Quantifying the Quality of AlphaFold2 Protein Structure Prediction With RMSD

Cynthia Pan, Tammie Tam, Patricia Tran

1 Introduction

1.1 Protein Structure Prediction

Proteins are the critical components that keep life running, whether they're transporting ions to maintain homeostasis or breaking down complex molecules to a simple usable form. All of these functions are made possible because of the protein's structural properties. Yet, as complex as proteins are in their function and structure, they are all made out of the same building blocks, amino acids. There are 20 types of amino acids that all of life uses, and each amino acid has intrinsic properties that allow for different interactions with each other and other biomolecules in the cellular environment.

Scientists have figured out how to obtain information about the amino acids sequences from genes in DNA. However, one long-standing challenge in biology has been about trying to resolve a protein's structure. With a clearer picture of a protein's structure, scientists can better design experiments to more deeply explore its functions. Furthermore, understanding how proteins work can aid in advancing protein designs for medical treatments.

Currently, the gold standard methods to experimentally resolve a protein's structure are X-ray crystallography, NMR spectroscopy, and electron microscopy, which are all slow and tedious processes [1]. Hence, to date, there has only been approximately 180,000 protein structures experimentally verified and registered in the Protein Data Bank (PDB) out of the many millions of existing proteins [2]. To accelerate this process, scientists have tried to turn to computational methods in order to predict protein structures. Some have attempted to use existing templates from the PDB to guide protein structure prediction. However, one challenge with this approach has been the lack of experimentally verified protein structures up until recent times [3]. With few available templates, there are less chances of finding one that is most evolutionary related and similar to the protein of interest. Others have tried to bypass this issue by developing methods to construct protein structures without any templates. However, it has been proven challenging to determine the lowest energy state of proteins—especially those of larger size—due to computational limits in determining all possible conformations and their energy state [3].

Recently, a breakthrough with artificial intelligence offering near-experimentally accurate protein structure predictions has emerged as AlphaFold2.

1.2 AlphaFold2

AlphaFold2 is an artificial intelligence system that utilizes deep learning techniques to accurately predict protein structures. It has allowed for significant advancement in solving the structure prediction piece of the protein folding problem: how a protein's one-dimensional amino acid sequence determines its three-dimensional structure [4, 5]. Previous attempts at solving this problem have not been successful as they require similar protein structures to be known in order to predict accurately, something that AlphaFold2 does not require [6]. What used to be computationally expensive to do is now possible in only a few hours using AlphaFold2.

The AlphaFold2 pipeline is shown below in Figure 1. It is divided into two main modules: the Evoformer and the structure module, that link together and utilize neural network architectures to perform protein structure prediction. To prepare the given input sequence for the Evoformer, AlphaFold2 first starts by finding sequences homologous to the input sequence by constructing a multiple sequence alignment. It also looks for templates which are proteins with similar structures to that of the input sequence. Using these templates, it constructs a pair representation that looks at which amino acids within a template input pair are likely to associate with each other. After this preprocessing step is complete, AlphaFold2 passes the generated multiple sequence alignment and pair representation to the Evoformer. The Evoformer then processes this information through 48 repeated layers and outputs a graphical representation of the useful information. The Evoformer then passes its output into the structure module, which constructs a three-dimensional structure of the protein. After a structure has been constructed, it gets passed back to the Evoformer and the whole process repeats three times. This recycling process allows the system to fine-tune its prediction and improve its accuracy. AlphaFold2 outputs a total of five different structure predictions which are ranked using the Local Distance Difference Test (IDDT).

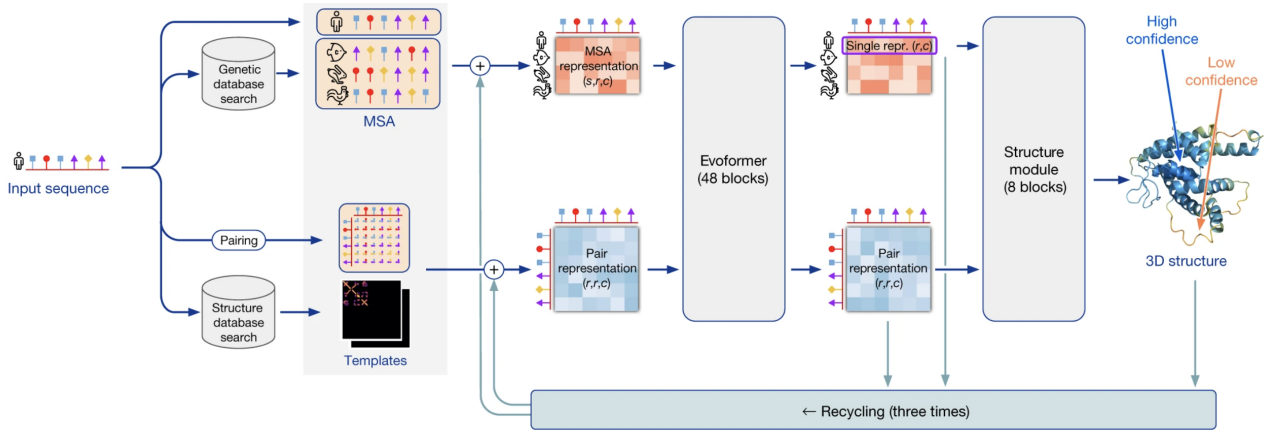

Figure 1: AlphaFold2 Pipeline [6]

1.3 RMSD

To determine the quality of a protein structure prediction, root mean square deviation (RMSD) is used to compare protein structures. RMSD is the most commonly used quantitative measure of the similarity between two superimposed atomic coordinates. It is used for measuring the difference between the backbones of a protein from its initial structural confirmation to its final position [7].

According to "Statistical Validation of the Root-Mean-Square-Distance, a Measure of Protein Structural Proximity" by Oliviero Carugo, there are some drawbacks that are worth mentioning in using RMSD [8]:

1. Small deviations in one part of a protein can create large RMSD values, which implies that the two structures are different. However, that may not always be the case.

2. The RMSD is not the best metric when comparing the structures that are not similar to each other.

3. RMSD values are smaller for protein structure pairs at better resolution. On the other hand, RMSD values are observed to increase if the two proteins are at different resolutions.

Having said that, RMSD is still very common for structural biologists to use and publish RMSD values because:

1. When applied to proteins of similar form, RMSD values are considered reliable RMSD values are considered reliable indicators of variability [9].

2. RMSD is a good indicator for structural identity [9]. When RMSD is 0, the structures are identical. When it increases, the two structures become more different [8].

3. In protein structure prediction, the RMSD value between predicted and experimentally determined structures can be considered a successful prediction when the RMSD is small (< 3 å; a typical RMSD for homologous proteins) [10].

1.4 Presentation of Problem

We are interested in determining how well protein structure prediction software is at predicting the structure of a protein sequence. For our project, we will use AlphaFold2 with the default settings to generate structure predictions of two different protein sequences before comparing them to a gold standard. We will focus on two different protein sequences: Fimbrial adhesin | Proteus mirabilis (strain HI4320) (529507) and CST complex subunit CTC1 | Homo sapiens (9606), referring to them as sequence 1 and sequence 2 respectively. To assess the accuracy of the protein structure prediction, we will be comparing the models to the gold standard structure with RMSD. Additionally, to assess the precisions of the protein structure prediction model outputs, we will compare the five models that AlphaFold2 generates to each other with RMSD.

2 Methods

Our program consists of three Python subprograms, **rmsd.py**, **accuracy.py**, and **precision.py**.

2.1 Running the Program

Replace the file path placeholders in the commands below with the file paths on your machine to the structures you want to compare.

2.1.1 Before Running

Install the Biopython library using the following command:

```
pip install biopython
```

2.1.2 Running rmsd.py

To run rmsd.py, use the following command:

```
python3 src/rmsd.py struct1_file_path struct2_file_path
```

2.1.3 Running accuracy.py

This program expects 6 different structures as input, with the first being the gold standard.
To run accuracy.py, use the following command:

```
python3 src/accuracy.py gold_standard_file_path struct1_file_path struct2_file_path
struct3_file_path struct4_file_path struct5_file_path
```

2.1.4 Running precision.py

This program expects 5 different structures as input, specifically the 5 structures that AlphaFold2 outputs. To run precision.py, run the following command:

```
python3 src/precision.py struct1_file_path struct2_file_path struct3_file_path
struct4_file_path struct5_file_path
```

2.2 rmsd.py

The program rmsd.py takes two pdb files as input and outputs the Root Mean Square Deviation (RMSD) between their alpha carbon (CA) atoms. It contains three functions: create_structure(), get_coordinates(), and calc_RMSD(). The first function create_structure() creates a structure object from a given pdb file by using the Bio.PDB package from the Biopython library. This package parses pdb files, allowing for easy data extraction. The second function get_coordinates() takes in a structure object and outputs a list of three dimensional coordinates. It utilizes the same Bio.PDB package used previously to isolate the lines corresponding to the CA atoms for the given structure and extracts their coordinates. The third function calc_RMSD() computes the best fit RMSD using

the algorithm specified in "Using Quaternions to Calculate RMSD" by Coutsias et al. [11]. It takes the coordinates of two structures to be compared as input. We call the coordinates of the first structure x_k and the coordinates of the second structure y_k. When this function is executed, we start by first shifting the set of coordinates of both structures by their respective barycenters: $\bar{x}$ and $\bar{y}$ and call the set of these shifted coordinates $\tilde{x_k}$ and $\tilde{y_k}$. Next, following equation 2, we calculate R_{ij} values. Using equation 3, we use the resulting R_{ij} values from the previous step to compute a 4x4 matrix called $\mathscr{F}$. Using the linalg package from the numpy library, we calculate the maximum eigenvalue of $\mathscr{F}$. Lastly, we use equation 4 to calculate the best fit RMSD value. As the value of the summation can be smaller than 2 times the maximum eigenvalue, it is possible for the numerator to be negative. Because negative values cannot be square rooted, if the numerator is less than the threshold of 0.01, we set its value to zero.

2.2.1 Equations

$$\tilde{x_k} := x_k - \bar{x}, \;\; \tilde{y_k} := y_k - \bar{y} \tag{1}$$

$$R_{ij} = \sum_{k=1}^{N} x_{ik}\, y_{jk}, \;\; i,j = 1,2,3 \tag{2}$$

$$\mathscr{F} = \begin{bmatrix} R_{11}+R_{22}+R_{33} & R_{23}-R_{32} & R_{31}-R_{13} & R_{12}-R_{21} \\ R_{23}-R_{32} & R_{11}-R_{22}-R_{33} & R_{12}+R_{21} & R_{13}+R_{31} \\ R_{31}-R_{13} & R_{12}+R_{21} & -R_{11}+R_{22}-R_{33} & R_{23}+R_{32} \\ R_{12}-R_{21} & R_{13}+R_{31} & R_{23}+R_{32} & -R_{11}-R_{22}+R_{33} \end{bmatrix} \tag{3}$$

$$e = \sqrt{\frac{\sum_{k=1}^{N} (\tilde{x_k^2} + \tilde{y_k^2}) - 2\lambda_{max}}{N}} \tag{4}$$

2.3 accuracy.py

To see how accurate AlphaFold2 is at predicting the structure of a protein sequence, we compare the five models that it outputs to the corresponding gold standard by computing the RMSD values between them. The program accuracy.py consists of a single function called accuracy() that does exactly this. It reads in a list of coordinates that correspond to each structure with the first set of coordinates belonging to the gold standard and the remaining belonging to AlphaFold2 models 1 through 5 respectively. It computes the RMSD value between each model and the gold standard, and stores this value in a list that it outputs after all comparisons are completed.

2.4 precision.py

To see how precise AlphaFold2 is, i.e. how similar or different the output models of AlphaFold2 are to each other, we compare each pair of models by computing the RMSD between them. The program precision.py utilizes the functions create_structure(), get_coordinates(), and calc_RMSD() defined in rmsd.py to do so. It takes the five models that were obtained by running AlphaFold2 as inputs and calls the create_structure() function defined in rmsd.py to create structure objects

for each model. It then calls the get_coordinates() function that is also defined in rmsd.py to get the coordinates of the CA atoms for each model and stores them in a list. The program contains a single function called precision() that then takes in this list of coordinates and iterates over one pair of coordinates at a time, calling calc_RMSD() each time. The resulting output is a symmetric matrix of RMSD values.

2.5 Brief Analysis

To verify that our main program rmsd.py works correctly, we compared each structure to itself. When the two structures that are being compared are identical, the RMSD value should be equal to zero. We used the gold standards as well as all of the AlphaFold2 outputs for both sequence 1 and sequence 2 as our test cases. As expected, the program works as intended, providing an RMSD value of zero for each test.

3 Results

3.1 PyMol Visualization of Sequences

Figure 2: Fimbrial adhesin | Proteus mirabilis (strain HI4320) (529507)

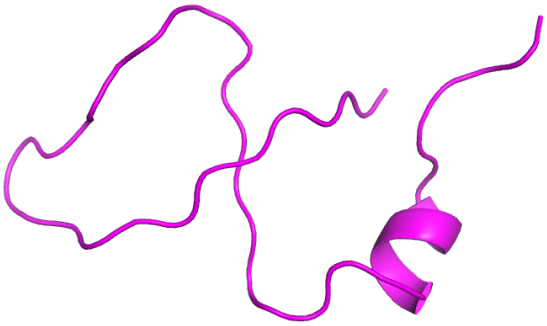

Figure 3: CST complex subunit CTC1 | Homo sapiens (9606)

3.2 Accuracy

We compared each model that AlphaFold2 outputted to the gold standard.

Sequence	Model 1	Model 2	Model 3	Model 4	Model 5
1	3.84629	3.85993	3.9042	3.82965	3.95642
2	12.40201	12.46717	12.37333	12.19234	12.17319

Table 1: RMSD Between Gold Standard and Prediction Models

3.3 Precision

We compared each model that AlphaFold2 outputted to each other.

Sequence 1	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	0.00000	0.40413	0.37964	0.31701	0.62855
Model 2	0.40413	0.00000	0.31305	0.42827	0.55716
Model 3	0.37964	0.31305	0.00000	0.39400	0.62596
Model 4	0.31701	0.42827	0.39400	0.00000	0.68198
Model 5	0.62855	0.55716	0.62596	0.68198	0.00000

Table 2: RMSD Between Prediction Models for Sequence 1

Sequence 2	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	0.00000	0.39962	0.46051	0.47615	0.64000
Model 2	0.39962	0.00000	0.37723	0.54118	0.66701
Model 3	0.46051	0.37723	0.00000	0.60627	0.64432
Model 4	0.47615	0.54118	0.60627	0.00000	0.39344
Model 5	0.64000	0.66701	0.64432	0.39344	0.00000

Table 3: RMSD Between Prediction Models for Sequence 2

4 Discussion

4.1 Accuracy of AlphaFold2 Predicted Models

4.1.1 How well does AlphaFold2 predict sequence 1 vs sequence 2?

Using AlphaFold2, we found all prediction models of sequence 1 to have similarly low RMSD values at around 3.82 - 3.95 [Table 1]. Such a low RMSD value indicates that the prediction model is very similar to the gold standard, which can be seen in Figures 5 and 7. Compared to sequence 1, the AlphaFold2 prediction models of sequence 2 had high RMSD values at around 12.173 - 12.467 [Table 1]. Such a high RMSD value indicates that the prediction model is very different from the gold standard, which can be seen in Figure 9. One reason why AlphaFold2 may be better at predicting sequence 1 structure than sequence 2 structure could be because the longer sequence 1 can provide more contextual information to inform structure. To determine whether this is the case, researchers can test AlphaFold2 on a larger sample of sequences of varying length and calculate the correlation between sequence length and RMSD score.

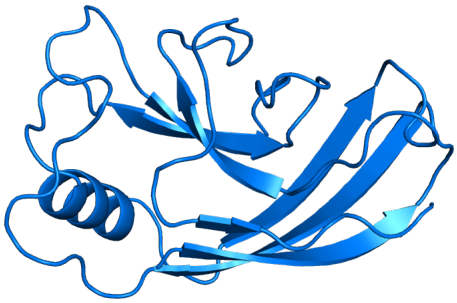

Figure 4: Seq 1 - Model 5

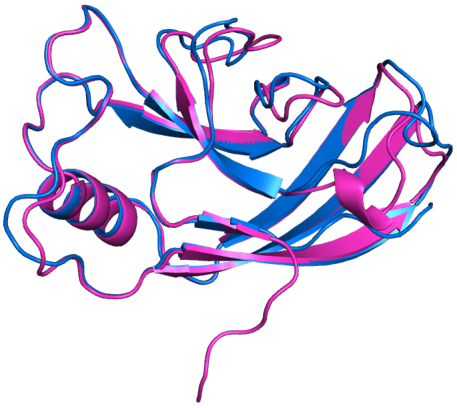

Figure 5: Seq 1 - Gold + Model 5 Alignment

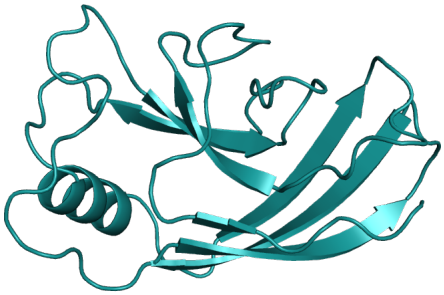

Figure 6: Seq 1 - Model 4

Figure 7: Seq 1 - Gold + Model 4 Alignment

Figure 8: Seq 2 - Model 5

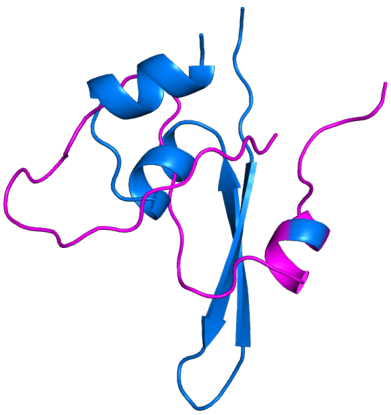

Figure 9: Seq 2 - Gold + Model 5 Alignment

4.1.2 Is the model ranked first in confidence by AlphaFold2 actually the best?

AlphaFold2 outputs five prediction models of the same proteins and ranks them using the Local Distance Difference Test. RMSD can be used as another standard to determine whether their ranking holds up and is trustworthy.

For sequence 1, AlphaFold2 assigned model 5 to be the most confident model. However, by RMSD calculations, we found model 5 to have the highest RMSD value, making it the model least similar to the gold standard. Instead, model 4 is actually the best model with the lowest RMSD value of 3.8296. RMSD values indicate the model from best to worst to be model 4, model 1, model 2, model 3, model 5. Thus, overall, the best to worst model ranking as determined by RMSD does not follow the ranking of AlphaFold2 for sequence 1.

For sequence 2, AlphaFold2 assigned the models from most confidence to least confidence as model 5, model 3, model 2, model 1, and model 4. Unlike sequence 1, model 5 is in fact the best model with the lowest RMSD value at 12.17. However, while model 4 was deemed by AlphaFold2 to be the model it was least confident in, model 4 is actually the second most similar compared to the gold standard model. RMSD values indicate the model from best to worst to be model 5, model 4, model 3, model 1, model 2. Again, overall, the best to worst model ranking as determined by RMSD does not follow the ranking of AlphaFold2 for sequence 2.

Therefore, given both test cases of sequence 1 and sequence 2, we recommend using RMSD to determine which AlphaFold2 model is the best, instead of relying on the first model provided by AlphaFold2.

4.2 Precision of AlphaFold2 Predicted Models

The RMSD values for all comparisons of models for both sequence 1 and sequence 2 are all less than 1 which indicates that overall, the models that AlphaFold2 outputs seem to be very similar to each other. Since the models are all so similar to each other, we wanted to see which models exactly were the most similar and whether or their rankings also coincide with similarity. We predicted that the models that are closer in rank are more similar to each other i.e. they have a lower RMSD value.

For sequence 1, models 2 and 3 were the most similar to each other, followed by models 1 + 4, 1 + 3, 3 + 4, 1 + 2, 2 + 4, 3 + 5, 1 + 5, and 4 + 5. As stated in the accuracy section, AlphaFold2 assigned model 5 to be the most confident model and model 4 to be the second most confident model. We predicted that because models 4 and 5 were considered the best models by AlphaFold2, they would be the most similar to each other. Yet, this does not seem to be the case as models 4 and 5 are actually the least similar out of all the models according to RMSD.

For sequence 2, models 2 and 3 were also the most similar to each other, followed by models 4 + 5, 1 + 2, 1 + 3, 1 + 4, 2 + 4, 3 + 4, 1 + 5, 3 + 5, and 2 + 5. AlphaFold2 ranked model 3 as the second best model and model 2 as the third best model. In this case, models 2 and 3 are the

most similar to each other and they are also closest in rank. AlphaFold2 assigned model 5 to be the most confident model and model 4 to be the least confident model. According to the RMSD comparisons, models 4 and 5 were the second most similar to each other. It does not seem to be the case that the models that are more similar to each other are closer in rank.

We cannot conclude that the models that are closer in rank are also more similar to each other.

5 Conclusion

From the research for our project, on average, the range for precision is between 0 to 1 for both sequence 1 and sequence 2. The range for accuracy for sequence 1 is 3.5 to 4 and for sequence 2 is 12 to 12.5. Looking at Figures 5 and 7 for sequence 1, we could see that the Gold + Model 5 Alignment and Gold + Model 4 Alignment are overall similar. However, towards the C-terminus of the protein sequence, the gold sequence protein's tail diverges from model 4 and model 5, which may have resulted in a higher RMSD value. Thus, in the future, we are interested in examining whether removing a portion of the C-terminus affects the RMSD value.

Regarding the information related to RMSD, multiple reliable sources, as cited below, assessed that when the RMSD is less than 3, then it is considered a successful prediction [10]. From the output of our project, our results are comparable with those from reliable sources.

In summary, we can safely conclude that AlphaFold2 is more precise than it is accurate.

6 Bibliography

1. "PDB101: Learn: Guide to Understanding PDB Data: Methods for Determining Structure." n.d. RCSB: PDB-101. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure.

2. AlQuraishi, Mohammed. 2021. "Protein-Structure Prediction Revolutionized." Nature 596 (7873): 487–88. https://doi.org/10.1038/d41586-021-02265-4.

3. Zhang, Yang. 2008. "Progress and Challenges in Protein Structure Prediction." Current Opinion in Structural Biology 18 (3): 342–48. https://doi.org/10.1016/j.sbi.2008.02.004.

4. Dill, Ken A., S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. 2008. "The Protein Folding Problem." Annual Review of Biophysics 37 (June): 289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558.

5. "AlphaFold: a solution to a 50-year-old grand challenge in biology." n.d. Deepmind. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

6. Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." Nature 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

7. Aier, Imlimaong, Pritish Kumar Varadwaj, and Utkarsh Raj. 2016. "Structural Insights into Conformational Stability of Both Wild-Type and Mutant EZH2 Receptor." Scientific Reports 6 (1): 34984. https://doi.org/10.1038/srep34984.

8. Carugo, Oliviero. 2007. "Statistical Validation of the Root-Mean-Square-Distance, a Measure of Protein Structural Proximity." Protein Engineering, Design and Selection 20 (1): 33–37. https://doi.org/10.1093/protein/gzl051.

9. Carugo, Oliviero, and Sándor Pongor. 2008. "A Normalized Root-Mean-Square Distance for Comparing Protein Three-Dimensional Structures." Protein Science 10 (7): 1470–73. https://doi.org/10.1110/ps.690101.

10. Reva, Boris A, Alexei V Finkelstein, and Jeffrey Skolnick. 1998. "What Is the Probability of a Chance Prediction of a Protein Structure with an Rmsd of 6 å?" Folding and Design 3 (2): 141–47. https://doi.org/10.1016/S1359-0278(98)00019-4.

11. Coutsias, Evangelos A., et al. "Using Quaternions to Calculate RMSD." Journal of Computational Chemistry, vol. 25, no. 15, Nov. 2004, pp. 1849–57. DOI.org (Crossref), https://doi.org/10.1002/jcc.20110.