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

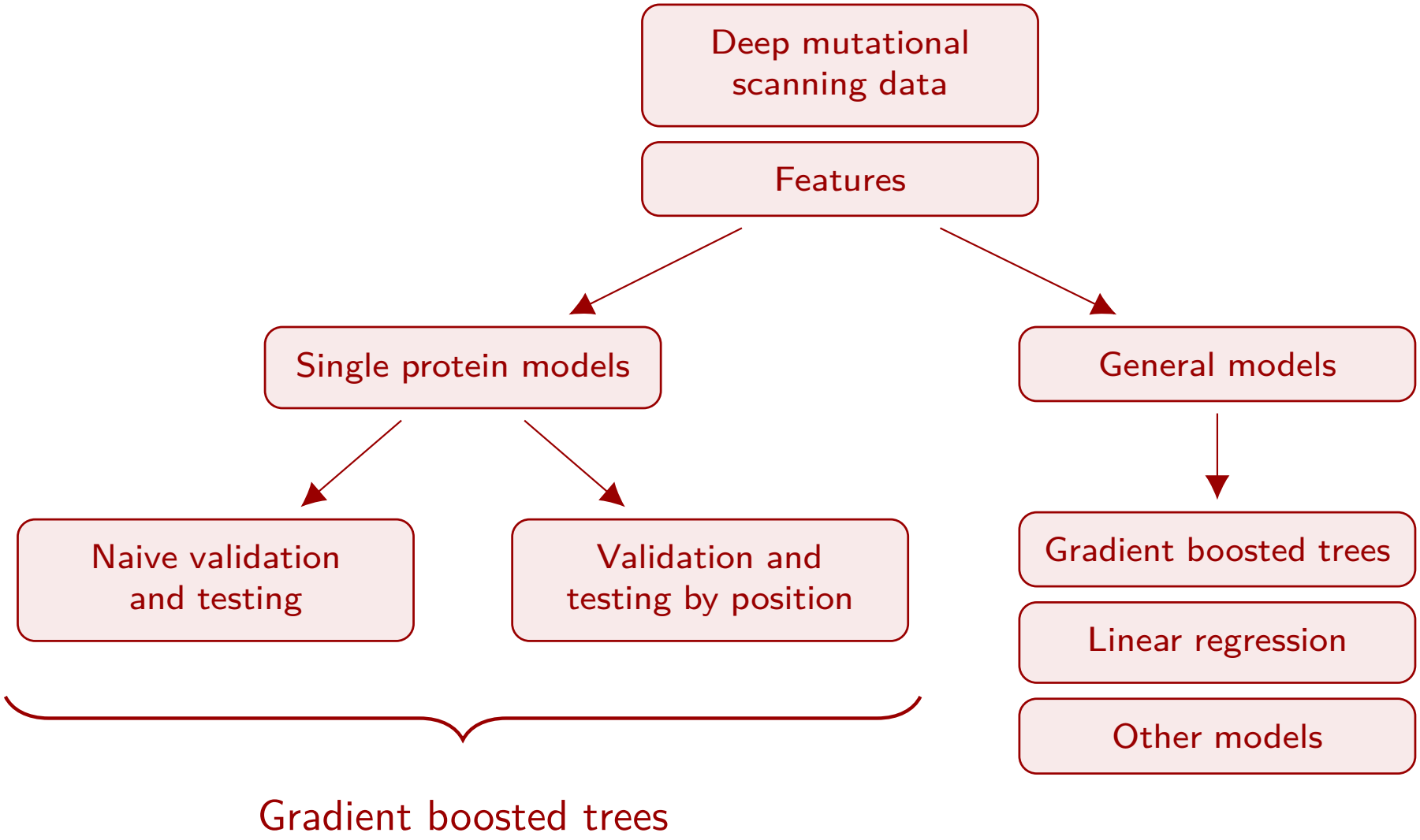
University of Bologna — Master Thesis in Bioinformatics

Pierotti Saul

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

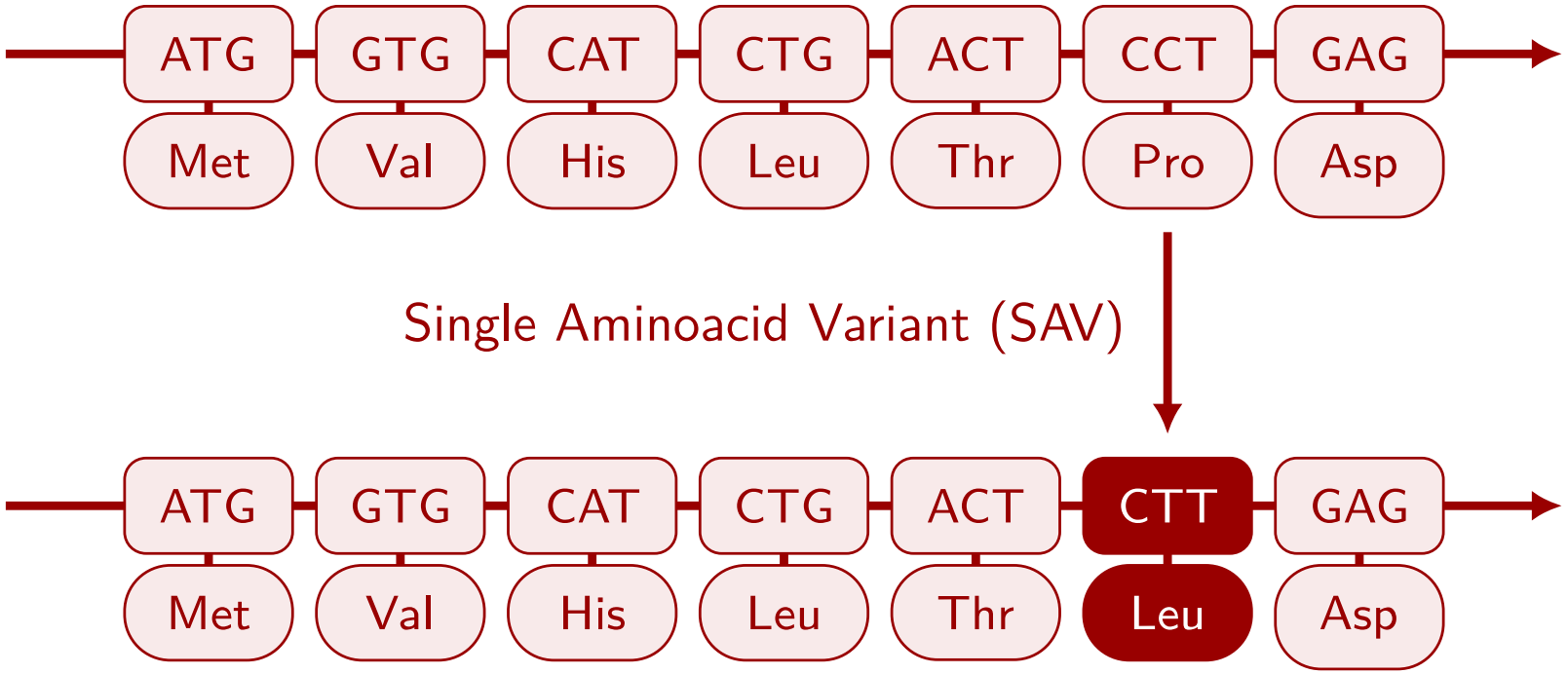
July 19, 2021

Structure of the project

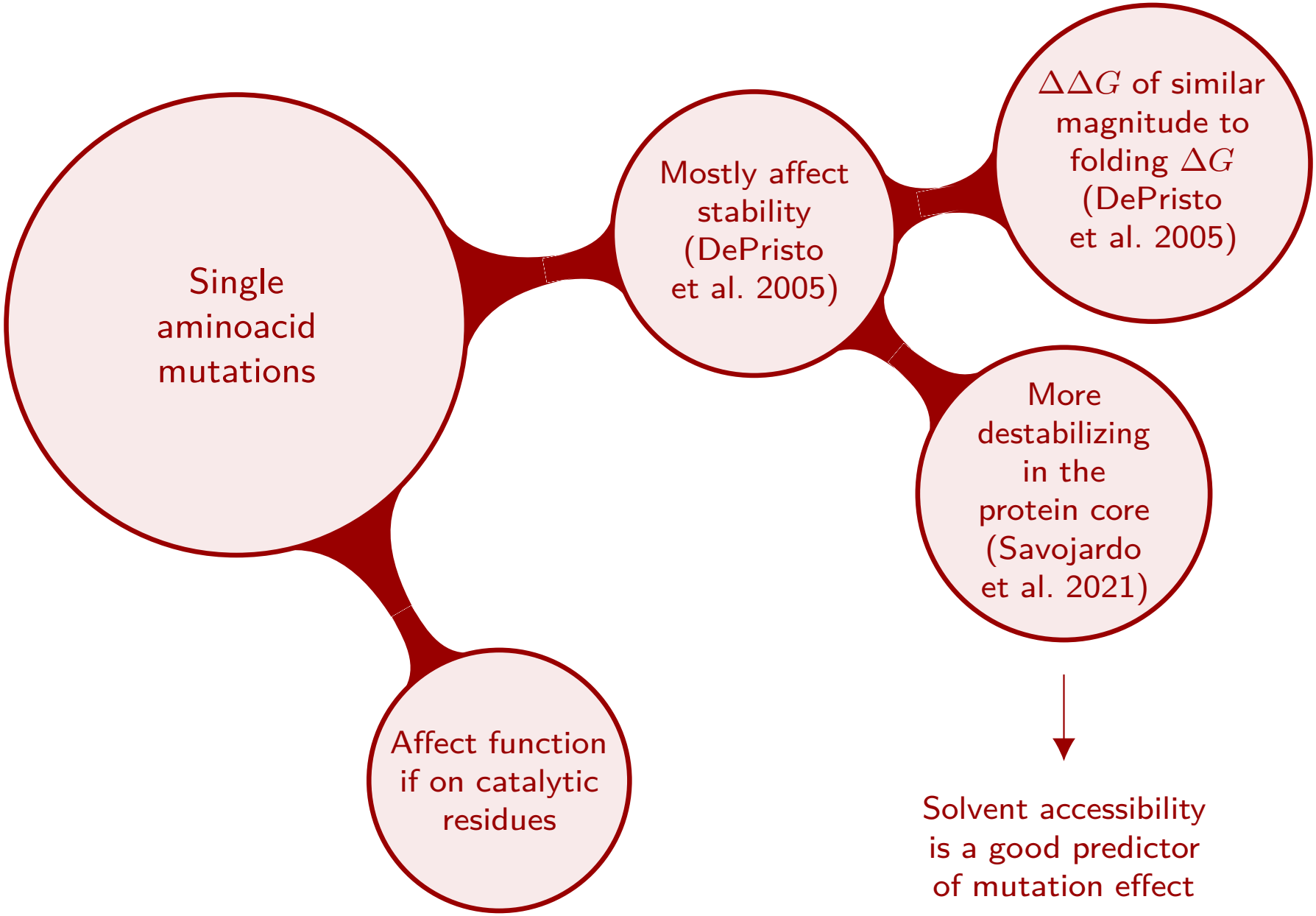


Single aminoacid variants

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

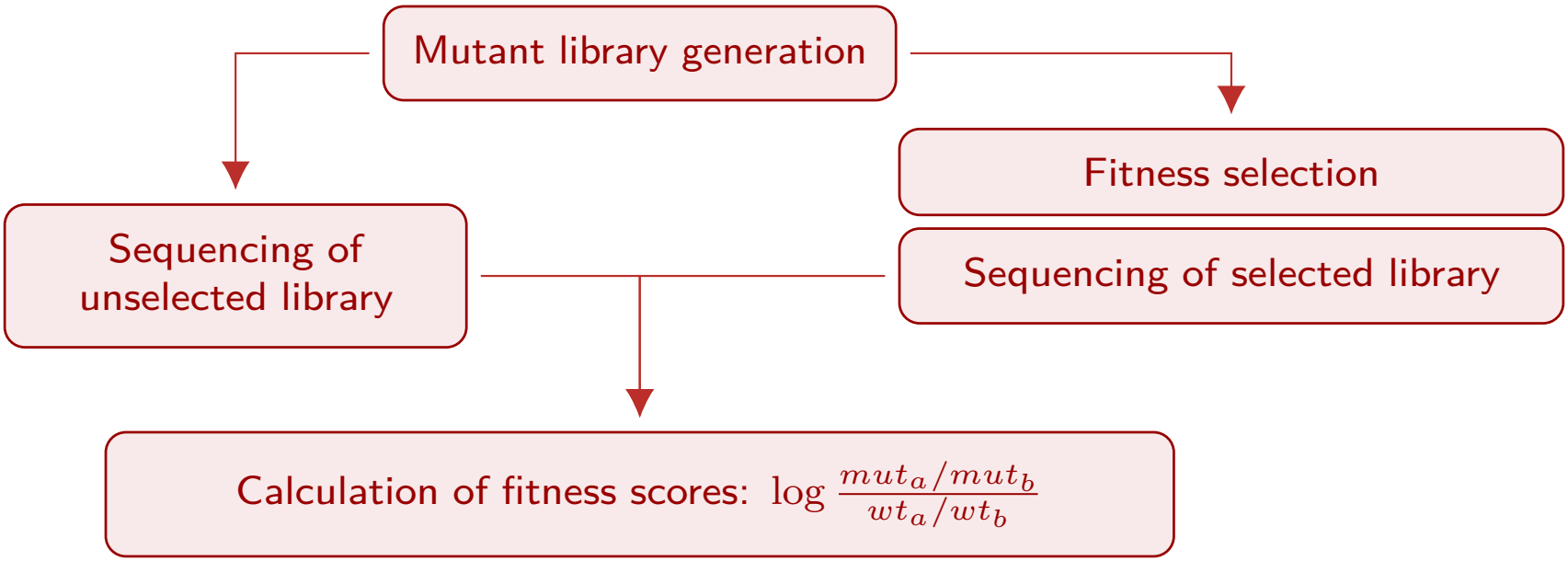


Effect of single aminoacid mutations in proteins

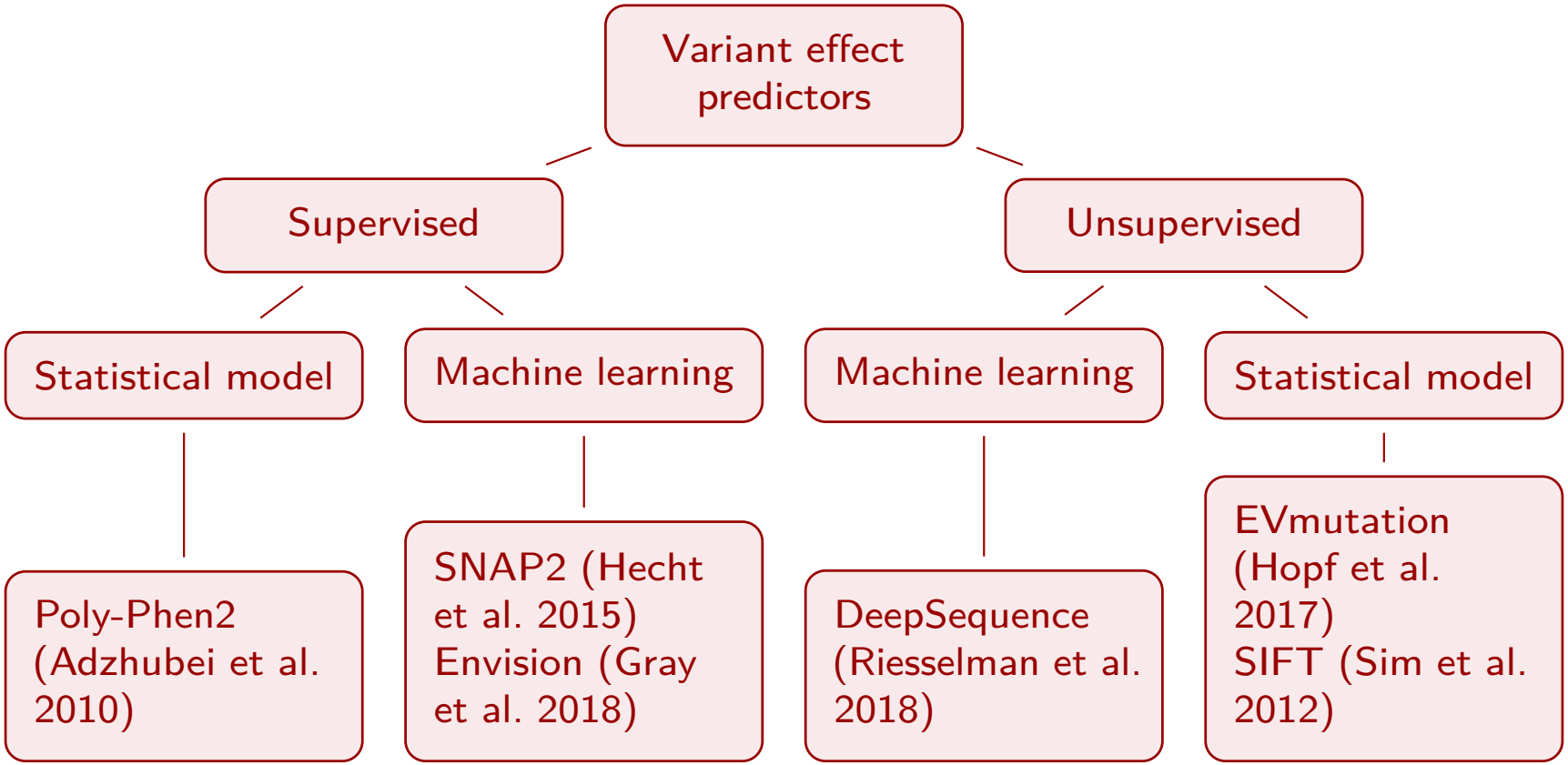


Deep mutational scanning

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



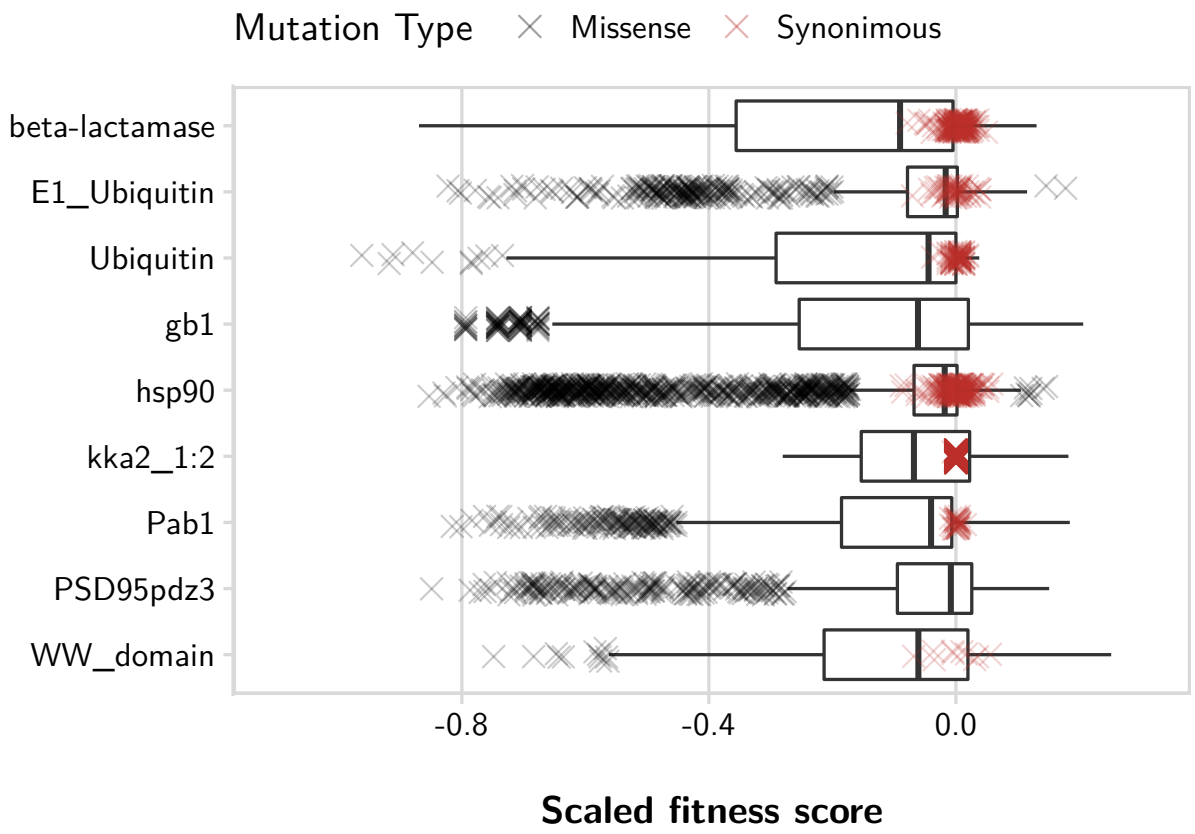
Variant effect predictors

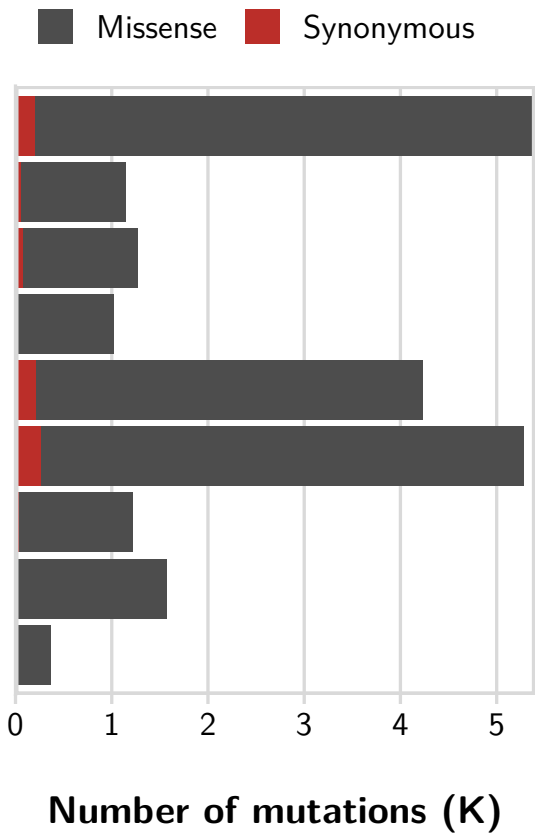


Envision, EVmutation, and DeepSequence provided quantitative predictions. Envision was trained on deep mutational scanning data while the others are either unsupervised or trained on SNP annotations.

Training data

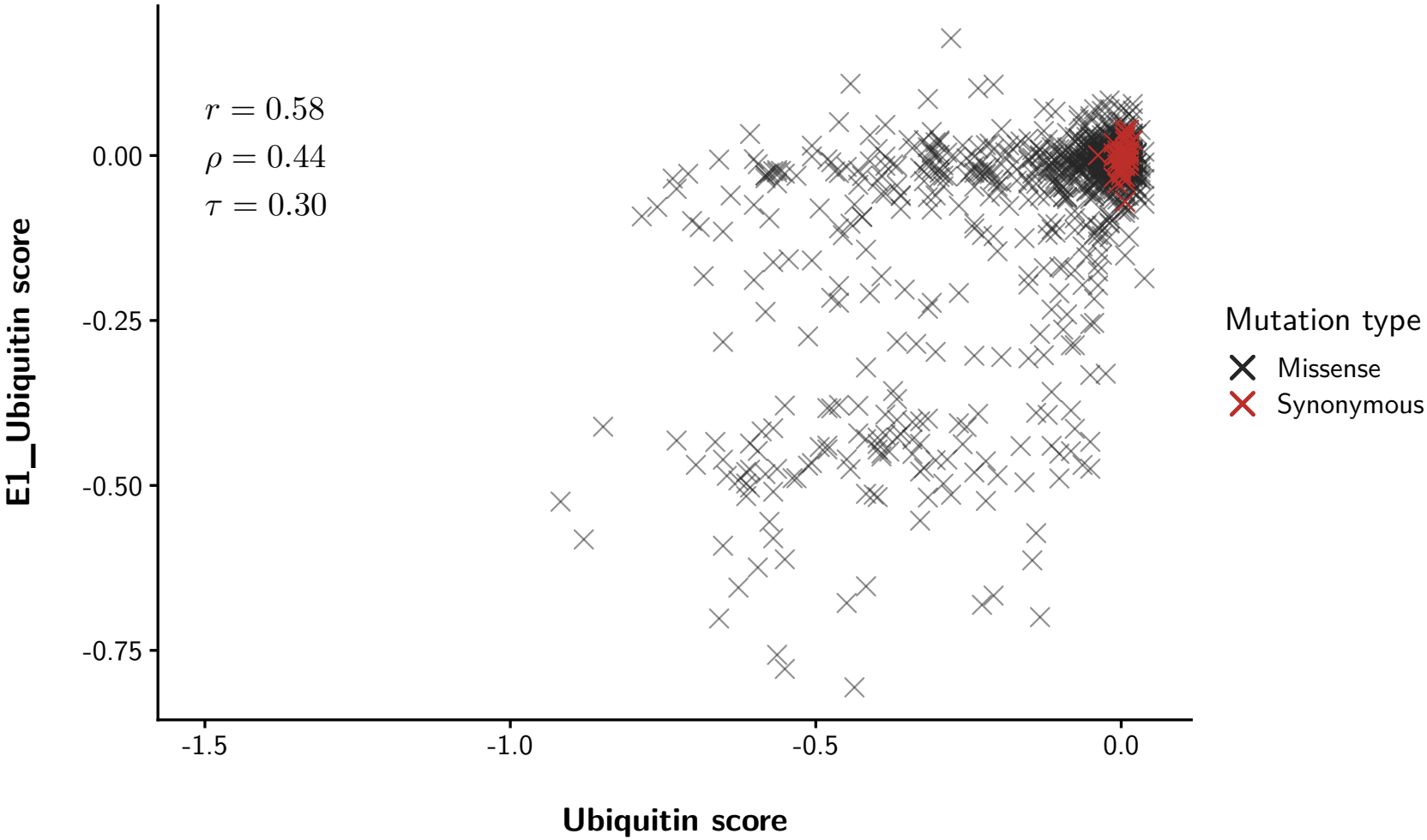
I used the training dataset of Envision (Gray et al. 2018), composed of nine independent experiments on eight different proteins. The distribution of fitness scores is bimodal and very variable across datasets.





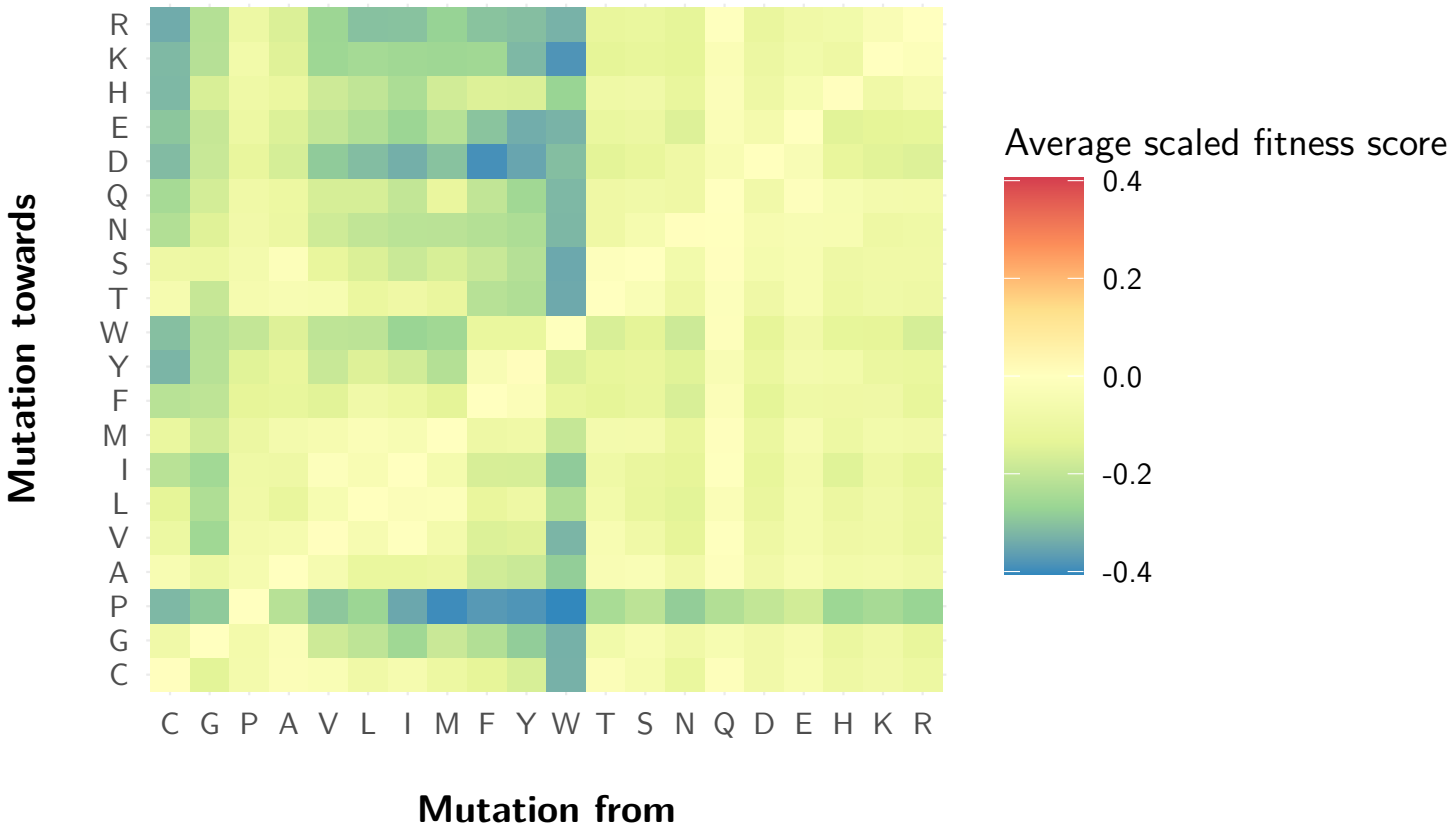
Poor correlation among experimental results

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

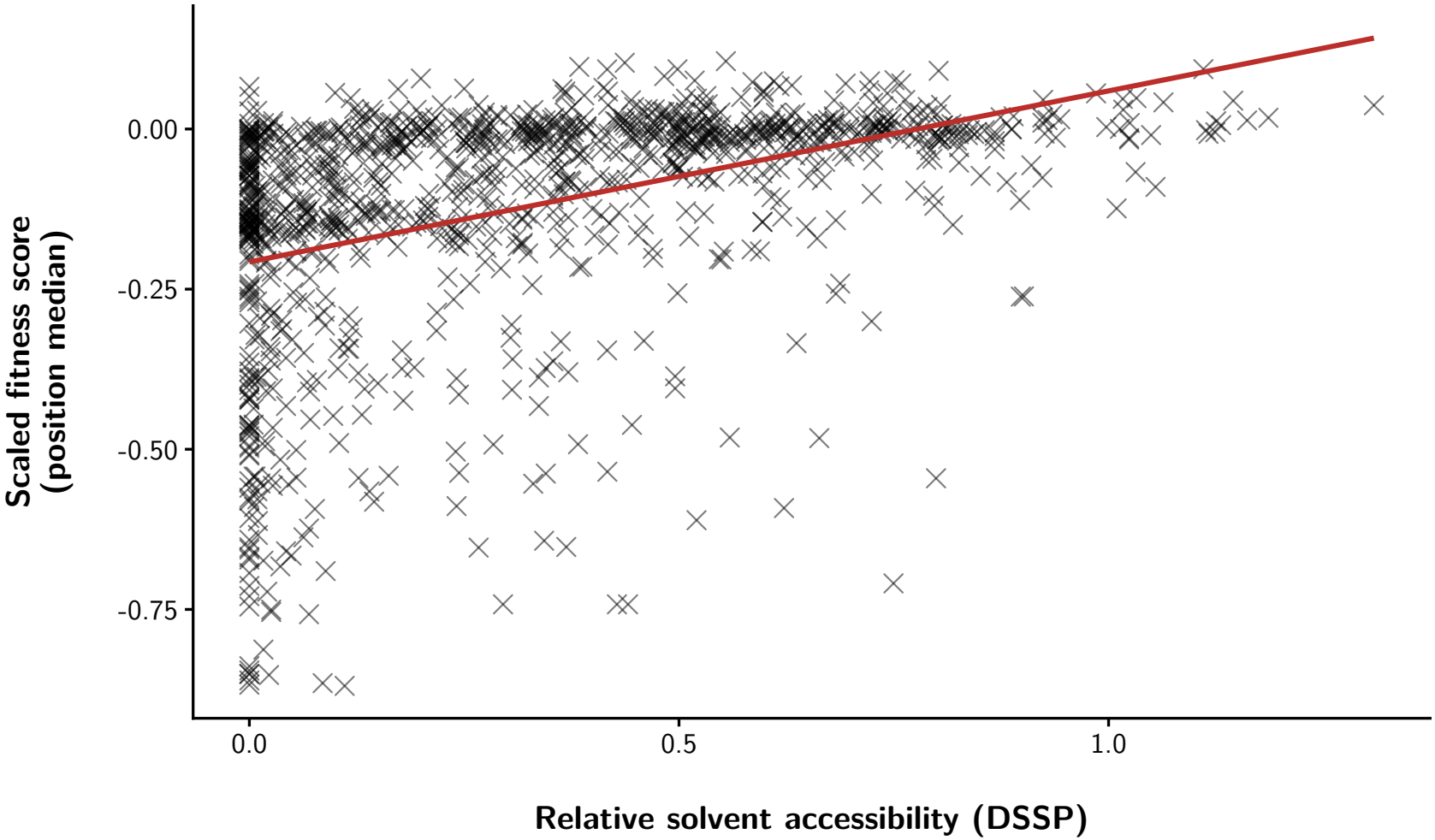


The effect of a mutation is strongly influenced by the identity of the wild-type and mutant residues

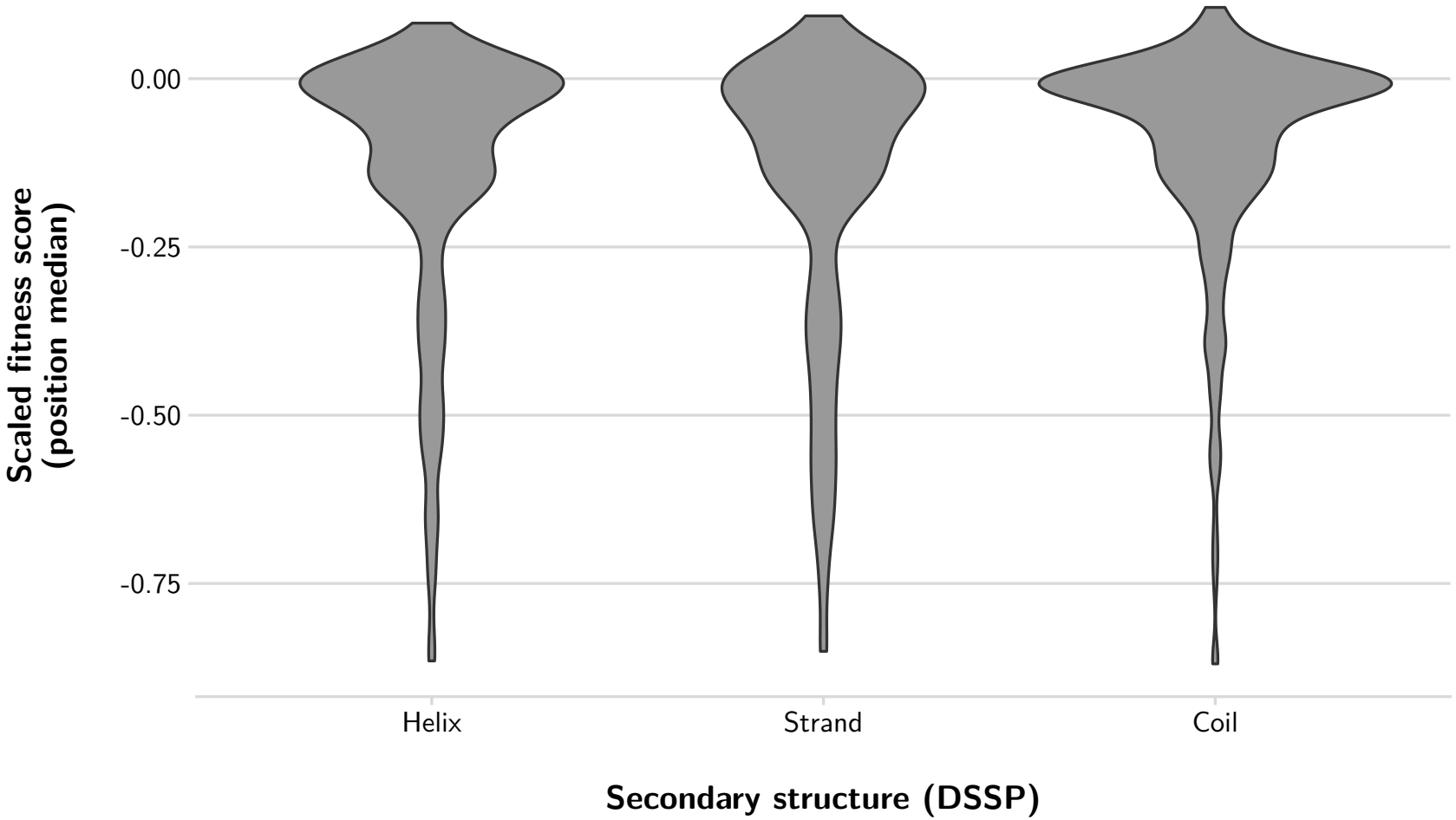
Mutations from polar residues tend to be less detrimental than mutations from apolar residues. This effect, however, disappears when filtering the mutations by relative solvent accessibility.



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.

- 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