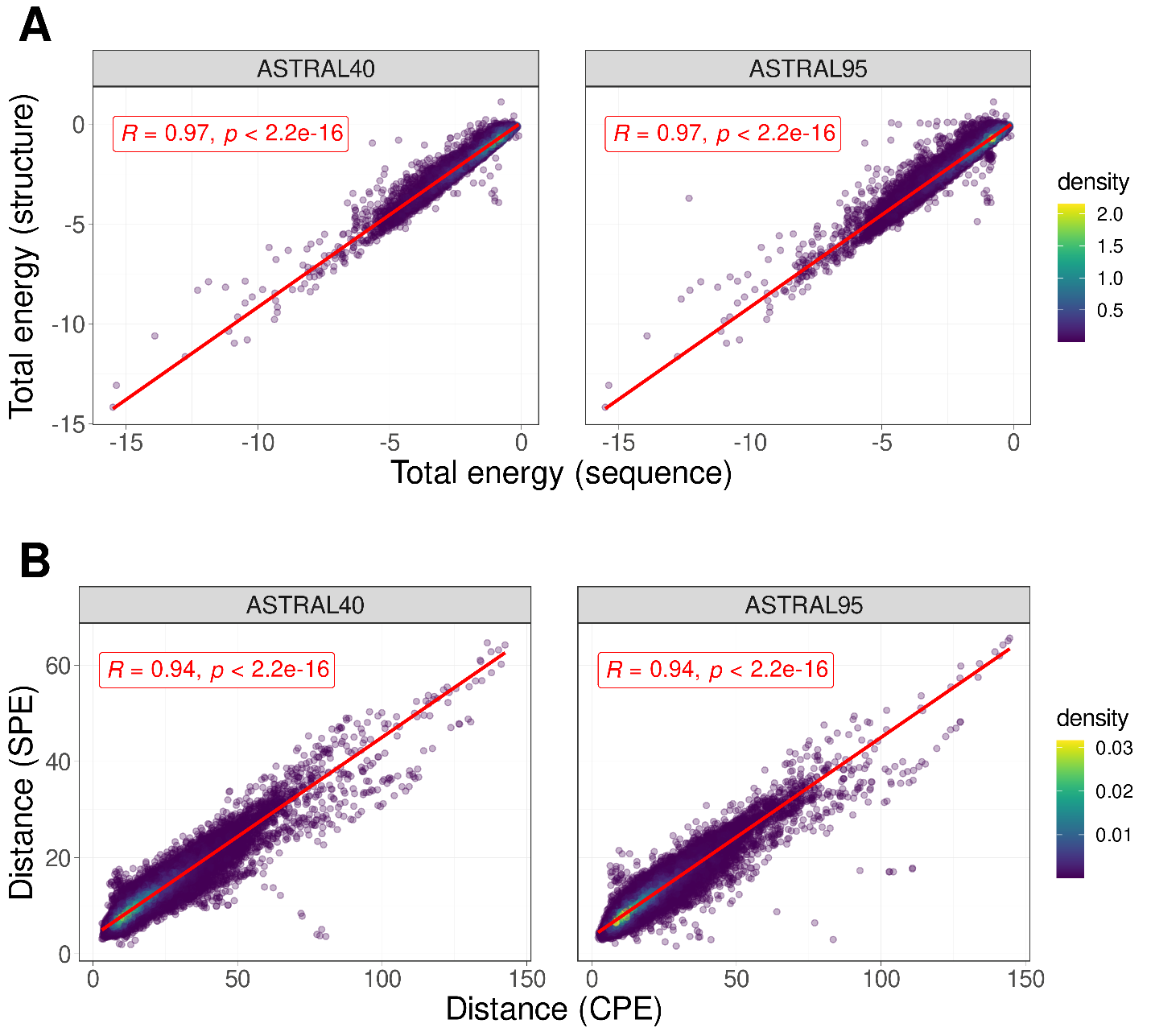
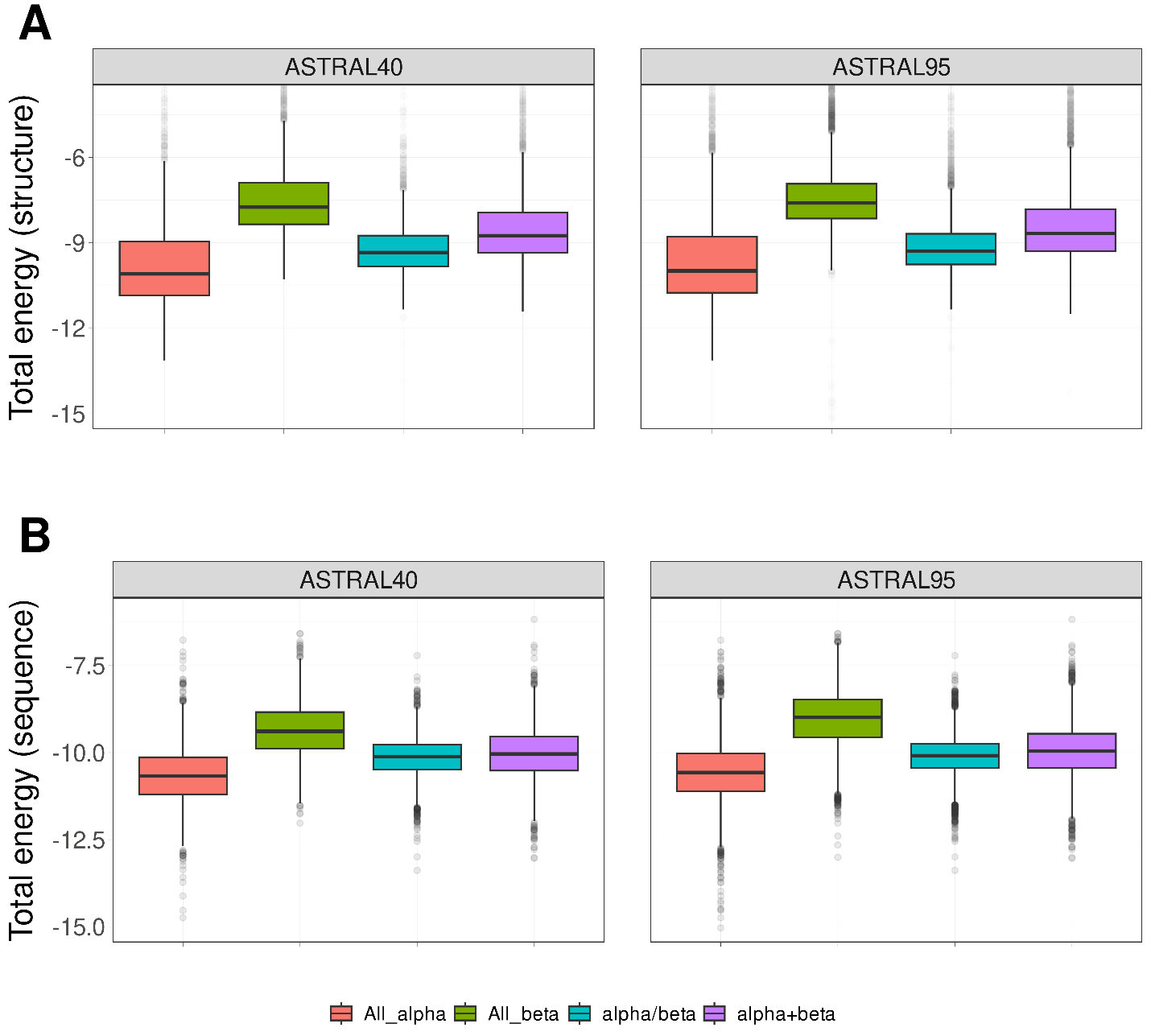
**Figures and Tables**

**Figure 1.**



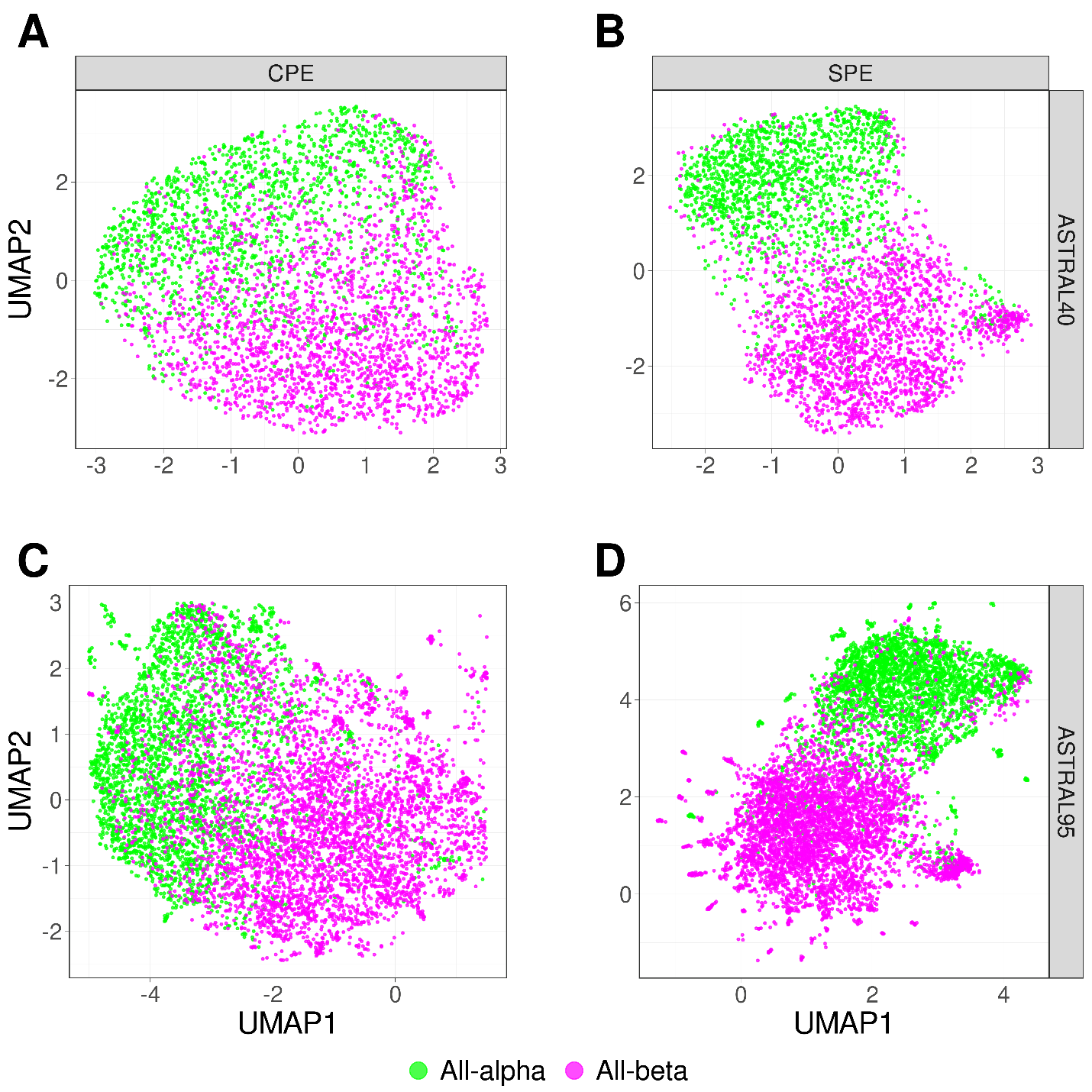
A) The correlation between total energy estimates derived from protein structure and sequence for protein domains within ASTRAL40(left) and ASTRAL95(right) data sets. B) The correlation between the distances of profile of energy estimated from sequence and structure for all pairs of domains in ASTRAL40(left) and ASTRAL95(right).

**Figure 2.**



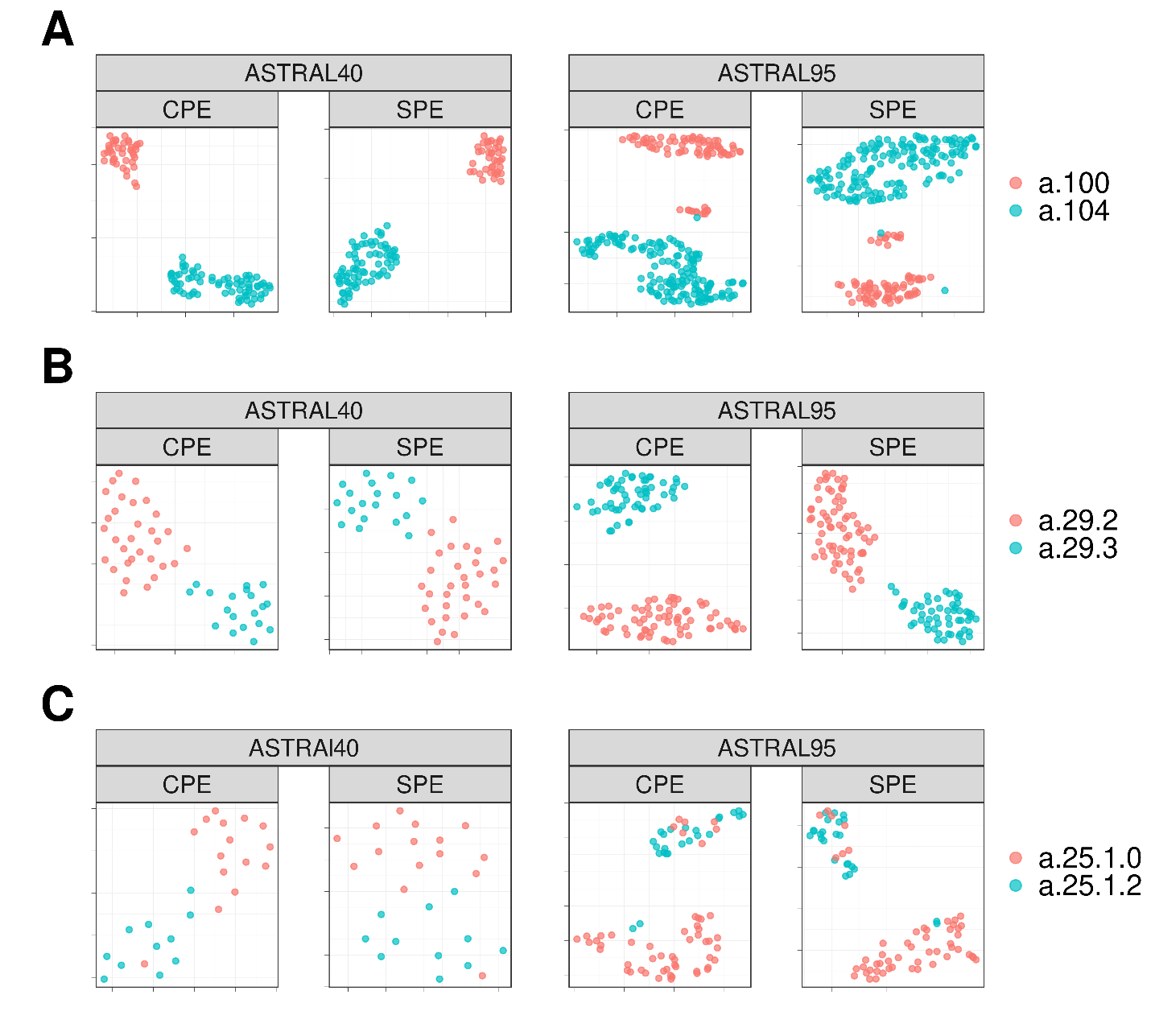
The distribution of normalized total energy in protein domains from ASTRAL40 and ASTRAL95 datasets based on protein structure (A) and sequence (B) across various structural scope classes. In the ASTRAL40 dataset, there are 2644, 3059, 4463, and 3653 protein domains in the all-alpha, all-beta, alpha+beta, 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.

**Figure 3.**



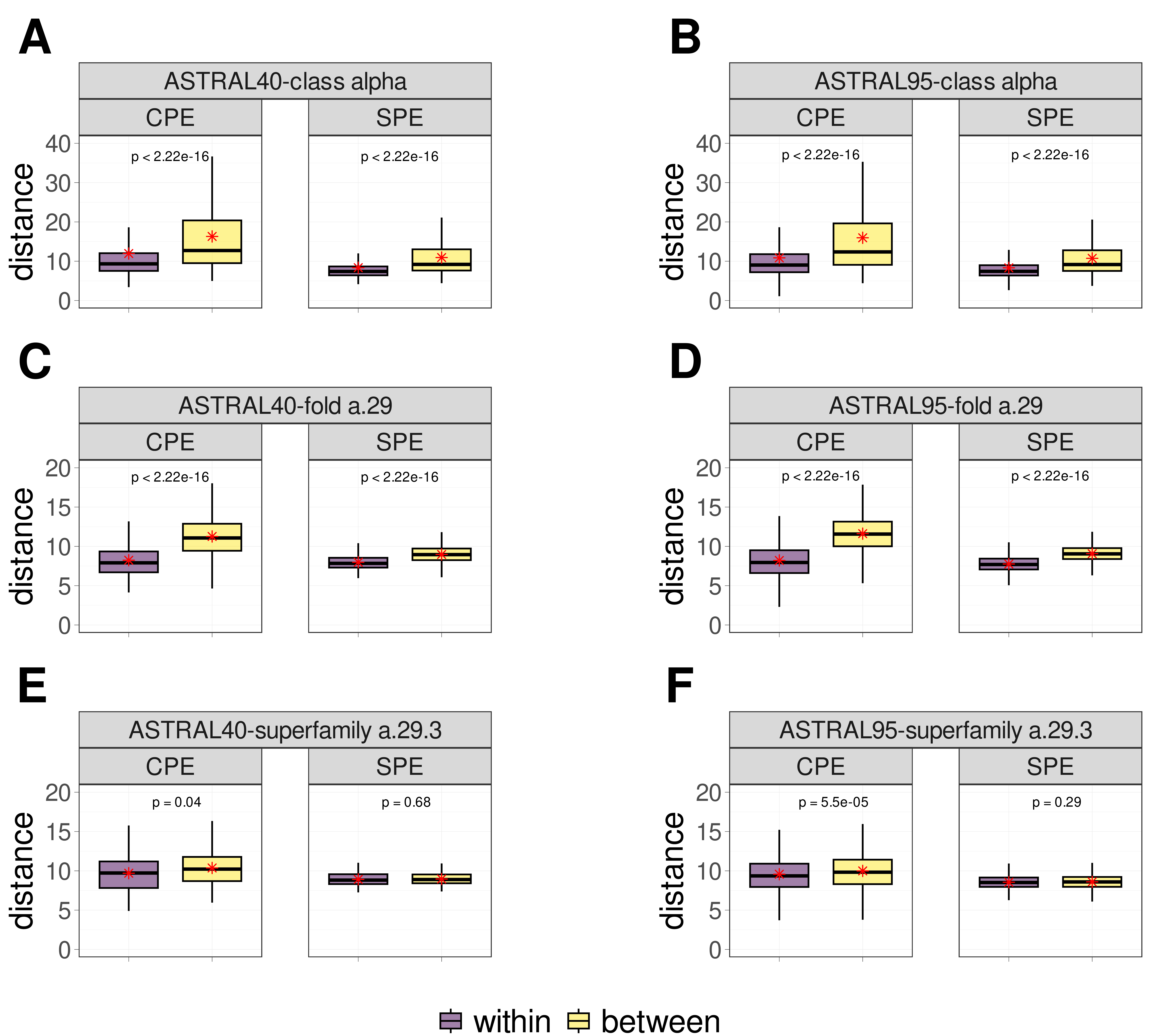
UMAP projection of SPE and CPE shows the separation of the all-alpha (green point) and all-beta (pink point) proteins selected from the ASTRAL40 and ASTRAL95 datasets. A) CPE of ASTRAL40, B) SPE of ASTRAL40, C) CPE of ASTRAL95, and D) SPE of ASTRAL95. Dots represent two dimensional UMAP projection of SPE(CPE) for individual sequences. UMAP plots were generated by parameters n\_neighbors = 20 and min\_dist = 0.1.

**Figure 4.**



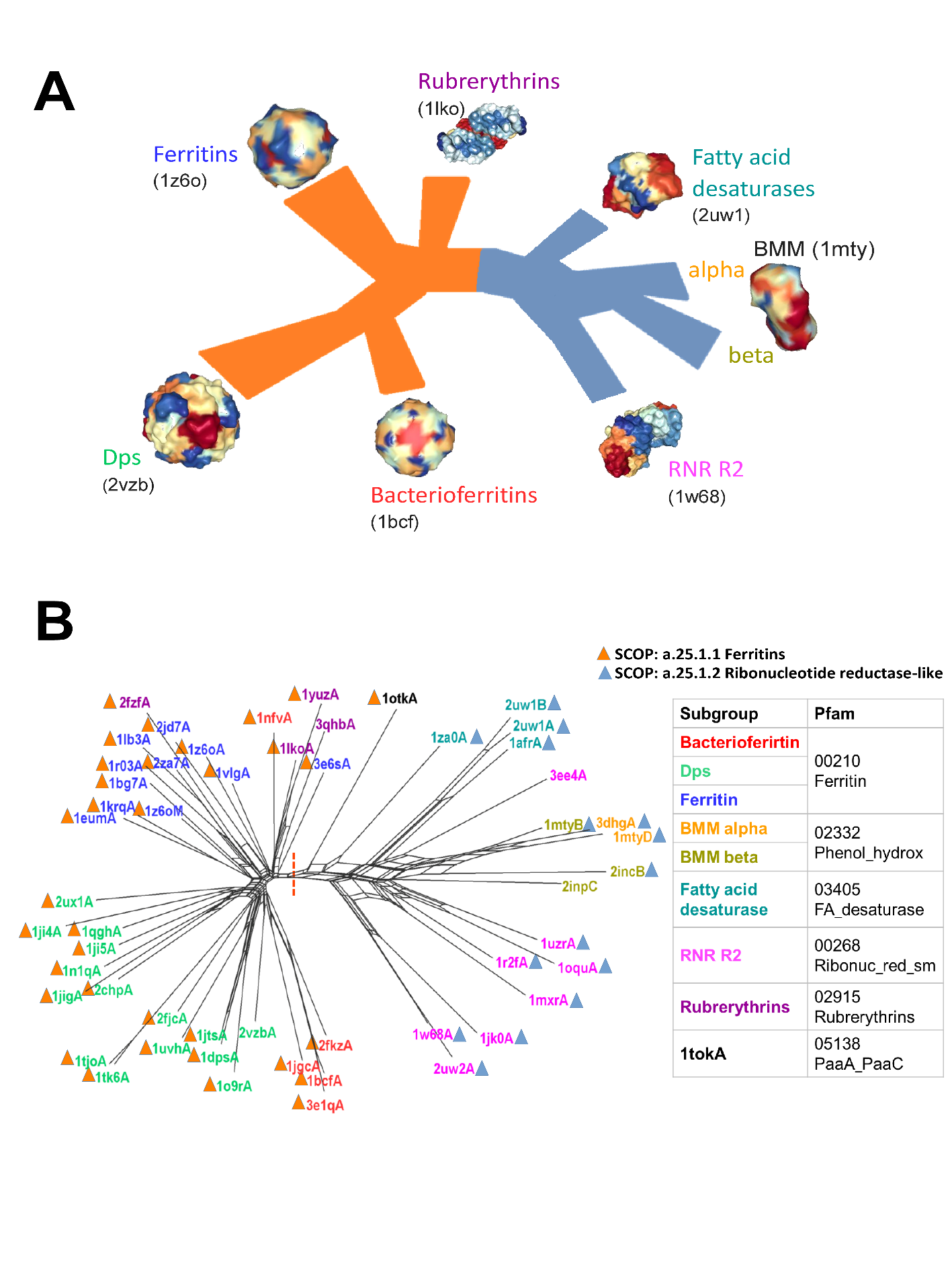
The UMAP projection of Structural Energy Profiles (SPE) and Compositional Energy Profiles (CPE) of protein domains from ASTRAL40 and ASTRAL95 represents the structural information embedded in energy profiles across hierarchical levels of SCOP; each panel includes two figures, one generated by CPE (left 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.

**Figure 5.**



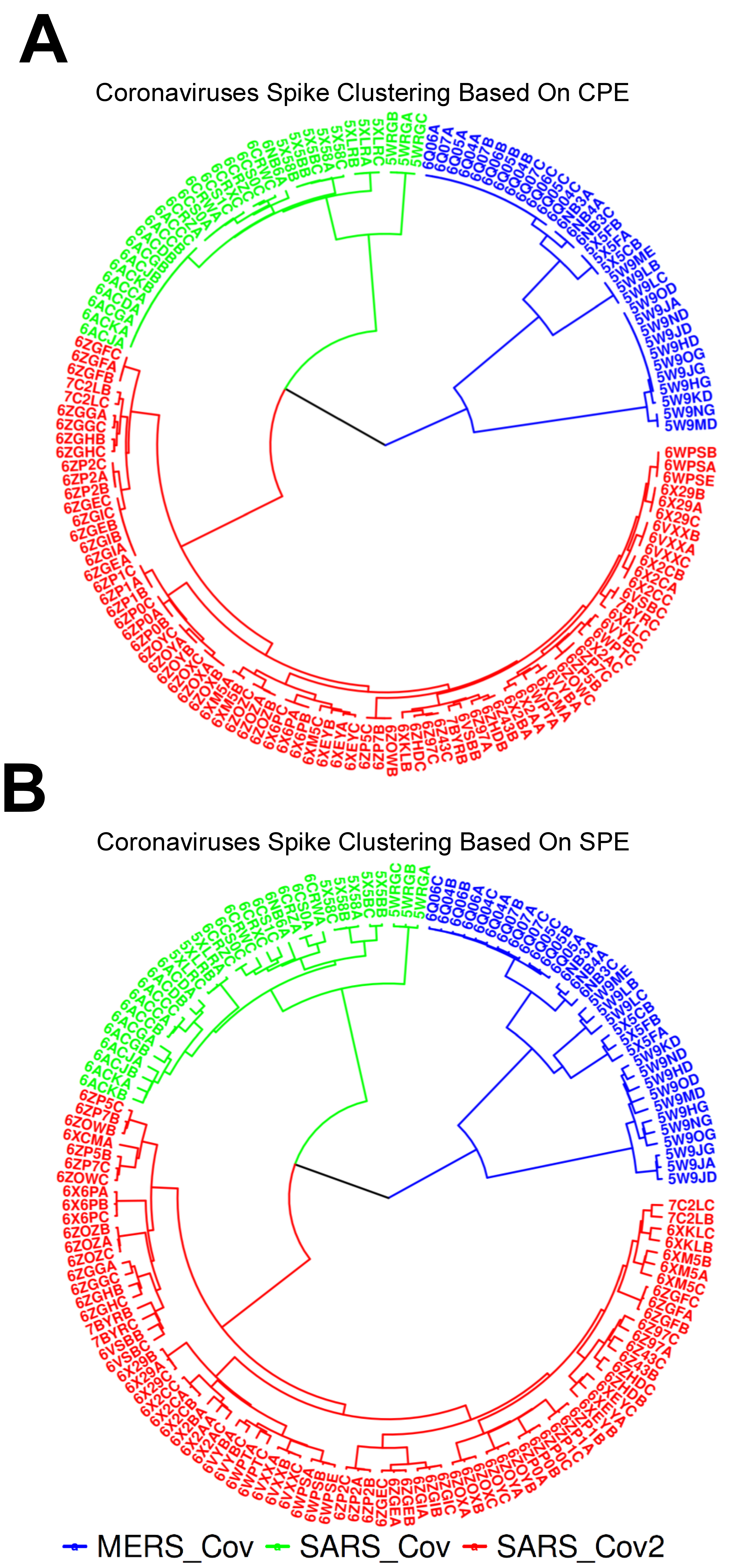
Comparative Boxplots of Pairwise Distances among Energy Profiles in ASTRAL40 and ASTRAL95, depicting A-B) intraclass distances within the all-alpha class (in purple) versus interclass distances (in yellow), C-D) intraclass distances within the a.29 fold (in purple) versus distances from protein domains 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).

**Figure 6.**



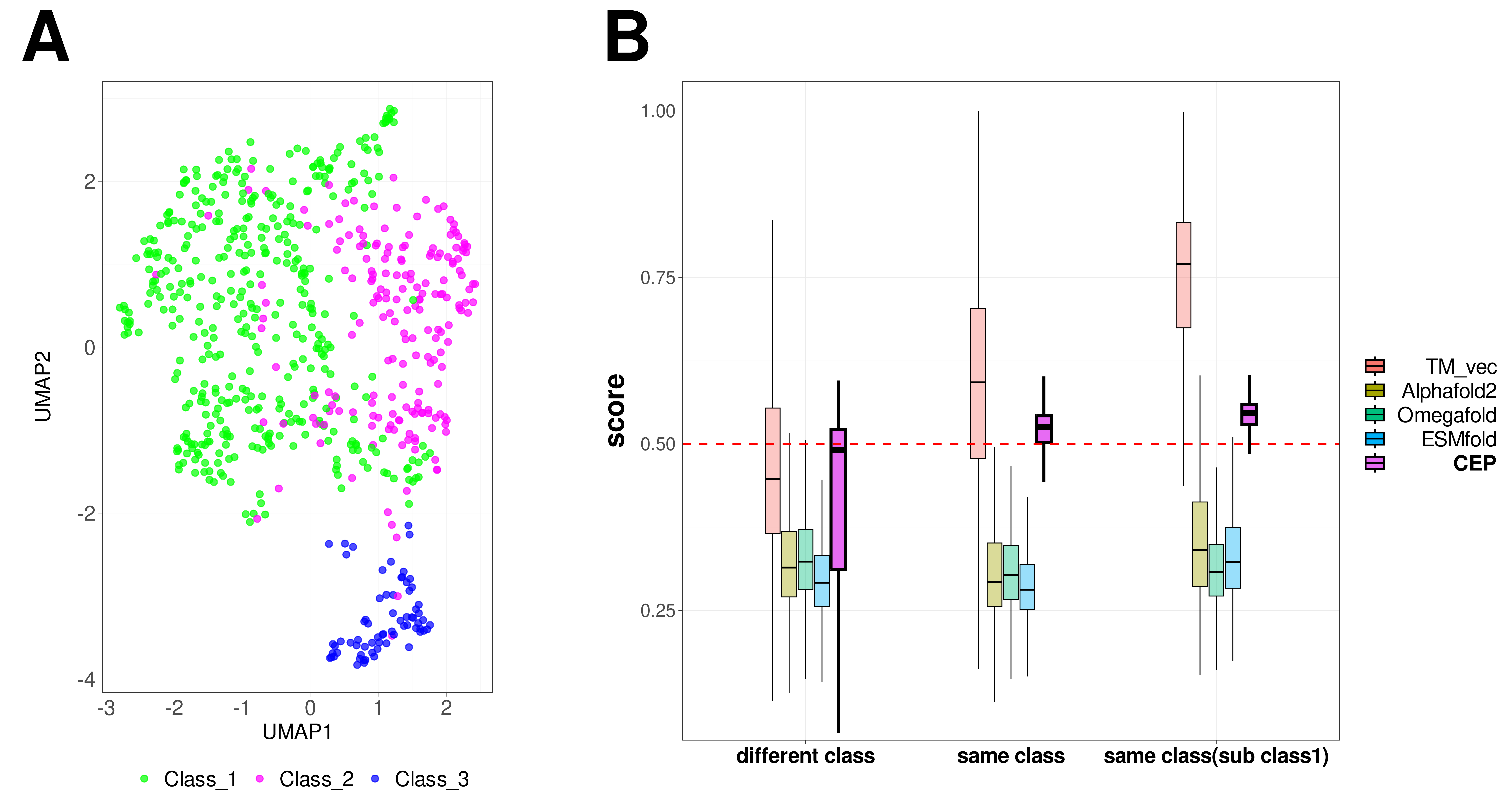
Energy-based phylogenetic network for the ferritin-like superfamily. A) Schematic view of the relationships between the major ferritin-like protein families. B) The network demonstrates the distinct separation (red dotted line) of two SCOP families: ferritins (SCOP id a.25.1.1), which includes Bacteri, 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.

**Figure 7.**



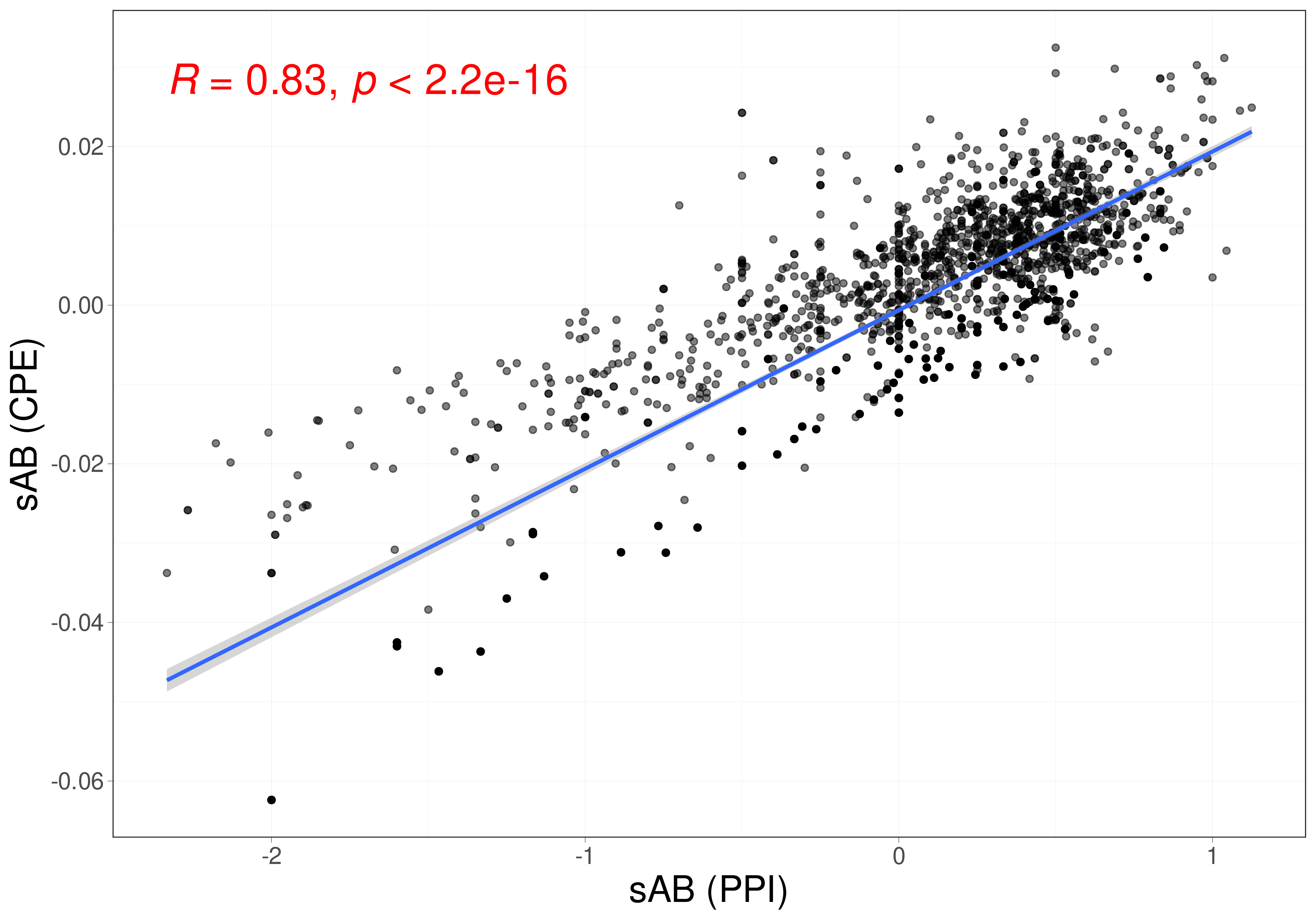
The dendrograms illustrate the clustering of spike glycoprotein structures from three viruses SARS-CoV, SARS-CoV-2, and MERS-CoV. The clustering is based on pairwise distances of energy profiles derived from A) protein structure and B) protein sequence. Each leaf on the dendrogram is labeled with the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

**Figure 8.**



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 CEP distances (CEP\_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. CEP\_dis is normalized by min-max normalization.

**Figure 9.**



The correlation between separation distances estimated by protein-protein interaction network (X-axis) and the distance between profiles of energies (Y-axis).

**Table 1.** 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** | **Accuracy** | **F1 Measure** | | | | |
| **wigend\_helix** | **PH.domain-like** | **NTF-like** | **Ubiquitin-like** | **Immunoglobulins** |
| **1NN** | 0.98 | 0.98 | 0.96 | 0.99 | 0.99 | 0.99 |
| **RF** | 0.99 | 0.97 | 0.97 | 0.99 | 0.99 | 0.99 |

**Table 2.** The accuracy and computation time for 1-NN classifier based on GR-Align, RMSD, TM-score, Yau-Hausdorff distance, and the distance between profiles of energy (CPE) as a measure of protein dissimilarity.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Method** | **GR-Align** | **RMSD** | **TM-Score** | **YH (10Rotation)** | **YH (2500Rotation)** | **CPE** |
| **Accuracy** | 62.3% | 59.2% | 61.5% | 70.8% | 81.5% | 97% |
| **Time** | 2 min | 1 h | 9h 20 min | 10 min | 4h 10 min | 3 min |