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

University of Bologna — Master Thesis in Bioinformatics

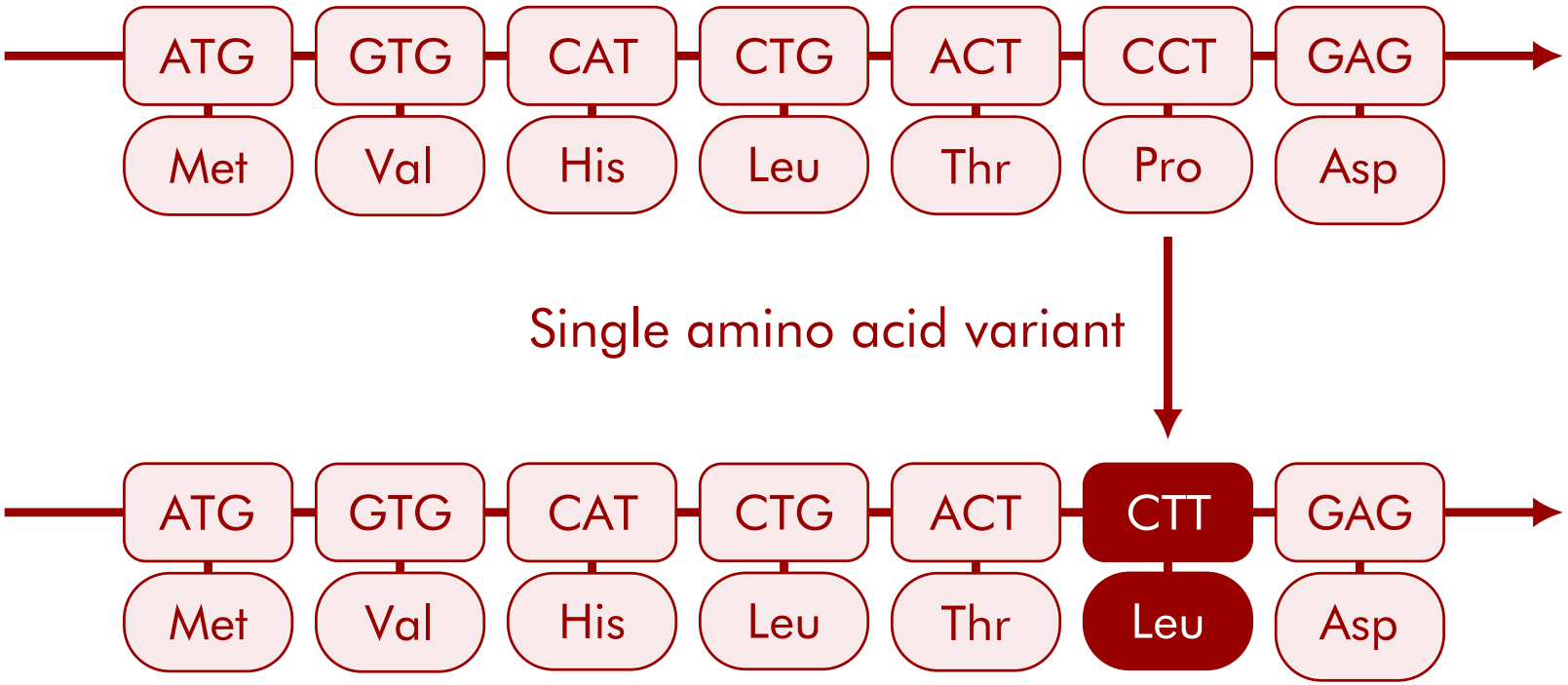
Pierotti Saul

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

July 19, 2021

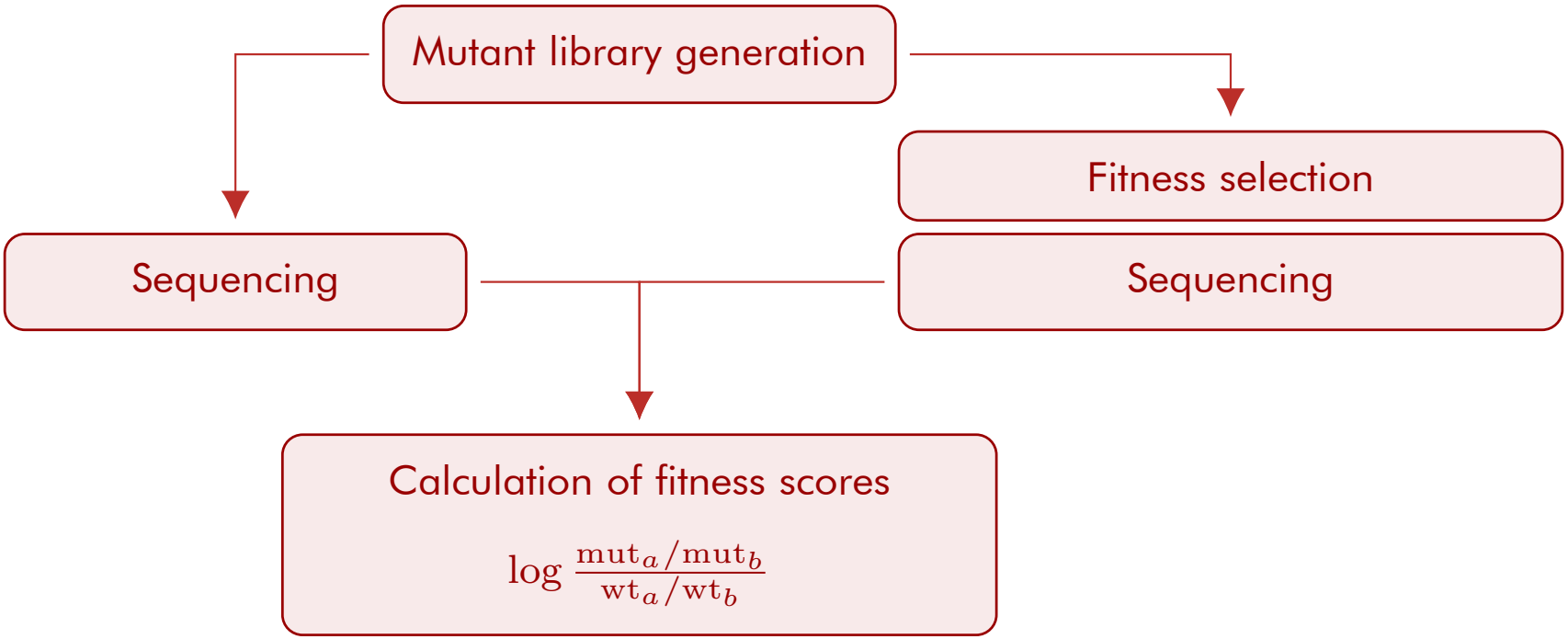
Single Amino Acid Variant

A mutation that replaces exactly one amino acid in a protein



Deep Mutational Scanning

A technique for obtaining fitness information on a large number of mutations in parallel



Predicting the Effect of Mutations

Why is it useful?

- Targeted medical treatments
- Protein engineering

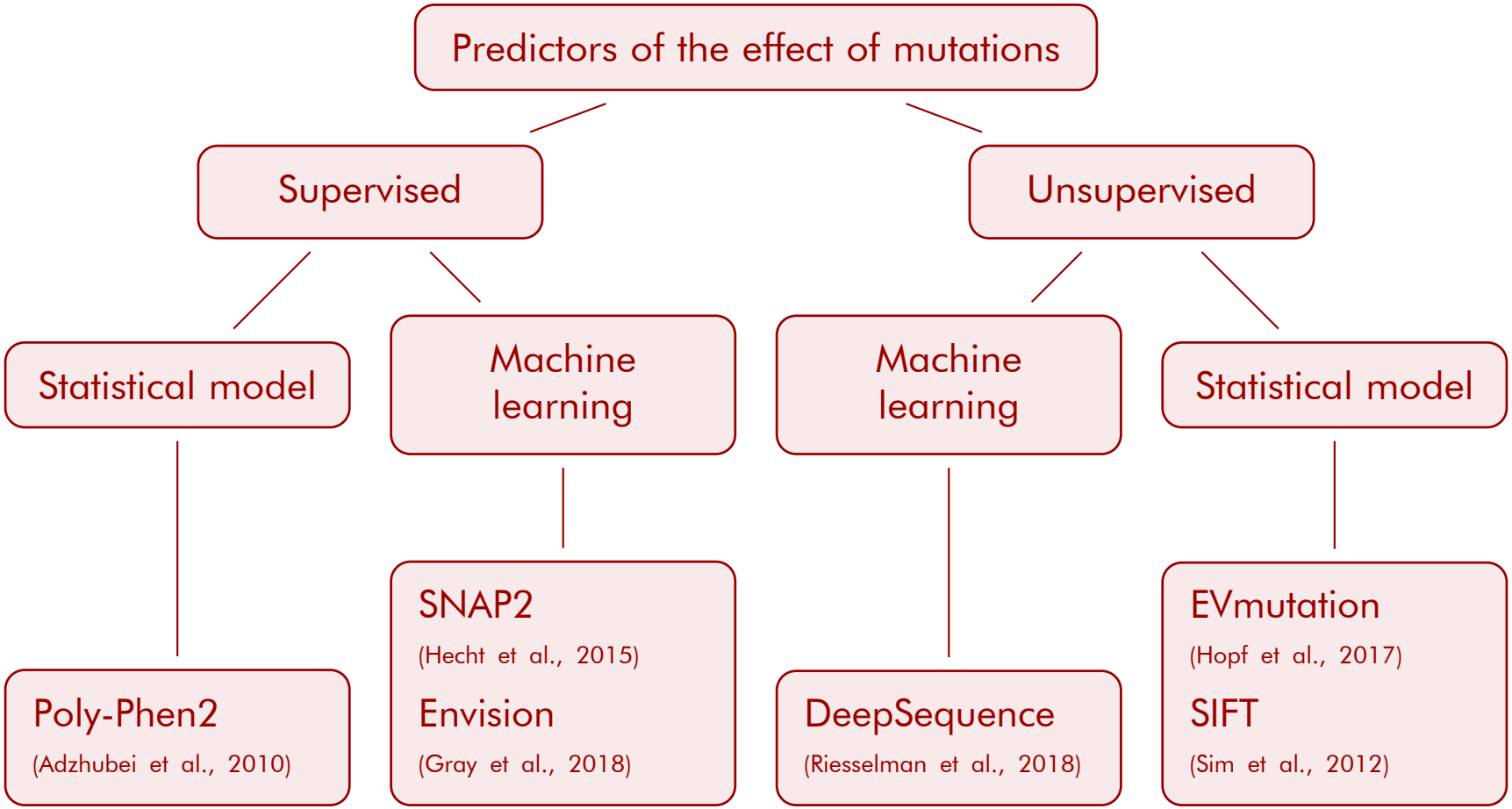
Why is it needed?

- Experiments are insufficient
- Experiments are expensive

How can it be done?

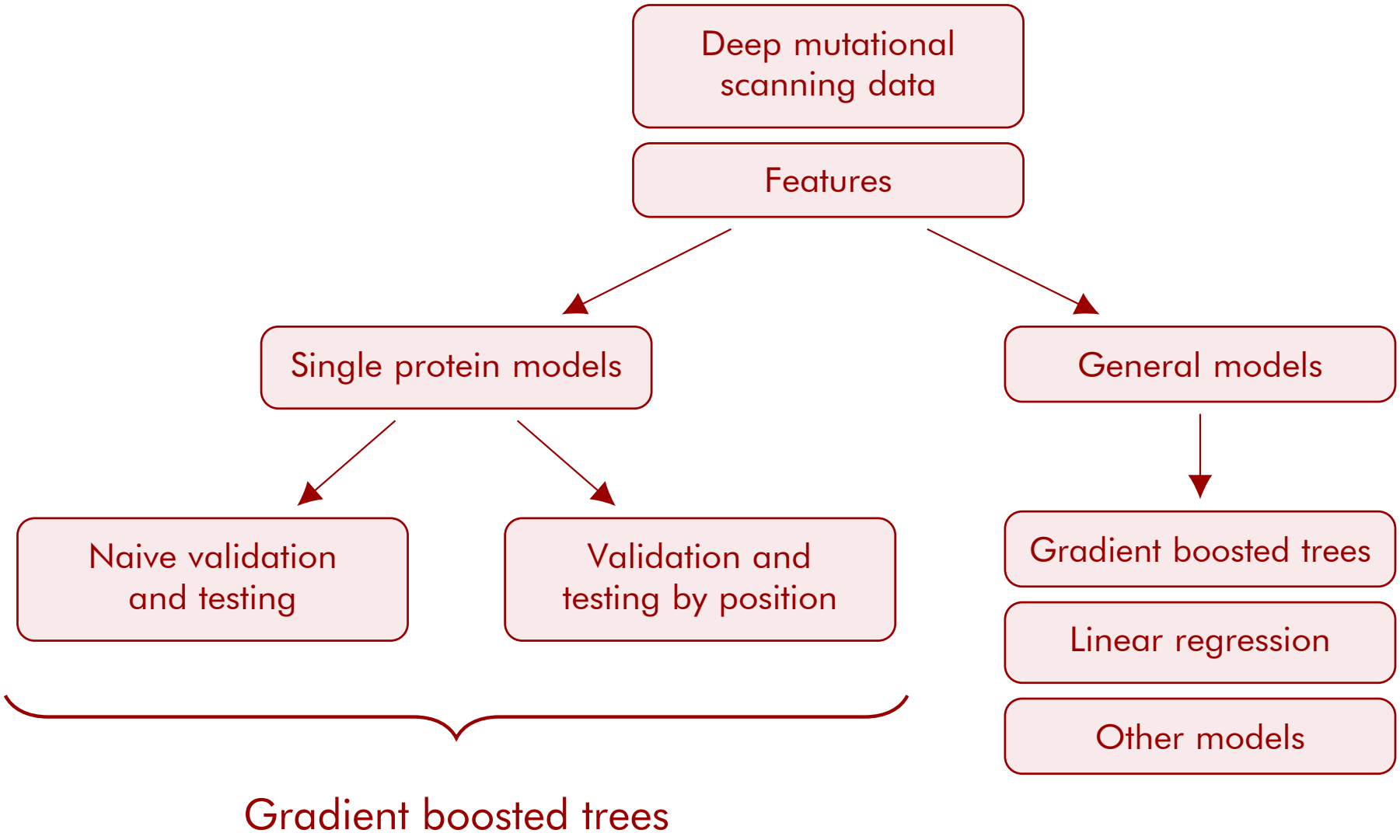
- Machine learning
- Statistical models

Previous Work in the Field



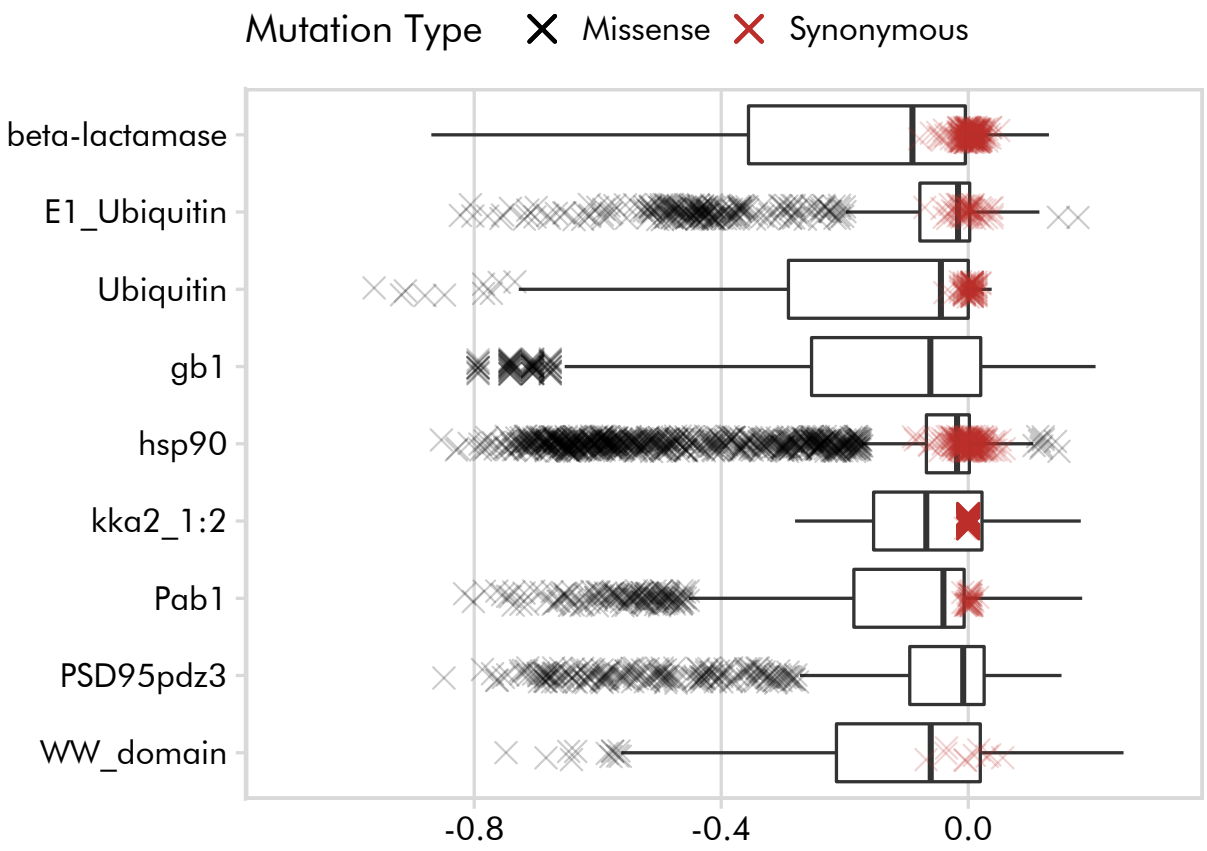
Among these predictors only Envision was trained on deep mutational scanning data

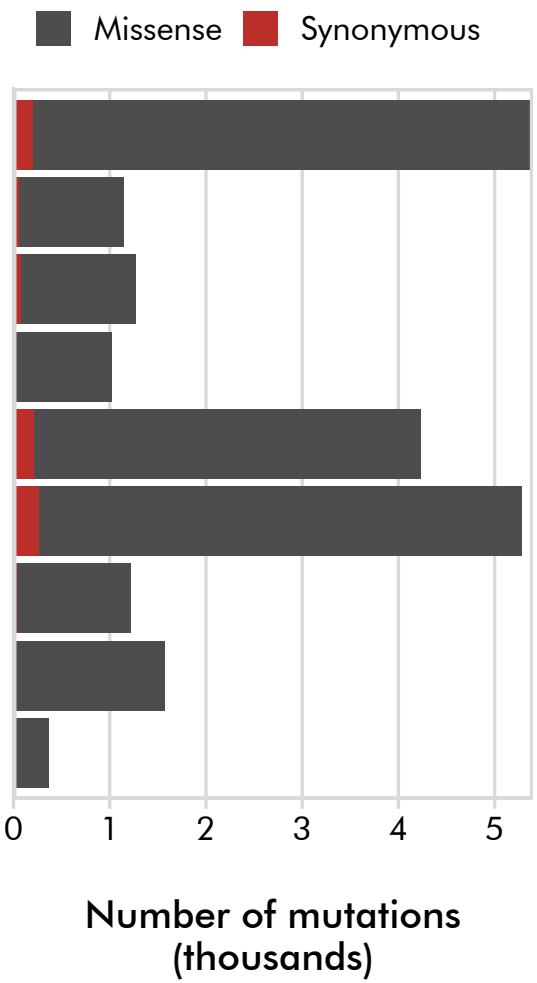
My Approach


Deep mutational scanning data
Features
Single protein models
General models
Naive validation and testing
Validation and testing by position
Gradient boosted trees
Linear regression
Other models
Gradient boosted trees

Training Data

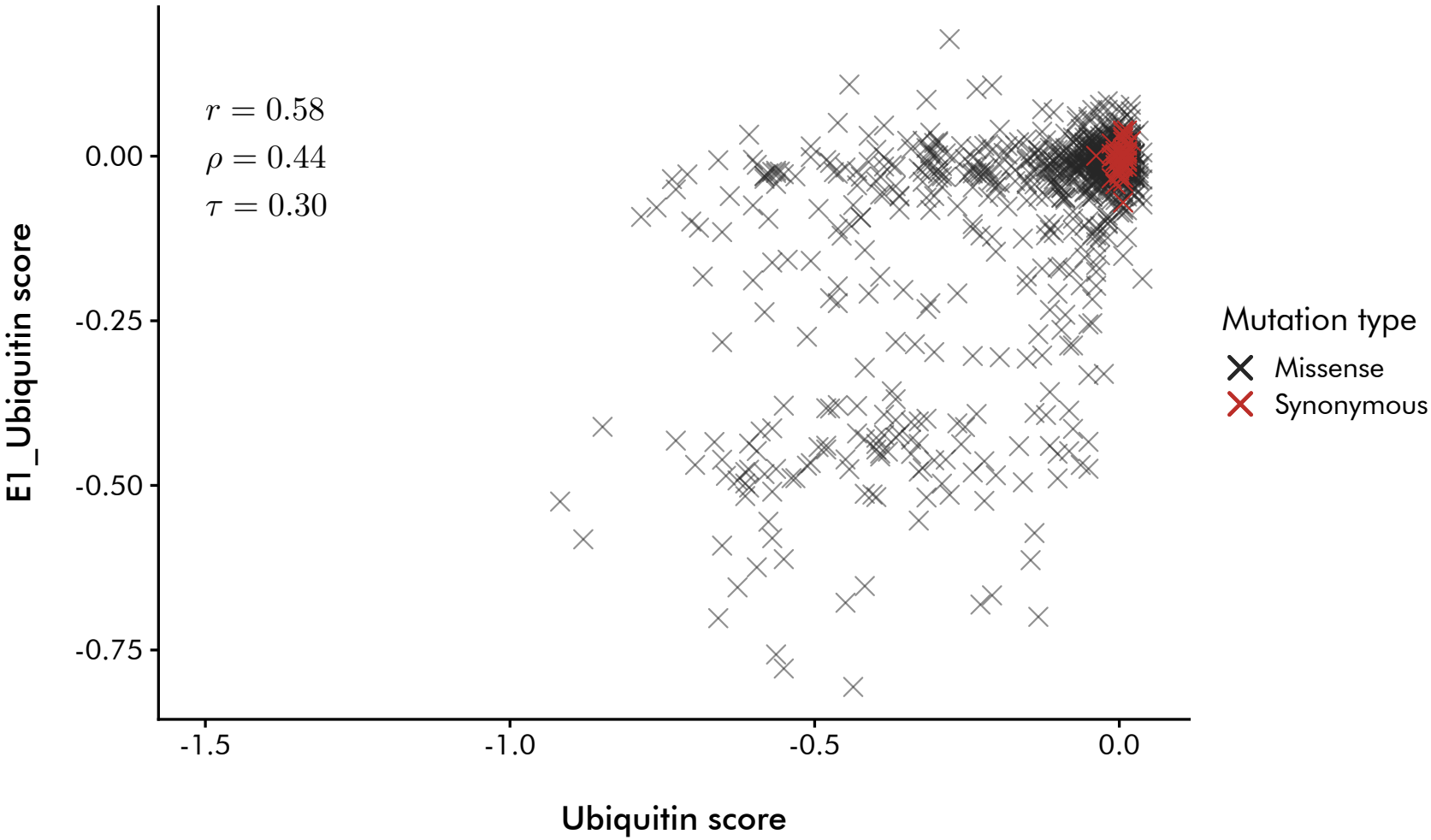
I borrowed the training dataset of the predictor Envision (Gray et al., 2018)



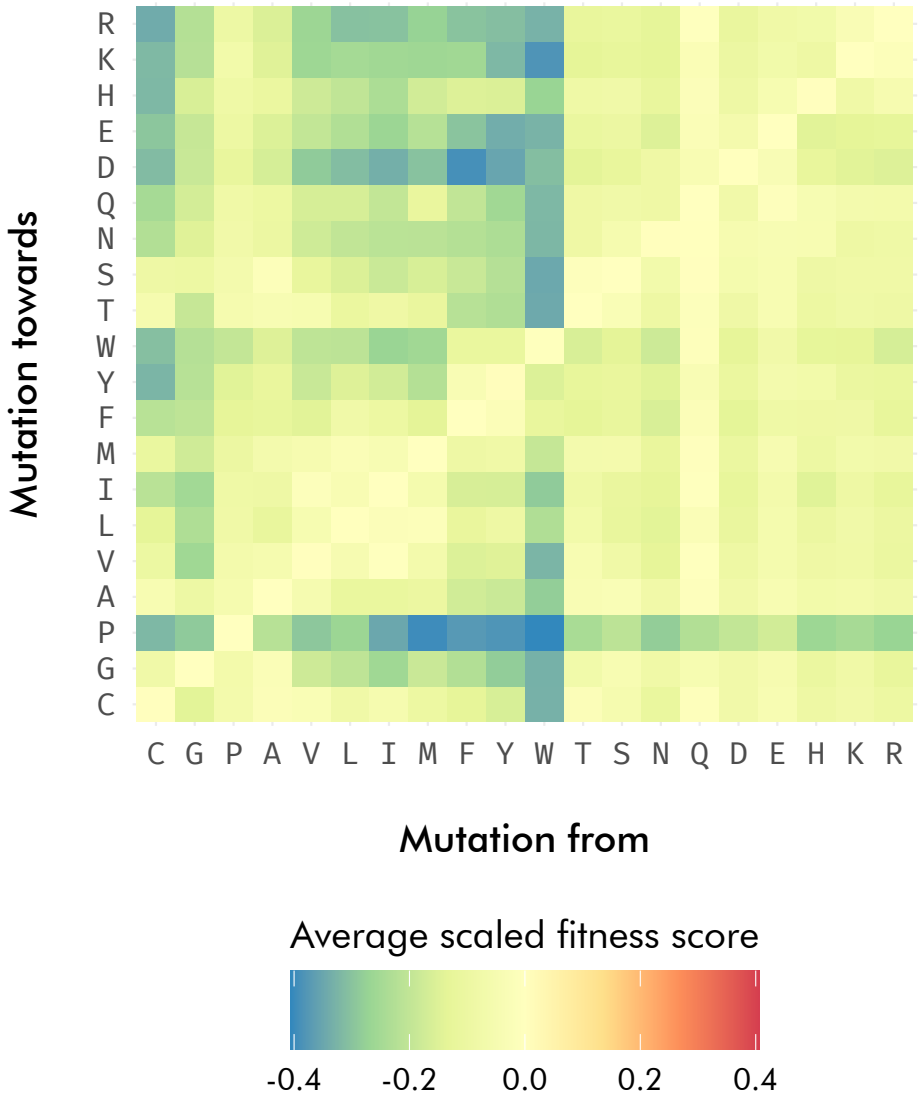


Experiments Do Not Agree Much with Each Other

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



Interesting Patterns in Mutation Sensitivity

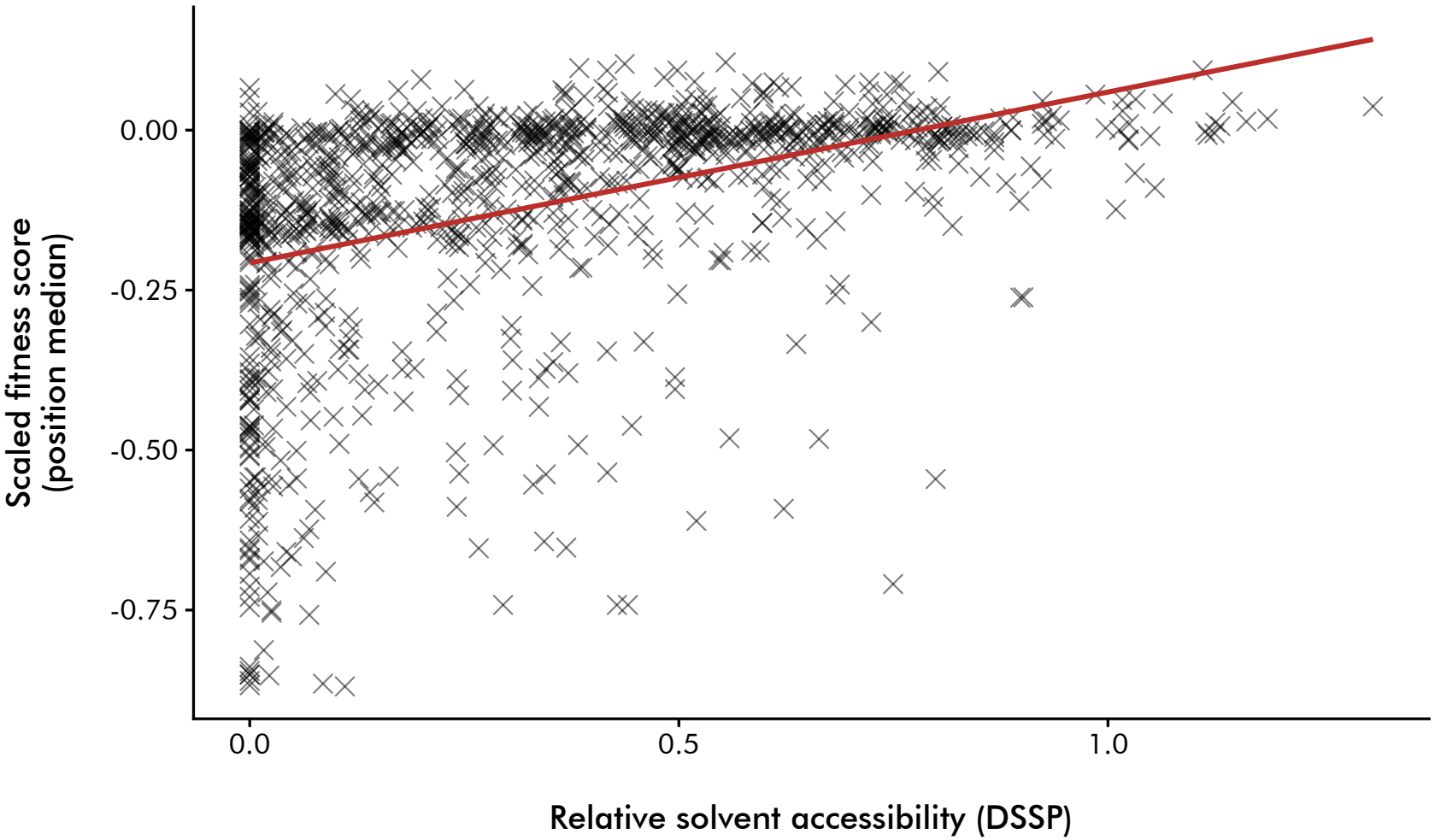


- Polar residues seem more tolerant to mutations than hydrophobic residues
- Proline (P) is the most disruptive residue
- Tryptophan (W) is hard to replace

Exposure Explains the Mutability of Polar Residues

When filtering by Relative Solvent Accessibility (RSA) apolar residues are ***not*** more sensitive to mutations than polar residues

Buried Residues Are More Conserved

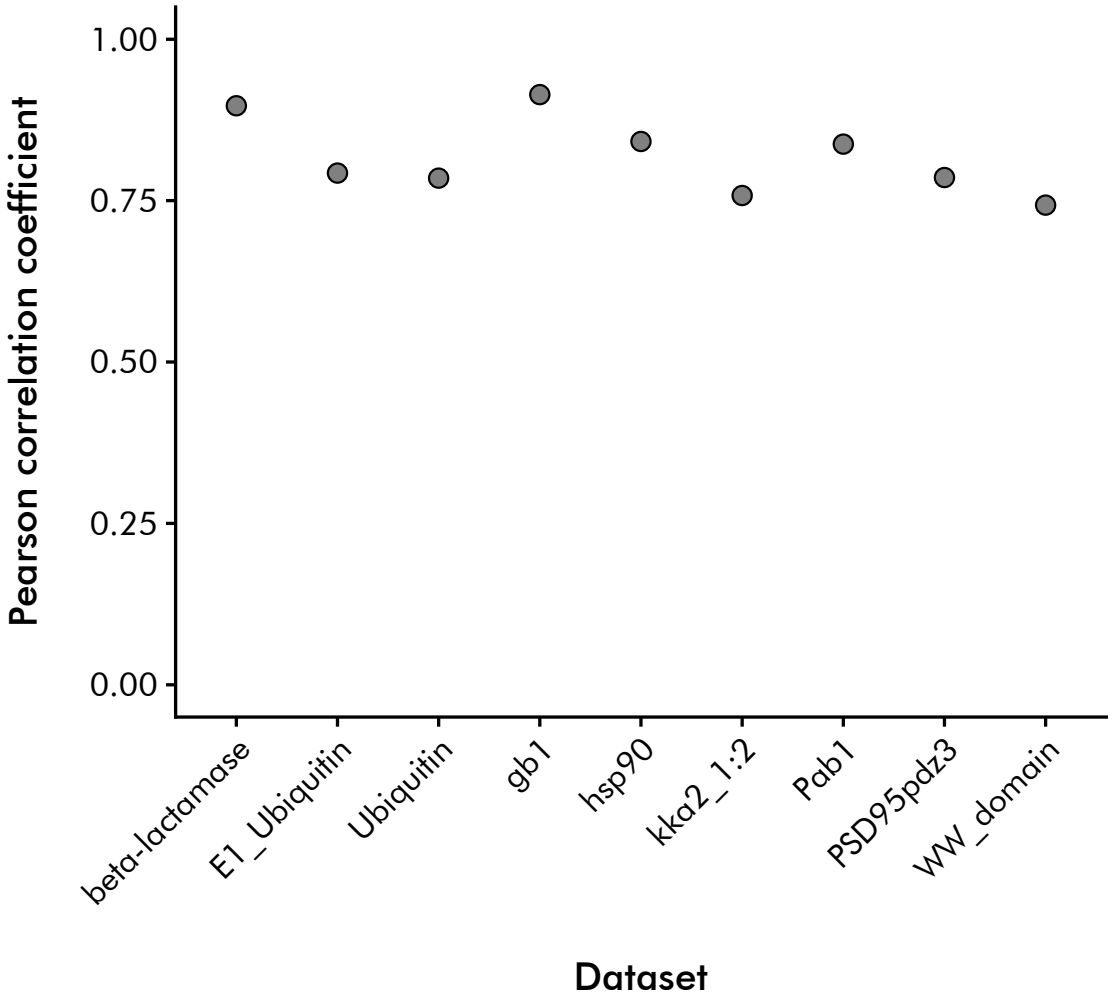


Features Used by the Predictors

All the features are derived from the sequences: I did ***not*** use any structural information

- Mutation identity
- EVmutation predictions (Hopf et al., 2017)
- NetsurfP-2 predictions (Klausen et al., 2019)
- HMMER emission probabilities (Eddy, 2011)
- trRosetta predicted contacts (Yang et al., 2020)

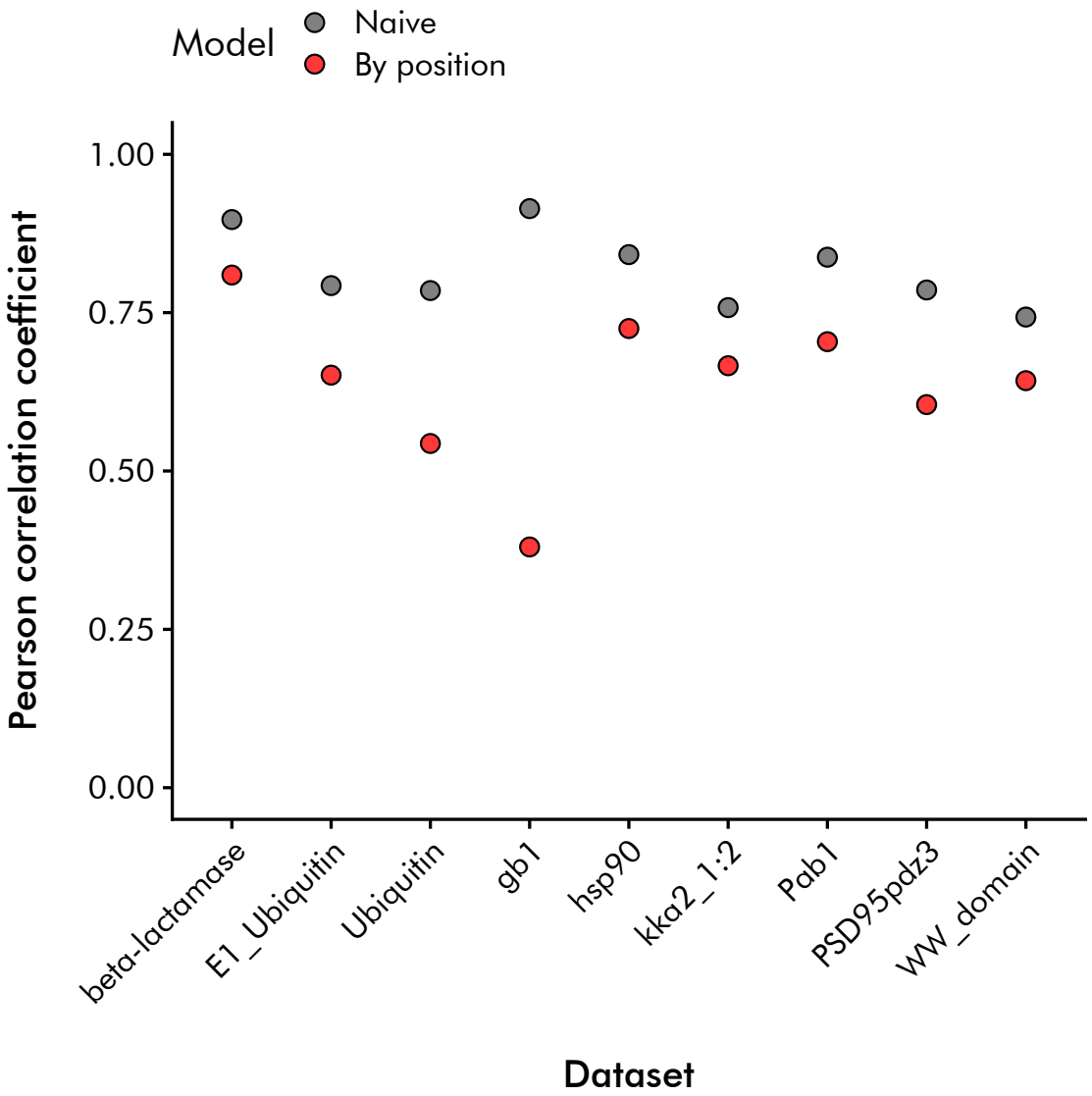
Single Protein Models



Naive approach

- A different model trained for each protein
- Half of the mutations used for testing and half for cross-validation
- Too good to be true

Single Protein Models



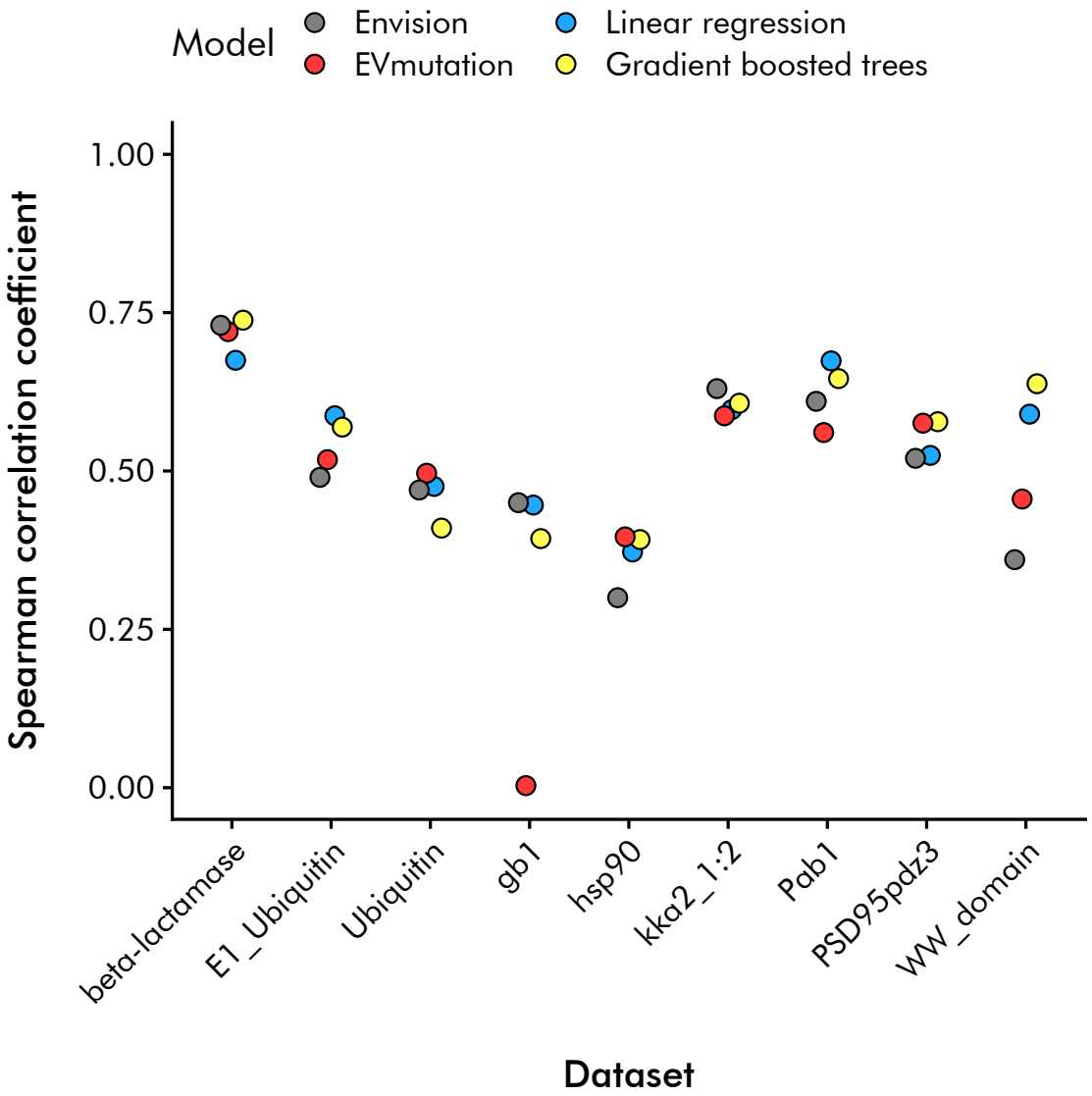
Naive approach

- A different model trained for each protein
- Half of the mutations used for testing and half for cross-validation
- Too good to be true

Segregating protein positions

- Same as above but mutations in the same position segregated in the training or testing sets
- Performances are more realistic

Leave-One-Protein-Out (LOPO) Models



- Models trained on the whole dataset while leaving one protein out
- For the left-out protein, half of the mutations used for testing and half for validation
- Spearman correlation coefficient used for evaluation

Discussion and Future Directions

What I learned

- The testing strategy is crucial
- Good performances without structural features
- Strong variability between datasets
- Complex models not necessarily better

Ideas for the future

- Using residue contacts in a graph convolutional neural network
- Training on more deep mutational scanning studies
- Finding a better normalization strategy

Questions?

Bibliography I

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Eddy, S. R. (2011). Accelerated profile HMM searches (W. R. Pearson, Ed.). *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Systems*, *6*(1), 116–124.e3. https://doi.org/10.1016/j.cels.2017.11.003

Hecht, M., Bromberg, Y. & Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, *16*(S8). https://doi.org/10.1186/1471-2164-16-s8-s1

Hopf, T., Ingraham, J., Poelwijk, F., Schärfe, C., Springer, M., Sander, C. & Marks, D. (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, *35*(2), 128–135. https://doi.org/10.1038/nbt.3769

Bibliography II

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B. & Marcatili, P. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, *87*(6), 520–527. https://doi.org/10.1002/prot.25674

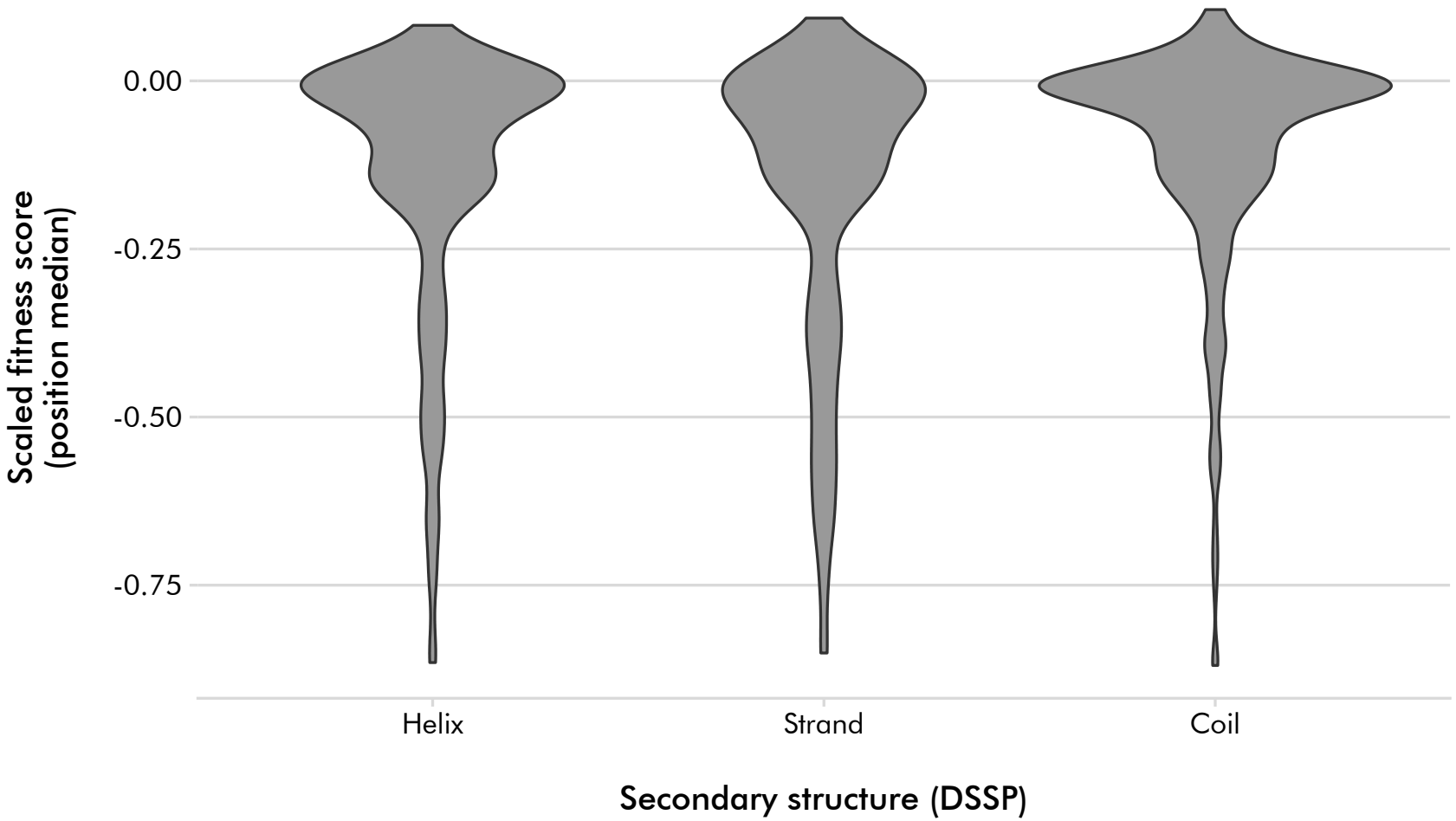
Riesselman, A. J., Ingraham, J. B. & Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, *15*(10), 816–822. https://doi.org/10.1038/s41592-018-0138-4

Sim, N.-l., Kumar, P., Hu, J., Henikoff, S., Schneider, G. & Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*(W1), W452–W457. https://doi.org/10.1093/nar/gks539

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S. & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, *117*(3), 1496–1503. https://doi.org/10.1073/pnas.1914677117

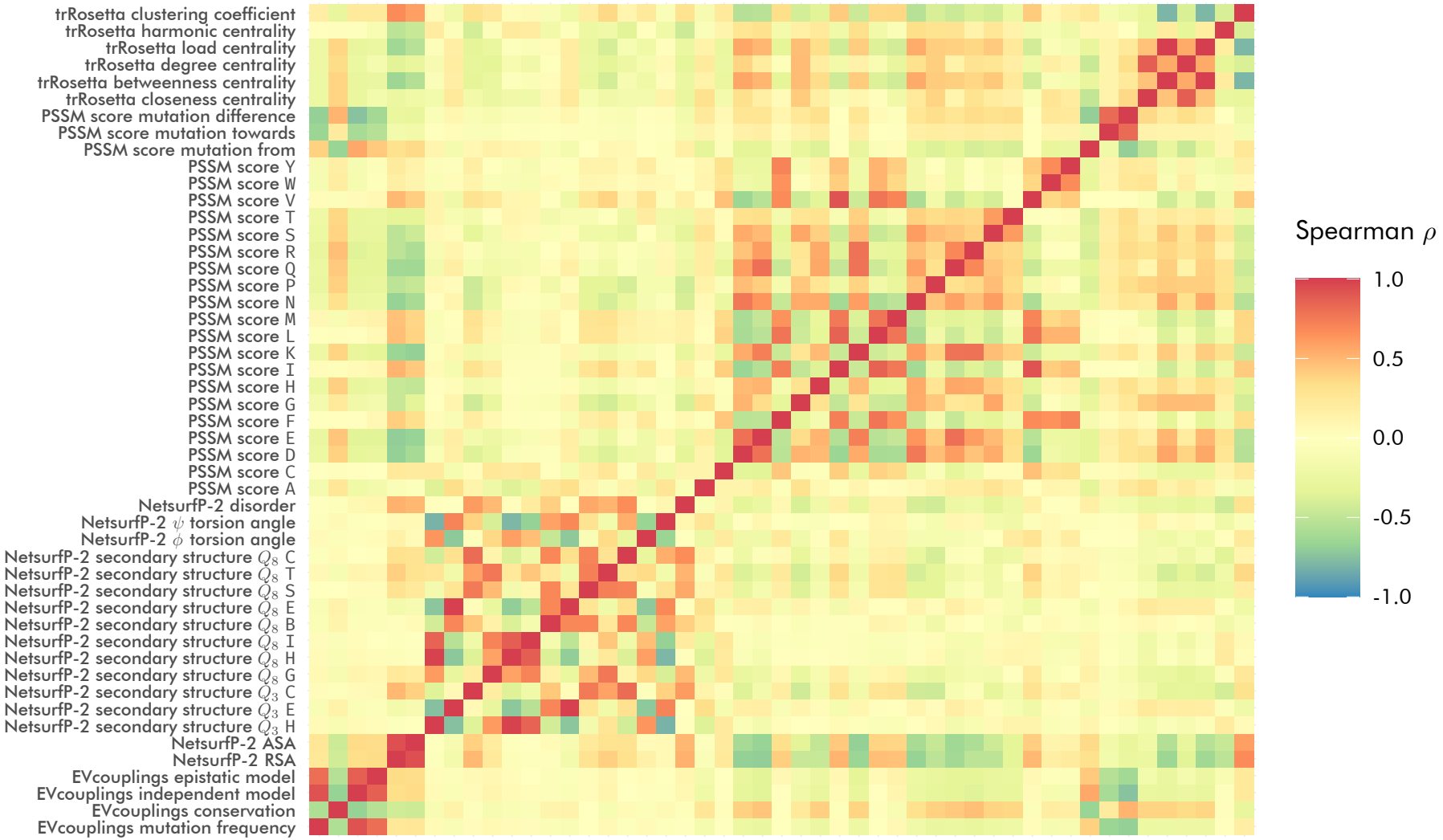
Supplementary Material

Secondary structure is of limited importance in the discrimination of damaging mutations



Supplementary Material

The correlation among features follows predictable patterns



Supplementary Material

Precision of trRosetta (Yang et al., 2020) in predicting residue contacts

	Medium-range ($s \geq 12$)			Long-range ($s \geq 24$)		
Dataset	Top $L/5$	Top $L/2$	Top L	Top $L/5$	Top $L/2$	Top L
beta-lactamase	1.00	0.92	0.86	0.96	0.93	0.76
WW_domain	0.95	0.90	0.83	0.90	0.87	0.75
PSD95pdz3	0.96	0.92	0.80	0.92	0.81	0.70
kka2_1:2	1.00	1.00	0.96	1.00	1.00	0.89
hsp90	1.00	1.00	0.96	1.00	1.00	0.89
Ubiquitin	0.98	0.92	0.82	1.00	0.90	0.70
Pab1	0.80	0.72	0.67	0.87	0.74	0.60
E1_Ubiquitin	0.82	0.86	0.77	0.91	0.75	0.54
gb1	1.00	0.85	0.46	0.63	0.40	0.22

Quality of the predicted structural features from NetsurfP-2 (Klausen et al., 2019)

Feature	Evaluation metric	Score
Relative solvent accessibility	Pearson r	0.79
Accessible surface area	Pearson r	0.80
Q_3 secondary structure	Q_3 accuracy	0.85
Q_8 secondary structure	Q_8 accuracy	0.72
ϕ torsion angle	Circular correlation	0.73
ψ torsion angle	Circular correlation	0.87

Supplementary Material

Relationship between the features used in the models and the fitness scores

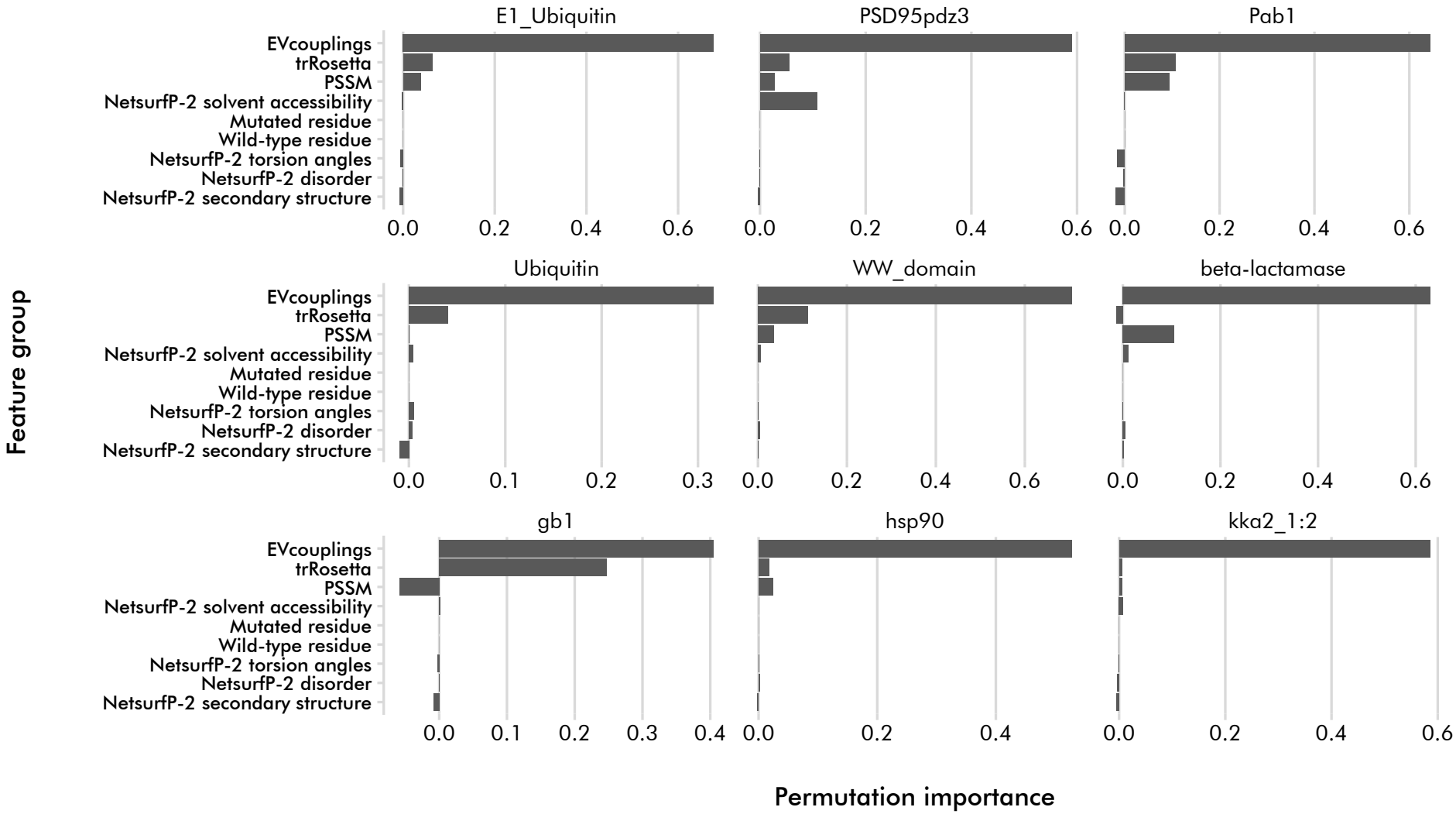
Feature	Pearson r	Spearman ρ	Kendall τ
PSSM mutation score	−0.29	−0.28	−0.19
Netsurf predicted RSA	0.34	0.37	0.25
Netsurf predicted ASA	0.32	0.35	0.24
Netsurf predicted disorder	0.06	0.18	0.12
EVcouplings epistatic model	0.46	0.50	0.34
EVcouplings independent model	0.44	0.44	0.30
EVcouplings frequency	0.19	0.35	0.24
EVcouplings conservation	−0.32	−0.33	−0.23
Closeness centrality (trRosetta predicted contacts)	−0.16	−0.17	−0.11
Betweenness centrality (trRosetta predicted contacts)	−0.20	−0.29	−0.19
Degree centrality (trRosetta predicted contacts)	−0.12	−0.13	−0.09
Load centrality (trRosetta predicted contacts)	−0.20	−0.29	−0.19
Harmonic centrality (trRosetta predicted contacts)	−0.19	−0.20	−0.14
Clustering coefficient (trRosetta predicted contacts)	0.23	0.25	0.17

	Linear-circular correlation
Netsurf predicted ϕ torsion angle	0.01
Netsurf predicted ψ torsion angle	0.02

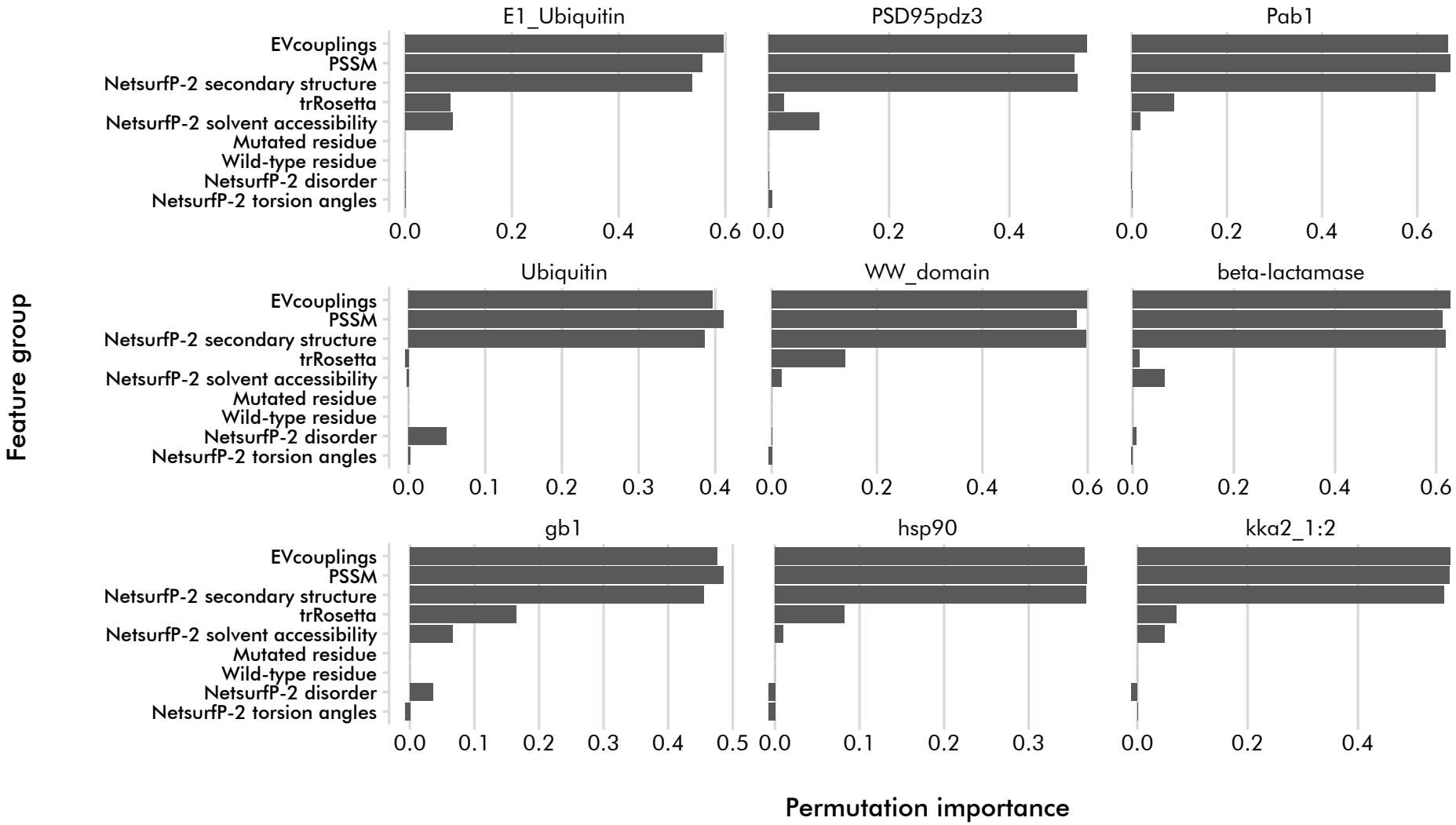
	Kruskal-Wallis χ^2	p-value
Wild-type residue	1482.40	$< 2.20 \cdot 10^{-16}$
Mutated residue	708.53	$< 2.20 \cdot 10^{-16}$
Netsurf predicted Q_3 secondary structure	215.33	$< 2.20 \cdot 10^{-16}$
Netsurf predicted Q_8 secondary structure	351.97	$< 2.20 \cdot 10^{-16}$

Supplementary Material

Feature importances for the gradient boosted tree general models

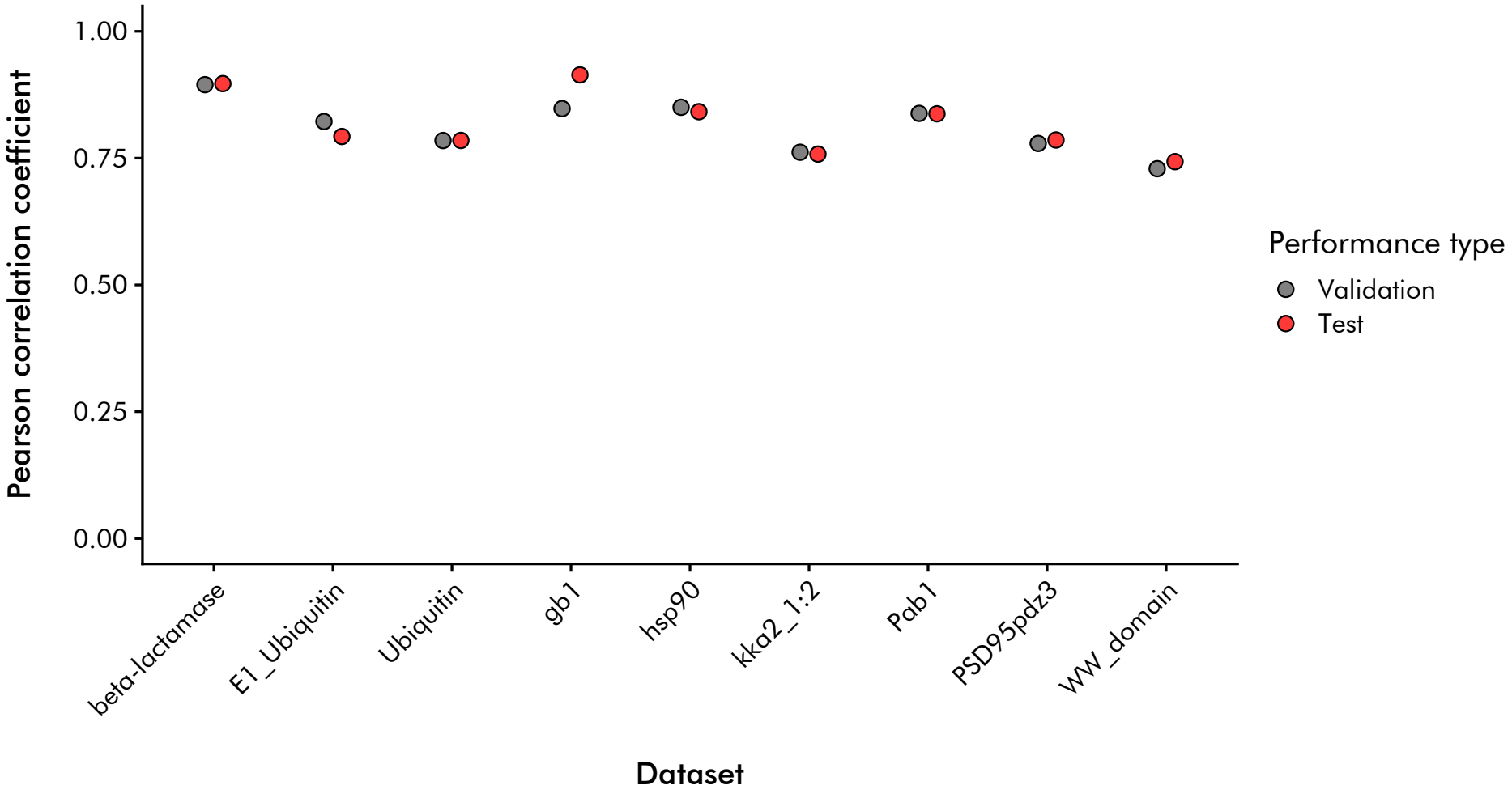


Feature importances for the linear regression general models



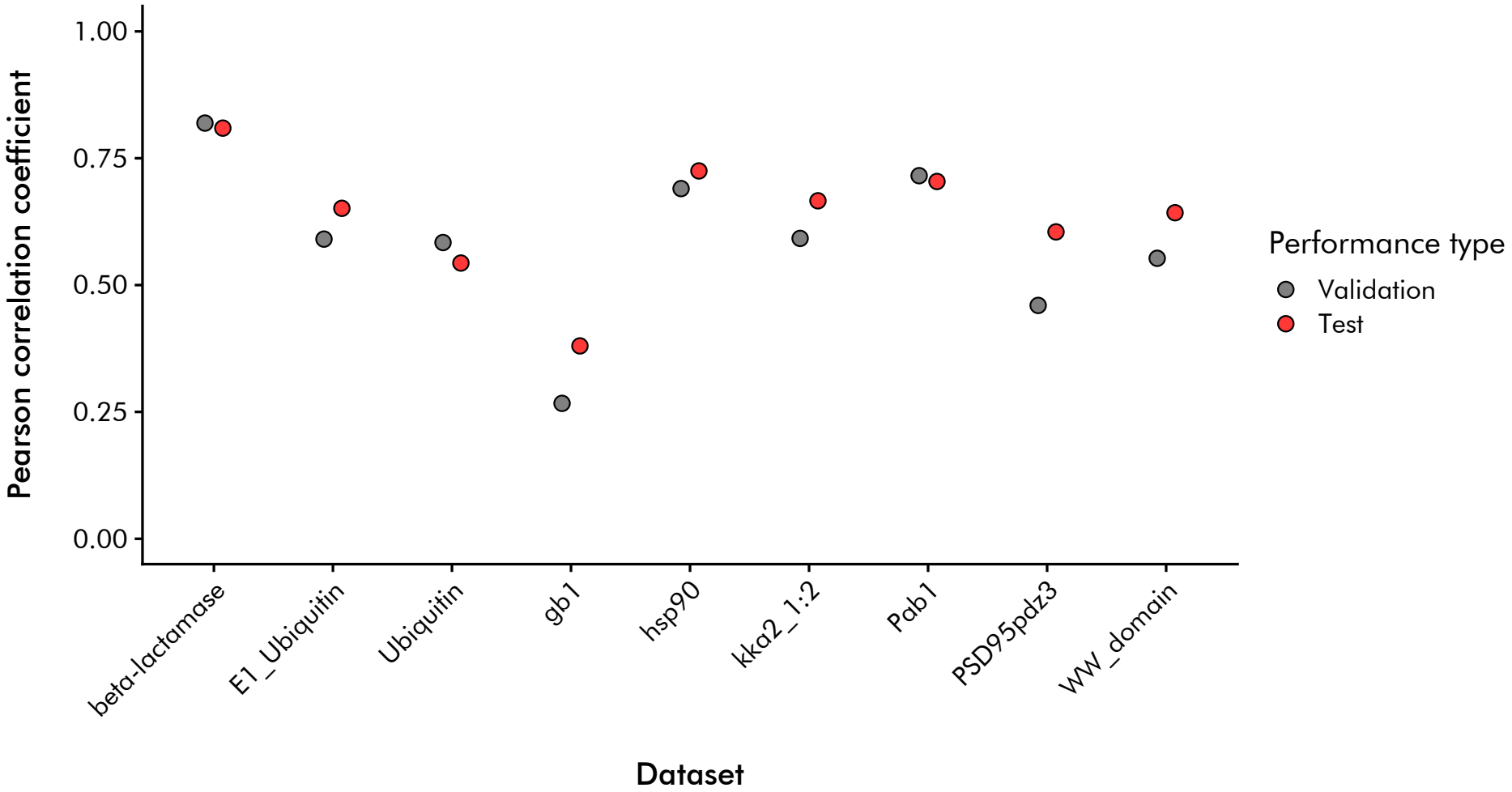
Supplementary Material

Validation and testing performances for the single protein models trained with the naive approach



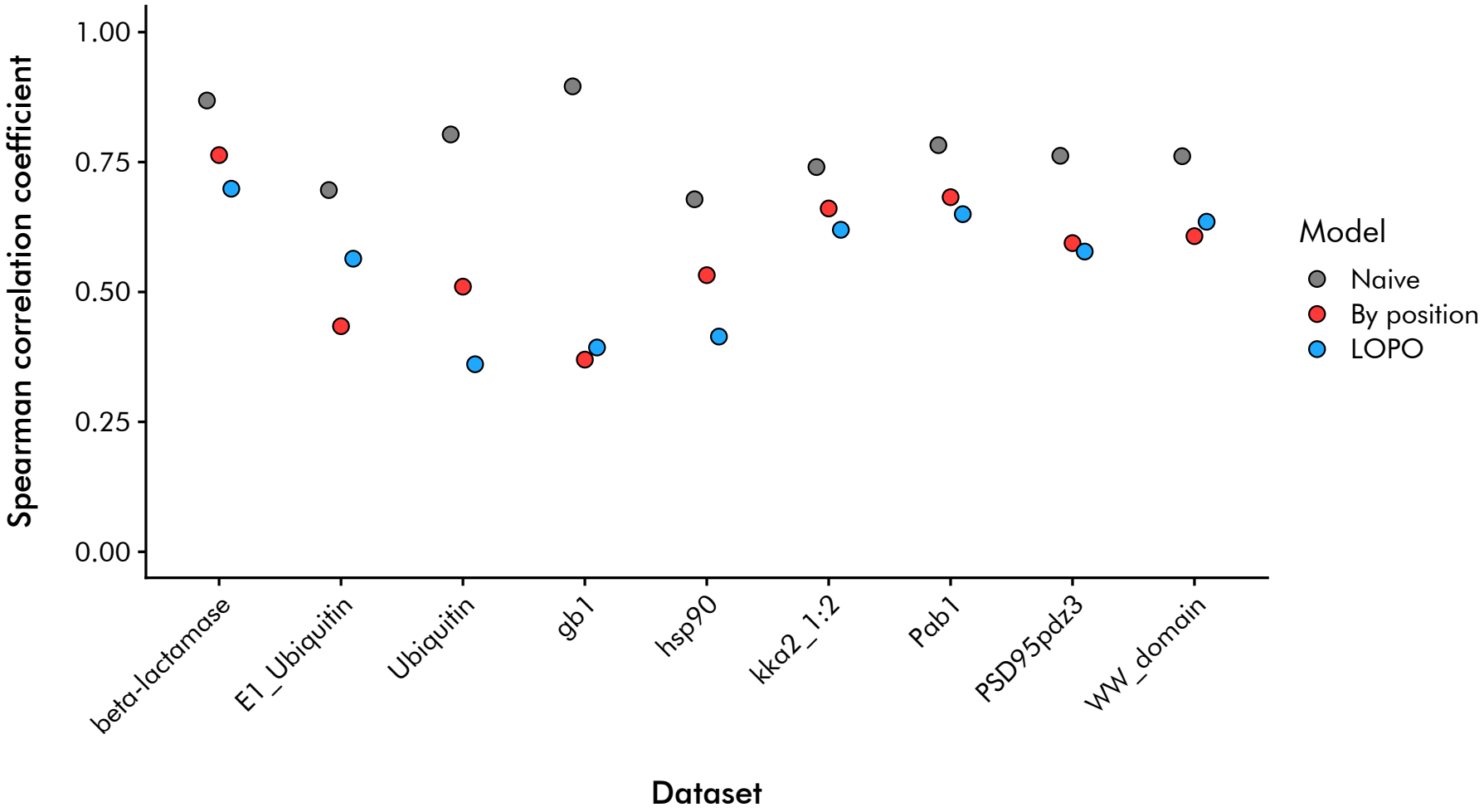
Supplementary Material

Validation and testing performances for the single protein models trained by segregating protein positions



Supplementary Material

Comparison of the performances of single protein models and general models



Confidence intervals in prediction performances

Dataset name	Model	95 % C.I. (Pearson)	95 % C.I. (Spearman)
beta-lactamase	Naive	0.89 to 0.91	0.86 to 0.88
beta-lactamase	By position	0.79 to 0.83	0.75 to 0.78
beta-lactamase	LOPO	—	0.68 to 0.72
WW_domain	Naive	0.67 to 0.82	0.70 to 0.84
WW_domain	By position	0.57 to 0.73	0.52 to 0.72
WW_domain	LOPO	—	0.56 to 0.73
PSD95pdz3	Naive	0.74 to 0.83	0.73 to 0.80
PSD95pdz3	By position	0.55 to 0.67	0.54 to 0.65
PSD95pdz3	LOPO	—	0.53 to 0.63
kka2_1:2	Naive	0.74 to 0.78	0.72 to 0.76
kka2_1:2	By position	0.65 to 0.69	0.64 to 0.68
kka2_1:2	LOPO	—	0.60 to 0.64
hsp90	Naive	0.82 to 0.87	0.65 to 0.71
hsp90	By position	0.69 to 0.76	0.50 to 0.57
hsp90	LOPO	—	0.38 to 0.45
Ubiquitin	Naive	0.75 to 0.83	0.78 to 0.83
Ubiquitin	By position	0.49 to 0.60	0.46 to 0.57
Ubiquitin	LOPO	—	0.30 to 0.43
Pab1	Naive	0.80 to 0.87	0.75 to 0.82
Pab1	By position	0.65 to 0.76	0.64 to 0.73
Pab1	LOPO	—	0.60 to 0.70
E1_Ubiquitin	Naive	0.75 to 0.85	0.65 to 0.75
E1_Ubiquitin	By position	0.59 to 0.72	0.36 to 0.51
E1_Ubiquitin	LOPO	—	0.50 to 0.63
gb1	Naive	0.90 to 0.93	0.88 to 0.92
gb1	By position	0.31 to 0.46	0.29 to 0.45
gb1	LOPO	—	0.32 to 0.47

Supplementary Material

Statistical significance of performance differences. Starred values are significant with Bonferroni correction.

$$\alpha = \frac{0.05}{27} = 0.00185185$$

Dataset name	Model 1	Model 2	p-value
beta-lactamase	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
beta-lactamase	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
beta-lactamase	EVmutation	Gradient boosted trees	$1 \cdot 10^{-4}$ *
WW_domain	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
WW_domain	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
WW_domain	EVmutation	Gradient boosted trees	$1 \cdot 10^{-4}$ *
PSD95pdz3	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
PSD95pdz3	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
PSD95pdz3	EVmutation	Gradient boosted trees	0.51
kka2_1:2	Linear regression	Gradient boosted trees	0.00
kka2_1:2	Linear regression	EVmutation	0.01
kka2_1:2	EVmutation	Gradient boosted trees	$1 \cdot 10^{-4}$ *
hsp90	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
hsp90	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
hsp90	EVmutation	Gradient boosted trees	0.24

The table continues on the next slide

Supplementary Material

Statistical significance of performance differences. Starred values are significant with Bonferroni correction.

$$\alpha = \frac{0.05}{27} = 0.00185185$$

The table continues from the previous slide

Dataset name	Model 1	Model 2	p-value
Ubiquitin	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
Ubiquitin	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
Ubiquitin	EVmutation	Gradient boosted trees	$1 \cdot 10^{-4}$ *
Pab1	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
Pab1	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
Pab1	EVmutation	Gradient boosted trees	$1 \cdot 10^{-4}$ *
E1_Ubiquitin	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
E1_Ubiquitin	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
E1_Ubiquitin	EVmutation	Gradient boosted trees	$1 \cdot 10^{-4}$ *
gb1	Linear regression	Gradient boosted trees	$1 \cdot 10^{-4}$ *
gb1	Linear regression	EVmutation	$1 \cdot 10^{-4}$ *
gb1	EVmutation	Gradient boosted trees	$1 \cdot 10^{-4}$ *