AlphaFold for the Study of Rare Genetic Diseases

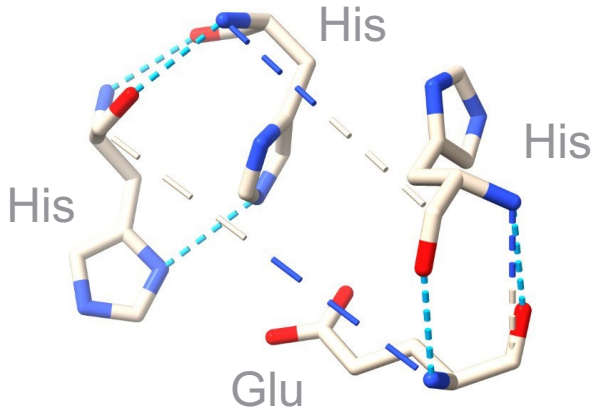
Computational Health Laboratory a.y. 2023/24

- Irene Dovichi
- Giacomo Lagomarsini
- Marco Lavorini
- Alice Nicoletta

Introduction

Alkaptonuria (AKU) is a recessive genetic disease caused by a mutation in the HGD gene.

Common symptoms: damage to cartilage and heart valves, kidney stones, and more.

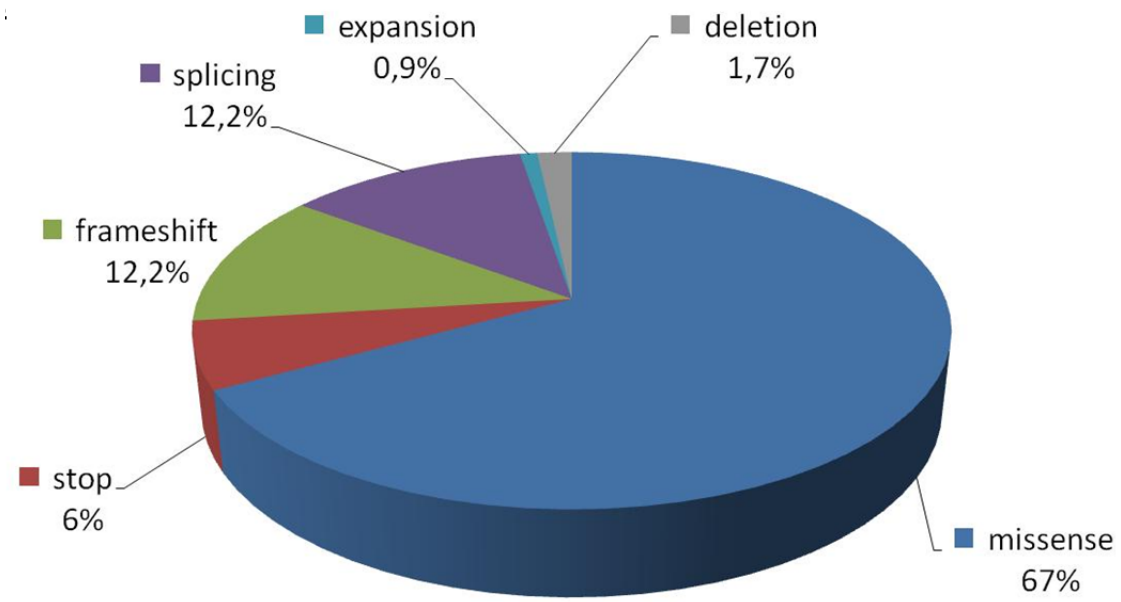


To date, a **genotype/phenotype correlation** is not recognized, and even siblings with the same genetic mutations and a similar diet show different ages of onset and different severity of symptoms.

One of the main obstacles for studying AKU is the **rarity** of patients and the **invasive** techniques required for sample collection.

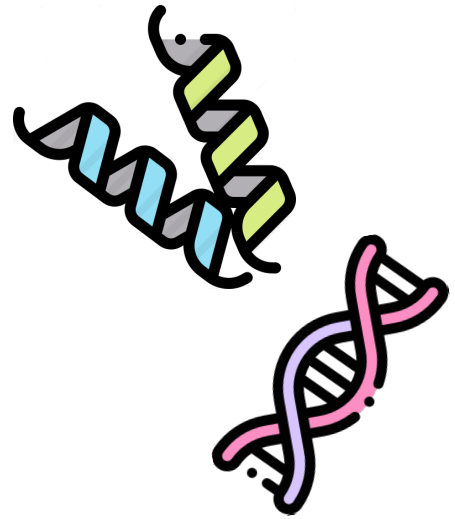
Types of Variants

The classification of variants is related to how they affect the process of transcription (DNA → mRNA) and translation (mRNA → protein), which ultimately influences protein function.



- **Missense**: a single nucleotide change results in the substitution of one amino acid for another in the protein.
- **Frameshift**: is caused by insertions or deletions of nucleotides that are not in multiples of three.
- **Stop**: a change in the DNA sequence introduces a premature stop codon.
- **Deletion** and **Expansion**: are caused by the removal or repetition of nucleotides.

Protein Sequences



- Missense example: **G161R** (DNA change `c.481G>A`)

Wild-type	...IVPQK**G**NLLI...
Mutated	...IVPQK**R**NLLI...

- Frameshift example: **G11fs** (DNA change `c.31_32delGGinsATT`)

...TTT **GG**G AAT GAG T...	Wild-type	MAELKYISGF~~GNECSSED...~~
...TTT **ATT** GAA TGA GT...	Mutated	MAELKYISGFIE
F I E **STOP**		

Project Purpose and Tools

What?

The project aims to **predict the severity** of alkaptonuria based on the 3D structure of mutated proteins.

How?

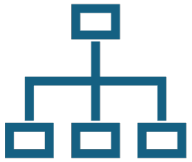
Using **AlphaFold**, we will predict the 3D structures of the mutated proteins.

→

We aim to **extract** features representative of the spatial structure and add them to the patient dataset.

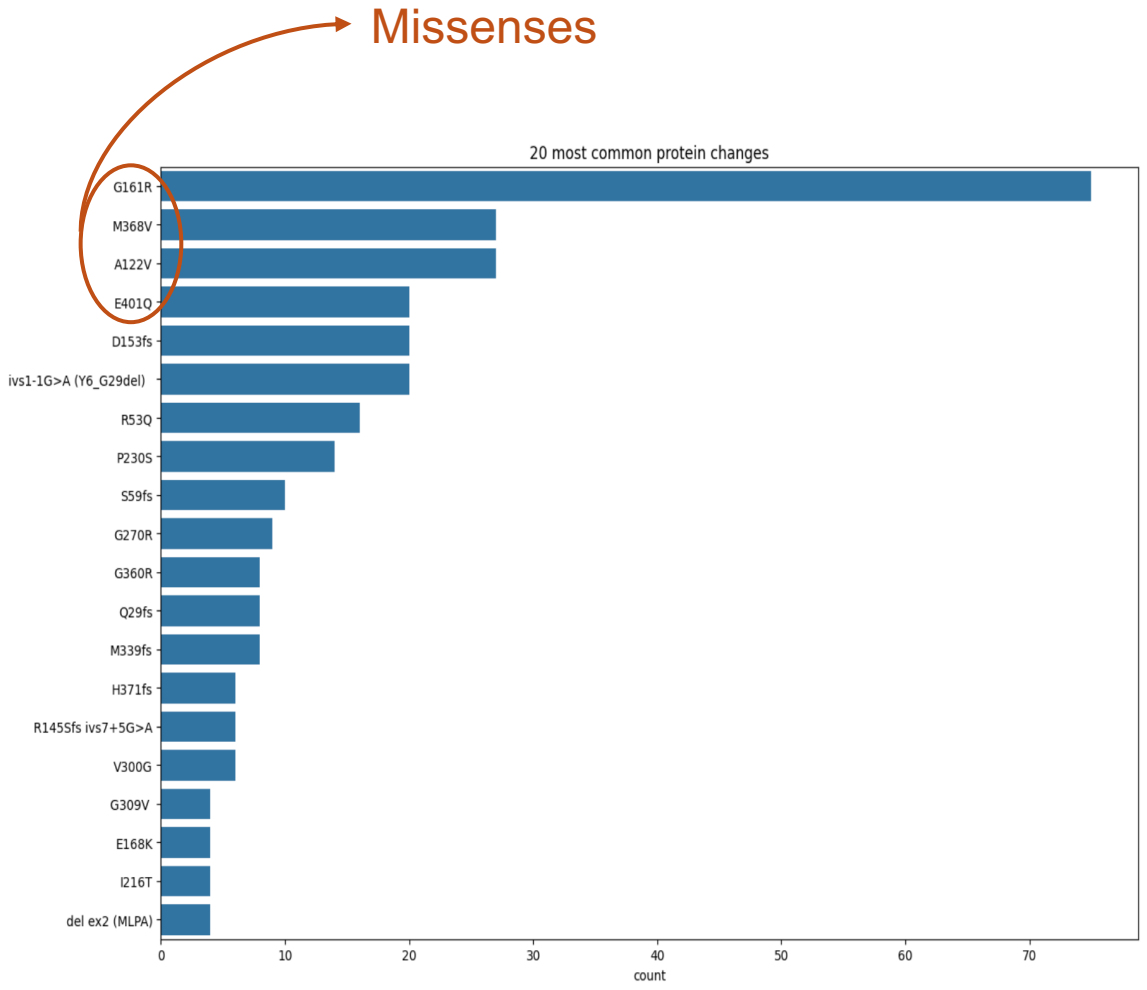
→

We will then use **Machine Learning** algorithms to search for correlations between the different disease severity metrics and the variants.



Clinical Dataset

- **219** patients affected with alkaptonuria.
- Each person carries **two alleles** for the HGD gene, one on each of the two homologous chromosomes (maternal and paternal).
- These alleles are **both mutant**: most patients are double heterozygous, i.e. carriers of two different variants, and only a minority are homozygous.
- **Problem**: we have to treat each patient as a couple of mutations.



AlphaFold

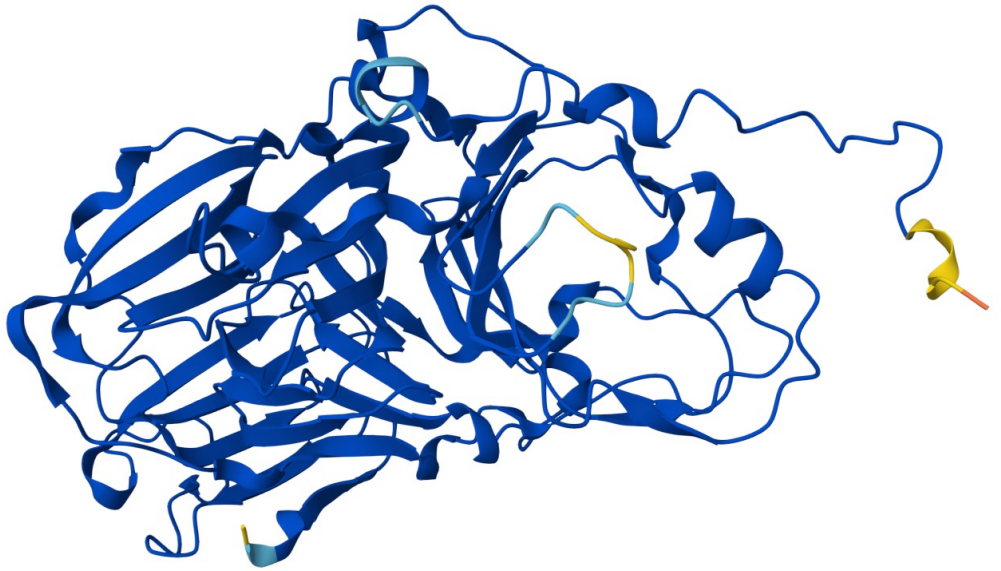
AlphaFold:
From primary to 3D structure.

Deep Learning model that uses graph neural networks.

Input: a sequence of amino acids

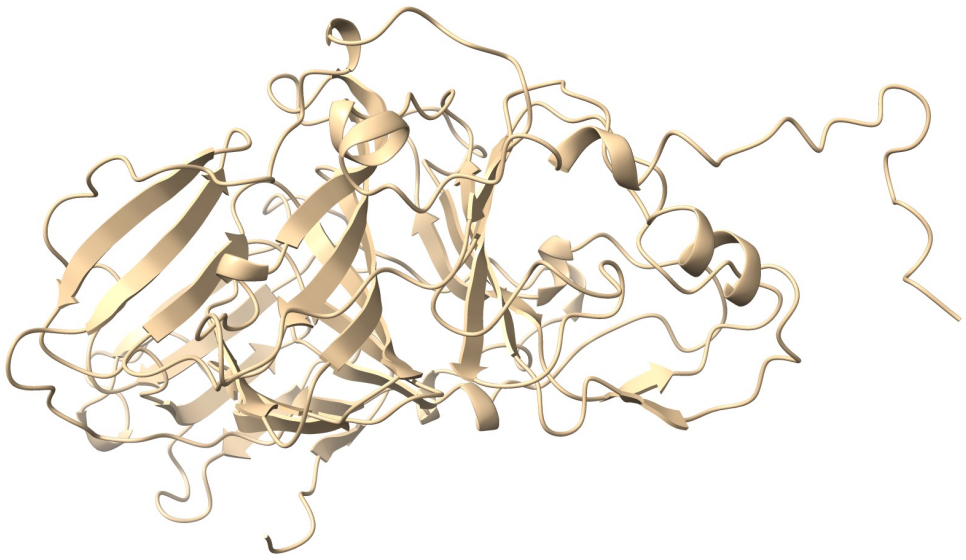
Output: 5 PDB files that map each atom in the 3D space, confidence scores, and more.

MAELKYISGFGN…

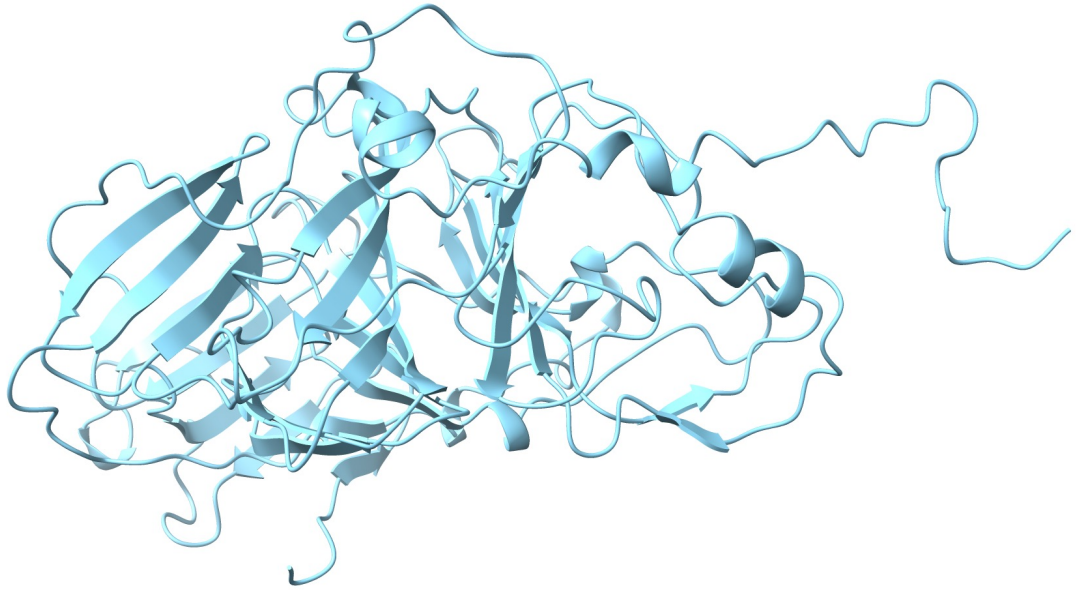


Example of Missense Mutation

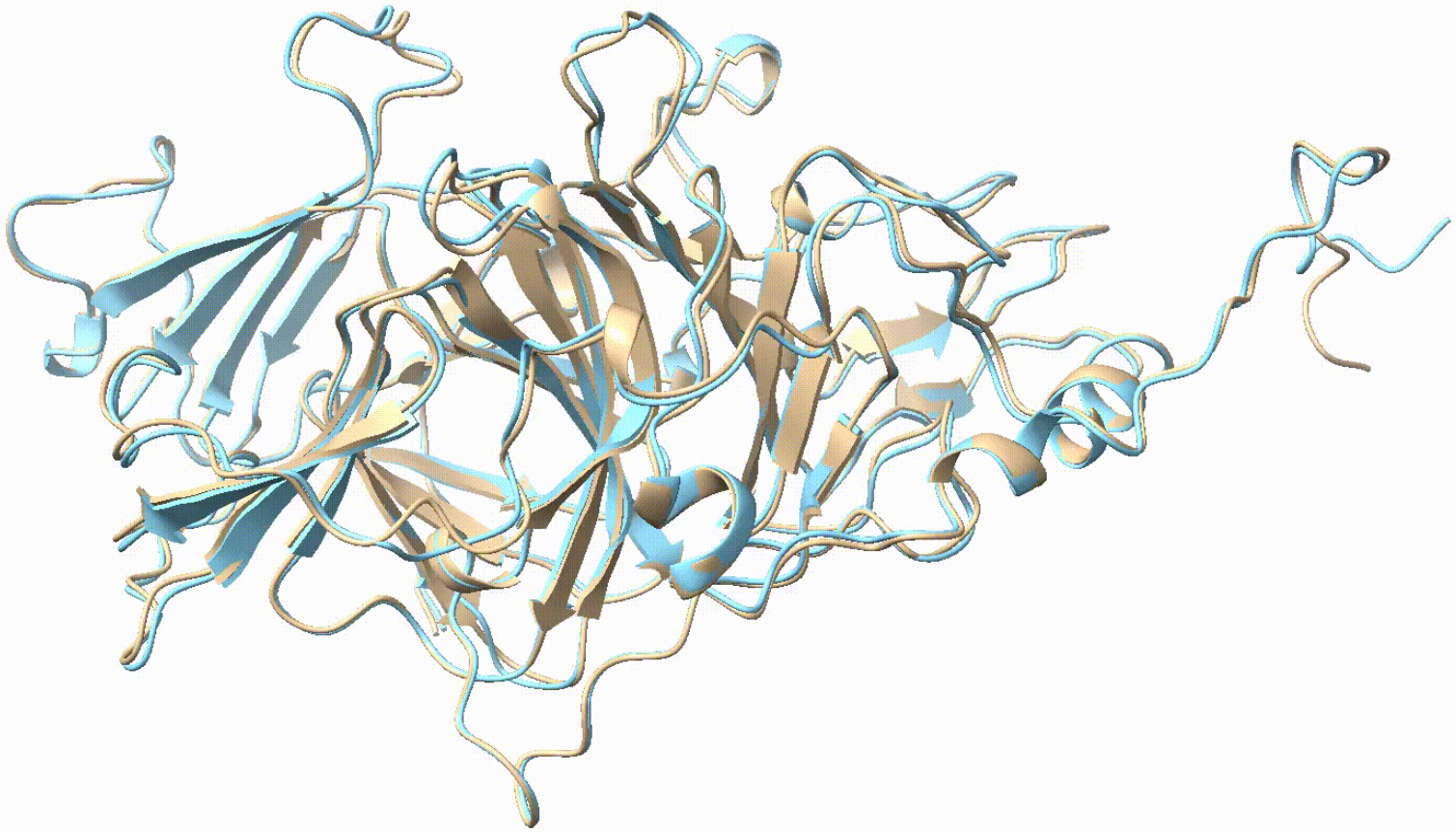
Original



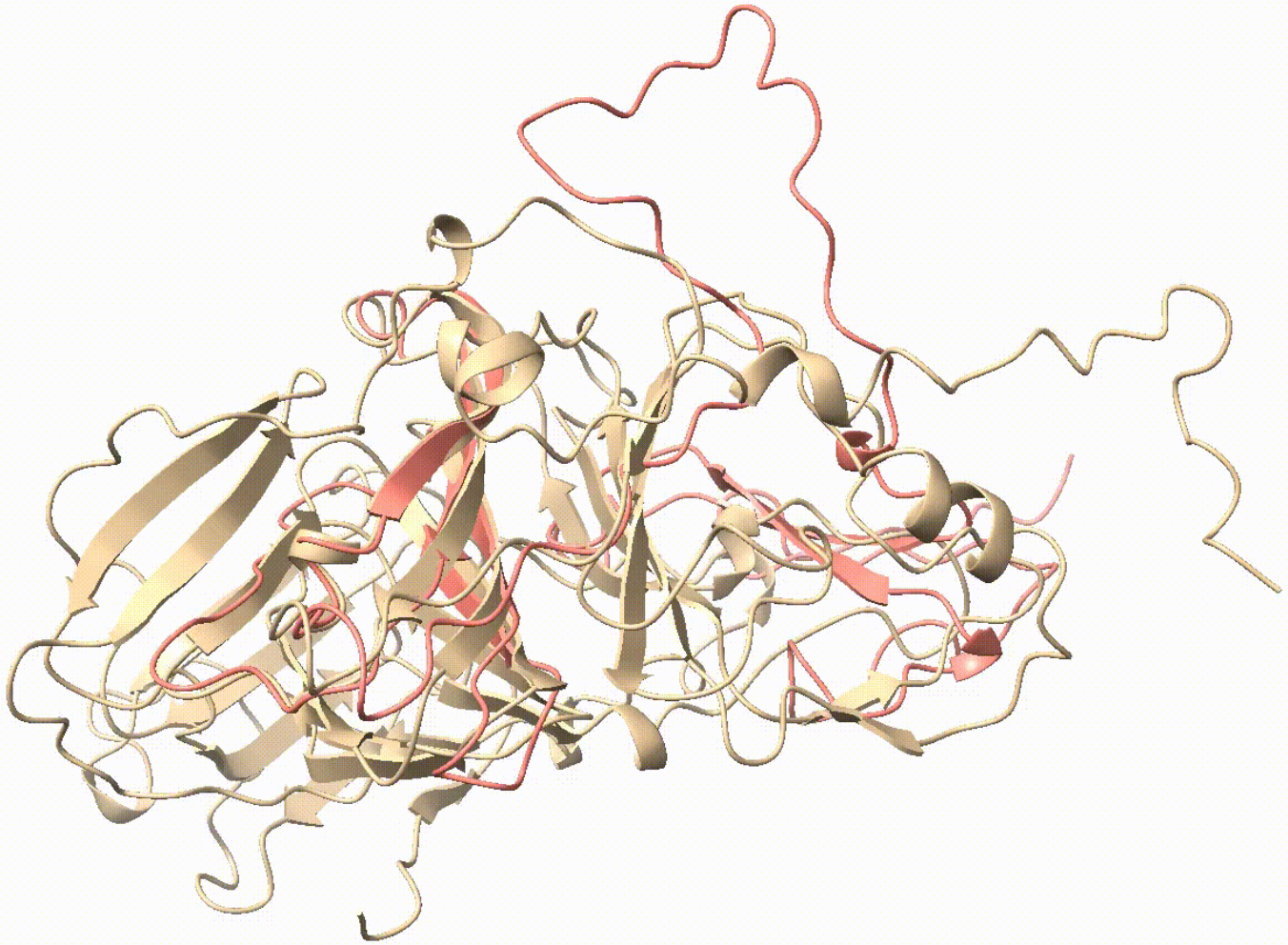
G161R



Original G161R



Original D153fs

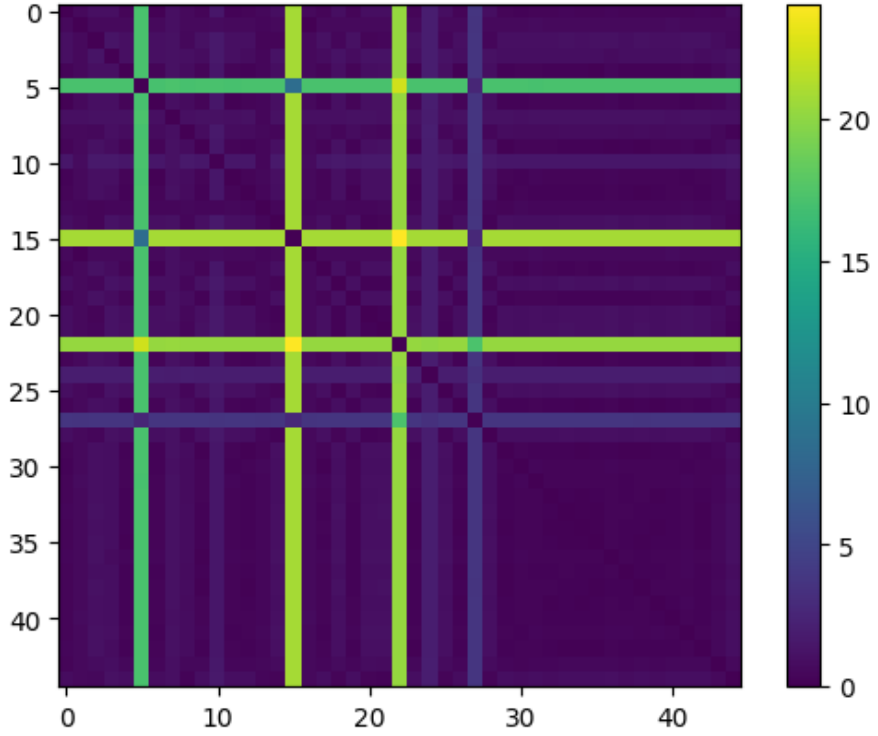


RMSD

Root Mean Square Deviation (RMSD):

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

- Requires two proteins to be superimposed
- We limit comparisons to common backbone atoms
- Each mutation is compared to each other



Clustering

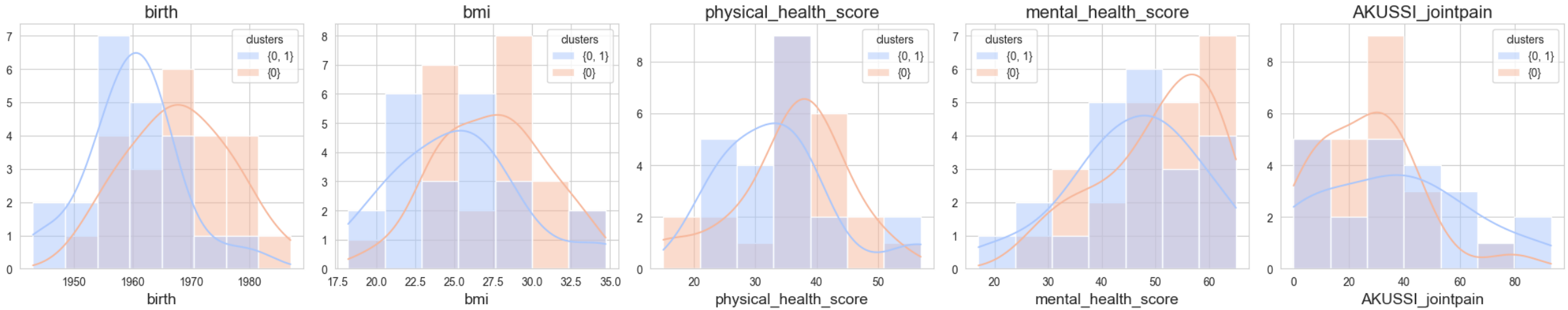
Clustering on the 3D structure could help identify which types of mutations are most harmful or functional.

Technique:

- K-Medoids clustering on the RMSD matrix
- *k = 2* chosen based on the SSE curve

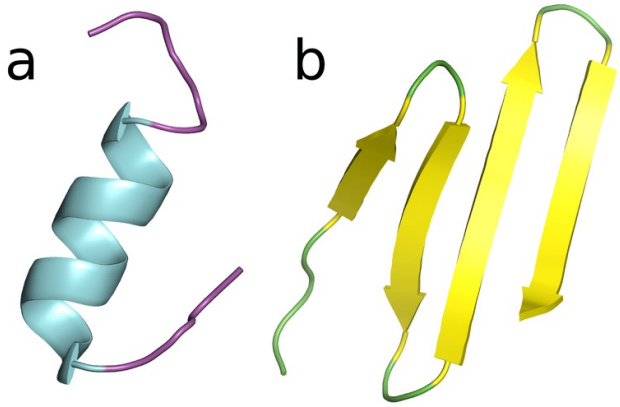
Clustering Results

- The most notable differences were observed between patients with both mutations in the first cluster {0}, and those with one mutation in each cluster {0, 1}.
- However, the statistical significance of these differences remains unclear based on **p-value** tests.



New Features I

- **Secondary structure** represented as a list of features: #α-helixes (a), #β-sheets (b), etc.
- More structural metrics via Needleman-Wunsch global alignment algorithm on **ChimeraX**.
  - Alignment Score
  - RMSD w.r.t. wild-type
- The metric reaches its maximum value of 2.348 when comparing the wild-type HGD to itself.



S--SS-GGGPEEEEEES

↓

S	-	G	E
4	3	3	6

New Features II

Variant Class	Destabilise Monomer?	Destabilise Hexamer?	Comments	Dist. to Interface	Dist. to Fe (within protomer)	Dist. to Fe (neighbouring protomer)	Dist. to Substrate (within protomer)	MONOMER		HEXAMER
mCSM proposed effect				(angstroms)	(angstroms)	(angstroms)	(angstroms)	mCSM (ΔΔG Kcal/mol)	protomer_DUET (ΔΔG Kcal/mol)	mCSM-PPI (ΔΔG Kcal/mol)
unknown	No	No		11.573	36.250	35.361	33.027	-0.059	0.082	-0.355
Protomer destabilisation	Yes	No		11.254	36.396	31.566	33.417	-1.243	-1.200	-0.114
Hexamer disruption	No	Yes		2.683	35.648	18.326	33.865	0.446	0.700	-1.562
Protomer destabilisation, Hexamer disruption	Yes	Yes		7.825	45.807	25.671	43.374	-1.522	-1.550	-1.004
Protomer destabilisation, Hexamer disruption	Yes	Yes		7.825	45.807	25.671	43.374	-0.611	-0.671	-0.859
Protomer destabilisation, Hexamer disruption	Yes	Mildly		4.255	45.388	20.240	43.331	-1.382	-1.515	-0.694

*Only for missense mutations.

Symptoms Prediction

TensorFlow

scikit learn

Features

- From genotype: RMSD, Secondary Structure information, Alignment score
- From phenotype: age, gender, bmi

Models

- Ridge Regression
- Random Forest Regressor
- Neural Network

Goal

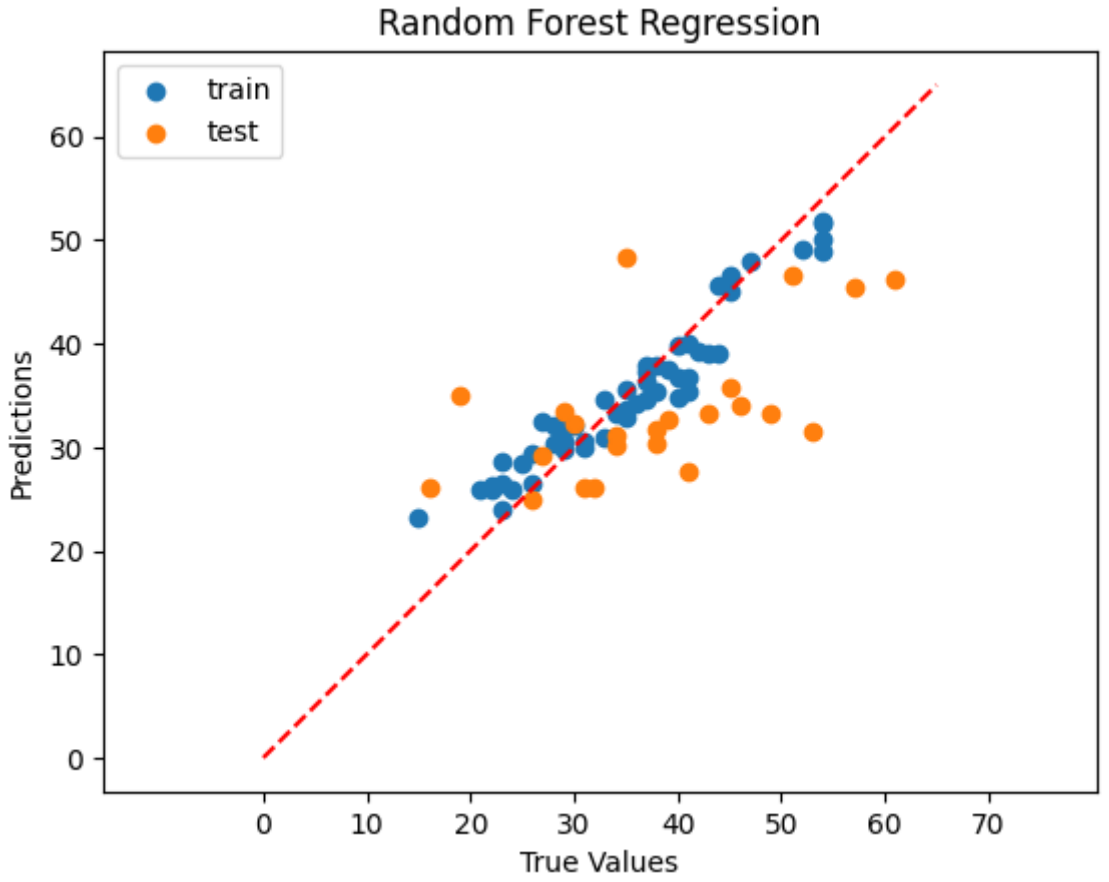
Predicting severity metrics based on given mutations

Results

We present the prediction outcomes for the Physical Health Score

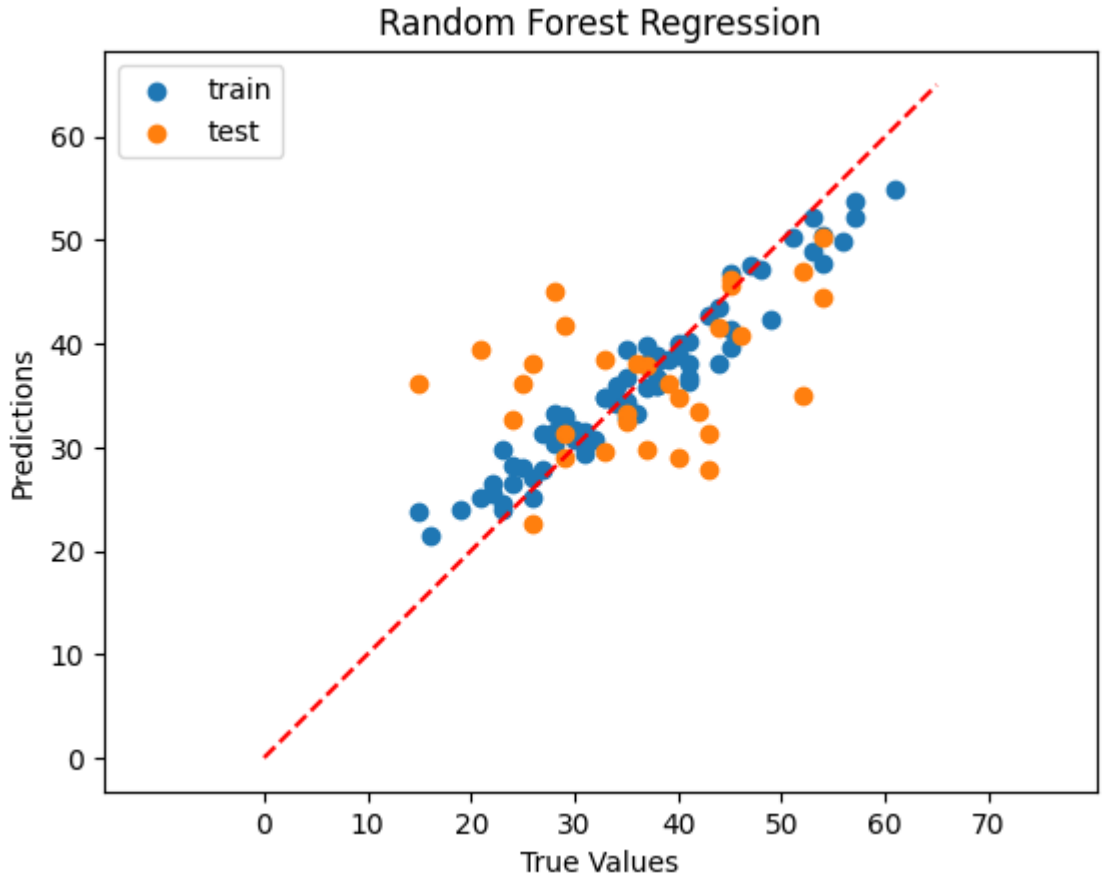
Regression Results I

- Prediction of Physical Health Score
- Feature used: RMSD, Alignment score, Secondary Structure list, Table scores
- **Random Forest**
- **MAE** of 8.7
- To include table results, we had to take only missenses from the dataset



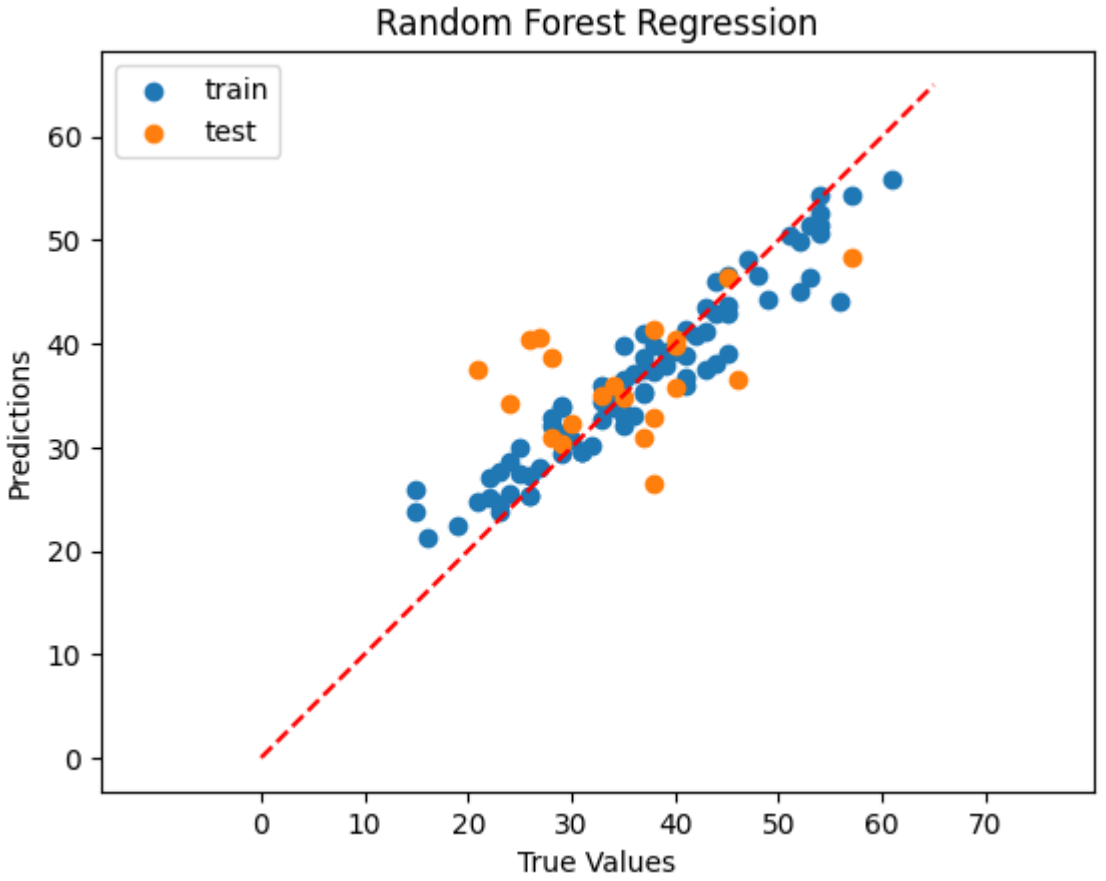
Regression Results II

- Prediction of Physical Health Score
- Feature used: RMSD, Alignment score, Secondary Structure list
- **Random Forest**
- **MAE** of 7.4
- Slightly better prediction



Regression Results III

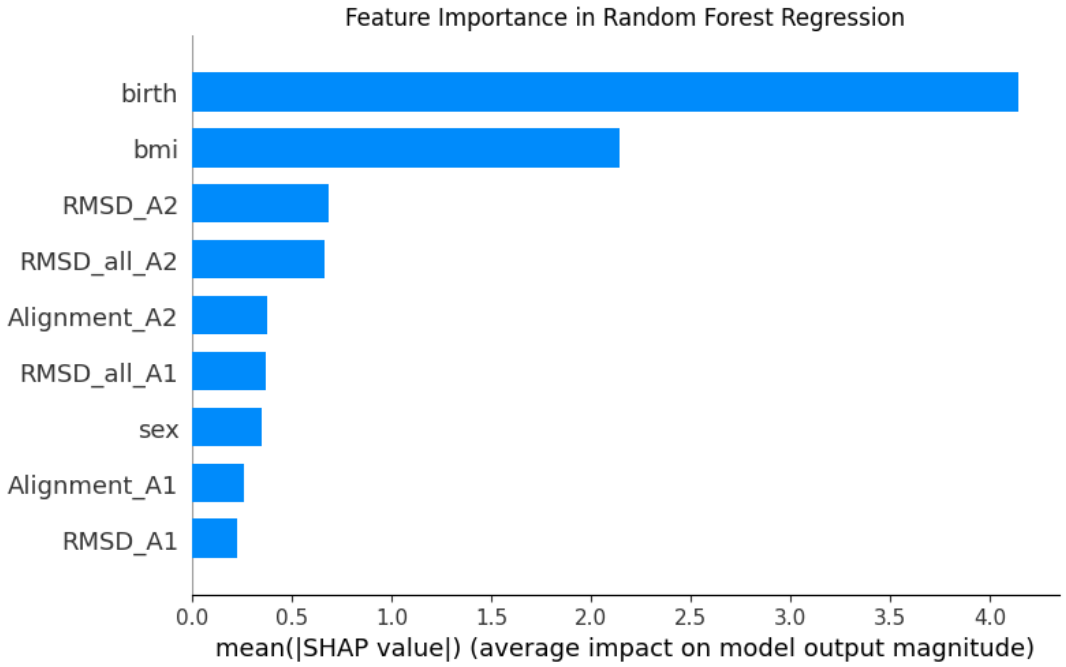
- Prediction of Physical Health Score
- Feature used: RMSD, Alignment score
- **Random Forest**
- **MAE** of 6.3
- Good positive correlation

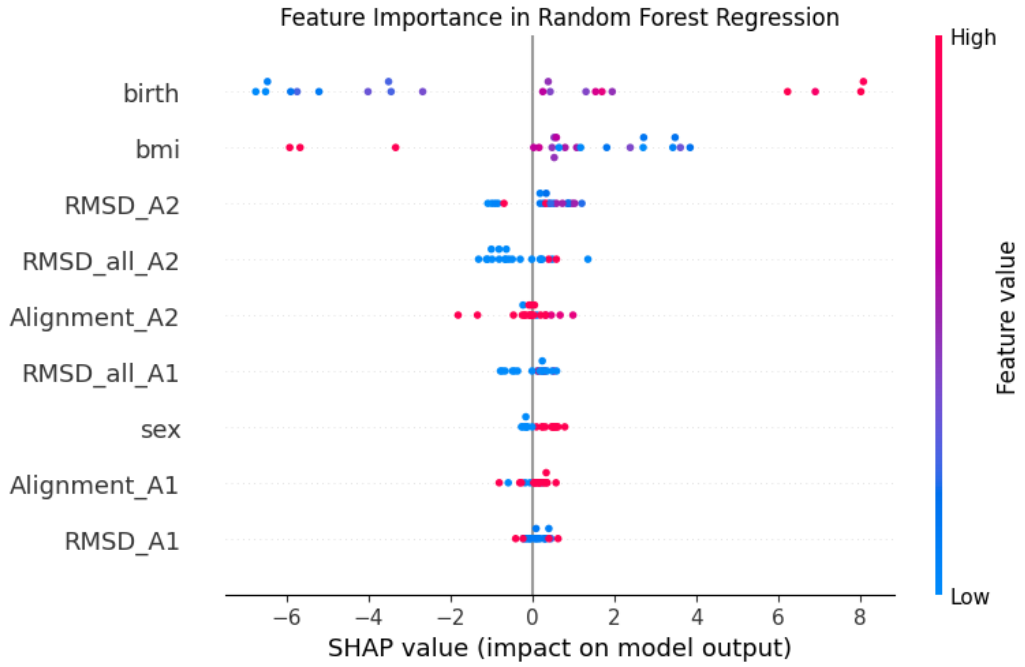


SHAP Analysis: Regression III

Key Insights:

- Higher SHAP values indicate greater influence on prediction outcomes.
- Positive/negative SHAP values show the direction of the feature's impact.





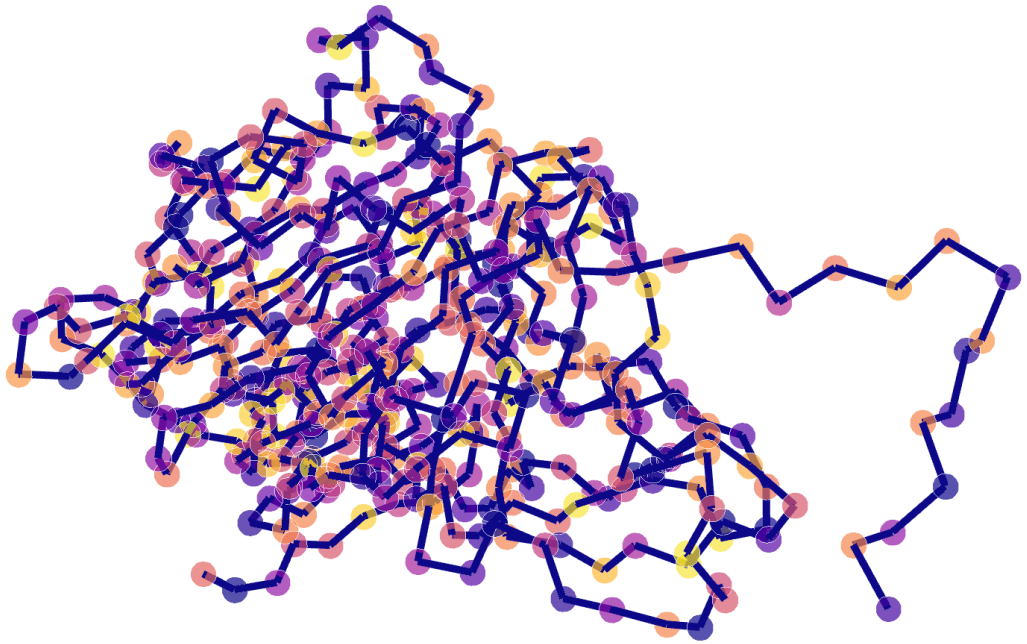
Conclusions

- Predicting **Physical** and **Mental Health Score** produces good results. **RMSD** proved to be a good way to embed structural information
- **Limitations**:
  - Many variants and a few representatives for each of them
  - Two mutations are assigned to a single score, and most of the times the two mutations are different
- **Future works**: deeper biological analysis focusing on active sites and graph representation.

Further Analysis with Graphein

To simplify the analysis of 3D structures we thought to convert the PDBs to NetworkX graphs:

- **Nodes** represent the amino acids
- **Edges** represent the kind of bonds between them (peptide, aromatic, ionic, and more)

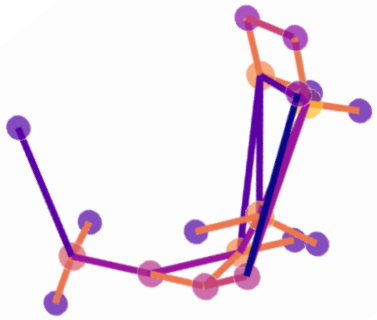


Deeper Active Site Analysis

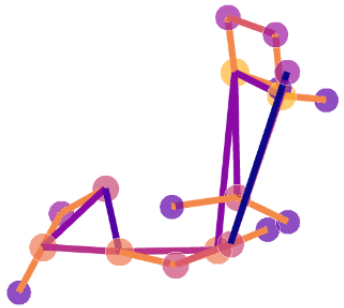
We analyzed the subgraph of the active sites and its neighbors:

- Some mutations were labelled as '**active site disruption**' (e.g. K353Q)
- Others did not present the involved amino acids due to early stop codon (e.g. **fs**)

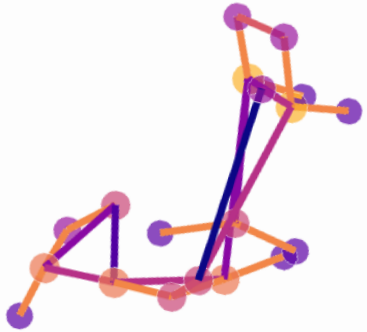
HGD



G161R



K353Q



References

[1] **HGD Mutation Database**, https://hgddatabase.cvtisr.sk/home.php?select_db=HGD

[2] **ColabFold**, https://github.com/sokrypton/ColabFold

[3] **ChimeraX**, https://www.cgl.ucsf.edu/chimerax/

[4] **Biopython,** https://biopython.org/

[5] **Alphafold**, https://alphafold.ebi.ac.uk/

[6] **Graphein**, https://graphein.ai/

[7] Zatkova A, Ranganath L, Kadasi L., *Alkaptonuria: Current Perspectives. The application of clinical genetics*. 2020; 13, 37.

[8] Bryony Langford et al, *Alkaptonuria Severity Score Index Revisited: Analysing the AKUSSI and Its Subcomponent Features*, JIMD Rep. 2018; 41: 53–62.

[9] Ottavia Spiga et al, *Machine learning application for patient stratification and phenotype/genotype investigation in a rare disease,* Briefings in Bioinformatics, 00(00), 2021, 1–13.

[10] David B. Ascher, *Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype–phenotype correlations in the largest cohort of patients with AKU,* European Journal of Human Genetics (2019) 27:888–902.

[11] Forbes J. Burkowski, *Structural Bioinformatics, An algorithmic approach.*

Thank you!