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

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

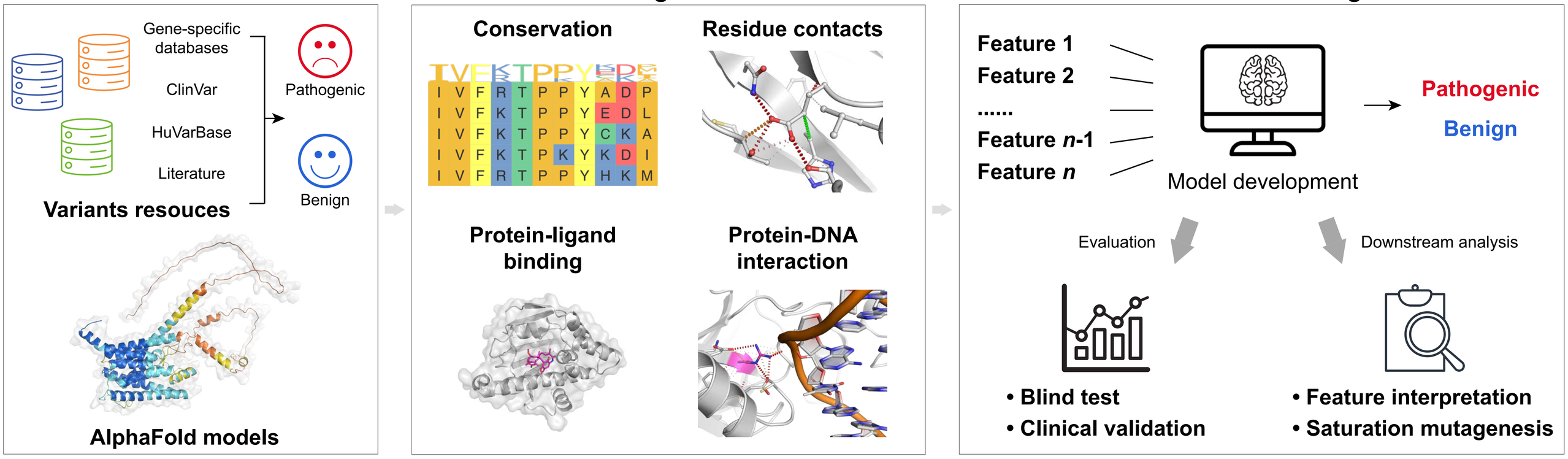
[1]*School of Chemistry and Molecular Bioscience, University of Queensland, Brisbane Queensland 4072, Australia*

[2]*Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne Victoria 3004, Australia*

BACKGROUND

- The effects of substantial genetic variant data, especially their links to diseases, are often unclear.
- Pioneering "gold standard" methods to quantify the effects of these variants rely primarily on gene/protein sequences, showing limited performance and a bias on the deleterious variants.
- Here, we present a machine learning-based approach which uses computational structural and biophysical tools to better predict clinical pathogenicity caused by missense mutations.

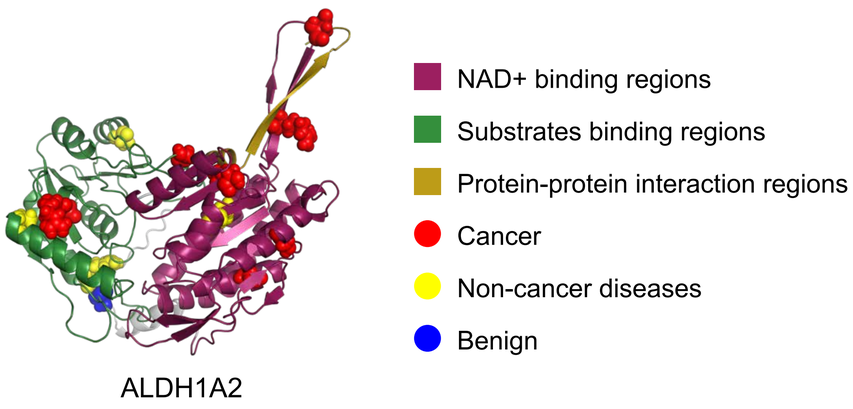
METHODOLOGY



RESULTS

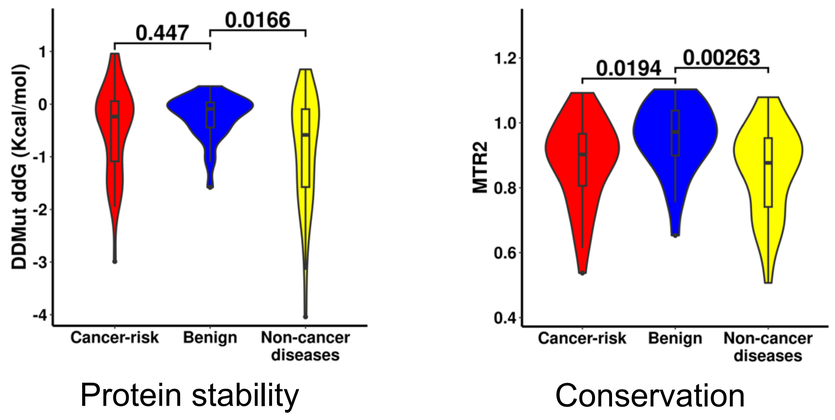
▶ Identifying molecular drivers of Aldehyde Dehydrogenase

Distribution of mutations on protein structures



ALDH1A2

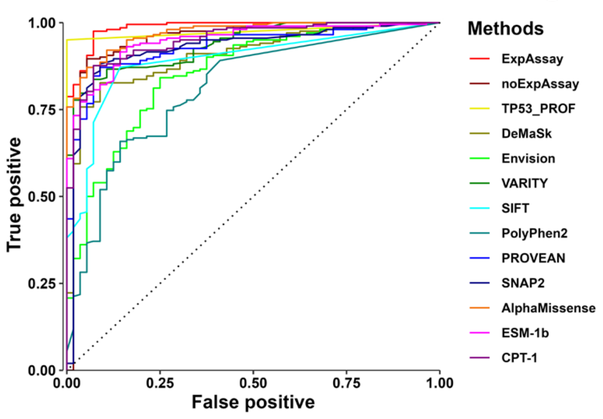
Qualitative analysis



Statistical tests indicate **the change of protein stability** and **conservation at the mutation site** are two risk factors to diseases.

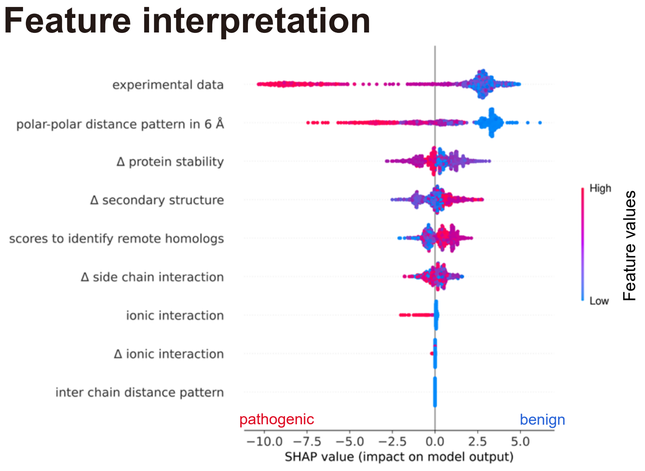
▶ Characterizing oncogenic effects of variants in p53

Performance of machine learning



	Blind test	Clinical
F1-score	0.98	0.96
MCC	0.88	0.83
Recall	0.99	1.00
Precision	0.96	0.93

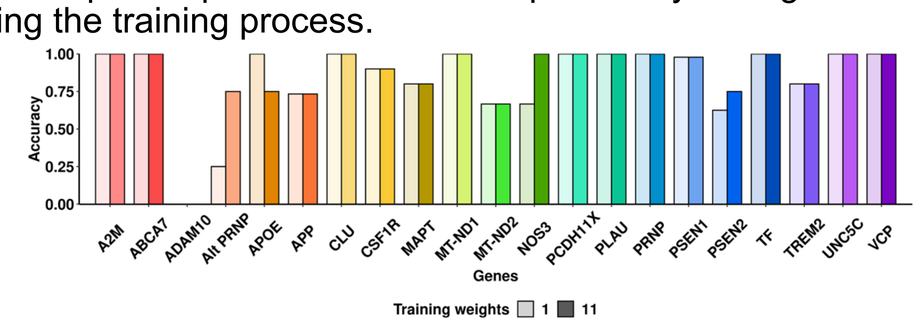
Feature interpretation



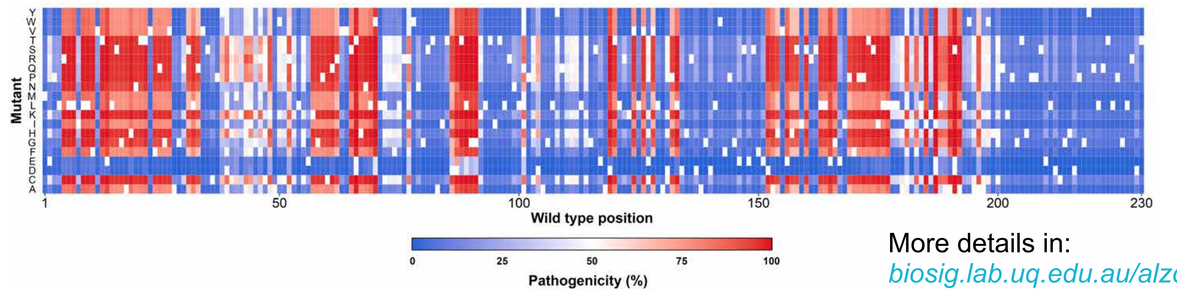
Feature interpretation reveals that intact p53 function is strongly reliant on **experimental residue activity**, **polar-polar atom pairs within 6Å of each other**, and **the change of protein stability upon mutation**.

▶ Identifying pathogenic variants leading to Alzheimer's Disease

Protein-specific performance was improved by tuning the weights during the training process.



In silico saturation mutagenesis



More details in:
biosig.lab.uq.edu.au/alzdiscovery

CONCLUSION

- Our structure-based mutation analysis can accurately characterise the pathogenic missense variants towards cancer and Alzheimer's Disease.
- Feature interpretation offers not only a better understanding on machine learning but different biological insights, such as protein stability and local mutation environment.

CONTACT

Qisheng Pan

QishengPan

qisheng.pan@uq.net.au