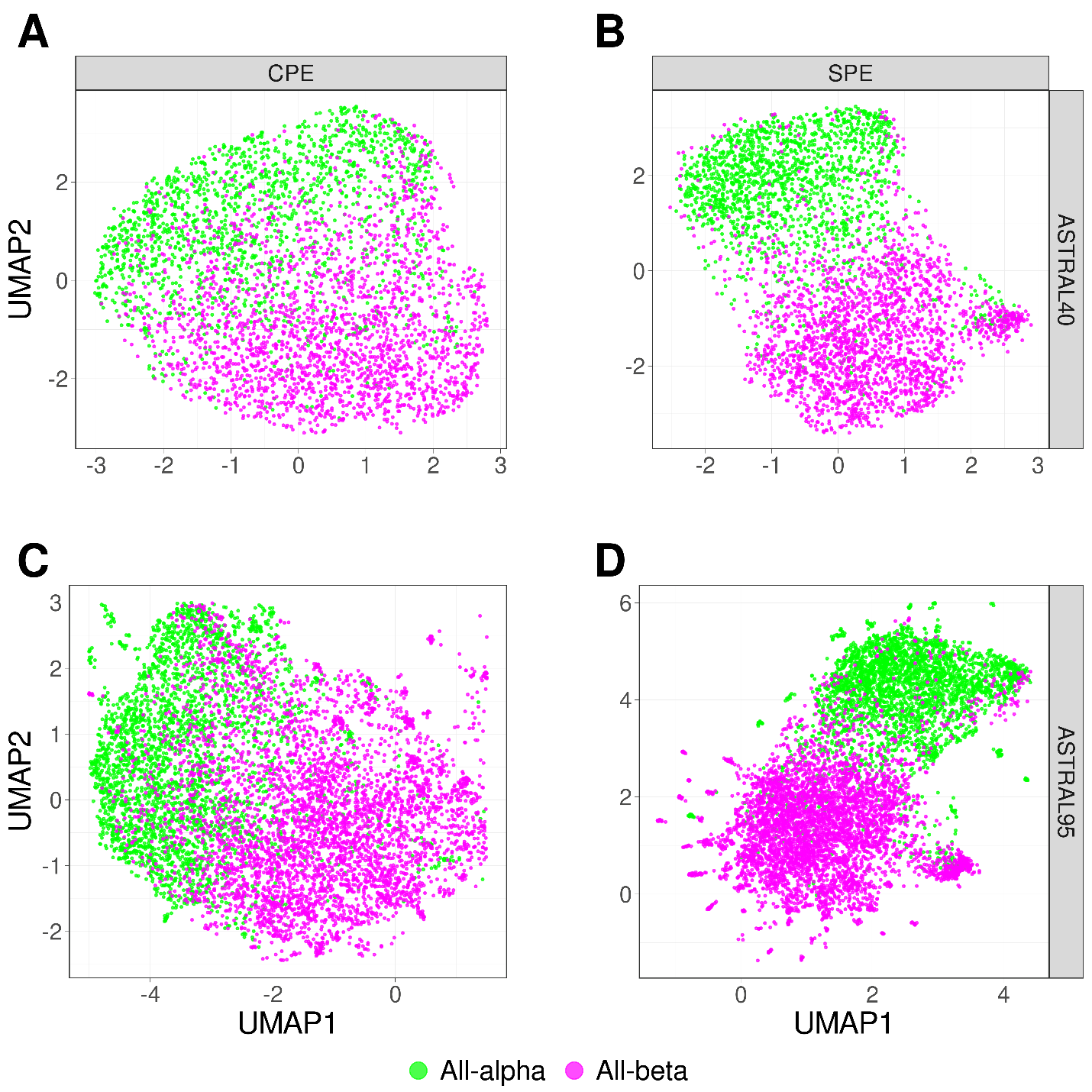


**Fig. 1 | Sequence-Structure relationship.** A) The correlation between total energy estimates derived from protein structure and sequence for protein domains within ASTRAL40(left) and ASTRAL95(right) data sets. B) The correlation between the distances of profile of energy estimated from sequence (CPE)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

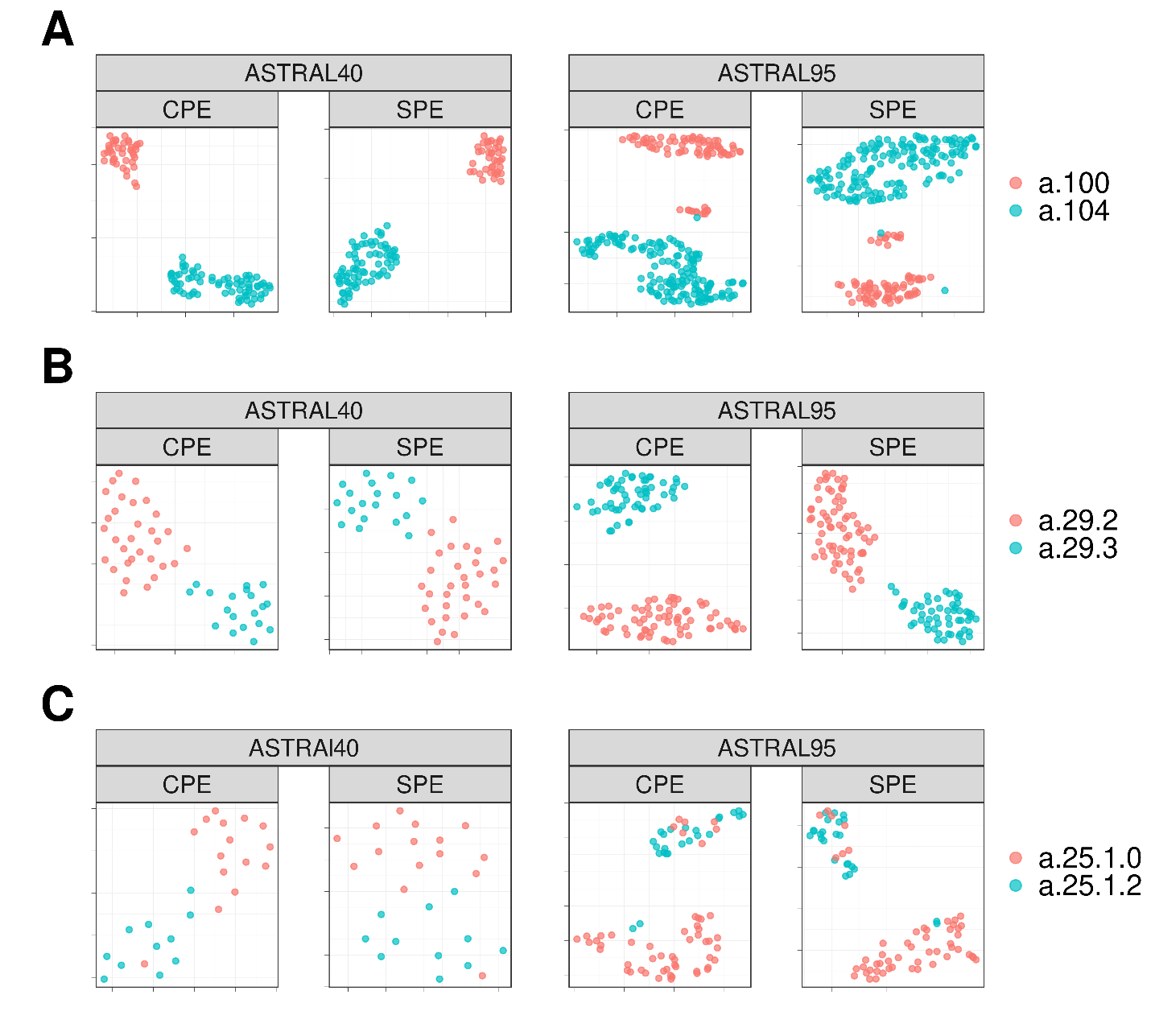
A group of colorful boxes

Description automatically generated with medium confidence

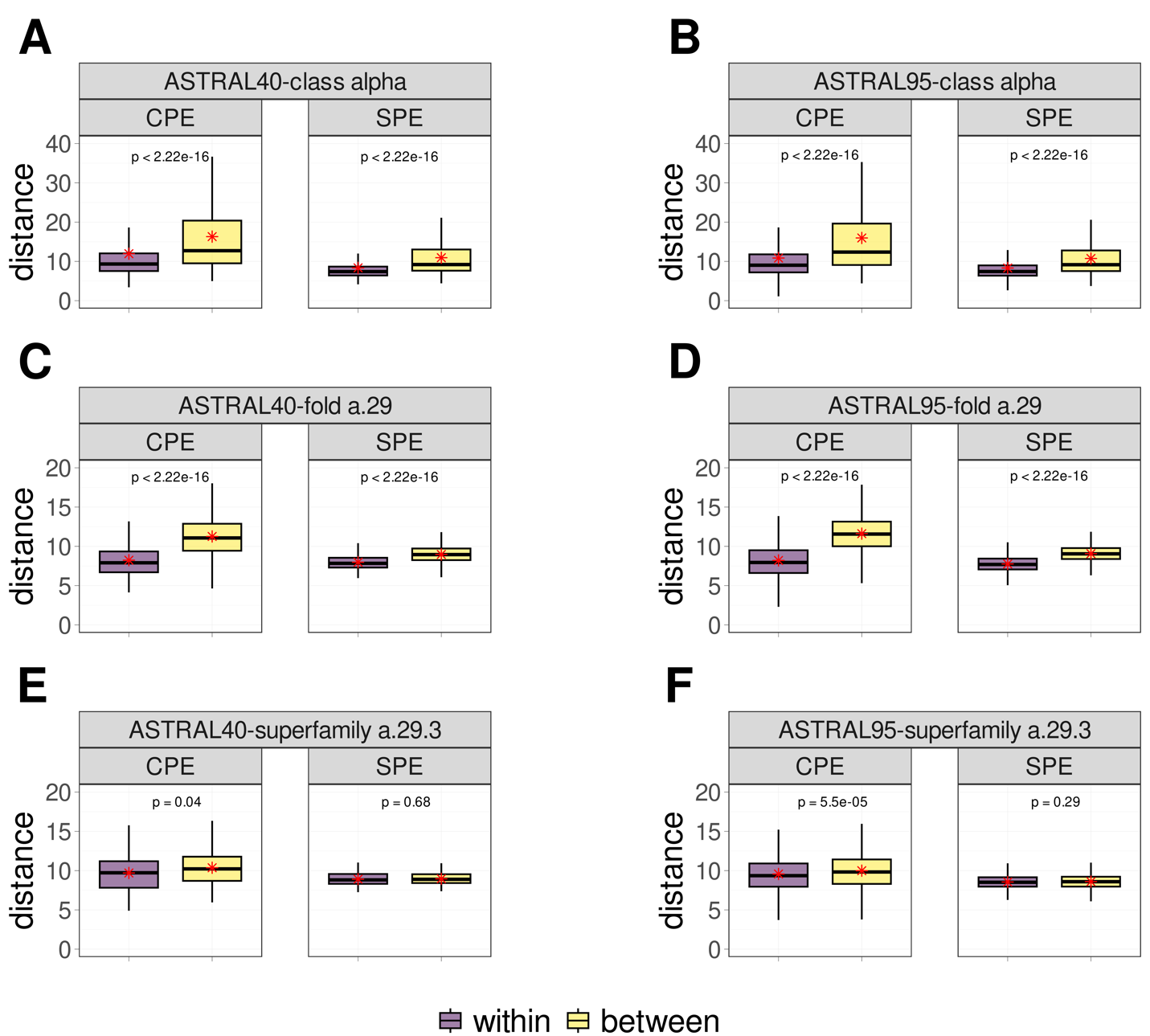
**Fig. 2 | Energy Distribution in Protein Domain Structural Classes.** The distribution of normalized total energy in protein domains from ASTRAL40 and ASTRAL95 datasets based on protein structure (A) and sequence (B) across various structural scope classes. In the ASTRAL40 dataset, there are 2644, 3059, 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.



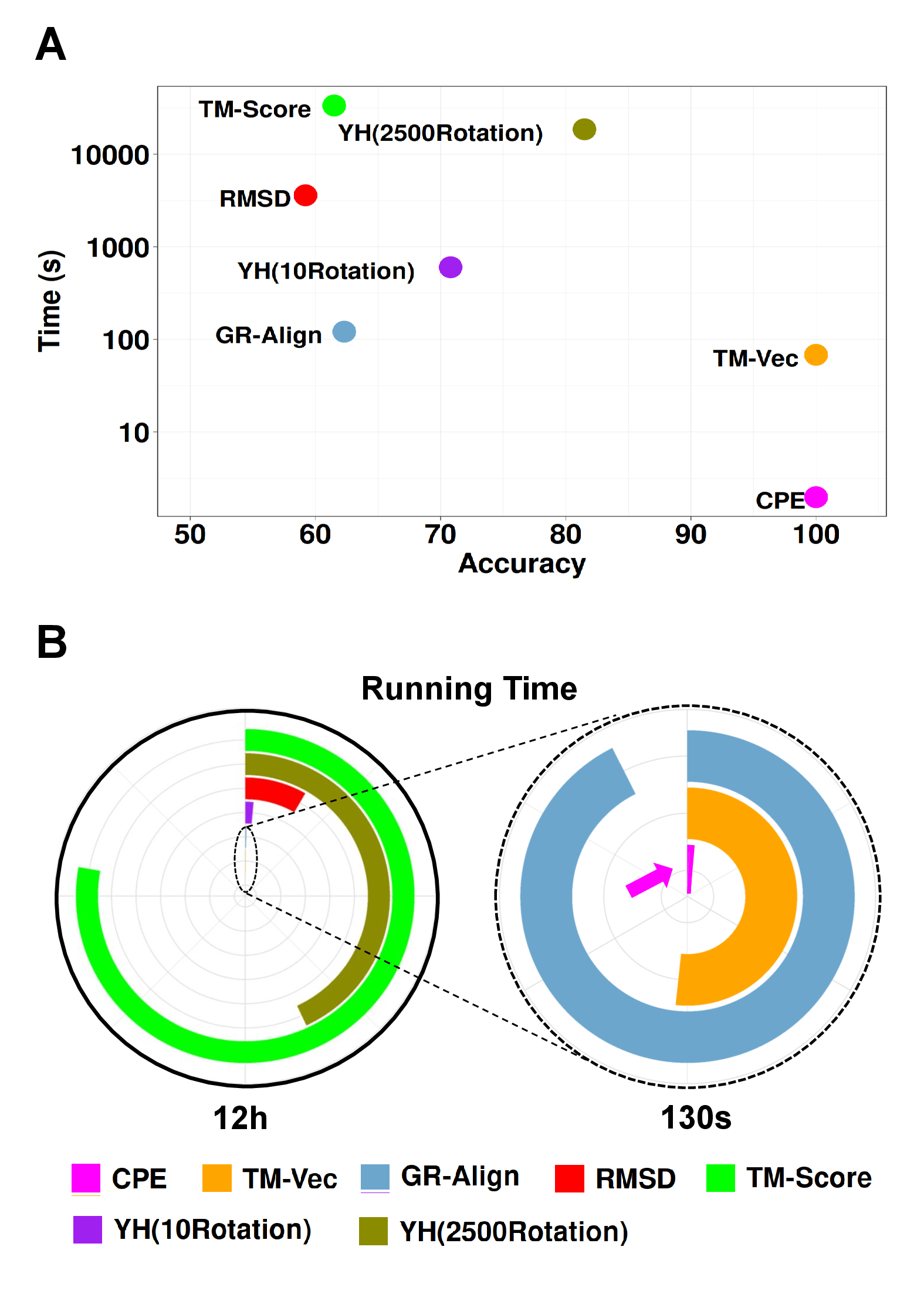
**Fig. 3 | UMAP Visualization of Energy Profiles in All-Alpha and All-Beta Domains from ASTRAL40 and ASTRAL95 Datasets.** UMAP projection of SPE and CPE shows the separation of the all-alpha (green point) and all-beta (pink point) proteins selected from the ASTRAL40 and ASTRAL95 datasets. A) CPE of ASTRAL40, 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.



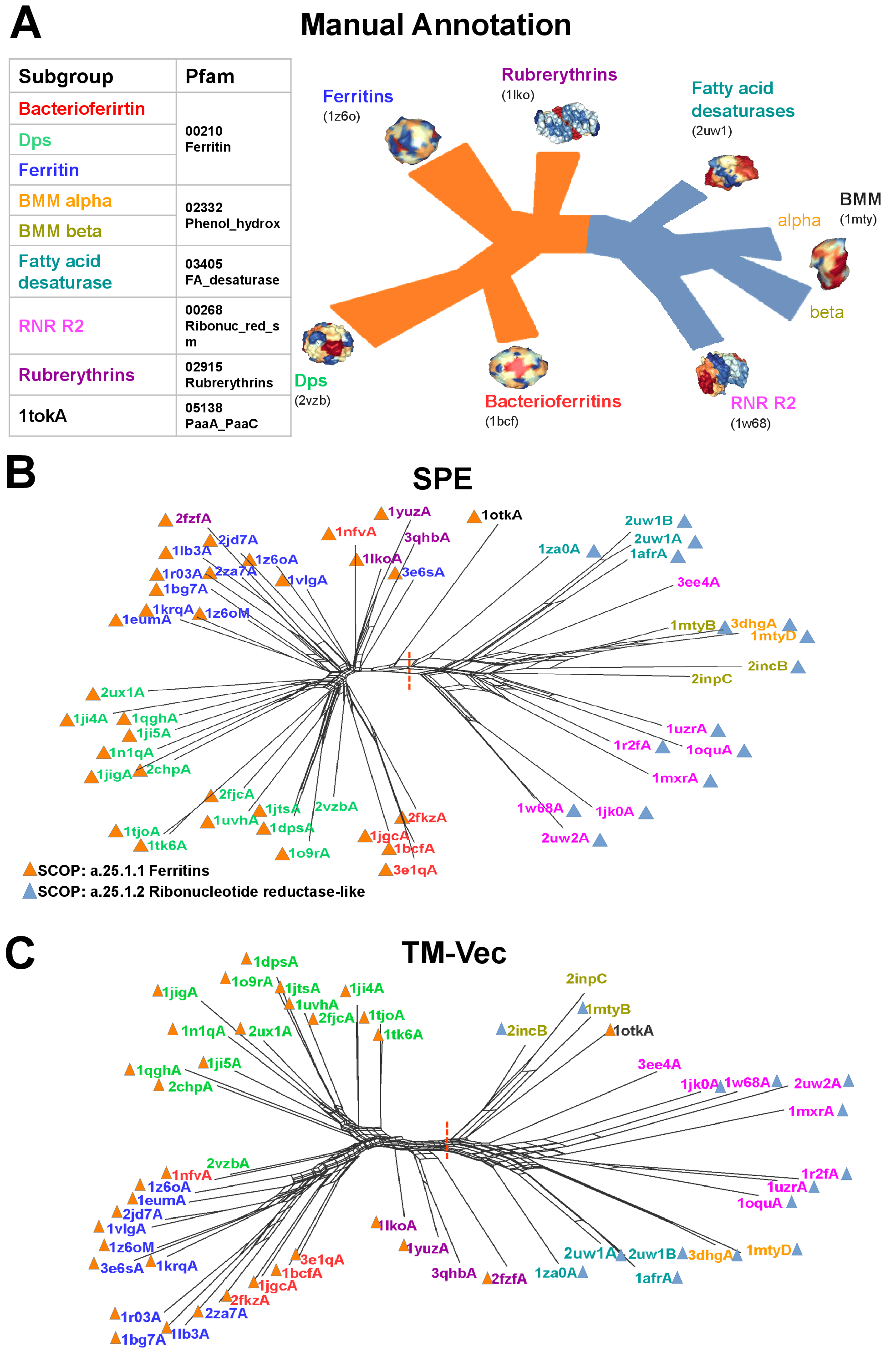
**Fig. 4 |** **UMAP Visualization of Energy Profiles.** The UMAP projection of Structural Energy Profiles (SPE) and Compositional Energy Profiles (CPE) of protein domains from ASTRAL40 and ASTRAL95 represents the structural information embedded in energy profiles across hierarchical levels of SCOP; each 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.



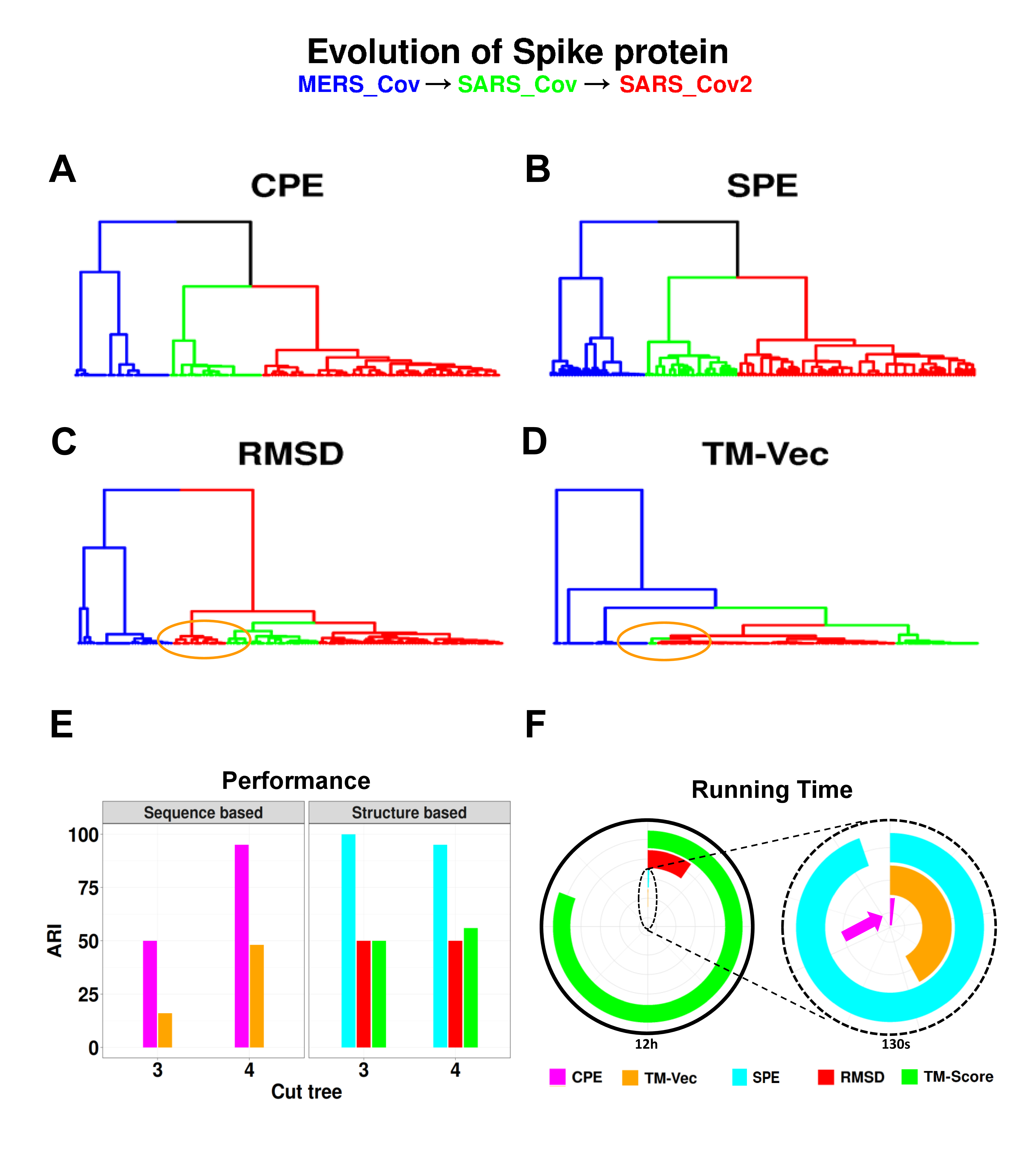
**Fig. 5 | Comparative Boxplots of Pairwise Distances among Energy Profiles in ASTRAL40 and ASTRAL95.** Comparative Boxplots of Pairwise Distances among Energy Profiles in ASTRAL40 and ASTRAL95, depicting A-B) intraclass distances within the all-alpha class (in purple) versus interclass distances (in 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).



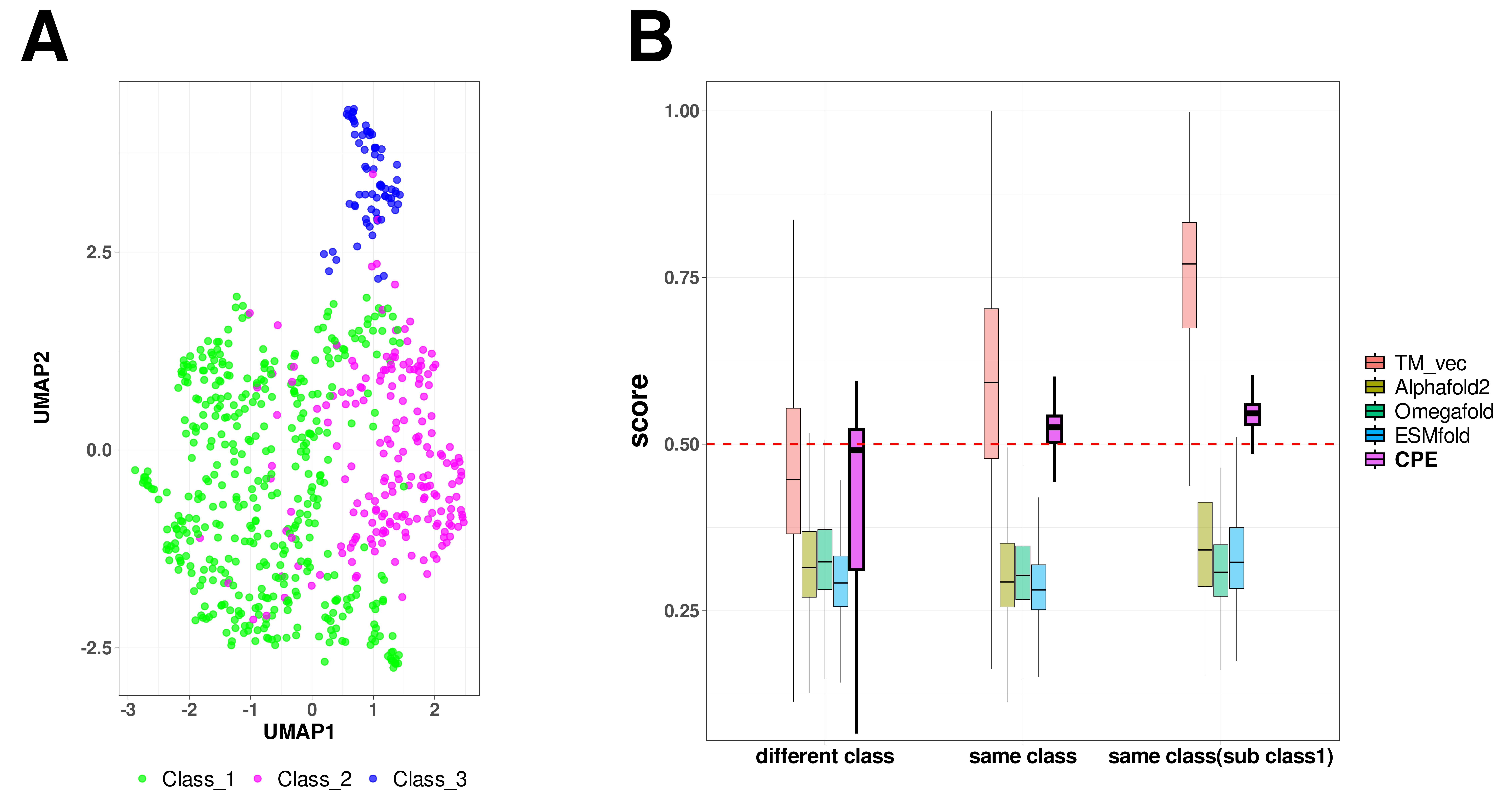
**Fig. 6 | Performance and Computational Efficiency of Protein Dissimilarity Measures**. A) Time versus accuracy for the 1-NN classifier using GR-Align, RMSD, TM-score, Yau-Hausdorff distance, TM-Vec, and the distance between energy profiles (CPE) as measures of protein dissimilarity. B) Running times 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

A graph of black dots

Description automatically generated

**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 |

**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 |