Characterisation of the pathogenic effect of missense mutations in proteins via machine learning

Qisheng Pan[1,2], Georgina Becerra Parra[1], Stephanie Portelli[1,2], Thanh Binh Nguyen[1,2], David B. Ascher[1,2]

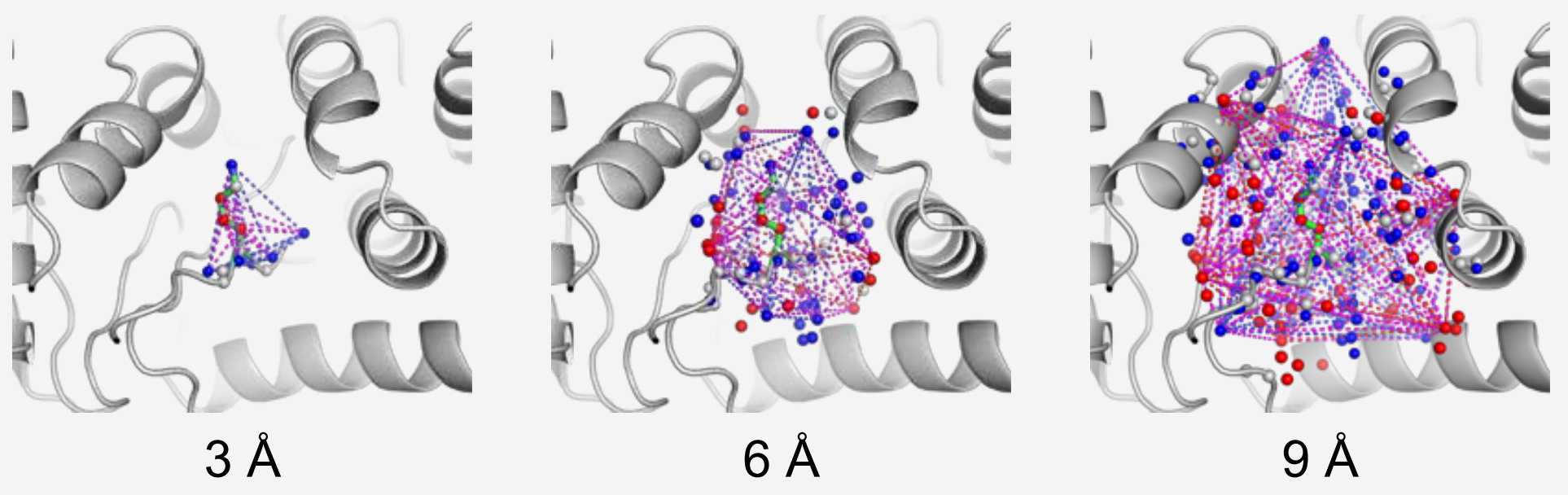
[1]School of Chemistry and Molecular Bioscience, University of Queensland, Brisbane Queensland 4072, Australia
[2]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne Victoria 3004, Australia

Qisheng Pan
QishengPan
qisheng.pan@uq.net.au

BACKGROUND

- Proteins control most fundamental cellular and biological processes, but small changes in the protein sequence can **alter these tightly regulated functions**, and may be associated with **a wide range of diseases**.
- It is time-consuming to experimentally experimentally elucidate the effects of all possible missense variants.
- Pioneering "gold standard" methods to quantify the effect of these variants rely primarily on **gene/protein sequences**, showing **limited performance** and **a bias on the deleterious variants**.
- To improve the capability of characterising missense mutations, we aimed to develop next-generation *in-silico* tools by leveraging protein information from **both sequence and structure**.

- **Graph-based signatures**: atomic pair patterns

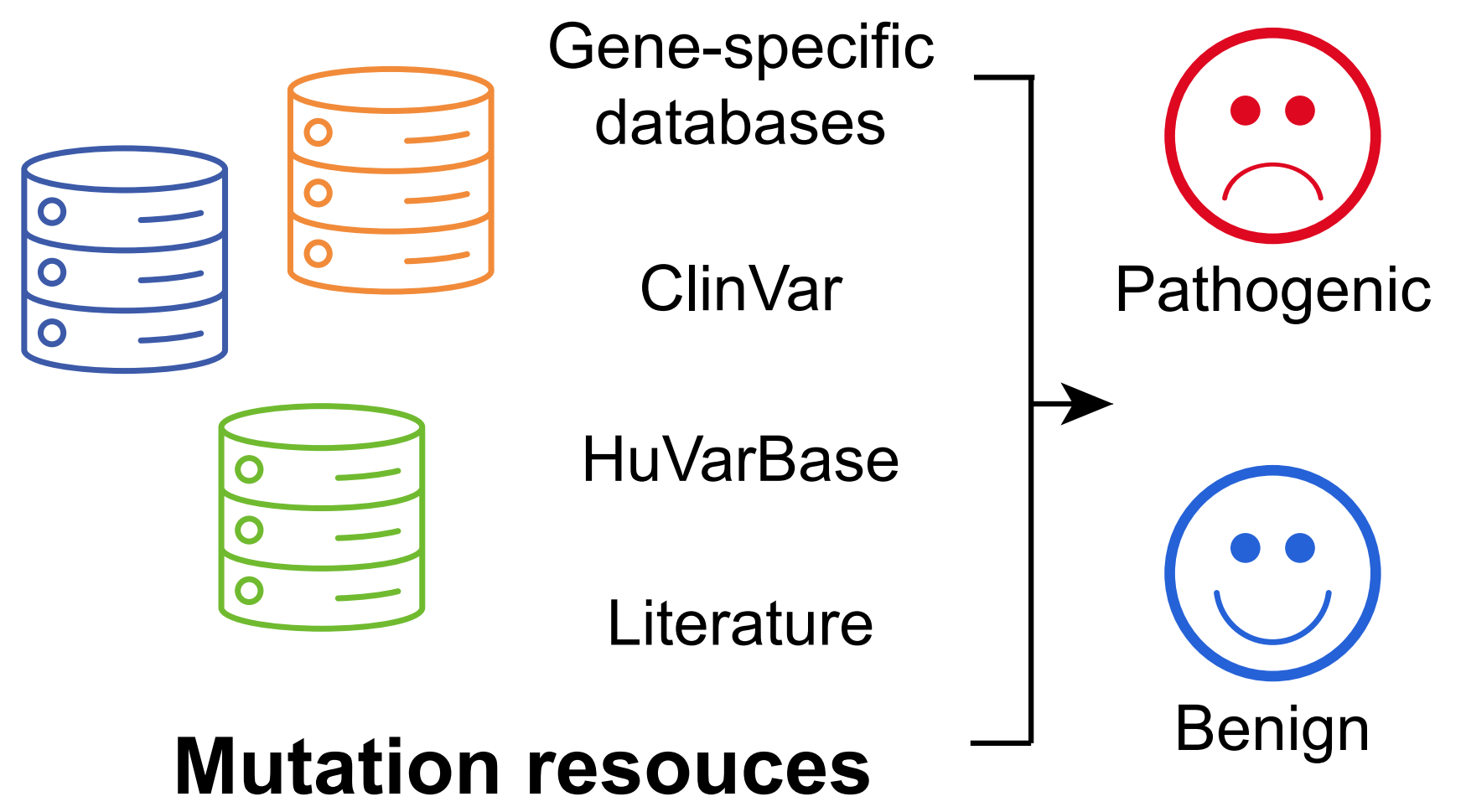


Red: hydro:hydro
Blue: polar:polar
Magenta: hydro:polar

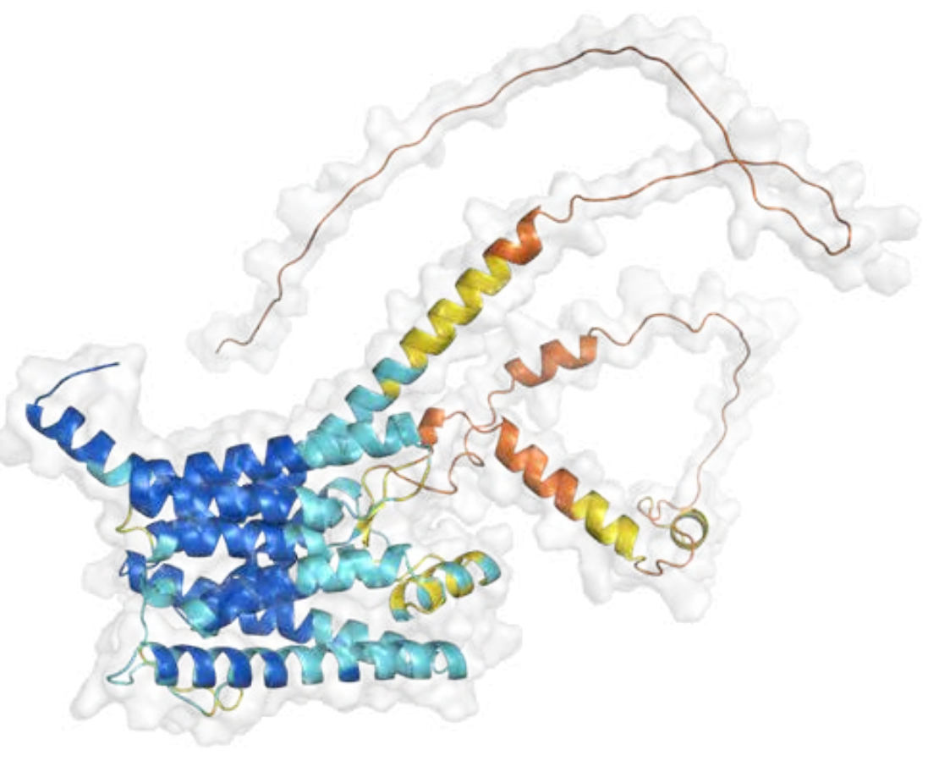
- ***Nodes***: atoms with different pharmacophores
- ***Edges***: atom pairs within a certain distance cutoff

METHODOLOGY

Data curation



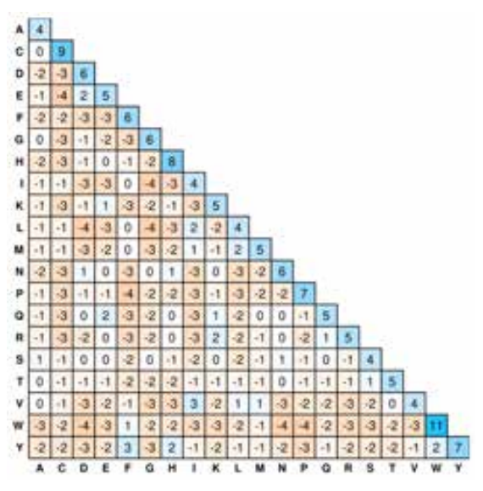
Mutation resouces



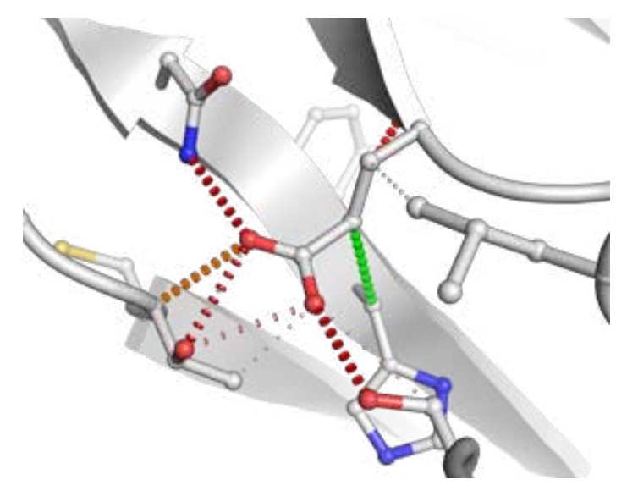
AlphaFold structures

Feature generation

Biochemical properties

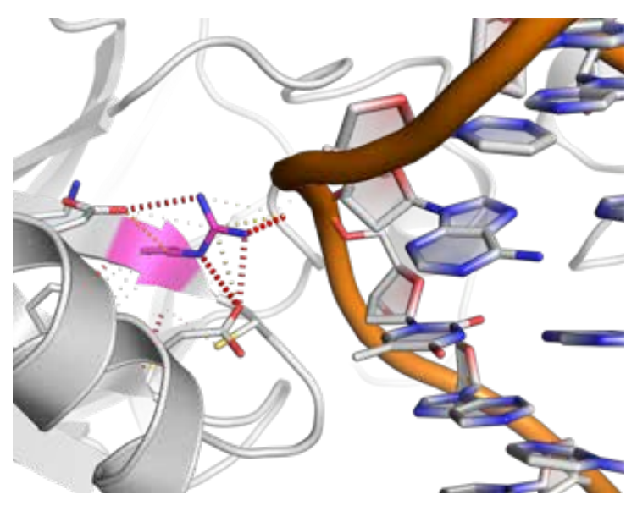


Residue contacts



Functional scores

Protein-DNA interaction

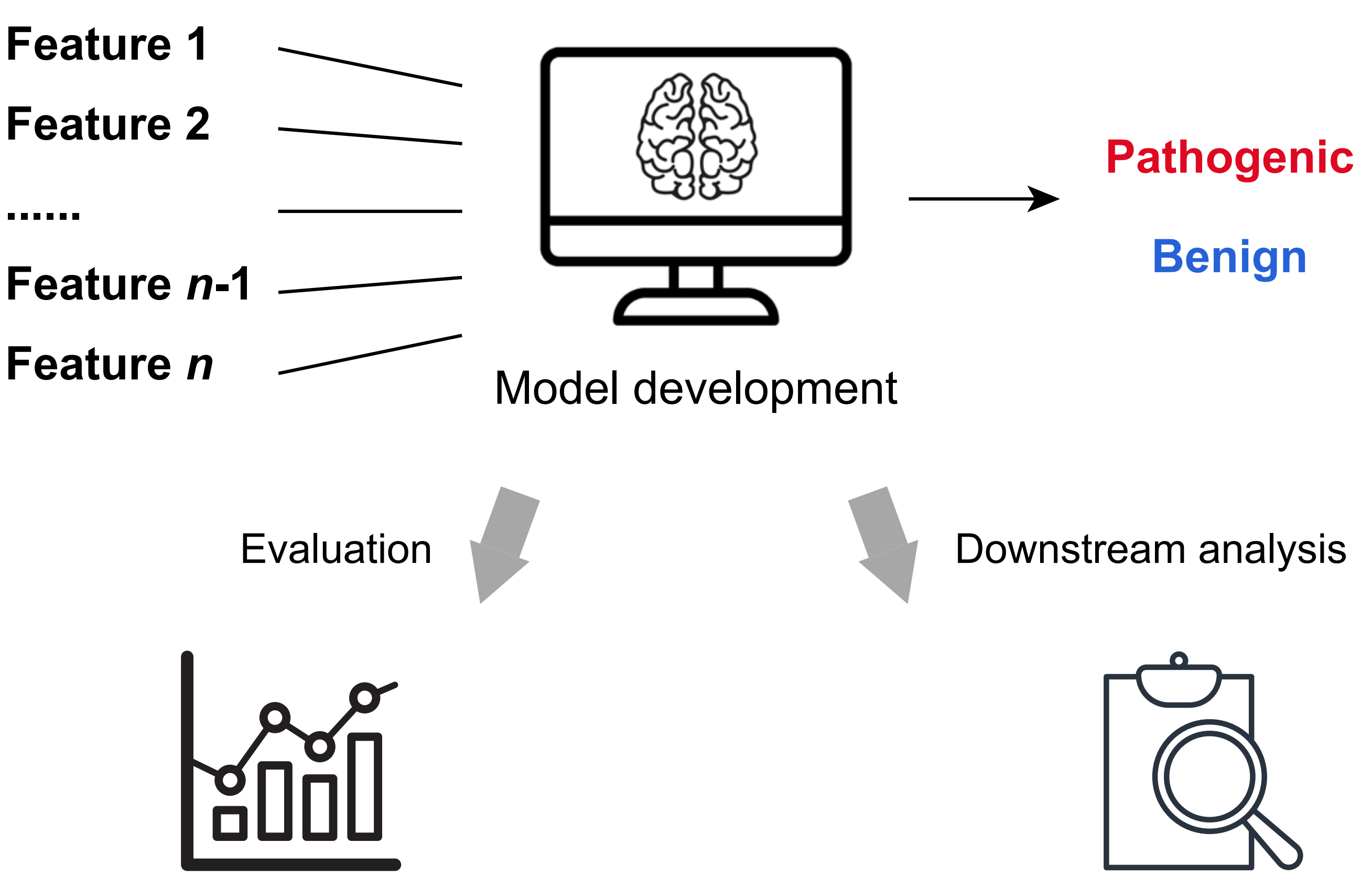


aliphatic long polar chain-break

… C Q E V Y P E L Q I T N V V E A N Q P V T …

Neighbour amino acid frequency

Machine learning



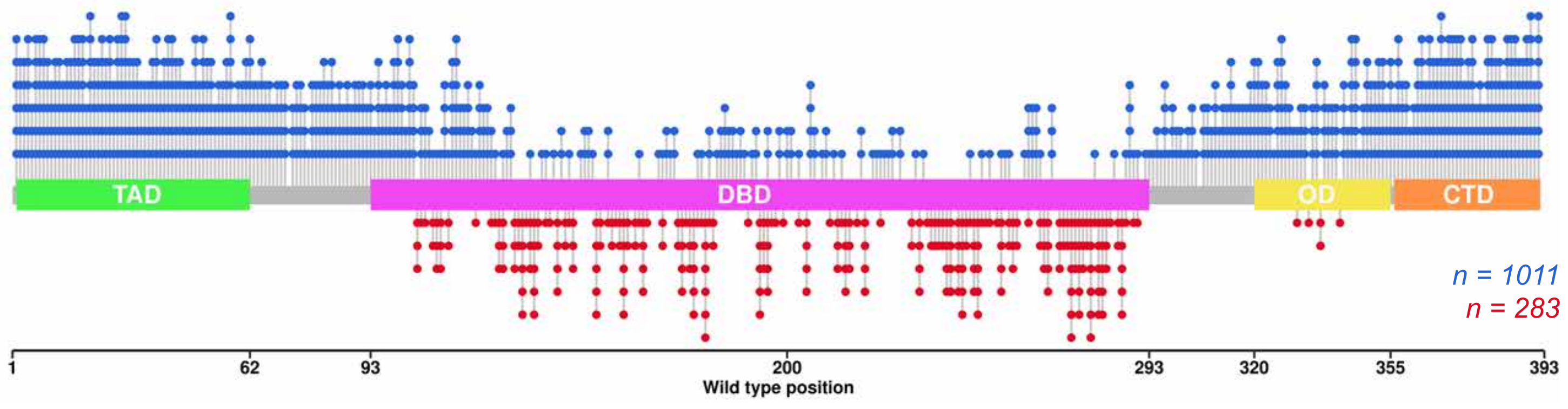
- **Blind test**
- **Clinical validation**

- **Feature interpretation**
- **Saturation mutagenesis**

CASE 1: mutations leading to cancer

- **Over 50% of cancers** are associated with the missense mutations in tumour suppressor protein p53.
- p53 plays a crucial role in DNA damage-induced activation by repairing erroneous replication and activating cellular apoptosis.

Mutation distribution of p53



- **Non-functional mutations** mainly distrbuted in the **DNA-binding domain** and sporadically distributed in the **Oligomerization domain**.

Blind test

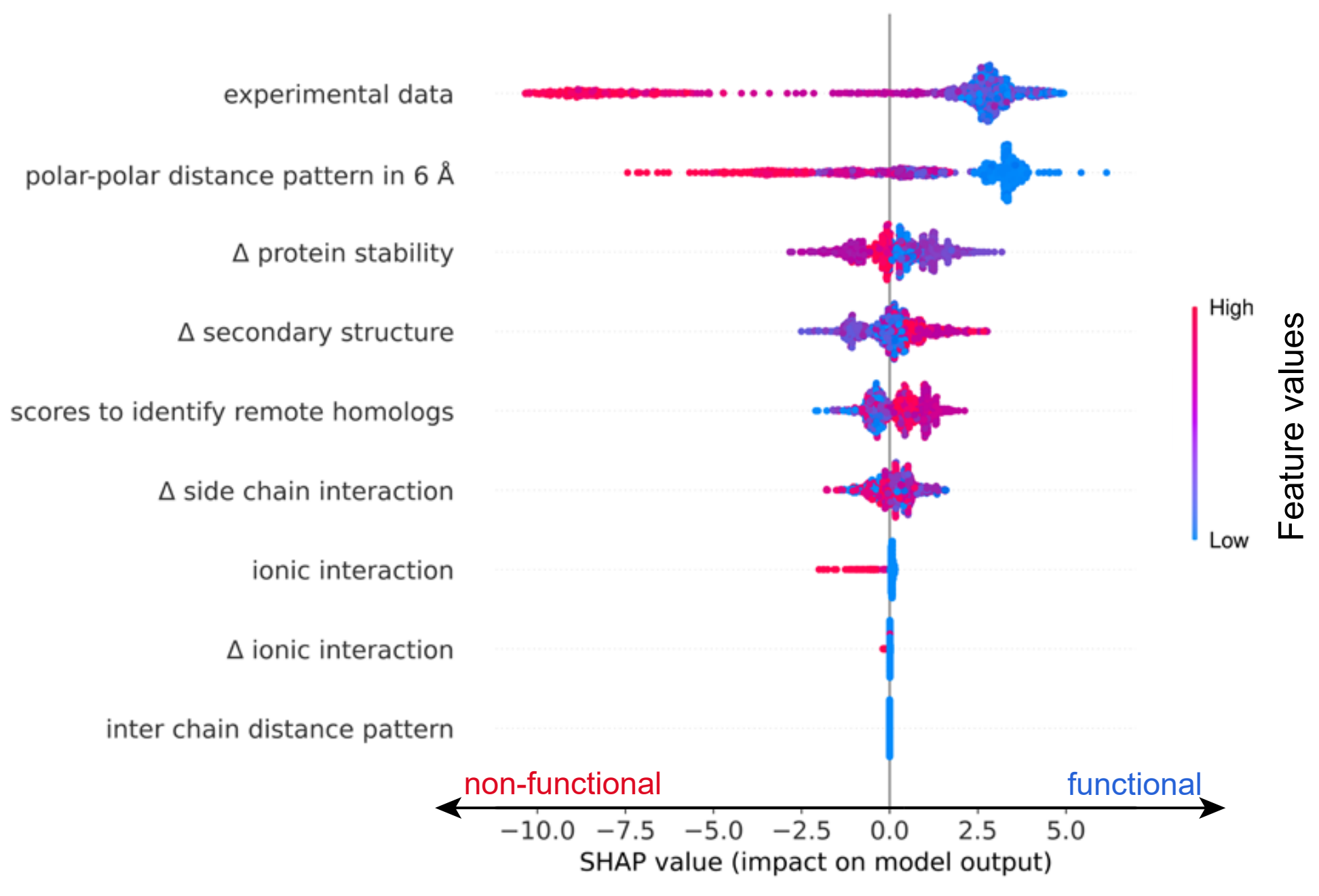
Method	MCC	Recall	Precision
our Model	***0.88***	***0.99***	***0.96***
TP53_PROF	**0.88**	**0.91**	**0.90**
Envision	0.60	0.95	0.89
VARITY	0.72	0.91	0.96
SIFT	0.46	0.59	0.98
PolyPhen2	0.42	0.54	0.98

Clinical validation

Method	MCC	Recall	Precision
our Model	***0.83***	***1.00***	***0.93***
TP53_PROF	**0.83**	**1.00**	**0.93**
Envision	0.58	0.96	0.88
VARITY	0.78	0.98	0.93
SIFT	0.58	0.75	0.98
PolyPhen2	0.47	0.64	0.97

MCC: Matthew's Correlation Coefficient / Recall: True positive rate / Precision:1 - False discovery rate

- Our model showed **robust performance** on both blind test and clinical vaidation.

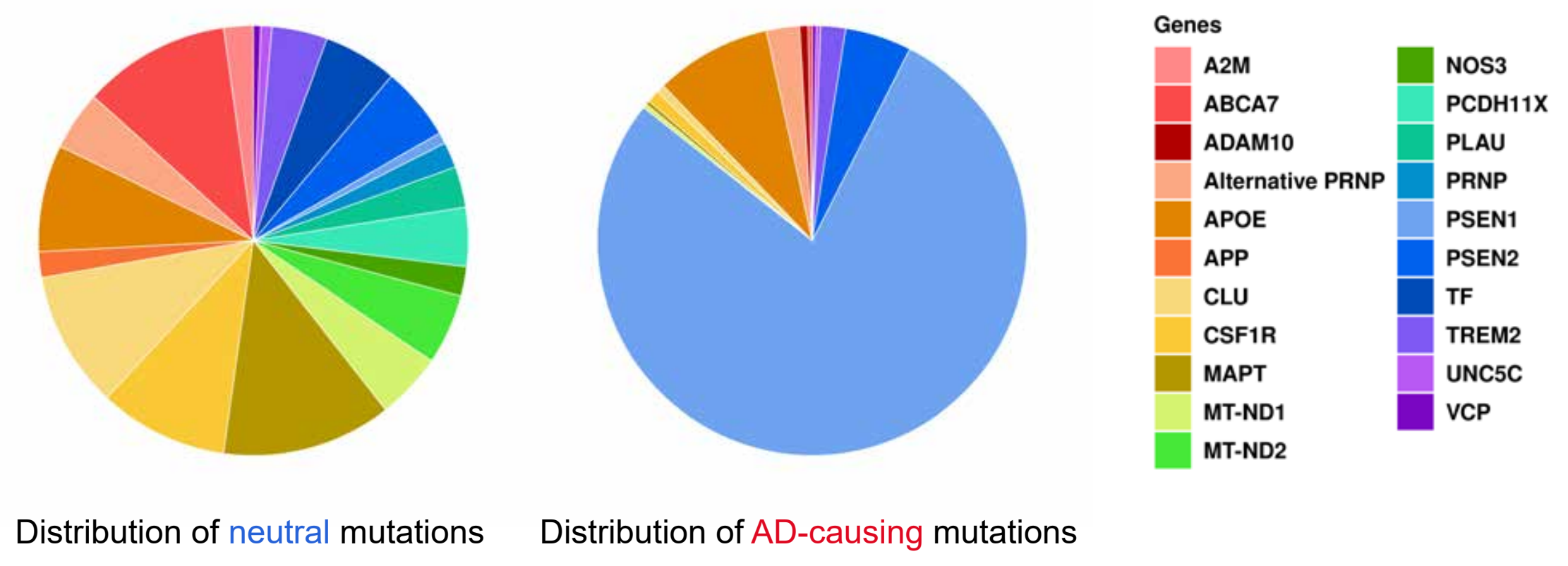


- Feature interpretation reveals that intact p53 function is strongly reliant on **experimental residue activity, the number of polar-polar atom pairs within 6 Å**, and **the change of protein stability upon mutation**

CASE 2: mutations leading to Alzheimer's Disease

- **Alzheimer's Disease (AD)** is one of the most common **neurodegenerative disease**.

Mutation distribution of a multi-gene dataset



- **680 missense mutations of 21 proteins related to AD** were collected from different databases.
- Mutation with different phenotypes were **highly imbalanced** in protein level.

Machine learning analysis

Generic performance

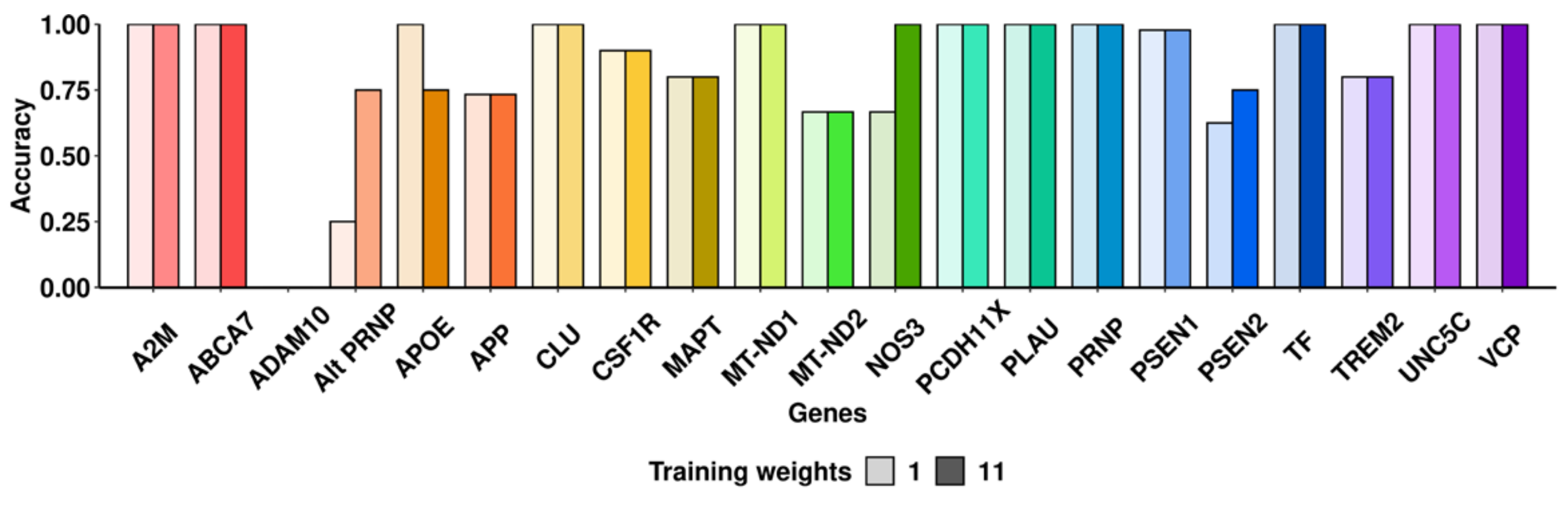
Predicted	Actual: Neutral	Actual: AD-causing
Neutral	63	9
AD-causing	8	56

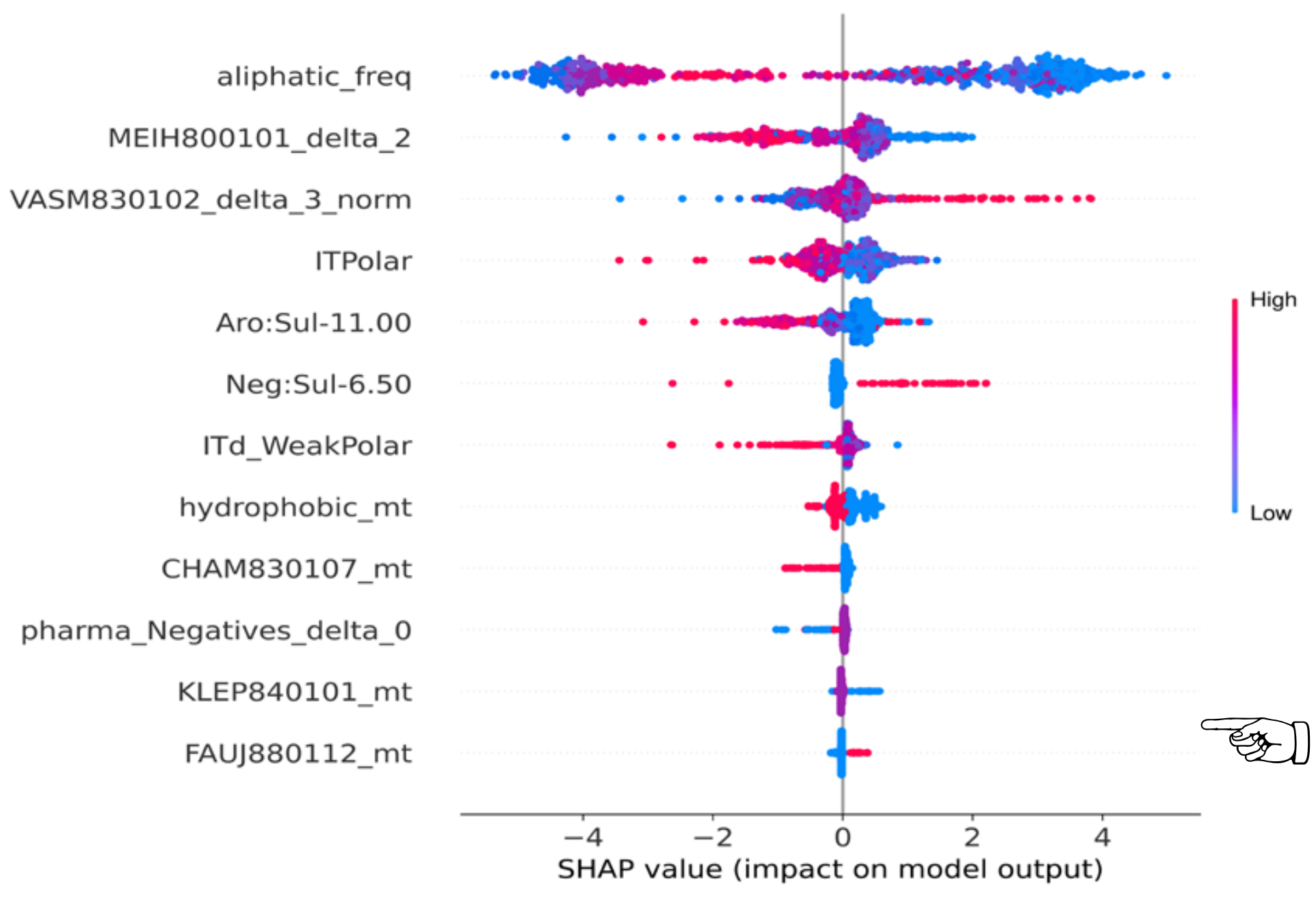
Raw MCC = 0.75

Predicted	Actual: Neutral	Actual: AD-causing
Neutral	65	8
AD-causing	6	57

Weighted MCC = 0.79

Protein-specifc performance





- By tunning the **sample weights** during the training process, we improved the ability of identify pathogenic mutations into **protein-specific level**.
- Feature interpretation presented **both the sequenced and structural residue environment**, **residue interaction**, and **the properties of mutant** are essential to the risk of Alzheimer's Disease.

CONCLUSION

- By integrating structure-based features, our models **accurately characterise** the oncogenic effects of **all possible missense mutations** in p53 and identify missense mutations increasing risks of Alzheimer's Disease, with a comparable performance to state-of-the-art methods.
- The mutation analysis of p53 offers **clinical diagnostic utility**, which is crucial for patient monitoring, and the development of personalised cancer treatment.
- Our multi-gene studies on AD not only provide clinically relevant tools, but also a better foundation to understand **the protein sequence-structure-function-pathogenicity relationships**.