**Comments from the Reviewers:**

**Reviewer 1:**

The authors presented a method to characterize protein-protein similarity using an energy profile. Although the proposed method is interesting, more systematic evaluations using larger datasets are needed. In addition, there are no details about how the authors obtained the energy predictor matrix to calculate sequence-based energy for each evaluation test presented in the manuscript.

*ANSWER: We thank the reviewer for appreciating our work and for very constructive feedback. We have carefully considered the comments and provide our detailed point-by-point responses below. To facilitate the review, we have included new data and figures/tables into our point-by-point responses below and updated the revised manuscript accordingly.*

1) More details about how the authors obtained the energy predictor matrix for calculating sequence-based energy for each evaluation test are needed. For instance, the testing dataset shouldn't be used to estimate the energy predictor matrix.

*ANSWER*: We thank the reviewer for this suggestion and have now excluded the 129 proteins that overlapped with the test sets from the training set. The knowledge-based potential function and energy predictor matrix are now recalculated using only the remaining training set proteins, ensuring that the test data is not involved in the matrix estimation process. In the revised manuscript, we have also updated the Methods section (Page 18-20) and added new **Fig. 1**, which illustrates the steps for estimating the predictor matrix and energy profile.

The recalculated predictor matrix showed a correlation of 0.99990 with the original one, indicating that the exclusion of overlapping proteins had a negligible impact on the energy predictions. In addition, we regenerated all figures and reran the method across all test sets, with results remaining consistent with our previous findings, confirming the robustness of our approach.

2) All the evaluations are based on small datasets. Evaluations and comparisons with other methods on larger datasets are needed.

*ANSWER*: We have now conducted an additional evaluation using a larger dataset and compared our results with the TM-Vec method. Given that structure-based methods such as RMSD and TM-score are time-intensive, we focused on TM-Vec for its balance of efficiency and accuracy. For the larger dataset, we employed the coronavirus dataset previously generated and analyzed by Freiberger et al., as suggested by Reviewer 2. These new results are now included in **Section 3.8** of the manuscript as displayed below:

**3.8 Large-Scale Application of Family Detection in Coronaviruses**

To evaluate our method on a larger dataset, we utilized a coronavirus dataset that was previously generated and analyzed by Freiberger et al.27 Initially, they retrieved homologous sequences from the SARS-CoV-2 reference genome MN985325, with low-quality and non-Coronaviridae sequences excluded37. Then sequences aligned using MAFFT software38, with non-structural protein sequences trimmed to focus on regions specifically relevant to SARS-CoV-2. These aligned sequences were then clustered using CD-Hit39 to ensure high similarity within each cluster. Structural models for these sequences were generated using AlphaFold2, retaining only those that met stringent quality criteria. The high-quality models were subsequently grouped into subfamilies using S3Det software40, allowing for the identification of Specificity Determining Positions (SDPs) within the proteins. This meticulous process resulted in a final set of 28 high-quality protein families, comprising 4,405 protein models. For each protein in this dataset, we computed the energy profiles CPE and SPE, along with the distance between these profiles and the cosine similarity between pairs of TM-Vec representations. The 1-nearest neighbor (1-NN) method was then utilized to classify the proteins into different families. The results, shown in Table 5, include metrics for accuracy and F1-score, demonstrating the effectiveness of our model. Both methods achieved performance levels near 100%, with CPE offering faster performance. Detailed outcomes of the 1-NN classification are provided in Supplementary Tables S3-S5, and the UMAP projections of SPE, CPE, and TM-Vec representations are displayed in Supplementary Figures S6-S8.

**Table 4 |** Large-scale analysis of the SARS-CoV-2 proteome across 28 families, covering 4,405 proteins. The overall accuracy, F1-score, and computation time for detecting families using the 1-NN classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Time (Sec)** | **Accuracy** | **F1 Measure** |
| **CPE** | 49 | 0.9968 | 0.9946 |
| **SPE** | 2715 | 0.9973 | 0.9903 |
| **TM\_Vec** | 685 | 0.9966 | 0.9925 |

**A screen shot of a chart

Description automatically generated**

**Fig. S6| UMAP Visualization of Energy Profiles in Large-Scale SARS-Cov2 data set.** The UMAP projection of Structural Energy Profiles (SPE) on 28 protein families with a total of 4,405 protein models. UMAP plot was generated using parameters n\_neighbors = 150 and min\_dist = 0.1.

A screen shot of a chart

Description automatically generated

**Fig. S7| UMAP Visualization of Energy Profiles in Large-Scale SARS-Cov2 data set.** The UMAP projection of Compositional Energy Profiles (CPE) on 28 protein families with a total of 4,405 protein models. UMAP plot was generated using parameters n\_neighbors = 150 and min\_dist = 0.1.

**A screen shot of a chart

Description automatically generated**

**Fig. S8| UMAP Visualization of Energy Profiles in Large-Scale SARS-Cov2 data set.** The UMAP projection of TM-Vec on 28 protein families with a total of 4,405 protein models. UMAP plot was generated using parameters n\_neighbors = 150 and min\_dist = 0.1.

The Papain-like Protease (PLPro) domain plays a crucial role in viral replication by catalyzing the proteolysis of viral polyproteins. In addition, PLPro interacts with two host proteins, ubiquitin (Ub) and the ubiquitin-like interferon-stimulated gene 15 protein (ISG15), allowing the virus to evade or weaken the host immune response. In this dataset, PLPro is divided into four subfamilies aligned with the Betacoronavirus subgenera: Sarbecovirus (n = 31), Nobecovirus (n = 11), Merbecovirus (n = 35), and Embecovirus (n = 45). The UMAP representation for both CPE and SPE is shown in Fig. 13, where proteins from each subfamily are clearly clustered together.



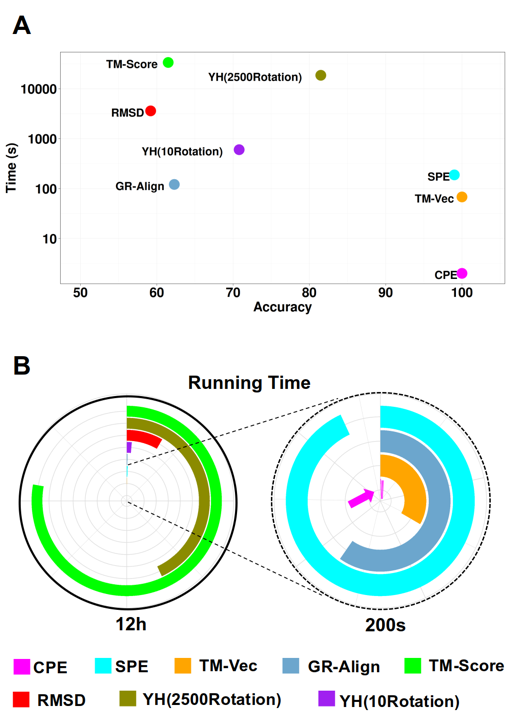
**Fig. 13| Papain-like Protease (PLPro) domains across Betacoronavirus subgenera.** The UMAP projection shows clustering of PLPro domains from Sarbecovirus (n = 31), Nobecovirus (n = 11), Merbecovirus (n = 35), and Embecovirus (n = 45) using TMVec, CPE and SPE representations. n\_neighbors = 13, min\_dist = 0.5.

3) Structural-based (SPE) and sequence-based (CPE) methods should be compared in each evaluation test. For example, CPE is compared with other methods in Figure 6 but not SPE. SPE is compared with other methods in Figure 7 but not CPE.

ANSWER: We have now included comparisons between **CPE** and **SPE** across the appropriate evaluation tests. Results are visualized on new **Fig. 7** (previously **Fig. 6**) and in **Table 1**. We find that (Table 1 and new Fig. 7) CPE method achieves a computation time faster than TM-Vec and that CPE and SPE methods significantly outperform GR-Align, RMSD, TM-Score, and YH in terms of accuracy and efficiency, highlighting advantages of CPE and SPE.

**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 SPE and CPE as a measure of protein dissimilarity.

|  |  |  |
| --- | --- | --- |
| **Method** | **Accuracy** | **Time** |
| **GR-Align** | 62.3% | 120 sec |
| **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 |
| **SPE** | 99% | 187 sec |



**Fig. 7 | 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 SPE and 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.

We have also updated Table 2 to include the SPE method results for distinguishing between 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).

**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) based on CPE, SPE, and TM-Vec.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |
| **SPE (1NN)** |  | 0.98 | 0.98 | 0.96 | 0.99 | 0.98 | 0.98 |
| **SPE (RF)** |  | 0.98 | 0.97 | 0.95 | 0.99 | 0.98 | 0.99 |
| **TM\_Vec** **(1NN)** | 955 Sec | 0.99 | 0.99 | 1 | 1 | 0.99 | 0.99 |

We have also included the CPE results for the phylogenetic network reconstruction of the ferritin-like superfamily in Fig. 9 (previously Fig. 7). A more detailed comparison of the performance of these methods is provided in the response to C5.

4) For the spike protein classification, the author may want to compare with a sequence similarity-based method. In addition, the order of the clusters in hierarchical clustering doesn’t represent the similarity among the clusters. If the authors want to make a statement about MERS\_Cov involved to SARS\_Cov, then SARS\_Cov2, a better way is to examine the similarity among these three groups.

*ANSWER*: Thank you for this very helpful comment. Section 3.5 has now been revised to meet the requirements. In response, we have added a phylogenetic tree constructed using sequence similarity-based methods. The multiple sequence alignment of spike proteins was performed using the **msa** package in R with the ClustalW method, and protein distances were calculated using the **seqinr** package with the number of identities as the distance metric. Phylogenetic trees were then built using the UPGMA method in the **phangorn** package. The resulting tree, based on sequence similarity, is presented in new **Fig. 10B**. To address the point about the order of clusters in hierarchical clustering, we recognize that the clustering order does not always accurately represent evolutionary relationships. Therefore, we have placed greater emphasis on using the phylogenetic trees from various methods (CPE, protein sequence, TM-Vec, SPE, RMSD, and TM-Score) to assess the relationships between these virus groups. In particular, the phylogenetic trees allow us to more accurately examine the similarity among SARS-CoV, SARS-CoV-2, and MERS-CoV.

Additionally, we computed the average distance between these three virus groups across all five methods—SPE, CPE, Sequence similarity-based, TM-Vec, and RMSD—to ensure consistency in our findings. All methods consistently show that SARS-CoV-2 is more closely related to SARS-CoV than to MERS-CoV (new **Fig. 10I**). Furthermore, the distance between the SARS-CoV-2 and MERS-CoV groups is nearly identical to the distance between SARS-CoV and MERS-CoV, confirming that both groups are similarly distant from MERS-CoV.

**Table 4 |** Comparison of clustering results using Adjusted Rand Index (ARI)

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Type** | **Cut Tree** | **ARI** |
| **CPE** | Sequence-Based | 4 | 0.95 |
| **MSA** | Sequence-Based | 3 | 0.49 |
| **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 |

A collage of different types of graphs

Description automatically generated

**Fig. 10 | Clustering analysis of spike glycoprotein structures from SARS-CoV, SARS-CoV-2, and MERS-CoV.** The dendrograms depict the clustering of spike glycoprotein structures from the three viruses: SARS-CoV, SARS-CoV-2, and MERS-CoV. The clustering is based on pairwise distances calculated from different methods: **A)** CPE, **B)** protein sequence, **C)** TM-Vec, **D)** SPE, **E)** RMSD, and **F)** TM-Score. The leaves of each tree are color-coded to indicate the originating virus for each spike glycoprotein structure. **G)** Displays the ARI values for each method, **H)** shows the running time associated with each method, and **I)** presents the average distance between the three virus groups as calculated by each method.

5) In new Figure 7, TM-Vec seems to do a better job than SPE. The authors may want to provide summary statistics to compare the performance of the two methods.

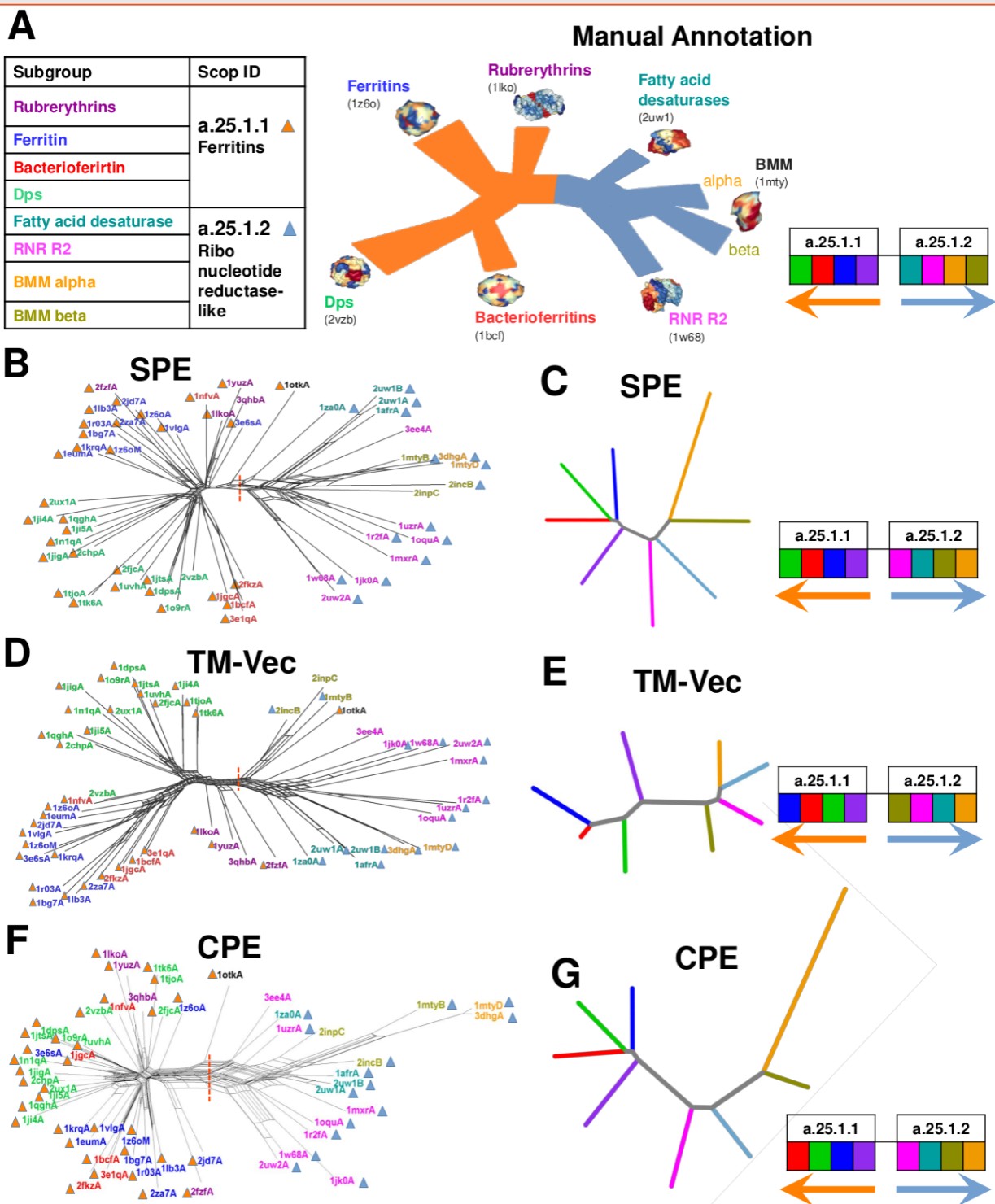
*ANSWER:* Thank you for your valuable comment. In line with the suggestion from **comment 3**, we have included the results for the **CPE** method in the comparison. To provide a comprehensive evaluation, we computed the average distances between proteins from each of the eight subfamily groups using three methods: CPE, SPE, and TM-Vec. These distance matrices were used to construct neighbor-joining phylogenetic trees, as shown in new **Fig. 9 C, E, and G**. To assess the accuracy of the subfamily arrangements, we compared the inferred phylogenetic trees with the manually annotated subfamily relationships, with color-coding corresponding to subgroups in **Fig. 9A**. The order of subgroups in the manually annotated network was then compared with the order produced by SPE, TM-Vec, and CP**E** (**Fig. 9 C, E, G**).

We found that while TM-Vec performs well in clustering similar proteins, it struggles to accurately reflect the evolutionary relationships between subfamilies. For instance, in the a.25.1.2 family, TM-Vec incorrectly places the subgroup BMM-b first, whereas it should appear last in the manual network. Similarly, in the a.25.1.1 family, although TM-Vec correctly identifies the initial branch, it misorders the remaining subgroups. To quantify these deviations, we calculated the Spearman correlation between the predicted branching orders from each method and the manual network. These correlations, summarized in Table 3, demonstrate that both SPE and CPE outperform TM-Vec in aligning with the manually annotated network. Notably, SPE achieves a perfect correlation in the a.25.1.1 family, while TM-Vec shows a negative correlation in the a.25.1.2 family, illustrating its limitations.

**Table 3 |** Spearman Correlation Between the Manual Network and Predicted Branching Orders by SPE, TM-Vec, and CPE.

|  |  |  |
| --- | --- | --- |
| **Method** | **a.25.1.1** | **a.25.1.2** |
| **CPE** | 0.8 | 0.6 |
| **SPE** | 1 | 0.6 |
| **TM\_Vec** | 0.2 | -0.4 |

These updates provide a more comprehensive comparison between the sequence-based and structure-based approaches, as requested, offering a clearer understanding of their relative performance across all tests.



**Fig. 9 | Phylogenetic network reconstruction of the ferritin-like superfamily.** A) Schematic representation of the relationships among major ferritin-like protein families. B) Phylogenetic tree reconstructed using SPE. C) Neighbor-joining tree generated based on the average distances between subgroups using SPE. D) Phylogenetic tree reconstructed using TM-Vec. E) Neighbor-joining tree generated using average distances between subgroups with TM-Vec. F) Phylogenetic tree reconstructed using CPE. G) Neighbor-joining tree based on the average distances between subgroups using CPE. The red dotted line highlights the clear separation between two SCOP families: ferritins (SCOP ID a.25.1.1), which includes the Bacterioferritin, Ferritins, Dps, and Rubrerythrin subgroups, and the Ribonucleotide Reductase-like family (SCOP ID a.25.1.2), which includes the BMM-alpha, BMM-beta, Fatty\_acid, and RNRR2 subgroups. The inferred arrangement of subfamilies using each method is shown to the right of C, E, and G.

**Reviewer 2:**

This is an interesting paper highlighting the use of energy profile-based analysis to examine evolutionary questions. The energy profile as a way to analyze sequence structure mapping was introduced by Eisenberg in Science 253,164(1991). Soon after Wolynes and coworkers showed how such profiles could be optimized using energy landscape theory in PNAS89,9029(1992) achieving success comparable to what is discussed here, but without the evolutionary aspects. These pioneering works should be cited.

It seems also the evolutionary analysis resembles the FRUSTRAEVO work by Ferreiro and coworkers that recently appeared in Nature Communications. This work should be compared and cited. The examples chosen are quite nice.

ANSWER: We would like to thank the reviewer for thoughtful and constructive feedback. We agree that the pioneering studies by Eisenberg (1991) and Wolynes et al. (1992), should be cited and contextualized properly and have done so in the revised version. We have now incorporated the following discussion into the introduction (Page 3 of the revised ms):

"The concept of using energy profiles to evaluate protein structures was initially introduced by Eisenberg10, who developed a method for mapping amino acid sequences to structural folds based on energy profiles. This approach enabled an early computational framework for assessing the compatibility of protein sequences with specific structural conformations. Soon after, Wolynes and co-workers11 expanded this study by applying energy landscape theory, utilizing optimized Hamiltonians to predict protein folding pathways. They introduced the spin-glass model to navigate the complex energy landscape of protein folding, ensuring that the native fold represents a global energy minimum. These pioneering methods laid the groundwork for modern approaches that predict protein structures using energy-based techniques."

Regarding your second point about the evolutionary analysis resembling the FRUSTRAEVO work by Ferreiro and coworkers, we acknowledge the relevance of this recent study.

Ferreiro et al. developed a method to assess frustration levels across multiple sequence alignments (MSAs) using various frustration indices, such as the Single Residue Frustration Index and pairwise interaction indices, to identify conserved frustration patterns. This approach is implemented through tools like FrustraEvo and FrustratometeR, which provide detailed analyses of evolutionary divergence within protein families. They demonstrated how local frustration can influence protein foldability and function, with case studies on α and β globins, as well as the SARS-CoV-2 dataset. By assigning a frustration score to each position, their method allowed for a comprehensive analysis of the SARS-CoV-2 proteome, identifying functionally relevant sites within viral proteins based on conserved frustration patterns. In contrast, our study assigns an energy profile to each protein rather than generating a score for individual positions. Our method represents each protein as a comprehensive energy profile, asserting that proteins within different families, superfamilies, or folds exhibit different energy profiles. While both approaches share the assumption that local energetic frustration reflects evolutionary and functional constraints, the key difference lies in implementation: Ferreiro et al.'s method assigns frustration scores to each position, enabling the identification of critical residues for protein function. In our approach, however, we analyze the overall energy profile of proteins, focusing on broader structural and evolutionary patterns rather than pinpointing individual functional residues.

We utilized a coronavirus dataset generated and analyzed by Freiberger et al. In the revised manuscript, we have included a new section (Section 3.8) titled "Large-Scale Application of Family Detection in Coronaviruses," which presents the results of our method in comparison with TM-Vec.

**3.8 Large-Scale Application of Family Detection in Coronaviruses**

To evaluate our method on a larger dataset, we utilized a coronavirus dataset that was previously generated and analyzed by Freiberger et al.27 Initially, they retrieved homologous sequences from the SARS-CoV-2 reference genome MN985325, with low-quality and non-Coronaviridae sequences excluded37. Then sequences aligned using MAFFT software38, with non-structural protein sequences trimmed to focus on regions specifically relevant to SARS-CoV-2. These aligned sequences were then clustered using CD-Hit39 to ensure high similarity within each cluster. Structural models for these sequences were generated using AlphaFold2, retaining only those that met stringent quality criteria. The high-quality models were subsequently grouped into subfamilies using S3Det software40, allowing for the identification of Specificity Determining Positions (SDPs) within the proteins. This meticulous process resulted in a final set of 28 high-quality protein families, comprising 4,405 protein models. For each protein in this dataset, we computed the energy profiles CPE and SPE, along with the distance between these profiles and the cosine similarity between pairs of TM-Vec representations. The 1-nearest neighbor (1-NN) method was then utilized to classify the proteins into different families. The results, shown in Table 5, include metrics for accuracy and F1-score, demonstrating the effectiveness of our model. Both methods achieved performance levels near 100%, with CPE offering faster performance. Detailed outcomes of the 1-NN classification are now provided in new Supplementary Tables S3-S5, and the UMAP projections of SPE, CPE, and TM-Vec representations are displayed in new Supplementary Figures S6-S8.

**Table 4 |** Large-scale analysis of the SARS-CoV-2 proteome across 28 families, covering 4,405 proteins. The overall accuracy, F1-score, and computation time for detecting families using the 1-NN classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Time (Sec)** | **Accuracy** | **F1 Measure** |
| **CPE** | 49 | 0.9968 | 0.9946 |
| **SPE** | 2715 | 0.9973 | 0.9903 |
| **TM\_Vec** | 685 | 0.9966 | 0.9925 |

**A screen shot of a chart

Description automatically generated**

**Fig. S6| UMAP Visualization of Energy Profiles in Large-Scale SARS-Cov2 data set.** The UMAP projection of Structural Energy Profiles (SPE) on 28 protein families with a total of 4,405 protein models. UMAP plot was generated using parameters n\_neighbors = 150 and min\_dist = 0.1.

A screen shot of a chart

Description automatically generated

**Fig. S7| UMAP Visualization of Energy Profiles in Large-Scale SARS-Cov2 data set.** The UMAP projection of Compositional Energy Profiles (CPE) on 28 protein families with a total of 4,405 protein models. UMAP plot was generated using parameters n\_neighbors = 150 and min\_dist = 0.1.

**A screen shot of a chart

Description automatically generated**

**Fig. S8| UMAP Visualization of Energy Profiles in Large-Scale SARS-Cov2 data set.** The UMAP projection of TM-Vec on 28 protein families with a total of 4,405 protein models. UMAP plot was generated using parameters n\_neighbors = 150 and min\_dist = 0.1.

The Papain-like Protease (PLPro) domain plays a crucial role in viral replication by catalyzing the proteolysis of viral polyproteins. In addition, PLPro interacts with two host proteins, ubiquitin (Ub) and the ubiquitin-like interferon-stimulated gene 15 protein (ISG15), allowing the virus to evade or weaken the host immune response. In this dataset, PLPro is divided into four subfamilies aligned with the Betacoronavirus subgenera: Sarbecovirus (n = 31), Nobecovirus (n = 11), Merbecovirus (n = 35), and Embecovirus (n = 45). The UMAP representation for both CPE and SPE is now illustrated in new Fig. 13, where proteins from each subfamily are clustered together.



**Fig. 13| Papain-like Protease (PLPro) domains across Betacoronavirus subgenera.** The UMAP projection shows clustering of PLPro domains from Sarbecovirus (n = 31), Nobecovirus (n = 11), Merbecovirus (n = 35), and Embecovirus (n = 45) using TMVec, CPE and SPE representations. n\_neighbors = 13, min\_dist = 0.5.

**Reviewer 3:**

The study presents a new approach for comparing protein structures using energy profiles derived from amino acid sequences, which offer a faster and computationally efficient alternative to traditional structural alignment methods. By generating 210-dimensional profiles based on pairwise amino acid interactions, the method accurately reflects structural and evolutionary relationships among proteins, as validated by strong correlations with benchmark structural data. It effectively distinguishes protein families across hierarchical levels and has shown success in clustering proteins from viruses and bacteriocins. Additionally, the method's efficiency extends to applications in drug combination prediction, highlighting its versatility and potential in large-scale protein analysis. The manuscript is commendable clear and the logical order in which the concepts and findings are presented greatly facilitates understanding.

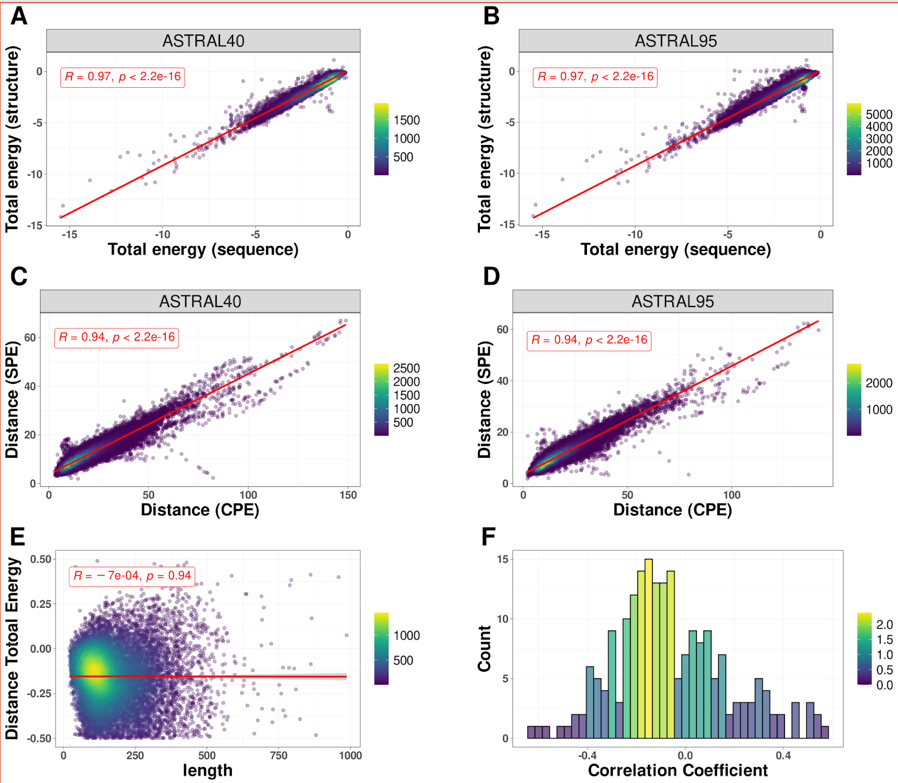
ANSWER: We would like to thank the reviewer for the positive feedback and for recognizing the value of our work.

1) On page 5, the authors presented a useful comparison between structural and sequence-based energy calculations, suggesting that sequence-based energy estimation serves as a reliable approximation. However, it would be beneficial to address potential limitations or assumptions in this comparative analysis. In particular, discussing how sequence length and protein complexity might impact the accuracy of energy estimates would provide a more comprehensive understanding of the strengths and weaknesses of each method.

ANSWER: We thank the reviewer for the valuable suggestion to address the limitations and assumptions in our comparison between structural and sequence-based energy calculations. We have added the following section and discussion to the revised version of the manuscript at page 6.

" We also computed the total energy for protein sequences and their corresponding structures using protein domains from the ASTRAL40 dataset. We then analyzed the differences between these two energy estimates. As illustrated in new Fig. 2C, we specifically investigated the correlation between these energy differences and protein length. Our results show no significant correlation, indicating that the accuracy of sequence-based total energy estimates is not affected by protein length. This suggests that sequence-based energy calculations provide a reliable approximation of structural energies, even for proteins of varying lengths. To further support our findings, we extended our analysis to examine energy discrepancies for all interaction types. For each interaction type, we calculated the difference between energy estimates derived from sequence and structure. As shown in new? Fig. 2D, for 96% of interaction types, the correlation between these energy differences and protein length was less than 0.5. This indicates that protein length has no effect on the accuracy of energy estimates across most interaction types, reinforcing our conclusion that sequence-based energy approximations are robust across different protein interactions. The scatter plots for all 210 interaction types are provided in new Supplementary Figure S1.

However, while protein length does not seem to influence the accuracy, we recognize that protein complexity—such as folding patterns, structural heterogeneity, and conformational dynamics—may indeed play a role in the precision of energy estimates. Therefore, exploring the impact of protein complexity on energy estimations will be a valuable direction for future research. We have now added this angle into the revised Discssuion section (when you feel it is relevant to add, with half or maybe with one-two sentences).



**Fig. 2 | Sequence-Structure relationship.** The correlation between total energy estimates derived from protein structure and sequence for protein domains within A) ASTRAL40 and B) ASTRAL95 data sets. The correlation between the distances of profile of energy estimated from sequence (CPE) and structure (SPE) for all pairs of domains in C) ASTRAL40 and D) ASTRAL95. E) The correlation between the difference in total energy (from sequence and structure) and protein length. F) Histogram showing the distribution of correlation coefficients between the difference in energy estimates (from sequence and structure) and protein length across all 210 pairwise interactions.

2) The manuscript effectively compares the new method with established measures (GR-Align, RMSD, TM-Score, Yau-Hausdorff distance, TM-Vec). However, additional details regarding the specific conditions or parameters used to adjust each benchmarked method, as well as any inherent biases or limitations, would enhance the reader's understanding of the comparative evaluation.

ANSWER: We would like to thank the reviewer for very helpful suggestions. We have revised the manuscript accordingly to include these details. We have added the following discussion to the revised version of the manuscript at page 7-8.

“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)21 quantifies the average spatial variance between corresponding atoms or components within superimposed proteins, providing a fundamental measure of structural deviation. For our analysis, RMSD calculations were performed by superimposing corresponding atomic coordinates using a least-squares fitting procedure implemented in R, focusing on backbone Cα atoms to assess the overall fold without the influence of side-chain orientations. While RMSD is widely used, it heavily relies on direct spatial overlap and is sensitive to outlier regions. This sensitivity often penalizes flexible regions or domain movements, such as hinge motions or flexible loops, potentially obscuring meaningful similarities in overall protein fold when local variations are present. The TM-score (Template Modeling score) 22 evaluates similarity by considering both residue-level alignment and overall topology, offering a nuanced assessment of structural resemblance. Unlike RMSD, TM-Score is less sensitive to domain-level movements but has its own limitations. Specifically, when comparing proteins with highly variable structures or multi-domain proteins, TM-Score tends to favor the alignment of larger structural elements, potentially leading to lower scores for proteins with significant domain rearrangements despite sharing similar overall folds. Additionally, TM-Score may not adequately account for functional sites that involve small but critical local structural differences. TM-Vec 23, a recent advancement, employs deep learning techniques trained on diverse protein structures to enhance accuracy and efficiency in similarity assessment. TM-Vec maps protein structures into a continuous vector space, allowing comparisons based on the distances between their vector representations. While highly accurate in detecting remote homology and structural similarities, TM-Vec’s reliance on training data introduces inherent biases. These biases are particularly evident when evaluating proteins with rare or novel folds that are underrepresented in the training set. Furthermore, as a black-box model, TM-Vec offers limited interpretability regarding the specific structural features contributing to the similarity scores, which can be a limitation when detailed structural insights are required. On the alignment front, we utilized GR-Align24 with its default parameters, employing the graphlet degree similarity (GDS) metric to capture topological similarities between protein structures. The GDS metric compares the distributions of small connected subgraphs (graphlets) within the protein structures. GR-Align is robust in identifying proteins with similar topological arrangements, but its sensitivity to minor structural variations can lead to higher false positives when comparing proteins with subtle differences in their tertiary structures. Additionally, GR-Align does not incorporate sequence information or account for conformational flexibility, which may limit its ability to discern functionally relevant structural variations, particularly those involving dynamic regions of proteins. Finally, the Hausdorff distance25 provides a measure of dissimilarity between sets of points, offering further insight into structural comparisons. In our study, the Yau-Hausdorff distance was calculated by comparing point sets representing the protein structures, specifically using distances from the structural centroids to the Cα atoms. This method captures overall geometric differences between structures by measuring the maximal deviation between point sets in a bidirectional manner. However, it may not fully account for local structural variations or conformational changes, such as those occurring at active sites or ligand-binding regions, which are crucial for functional similarity. Moreover, the method assumes that global geometric similarity correlates with functional similarity, which may not always hold true, especially for proteins whose function is dictated by specific local conformations.”

3) The reported computational efficiency of the proposed method is notable. Nevertheless, a discussion on the scalability of the method would be beneficial. Specifically, information on how computational time scales with the number of proteins or the complexity of the dataset would provide a clearer perspective on the practical applicability of the method in larger and more complex scenarios.

ANSWER: We appreciate the reviewer’s suggestion. To address this, we performed an additional analysis and added the results at page 17 on new Figure 14.

To evaluate the scalability of the CPE method, we performed an additional analysis by randomly selecting protein subsets from the Astral95 dataset, ranging in size from 1,000 to 30,000 proteins (at intervals of 5,000). For each subset, we computed the pairwise distances between proteins using both the CPE and TM-Vec methods, while also recording the processing time per amino acid. This metric represents the total computation time divided by the cumulative number of amino acids across all analyzed protein domains. As shown in **Fig. 14**, both methods demonstrate a linear increase in computation time per amino acid as the dataset size expands. However, the CPE method exhibits a gentler slope compared to TM-Vec, indicating superior scalability. These findings underscore the efficiency of the CPE method, particularly for handling large, complex datasets, making it suitable for high-throughput computational studies.

A graph with orange and purple dots

Description automatically generated

**Fig. 14 | Scalability of CPE and TM-Vec.** Processing time per amino acid for subsets from the Astral95 dataset, ranging in size from 1,000 to 30,000 proteins (at intervals of 5,000).

4) The results indicate that the energy profile-based model effectively reconstructs the phylogenetic tree. To strengthen these findings, the inclusion of statistical validation measures, such as confidence intervals or statistical significance tests, would be valuable. This addition would provide more robust supporkflow to the full SARS-CoV-2 or for the accuracy of the phylogenetic reconstructions.

ANSWER: We agree.. In response we have implemented a bootstrap analysis with B = 100 replicates to assess the robustness of the phylogenetic tree reconstructions using the CPE, SPE, TM-Vec, and MSA methods. Additionally, we calculated confidence intervals to statistically validate the accuracy of our results. The bootstrap analysis and confidence intervals for the main branches that distinguish the three species are now presented in the relevant figures within the manuscript, providing stronger evidence for the reliability of the energy profile-based model in reconstructing phylogenetic trees. We have also updated Section 3.5 to address this requirement. The complete bootstrap results for all branches can be found in new Supplementary Figures S2-S5. We hope these revisions address the reviewers suggestion and strengthen our findings.

5) The manuscript reports a discrepancy between the text and Table 1 regarding the accuracy and computation time of the 1-NN classifier based on the distance between profiles' energy (CPE). The text states an accuracy of 97% and a computation time of approximately 3 minutes, whereas Table 1 shows an accuracy of 100% and a computation time of 1 second. The authors should reconcile this discrepancy to ensure consistency and accuracy in the reporting of results.

ANSWER: We would like to thank the reviewer for the rigour. We confirm that the correct accuracy for the 1-NN classifier based on the distance between profiles' energy (CPE) is indeed **100%**, and the computation time is **1 second**, as correctly indicated in Table 1. The previously mentioned accuracy of 97% and computation time of approximately 3 minutes in the text was an error and we apologize for that. We have now updated the text (Page 8) to reflect the correct values as presented in Table 1

**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 SPE and 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 |
| **SPE** | 99% | 187 sec |

6) The authors demonstrate that the CPE method not only matches TM-Vec in accuracy but also exhibits superior computational efficiency and interpretability. The reliance of CPE on energy profiles and pairwise amino acid interactions provides a clearer understanding of protein similarities compared to the black-box nature of deep learning techniques used in TM-Vec. Further commentary on how the CPE approach achieves this enhanced interpretability would be valuable. Specifically, elucidating how the reliance on energy profiles and amino acid interactions facilitates a more straightforward understanding of protein relationships compared to TM-Vec’s deep learning model would be beneficial.

ANSWER: We would like to thank the reviewer for highlighting this and do agree with the suggestion.  We have now added the following discussion into the Discussion section of the manuscript.

"The CPE method leverages energy profiles derived from pairwise amino acid interactions, where each dimension of the energy vector corresponds to specific amino acid pairs. This structured, physically grounded representation makes the data more intuitive and interpretable because the calculated energy values are likely to directly reflect biologically meaningful aspects of protein functions?, such as stability, folding, and molecular dynamics. This clarity allows researchers to trace protein similarities back to the underlying energy landscapes, which are well-established in protein science.

In contrast, TM-Vec utilizes deep learning embeddings that, although effective, function as a "black box." While TM-Vec can identify remote homologies and structural similarities, the embeddings it generates are abstract and difficult to deconstruct in a biologically meaningful way. This limits the ability to draw direct connections between the model's outputs and the physical or functional properties of proteins.

The CPE method, by focusing on energy profiles, offers distinct advantages in terms of interpretability. Energy profiles are inherently linked to protein folding, stability, and interaction networks, which are fundamental to biological function. For example, an increase in energy in specific pairwise interactions might suggest destabilizing mutations or conformational shifts that affect protein function. This explicit connection between energy and structural features allows for more transparent insights into how variations in energy impact the overall behavior and evolutionary relationships of proteins. By directly correlating these energy states with functional classifications such as folds, superfamilies, or evolutionary relationships, CPE provides clearer, actionable insights for researchers.

Additionally, because energy profiles can be tied to specific biophysical principles—such as electrostatic interactions, hydrophobic effects, or van der Waals forces—CPE offers a mechanistic understanding of protein relationships that is often lacking in machine learning models like TM-Vec. In fields such as drug discovery or protein engineering, where understanding the precise molecular interactions is crucial, CPE provides a significant advantage in generating interpretable and actionable data.

In conclusion, CPE’s reliance on energy profiles provides not only a more interpretable but also a biophysically grounded model of protein similarity. This contrasts with TM-Vec’s deep learning approach, which, while powerful, offers less transparency and explainability. CPE’s approach is particularly valuable in contexts where understanding the biological and structural principles behind protein behavior is critical, such as in evolutionary studies, disease-related mutation analysis, and drug development."

7) The methods section in general would benefit from a more comprehensive description to facilitate a clearer understanding of the methodology and enable others to replicate the study more effectively.

ANSWER: We acknowledge the importance of providing a more detailed and comprehensive Methods section to ensure clarity and reproducibility of our study. We have now revised the Method section and have incorporated a schematic flowchart (Fig. 1) in the revised manuscript.

#### “Dataset Preparation

A curated dataset of non-redundant protein chains was generated using PISCES41 from the Protein Data Bank (PDB). The dataset was selected based on the following criteria:

* **Pairwise sequence identity:** Less than 50% to ensure non-redundancy.
* **Resolution:** Higher than 1.6 Å to guarantee structural accuracy.
* **R-factor:** Below 0.25 to ensure reliable crystallographic data.
* **Protein length:** Between 40 and 1,000 residues to include proteins of varying sizes while excluding excessively short or long chains.
* **Overlap:** Proteins overlapping with the test sets from this manuscript were removed from the training set.

These filtered proteins were utilized to train and calculate the knowledge-based potential function as follows.

**Pairwise Distance-Dependent Knowledge-Based Potential**

Knowledge-based potentials are derived from databases of known protein structures and are essential for estimating the energies of pairwise interactions. These potentials can be based on various factors, including distance dependencies, dihedral angles, and accessible surface areas. In this study, we employed a distance-dependent potential function where atomic contacts were identified using the tessellation method as follows:

**Contact Identification:**

1. **Representation:** All amino acids in each protein chain were represented by their heavy atoms (excluding hydrogen atoms).
2. **Delaunay Tessellation:** A Delaunay tessellation of the resulting point set was computed using Qhull42, identifying neighboring atoms based on spatial proximity.
3. **Defining Contacts:** Two atoms were considered to be in contact if they are connected by an edge in the Delaunay triangulation. This implies that they are not shielded from each other by other atoms, ensuring direct interaction without obstruction.
4. **Distance Shells:** The distances between contacting atoms were divided into 30 discrete shells, starting at 0.75 Å with each shell having a width of 0.5 Å. This binning allows for the extraction of distance-dependent interaction potentials.
5. **Interaction Range:** Only atoms separated by less than 6 Å were considered to interact. If a third atom exists between two close atoms, preventing direct contact, the interaction was excluded. Additionally, all pairwise interactions within the same residue were omitted to focus on inter-residue interactions.

**Atom Types:** A total of 167 atom types were considered by treating non-hydrogen atoms as distinct based on their specific amino acid residues.

**Energy Calculation:** The potential energy between two atoms    and   at distance was calculated using the following equation:

(1)

where *RT* is constant and equal to 0.582 kcal/mole.  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 7, 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)

where the sum is over all atom pairs in contact, identified via the Delaunay triangulation method.

**Structural Profile of Energy (SPE):** Given the 20 standard amino acids, there are 210 unique amino acid-amino acid interaction types. For each protein structure, a 210-dimensional vector was created to represent the distance-dependent energy interactions between residues. Each dimension corresponds to the energy interaction between a specific pair of amino acid types. This vector is referred to as the **Structural Profile of Energy (SPE)**.

**Pairwise Energy Content from Amino Acid Composition**

While the knowledge-based potential function relies on having the three-dimensional structure of a protein, many protein structures remain undetermined experimentally. To address this, we developed a method to estimate pairwise energy content based solely on amino acid composition.

**Energy Estimation:** For each protein in the training set:

 denotes the energy of interactions between all residues of type and all other amino

acids in protein .

 The estimated energy ​ is calculated using:

where, represents the frequency of amino acid type in the structure , and , is calculated as the ratio of to the total number of amino acids in , denoted by , and is the energy predictor matrix, delineating the dependence of amino acid 's energy on the th element within the amino acid composition.

**Parameter Optimization:** The parameters of each row of matrix P were optimized by minimizing the following objective function for each amino acid type

By setting the partial derivatives for all , a system of linear equations was obtained and solved using the Symbolic Math Toolbox in MATLAB.

**Compositional Profile of Energy (CPE):** For each pair of amino acid types and , the energy was estimated based on amino acid sequence composition:

where is the energy predictor matrix estimated using equation 4. This results in a 210-dimensional vector representing energy interactions between amino acid types based on composition alone. This vector is termed the **Compositional Profile of Energy (CPE)** and is normalized according to protein length.

”

A diagram of a cell block

Description automatically generated

**Fig. 1** | A) Construction of the knowledge-based potential function. B) Estimation of the predictor matrix P. C) Construction of the structural profile of energy (SPE) based on protein structure. D) Construction of the compositional profile of energy (CPE) based on protein sequence.

8) The GITHUB repository contains all the described datasets and code; however, it would benefit from additional documentation. Specifically, further explanation on how to execute and run the scripts would be valuable. Additionally, guidance on how to interpret the outputs of these scripts is necessary for users to fully understand and utilize the provided tools. The repository and the manuscript should include a comprehensive table listing all the datasets used in the manuscript. This table should detail the exact names of the datasets and provide direct pointers to the files containing them. This improvement would greatly enhance the usability and transparency of the repository.

ANSWER: We have now….