Systematic evaluation of computational tools to predict the effects of mutations on protein stability in the absence of experimental structures

Qisheng Pan[1,2], Thanh Binh Nguyen[1,2], David B. Ascher[1,2,3], Douglas E.V. Pires[2,3,4]

[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*
[3]*Systems and Computational Biology, Bio21 Institute, University of Melbourne, 30 Flemington Rd, Parkville VIC 3052, Australia*
[4]*School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3053, Australia*

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

BACKGROUND

- Investigating the effect of mutations on **protein thermodynamic stability** is essential to the characterisation of **genetic variants and protein engineering**.
- Over the last two decades, pioneering methods have been developed to try to estimate the effects of missense mutations on protein stability, leveraging the growing availability of protein 3D structures.
- Most of these approaches were developed and validated using experimentally derived structures and biophysical measurements, but many protein structures remain to be experimentally elucidated.
- There has been **no systematic evaluation** of the reliability of these tools **in the absence of experimental structural data**.
- To fill this gap, we therefore investigated the performance and robustness of ten widely used structural methods using homology models and the AlphaFold2 structures

METHODOLOGY

Datasets building

Target

>protein_sequence
AKLNESHRLHQFFR
DMDDEESWIKEKKL
LVSSEDKHKRLEAE
LAAHEPAIQGVLDT
GKKLSDDNTIGKEE
IQQRLAQFVDHWK
ELKQLAAARGQRLE

Templates

Identity
15% - 25%
25% - 35%
.....
.....
75% - 85%
85% - 95%

Homology models

AlphaFold models

In-silico prediction

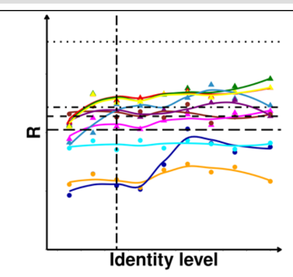
Ten widely used structure-based methods:

- Energy-based and dynamics: FoldX, ENCoM
- Knowledge-based and statistical: SDM, DDGun
- Machine learining: I-Mutant 2.0, MAESTRO, mCSM-Stability, DUET, DynaMut1, DynaMut2

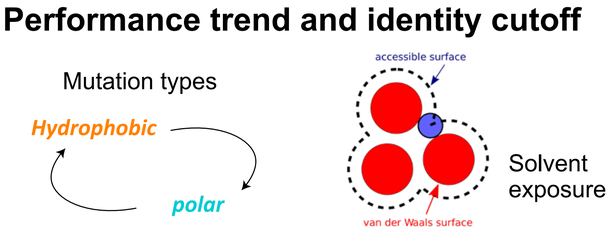
DUET · FoldX · I-Mutant · MAESTRO · DynaMut

Sequence-based methods (baseline)

Performance analysis



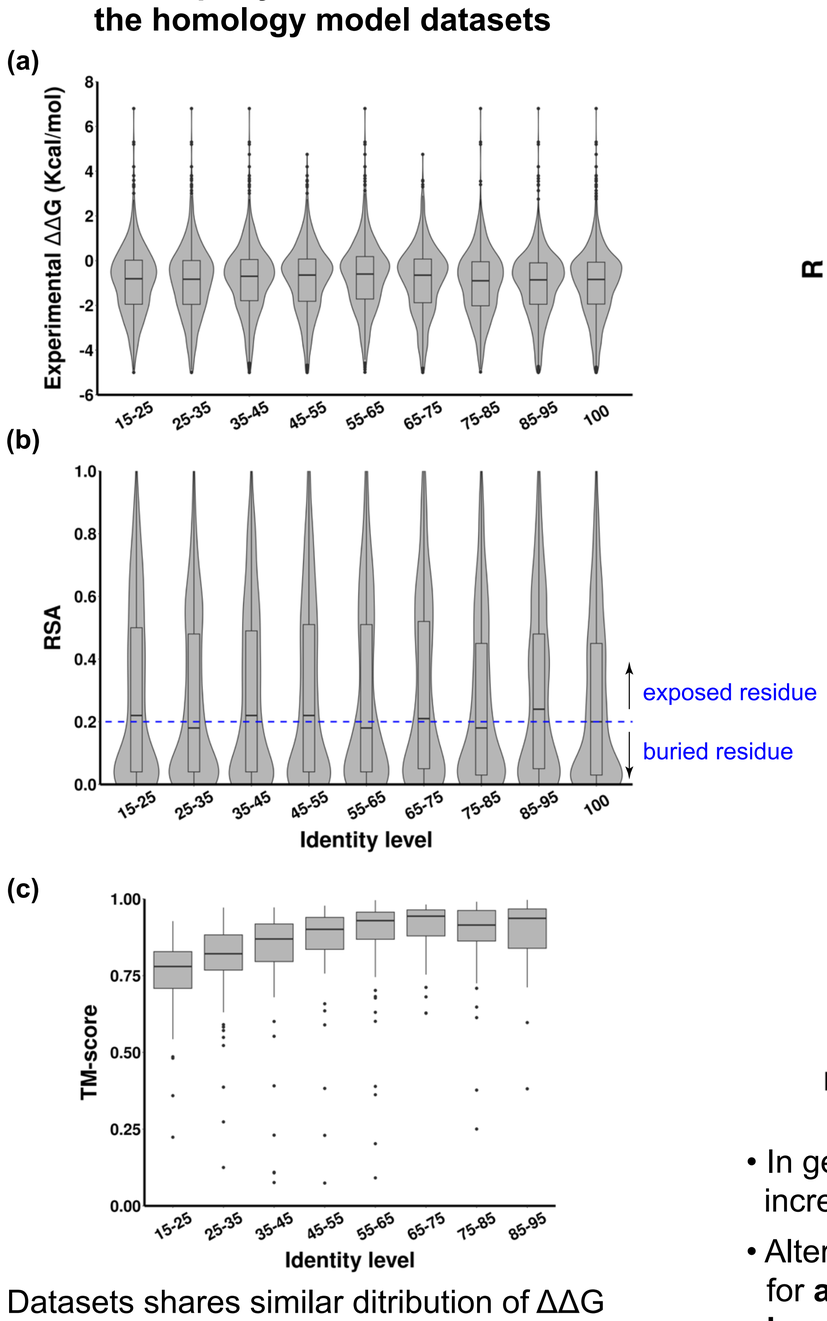
Performance trend and identity cutoff



Detailed analysis: sequenced and structural properties

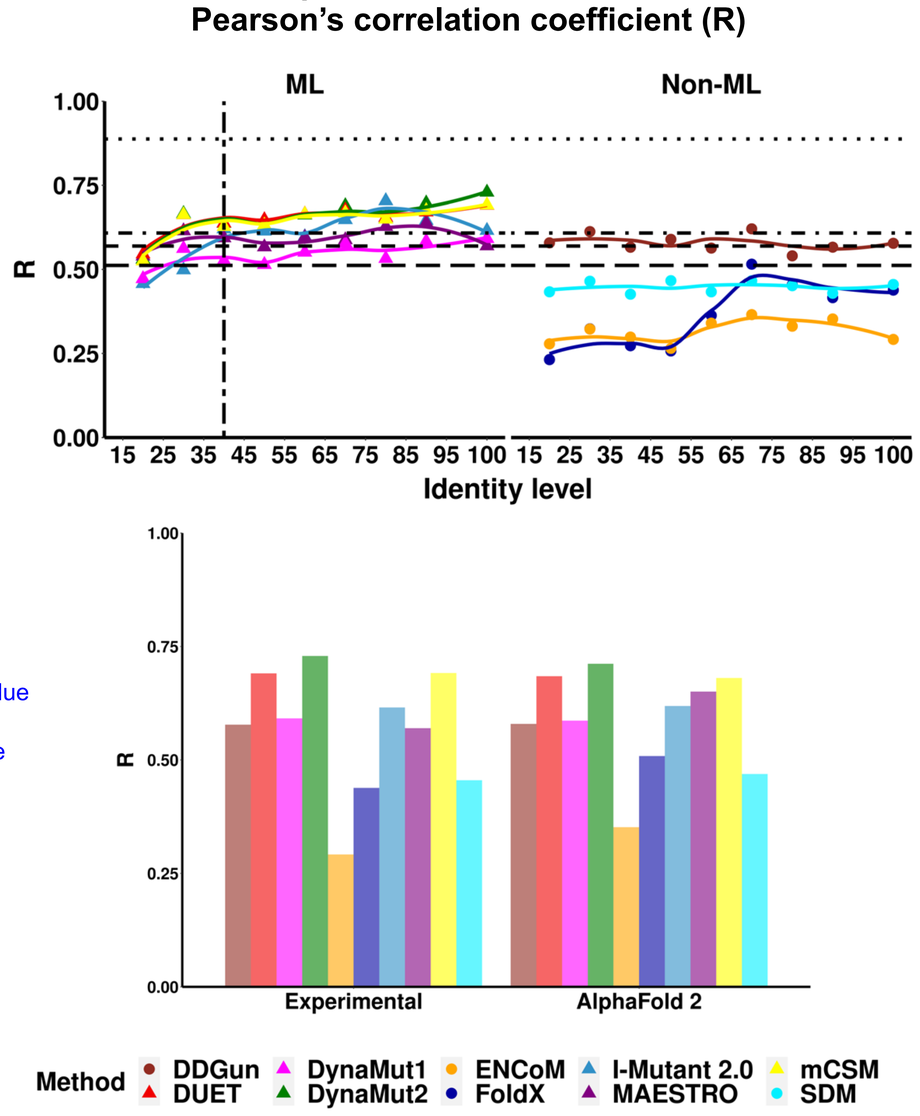
RESULTS

Property distribution of the homology model datasets



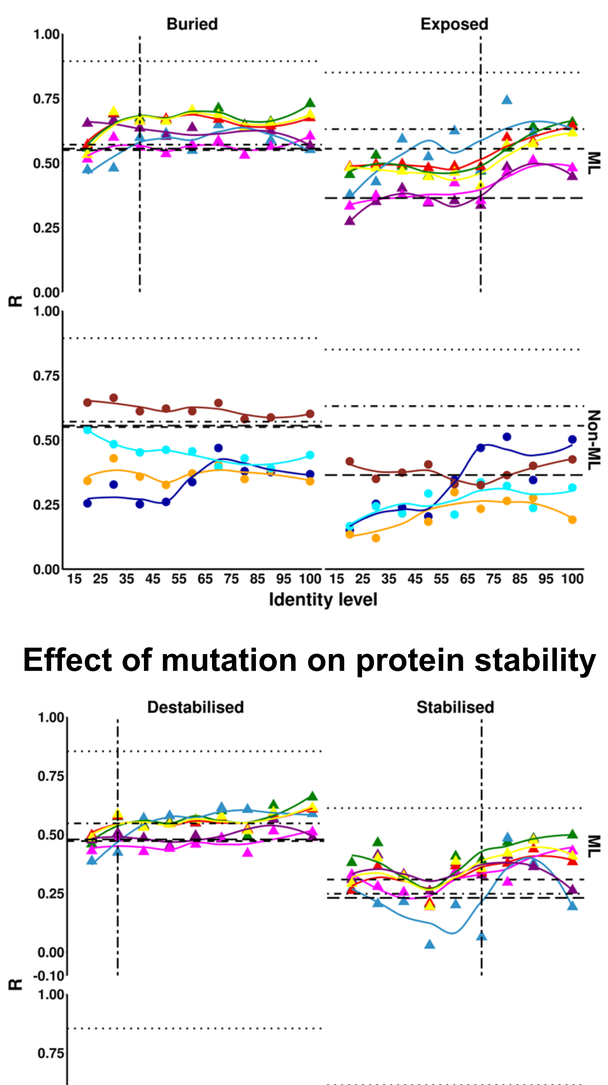
- Datasets shares similar ditribution of ΔΔG values (a) and solvent accessibility (b).
- The higher the target-template identity is, the better quality of homology models are (c).

Overall performance trends based on Pearson's correlation coefficient (R)

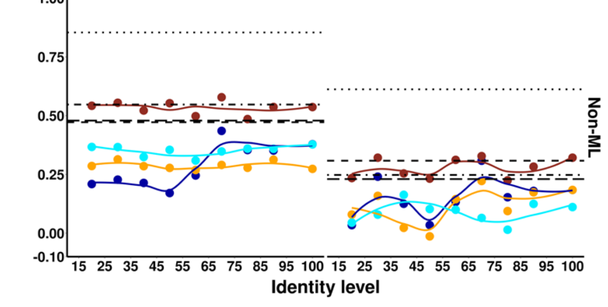


- In general, the predictive performance of the evaluated methods increases with target-template identity.
- Alternatively, we observed **a consistent performance deterioration** for **all structure-based methods**, particularly in **machine learning based methods** and FoldX, when the sequence identity of the homology modelling template dropped.
- Performance of most methods on AlphaFold2 models is close to those obtained on experimental structures.

Buried vs Exposed



Effect of mutation on protein stability



- A larger performance deterioration can be observed on the prediction of **buried residues** and **stabilising mutations**.

CONCLUSION

- Considering the consistent performance deterioration for the structure-based methods, we suggest **a target-template identity cutoff of 40%** for homology modelling when users base their conclusions in the absence of experimental structures, which differs from the conventional standard (30%).
- This work provides **a detailed guideline** for *in silico* mutation analysis, which will assist users in **appropriately using and interpreting prediction results**, and offer supports in the study of mutations in protein design and in genetic diseases.