Improving enzyme engineering through learning conditional functional landscapes

Sergio Garcia Busto[1,2,3*], Ethan Eschbach[1*], Thomas A. Hopf[4], Ben Lehner[2], Debora S. Marks[1]

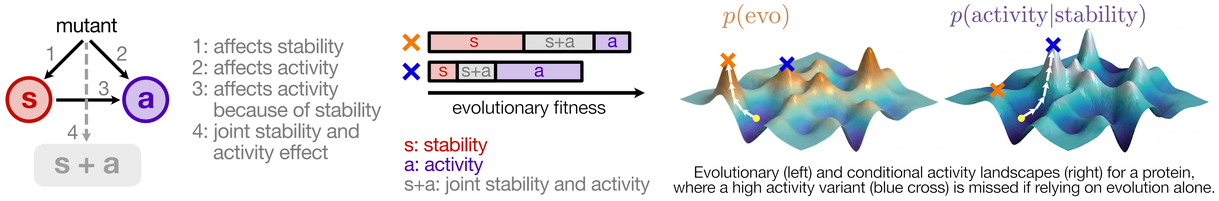
[1]Department of Systems Biology, Harvard Medical School, Boston, USA
[2]Wellcome Sanger Institute, Cambridge, UK
[3]University of Cambridge, Cambridge, UK
[4]Thomas Hopf Scientific Consulting, Bad Schönborn, Germany
*These authors contributed equally to this work.

Background

Evolutionary models learn a **convolution of protein functions,** including **stability and activity**

But for **enzyme design**, we want to sample the ***functional* landscape**

Optimising evolutionary fitness alone favours variants that are active *because* they are stable



Evolutionary (left) and conditional activity landscapes (right) for a protein, where a high activity variant (blue cross) is missed if relying on evolution alone.

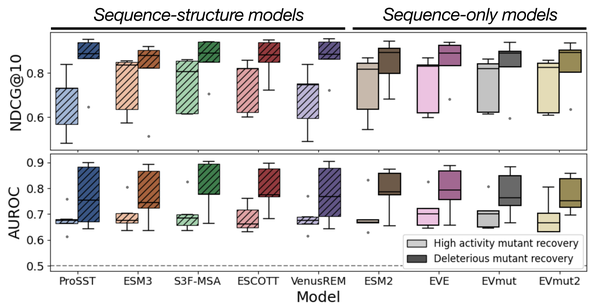
Aim: control for stability effects using structure to learn a *conditional* distribution of protein function

Multimodal approaches do not learn conditional activity

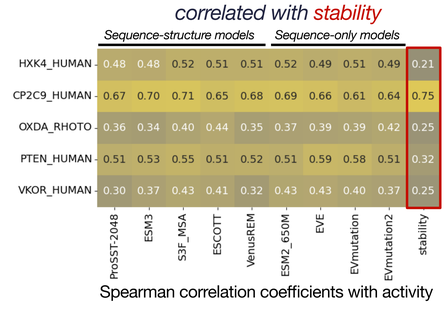
We used a panel of 5 well-characterised **enzymes with multidimensional deep mutational scans** to inspect the ability of leading multimodal and sequence-based approaches to predict *unconditional* (raw) and *conditional* activity.

How well do models capture activity?

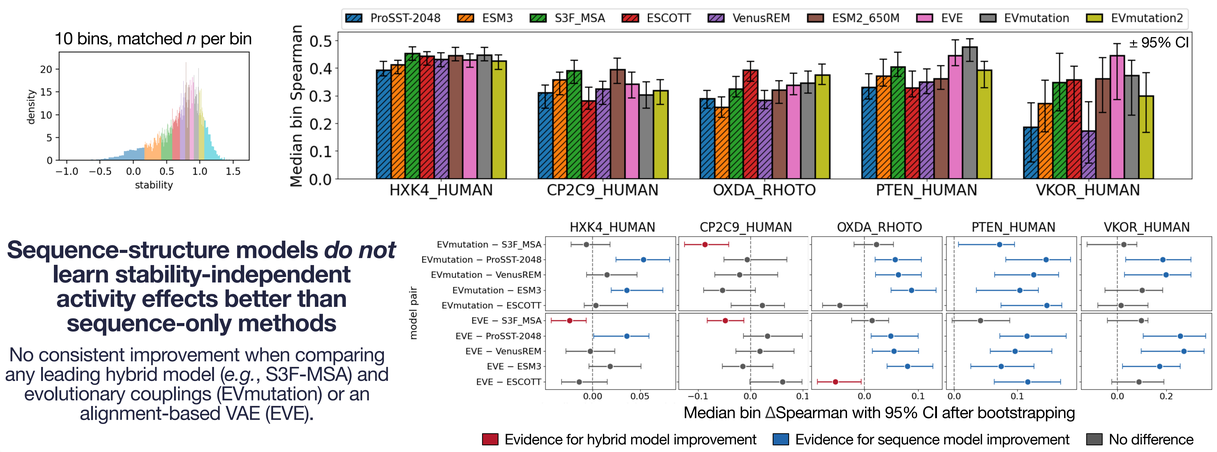
All models capture the top and bottom 10% of activity similarly



Activity is predicted best when correlated with stability



Conditional activity: at similar stability regimes, how well do models recover activity ranks?

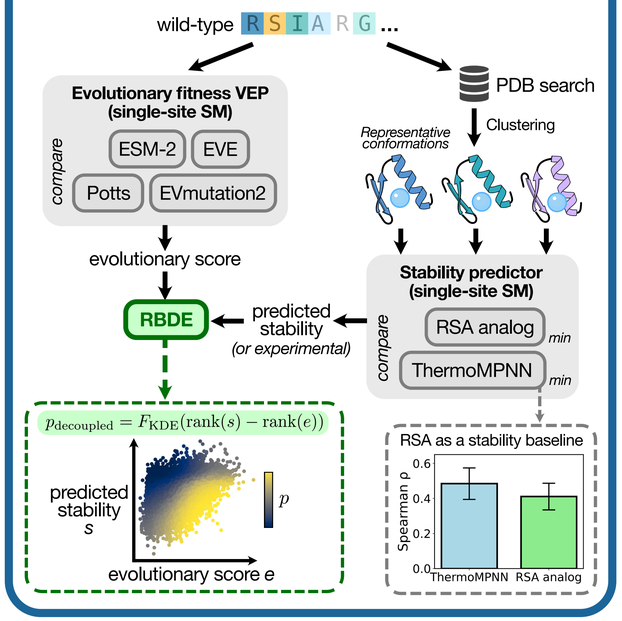


Sequence-structure models *do not* learn stability-independent activity effects better than sequence-only methods

No consistent improvement when comparing any leading hybrid model (*e.g.*, S3F-MSA) and evolutionary couplings (EVmutation) or an alignment-based VAE (EVE).

Investigating the signal with an independent predictor

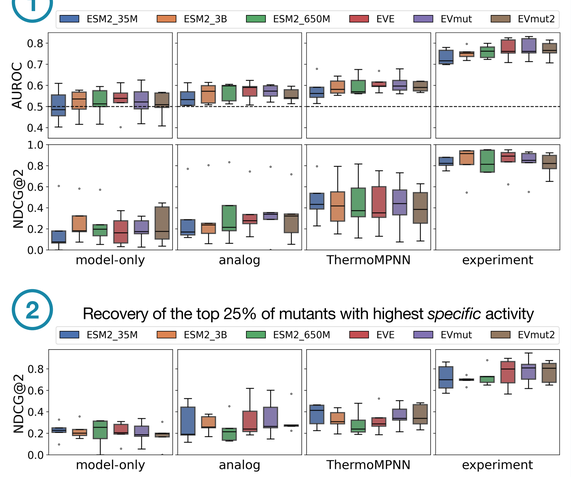
By using **sequence-based evolutionary models** and **stability predictors**, we investigated the *decoupled* signal through building *joint* models via a **rank-based density estimator (RBDE).**



Evaluation: recovery of decoupled activity effects

How well do joint models recover and rank the *most decoupled* mutants?

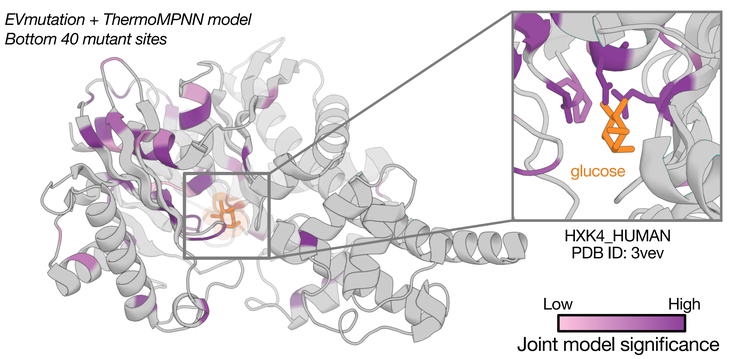
1. **Stable-but-inactive** as defined by WT ranges
2. High **specific activity** $a_{\text{specific}} = z(a) - z(s)$



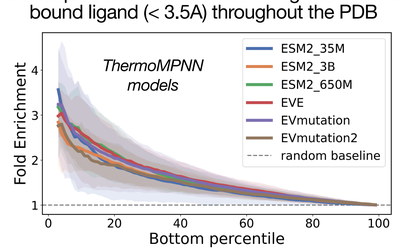
Results using **computational stability predictors differ from experiments**, but appear superior to using evolutionary models alone.

The decoupled signal can be used to identify functional sites

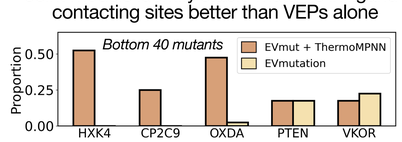
Stability-conditioned fitness scores from joint models correspond to sites that are **distributed throughout the protein** and are **enriched for ligand binding sites.**



Decoupled scores allow recovering sites with a bound ligand (< 3.5A) throughout the PDB



Joint models identify *stable* mutants in ligand-contacting sites better than VEPs alone



Conclusions

1. Multimodal approaches **do not capture stability-independent enzyme activity** effects better than sequence-only models
2. **Independent predictors can be used to deconvolve activity effects**, but results differ from using experimental stability
3. Conditioning on stability allows **identifying sites that modulate activity** and **bind relevant ligands**
4. Building sequence-structure models with better inductive biases will enable learning functional landscapes for design

PDF with references

References

Data

Sarah Gersing, Matteo Cagiada, Marinella Gebbia, et al. A comprehensive map of human glucokinase variant activity. Genome Biology, 24(1):97, 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02935-8.
Sarah Gersing, Thea K Schulze, Matteo Cagiada, et al. Characterizing glucokinase variant mechanisms using a multiplexed abundance assay. Genome Biology, 25(1):98, 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03238-2.
Clara J Amorosi, Melissa A Chiasson, Matthew G McDonald,et al. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. The American Journal of Human Genetics, 108(9):1735–1751, 9 2021. ISSN0002-9297. doi: 10.1016/j.ajhg.2021.07.001.
Kenneth A Matreyek, Jason J Stephany, Ethan Ahler, et al. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. Genome Medicine, 13(1):165, 2021. ISSN 1756-994X. doi: 10.1186/s13073-021-00984-x.
Rosario Vanella, Christoph Küng, Alexandre A Schoepfer, et al. Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. Nature Communications, 15(1):1807, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45630-3.
Melissa A Chiasson, Nathan J Rollins, Jason J Stephany, et al. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. eLife, 9:e58026, 2020. ISSN 2050-084X. doi: 10.7554/eLife.58026.

Models (see ProteinGym for full list)

Pascal Notin, Aaron W Kollasch, Daniel Ritter, et al. ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. bioRxiv, page 2023.12.07.570727, 1 2023. doi: 10.1101/2023.12.07.570727.
Thomas A Hopf, John B Ingraham, Frank J Poelwijk, et al. Mutation effects predicted from sequence co-variation. Nature Biotechnology, 35(2):128–135, 2017. ISSN 1546-1696. doi: 10.1038/nbt.3769.
Jonathan Frazer, Pascal Notin, Mafalda Dias, et al. Disease variant prediction with deep generative models of evolutionary data. Nature, 599(7883):91–95, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04043-8.