PROTEIN COMPLEX PREDICTION WITH ALPHA FOLD-MULTIMER

A report submitted in the partial fulfillment of the requirements for

Course Code GE- 511 of M.S. (Pharm.)

Submitted by

Triveni Navuluri

Regd.No.21PIM3493

M.S. (Pharm.) semester-I

Department of Pharmacoinformatics

National Institute of Pharmaceutical Education and Research,

Sector-67, S.A.S. Nagar, Punjab-160062.

December 2021

Table of content

1.Introduction :

Introduction to alphaFold:

In his recognition speech for the 1972 Nobel prize in chemistry, christian anfinsen postulated that, a protein's amino acid series need to fully decide its shape. This speculation sparked a 5 decade quest on the way to computationally predict a protein's 3D structure based totally totally on its 1d amino acid series.

The formation of permanent and temporary protein complexes underpins most of the biological processes and understanding of these protein complexes is a key step towards understanding and modifying their functions.

The vast majority of well-established single protein chains are expected to excessive accuracy because of the current alphafold [1] version, the prediction of multi-chain protein complexes remains a challenge in many cases.

Though AlphaFold is trained on single protein chains, including many proteins whose structures was solved in complex with other proteins, it showed remarkable ability to predict protein structures with co-bound factors or proteins stabilised by its multimeric interactions. subsequent work has shown that by providing pseudo-multimer inputs (for e.g. residue gap insertion or chains joined with a flexible linker) to the single chain AlphaFold model is successful at predicting multimer interactions. These papers shown surprising generalization performance of the original trained AlphaFold model but they leave the open the question of how much accurate AlphaFold is when it is training adapted for multimeric inputs.

The primary metric used by CASP to measure the accuracy of the predictions is the Global distance test which will range from 0-100. It is a measure of similarity between two protein structures with known amino acid sequences.

According to Professor_Moult, a score of around 90 GDT is informally considered to be competitive with results obtained from experimental method.

Finding out protein's shapes take at least weeks or months in lab. Alphafold can predict the shapes to the nearest atom in a day or two. In this work, they display that an Alphafold model trained especially for multimeric inputs of acknowledged stoichiometry, which we name Alphafold-multimer, notably will increase the accuracy of expected multimeric interfaces over input-adapted single-chain Alphafold while preserving high intra-chain accuracy. On a benchmark dataset of 17 heterodimer proteins with out templates, they have achieved at the least medium accuracy (dockQ (0.49) on 14 targets and high accuracy (dockQ 0.8) on 6 goals, in comparison to 9 targets of as a minimum medium accuracy and 4 of high accuracy for the previous state of the art system (an Alphafold-primarily based system). In addition, they also predicted structures for a huge dataset of 4,433 current protein complexes, from which they score all non-redundant interfaces with low template identity. For heteromeric interfaces they effectively predicted the interface (dockQ 0.23) in 67% of cases and convey high accuracy predictions (dockQ 0.8) in 23% of cases, an improvement of +25 and +11 percent factors over the flexible linker modification of Alphafold respectively. For homomeric interfaces, they efficaciously expect the interface in 69% of instances, and produce high accuracy predictions in 34% of cases, an development of +five percent points in each times.

Working of AlphaFold:

The alphafold system combines records from the amino acid sequence, Multiple Sequence Alignments and homologous structures with a view to predict the shape of individual protein chains. The core part of the neural network, known as evoformer, includes a neural representation of the Multiple Sequence Alignment (MSA) and pairwise relations between the different amino acids in the protein. These two representations are mixed and processed via a group of neural network modules.

The pair illustration may be concept of as containing information about the relative positions of amino acids within the chain. This is used to predict the relative distances between the amino acids inside the chain through a binned distance distribution. The primary row of the multiple sequence alignment embedding is then used together with the pair embedding to predict the very last structure. The model is trained quit-to-end with gradients propagating from the expected shape through the entire network.

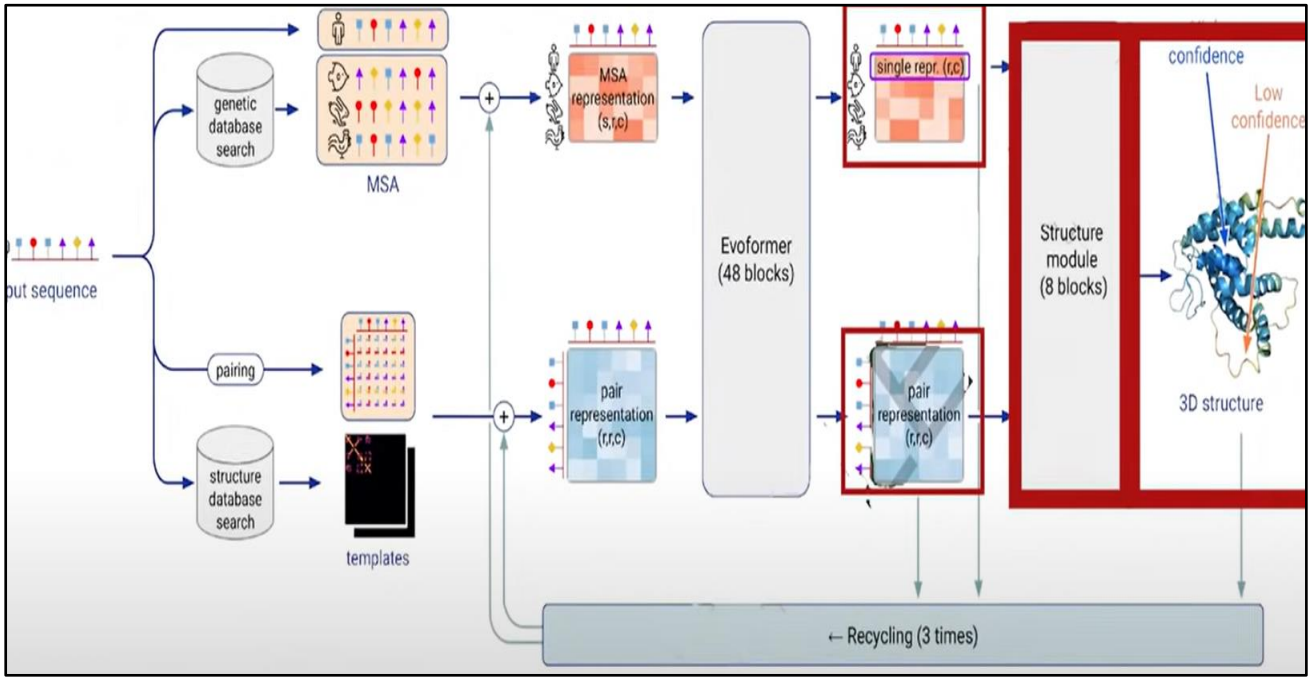

Fig.1: The process inside the Alphafold, Reproduced with permission.[8]

2.Methods:

some changes are made in the alphafold system to train it on multiple complexes which also includes making some adjustments in the structure losses and the model architecture.

2.1. Multi-chain permutation alignment:

While a protein of a given sequence appears multiple times in the complex, the mapping between the predicted and the ground truth coordinates is arbitrary and so the model cannot be assumed to predict chains within the identical order as in the ground reality. To account for this they aimed to pick out the optimal permutation of predicted homomer chains that pleasant fits the ground truth. The complexity of optimizing over all variations grows combinatorially so we hire a easy heuristic that greedily attempts to discover a right permutation.

Stoichiometry need to be cautiously accounted for when scoring multimer structure predictions. In a prediction for a2b, both orderings of the chains are similarly valid, regardless of their ordering within the ground truth. Specifically, if this is not accounted for in the loss then accurate predictions can be unfairly penalized and the network will fail to train properly.

To address this difficulty, earlier than scoring a multimer prediction they first permuted chains with identical sequences such that they're best effort aligned with those of the prediction. One could imagine considering the pleasant alignment over all possible permutations, however this fastly becomes intractable, so a heuristic method is needed. Our approach is a simple greedy heuristic that may be performed effectively on TPU.

2.2. Cross chain Genetics:

Multiple Sequence Alignment (MSA) is a useful characteristic in predicting contacts and 3-D which is in generally ambiguous [2]. In this work they supplied explicit aligned sequences to the network following the method of zhou et al [3]. They paired sequences with the use of the UniProt species annotation and resolve ambiguities by using the following process. If the target protein is prokaryotic, the within-species pairing is completed using the smallest genetic distance (approximated by using the difference between their UniProt accession ids). If the target protein is eukaryotic, we rank the candidate rows of each chain through similarity to their respective goal sequence.

2.3. Multi chain cropping:

The number of residues that Alphafold-multimer may be skilled on is constrained by memory and also compute Considerations as both memory use and also computer use will increase rapidly with the total number of the amino Acids in the complex. To ameliorate this, the Alphafold system is trained on cropped segments of proteins, where in those cropped areas are contiguous blocks of residues as much as 384 residues. They modified this method whilst training on complexes because the cropped regions need to involve more than one chains in a given complex, and binding interfaces among chains are important for modelling protein complexes. Consequently, they designed a cropping process that maximizes chain coverage and Crop variety even as making sure good stability between interface and non-interface areas.

2.4. Architecture and Losses:

Alphafold makes use of a Frame Aligned Point Error (FAPE) loss, whereby the distances between ground truth and predicted atoms are computed inside the local reference frame of every residue. In Alphafold, this Loss has clamped at 10 Å . For training the multimer model, they made modifications to the loss function used. For the intra-chain aminoacid pairs of the complex, they keep the identical 10 Å clamping. For the inter-chain pairs, they have used an unclamped FAPE loss. This affords a better gradient signal for wrong interfaces. Moreover, they add more positional encodings denoting whether a given pair of amino acids corresponds to distinct chains and whether they belong to distinctive homomer or heteromer chains. They also made numerous small modifications to the model and implementation in order to facilitate inference of large proteins for a given quantity of memory.

2.5. Training Regimen:

Alphafold-Multimer has trained in a completely similar way to Alphafold. The training dataset comprised of structures from the Protein Data Bank (PDB) [4]. Chains had been sampled in inverse proportion to cluster length and their corresponding mmCIFs selected as input to the data pipeline, meaning that the rest of the chains within the bio-assembly will be included.

This means that the probability of sampling a specific PDB entry is proportional to the sum of the probabilities of the individual chain clusters for all chains within the report. The chain clusters are forty percent identification clusterings of the Protein Data Bank with MMSeqs2 [5].

They randomly crop to 384 residues according to the process. They have trained the model to convergence (about 10M samples, for 2 weeks) across 128 TPUv3 cores with a batch size of 1 per TPU core. Then they halved the studying rate and double the number of sequences fed into the MSA stack earlier before running two separate fine-tuning stages (one further day of training each), the first with the experimentally resolved and the predicted LDDT(pLDDT) heads switched on, and the second one with violation losses enabled. Those extra heads and losses are equal to those used within the Alphafold paper. They have trained three models for the first stage; the great model at the validation has then fine tuned with 5 different random seeds for the 2 fine tuning stages, producing five models in total.

2.6. Inference Regimen:

At inference time, they run all 5 trained models and select best model on each target according to the model confidence. The only difference to Alphafold-multimer from Alphafold is that alphafold multimer model confidence metric is slightly modified to accuracy on interfaces.

2.7. Model confidence:

The aphafold model accuracy is predicted by using TM-score (an algorithm to calculate the similarity of topologies of two proteins/models).They provided a similar metric for Alphafold-Multimer model, but has modified to score interactions between the residues of different chains. They call it as metric interface pTM, or ipTM.

In practice, they have used a weighted combination of pTM and ipTM as the model confidence metric, in order to account for some intra-chain confidence in this modelling ranking:

$$\text{Model confidence} = 0.8.\ \text{ipTM} + 0.2\ .\ \text{pTM}$$

3.Datasets:

The recent Protein Data Bank-Multimers set consists of all the targets in Protein Data Bank released between 2018-04-30 to 202-08-02. This set is filtered to proteins with more than one chain, less than 9 chains and less than 1,536 of total residues. It is also clustered with the following approach, which yields a set of 4,433 protein complexes:

1. Assign each chain to its 40% cluster overlap (using the clusters provided by Protein Data Bank).
2. Assign each protein complex single cluster identifier that was the union of all the single chain cluster IDs from step 1.
3. Then randomly pick a single protein complex from every full-complex cluster.

Predictions are made on full complexes, however for recent PDB-Multimers were computed for each chain pair separately. The chains are selected if they are in contact with the ground truth, defined as any heavy atom of one chain being within 5A of any heavy atom of other chain. They were then clustered to remove redundancy and the chain pairs are greedily selected if the cluster id pair was unique amongst those already selected. Finally the data is filtered such

that no chain has greater than 40% template identity to the training set. This has resulted in 2,603 unique in-contact chain pairs.

4.Results:

They have compared Alphafold-Multimer on two datasets: 1) benchmark with other relevant systems [6]. 2)Recent PDB multimers. This benchmark is a set of seventeen heterodimers from Protein Data Bank, they have selected the targets such that there no homologous complexes from Protein Data Bank (PDB).

Recent PDB Multimers is a homology reduced set of 4,333 recent protein complexes from Protein Data Bank. The predictions were made on full complexes before splitting into the in-contact chain pairs (according to ground truth) for the analysis. These pairs are clustered to remove the redundancy and are filtered so that neither chain in the pair had greater than 40% template identity to training set, yielding about 2,603 unique pairs upon metric is reported. For both datasets, dockQ score is measured [7]. DockQ measures the quality of interface and yields a score in the range (0,1). Interfaces are in the score of <0.23 are considered as incorrect and scores >0.8 are of highest quality [8].

4.1 Benchmarks:

They have compared Alphafold-Multimer to different models:

- Alphafold-Linker- In this, they have added a 21residue repeated Glycine-Glycine-Serine linker between each chain before running it as a single chain through Alphafold model [8].

- Alphafold-Gap (colabfold) [9]- This is a third party google colab, that runs Alphafold with a 200 residue gap in the residue index between chains. It uses MMseqs2 for genetics, including the MSA pairing and do not include templates.

- Cluspro – This setup runs the Alphafold on individual chains before docking them together using Cluspro.

- Alphafold-refined cluspro- The ClusPro predictions are refined by feeding them back into the AlphaFold model as templates. The resulting 10 predictions are re-ranked according to the Interface predicted aligned error(PAE) score.

- *AlphaFold refined ClusPro plus AlphaFold*: This pools the *AlphaFold-Gap* predictions with those of the *AlphaFold refined ClusPro* and re-ranks all the 15 predictions using interface PAE.

The figure (2) compares the average dockQ scores between different systems. Alphafold multimer system has an average dockQ score of 0.65, while the cluspro refined Alphafold system has a score of 0.49.

The figure (3) represents the relative difference in dockQ score in between each of the above mentioned systems and Alphafold Multimer and there was 95% confidence intervals over the targets.

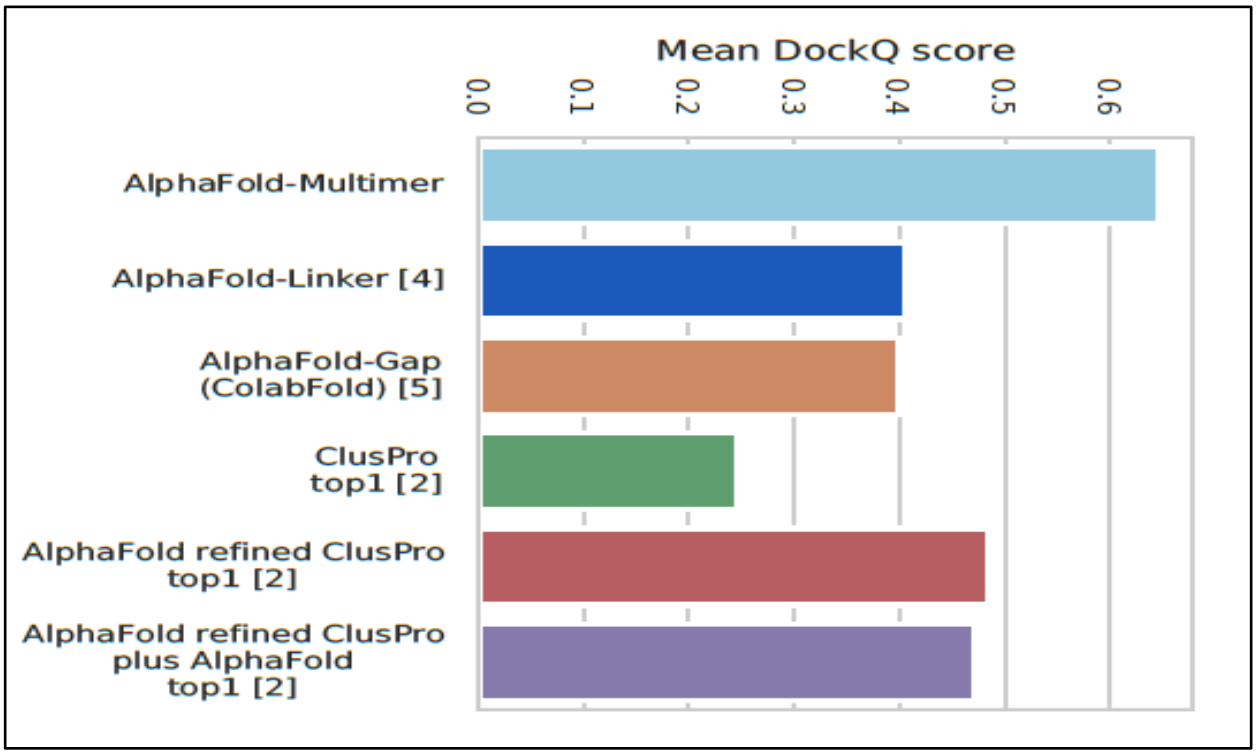

Fig.2.Represents the comparison of DockQ score between AlphaFold-Multimer and related systems.[8]

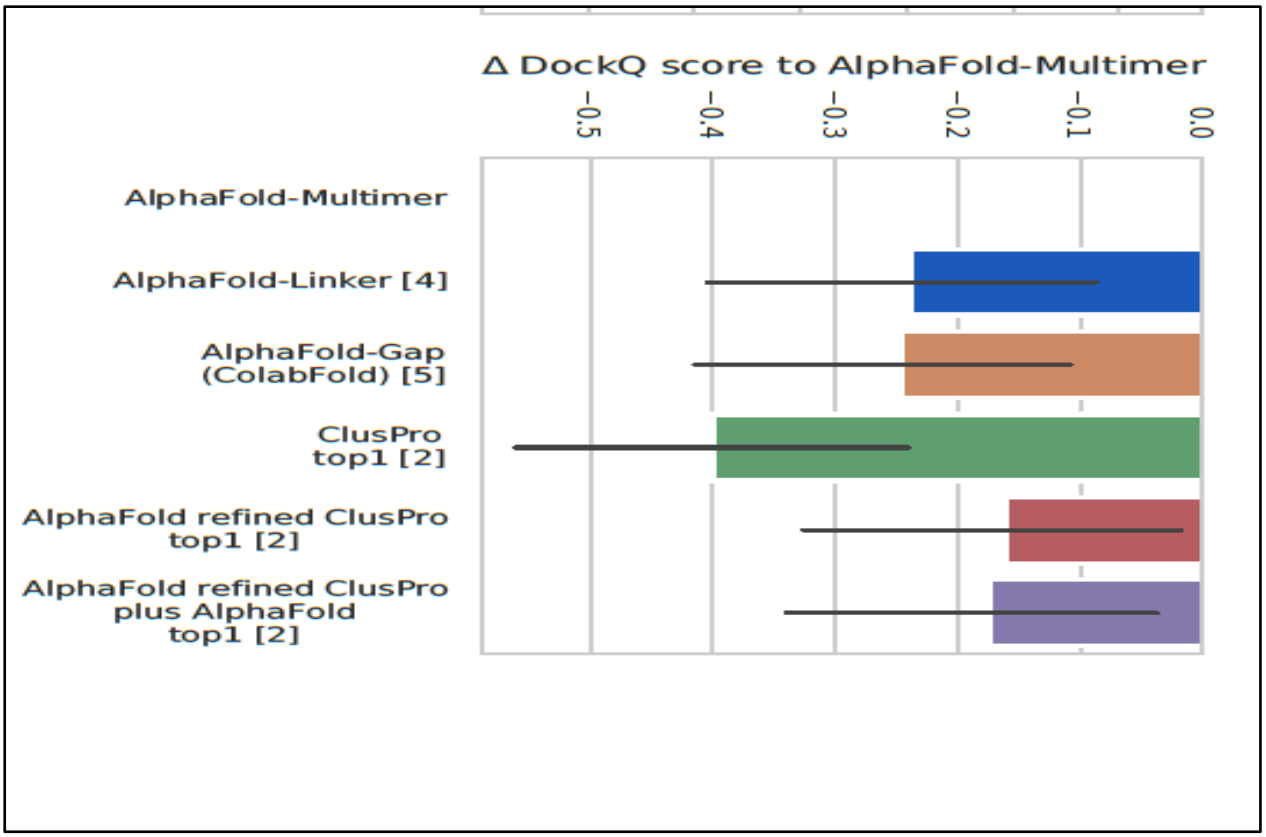

Fig.3: comparison of relative differences in dockQ between Alphafold-Multimer and other systems.[8]

4.2 Recent PDB Multimers:

When the dockQ scores are compared between Alphafold-Multimer and Alphafold Linker, it was found that Alphafold-Multimer is more accurate on both homomeric and heteromeric interfaces. In the homomeric case, the improvement is relatively modest with mean +0.05 mean dockQ score, in heteromeric case, it is more pronounced with +0.19 mean dockQ score compared to Alphafold Linker.

On an average, the performance on homomeric interfaces is better than heteromeric interfaces with Alphafold-Multimer. Alphafold-Multimer system successfully predicts (DockQ $\geq$ 0.23)

heteromeric interfaces in 67% of the cases and successfully predicts the homomeric interfaces in 69% of the cases.

They have compared the performance of Alphafold-Multimer and Alphafold-Linker under two regimes. In Alphafold-Multimer, full complex is folded before extracting the individual chains, so that the system can see the additional context. In the second regime, as Alphafold-Multimer as monomer and Alphafold, the individual chains are folded in isolation. In the case of heteromers, the Alphafold-Multimer is less accurate than the Alphafold when give the single chains with incorrect monomer stoichiometry, but it is more accurate than the Alphafold when single chains are predicted as a part of full complex. This similar effect was not observed in Alphafold- Linker system.

The following fig.5. show number of examples where the Alphafold-Multimer successfully predicted the right multimeric structures.

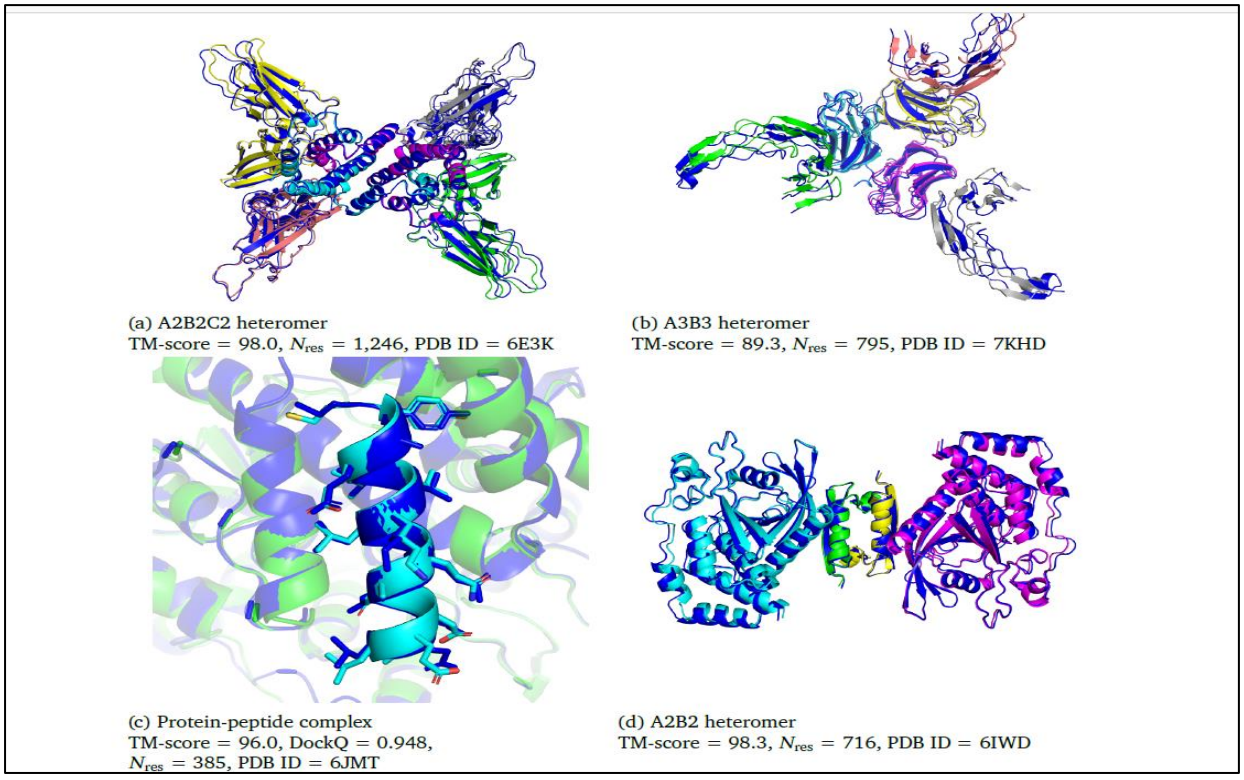

Fig.5: Structures predicted by Alphafold-Multimer. Visualized are ground truth structures (blue) and predicted structures (green).[8]

Method	Mean DockQ Score	Incorrect Count	Acceptable Count	Medium Count	High Count
AlphaFold-Multimer	0.65	3	0	8	6
AlphaFold refined ClusPro	0.49	7	1	4	5
AlphaFold refined ClusPro plus AlphaFold	0.47	7	1	5	4
AlphaFold-Linker	0.41	7	1	5	4
AlphaFold-Gap (ColabFold)	0.40	8	1	6	2
ClusPro	0.25	10	3	4	0

Table S1 | Performance on the Benchmark 2 dataset from [2] consisting of 17 heterodimers with low training set similarity. The following CAPRI definitions were used:
Incorrect: $0 \leq \text{DockQ} < 0.23$
Acceptable: $0.23 \leq \text{DockQ} < 0.49$
Medium: $0.49 \leq \text{DockQ} < 0.80$
High: $0.80 \leq \text{DockQ}$

Table S1 represents the performance of AlphaFold-Multimer system when compared to other systems.[8]

Target	DockQ Score
5ZNG	0.02
6A6I	0.05
6GS2	0.50
6H4B	0.80
6IF2	0.74
6II6	0.73
6ONO	0.64
6PNQ	0.56
6Q76	0.91
6U08	0.93
6ZBK	0.79
7AYE	0.88
7D2T	0.78
7M5F	0.89
7N10	0.86
7NLJ	0.06
7P8K	0.86

Table S2 | Results per target for AlphaFold-Multimer, on the 17 heterodimer Benchmark 2 dataset from [2]

Table S2 represents the performance of AlphaFold-Multimer on 17 heterodimer proteins.[8]

(a) Homomeric Interfaces

Method	Mean DockQ Score	Incorrect %	Acceptable %	Medium %	High %
AlphaFold-Linker	0.476	35.7	10.1	25.4	28.8
AlphaFold-Multimer	0.523	30.7	9.83	25.1	34.3

(b) Heteromeric Interfaces

Method	Mean DockQ Score	Incorrect %	Acceptable %	Medium %	High %
AlphaFold-Linker	0.290	57.5	10.4	20.6	11.4
AlphaFold-Multimer	0.479	32.7	11.9	33.1	22.3

Table S3 represents the performance of AlphaFold-Multimer on homomeric and heteromeric interfaces.[8]

5.Discussion:

By modification of Alphafold system architecture to natively handle multimers, they are able to provide high accuracy predictions for large fraction of Protein Data Bank complexes, surpassing the accuracy of inference-only modifications to the Alphafold system.

They have observed that the performance is generally for homomeric interfaces than for heteromeric interfaces. Since in the homomeric case, the MSA will readily encode the evolutionary information about complex interfaces and while this information is limited and harder to access in case of the heteromeric interfaces.

As a limitation, they observed that Alphafold-Multimer system is generally not able to predict the binding of antibodies and this remains as an area for further future work. They have shown that the confidence metrics provided by the model correlate well with the true accuracy, something which is vital for useability of a structure prediction model. By allowing the accurate prediction of the protein complexes, we hope that this method will enable biologists to further accelerate the recent progress in the structural bioinformatics.

6.Conclusion:

In this study, they have not yet implemented multimer templates or self-distillation of multimer predictions, so there is a likely substantial scope for the future accuracy predictions.

And as a limitation, they also observed that Alphafold-Multimer is generally unable to predict the binding of antibodies and this remains as an area for future work.

7.References:

1.Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583-589.

2.Wall, D.; Fraser, H.; Hirsh, A., Detecting putative orthologs. *Bioinformatics* **2003**, *19*, 1710-1711.

3.Zhou, T.-m.; Wang, S.; Xu, J., Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis. *bioRxiv* **2018**, 240754.

4.Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J. M., RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research* **2021**, *49*, *D437-D451*.

*5.Steinegger, M.; Söding, J., Clu*stering huge protein sequence sets in linear time. *Nature communications* **2018**, *9*, 1-8.

6.Ghani, U.; Desta, I.; Jindal, A.; Khan, O.; Jones, G.; Kotelnikov, S.; Padhorny, D.; Vajda, S.; Kozakov, D., Improved docking of protein models by a combination of AlphaFold2 and ClusPro. *bioRxiv* **2021**.

7.Basu, S.; Wallner, B., DockQ: a quality measure for protein-protein docking models. *PloS one* **2016**, *11*, e0161879.

8.Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A. W.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., Protein complex prediction with AlphaFold-Multimer. *Biorxiv* **2021**.

9.Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M., ColabFold-Making protein folding accessible to all. **2021**.