Prediction of the Effect of Single Amino Acid Protein Variants Using Deep Mutational Scanning Data

University of Bologna — Master Thesis in Bioinformatics

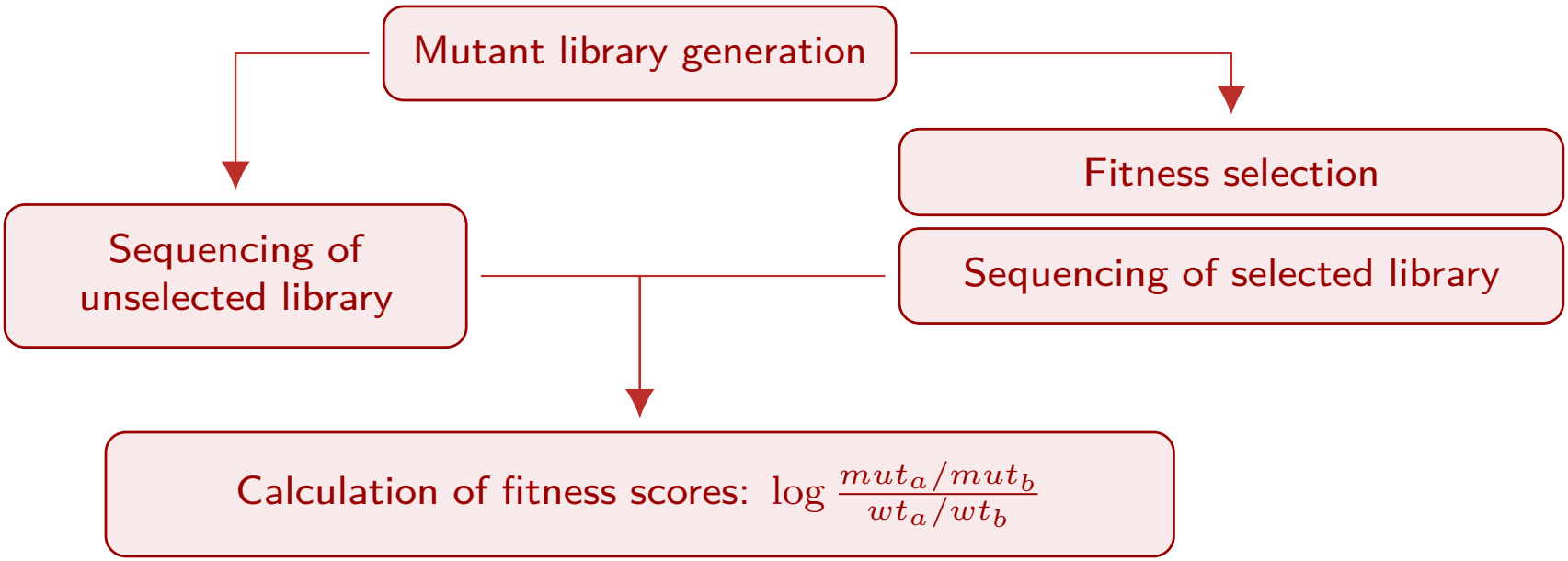
Pierotti Saul

Internal Advisor: Prof. Pietro Di Lena
External Advisor: Prof. Arne Elofsson (Stockholm University)

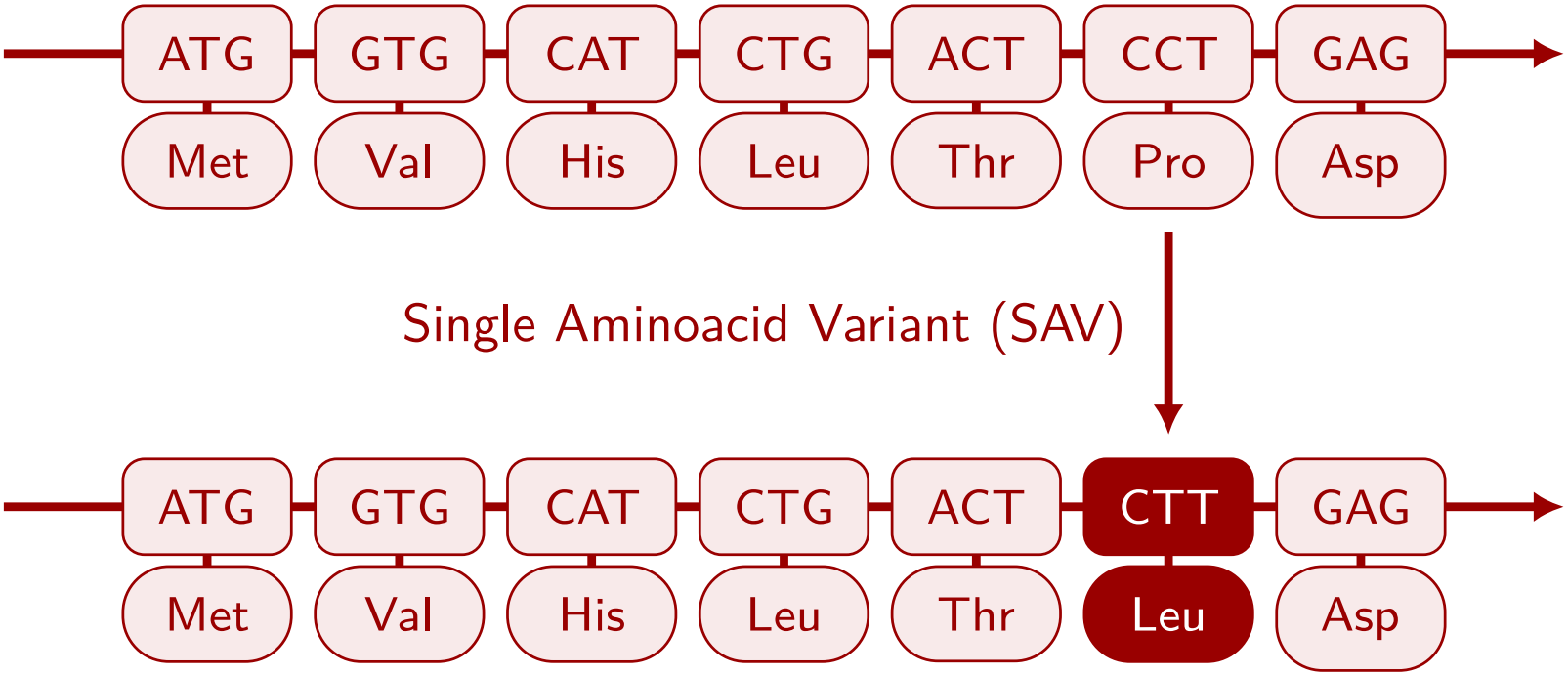
July 19, 2021

Deep mutational scanning

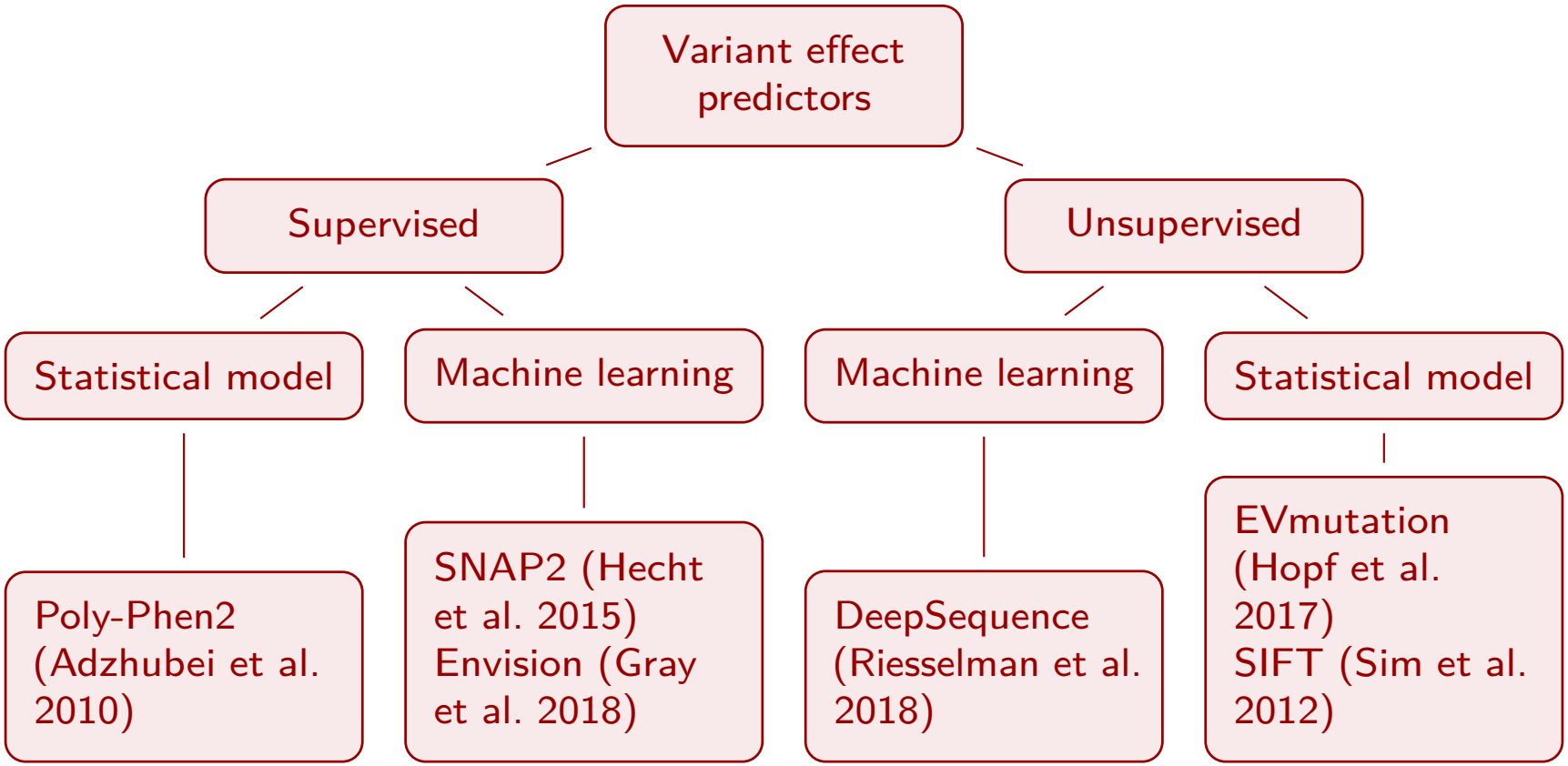
High-throughput technique for obtaining fitness information on a large number of mutations



I considered only single aminoacid variants

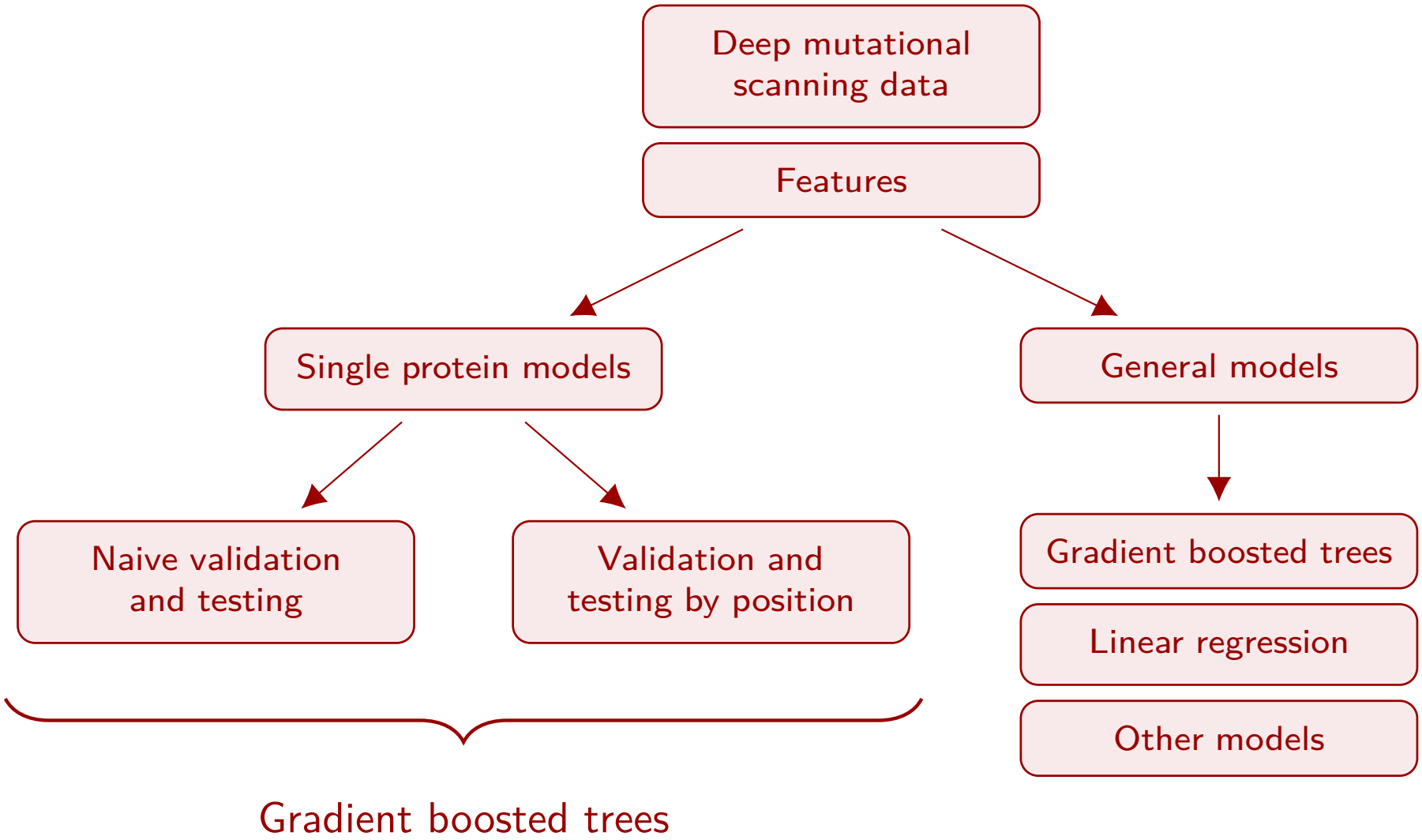


Some notable variant effect predictors

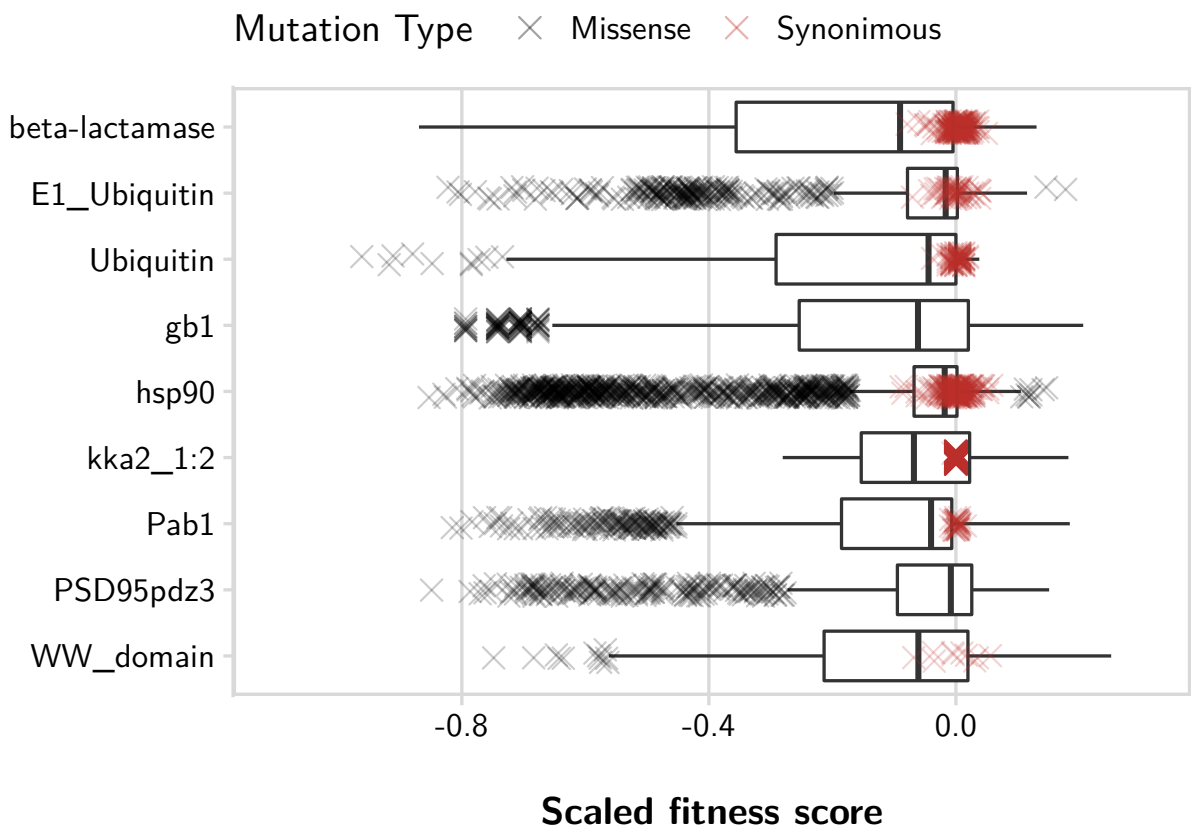


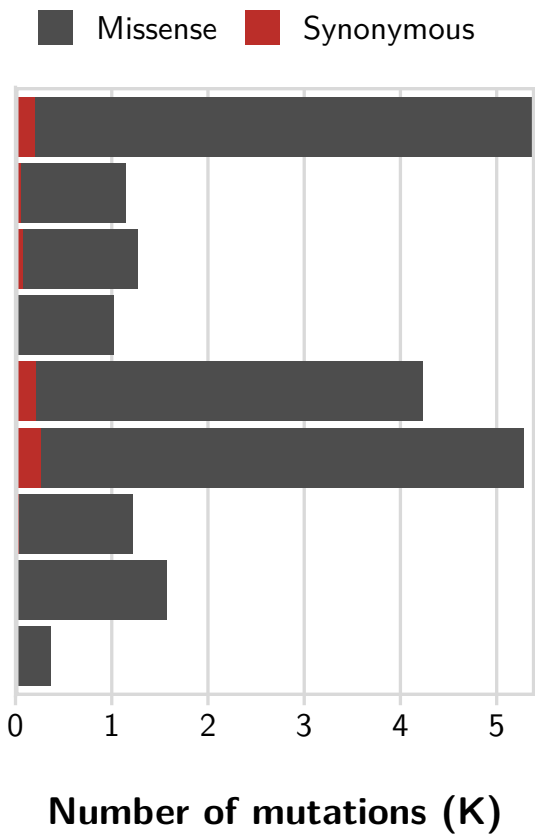
Among these predictors only Envision was trained on deep mutational scanning data

Structure of the project



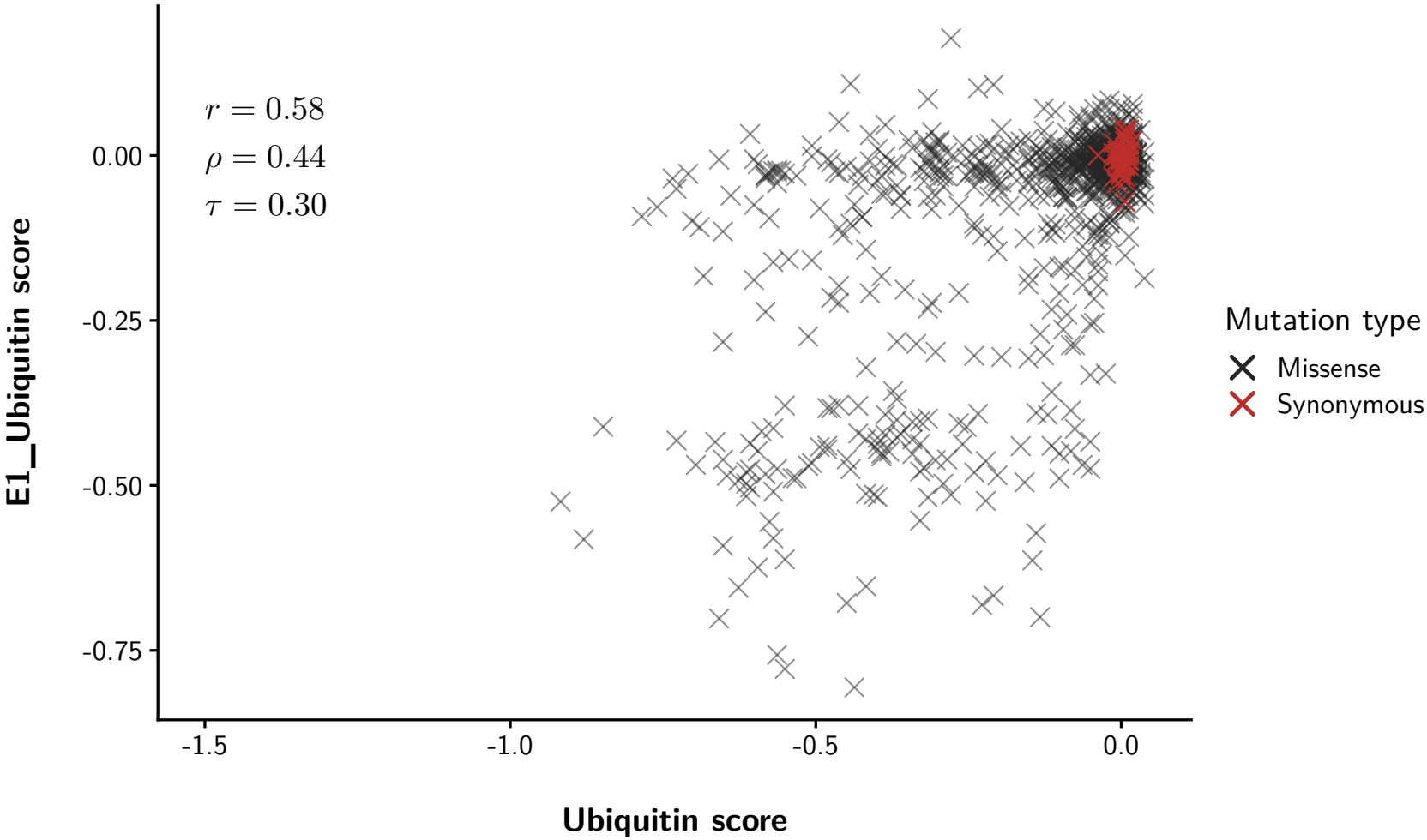
I used the training dataset of Envision (Gray et al. 2018)





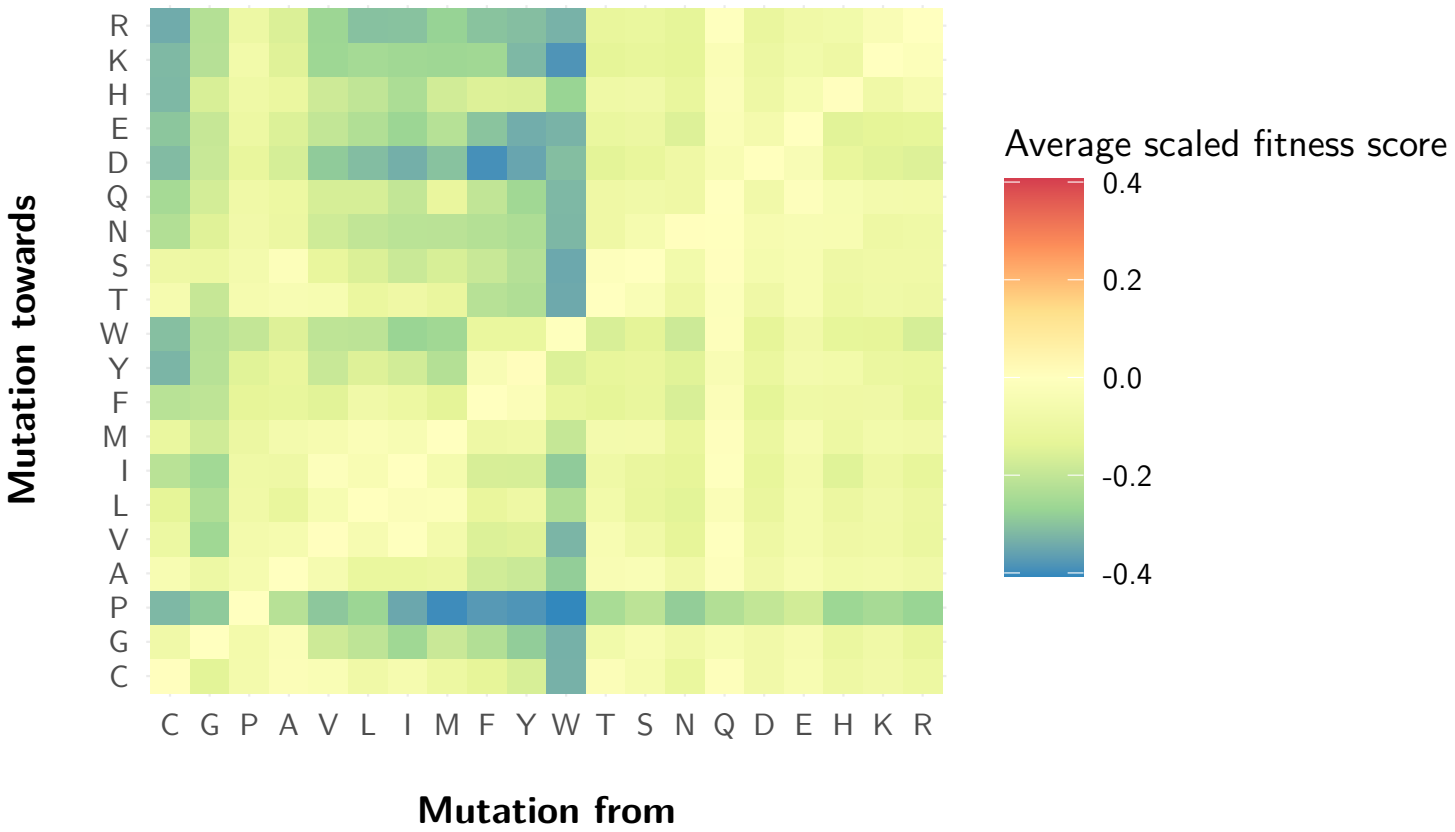
Poor correlation among experimental results

Two independent deep mutational scanning experiments on Ubiquitin are present in the training dataset. Their correlation is low.

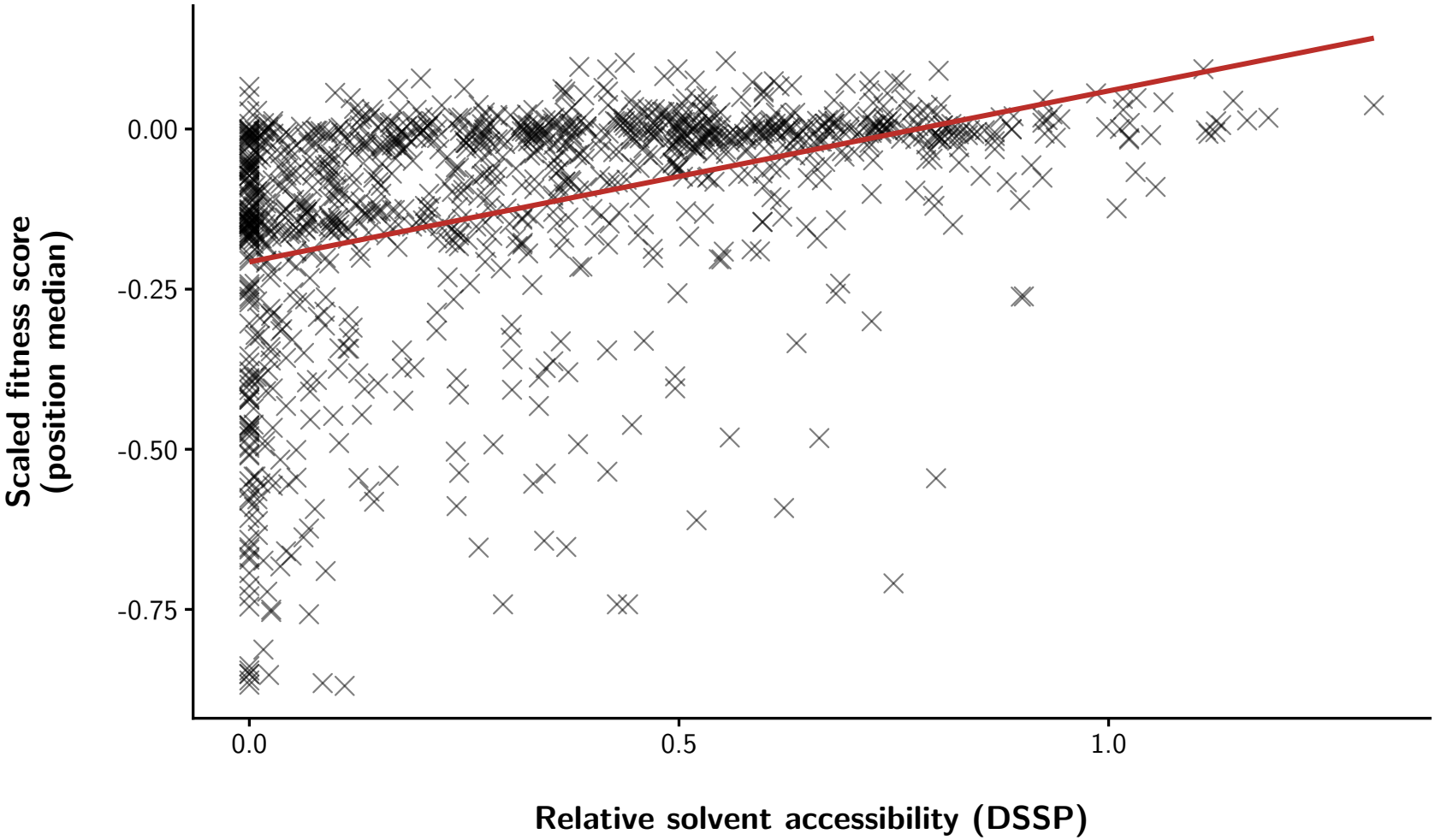


The identities of the mutated residues are important

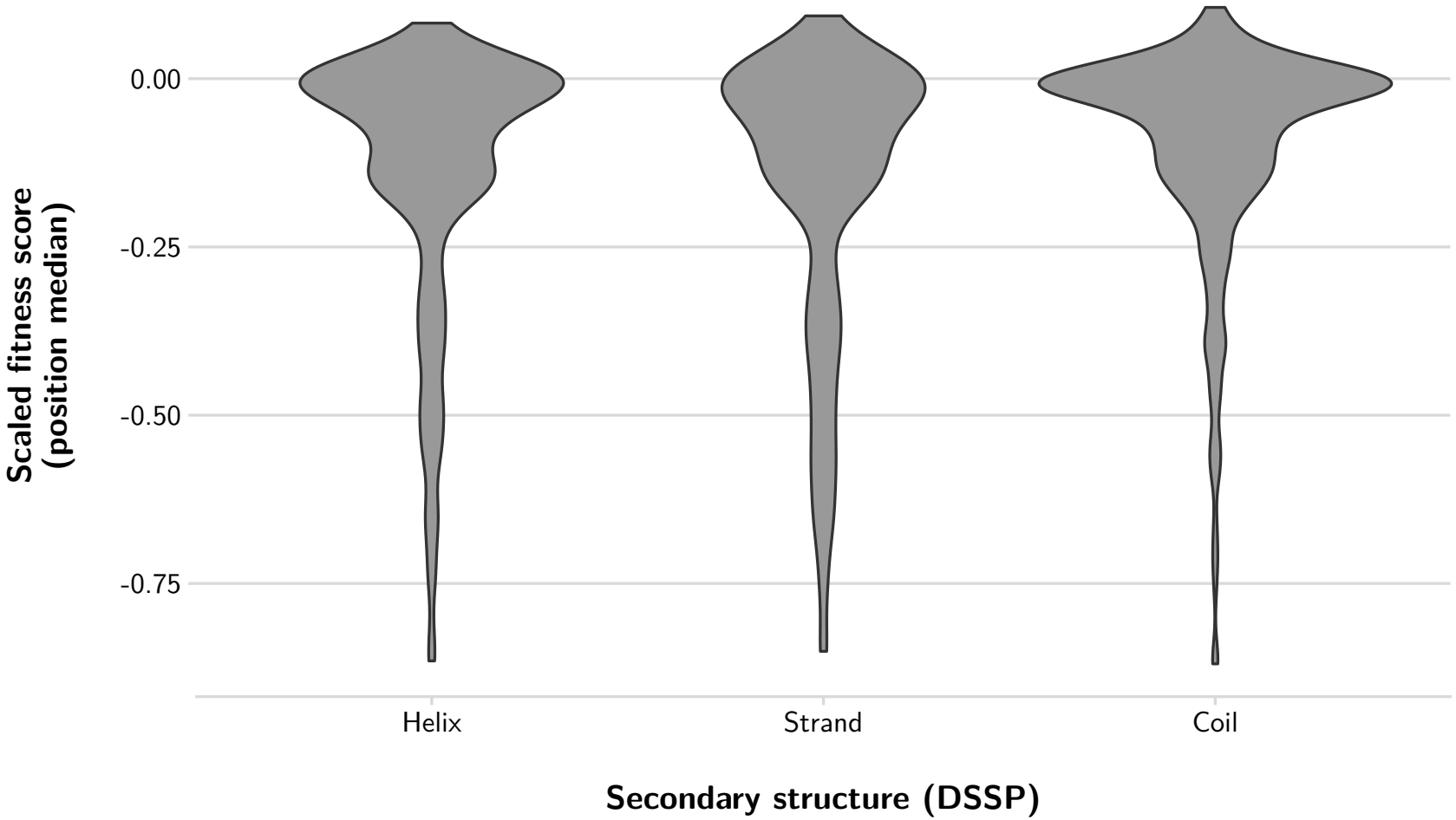
Mutations from polar residues are less detrimental than other mutations. This effect disappears when filtering by relative solvent accessibility (not shown here).



Buried residues are more conserved



Secondary structure is not critical



I did not use structural information

Mutation identity

Wild-type residue

Mutated residue

EVmutation predictions (Hopf et al. 2017)

Epistatic model

Independent model

Conservation

Mutation frequency

NetsurfP-2 predictions (Klausen et al. 2019)

Secondary structure

Solvent accessibility

Disorder

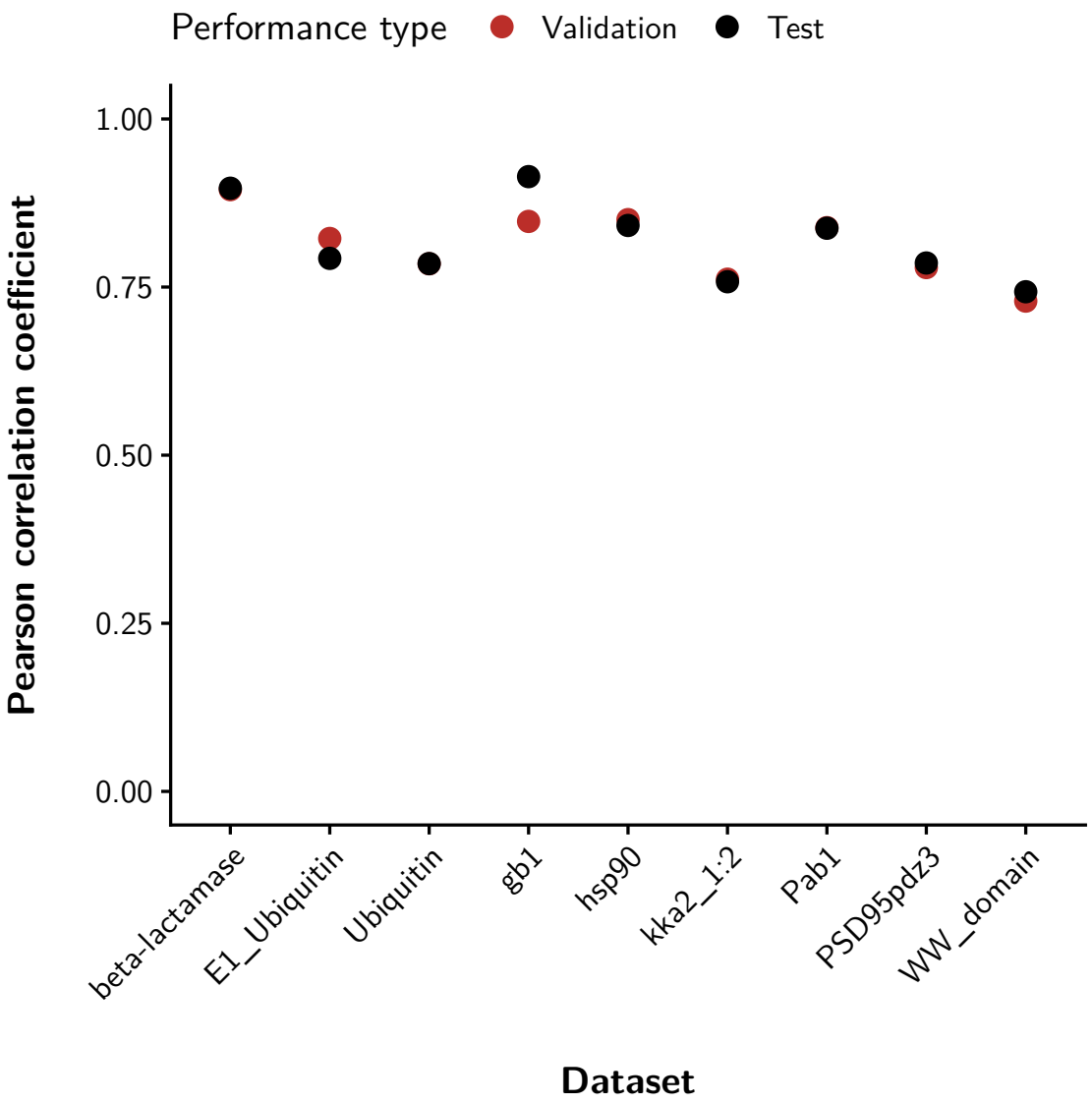
Torsion angles

HMMER (Eddy 2011) emission probabilities

trRosetta predicted contacts (Yang et al. 2020)

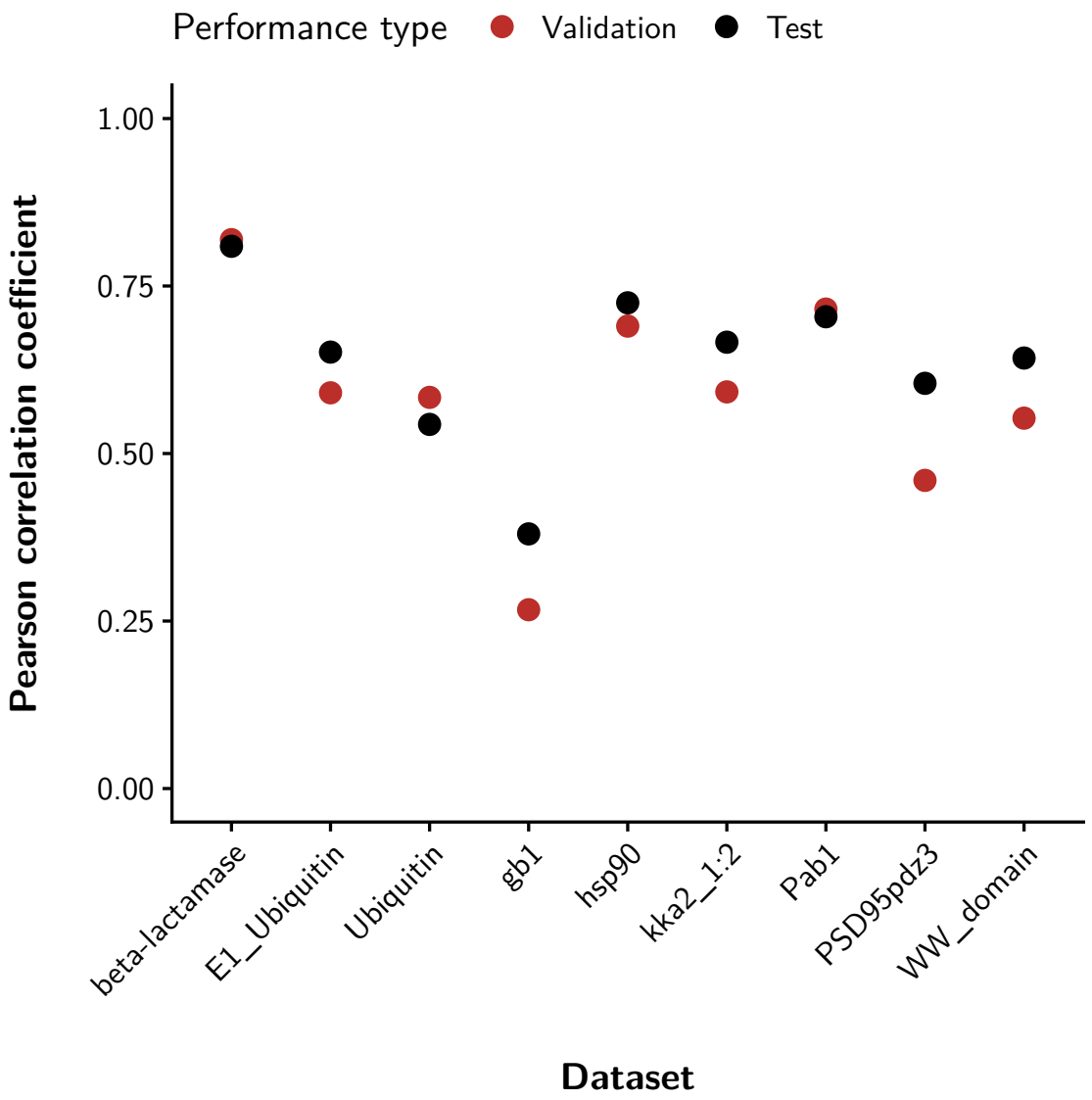
Centrality metrics

Single protein models with naive testing



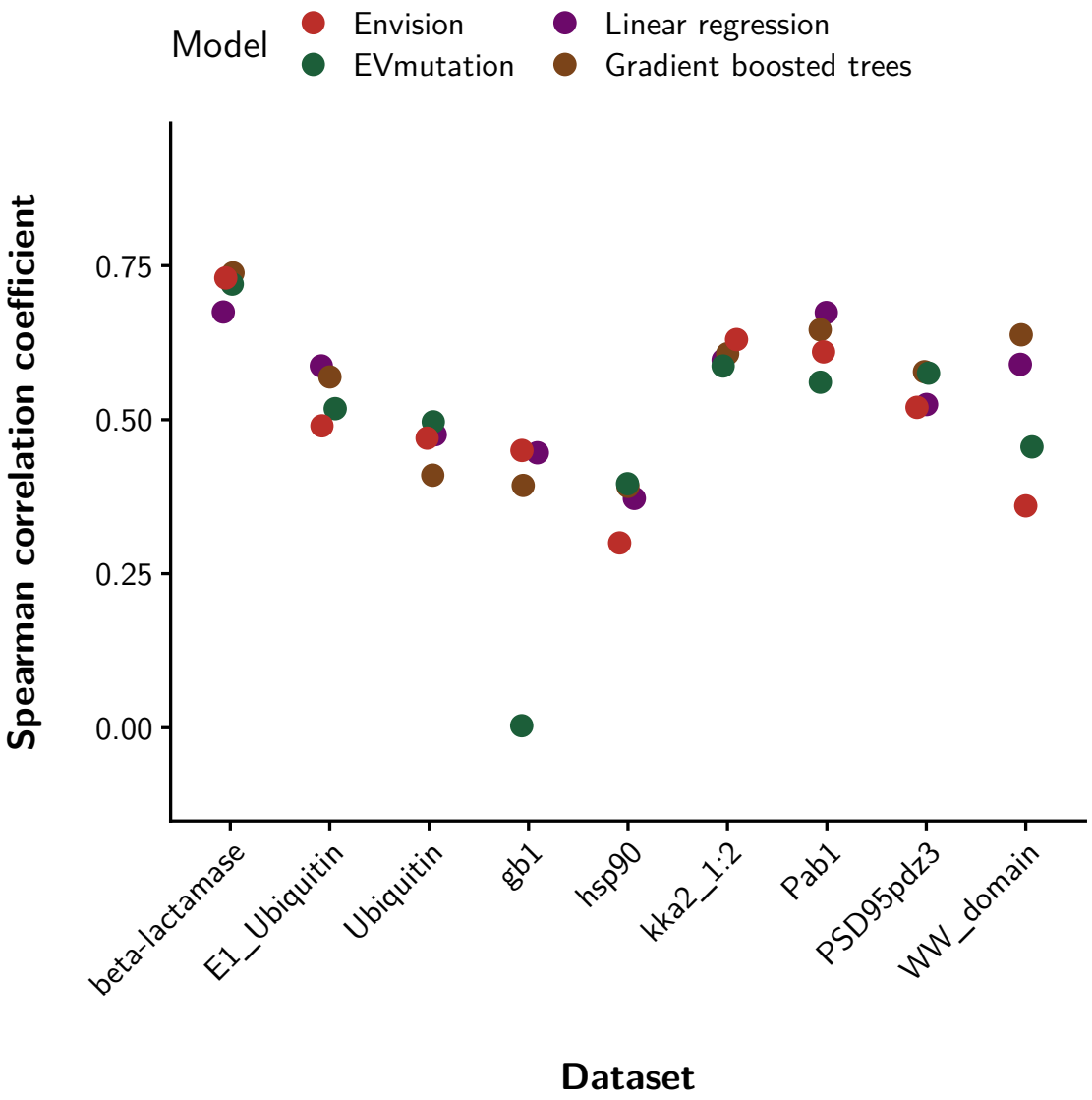
- Half of the mutations in a protein randomly set aside for testing
- Hyperparameters optimized in the remaining half with 5-fold cross-validation
- Good results but likely overfitting the testing set

Single protein models with testing by position



- Half of the mutations in a protein set aside for testing but avoiding mutations in the same protein position to end up in different splits
- Hyperparameters optimized in the remaining half with 5-fold cross-validation
- Performance more realistic

Leave-One-Protein-Out (LOPO) models



- For the left-out protein, half of the mutations used for testing
- Spearman correlation coefficient used for evaluation
- Small difference among gradient boosted trees and linear regression
- Performances comparable to those of Envision

Discussion

Complex models do not improve much on linear regression

Unsupervised models perform similarly to supervised models

There is strong variability between datasets

How validation and testing are performed is crucial

Performances on par with other predictors can be reached without structural features

Future directions

Unsupervised models seem promising and may be worth exploring more

Training on more deep mutational scanning studies

Tuning the set of features

Trying different models

Using residue contacts in a graph convolutional neural network

Finding a better normalization strategy for the scores from different experiments

Bibliography I

Adzhubei, Ivan A. et al. (Apr. 2010). 'A method and server for predicting damaging missense mutations'. In: *Nature Methods* 7.4, pp. 248–249. DOI: 10.1038/nmeth0410-248.

Eddy, Sean R. (Oct. 2011). 'Accelerated Profile HMM Searches'. In: *PLoS Computational Biology* 7.10. Ed. by William R. Pearson, e1002195. DOI: 10.1371/journal.pcbi.1002195.

Gray, Vanessa E. et al. (Jan. 2018). 'Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data'. In: *Cell Systems* 6.1, 116–124.e3. DOI: 10.1016/j.cels.2017.11.003.

Hecht, Maximilian, Yana Bromberg and Burkhard Rost (June 2015). 'Better prediction of functional effects for sequence variants'. In: *BMC Genomics* 16.S8. DOI: 10.1186/1471-2164-16-s8-s1.

Hopf, Thomas et al. (Jan. 2017). 'Mutation effects predicted from sequence co-variation'. In: *Nature Biotechnology* 35.2, pp. 128–135. DOI: 10.1038/nbt.3769.

Bibliography II

Klausen, Michael Schantz et al. (Mar. 2019). 'NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning'. In: *Proteins: Structure, Function, and Bioinformatics* 87.6, pp. 520–527. DOI: 10.1002/prot.25674.

Riesselman, Adam J., John B. Ingraham and Debora S. Marks (Sept. 2018). 'Deep generative models of genetic variation capture the effects of mutations'. In: *Nature Methods* 15.10, pp. 816–822. DOI: 10.1038/s41592-018-0138-4.

Sim, Ngak-leng et al. (June 2012). 'SIFT web server: predicting effects of amino acid substitutions on proteins'. In: *Nucleic Acids Research* 40.W1, W452–W457. DOI: 10.1093/nar/gks539.

Yang, Jianyi et al. (Jan. 2020). 'Improved protein structure prediction using predicted interresidue orientations'. In: *Proceedings of the National Academy of Sciences* 117.3, pp. 1496–1503. DOI: 10.1073/pnas.1914677117.