Explainable AlphaFold-Multimer

Luke Mozarsky
Department of Computer Science
Yale University
luke.mozarsky@yale.edu

Michael Ying
Department of Computer Science
Yale University
michael.ying@yale.edu

Abstract

The task of predicting and understanding protein structures has been a challenge for a wide range of researchers across many disciplines. With the recent wave of developments in computing and machine learning, deep learning models for protein structure prediction have surfaced. In this work, we explore AlphaFold-Multimer, a model derived from the prior AlphaFold that was built to predict structures of protein complexes. In particular, we provide an explainability framework to understand how the model produces outputs and compare results for both AlphaFold and AlphaFold-Multimer, as well as for general protein complexes and the class of antibody-antigen binding pairs.

1 Introduction

The prediction of protein structures from amino acid sequences is an important field of study in the world of bioinformatics that has seen much attention in the research community, especially in recent years with the age of machine learning and expanded computational capabilities. Knowledge of protein structures is key in understanding the the way proteins function and interact with other molecules. As a result, understanding protein structures is important to a broad range of biomedical researchers, from those working in pharmaceutical drug development to those designing new vaccines.

Traditional wet-lab approaches to determining protein structures at atomic resolution involve methods like crystallization and x-ray diffraction, which are costly and time-consuming processes. With the rise of machine learning, many computational techniques have been proposed to predict the structures of proteins. Among such models is AlphaFold, a model developed and released by researchers at DeepMind that is capable of predicting protein structures with groundbreaking accuracy. Not long after the release of AlphaFold, the researchers at DeepMind released AlphaFold-Multimer, a model based off AlphaFold with some small changes designed to make it amenable to predicting protein complexes.

As AlphaFold and AlphaFold-Multimer continues to be improved, it is important to investigate the models' inner workings to understand how it generates predictions. Such insights can help us trust the model's predictions before applying them to real-world tasks. In this project, we apply explainability methods to AlphaFold-Multimer in hopes of understanding the models' inner workings and learning more about the way proteins bind to each other to form complexes. In particular, the original authors of the AlphaFold-Multimer paper remark that the model does not predict antibody-antigen complexes very well. The class of antibody-antigen protein complexes is a critical area of study in the field of immunology, and progress can help us understand the nature of diseases and our immune systems. As such, we also look at explainability results for this special class of protein complexes to gain insights into both model predictions and performance.

2 Related Works

AlphaFold is the base protein structure prediction model we would like to investigate in this project. AlphaFold incorporates physical, biological, and evolutionary knowledge about protein sequences and structures to make its predictions. AlphaFold utilizes multi-sequence alignments and pairwise features as inputs into a core network consisting of Evoformer blocks which refine these features before sending the results to the structural prediction phase of the network [3]. The model returns 3D coordinate predictions in Protein Data Bank (PDB) format alongside some metrics to quantify model certainty in its outputs.

Alphafold Multimer incorporates some changes designed to improve the model's capabilities in predicting protein complexes [2]. These changes include special multi-sequence alignment construction given the multiple input chains of amino acids and some changes to the loss and training stage designed for the protein complex task.

There are many other models that have been proposed to predict protein structures. RoseTTAFold is one such model that employs a three-track model for processing sequence, distance, and coordinate information in parallel [1]. IgFold is another model that uses pre-trained language models and was designed specifically for predicting antibody structures with high accuracy and efficiency [5].

While there are many models which seek to predict protein structures, there is a lack of explainability research on these models in the literature. To our knowledge, there is one explainability paper describing ExplainableFold, a framework of explaining the structural predictions of AlphaFold [7]. ExplainableFold treats AlphaFold as a blackbox and seeks to find necessary and sufficient explanations for a structural prediction. They employ amino acid deletion and substitution strategies to optimize and find both the minimal perturbation possible while drastically changing the model's outputs (ie. necessity) and the maximal perturbation possible without major impacts to the model's outputs (ie. sufficiency).

While ExplainableFold already analyzes AlphaFold with perturbation strategies, we seek to further explainability research on AlphaFold with similar deletion perturbation methods by analyzing AlphaFold-Multimer, comparing results with the baseline AlphaFold model, and also looking at explanability in the context of antibody-antigen binding predictions, all of which are novel pursuits to our knowledge.

3 Proposed Approaches

3.1 Explainability Through Deletion Perturbation

The core of our approach is to generate explainability through deletion perturbation approaches. Following the general black-box strategy of perturbing inputs and observing impacts on outputs, we seek to observe the impact on the outputs of AlphaFold and AlphaFold-Multimer when we iterate through the amino acids constituting a query protein sequence and delete them one at a time. By observing the impact on model outputs when any given amino acid is deleted, we can measure how important that amino acid was with respect to the model's prediction. We run this instance-level explainability method on a range of input protein sequences and compare results.

3.2 Explainability Metrics

To quantitatively measure changes in the models' outputs following perturbation, we use the following three output metrics given by AlphaFold and AlphaFold-Multimer. The first metric we utilize as a baseline is the predicted template modeling (pTM) score. For each prediction that AlphaFold makes, it returns a pTM score describing the model's confidence as to how accurate the overall structure of the output protein matches an experimental template from, say, the PDB. In essence, it is a prediction of the true template modeling (TM) score. If a perturbation drastically impacts the structural prediction, we would expect to see a drop in the pTM score. We use this metric as our baseline since it is a direct measure of how well AlphaFold believes it has predicted the true structure of the protein.

The second metric is the predicted Local Distance Difference Test (pLDDT) score. The Local Distance Difference Test (LDDT) score measures how well local distances among the amino acids in

a predicted structure match with the "true" local distances. While the pLDDT score also measures how confident AlphaFold is in its output, it functions at a more localized level. In fact, AlphaFold computes a pLDDT score for each amino acid representing this confidence in structure prediction at the respective local regions, and then it averages the pLDDT scores to return an aggregate metric that we use to quantify model prediction. We compare this metric with the baseline pTM to see whether a metric that relies more on local structures will behave differently upon perturbation tests.

The third and final metric we use is specific to AlphaFold-Multimer and incorporates the model's objective of predicting protein complexes. The interface predicted template modeling (ipTM) score is again a prediction of the quality of the protein complex structure prediction, except averaged over the interface at which the proteins interact to form the complex [9]. As a result, ipTM is expected to be more sensitive to structural changes that impact the interface and thus binding of proteins. We thus expect the ipTM score to behave differently than the baseline pTM score upon perturbation and explainability evaluation.

3.3 Proposed Applications of Perturbation Explainability

We utilize pTM, ipTM, and predicted LDDT scores in the following manner. For a given input sequence, we compute the model prediction for the resulting structure and extract each score (pTM and LDDT for monomers; pTM, ipTM, and LDDT for multimers). This can be denoted as the baseline score for the given input protein or protein complex. We then delete one amino acid from the original input, and run the model again on this perturbed input to get new scores. We repeat the deletion and prediction process for each amino acid in the original sequence. After all scores have been extracted, we subtract the score we get from the modified input from the baseline score. Concretely, for the j-th amino acid in the i-th input sequence, we get a set of "delta scores" where the k-th delta score is computed as

$$(\text{Delta score})_{i,j,k} = (\text{Baseline sequence score})_{i,k} - (\text{Modified sequence score})_{i,j,k}. \quad (1)$$

The "delta score" for a particular metric and amino acid of an input sequence is therefore a measure of how the model's confidence in its predictions changes when that amino acid is deleted. In other words, we use the delta score to measure how important an amino acid is to the model's prediction. Amino acids with small delta scores are less important for the model in forming its prediction because deleting them doesn't change the output very much, whereas amino acids with large delta scores are more important for the model. With this set up, we attempt to explain the models' inner workings in the following settings.

First, we compare the delta scores for the three metrics (where possible) to see whether amino acids tend to affect the three metrics in the same manner. The three metrics measure model confidence in output with three slightly different objectives, so we explore whether certain amino acids tend to impact certain metrics more than others. We inspect correlations of delta scores to look into this objective.

One might also expect a monomer to have different structural properties as a monomer versus as part of a protein complex. As a result, we compare the explainability of multimers versus the constituent monomers. For this comparison, we measure the sensitivity of AlphaFold-Multimer's output for a given multimer input to perturbations. We then measure the sensitivity of AlphaFold's predictions for each of the individual monomeric sequences that composed the multimeric input for AlphaFold-Multimer. Then, we see whether sensitivity properties differ across the sequence of amino acids of the input proteins. Such a comparison allows us to understand differences in how AlphaFold-Multimer treats a protein when it is expected to bind to other proteins, compared to how AlphaFold would predict the structure of the protein on its own.

With these two broad classes of tests (metric against metric, and monomer against multimer), we also have two main sets of data we are interested in. We have a collection of regular protein monomers and multimers and a collection of antibody-antigen pairs. By running the above explainability methods on both regular protein dimers and antibody-antigen pairs, we can compare and understand how AlphaFold and AlphaFold-Multimer generate outputs for normal proteins and for specifically antibody-antigen pairs.

4 Experiments and Results

4.1 Dataset

For predicting the structure of general multi-protein complexes (not necessarily antigen-antibody pairs), we utilized a benchmark dataset used in [9] to evaluate the performance of AlphaFold-Multimer. This dataset was downloaded from the biounits sections of the PDB. While this dataset contained protein complexes with as many as six bound sequences, to simplify the explainability objective we restricted ourselves to dimers, or pairs of proteins. The choice to focus on explaining dimers additionally allows for easy comparison to the explainability results for antigen-antibody pairs also explored in this work. The original dataset contained 13,136 dimers, and we randomly sampled fifteen heteromers (dimers with different constituent proteins) and fifteen homomers (dimers with identical constituent proteins) with total amino acid length less than 200. In particular, we used the FASTA files linked to each protein complex identifier provided in the dataset. FASTA is a commonly used text format for representing amino acid sequences [4]. Each FASTA file contains the individual sequences that compose the given dimer, as well as the PDB four-digit identifier for each sequence. We note that the sequences in this benchmark dataset were withheld from the AlphaFold-Multimer training process; however, since this work focuses on the explainability of AlphaFold-Multimer predictions, whether or not the sequences we use were seen during training should not have a major impact on our results.

For predicting the structure of antibody-antigen pairs, we used 15 random antibody-antigen from a benchmark dataset of 88 pairs used in [8] to evaluate the performance of AlphaFold. We use 15 pairs, all of which had antibody and antigen lengths of roughly 100 to 120 amino acids long, to match that of the general protein complex dataset and for computational complexity concerns. These sequences were provided in a dataframe format, so no further processing was required. This dataset also contains antibody-antigen systems published on Jan 1, 2022, which is after the cutoff date for training data for AlphaFold 2.3.0 and the version of AlphaFold-Multimer we used.

4.2 Perturbation Explainability Results

As an initial test of our explainability methods, we observe the distributions of per-amino acid delta scores for the different metrics and across all input protein types. Figure 1 shows the distribution of the delta scores for the pTM metric on AlphaFold-Multimer.

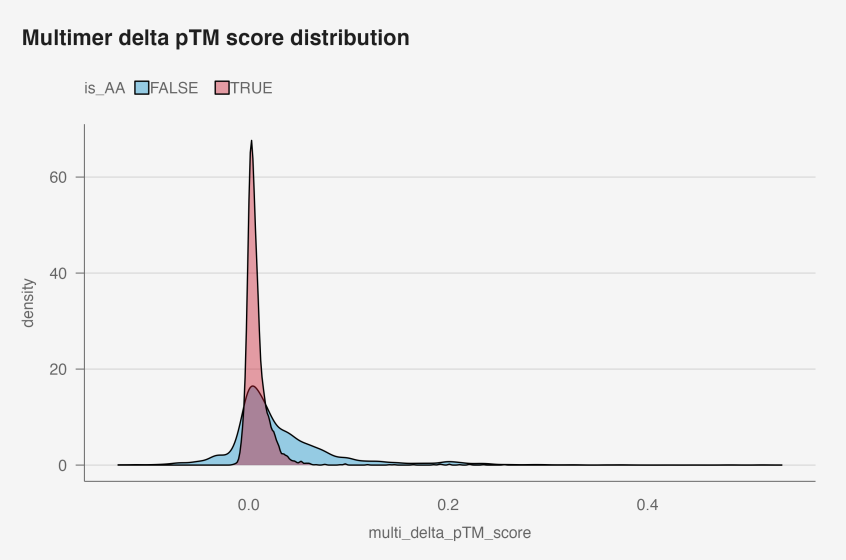


Figure 1: Delta score distribution for pTM for AlphaFold Multimer

We notice that many of the delta scores are near zero, indicating that those corresponding amino acids do not significantly change the model's pTM score. We also see some interesting properties of the distribution. For amino acids (is_AA = TRUE), the delta scores are more centered around

zero with lower variance, potentially indicating that the model prediction is more stable for antibody-antigen pairs. This does not directly contradict the finding that AlphaFold-Multimer performs poorly on antibody-antigen pairs. In fact, this could indicate that the outputs of AlphaFold-Multimer for antibody-antigen inputs are not very sensitive to perturbations in the input, possibly because the model is not very competent in modelling antibody-antigen complexes.

In figure 2, we observe some moderate correlation between the pTM delta and ipTM delta scores. This shows that pTM and ipTM are somewhat similar in the way they capture output changes due to input perturbations. However, we do see that the two delta scores are not always aligned. We also confirm the trend that for antibody-antigen complexes, the scores tend to clump around zero more tightly, with some exceptions where the delta pTM score is around 0.2.

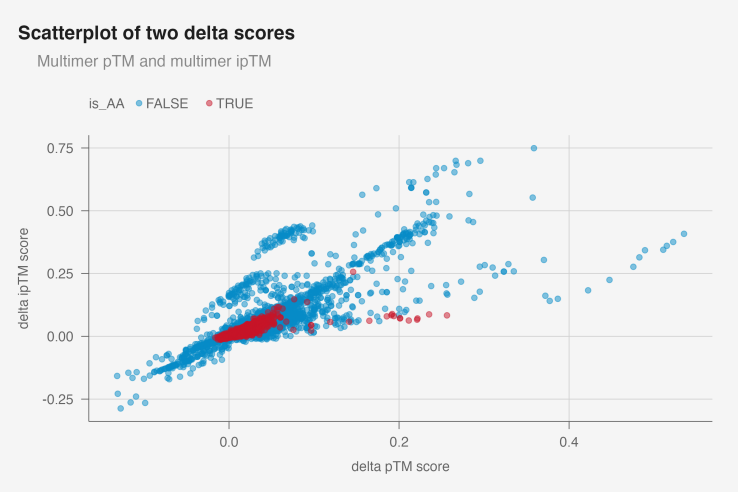
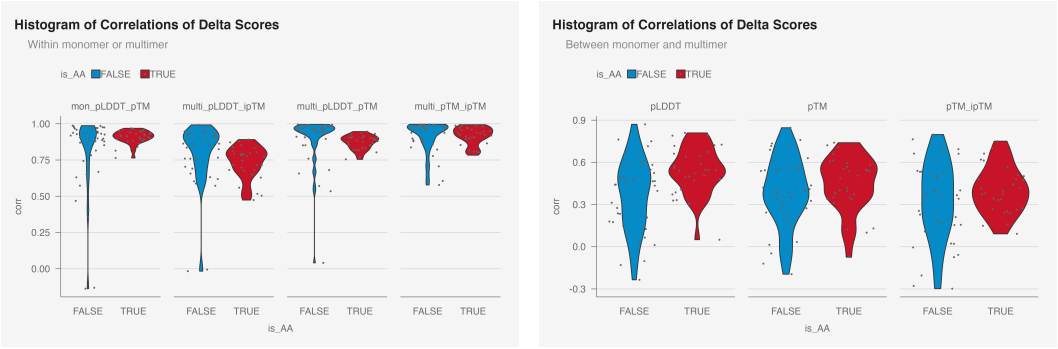


Figure 2: Scatterplot of pTM and ipTM scores for AlphaFold-Multimer

In figure 3, we detail several correlation plots. Since each input protein has a delta score for each metric and each of its amino acids, we compute the correlations between pairs of lists of delta scores. Each data point is one input sequence and has a correlation value. For example, in the leftmost facet of subfigure 3a, each data point is the correlation between the pLDDT scores and pTM scores of its amino acids. We then use violin plots to visualize the distributions of correlations.



(a) Within correlations

(b) Cross correlations

Figure 3: Correlations of delta scores

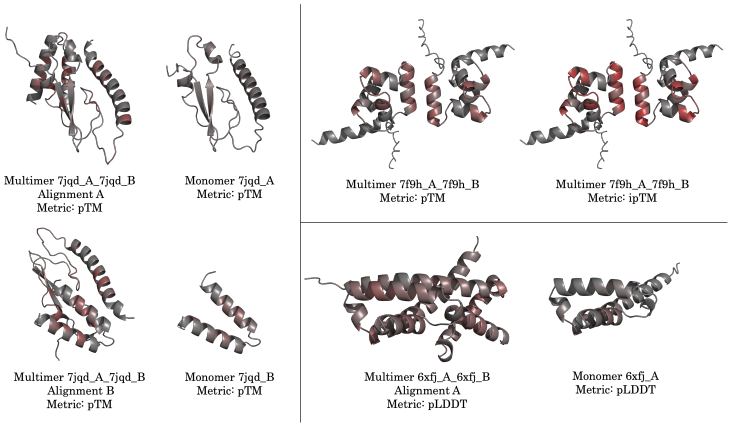
We see that the plots with pLDDT scores tend to contain outliers with near zero correlation, indicating that there are examples where pLDDT is uncorrelated with pTM and ipTM. This could be caused by proteins where the sensitivity of localized structures has little to no relationship to the sensitivity of the entire scruture.

In figure 3b, we notice that cross correlations between the delta scores for monomers and delta scores for those same monomers within complexes tend to be lower compared to the within correlations

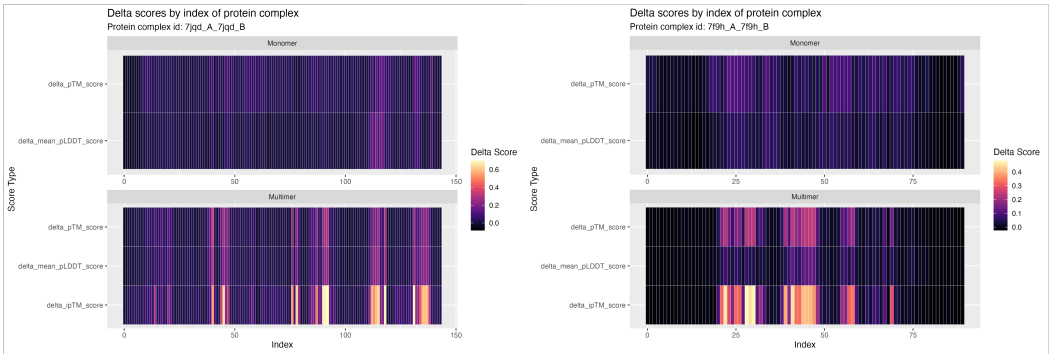
in plot 3a. This indicates that AlphaFold and AlphaFold Monomer sometimes provide different explanations, likely depending on the task at hand. In other words, certain amino acids that are important in structure prediction for a monomeric protein are not necessarily as important when predicting the structure of that protein within a complex. This is an illuminating result that supports the idea of how protein structures can change depending on whether the protein exists independently or as part of a protein, and how the prediction tasks of AlphaFold and AlphaFold-Multimer are different.

4.3 Heatmaps and Molecule Visualization

To visualize the structures predicted by AlphaFold and AlphaFold-Multimer, we used PyMOL [6], an open source protein structure visualization software. Using PyMOL, we were able to view the structures predicted by the model and color each amino acid by a delta score. Figure 4a shows a number of predicted dimer structures colored by delta score. The leftmost panel shows a multimer and its constituent monomers colored by delta pTM score; the upper right panel shows a multimer colored by delta pTM and delta ipTM score; and the lower rigth panel shows a multimer and one of its mononumeric sequences colored by pLDDT. Additionally, figure 4b shows cross correlations among amino acids and delta scores for two of the protein complexes in figure 4a.



(a) PyMOL protein visualizations of general dimers

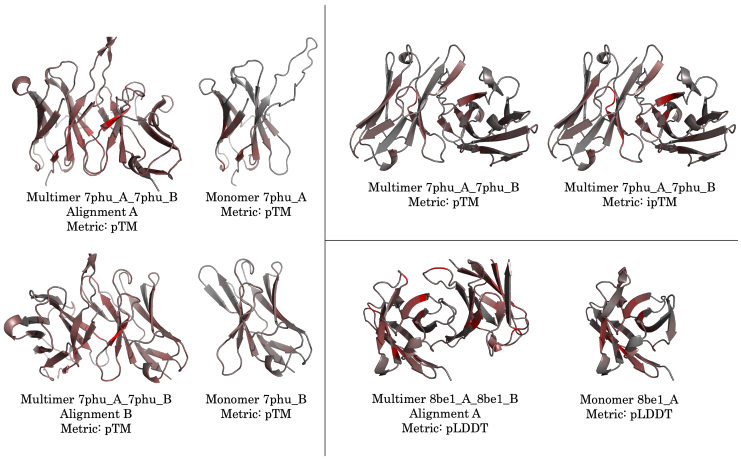


(b) Heatmaps of delta scores

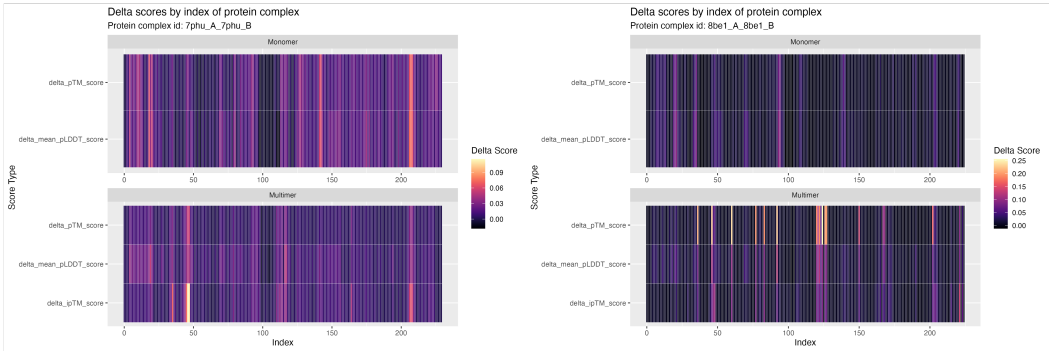
Figure 4: Visualization of delta scores for general proteins

In general, we found that amino acids near the multimer binding interface tend to have a larger delta pTM score when comparing to the same amino acids in monomers alone. This indicates that AlphaFold-Multimer places more importance on amino acids that are responsible for binding when it forms its predictions. We also found that delta ipTM score is usually as large, if not larger than pTM score per amino acid. This observation makes sense, since we expect AlphaFold-Multimer to place more value on a score that is designed to be more sensitive to changes along the protein-protein interface. Finally, we also found that delta pLDDT score can be less correlated with other scores.

This is particularly true for multimers, possibly due to an "averaging out" of the pLDDT score over the large number of amino acids.



(a) PyMOL protein visualizations of antigen-antibody pairs



(b) Heatmaps of delta scores

Figure 5: Visualization of delta scores for antigen-antibody

For antibody-antigen complexes, we notice similar trends as that of the general proteins. In figure 5 and over many of our dataset proteins, we notice that the sensitivity of all metrics tended to be smaller for antibody-antigen complexes, as indicated by the weaker delta score signals in the heat maps and on the molecule diagrams.

5 Conclusion

Through our deletion perturbation framework of explainability, we find some interesting results that give insight into the models AlphaFold and AlphaFold-Multimer as well as the broader problem of protein complex and antibody-antigen pair structure prediction. We find that AlphaFold-Multimer tends to be less sensitive to antibody-antigen inputs compared to regular proteins. We also find that pTM and ipTM are correlated but relatively robust metrics with strong signals that can be used for explainability. Upon inspection of heat maps and generated molecule images, we see that ipTM tends to provide a stronger signal that is more sensitive to input perturbations, particularly at the interface of a protein complex. With further research into the explainability of AlphaFold and AlphaFold-Multimer, the research community may be able to learn more about these impactful models and the world of protein structure prediction.

6 Reproducibility

See our github for code used: https://github.com/mying2002/Explainable-AlphaFold-Multimer/

References

[1] Minkyung Baek et al. "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* (2021), pp. 871–876. DOI: 10.1126/science.abj8754.

[2] Richard Evans et al. "Protein complex prediction with AlphaFold-Multimer". In: *bioRxiv* (2022). DOI: 10.1101/2021.10.04.463034.

[3] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* (2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

[4] W. R. Pearson and D. J. Lipman. "Improved tools for biological sequence comparison". In: *Proc. Natl. Acad. Sci. USA* (1988).

[5] Jeffrey A. Ruffolo et al. "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies". In: *Nature Communications* (2023), p. 2389. DOI: 10.1038/s41467-023-38063-x.

[6] Schrödinger, LLC. "The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8". Nov. 2015.

[7] Juntao Tan and Yongfeng Zhang. *ExplainableFold: Understanding AlphaFold Prediction with Explainable AI*. 2023. arXiv: 2301.11765 [cs.AI].

[8] Rui Yin and Brian G Pierce. "Evaluation of AlphaFold Antibody-Antigen Modeling with Implications for Improving Predictive Accuracy". In: *bioRxiv* (2023). DOI: doi:10.1101/2023.07.05.547832.

[9] Wensi Zhu et al. "Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes". In: *Bioinformatics* (2023). DOI: doi:10.1093/bioinformatics/btad424.