End-semester Report

Semester Project

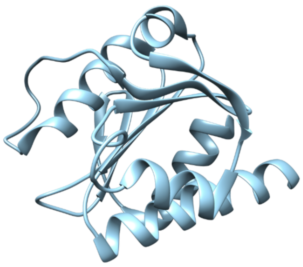
Aryan Atul Surana
under the supervision of
M.S. Madhusudhan

This is a comprehensive report of all the work carried out as part of a semester project between January to April 2025 on the apo-holo project.

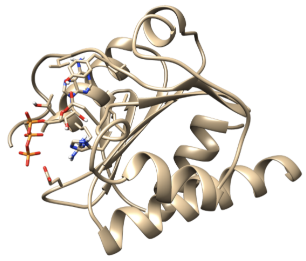
Aim of the Project: To develop a model to predict the conformational changes that a protein undergoes when it binds a ligand.

Abstract: Proteins when bound to small molecule ligands undergo certain conformational changes, especially in regions around the ligand binding site. This could lead to a loss/gain in function or efficiency of the protein. We want to study proteins in their bound and unbound states and see the changes that occur in RMSD between the two. Using this data we can model the changes that occur and use Molecular Dynamic simulations to see if we can try and predict these changes. If our model is successful in providing us with predictive binding sites and their size this can be used for drug design.

Apo and Holo proteins: Proteins that do not have a ligand bound to them are called Apo proteins and those with a bound ligand are called Holo proteins. To predict bidning sites in apo proteins we only use proteins for which we have both apo and holo forms.



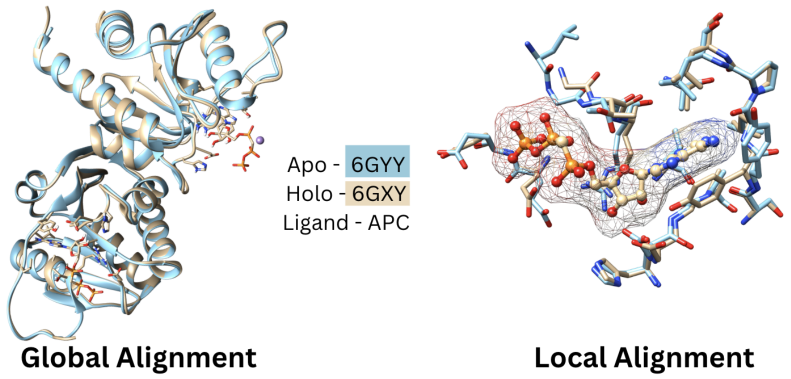
Apo protein



Holo Protein

Dataset Filtration: The dataset was filtered by my mentor Vipul Nilkanth. He looked for Uniprot PDB IDs with both the presence and absence of HETATM. If a line in a PDB files starts with HETATM it might have a ligand and although it can also have a non-standard amino acid residue in some cases we will eliminate those cases further ahead. Then we filter out NMR structures since those are dynamic and therefore it difficult to align their different poses. A resolution cutoff of 3.5 Å is also applied to ensure our RMSD calculations are precise. Only proteins with ligands in the range of 100-1000 Da are kept as this is the range relevant for drug design. After this filtering we get approximately 798 pairs of Apo-Holo proteins.

Global and Local RMSD: Now that we have the pairs, we will align them and see what changes there are between the 2. To analyse what changes happen due to ligand binding we will look at the change in Root Mean Square Deviation after aligning all the common atoms. RMSD thus obtained is the Global RMSD. However to predict ligand binding we are more interested in cases where the local change is significant enough so protein function will be affected. For this we define a binding pocket, which is defined for the holo form as all residues inside a 5 Å radius of the ligand. For the apo form it is the corresponding atoms. The RMSD between these common atoms is called the local RMSD. Note that we need to consider global RMSD as in some cases, there will be chains that are spatially apart despite the 2 proteins having the exact same sequence. We want to eliminate these cases as alignment will be improper in these sequences. For the alignment, we use a 3D least square algorithm.



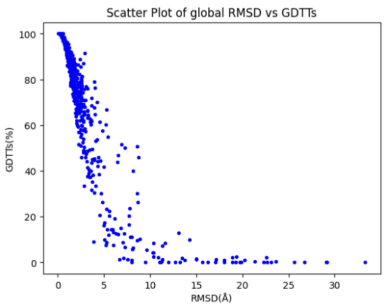
3D Least Square Algorithm: We input lists of coordinates with atom-atom correspondences, in this case the residues from the common Uniprot IDs atom wise will be aligned. Only if a certain atom of a residue is present in both the files do we add the coordinates to an array. Only atoms from chains that are corresponding to Uniprot ID are used to populate the array in cases where a protein has multiple chains. The two arrays generated from the apo and holo file are then fitted by finding a translation vector T and a rotation vector R which minimises the R.M.S.D between corresponding atoms. This calculated R and T are then applied to all the atom coordinates in the PDB file to see the superimposed proteins. It requires atom-atom correspondences which in our case are all the common atoms other than Hydrogen.

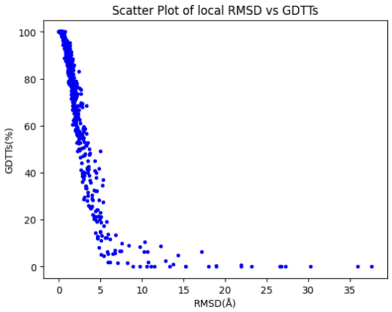
$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - (\mathbf{Y}_i\mathbf{R} + \mathbf{T})\|^2}$$

Structural Overlap and GDTTs score: The 3D least square fit algorithm also calculates the structural overlap which is the number of atoms overlapped, given a distance cutoff. If m out of n atoms are within the SODC (structural overlap distance cutoff) then the overlap is m/n*100. We calculate this Structural overlap at 3 different distance cutoffs i.e. 1Å, 2Å, 3.5Å and calculate a GDTT (Global distance test total) score which is essentially going to be an average of these.

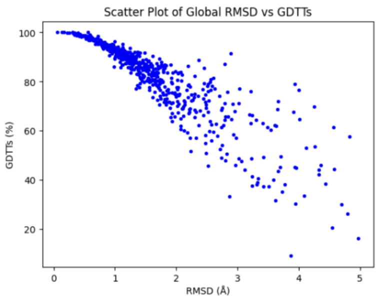
$$\mathrm{GDTTS} = \frac{1}{3}\sum_{i\in\{1\text{Å},2\text{Å},3.5\text{Å}\}} \mathrm{SO}_i$$

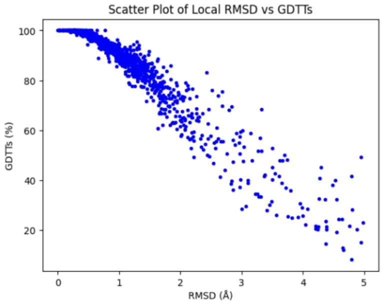
Results:



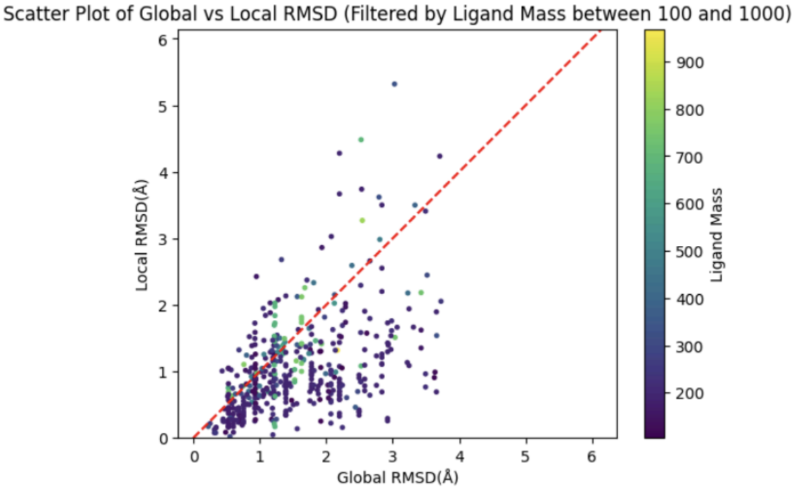


This is what the data for Local RMSD vs GDTTs and Global RMSD vs GDTTs looks like. This makes sense because as more atoms are out of the distance cutoff, RMSD would increase and conversely structural overlap would decrease so the first 2 scatter plots make sense. However the reason for extremely high RMSD in some of the global cases and correspondingly in the local cases is due to proteins with the same sequence having different orientations in space. This causes them to have a high RMSD. Below are plots where global and local RMSD are bounded below 5Å.

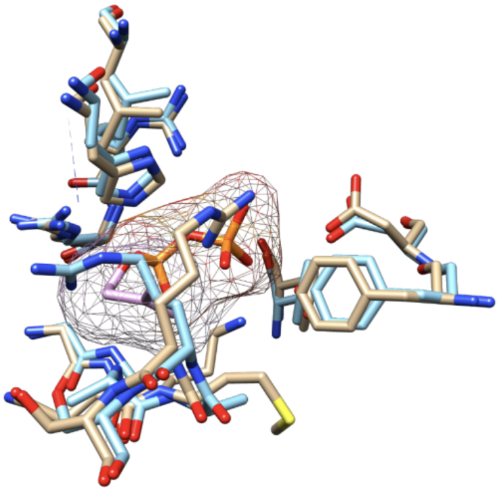




Now to see what cases have local RMSD more than Global we make the below plot which gives a metric of how much of the change is local compared to overall difference in the 2 proteins. The red dotted line is y=x and we can see there are numerous points above the lines which show that in some cases the local RMSD is higher than the global.

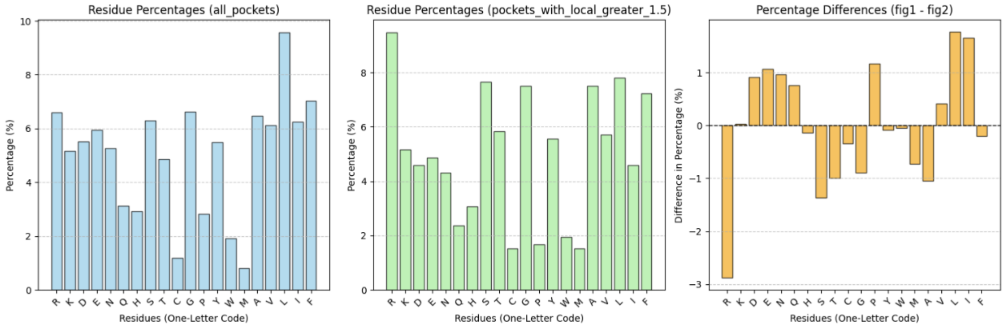


One of the many cases where this happens is shown below. These cases are very interesting to us since this means the change that is induced locally is more significantly due to the ligand.

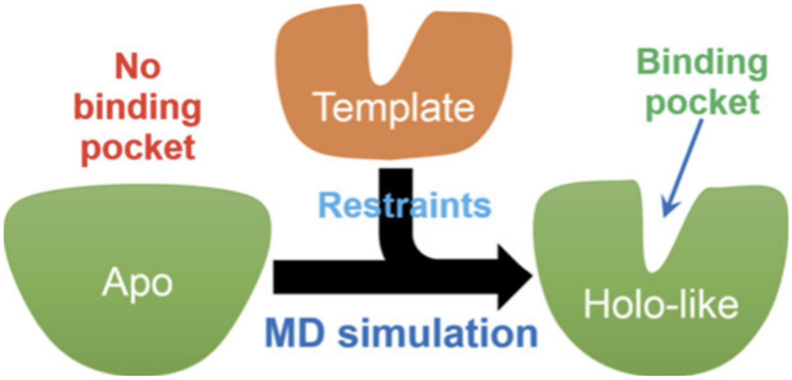


Ligand in this case is IPE.
3-METHYLBUT-3-ENYL TRIHYDROGEN DIPHOSPHATE
Apo PDB ID - 7S0H
Holo PDB ID - 7S0M

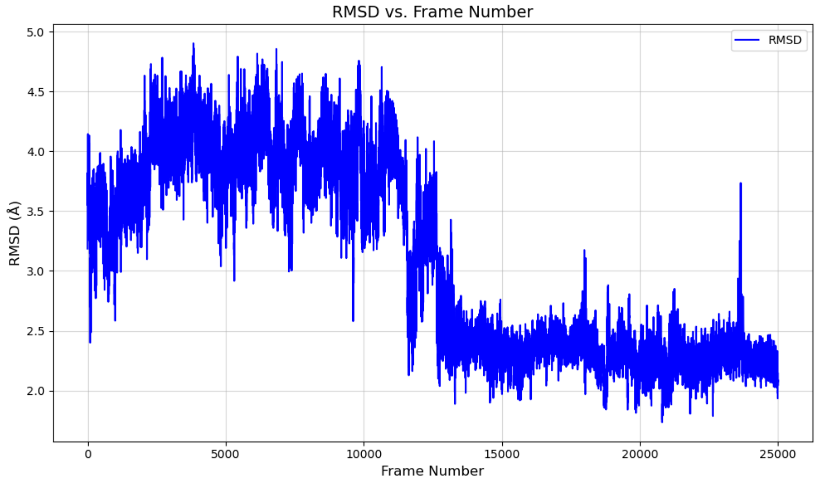
Residue specific changes: We decided to see which residues are abundant in the cases where the local change is high, which are the cases relevant to us. The cases where local RMSD is higher are better to study for drug discovery since these proteins are more likely to be functionally compromised. We look at the cases where local RMSD is over 1.5 Å, and see what residues are more prevalent in these cases versus all the cases that we have. One important difference between the 2 we see is Arginine (represented by an R), which lead us to believe side chains are a large reason for the high RMSD in all atom alignments. So we decided to go ahead with C alpha measurements thereafter as side chains are in fact very flexible and would be harder to predict. Note that the residues are arranged according to polarity with arginine being the most polar.

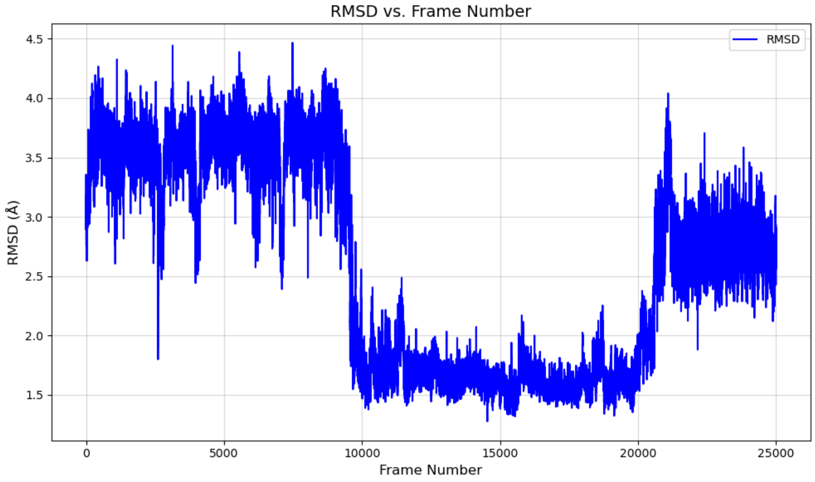


Molecular Dynamics Simulations: In my work on the apo-holo project I came across a paper by Jinse Zhang et al, who have the following work flow:

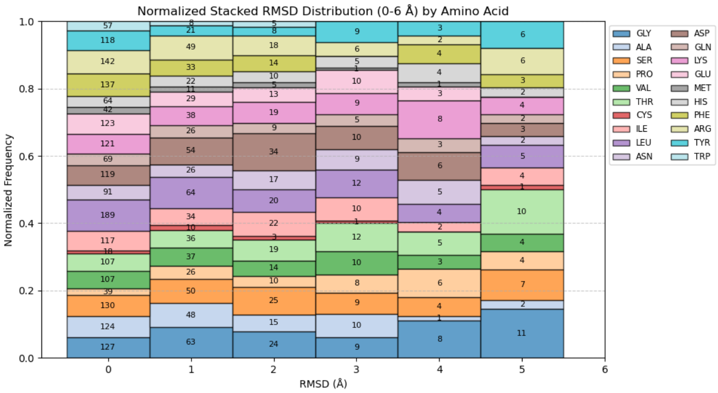


Following the work of Jinse Zhang et al. who run MD simulations on apo-structures for 32 pairs and predicted an improvement in 23 of 32 structures and an average improvement of 0.31 Å, we decided to run MD on one of our proteins from the dataset. We run 1X8E.pdb for 250ns for 2 runs and an average improvement of 1.205 Å was observed in the binding pocket relative to the holo form 2GC3.pdb. We have a much smaller dataset to make this claim but running MDs for the other proteins is more of a computational problem and can be done easily.

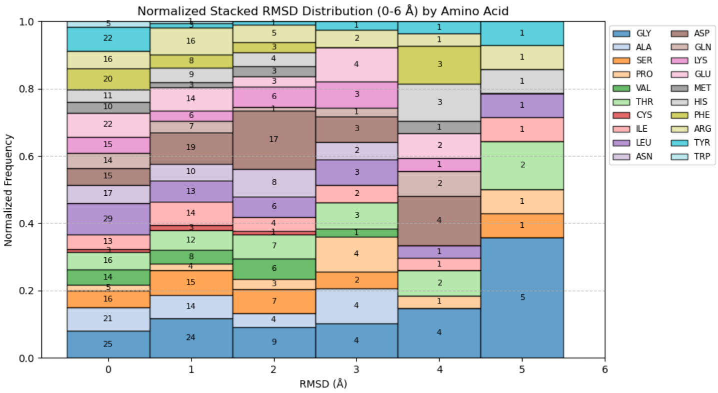




Variation of C α : Amino acids with larger side chains often fluctuate a lot more when aligned for all atoms, however a true representative atom of them, that is, the C α does not vary as much and hence an all atom alignment and an all atom RMSD is not a correct representation and hence a C α alignment is used henceforth. We plotted every residues variation in RMSD and plotted a stacked histogram going from 0-6 Å. Below is the plot for all the binding sites in question not accounting for high global RMSD and low local RMSD cases:

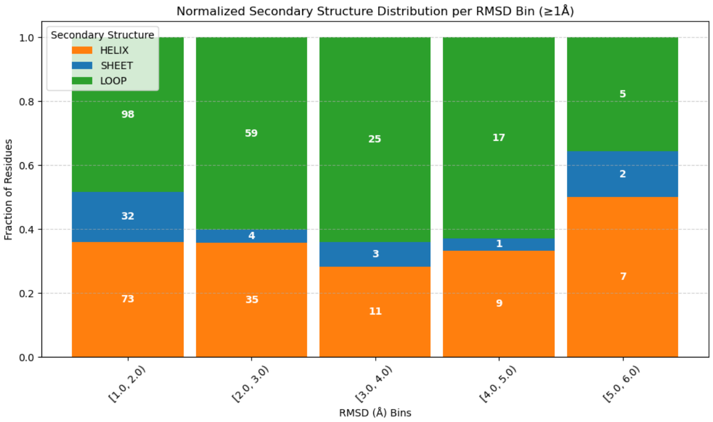


However the cases that are highly relevant to us include those with a low global RMSD (i.e. probably vary mostly in the binding site) and high local RMSD. So we also look at the 62 binding pockets, which have around 691 residues, where global RMSD < 3Å and local RMSD > 1Å.



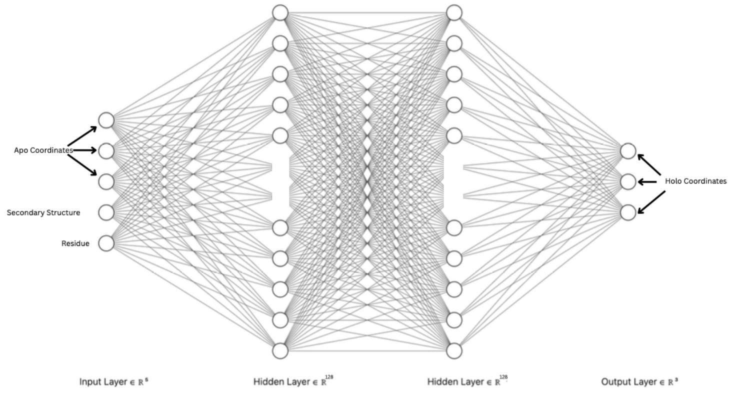
This histogram is incredibly useful to train our neural network. It shows that in cases where there is an appreciable RMSD, cases relevant to us, Alanine never varies more than 4Å. This is very relevant to our model as certain residues are more rigid and less likely to move and its something we need to take into account.

Secondary structures: The secondary structures our high-variation binding sites have in their residues are functionally relevant. The distribution of the secondary structures that prevail in the RMSD bins is as follows:

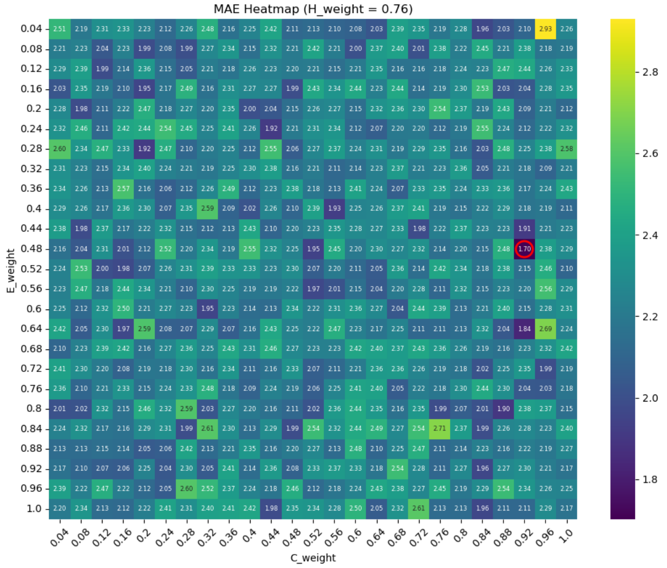


This will also help train our neural network as a residue in sheets, for example, will move more than 2Å only 10/42 times. However a residue in a loop will move over 2Å 106 times out of 204 cases. This is a vital distinction and something we predicted as we expected coils to be the most flexible followed by helices and sheets are something we expected to not only not fluctuate much but also be less common in sites with conformationally changing sites which is clear from the results.

The Neural network: Using the information so far about variation of C α of various residues and their secondary structure we will now design a neural network to make predictions of what conformational changes can happen in a binding site. Architecture of the network is as follows:



For this we have a neural network with 2 layers with 128 neurons each. It takes 5 inputs which include the 3 apo coordinates of the residue (x,y,z) , the secondary structure and which amino acid it is. It outputs the holo coordinates of that residue. Now we need to ensure we can check the validity of the network as well so we train on 80% of the dataset we have and test on the remaining 20%. We also cannot give equal weightage to sheets, coil and helix residues so we ensure the network iterates different weights to give to residues found in each of these secondary structures between 0 and 1 with a step of 0.04. We only want to see the heatmap of the case with the least MAE(Mean Aligned Error) which is shown below.



We can see that the best MAE we get is for high C-weight, H-weight and low E-weight. This makes sense because we have the highest number of cases for Coils so weighting it more is good and lowest cases for sheets so giving it a high weight is not a great idea. For this weights 10/138 of our predictions are within 0.5 Å, 28/138 are within 1Å and 60/138 are within 2Å which are results that could use improvements but a good place to start for sure.

Future work:

1. Account for opening and closing sites
2. Reduce overfitting by changing layers and neurons
3. Reduce MAE by refining the network
4. Predict conformational changes and use docking software to validate for cases with no holo form

References

1. Holo Protein Conformation Generation from Apo Structures by Ligand Binding Site Refinement Jinse Zhang et al. `https://pubs.acs.org/doi/10.1021/acs.jcim.2c00895`

2. On the orthogonal transformation used for structural comparisons. Simon Kearsley et al. `https://journals.iucr.org/paper?gr0023=`