**Prediction of structural changes from mutation in human AlphaB Crystallin**

Oriana Chen, Yujung Lin, Johnson Huang, Tina Ryu

# Introduction

AlphaB Crystallin is a small heat-shock protein that plays a critical role in maintaining the transparency and functionality of the human eye lens. This protein is not only fundamental for the lens's structural integrity but also in preventing the aggregation of unfolded or misfolded proteins, which is crucial for lens clarity throughout one’s life[1]. Mutations in AlphaB Crystallin can cause structural changes in its key domain and interfere with normal oligomerization, leading to aggregation or loss of chaperon functionality that are considered to be the main cause of age-related cataracts[2].

It is meaningful to develop effective methods to identify mutations that will cause structural changes in proteins like AlphaB Crystallin, as understanding how such mutations influence protein folding helps us uncover the broader mechanisms of cataracts and many other diseases related to protein misfolding. Computational prediction of protein structures is a perfect choice as we will be able to see whether a mutation changes protein structure without performing laboratory bioimaging. In this project, we designed a pipeline for performing structural prediction and analysis of mutated AlphaB Crystallin that can also be seamlessly applied to other proteins. This flexibility enhances the potential impact of our findings across various protein-related conditions.

The pipeline contains two python scripts, ESMFold.py and PDBanalysis.py. ESMFold.py performs mutations and structure prediction using ESMFold and the PDBanalysis.py performs structural analysis using the Biopython PDB module. In running ESMFold.py, the user will provide a protein sequence they wish to modify and introduce mutations into the sequence as input. The 3D structure of the mutated sequence will be generated from ESMFold prediction, which is a structural prediction tool that generates the protein structure based on the amino acid sequence of interest. After getting the predicted structures, users will have two options to perform structural analysis. One is to continue to the next section of this pipeline and compare two structures using PDBanalysis.py, which will produce the root mean square deviation (RMSD) as the measurement of the difference between the two protein structures. The other option is to visualize secondary protein structures using Pymol, a visualization tool for protein structure.

# Packages

## ESMFold

ESMFold is a transformer protein language model from the Meta Fundamental AI Research Protein Team (FAIR). The model allows users to perform structural prediction end to end directly from the sequence of their protein of interest.

The ESMfold module used in this project is get from the following github repository: <https://github.com/facebookresearch/esm/tree/main>

## Biopython PDB

Biopython is a set of free modules available for computational analysis of biological data in Python. Biopython includes Python libraries and applications that help for processing biological data files commonly encountered in bioinformatics, such as Fasta file and PDB structural file. Please refer to the following website for Biopython packages:

<https://biopython.org/wiki/Documentation>

In this pipeline specifically, Biopython is used for extracting structural information from PDB files and performing structural alignment for PDB objects.

Please refer to the following link for more detail about the package used:

https://biopython.org/wiki/The\_Biopython\_Structural\_Bioinformatics\_FAQ

<https://biopython.org/docs/1.75/api/Bio.PDB.Superimposer.html>

## DSSP

DSSP (Dictionary of Secondary Structure of Proteins) is a Python package under Biopython that helps in predicting protein secondary structure based on its three-dimensional atomic coordinates. The assignment is based on the H-bonding patterns indicated by the residue coordinates. The assignment is achieved through the *Define Secondary Structure of Proteins* Algorithm defined in the Pascal program in 1983, and the algorithm is currently used as the common method for determining protein secondary structure given a PDB file (such as in Pymol).

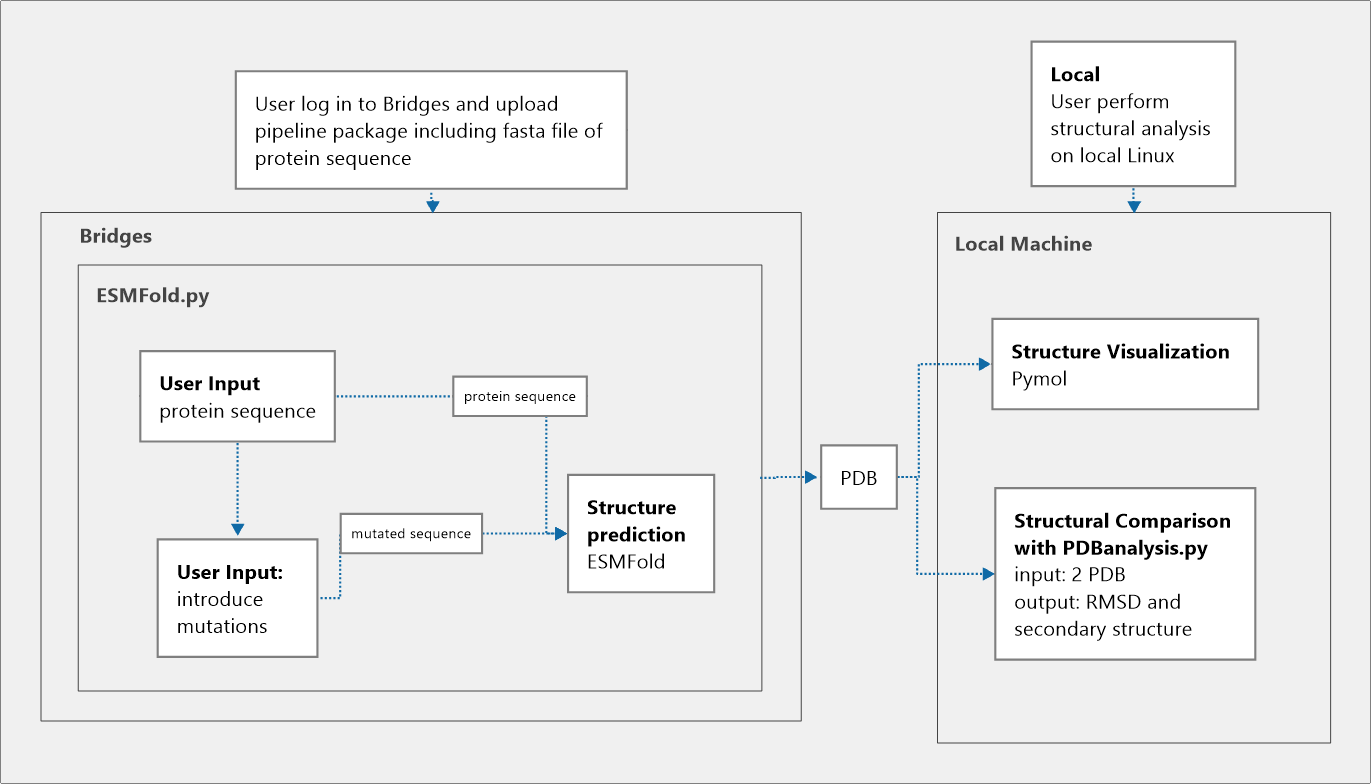
Please refer to the following link for the source code of the package:

<https://biopython.org/docs/1.75/api/Bio.PDB.DSSP.html>

<https://github.com/biopython/biopython/blob/master/Bio/PDB/DSSP.py>

# 

# Workflow



# Pipeline

The pipeline contains two separated scripts: one is responsible for introducing mutations to the target protein sequence and structure prediction using ESMFold. The other script is responsible for 3D structure comparison between two protein structures.

## Required input files

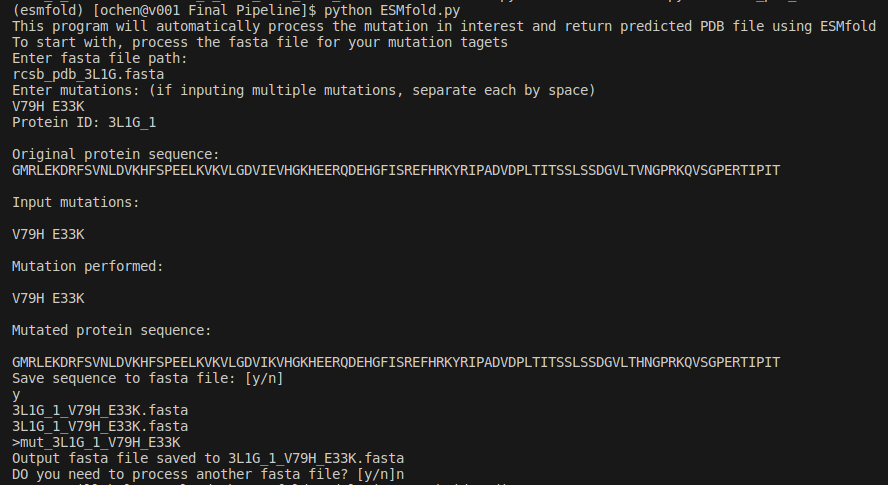
The required input files are different for the two sections. For the first section, the required input file is a file containing the protein sequence of interest in the fasta format. Fasta files containing protein sequences can be found on websites such as RSCB Protein Data Bank: <https://www.rcsb.org/> and UniProt: <https://www.uniprot.org/>.

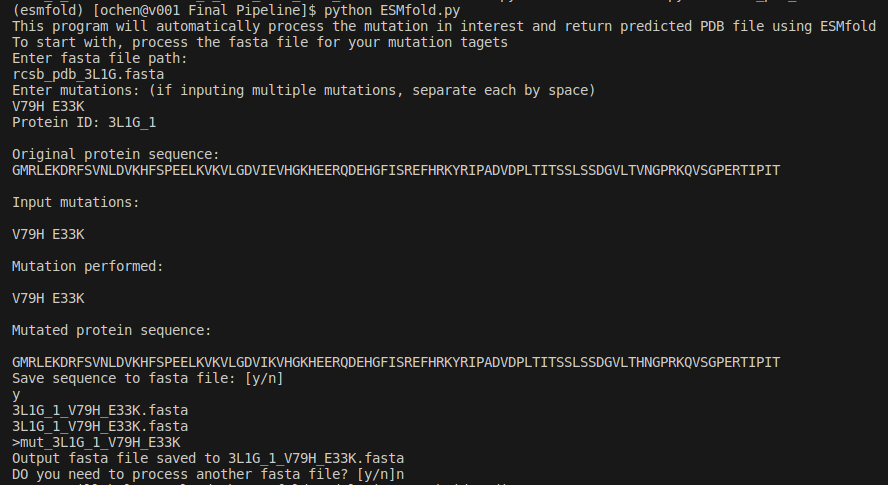
For the second section, the required input files are two PDB files containing the structure information of the proteins. PDB files used can be generated from the first section of the pipeline or can be found on websites such as RSCB Protein Data Bank: <https://www.rcsb.org/> and AlphaFold Protein Structure Database: <https://alphafold.ebi.ac.uk/>.

## ESMFold.py

The purpose of ESMFold.py is to allow users to introduce mutations to the protein sequence of interest and predict the structure of the proteins.

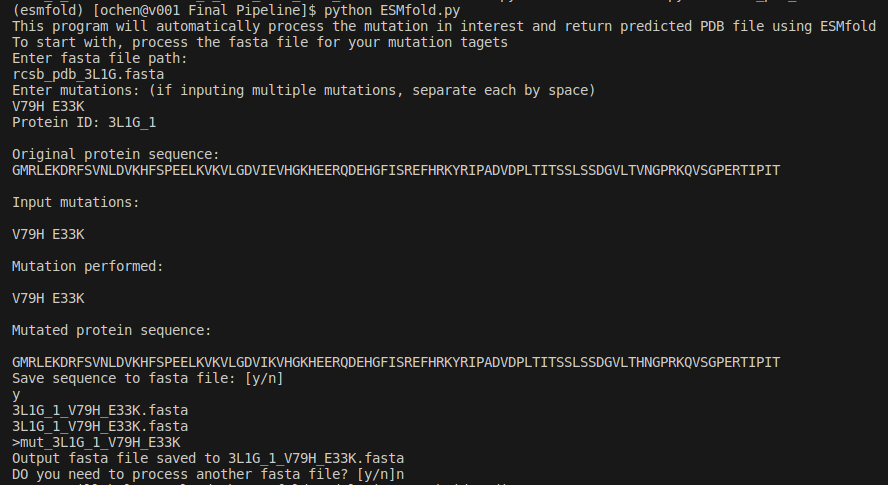
After initiating the script, users will be asked to input the path to the fasta file containing the target protein sequence. Complete path should be used if the target file is not under the same directory as the script.

After inputting the path, users will have the option to introduce one or multiple single point mutations to the sequence at the selected locations.



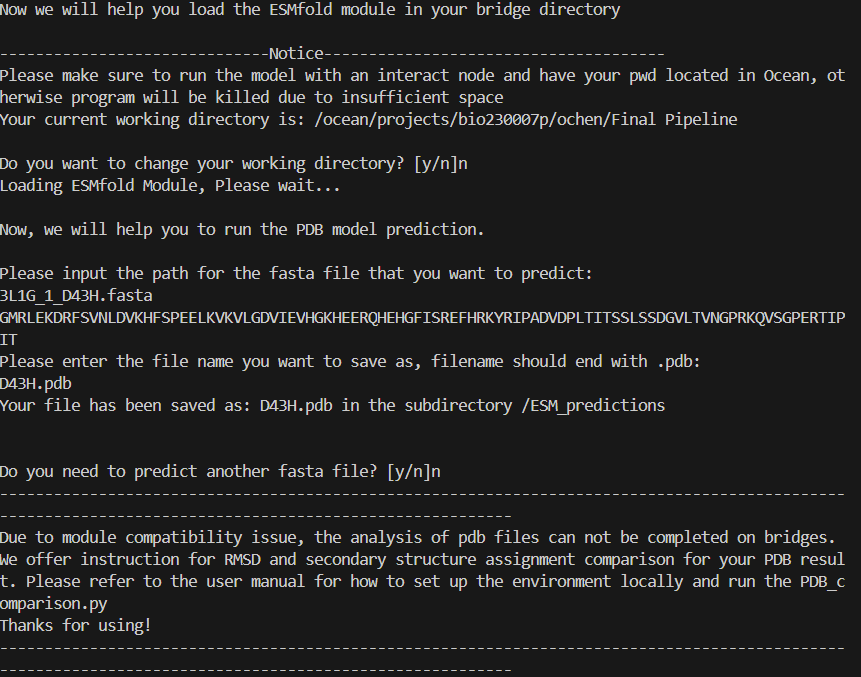
Only valid mutations will be processed and introduced to the sequence. Thus, if users input mutations that cannot be performed (i.e. locations outside of sequence length), such mutations will be skipped.

After the input mutations are processed, results will be printed to the terminal.



Then, users will be asked if they would like to save the results to a new fasta file. Saving the results is highly recommended since ESMFold requires a fasta file as the input.

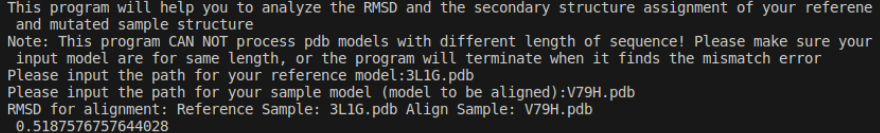
Before proceeding to the prediction, users will be asked to input the path to the fasta file they want to run the prediction on. After validating the path, the prediction will start and output a PDB file containing the structure information once it is done.



## PDBanalysis.py

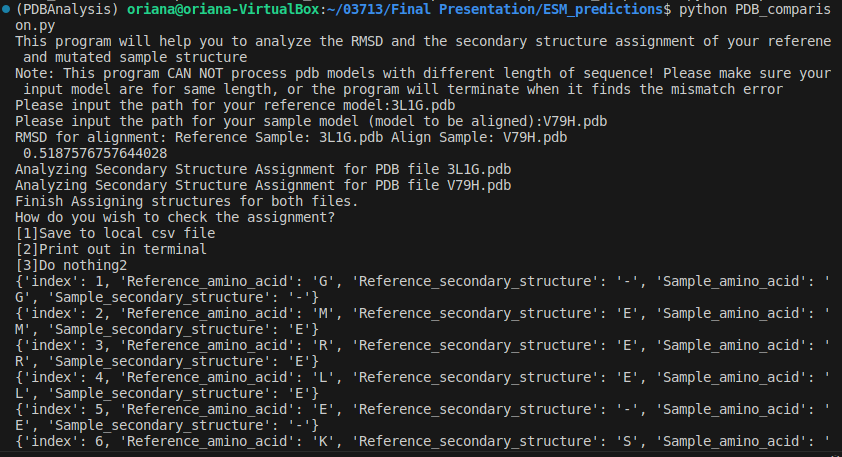
The purpose of PDBanalysis.py is to allow users to compare two protein structures. The comparison is done using the Biopython.PDB module and DSSP module. The script calculates the difference between the two input structures and quantifies the differences using root mean square deviation score (RMSD). The higher the RMSD score is, the more different the two structures are.

After initiating PDBanalysis.py, users will be asked to input two PDB files containing the structure information of the proteins they want to compare.



The RMSD score will be output to the terminal after the calculation is done.

At the same time, the pipeline will calculate the secondary structure assignment of the two input PDB files by using the coordinates of the atoms to calculate H-bonds patterns. Users will have the option to save the secondary structure assignment for each residue to a local csv file or have the results printed to the terminal for brief examination.



\*The notation for DSSP secondary structure assignment can be found in the user manual

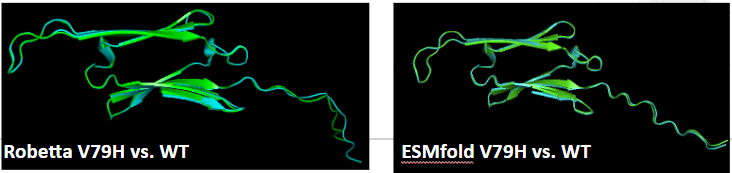
## Pymol

If users wish to visualize the predicted structure of the input protein, they can import the output PDB files into Pymol to observe the 3D structure. Pymol uses the same algorithm for assignment cartoons for the imported structure (DSSP), so the cartoon presented in Pymol should be consistent as the output of the pipeline.

# Results

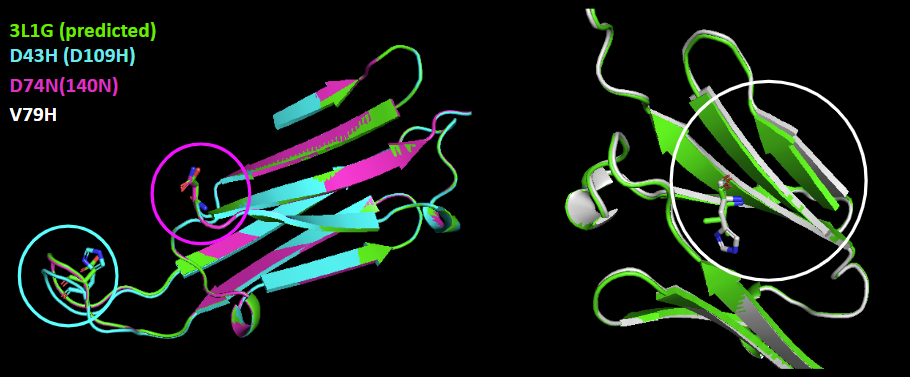
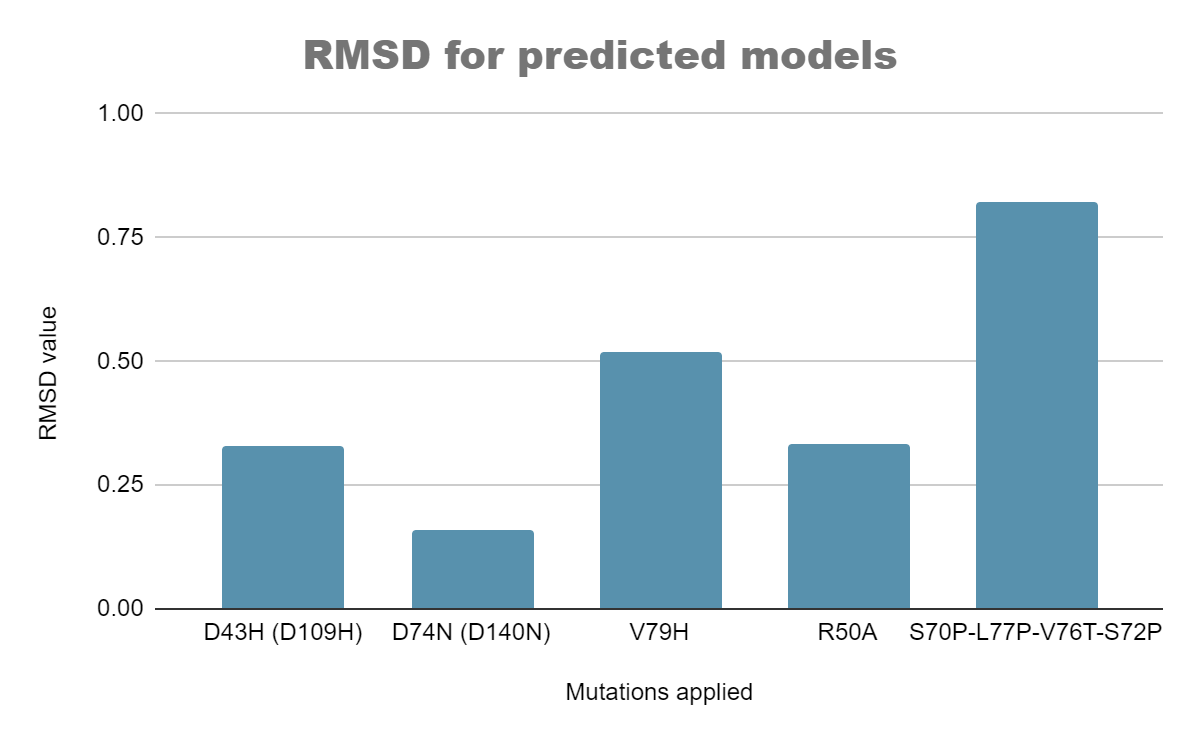
We used the pipeline to predict structural changes of AlphaB Crystallin carrying mutations recorded in literature and mutations that are proposed to bring structural changes. We performed structural prediction for two mutations that are mentioned in research to cause cataract: D43H (D109H) / D74N (D140N); one point mutation that we propose to disrupt the structure: V79H; and a group of mutations that likely to disrupt H-bondings within the protein: S70P-L77P-V76T-S72P.

We performed the same mutation prediction using Robetta in the previous version of the pipeline. We see a difference in the results given by Robetta and ESMfold. There is a better overall alignment for ESMfold compared to Ronetta, and ESMfold seems to provide more rigid predictions for the regions that are disoriented. At the same time, ESMfold also gives a larger regional shift around the inserted mutations compared to Robetta.



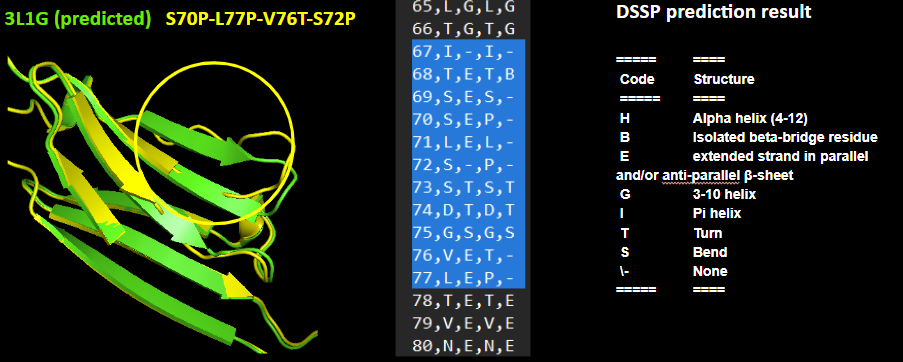
The below chart shows different mutations we get from the pipeline. Generally we see a larger RMSD when mutations like V79H significantly disrupt the hydrophobic core, and group mutations like S70P and L77P disrupt the anti-parallel beta sheet.

Mutations D43H and D74N recorded in literature are considered cataract-causing as they change secondary structure and tertiary structure of AlphaB Crystallin by disrupting the disulfide bridges. Our pipeline successfully captures a small shift around the residue for those two mutations, which further prove the ability of our pipeline to present structural changes in respond to mutations. However, due to the nature of ESMfold we are not able to predict the changes in oligomerization.



Some of the mutations are likely to change the secondary structure arrangement of the protein due to disrupted H-bondings between chains. We tried implementing a group of mutations consisting of S70P-L77P-V76T-S72P that simulates such a situation and see whether the pipeline can capture the difference.

Predicted structure from ESMfold effectively reflects the residue shift due to the mutations implemented. DSSP algorithms in our pipeline capture the change of anti-parallel beta-sheets into disoriented structures when the residues are replaced due to proposed mutations. The information given by secondary structure prediction combined with RMSD gives a comprehensive overview of type and consequence of structural changes ground by DSSP.



# Discussion

Our developed pipeline for 3D protein structure prediction allows for structural comparison of

two protein structures, given the amino acid sequences. In addition, the pipeline has the functionality to incorporate and predict effects from point mutations.

In developing our pipeline, we first compared the performance of Robetta and ESMfold to assess its reliability. Examining the predicted structures through Pymol, we observed that ESMfold gives a stable prediction on disoriented regions of the protein. Robetta caused unstable predictions in disoriented regions, overwhelming the relatively small changes in mutated regions when calculating the overall RMSD. The fact that ESMfold predictions only respond to mutation by raising local residue shifts makes it possible for us to conclude that the change of RMSD is an accurate representation of structural changes caused only by introduced mutations and the value of RMSD is comparable to the expected degree of influence from point mutations.

We demonstrated the use of our pipeline by applying it to study the structural changes of mutations AlphaB Crystallin. We saw the ability of our pipeline to effectively capture both regional shifts and secondary structural assignments caused by point or group mutations. As the reliability of our pipeline is demonstrated in the analysis of AlphaB Crystallin, we state that the pipeline has larger potential in cases where we do not have available crystallized structure for the target protein. In this case, our pipeline will provide a preview of the outcome of certain mutations and thus help improve the efficiency of related research by preselecting mutations that are most effective in bringing structural changes.

Our pipeline has certain limitations. In regards to the pipeline compatibility, one such limitation is that we have to set up two separate conda environments to successfully run our pipeline. The module used in PDB structural analysis requires a specific libboost module version that is not compatible with the environmental setup on Bridges. In regards to the capability of our pipeline, there is a limitation of the availability of types of mutations the user can perform on the amino acid sequence. Our pipeline only allows single point mutations and does not allow insertion and deletion since the comparison of protein structures is only accurate when the sequences have the same length. Another capability limitation is that the pipeline can only predict the monomer structure changes of the protein. The pipeline cannot predict how the mutation affects the oligomerization of the protein. As a result, the output cannot be used to provide a definitive conclusion on how the functionality of the protein will be affected.

One of our future improvements would be to make the pipeline in a more comprehensive version regarding the availability of the mutations input. Since our goal of this pipeline is to make a prediction of structural changes from mutation in any protein, we should account for the most common type of mutations in one protein instead of only single point mutations. Another improvement we want to make is to combine the structure prediction and comparison into one script, which will be smoother and easier for users to perform mutation, prediction, and analysis in one run.

# Reference

1. Laganowsky A, Benesch JL, Landau M, et al. Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. Protein Sci. 2010;19(5):1031-1043. doi:10.1002/pro.380
2. Budnar, P., Tangirala, R., Bakthisaran, R. et al. Protein Aggregation and Cataract: Role of Age-Related Modifications and Mutations in α-Crystallins. Biochemistry Moscow 87, 225–241 (2022). <https://doi.org/10.1134/S000629792203004X>