Prediction of the effect of single aminoacid protein variants using deep mutational scanning data

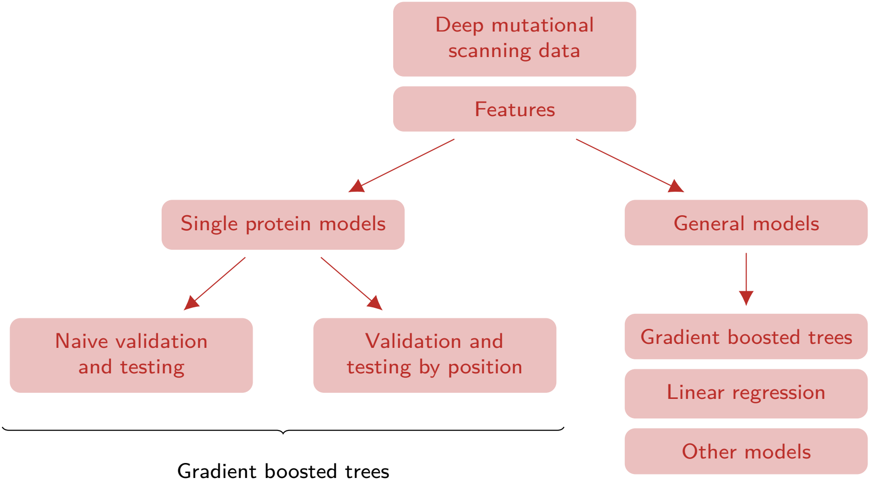
University of Bologna — Master Thesis in Bioinformatics

Pierotti Saul

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

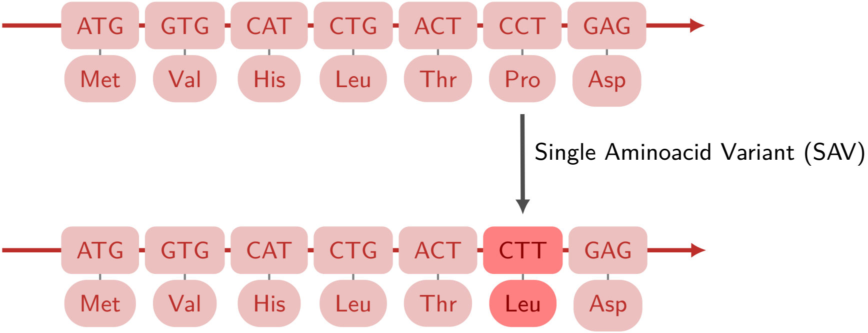
July 19, 2021

Structure of the project

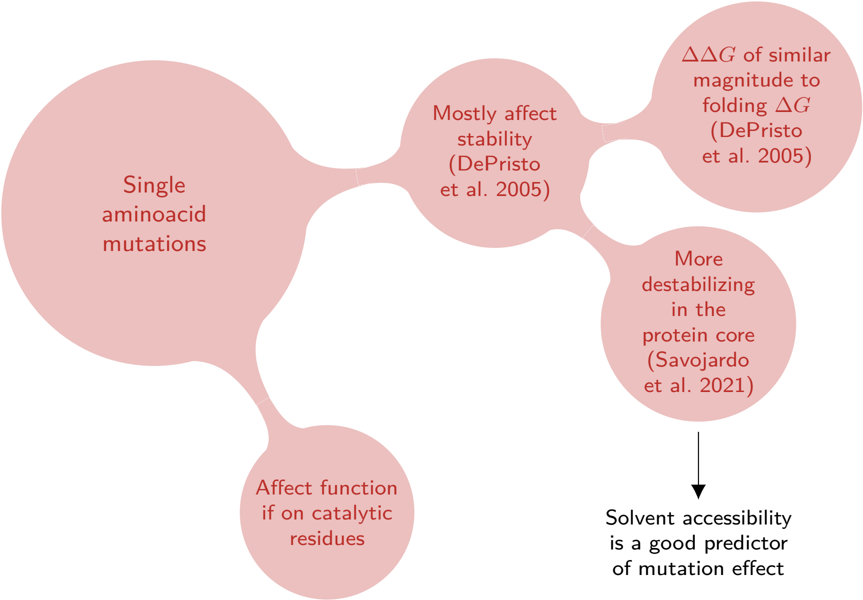


Single aminoacid variants

In this work I focused exclusively on point missense mutations. Nonsense mutations, indels, and mutations in non-coding regions were not considered.



Effect of single aminoacid mutations in proteins



Single aminoacid mutations

Mostly affect stability (DePristo et al. 2005)

$\Delta\Delta G$ of similar magnitude to folding ΔG (DePristo et al. 2005)

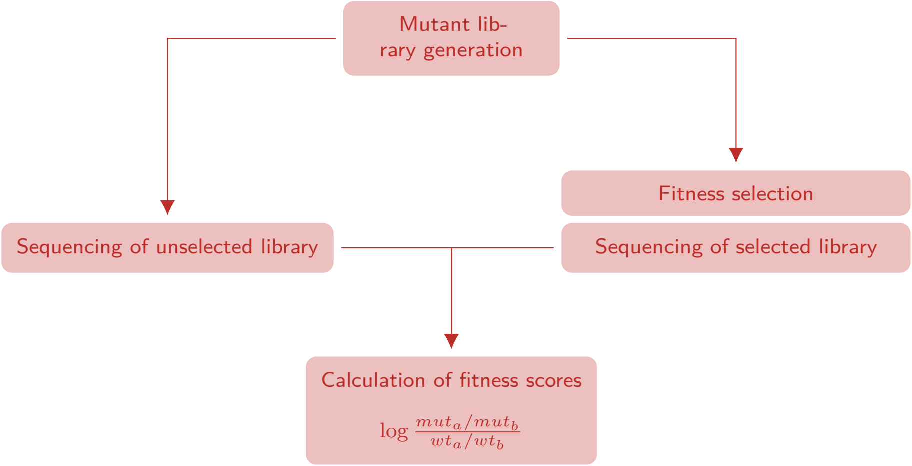
More destabilizing in the protein core (Savojardo et al. 2021)

Solvent accessibility is a good predictor of mutation effect

Affect function if on catalytic residues

Deep mutational scanning

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



Mutant library generation

Sequencing of unselected library

Fitness selection

Sequencing of selected library

Calculation of fitness scores

$$\log \frac{mut_a/mut_b}{wt_a/wt_b}$$

Variant effect prediction

Complete experimental coverage of the human proteome mutational landscape is not currently in reach, and it is a distant goal for model organisms. Computational predictions are needed.

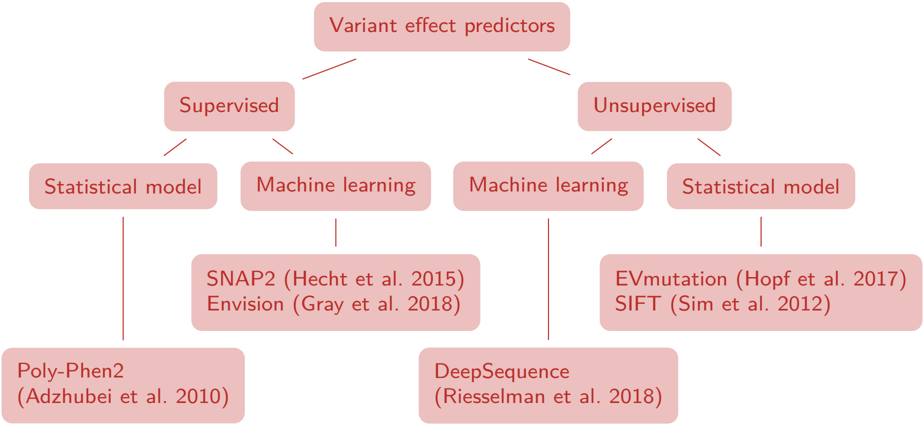
But why is it important to know the effect of a mutation?

- Precision medicine
- Protein engineering

How can it be done?

- Supervised or unsupervised
- Quantitative or as a classification problem
- Machine learning or simple statistical models

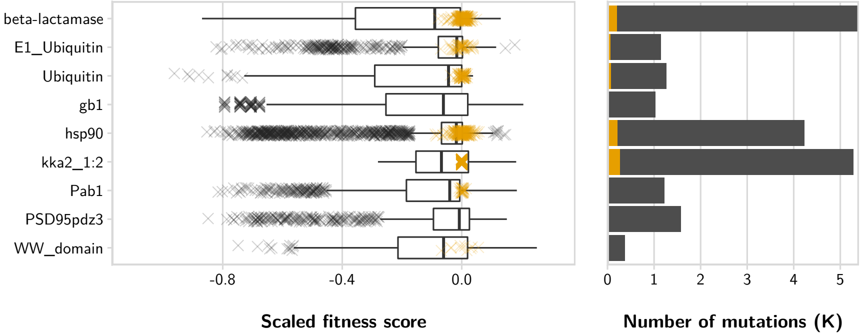
Variant effect predictors



Envision, EVmutation, and DeepSequence provided quantitative predictions. Envision was trained on deep mutational scanning data while the others are either unsupervised or trained on SNP annotations.

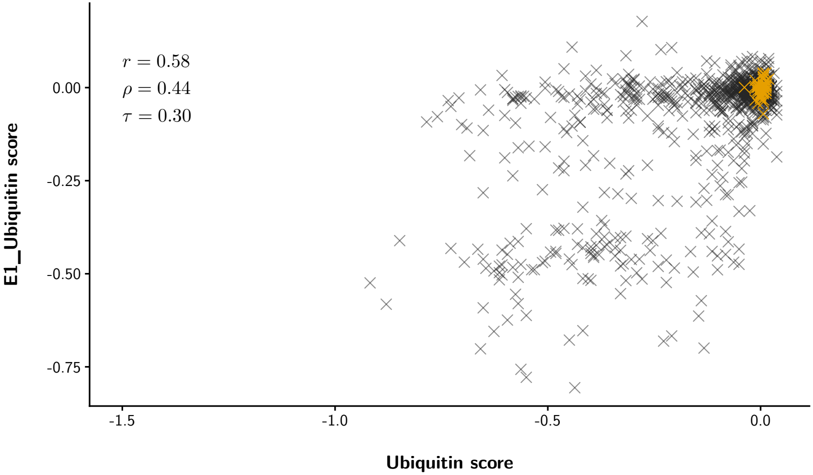
Training data

I used the training dataset of Envision (Gray et al. 2018), composed of nine independent experiments on eight different proteins. The distribution of fitness scores is bimodal and very variable across datasets.

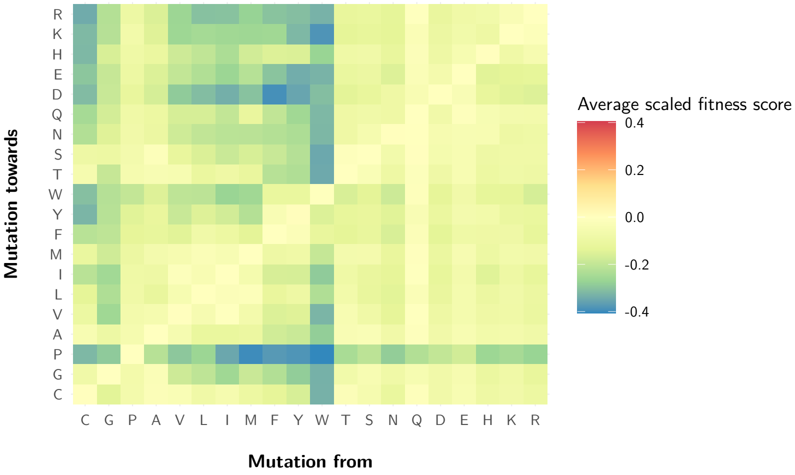


Poor correlation among experimental results

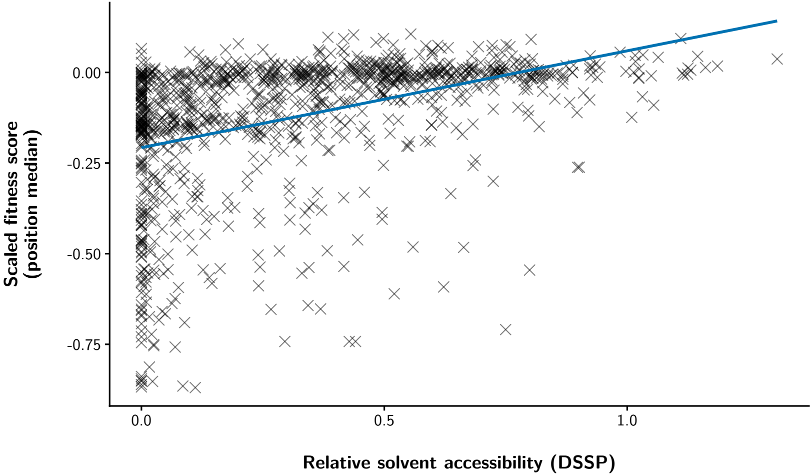
Two independent deep mutational scanning experiments on Ubiquitin are present in the aggregated dataset, but their correlation is quite low.



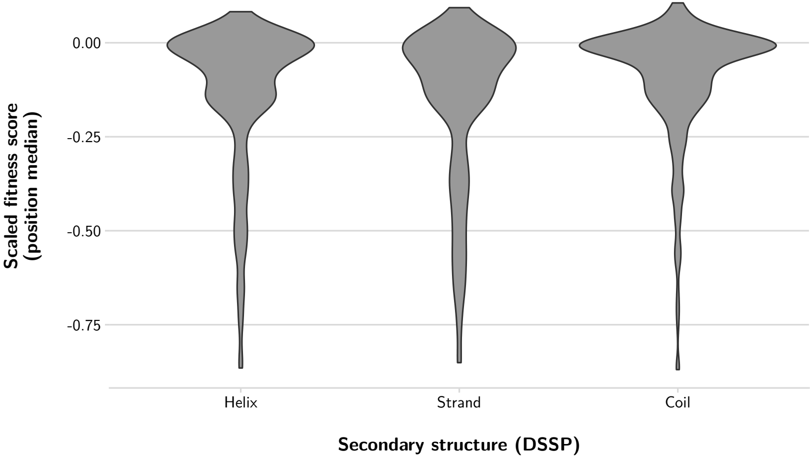
The effect of a mutation is strongly influenced by the identity of the wild-type and mutant residues



Solvent-accessible positions are more tolerant towards mutations



Mutation effect is not well discriminated by secondary structure



Features

I used only features that could be obtained from the sequence of the mutated proteins or from multiple sequence alignments. No structural features were used.

EVmutation predictions (Hopf et al. 2017)

- Epistatic model
- Independent model
- Conservation
- Mutation frequency

Mutation identity

- Wild-type residue
- Mutated residue

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

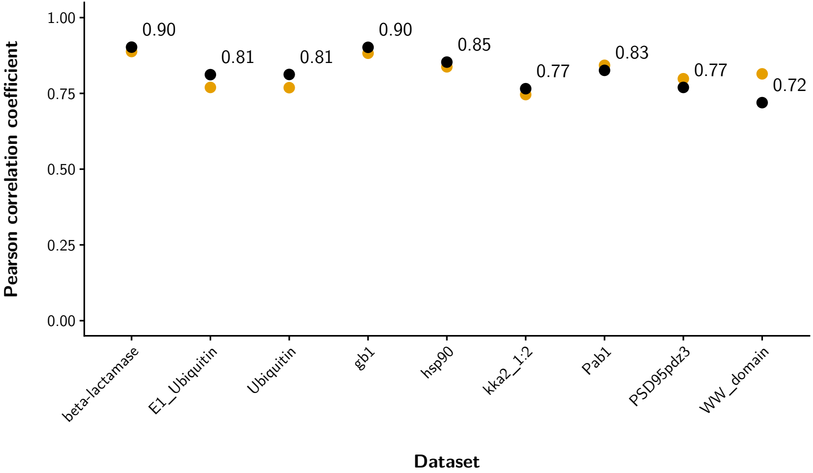
Single protein models

I trained two different gradient boosted tree models for each of the proteins in the training set.

- Naive cross-validation and testing
  - Half of the mutations in a protein randomly set aside for testing
  - Hyperparameters optimized in the remaining half with 5-fold cross-validation (randomly selecting mutations)
- Cross-validation and 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 (avoiding that mutations in the same protein position ended up in different folds)

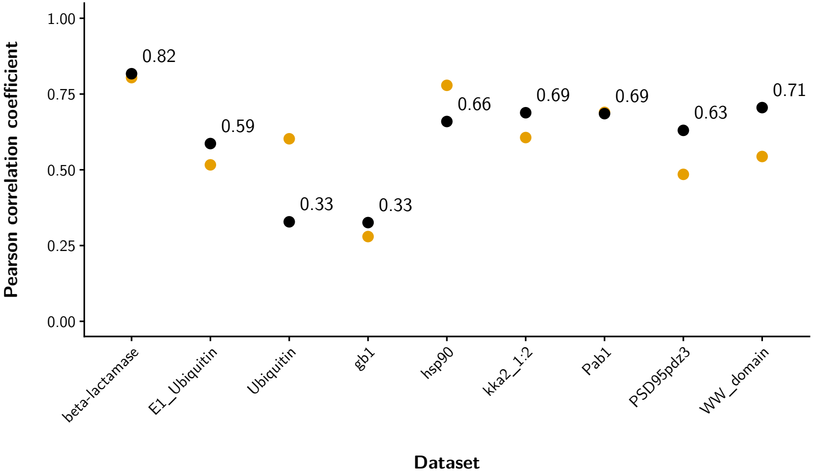
Naive single protein models

Superficially good results when training and testing randomly across mutations ...



Single protein models that segregate protein positions

...but overoptimistic, since when I separate protein positions the performances drop dramatically, particularly for some datasets.



General models

I trained a set of general predictors on the full dataset.

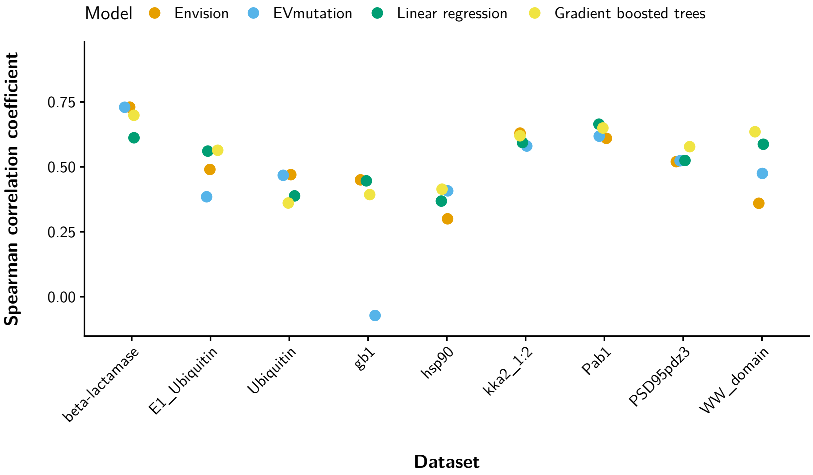
- Cross-validation by leaving a protein out at each iteration
- Of the left-out protein, half of the mutations were used in validation and the rest for testing

Given the diversity of deep mutational scanning datasets, aiming at predicting the ranks is more meaningful than predicting the absolute fitness scores

- Pairwise ranking loss used for gradient boosted trees
- Quantile normalization of the experimental fitness for the linear regression
- Models evaluated in terms of Spearman correlation coefficient instead of Pearson correlation coefficient

General models

Gradient boosted trees perform slightly better than linear regression, but the difference is minor. Performances comparable to those of Envision and EVmutation.



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.

DePristo, Mark A., Daniel M. Weinreich and Daniel L. Hartl (Aug. 2005). 'Missense meanderings in sequence space: a biophysical view of protein evolution'. In: *Nature Reviews Genetics* 6.9, pp. 678–687. DOI: 10.1038/nrg1672.

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.

Bibliography II

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.

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.

Savojardo, Castrense et al. (Jan. 2021). 'Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences'. In: *Frontiers in Molecular Biosciences* 7. DOI: 10.3389/fmolb.2020.626363.

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.

Bibliography III

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.