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

Peyman Choopanian1, 2, Jaan-Olle Andressoo1, 2, 3\*, and Mehdi Mirzaie1, 2\*

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

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

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

Table S1. The results of clustering using the Adjusted Rand Index (ARI) and class error

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Based** | **Method** | **Time** | **ARI** | | | | **Class Error** | | | |
| **3** | **4** | **5** | **6** | **3** | **4** | **5** | **6** |
| Sequence | **CPE** | **0.9 sec** | 0.50 | **0.95** | 0.94 | 0.92 | 0.22 | **0.08** | 0.1 | 0.11 |
| **TM Vec** | 89 sec | 0.16 | 0.48 | 0.87 | **0.86** | 0.34 | 0.22 | 0.12 | **0.14** |
| **MSA (ClustalW)** | 72 sec | 0.49 | 0.49 | 0.32 | 0.32 | 0.22 | 0.22 | 0.29 | 0.29 |
| Structure | **SPE** | **3 min** | **1** | **0.95** | 0.93 | 0.66 | **0** | **0.08** | 0.11 | 0.26 |
| **RMSD** | 70 min | 0.50 | 0.50 | 0.36 | **0.73** | 0.22 | 0.22 | 0.22 | **0.17** |
| **TM score** | 9.7 h | 0.50 | **0.56** | 0.40 | 0.40 | 0.22 | **0.17** | 0.24 | 0.24 |

Table S2. The PDB IDs of spike proteins.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **pdbID** | **length** | **virus** | **pdbID** | **length** | **virus** | **pdbID** | **length** | **virus** | **pdbID** | **length** | **virus** |
| **6XM5A** | **1057** | **SARS-CoV-2** | **6CRZA** | **1068** | **SARS-CoV** | **6Q04B** | **1159** | **MERS-CoV** | **6ZGFC** | **1060** | **SARS-CoV-2** |
| **6ZP2A** | **1097** | **SARS-CoV-2** | **6CRWA** | **1068** | **SARS-CoV** | **6ACKB** | **1065** | **SARS-CoV** | **6ZGEC** | **1098** | **SARS-CoV-2** |
| **6ZP1A** | **1017** | **SARS-CoV-2** | **5W9JA** | **463** | **MERS-CoV** | **6ACJB** | **1065** | **SARS-CoV** | **6Z97C** | **991** | **SARS-CoV-2** |
| **6ZP0A** | **1030** | **SARS-CoV-2** | **5XLRA** | **1022** | **SARS-CoV** | **6ACGB** | **1065** | **SARS-CoV** | **7BYRC** | **973** | **SARS-CoV-2** |
| **6ZOZA** | **1069** | **SARS-CoV-2** | **5X5FA** | **1141** | **MERS-CoV** | **6ACDB** | **1065** | **SARS-CoV** | **6X6PC** | **1017** | **SARS-CoV-2** |
| **6ZOYA** | **1021** | **SARS-CoV-2** | **5X58A** | **1054** | **SARS-CoV** | **6ACCB** | **1065** | **SARS-CoV** | **6Z43C** | **991** | **SARS-CoV-2** |
| **6ZOXA** | **1017** | **SARS-CoV-2** | **5WRGA** | **736** | **SARS-CoV** | **5W9OD** | **463** | **MERS-CoV** | **6X2CC** | **971** | **SARS-CoV-2** |
| **6XEYA** | **1034** | **SARS-CoV-2** | **6ZP7B** | **930** | **SARS-CoV-2** | **5W9ND** | **463** | **MERS-CoV** | **6X2AC** | **961** | **SARS-CoV-2** |
| **6ZGIA** | **1098** | **SARS-CoV-2** | **6ZP5B** | **943** | **SARS-CoV-2** | **5W9MD** | **457** | **MERS-CoV** | **6X29C** | **972** | **SARS-CoV-2** |
| **6ZGGA** | **1069** | **SARS-CoV-2** | **6ZOWB** | **930** | **SARS-CoV-2** | **5W9LB** | **726** | **MERS-CoV** | **6WPTC** | **952** | **SARS-CoV-2** |
| **6ZGFA** | **1060** | **SARS-CoV-2** | **6ZHDB** | **992** | **SARS-CoV-2** | **5W9KD** | **462** | **MERS-CoV** | **6WPSE** | **955** | **SARS-CoV-2** |
| **6ZGEA** | **1098** | **SARS-CoV-2** | **6XM5B** | **1056** | **SARS-CoV-2** | **5W9JD** | **463** | **MERS-CoV** | **6VYBC** | **960** | **SARS-CoV-2** |
| **6Z97A** | **995** | **SARS-CoV-2** | **6ZP2B** | **1097** | **SARS-CoV-2** | **5W9HD** | **463** | **MERS-CoV** | **6VXXC** | **972** | **SARS-CoV-2** |
| **6XCMA** | **966** | **SARS-CoV-2** | **6ZP1B** | **1017** | **SARS-CoV-2** | **5XLRB** | **1022** | **SARS-CoV** | **6VSBC** | **973** | **SARS-CoV-2** |
| **6X6PA** | **1017** | **SARS-CoV-2** | **6ZP0B** | **1030** | **SARS-CoV-2** | **5X5FB** | **1141** | **MERS-CoV** | **6Q07C** | **1159** | **MERS-CoV** |
| **6X2CA** | **971** | **SARS-CoV-2** | **6ZOZB** | **1069** | **SARS-CoV-2** | **5X5CB** | **1141** | **MERS-CoV** | **6Q06C** | **1159** | **MERS-CoV** |
| **6X2BA** | **963** | **SARS-CoV-2** | **6ZOYB** | **1021** | **SARS-CoV-2** | **5X5BB** | **1053** | **SARS-CoV** | **6Q05C** | **1159** | **MERS-CoV** |
| **6X2AA** | **966** | **SARS-CoV-2** | **6ZOXB** | **1017** | **SARS-CoV-2** | **5X58B** | **1053** | **SARS-CoV** | **6Q04C** | **1159** | **MERS-CoV** |
| **6X29A** | **972** | **SARS-CoV-2** | **6XEYB** | **1034** | **SARS-CoV-2** | **5WRGB** | **736** | **SARS-CoV** | **6NB3C** | **1169** | **MERS-CoV** |
| **6WPTA** | **945** | **SARS-CoV-2** | **6XKLB** | **976** | **SARS-CoV-2** | **6ZP7C** | **943** | **SARS-CoV-2** | **6ACCC** | **1065** | **SARS-CoV** |
| **6WPSA** | **955** | **SARS-CoV-2** | **7C2LB** | **1055** | **SARS-CoV-2** | **6ZP5C** | **930** | **SARS-CoV-2** | **6CS1C** | **1068** | **SARS-CoV** |
| **6VYBA** | **966** | **SARS-CoV-2** | **6ZGIB** | **1098** | **SARS-CoV-2** | **6ZOWC** | **943** | **SARS-CoV-2** | **6CS0C** | **1069** | **SARS-CoV** |
| **6VXXA** | **972** | **SARS-CoV-2** | **6ZGHB** | **1077** | **SARS-CoV-2** | **6ZHDC** | **990** | **SARS-CoV-2** | **6CRZC** | **1069** | **SARS-CoV** |
| **6Q07A** | **1159** | **MERS-CoV** | **6ZGFB** | **1060** | **SARS-CoV-2** | **6XM5C** | **1036** | **SARS-CoV-2** | **6CRXC** | **1069** | **SARS-CoV** |
| **6Q06A** | **1159** | **MERS-CoV** | **6ZGEB** | **1098** | **SARS-CoV-2** | **6ZP2C** | **1097** | **SARS-CoV-2** | **6CRWC** | **1068** | **SARS-CoV** |
| **6Q05A** | **1159** | **MERS-CoV** | **7BYRB** | **998** | **SARS-CoV-2** | **6ZP1C** | **1017** | **SARS-CoV-2** | **5W9OG** | **463** | **MERS-CoV** |
| **6Q04A** | **1159** | **MERS-CoV** | **6X6PB** | **1017** | **SARS-CoV-2** | **6ZP0C** | **1030** | **SARS-CoV-2** | **5W9NG** | **457** | **MERS-CoV** |
| **6NB6A** | **1052** | **SARS-CoV** | **6Z43B** | **992** | **SARS-CoV-2** | **6ZOZC** | **1070** | **SARS-CoV-2** | **5W9ME** | **726** | **MERS-CoV** |
| **6NB4A** | **1169** | **MERS-CoV** | **6X2CB** | **971** | **SARS-CoV-2** | **6ZOYC** | **1021** | **SARS-CoV-2** | **5W9LC** | **726** | **MERS-CoV** |
| **6NB3A** | **1169** | **MERS-CoV** | **6X29B** | **972** | **SARS-CoV-2** | **6ZOXC** | **1017** | **SARS-CoV-2** | **5W9JG** | **463** | **MERS-CoV** |
| **6ACKA** | **1065** | **SARS-CoV** | **6WPSB** | **955** | **SARS-CoV-2** | **6XEYC** | **1030** | **SARS-CoV-2** | **5W9HG** | **463** | **MERS-CoV** |
| **6ACJA** | **1065** | **SARS-CoV** | **6VXXB** | **972** | **SARS-CoV-2** | **6XKLC** | **976** | **SARS-CoV-2** | **5XLRC** | **1022** | **SARS-CoV** |
| **6ACGA** | **1065** | **SARS-CoV** | **6VSBB** | **973** | **SARS-CoV-2** | **7C2LC** | **1049** | **SARS-CoV-2** | **5X5BC** | **1053** | **SARS-CoV** |
| **6ACDA** | **1065** | **SARS-CoV** | **6Q07B** | **1159** | **MERS-CoV** | **6ZGIC** | **1098** | **SARS-CoV-2** | **5X58C** | **1052** | **SARS-CoV** |
| **6ACCA** | **1065** | **SARS-CoV** | **6Q06B** | **1159** | **MERS-CoV** | **6ZGHC** | **1080** | **SARS-CoV-2** | **5WRGC** | **736** | **SARS-CoV** |
| **6CS0A** | **1068** | **SARS-CoV** | **6Q05B** | **1159** | **MERS-CoV** | **6ZGGC** | **1067** | **SARS-CoV-2** |  |  |  |

Table S3. The results of 1-NN classification on SARS Proteome using CPE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SARS\_Proteome | ac | pr | re | f1 |
| E\_protein | 99.68 | 100 | 96.55 | 98 |
| N\_C-terminaldomain | 99.68 | 100 | 100 | 100 |
| N\_N-terminaldomain | 99.68 | 99.82 | 100 | 100 |
| NSP1\_protein | 99.68 | 100 | 95.83 | 98 |
| NSP10\_protein | 99.68 | 100 | 100 | 100 |
| NSP12\_protein | 99.68 | 100 | 100 | 100 |
| NSP13\_protein | 99.68 | 100 | 100 | 100 |
| NSP14\_protein | 99.68 | 100 | 100 | 100 |
| NSP15\_protein | 99.68 | 100 | 100 | 100 |
| NSP16\_protein | 99.68 | 99.61 | 100 | 100 |
| NSP2\_protein | 99.68 | 100 | 100 | 100 |
| NSP3\_cd21525\_SUD\_C\_SARS-CoV\_Nsp3 | 99.68 | 99.2 | 100 | 100 |
| NSP3\_cd21557\_Macro\_X\_Nsp3-like | 99.68 | 99.19 | 99.19 | 99 |
| NSP3\_cd21717\_TM\_Y\_SARS-CoV-like\_Nsp3\_C | 99.68 | 100 | 99.19 | 100 |
| NSP3\_cd21732\_betaCoV\_PLPro | 99.68 | 99.18 | 100 | 100 |
| NSP3\_cd21822\_SARS-CoV-like\_Nsp3\_NAB | 99.68 | 98.33 | 99.16 | 99 |
| NSP3\_cl00019\_Macro\_SF | 99.68 | 100 | 97.62 | 99 |
| NSP3\_cl13138\_SUD-M | 99.68 | 99.17 | 95.97 | 98 |
| NSP3\_cl13772\_DUF3655 | 99.68 | 100 | 100 | 100 |
| NSP5\_protein | 99.68 | 100 | 99.59 | 100 |
| NSP7\_protein | 99.68 | 100 | 100 | 100 |
| NSP8\_protein | 99.68 | 99.5 | 100 | 100 |
| NSP9\_protein | 99.68 | 98.21 | 99.4 | 99 |
| orf3a\_protein | 99.68 | 100 | 100 | 100 |
| orf7a\_protein | 99.68 | 100 | 100 | 100 |
| orf8\_protein | 99.68 | 100 | 100 | 100 |
| orf9b\_protein | 99.68 | 91.3 | 100 | 95 |
| S\_protein | 99.68 | 100 | 100 | 100 |
| 29 | 99.68 | 99.4110714 | 99.375 | 99.4642857 |

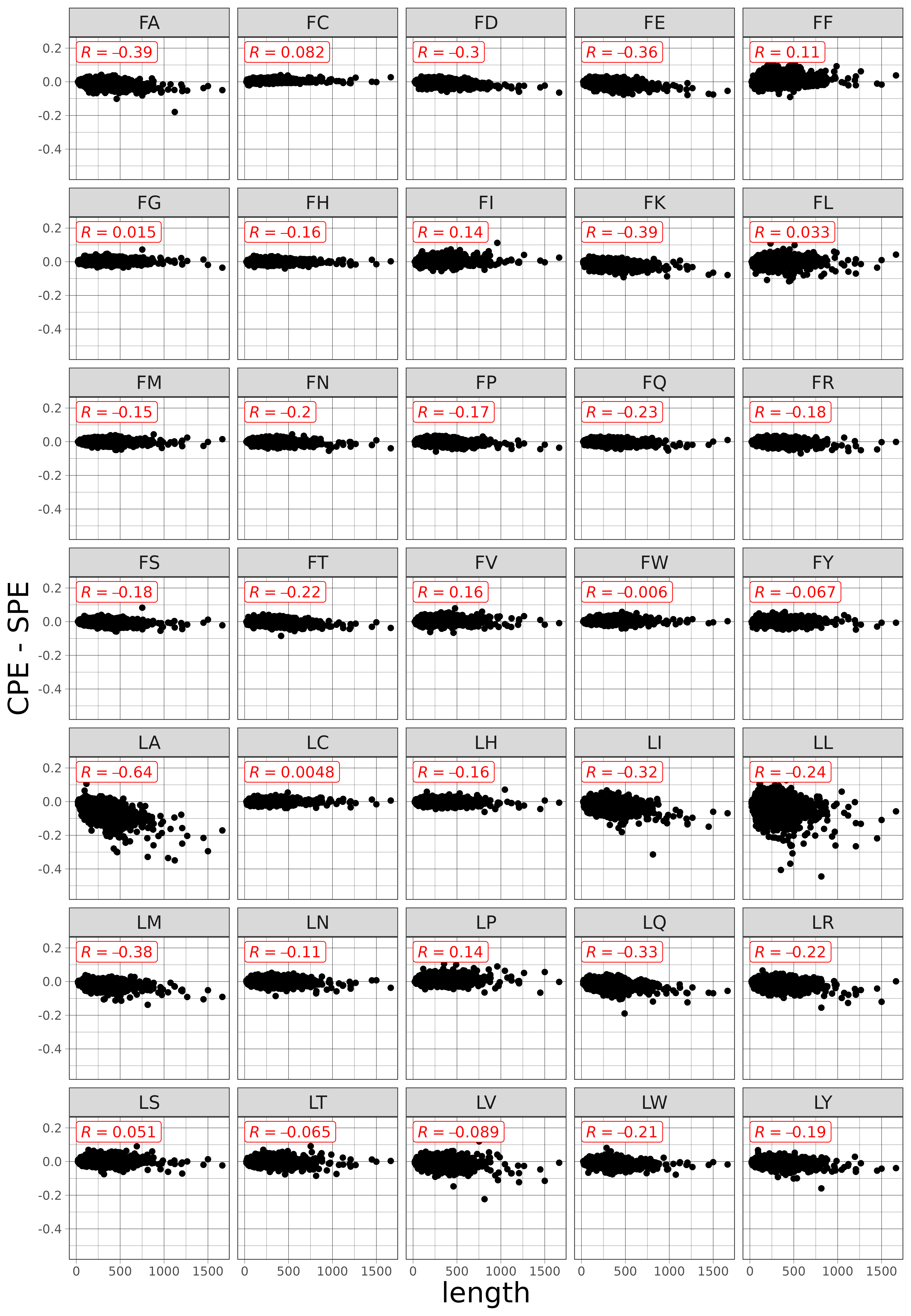
Table S4. The results of 1-NN classification on SARS Proteome using SPE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proteom | ac | pr | re | f1 |
| E\_protein | 99.73 | 100 | 96.55 | 98 |
| N\_C-terminaldomain | 99.73 | 100 | 99.82 | 100 |
| N\_N-terminaldomain | 99.73 | 100 | 100 | 100 |
| NSP1\_protein | 99.73 | 100 | 95.83 | 98 |
| NSP10\_protein | 99.73 | 99.55 | 100 | 100 |
| NSP12\_protein | 99.73 | 100 | 100 | 100 |
| NSP13\_protein | 99.73 | 100 | 100 | 100 |
| NSP14\_protein | 99.73 | 100 | 100 | 100 |
| NSP15\_protein | 99.73 | 100 | 100 | 100 |
| NSP16\_protein | 99.73 | 100 | 100 | 100 |
| NSP2\_protein | 99.73 | 100 | 100 | 100 |
| NSP3\_cd21525\_SUD\_C\_SARS-CoV\_Nsp3 | 99.73 | 96.88 | 100 | 98 |
| NSP3\_cd21557\_Macro\_X\_Nsp3-like | 99.73 | 100 | 100 | 100 |
| NSP3\_cd21717\_TM\_Y\_SARS-CoV-like\_Nsp3\_C | 99.73 | 100 | 100 | 100 |
| NSP3\_cd21732\_betaCoV\_PLPro | 99.73 | 100 | 100 | 100 |
| NSP3\_cd21822\_SARS-CoV-like\_Nsp3\_NAB | 99.73 | 98.35 | 100 | 99 |
| NSP3\_cl00019\_Macro\_SF | 99.73 | 97.22 | 83.33 | 90 |
| NSP3\_cl13138\_SUD-M | 99.73 | 100 | 99.19 | 100 |
| NSP3\_cl13772\_DUF3655 | 99.73 | 100 | 100 | 100 |
| NSP5\_protein | 99.73 | 100 | 100 | 100 |
| NSP7\_protein | 99.73 | 100 | 100 | 100 |
| NSP8\_protein | 99.73 | 100 | 100 | 100 |
| NSP9\_protein | 99.73 | 98.81 | 100 | 99 |
| orf3a\_protein | 99.73 | 100 | 100 | 100 |
| orf7a\_protein | 99.73 | 96.3 | 100 | 98 |
| orf8\_protein | 99.73 | 95 | 100 | 97 |
| orf9b\_protein | 99.73 | 100 | 100 | 100 |
| S\_protein | 99.73 | 100 | 100 | 100 |

Table S5. The results of 1-NN classification on SARS Proteome using TM-Vec.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ac | pr | re | f1 |
| Average | 99.66 | 99.4157143 | 99.1167857 | 99.25 |
| E\_protein | 99.66 | 98.31 | 100 | 99 |
| N\_C-terminaldomain | 99.66 | 100 | 99.82 | 100 |
| N\_N-terminaldomain | 99.66 | 100 | 100 | 100 |
| NSP1\_protein | 99.66 | 100 | 95.83 | 98 |
| NSP10\_protein | 99.66 | 100 | 99.55 | 100 |
| NSP12\_protein | 99.66 | 100 | 100 | 100 |
| NSP13\_protein | 99.66 | 100 | 100 | 100 |
| NSP14\_protein | 99.66 | 100 | 100 | 100 |
| NSP15\_protein | 99.66 | 100 | 100 | 100 |
| NSP16\_protein | 99.66 | 100 | 100 | 100 |
| NSP2\_protein | 99.66 | 100 | 100 | 100 |
| NSP3\_cd21525\_SUD\_C\_SARS-CoV\_Nsp3 | 99.66 | 94.62 | 99.19 | 97 |
| NSP3\_cd21557\_Macro\_X\_Nsp3-like | 99.66 | 100 | 100 | 100 |
| NSP3\_cd21717\_TM\_Y\_SARS-CoV-like\_Nsp3\_C | 99.66 | 100 | 99.19 | 100 |
| NSP3\_cd21732\_betaCoV\_PLPro | 99.66 | 100 | 100 | 100 |
| NSP3\_cd21822\_SARS-CoV-like\_Nsp3\_NAB | 99.66 | 98.32 | 98.32 | 98 |
| NSP3\_cl00019\_Macro\_SF | 99.66 | 100 | 97.62 | 99 |
| NSP3\_cl13138\_SUD-M | 99.66 | 99.17 | 96.77 | 98 |
| NSP3\_cl13772\_DUF3655 | 99.66 | 100 | 100 | 100 |
| NSP5\_protein | 99.66 | 100 | 100 | 100 |
| NSP7\_protein | 99.66 | 100 | 100 | 100 |
| NSP8\_protein | 99.66 | 100 | 100 | 100 |
| NSP9\_protein | 99.66 | 98.22 | 100 | 99 |
| orf3a\_protein | 99.66 | 100 | 96.67 | 98 |
| orf7a\_protein | 99.66 | 100 | 92.31 | 96 |
| orf8\_protein | 99.66 | 95 | 100 | 97 |
| orf9b\_protein | 99.66 | 100 | 100 | 100 |
| S\_protein | 99.66 | 100 | 100 | 100 |

Fig. S1. Corr210





**Fig. S2|** The Clustering analysis of spike glycoprotein structures from SARS-CoV, SARS-CoV-2, and MERS-CoV based on CPE method.



**Fig. S3|** The Clustering analysis of spike glycoprotein structures from SARS-CoV, SARS-CoV-2, and MERS-CoV based on SPE method.



**Fig. S4|** The Clustering analysis of spike glycoprotein structures from SARS-CoV, SARS-CoV-2, and MERS-CoV based on TM-Vec method.



**Fig. S5|** The Clustering analysis of spike glycoprotein structures from SARS-CoV, SARS-CoV-2, and MERS-CoV based on MSA method.



**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.



**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.



**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.