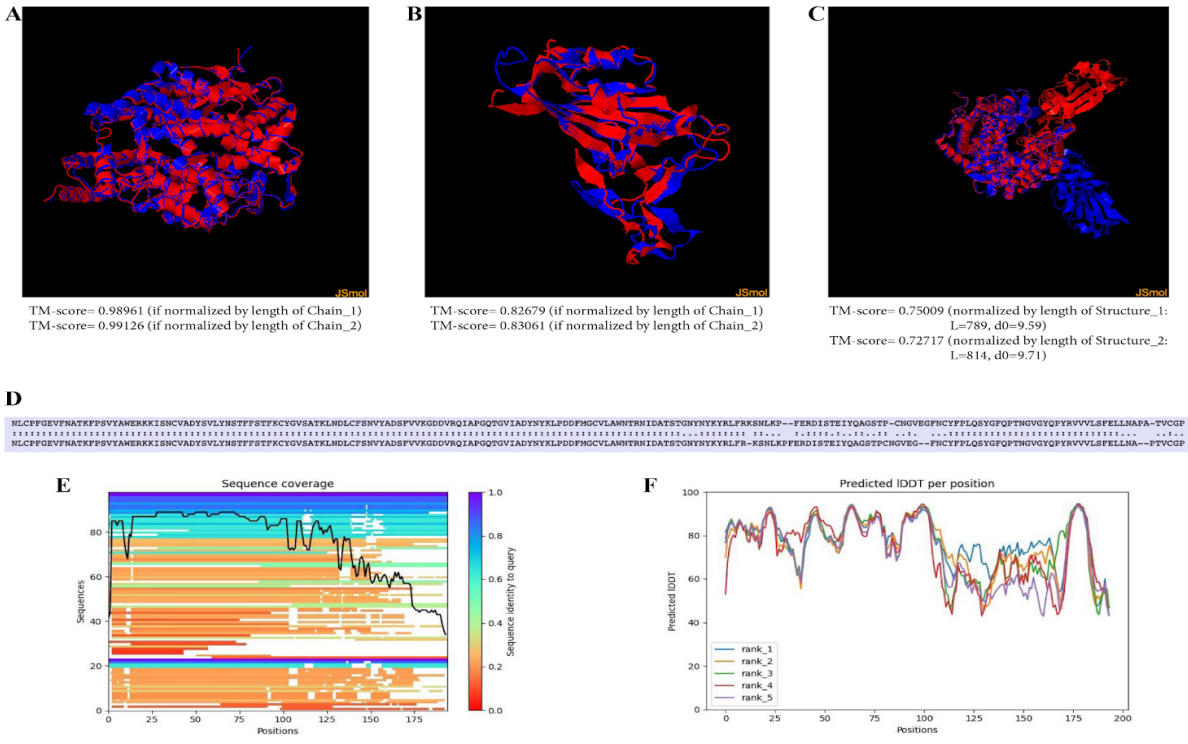
Investigating AlphaFold2 structural prediction accuracy for SARS-CoV-2 Receptor Binding Domain-Angiotensin Converting Enzyme 2 complex

Chi-Hung Shu, Joshua Huang, Lucy Tian, William Hou

Introduction - Artificial intelligence system AlphaFold2 (AF2) grasped global attention recently because of its major breakthrough in solving the protein folding challenge, and our team aimed to utilize this groundbreaking tool to predict novel proteins whose structures are not observed yet. Briefly, AF2 is a neural network embedded with multiple sequence alignments (MSAs) and pairwise features, and the network was trained by supervised learning of a large collection of protein sequences on the Protein Data Bank (PDB) to diversify the MSA database. The full network then used the primary amino acid sequence and aligned sequences as inputs to predict the 3D coordinates of all heavy atoms for a protein [1], [2]. Our team applied AF2 to explore how the newly evolved SARS-CoV-2 receptor binding domain (RBD) folds when the spike protein binds to the angiotensin converting enzyme 2 (ACE2) receptor on human cells.

Method - Primary amino acid sequences were sourced and compiled from Uniport and RCSB, and the corresponding PDB files were downloaded from the 6VW1 entry of the Protein Data Bank, cleaned using PyMol, and saved for comparison. CollabFold notebook was utilized to run basic functions of AF2 and the default settings were adjusted based on experimental needs, which will be detailed in later parts of the report. The resulting files were saved on a Google Drive folder, and the predicted PBD file was uploaded to TM-align for monomer structure alignment and MM-align for multimer structural alignment to compare with the PDB file downloaded from the public database. Both TM-align and MM-align produce an alignment score between 0 and 1, with larger values corresponding to smaller differences between structures.

Result



SARS-CoV-2 RBD and human ACE2 complex structure as well as single RBD were predicted by AF2 and compared with corresponding original structures (PDB entry: 6vw1). TM_align of the predicted complex structure resulted in a single ACE2 alignment with a score of 0.991

normalized to reference length (Fig 1.A). MM-align of the predicted complex structure gave rise to a score of 0.727, normalized by the length of the reference protein (Fig 1.C). TM_alignment of predicted RBD structure using primary sequence resulted in a score of 0.707. The investigation suggested a lack of structural coverage for parts of the primary RBD sequence, which were then deleted for a second run of AF2 where the remaining 194 nucleotides were inputted to improve prediction accuracy. The calculated TM-score increased to 0.831, indicating a more similar folding (Fig 1.B). Associated aligned residues of predicted RBD are shown where the absence of dots between two sequences indicated residues that were not aligned, while the presence of only one dot indicated aligned residues with a distance > 5.0A (Fig1.D). Sequencing coverage (Fig1.E) and predicted local distance difference test (lDDT) (Fig1.F) are also shown.

Discussion - AF2 achieves extremely high prediction accuracy for the ACE2 protein. However, the prediction for S-protein RBD is not as accurate, while the binding configuration between ACE2 and RBD is more inaccurate. Although the structure of the ACE2-RBD complex has been solved in 2020, which implies it is not in the training data of AF2, the structure of the ACE2 protein itself has been solved earlier. As a result, it might already be included in the training data. The SARS-CoV-2 S-protein RBD, however, is part of a new protein that was just discovered in 2020. According to Figure 1.E, its alignment coverage is also suboptimal. In addition, there were amino acids at the start and the end of the RBD sequence but not the structure. After removing those amino acids, the prediction accuracy of AF2 improved substantially from 0.707 to 0.831, though still well shy of the 0.991 prediction accuracy of the ACE2 protein. We can see that the predicted error by AF2 (Fig 1.F) concentrates at positions 100-175, which matches the misaligned residues from MMFold (Fig 1.D). Despite not being able to yield the correct structure, AF2 still produces good confidence levels. In complex prediction mode, AF2 mispredicted the binding configuration between the SARS-Cov-2 RBD and the ACE2 protein, which is a substantial error. This large error is possibly due to the prediction errors in the RBD structure, which indicates that the accuracy of complex prediction is highly sensitive to the accuracy of its monomer components.

Conclusion - As shown in our results, AF2 prediction accuracy is highly dependent on the sequence integrity of the measured structure and coverage of MSAs for the protein sequence. Even for sequences with suboptimal quality for both features, AF2 outputs predicted structure with confidence level reflective of the actual prediction error. In addition, the complex prediction mode of AF2 is highly sensitive to the prediction accuracy of its components. Future improvements to AlphaFold may gear towards reducing the reliance on MSAs, while improving the prediction of complexes with relatively fuzzy components.

References

1. J. Jumper, et al., “Highly accurate protein structure prediction with alphafold,” Nature, 2021.
2. M. Mirdita, K. Schütze, et al., “Colabfold - making protein folding accessible to all,” 2021.

Team Contribution Breakdown

Chi-Hung Shu ran TM_align comparisons, and gathered datas. **Joshua Huang** cleaned up the structure using PyMOL, and ran TM-align and MM-align comparisons. **Lucy Tian** ran CollabFold and gathered figures. **William Hou** gathered corresponding files, ran TM_align comparisons, and assembled the figures. All the members searched for novel target proteins and co-wrote the report.