HIGH SCHOOL EDITION
Journal of Student Research

Volume 14 Issue 1 (2025)

Deep Learning Application in Protein Structure Understanding–FoxP Series of Proteins as an Example

Zixuan Wang

BASIS International Guangzhou, USA

ABSTRACT

This study explores the application of deep learning (DL) in understanding protein structures, focusing on the FoxP2 protein as an example. The integration of DL tools, such as AlphaFold2 (AF2) and ChatGPT, enables analysis of FoxP2's structure, accuracy, and function, with implications for its role in speech and language development. AF2's predictions were compared to experimental data, revealing strengths in identifying key regions like DNA-binding domains but limitations in accuracy and reliability in less structured areas. Additionally, ChatGPT demonstrated effectiveness in providing supplementary biological insights, such as the impact of mutations like ARG553 and potential clinical implications. The findings highlight the utility of DL in accelerating protein research while emphasizing the need for caution in interpreting predictions.

Introduction

With the scientific revolution, scientists began to highlight using scientific and mathematical methods to gain accurate data and predictions from various fields in the natural world. Hundreds of years later, we have accumulated a huge amount of data and developed a high computing power that enables machine learning, in other words, deep learning (DL). Over time, the availability of larger datasets has allowed models to capture intricate data patterns, and advances in computational hardware, especially Graphics Processing Units (GPUs) and Neural Network Processing Units (NPUs) have enabled the training of larger DL models. Eventually, these large trained models begin to contribute to various practical fields.

Years later, DL stepped into science research, revolutionizing some fields of science and setting a fundamental base for AI4Science, which has become a new popular field in scientific research, as protein structure for example. Recently, a research group[9] used DL for de novo redesign of protein scaffolds and active sites to create an enzyme called luciferases with specific catalytic activities and a significantly greater reaction rate, involving an approach termed ”family-wide hallucination.” This method generates numerous idealized protein structures with different pocket shapes and designed sequences encoding these structures. Artificial luciferases were designed from these scaffolds. The process involved generating large numbers of small, stable protein scaffolds with suitable binding pockets, using a DL-based approach integrated with Rosetta. This highlighted the importance of the protein structure to its function. In the 15th Critical Assessment of Protein Structure Prediction (CASP 15), the data set [1] points out that the accurate An accurate protein structure is crucial when analyzing the function of a protein.

Without deep learning (DL), the most essential way to do this is to analyze proteins physically and chemically. For instance, X-ray Crystallography is the most widely used method for determining protein structures. The protein is first crystallized, and then X-rays are directed at the crystal producing a diffraction pattern. By analyzing this pattern, scientists can produce a three-dimensional picture of the density of electrons within the crystal, which in turn provides information about the positions of the atoms. The primitive challenge in X-ray crystallography is the ”phase problem.”[4] When X-ray crystallography is used to determine protein structures. Each spot from the diffraction pattern has an intensity (how bright the spot is) and a phase (a value that represents the position of the wave peak relative to a reference point). While the intensities of the diffracted waves can be directly measured from the diffraction pattern, their phases sometimes cannot. This makes it difficult to calculate electron density maps from diffraction data, which are essential for determining the 3D structure of proteins. The most frequently used technique to derive electron density maps is molecular replacement. This method is used when the molecule under study is similar to another molecule with a known structure, calculating phases for all the structure factors.

A team of researchers focused on determining the crystal structure of the Nmd4 protein with 252 residues (UniProt code: YLR363C) by using the traditional method for almost two years. After collecting a 2.45Å resolution result, they got stuck with the crystal structure of the protein even though they tried many methods including molecular replacement[2]. However, with the help of the recent advancements in protein structure prediction, especially by Alphafold2(AF2), they rapidly determine the structure of Nmd4 by molecular replacement accurately by DL. DL uses algorithms to enhance the accuracy and efficiency of interpreting X-ray diffraction data, improving the ability to determine the structures of molecules. Therefore, AF2 usually takes approximately only a few hours to predict a protein's structure with 252 residues[5]. AF2 self-prediction of the predicted protein structure is also pretty high.

DL has made a big accomplishment in AI4Science, though there are still some limitations. As a high school student, DL can also help the student who wants to study advanced subjects but cannot get research results from the laboratory, for instance, protein structure, let's take the FoxP series protein as an example.

Method

In this study, we choose AF2(no system message in all conversations) to study and analyze a concrete protein Foxp2 example. Figures showing the structures of the protein Foxp2 were captured using a snipping tool in the 3D structure viewer in RCSB Protein Data Bank (RCSB PDB)/AF2 (called 3D canvas in RCSB PDB) with water and ion component of the protein hidden by toggling the control panel on the left-hand side of the main panel on bot RCSB PDB. For better quality, we used Photoshop to adjust the contrast, brightness, and saturation of the figure. Some regions of the figures are highlighted/boxed by Preview on Mac since they are important structures and regions such as variant points, DNA binding regions, etc.

We also choose ChatGPT-Plus with the GPT-4 model (no system message in all conversations), using zero-shot promotion for all the questions. We want to test the original behavior without any promote engineering since this will affect the results significantly[10].

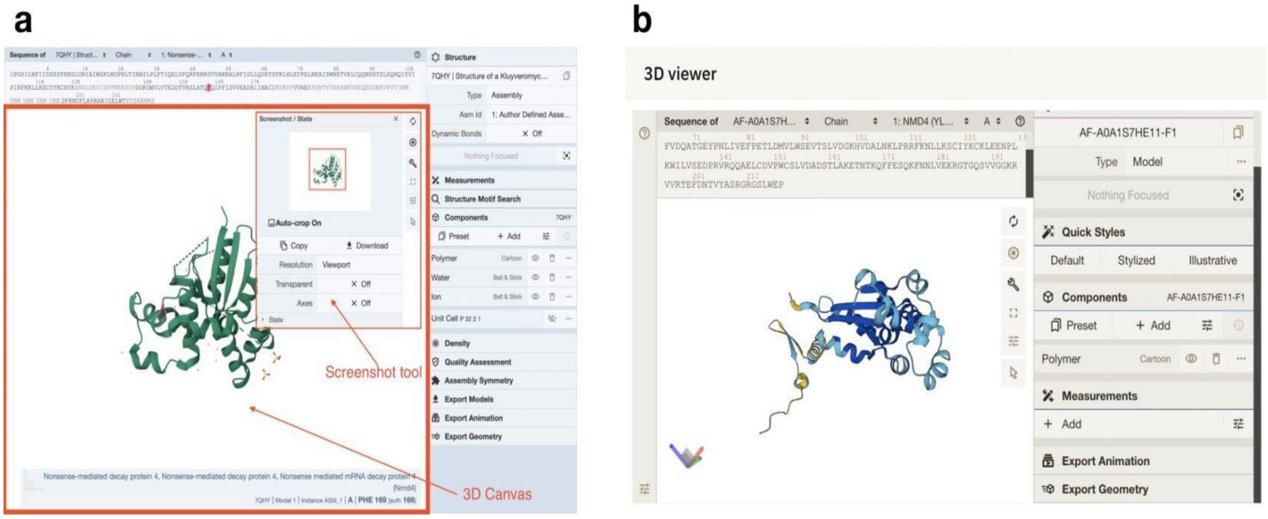


Figure 2.1. The main panel of the protein structure viewer in RCSB PDB(a), with frequently used tools

(screenshot and 3D canvas) boxed, and the main panel of the protein structure viewer in AF2(b).

Results

Structure Comparison and Exploration

The structure of Foxp2(Figure 3.1(a)) with DNA bound from the RCSB PDB is derived from an X-ray diffraction experiment from a research team[8]. The structure of Foxp2 is predicted by AF2(Figure 3.1(b)). Foxp2 is a protein with 715 residues, which takes approximately a few hours[5]. Although it is really fast for the model to predict the structure, let's consider the certainty. Although it is really fast for the model to predict, let's consider the uncertainty. DeepMind had developed a self-estimate system rating a confidence score between 0 and 100(per-residue model confidence score, also known as pLDDT), the bluer and darker the region, the more accurate it is, oppositely, if it is yellow or orange (Figure 3.1(c)). The predicted aligned error (PAE) diagram also shows the model confidence. For example, in Figure 3.2, residues from roughly 500-600 (boxed in red) are dark green, representing fewer errors. We can also see that the PAE diagram is symmetrical, meaning that this protein doesn't have too much beta-sheet structure. They are more uncertain with aligning and orientating in beta-sheet, causing the relative accuracy between residues to vary. Traditional self-estimating systems, which are based on a global superposition of carbon atoms, can be heavily influenced by domain movements and do not accurately assess the precision of local atomic details in the model. plDDT is a superposition-free score, meaning it doesn't rely on aligning the structures on top of each other to compare them. This makes it particularly more accurate for comparing structures with domain movements[6].

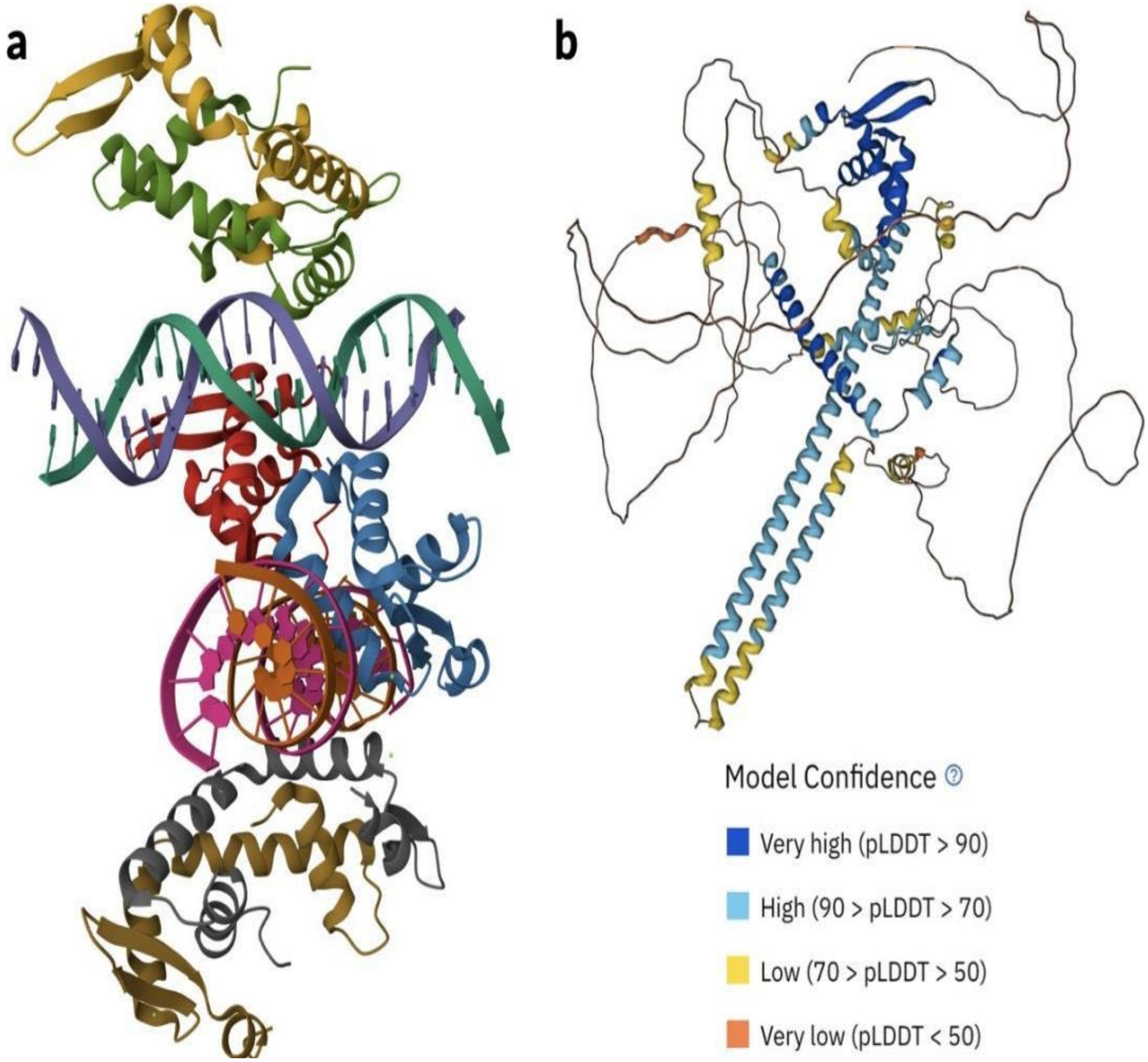


Figure 3.1. Global structure of FoxP2 obtained from (a) X-ray diffraction experiments (PDB ID: 2A07) colored by the rainbow color scheme (b) AlphaFold2 predicted structures (UniProt ID: AF-O15409-F1) colored by model confidence shown in the panel (b).

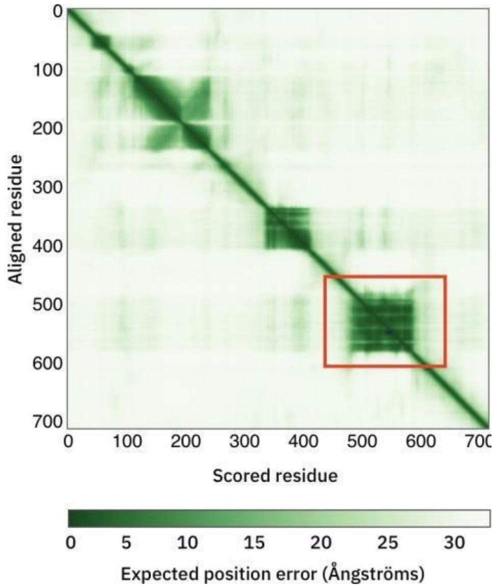


Figure 3.2. This figure shows the predicted aligned error (PAE) of Foxp2 on AF2. The color at coordinates (x, y) represents the predicted position error at residue x if the predicted and true structures were aligned on residue y. The darker regions correspond to good positions, whereas the light green regions indicate poor prediction.

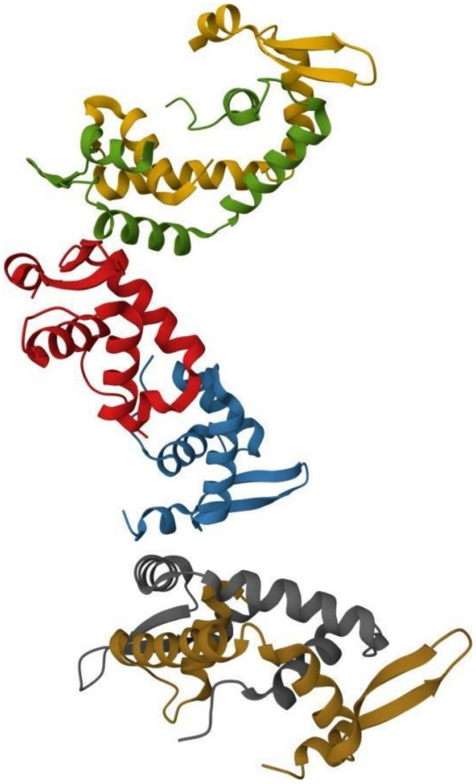


Figure 3.3. Global structure of FoxP2 obtained from (a) X-ray diffraction experiments(PDB ID: 2A07) colored by rainbow color scheme with DNA/RNA hidden.

However, when we compare FoxP2's structure with hidden DNA in RCSB PDB (Figure 3.3) to structure 4 in AF2, it should be very similar to each other, but they are quite different. So what causes this? First of all, the protein sequence inputted to the AF2 is not bound to the DNA/RNA, thus, the physical and chemical properties of the protein are different, such as the hydrogen bond. In some cases, the hydrogen bond interaction with the DNA/RNA is even greater than the reaction with the protein. For example, in Figure 3.4, the genome variant position formed a hydrogen bond with the DNA, and the distance is 2.99Å, it is shorter compared to the hydrogen bond to the protein structure, which distance is 3.46Å. A shorter distance means they are stronger interaction, which is important when determining the structure. This can explain biologically why they are quite different. Secondly, considering AF2's algorithm – multiple sequence alignments (MSA), during MSA, AF2 will find similar protein sequences in the training pool, sometimes it is possible that it cannot find a corresponding similar protein structure in the pool to predict the structure, causing lots of region remained unfolded. This is another possible explanation. There's a long alpha-helix structure in AF2's prediction, which is also blue, meaning high confidence. However, we can't see such a structure in PDB RCSB. This might be caused by the biological property mentioned before, and the property of the DL. AF2 will say this structure is accurate because it found a corresponding similar structure during the MSA, but it is not. So it is hard for researchers to choose reliable regions. For example, the unfolded structure region of Foxp2 is mostly orange and yellow, which is not predicted to be accurate (figure 3.4(b)). This is very obvious that this is not a reliable region. In Figure 3.4(a), it is an alpha helix structure and it is predicted as very accurate, but researchers can't know the exact accuracy of the prediction. The predicted accuracy is not objective. So this reflects the limits of AF2.

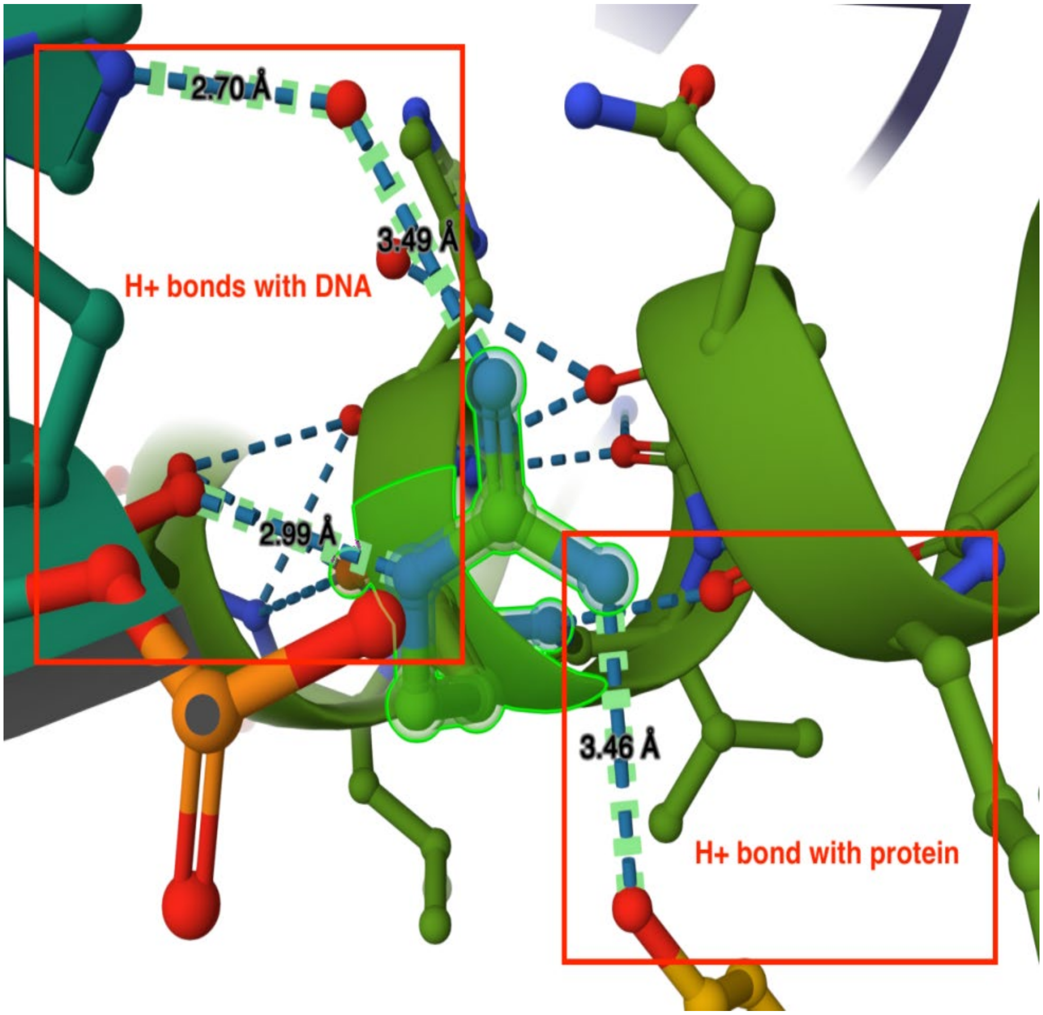


Figure 3.4. The local view of protein FOXP2's genome variant position: ARG553 from RCSB PDB visualized by ball and stick representation. The hydrogen bonding region is boxed and indicates the components that the ARG553 formed hydrogen bonds with. The length of each hydrogen bond is also indicated.

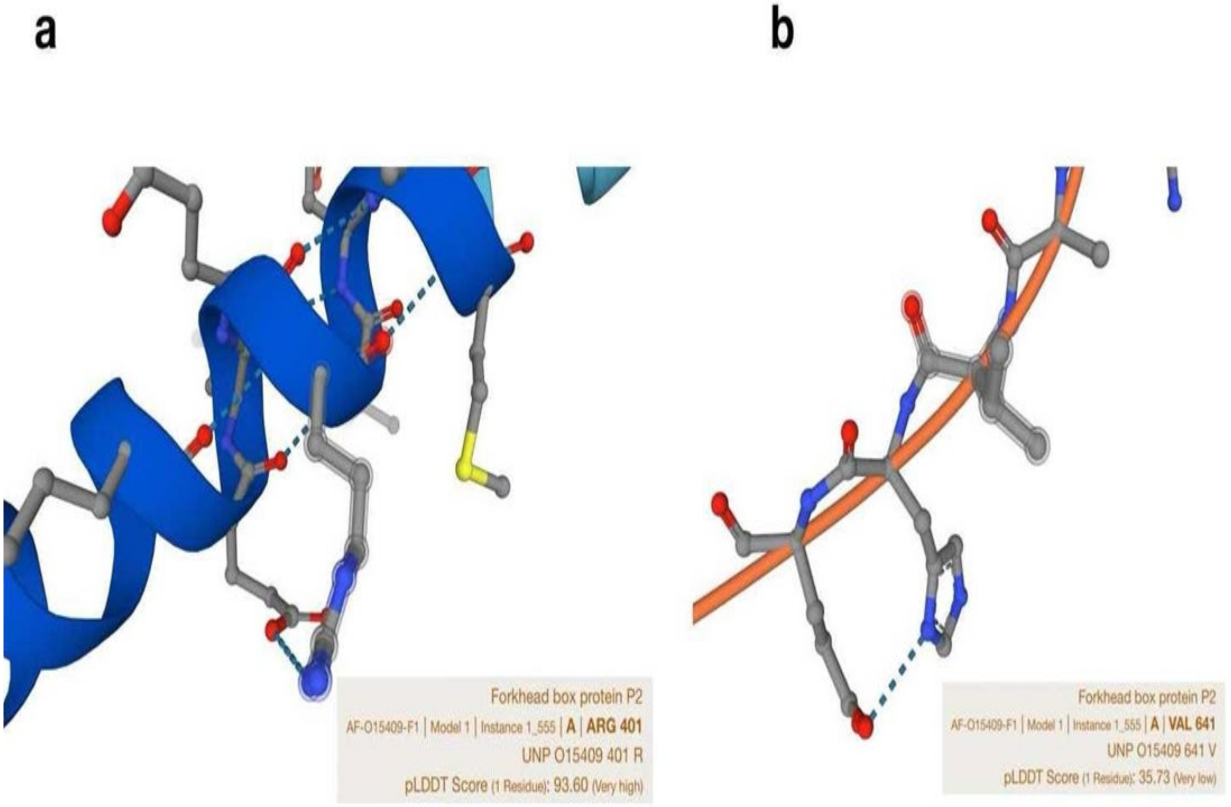

a
b
Forkhead box protein P2
AF-O15409-F1 | Model 1 | Instance 1_555 | A | ARG 401
UNP O15409 401 R
pLDDT Score (1 Residue): 93.60 (Very high)
Forkhead box protein P2
AF-O15409-F1 | Model 1 | Instance 1_555 | A | VAL 641
UNP O15409 641 V
pLDDT Score (1 Residue): 35.73 (Very low)

Figure 3.5. Local structures of FoxP2 from AF2 predicted structures (UniProt ID: AF-O15409-F1) colored by model confidence, showing the residue ARG401(a) and VAL641(b). ARG401's pLDDT score is 93.6, and VAL641's pLDDT score is 36.73.

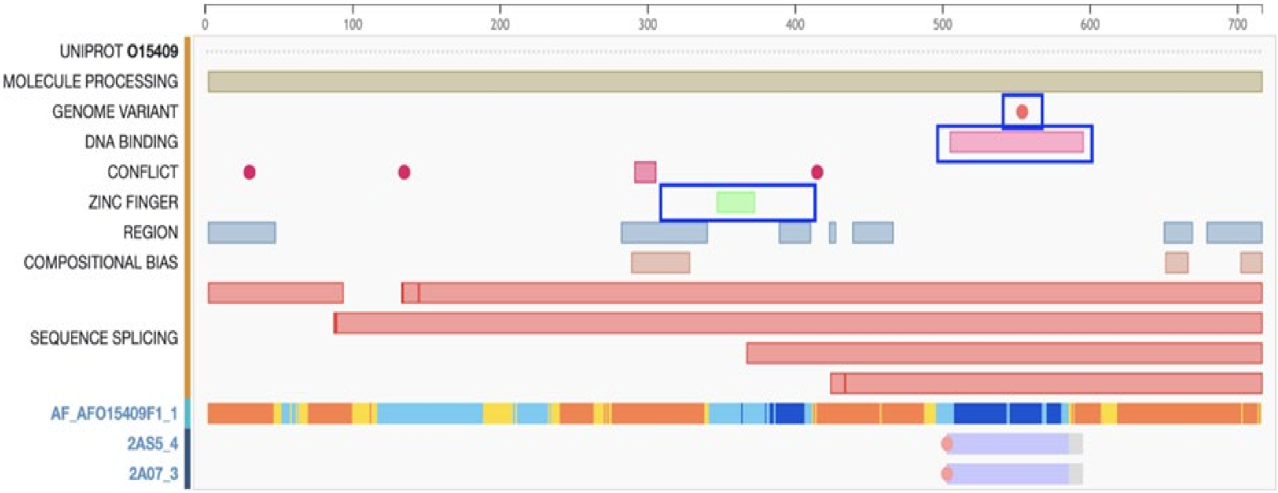


Figure 3.6. Sequence alignment on RCSB PDB of protein Foxp2 predicted structure by AF2(UniProt ID: AF-O15409-

F1) with important areas boxed.

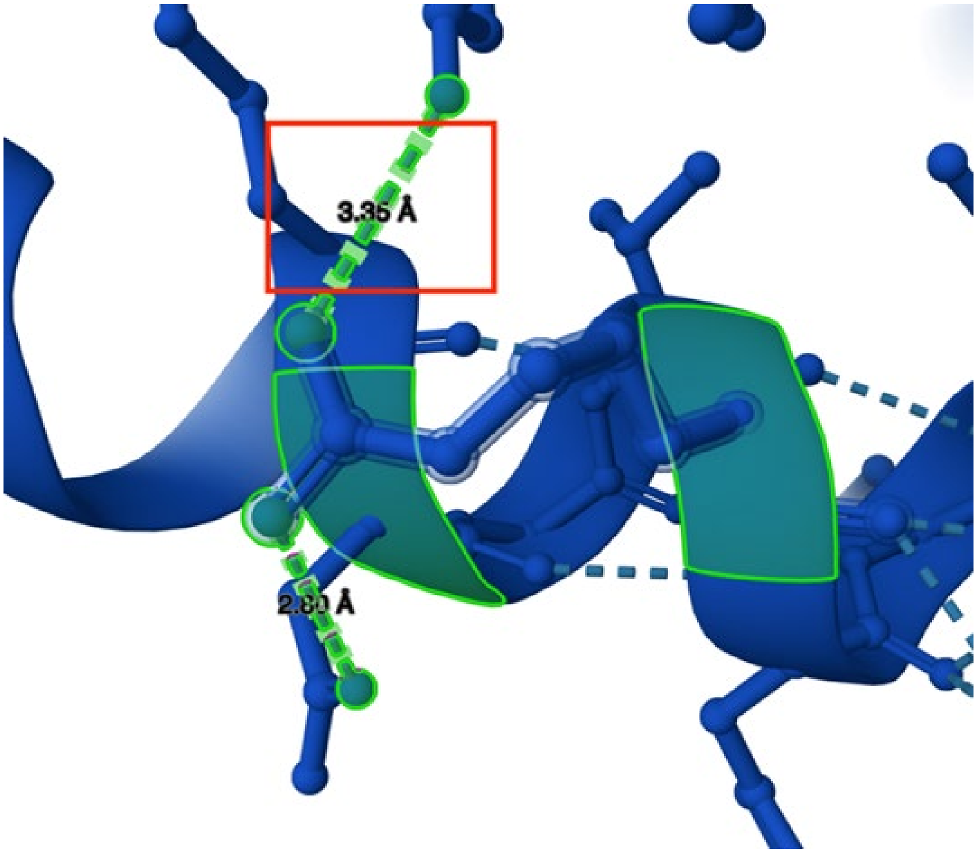
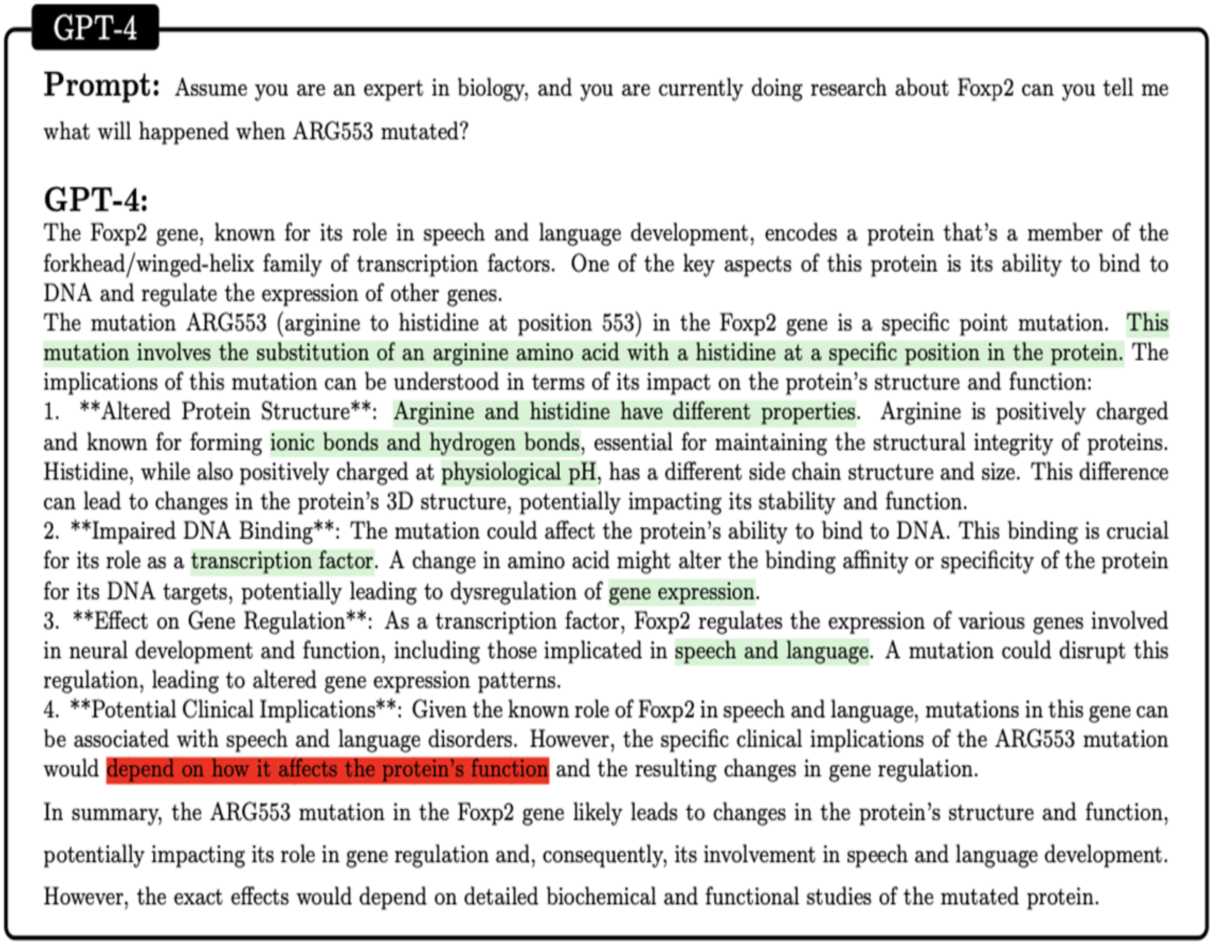


Figure 3.7. The local view of protein FOXP2's nome variant position: ARG553 predicted by AF2 (UniProt ID: AF-O15409-F1) and represented in ball and stick, with distance labeled on two hydrogen bonds. The boxed hydrogen bond is the same as the hydrogen bond bonds with the protein in Figure 3.4

The sequence alignment diagram(figure 3.6) indicates the sequences and regions that are involved in major characteristics. For example, the DNA binding region is from 504 to 594, and the genome variant site is 553. The majority of important areas are self-predicted as accurate, for example, the DNA binding region, shows a very high predicted accuracy. By comparing the genome variant site (ARG553) visually (Figure 3.7 and Figure 3.4), we can see their structures are roughly the same and the predicted error is very low, since it is dark blue, suggesting AF2 is very confident in the prediction of this variant site. Also, the length of the boxed hydrogen bond of the site is 3.36Å is not very different. Compared to the same hydrogen bond in RCSB PBD which is 3.46Å. This might not caused by inaccuracy in DL, since it lacked the interaction with the DNA. This indicates that this might help people to identify diseases related to this site without doing an actual experiment by only observing the surrounding environment of this site and the interaction with hydrogen bonds. The identity of the genome variant is the same as the lab results on RCSB PDB. The genome variant can mutate from codon ARG to codon HIS caused by a missense mutation. The genome variant site is bonded to the DNA. Therefore, the mutation of ARG553 disrupts the hydrogen bonding between the DNA, since the ARG codon will not interact with the DNA same as the HIS codon. Making Foxp2 dysfunctional. This will cause diseases related to speech, like Childhood apraxia of speech and speech-language disorder 1[8]. However, according to Ensembl Release, the Highest population (MAF) is less than 0.01%, meaning that this variant occurs in less than 1% of the population, Some are not very important regions compared to the DNA binding region, for example, the zinc finger. AF2 also predicts the structure of it. It often appears as a metal-binding domain in multi-domain proteins[3]. This region in AF2 shows majorly light blue(Figure 3.6), meaning acceptable accuracy. It can be a reliable structure to analyze the protein such as altering the DNA-binding specificity of zinc fingers[7], since the experiment results in RCSB PDB do not contain that region.

Interpretation and Information Collection Using ChatGPT-Plus (powered by GPT-4)

After the exploration of AF2, we see the advantages and limitations of AF2 in predicting the protein structure. However, it can't provide more information about the nature and function of this protein. ChatGPT, a large language model, could be helpful for us to collect additional information in medical applications. In this case(Figure 3.8), we ask for some information about the mutation of FoxP2 protein at the genome variant point (Arg553). This tests GPT-4's skill in gathering information and integrating. It did very well on them. It points out the causes like the changes in structure and pH and the effect(highlighted in green). However, in Potential Clinical Implications, it says it depends on how it affects the protein's function(highlighted in red). This response seems hasty, as we expected it could try to infer the specific effects and give us some examples, but it didn't mention the specific name of the disease.



GPT-4

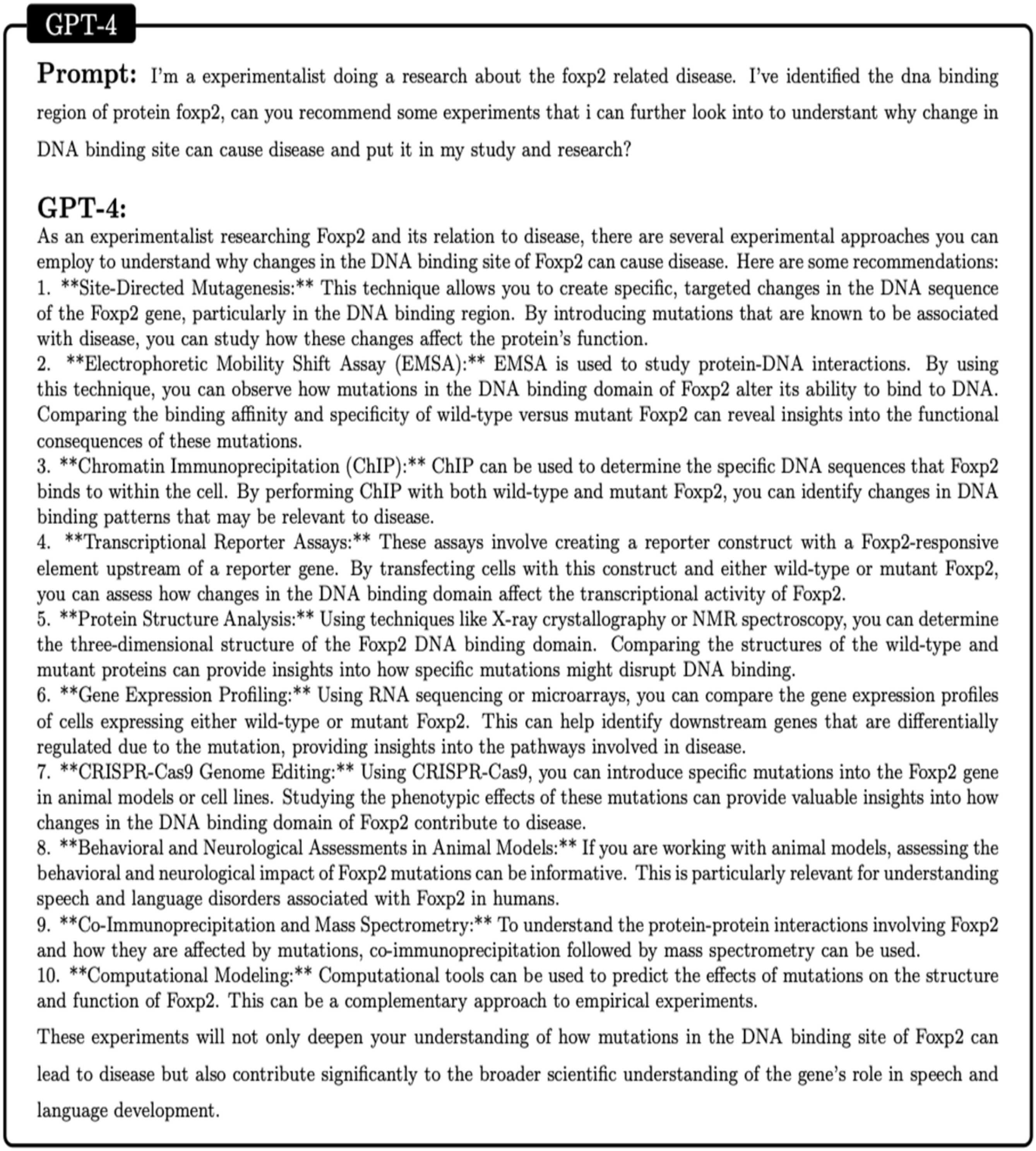
Prompt: Assume you are an expert in biology, and you are currently doing research about Foxp2 can you tell me what will happened when ARG553 mutated?

GPT-4:
The Foxp2 gene, known for its role in speech and language development, encodes a protein that's a member of the forkhead/winged-helix family of transcription factors. One of the key aspects of this protein is its ability to bind to DNA and regulate the expression of other genes.
The mutation ARG553 (arginine to histidine at position 553) in the Foxp2 gene is a specific point mutation. This mutation involves the substitution of an arginine amino acid with a histidine at a specific position in the protein. The implications of this mutation can be understood in terms of its impact on the protein's structure and function:
1. **Altered Protein Structure**: Arginine and histidine have different properties. Arginine is positively charged and known for forming ionic bonds and hydrogen bonds, essential for maintaining the structural integrity of proteins. Histidine, while also positively charged at physiological pH, has a different side chain structure and size. This difference can lead to changes in the protein's 3D structure, potentially impacting its stability and function.
2. **Impaired DNA Binding**: The mutation could affect the protein's ability to bind to DNA. This binding is crucial for its role as a transcription factor. A change in amino acid might alter the binding affinity or specificity of the protein for its DNA targets, potentially leading to dysregulation of gene expression.
3. **Effect on Gene Regulation**: As a transcription factor, Foxp2 regulates the expression of various genes involved in neural development and function, including those implicated in speech and language. A mutation could disrupt this regulation, leading to altered gene expression patterns.
4. **Potential Clinical Implications**: Given the known role of Foxp2 in speech and language, mutations in this gene can be associated with speech and language disorders. However, the specific clinical implications of the ARG553 mutation would depend on how it affects the protein's function and the resulting changes in gene regulation.

In summary, the ARG553 mutation in the Foxp2 gene likely leads to changes in the protein's structure and function, potentially impacting its role in gene regulation and, consequently, its involvement in speech and language development. However, the exact effects would depend on detailed biochemical and functional studies of the mutated protein.

Figure 3.8. Asking GPT-4 for information about the mutation of Foxp2 protein.

In the second case(Figure 3.9), we ask GPT-4 an open question, testing its ability of logical reasoning, inferring, and designing skills in a practical and specific example. GPT-4's response is unexpected to us. We didn't expect that it would give us an apropos response by zero-shot based on our experience. The response introduces 10 relevant experiences from multiple aspects like gene mutation and protein-protein interaction. However, we can’t know the feasibility of these experiments in this specific case from its response. We might need to do more academic research since it just speaking generally in the second response(Figure 3.10).

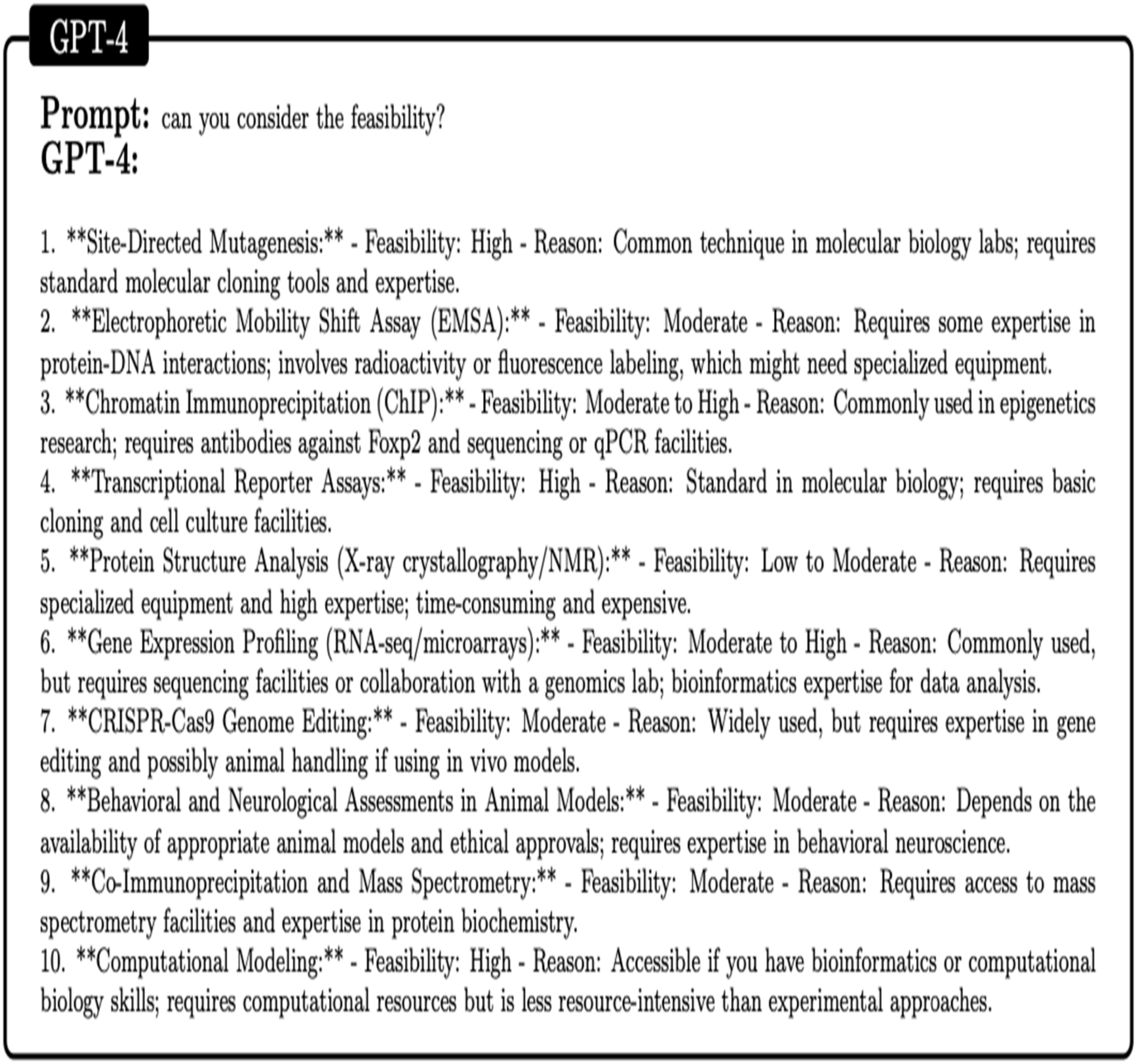


Prompt: I’m a experimentalist doing a research about the foxp2 related disease. I’ve identified the dna binding region of protein foxp2, can you recommend some experiments that i can further look into to understant why change in DNA binding site can cause disease and put it in my study and research?

GPT-4:
As an experimentalist researching Foxp2 and its relation to disease, there are several experimental approaches you can employ to understand why changes in the DNA binding site of Foxp2 can cause disease. Here are some recommendations:
1. **Site-Directed Mutagenesis:** This technique allows you to create specific, targeted changes in the DNA sequence of the Foxp2 gene, particularly in the DNA binding region. By introducing mutations that are known to be associated with disease, you can study how these changes affect the protein’s function.
2. **Electrophoretic Mobility Shift Assay (EMSA):** EMSA is used to study protein-DNA interactions. By using this technique, you can observe how mutations in the DNA binding domain of Foxp2 alter its ability to bind to DNA. Comparing the binding affinity and specificity of wild-type versus mutant Foxp2 can reveal insights into the functional consequences of these mutations.
3. **Chromatin Immunoprecipitation (ChIP):** ChIP can be used to determine the specific DNA sequences that Foxp2 binds to within the cell. By performing ChIP with both wild-type and mutant Foxp2, you can identify changes in DNA binding patterns that may be relevant to disease.
4. **Transcriptional Reporter Assays:** These assays involve creating a reporter construct with a Foxp2-responsive element upstream of a reporter gene. By transfecting cells with this construct and either wild-type or mutant Foxp2, you can assess how changes in the DNA binding domain affect the transcriptional activity of Foxp2.
5. **Protein Structure Analysis:** Using techniques like X-ray crystallography or NMR spectroscopy, you can determine the three-dimensional structure of the Foxp2 DNA binding domain. Comparing the structures of the wild-type and mutant proteins can provide insights into how specific mutations might disrupt DNA binding.
6. **Gene Expression Profiling:** Using RNA sequencing or microarrays, you can compare the gene expression profiles of cells expressing either wild-type or mutant Foxp2. This can help identify downstream genes that are differentially regulated due to the mutation, providing insights into the pathways involved in disease.
7. **CRISPR-Cas9 Genome Editing:** Using CRISPR-Cas9, you can introduce specific mutations into the Foxp2 gene in animal models or cell lines. Studying the phenotypic effects of these mutations can provide valuable insights into how changes in the DNA binding domain of Foxp2 contribute to disease.
8. **Behavioral and Neurological Assessments in Animal Models:** If you are working with animal models, assessing the behavioral and neurological impact of Foxp2 mutations can be informative. This is particularly relevant for understanding speech and language disorders associated with Foxp2 in humans.
9. **Co-Immunoprecipitation and Mass Spectrometry:** To understand the protein-protein interactions involving Foxp2 and how they are affected by mutations, co-immunoprecipitation followed by mass spectrometry can be used.
10. **Computational Modeling:** Computational tools can be used to predict the effects of mutations on the structure and function of Foxp2. This can be a complementary approach to empirical experiments.

These experiments will not only deepen your understanding of how mutations in the DNA binding site of Foxp2 can lead to disease but also contribute significantly to the broader scientific understanding of the gene’s role in speech and language development.

Figure 3.9. Asking GPT-4 for further information about experiments relating to the DNA binding site.



GPT-4

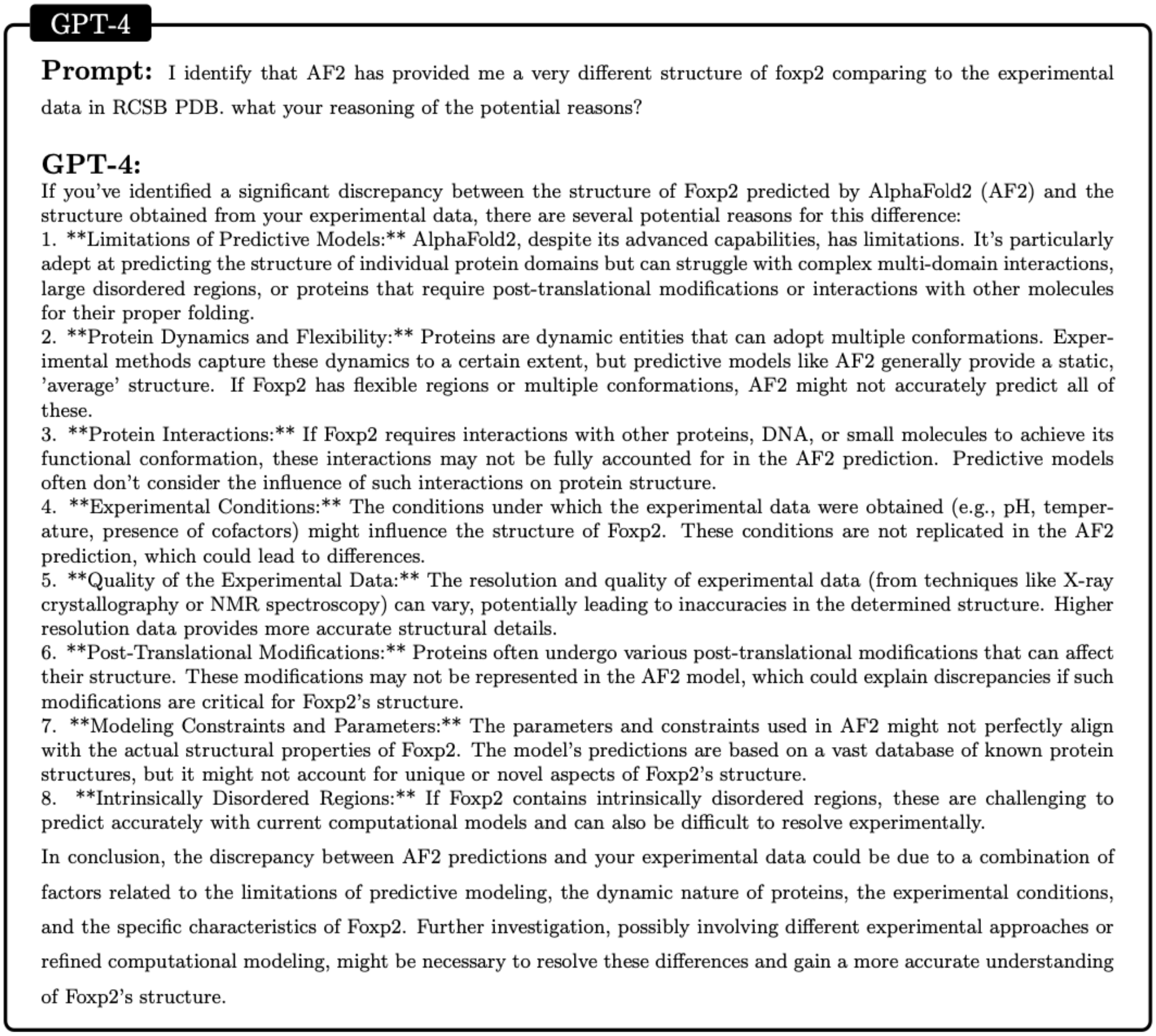
Prompt: can you consider the feasibility?

GPT-4:

1. **Site-Directed Mutagenesis:** - Feasibility: High - Reason: Common technique in molecular biology labs; requires standard molecular cloning tools and expertise.
2. **Electrophoretic Mobility Shift Assay (EMSA):** - Feasibility: Moderate - Reason: Requires some expertise in protein-DNA interactions; involves radioactivity or fluorescence labeling, which might need specialized equipment.
3. **Chromatin Immunoprecipitation (ChIP):** - Feasibility: Moderate to High - Reason: Commonly used in epigenetics research; requires antibodies against Foxp2 and sequencing or qPCR facilities.
4. **Transcriptional Reporter Assays:** - Feasibility: High - Reason: Standard in molecular biology; requires basic cloning and cell culture facilities.
5. **Protein Structure Analysis (X-ray crystallography/NMR):** - Feasibility: Low to Moderate - Reason: Requires specialized equipment and high expertise; time-consuming and expensive.
6. **Gene Expression Profiling (RNA-seq/microarrays):** - Feasibility: Moderate to High - Reason: Commonly used, but requires sequencing facilities or collaboration with a genomics lab; bioinformatics expertise for data analysis.
7. **CRISPR-Cas9 Genome Editing:** - Feasibility: Moderate - Reason: Widely used, but requires expertise in gene editing and possibly animal handling if using in vivo models.
8. **Behavioral and Neurological Assessments in Animal Models:** - Feasibility: Moderate - Reason: Depends on the availability of appropriate animal models and ethical approvals; requires expertise in behavioral neuroscience.
9. **Co-Immunoprecipitation and Mass Spectrometry:** - Feasibility: Moderate - Reason: Requires access to mass spectrometry facilities and expertise in protein biochemistry.
10. **Computational Modeling:** - Feasibility: High - Reason: Accessible if you have bioinformatics or computational biology skills; requires computational resources but is less resource-intensive than experimental approaches.

Figure 3.10. Continue asking GPT-4 about the feasibility of the experiments.

In the third case(Figure 3.11), we ask a more open-ended question to Chat-GPT. GPT-4 points out the weaknesses discussed widely in AF2. However, some points it mentioned are not very pertinence to AF2's algorithms and behavior.



GPT-4

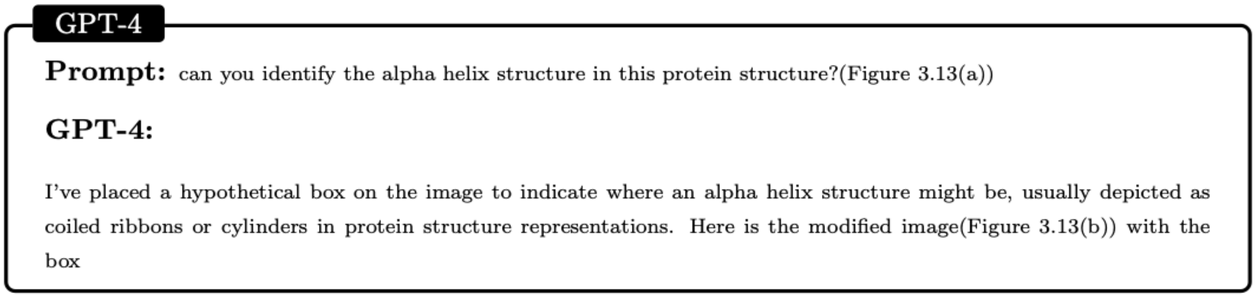
Prompt: I identify that AF2 has provided me a very different structure of foxp2 comparing to the experimental data in RCSB PDB. what your reasoning of the potential reasons?

GPT-4:
If you've identified a significant discrepancy between the structure of Foxp2 predicted by AlphaFold2 (AF2) and the structure obtained from your experimental data, there are several potential reasons for this difference:
1. **Limitations of Predictive Models:** AlphaFold2, despite its advanced capabilities, has limitations. It's particularly adept at predicting the structure of individual protein domains but can struggle with complex multi-domain interactions, large disordered regions, or proteins that require post-translational modifications or interactions with other molecules for their proper folding.
2. **Protein Dynamics and Flexibility:** Proteins are dynamic entities that can adopt multiple conformations. Experimental methods capture these dynamics to a certain extent, but predictive models like AF2 generally provide a static, 'average' structure. If Foxp2 has flexible regions or multiple conformations, AF2 might not accurately predict all of these.
3. **Protein Interactions:** If Foxp2 requires interactions with other proteins, DNA, or small molecules to achieve its functional conformation, these interactions may not be fully accounted for in the AF2 prediction. Predictive models often don't consider the influence of such interactions on protein structure.
4. **Experimental Conditions:** The conditions under which the experimental data were obtained (e.g., pH, temperature, presence of cofactors) might influence the structure of Foxp2. These conditions are not replicated in the AF2 prediction, which could lead to differences.
5. **Quality of the Experimental Data:** The resolution and quality of experimental data (from techniques like X-ray crystallography or NMR spectroscopy) can vary, potentially leading to inaccuracies in the determined structure. Higher resolution data provides more accurate structural details.
6. **Post-Translational Modifications:** Proteins often undergo various post-translational modifications that can affect their structure. These modifications may not be represented in the AF2 model, which could explain discrepancies if such modifications are critical for Foxp2's structure.
7. **Modeling Constraints and Parameters:** The parameters and constraints used in AF2 might not perfectly align with the actual structural properties of Foxp2. The model's predictions are based on a vast database of known protein structures, but it might not account for unique or novel aspects of Foxp2's structure.
8. **Intrinsically Disordered Regions:** If Foxp2 contains intrinsically disordered regions, these are challenging to predict accurately with current computational models and can also be difficult to resolve experimentally.

In conclusion, the discrepancy between AF2 predictions and your experimental data could be due to a combination of factors related to the limitations of predictive modeling, the dynamic nature of proteins, the experimental conditions, and the specific characteristics of Foxp2. Further investigation, possibly involving different experimental approaches or refined computational modeling, might be necessary to resolve these differences and gain a more accurate understanding of Foxp2's structure.

Figure 3.11. Asking GPT-4 information about reasons why the result from AF2 is different from the experimental result.

In the fourth case(Figure 3.12), we want GPT-4 to identify alpha-helix and beta-sheet structures on the uploaded protein structure, testing its image-identifying capability. It gives a correct answer with a clear and specific observation of the structure.



GPT-4

Prompt: can you identify the alpha helix structure in this protein structure?(Figure 3.13(a))

GPT-4:

I've placed a hypothetical box on the image to indicate where an alpha helix structure might be, usually depicted as coiled ribbons or cylinders in protein structure representations. Here is the modified image(Figure 3.13(b)) with the box

Figure 3.12. Asking GPT-4 to identify the Alpha-helix and beta-sheet structure in the image.

Conclusion

Based on the result, DL helped us to interpret the protein Foxp2 in various aspects including structure, accuracy, function, and future development of DL in biology. Overall, AF2's prediction is considerably poor, the majority of the structure is yellow, orange, and unfolded. However, most of the important regions discussed in the Result are predicted as accurate. This can probably help scientists to understand and analyze the pioneer concepts and function of a protein without deriving from the crystal structure. Despite the benefit, we need to recognize its disabilities such as being overconfident and falsely detecting in some regions, especially the alpha helix structure. On the other hand, ChatGPT, another fundamental model like AF2, shows good capabilities in interpreting information. ChatGPT's upper limit seems to have increased. The response is not as general and broad as before. It is a great tool to search and gather information that is already online, helping us to select useful and pertinent information. Therefore DL provides more capabilities to the scientific field, a tool assisting scientists on long-time time-consuming tasks, rather than a great threat to human scientists.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Mohammad AlOmari, Ahmad AlOmari, and Izzat Alsmadi. "CASP Dataset and Protein Structures Prediction". In: Available at SSRN 4315842 (2022).

[2] Barbarin-Bocahu, I., & Graille, M. (2022). The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report. *Acta crystallographica. Section D, Structural biology*, *78*(Pt 4), 517–531. https://doi.org/10.1107/S2059798322002157

[3] Berg, J. M. (1990). Zinc fingers and other metal-binding domains. J. biol. Chem., 265, 6513-6516.

[4] Cowtan, K. (2003). Phase problem in X-ray crystallography, and its solution. eLS.

[5]Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[6] Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics (Oxford, England), 29(21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

[7] Pabo, C. O., Peisach, E., & Grant, R. A. (2001). Design and selection of novel Cys2His2 zinc finger proteins. Annual review of biochemistry, 70(1), 313-340. https://doi.org/10.1146/annurev.biochem.70.1.313

[8] Stroud, J. C., Wu, Y., Bates, D. L., Han, A., Nowick, K., Paabo, S., Tong, H., & Chen, L. (2006). Structure of the forkhead domain of FOXP2 bound to DNA. Structure (London, England : 1993), 14(1), 159–166. https://doi.org/10.1016/j.str.2005.10.005

[9] Yeh, A. H. W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S. J., Evans, D., ... & Baker, D. (2023). De novo design of luciferases using deep learning. Nature, 614(7949), 774-780.

[10] Yong, G., Jeon, K., Gil, D., & Lee, G. (2023). Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Computer-Aided Civil and Infrastructure Engineering, 38(11), 1536-1554. https://doi.org/10.1111/mice.12954