Sequence- and structure-based prediction of protein stability change due to single mutations

Skoltech

Dmitry Ivankov

Protein design and redesign

Skoltech

Nobel prize 2024

Protein design

Daniela Röthlisberger[1]*, Olga Khersonsky[4]*, Andrew M. Wollacott[1]*, Lin Jiang[1,2], Jason DeChancie[6], Jamie Betker[3], Jasmine L. Gallaher[3], Eric A. Althoff[1], Alexandre Zanghellini[1,2], Orly Dym[5], Shira Albeck[5], Kendall N. Houk[6], Dan S. Tawfik[4] & David Baker[1,2,3]

Kemp elimination catalysts by computational enzyme design

doi:10.1038/nature06879

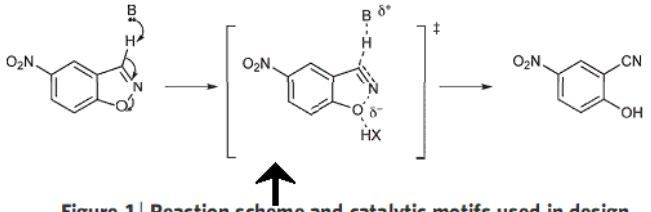
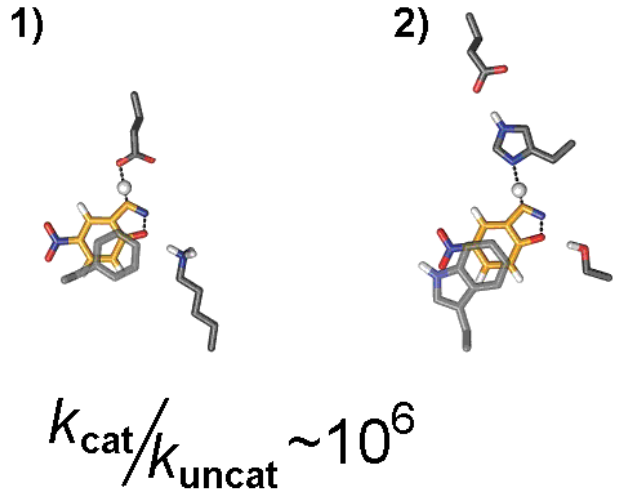


Figure 1 | Reaction scheme and catalytic motifs used in design.



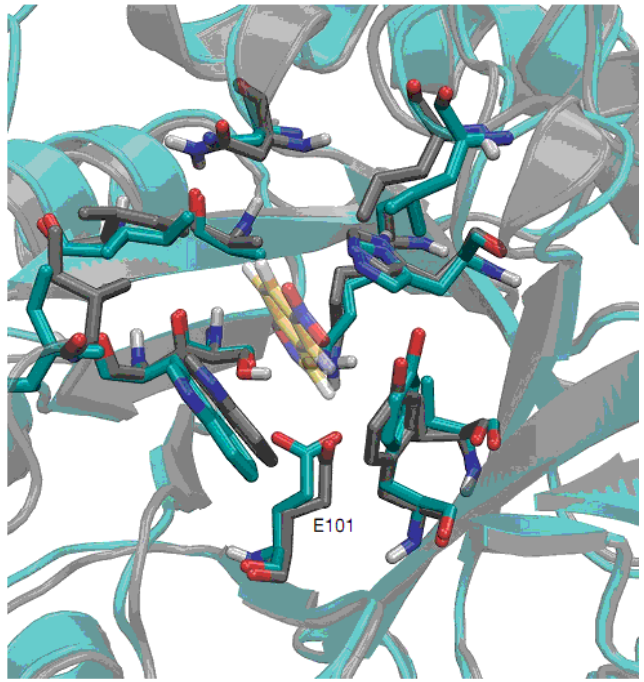
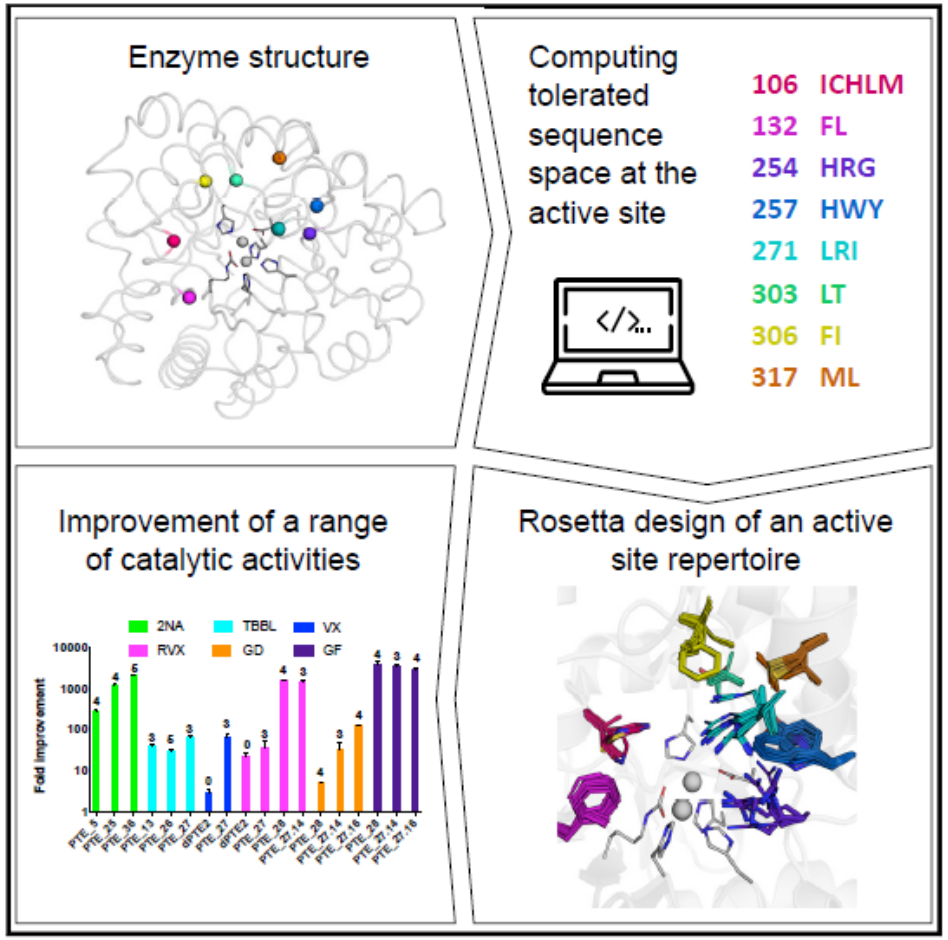
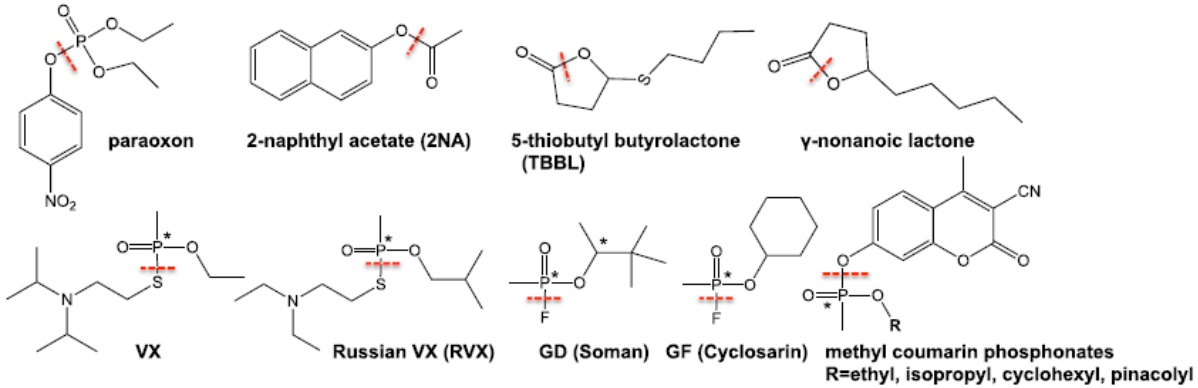


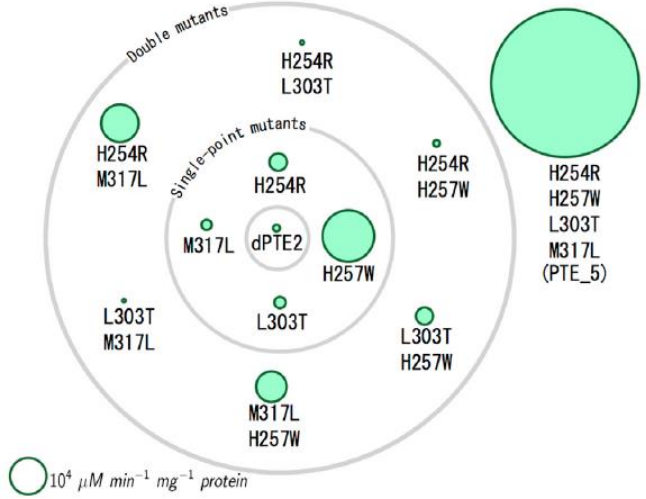
Figure 4 | Comparison of the designed model of KE07 and the crystal structure.

Skoltech

Protein redesign: phosphotriesterase







Skoltech

Khersonsky et al., Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. Mol. Cell, 2018.

Why protein redesign: enzymes in washing powder

- Enzymes added to washing powder:
  - Proteases – break down protein chains from stains;
  - Lipases – break down fats and oils in stains;
  - Amylases – break down starch;
  - Cellullases – break down cellulose;
  - Mannanases – break down mannans.
- Enzymes work at normal temperatures
- We need to increase their thermostability to allow for washing at higher temperatures

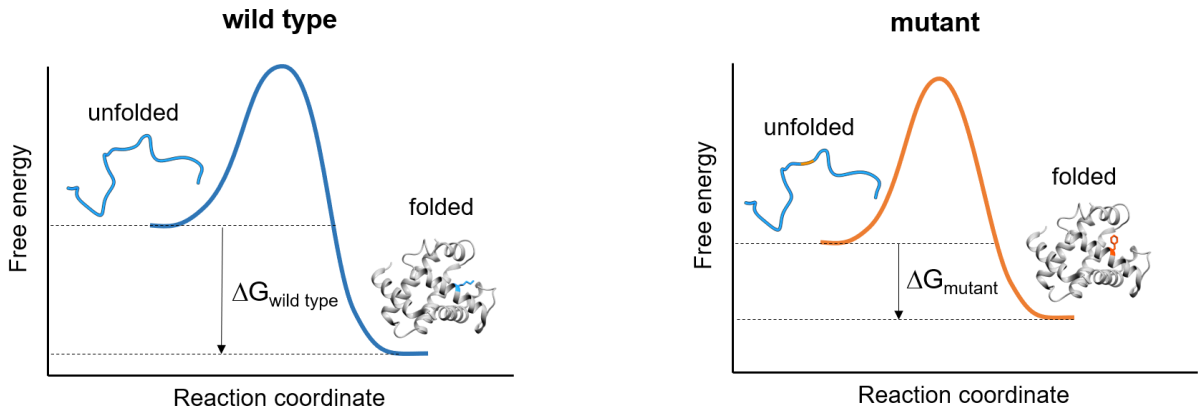
Skoltech

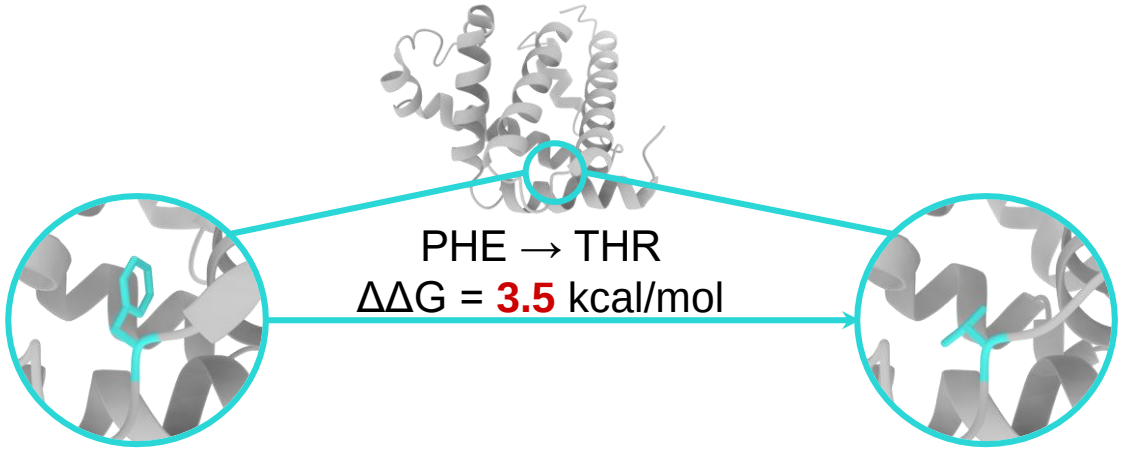
https://www.persil.com/uk/laundry/detergent/capsules.html

Change of protein stability on mutation

$$\Delta G = G_{folded} - G_{unfolded}$$

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\ type}$$



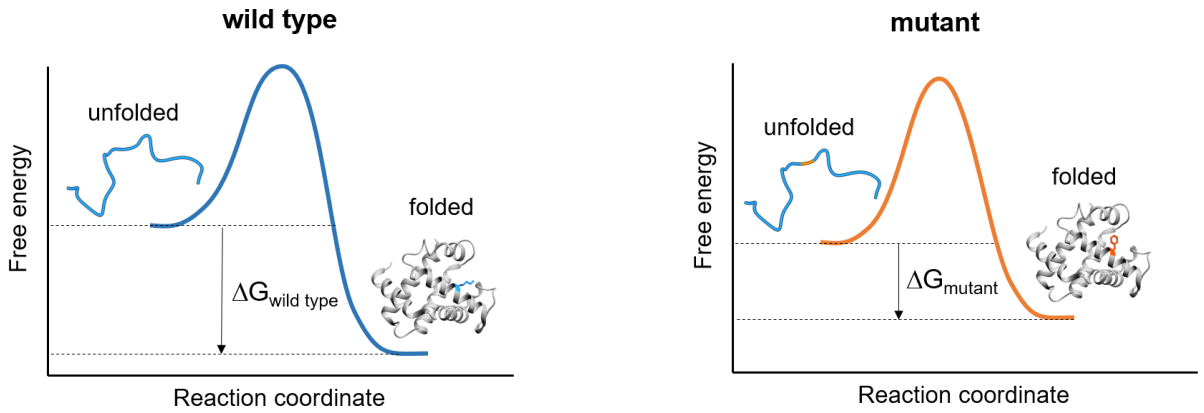


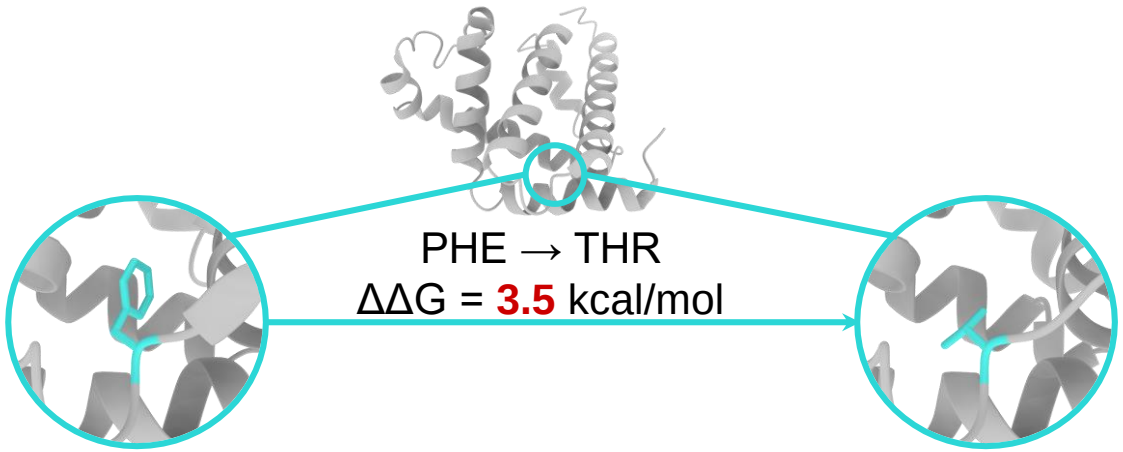
Skoltech

ΔΔG prediction: simplest task of protein design

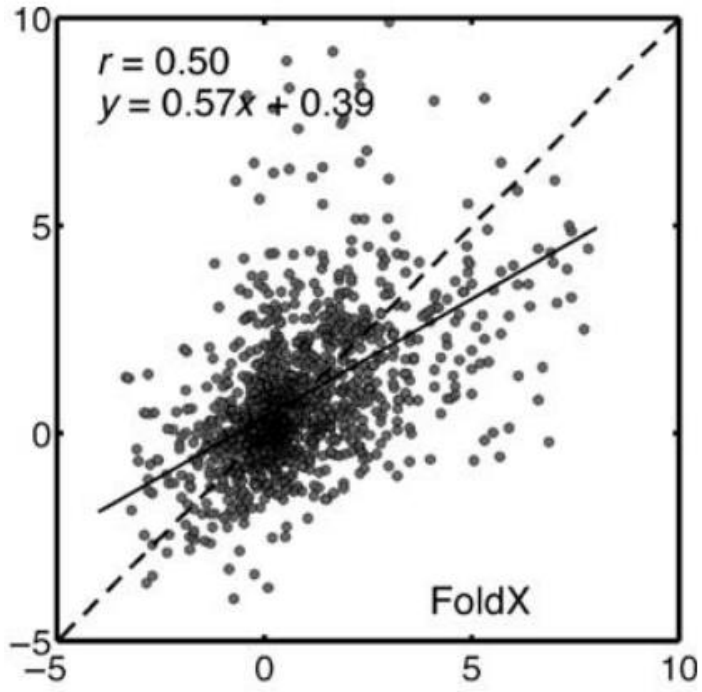
$$\Delta G = G_{folded} - G_{unfolded}$$

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\ type}$$





- Important for protein engineering
- Performance is ~ 50-60% (Pearson correlation)

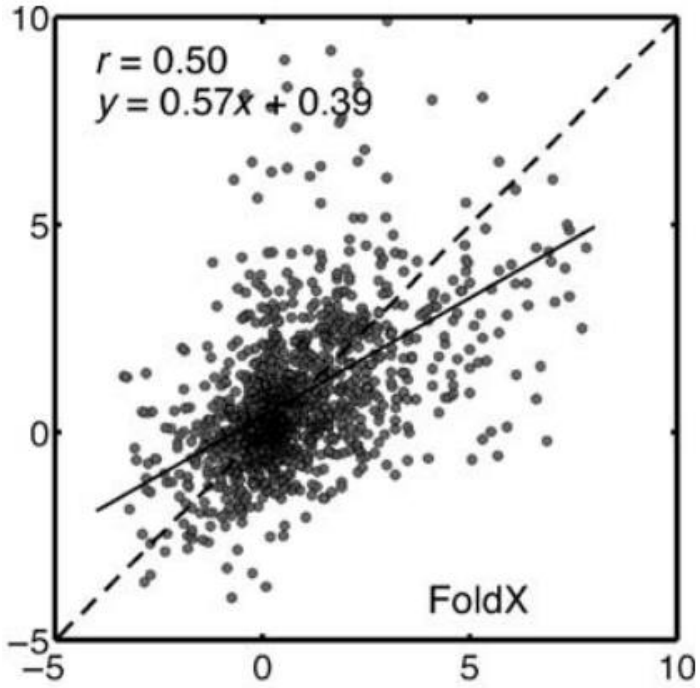


Skoltech

The number of predictors is 40+

PremPS
BoostDDG
I-Mutant ACDC
FoldX Maestro
HoTMuSiC INPS
SDM DynaMut mCSM
AUTO-MUTE DUET
STRUM PoPMuSiC
ThermoNet
DDGun EASE-MM
Eris Rosetta

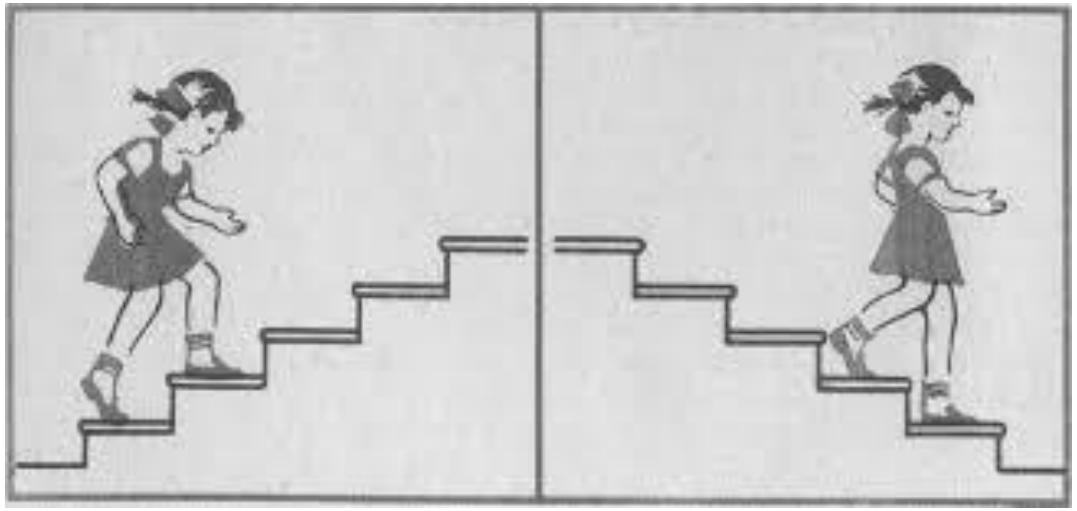
- Correlation ~ 50-60%



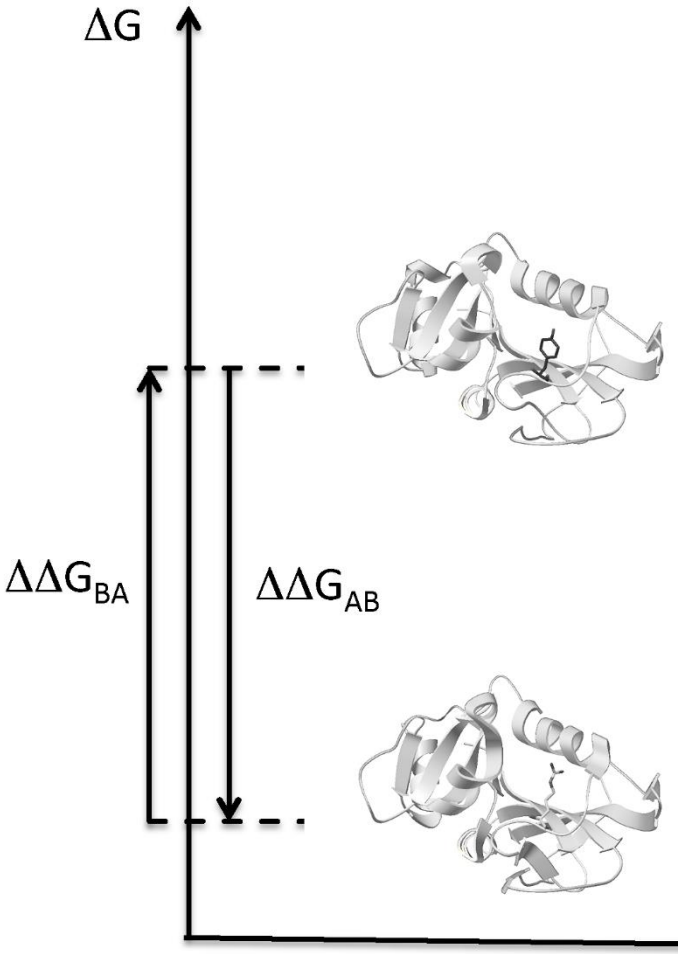
Skoltech

Predictors overestimate ΔΔG

- How to measure the overestimation?



Δh = 1m Δh = -1m

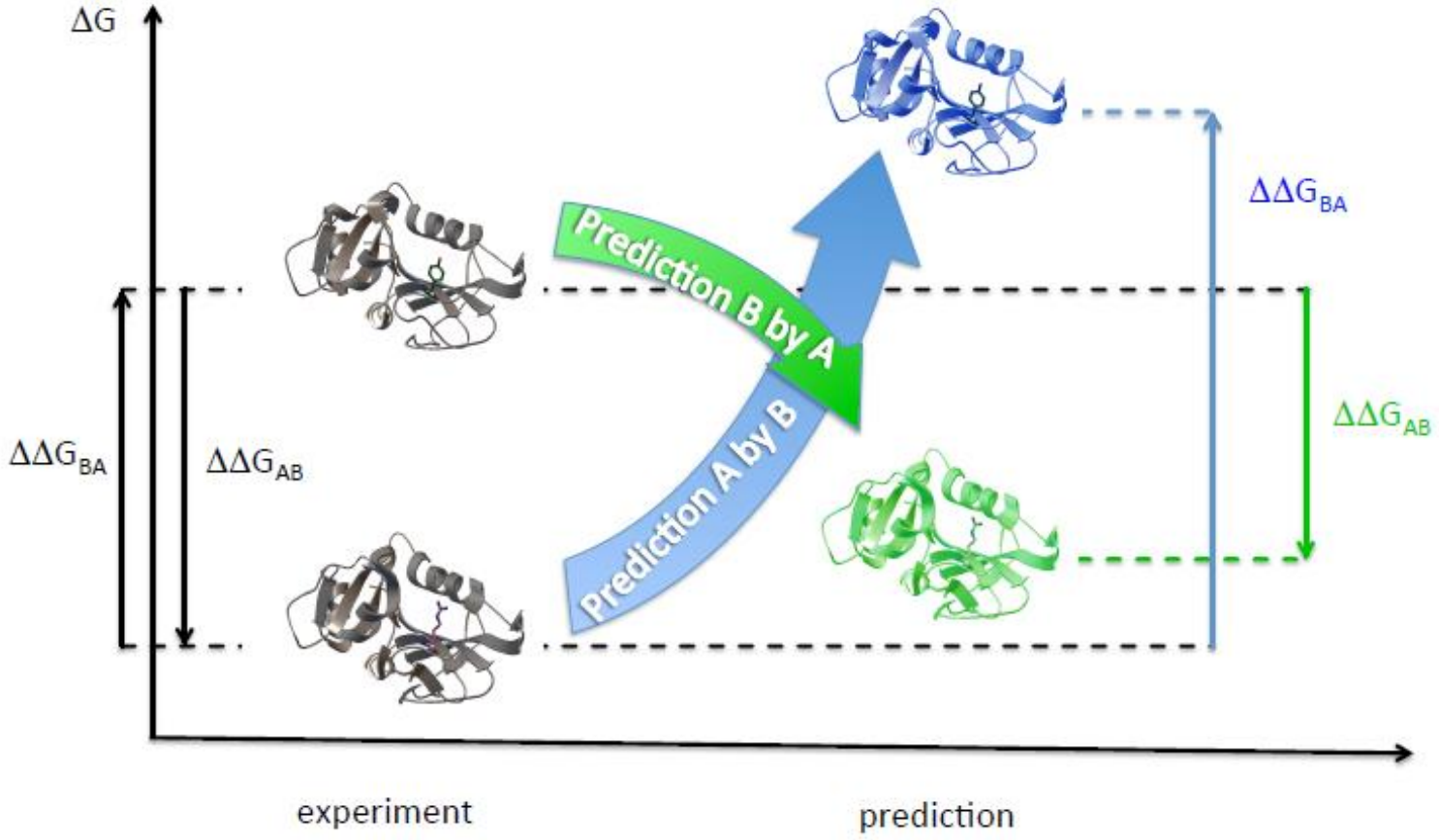


Skoltech

Usmanova et al., 2018, Bioinformatics

Self-consistency test

- We do not need experimental ΔΔG data!

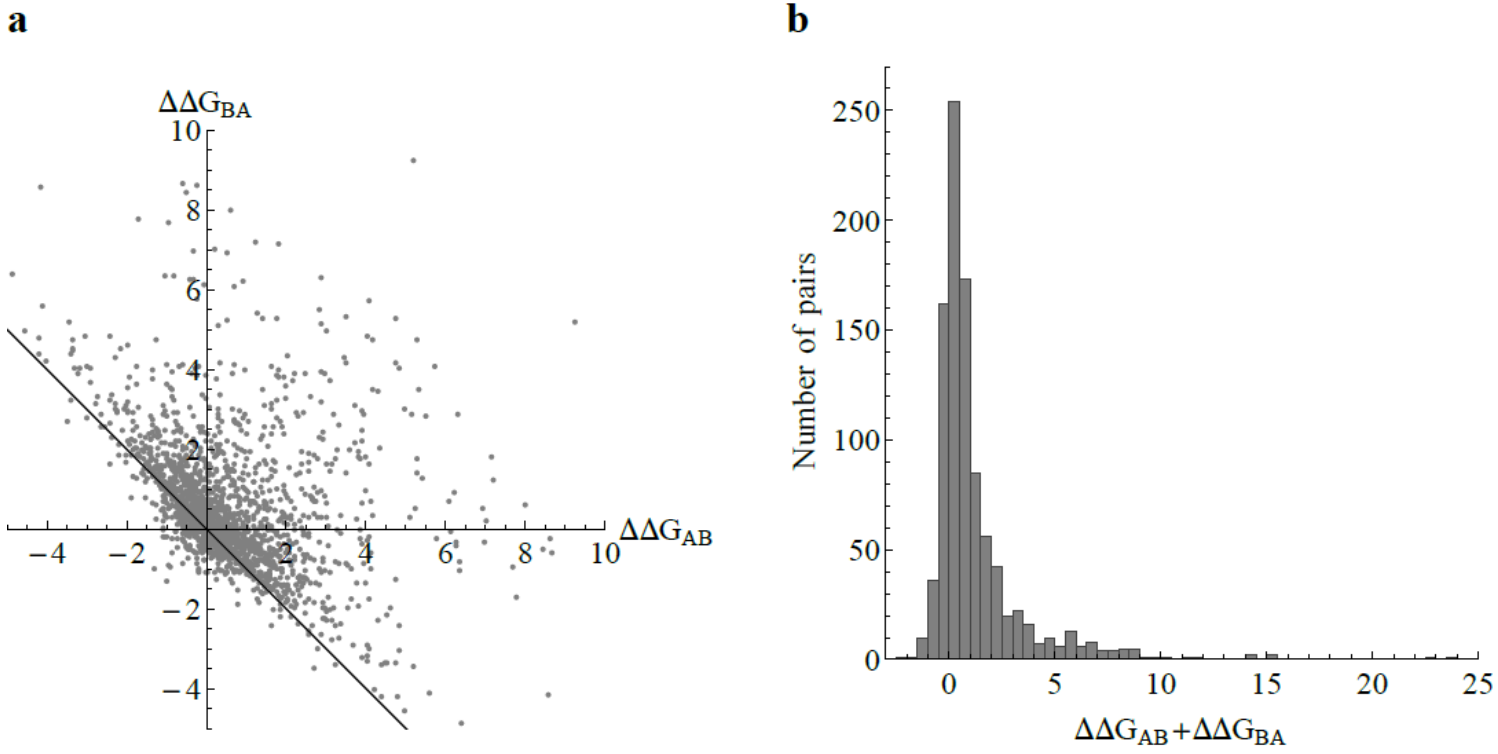


Skoltech

Usmanova et al., 2018, Bioinformatics

Bias for FoldX

- Equals 0.72 kcal/mol per single mutation
- Structure A is not optimal for new amino acid residue

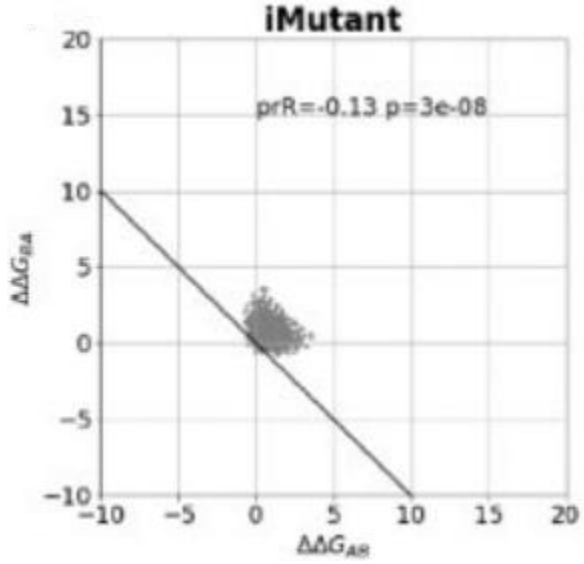


Skoltech

Usmanova et al., 2018, Bioinformatics

Bias for iMutant

- Equals 0.80 kcal/mol per single mutation
- Reflects the trend of the training dataset: most mutations are deleterious



Skoltech

Usmanova et al., 2018, Bioinformatics

How to exclude the bias? (1/2)

- Data symmetrization:

Myoglobin1	A13M	2kcal/mol
Myoglobin2	M13A	-2kcal/mol

- All new predictors after 2018 are symmetrized

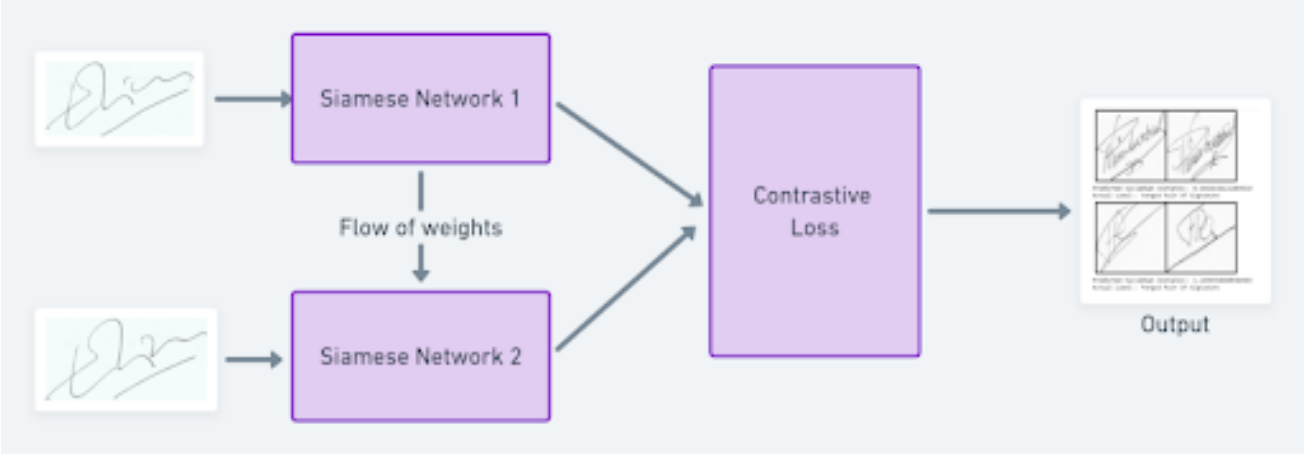
Skoltech

How to exclude the bias? (2/2)

- Predictor symmetrization during learning:

ADHase1	S123T	Xkcal/mol
ADHase2	T123S	-Xkcal/mol

- Siamese neural network architecture



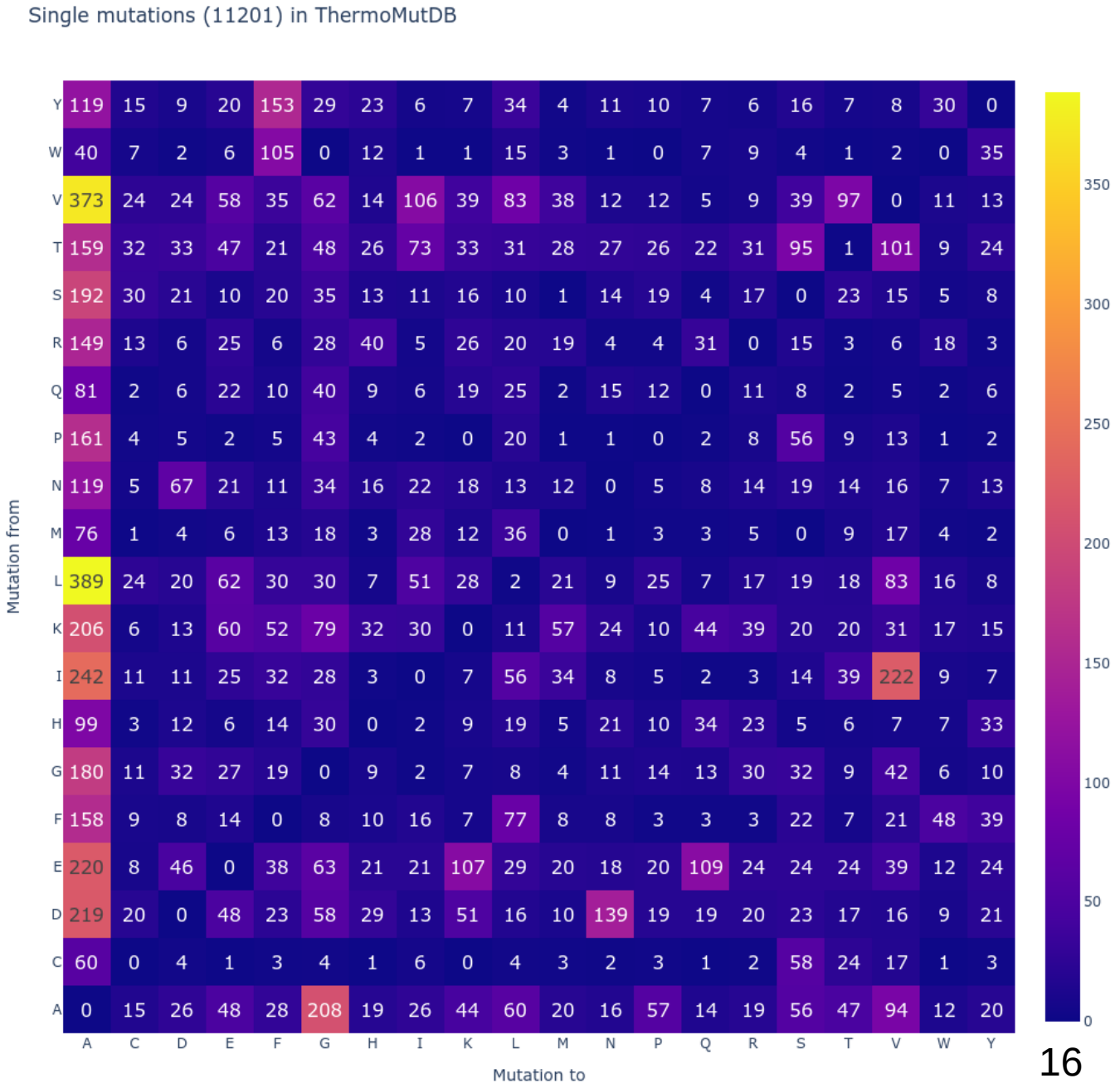
https://builtin.com/machine-learning/siamese-network

Skoltech

Experimental dataset is unbalanced

- ThermoMutDB

11 201 single mutations

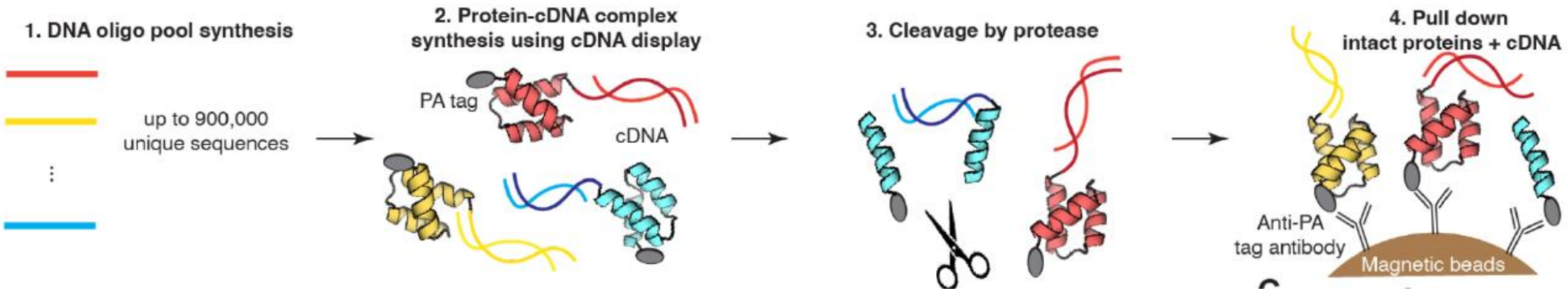


Skoltech

851 552 new mutations / 376 918 single

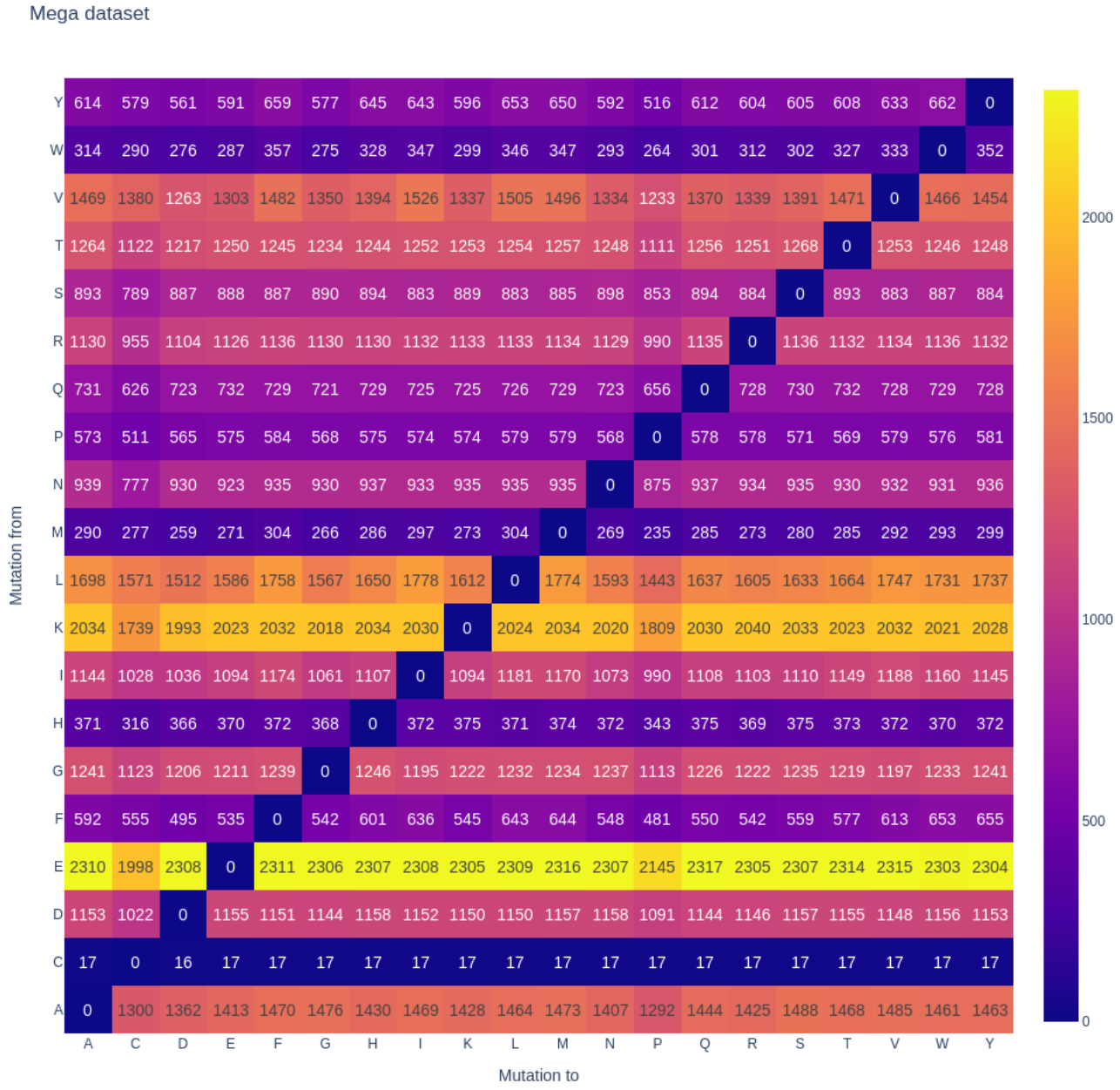
Mega-scale experimental analysis of protein folding stability in biology and protein design

Authors: Kotaro Tsuboyama[1,2,3], Justas Dauparas[4,5], Jonathan Chen[1,2], Niall M. Mangan[2,7], Sergey Ovchinnikov[8], Gabriel J. Rocklin[1,2] *



Skoltech

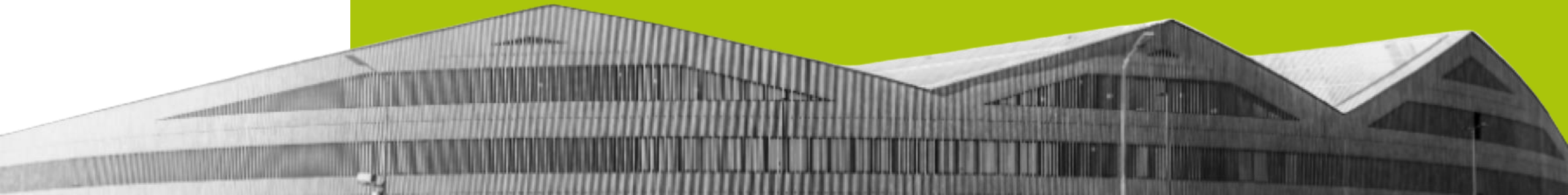
Statistics of single mutations for Mega-dataset



Skoltech

Sequence-based ΔΔG prediction

Skoltech



Dataset and Design

Data:

Mega dataset: All possible single-point mutations in 396 proteins

Tsuboyama et al. (2023). Nature, 620, 434.

Protein representation:

ESM-2 embeddings

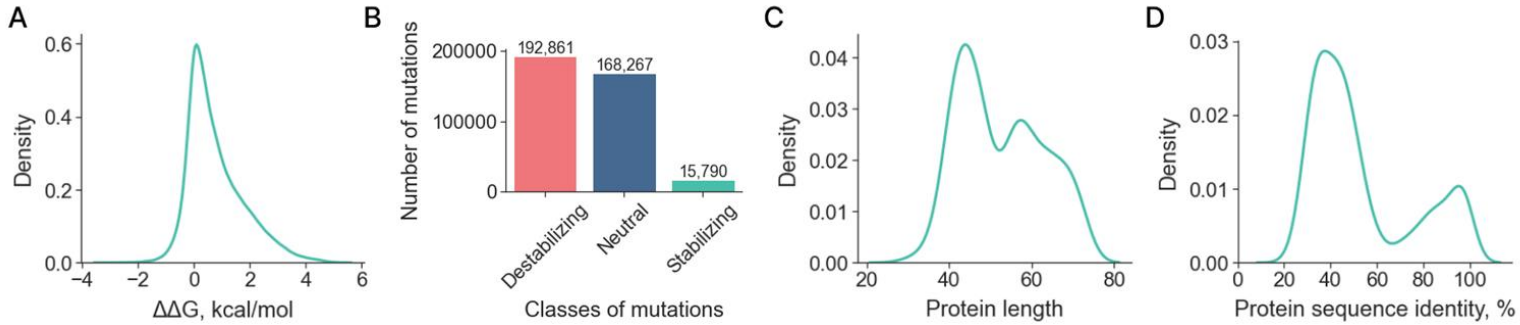
Lin et al. (2023). Science, 379, 1123.

Antisymmetry of ΔΔG prediction:

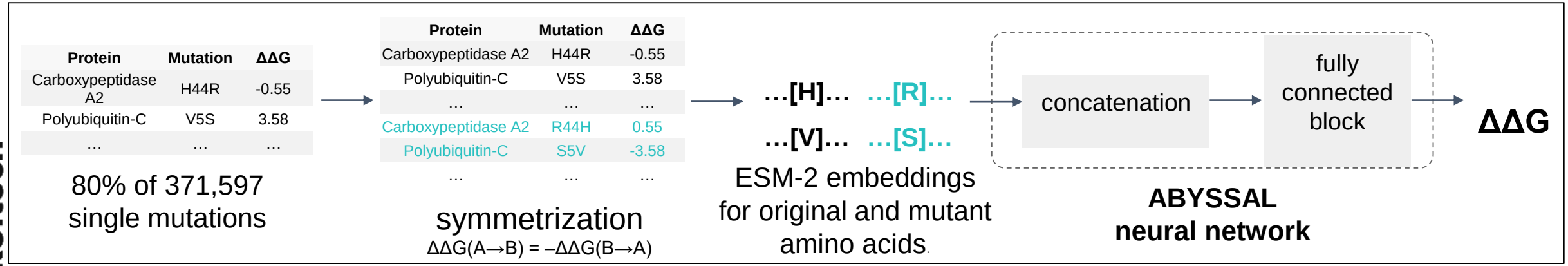
Dataset symmetrization

Siamese network

Bromley et al. (1993). International Journal of Pattern Recognition and Artificial Intelligence. 7, 669.



Description of the filtered Mega dataset



Design of the ΔΔG predictor

ABYSSAL performance

ABYSSAL outperformed other predictors on unseen subset of Mega dataset.

Predictor	PCC	SCC	MSE, kcal/mol	Accuracy
ABYSSAL	**0.76±0.01**	**0.71±0.01**	**0.67**	**0.75**
DeepDDG	0.70±0.01	0.58±0.01	1.01	0.72
INPS 3D	0.69±0.01	0.61±0.01	0.78	0.73
DDGun 3D	0.66±0.01	0.51±0.01	1.00	0.67
INPS	0.61±0.01	0.56±0.01	0.88	0.72

Performance of predictors on new data: Mega Holdout dataset (5321 mutations in 5 proteins)

Tsuboyama et al. (2023). Nature, 620, 434.

On old data ABYSSAL is comparable with top-performing predictors implying the ceiling of 50% PCC on this type of data.

Predictor	Symmetric data				PCC (f-r)	<δ>
	PCC	SCC	MSE, kcal/mol	Accuracy		
INPS-Seq	0.50±0.03	0.51±0.03	1.74	0.66	-0.99	0.00
ABYSSAL	**0.49±0.03**	**0.48±0.03**	**1.74**	**0.63**	**-0.98**	**0.02**
PremPS	0.49±0.03	0.48±0.03	1.75	0.67	-0.84	0.06
ACDC-NN3D	0.49±0.03	0.47±0.03	1.74	0.65	-0.98	-0.02
ACDC-NN	0.47±0.03	0.45±0.03	1.76	0.64	-1.00	0.00

Performance of predictors on old data: S669 dataset (420 mutations in 86 proteins)

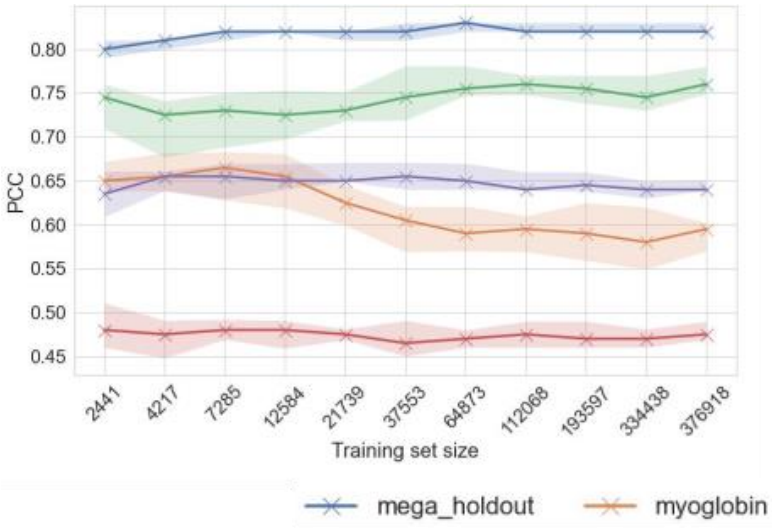
Pancotti et al. (2022). Briefings in Bioinformatics, 23(2).

Skoltech

Factors influencing performance

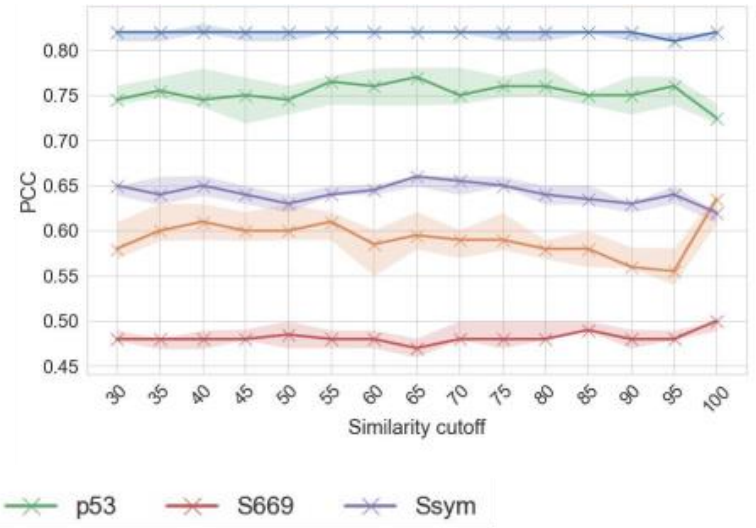
Data quality is the key factor influencing performance.

No significant change in performance when trained on a subset of Mega dataset as low as 2441 mutations.



Influence of training set size

Protein sequence identity cutoff for train-test split does not influence performance. Naive random split approach shows the same performance.



Influence of train-test splits by protein sequence identity

ABYSSAL ranks in the top-5 on Mega dataset when trained on old data of S2648.

	New data (Mega train)	Old data (S2648)
New data (Mega Holdout)	0.84±0.01	0.75±0.01
Old data (S669)	0.49±0.03	0.50±0.03

Dehouck, Y. et al. (2009). Bioinformatics, 25, 2537.

Influence of type of training data

Skoltech

Conclusion #1

- Transformer-based siamese network trained on symmetrized ESM-2 embeddings achieves top performance in ΔΔG prediction.

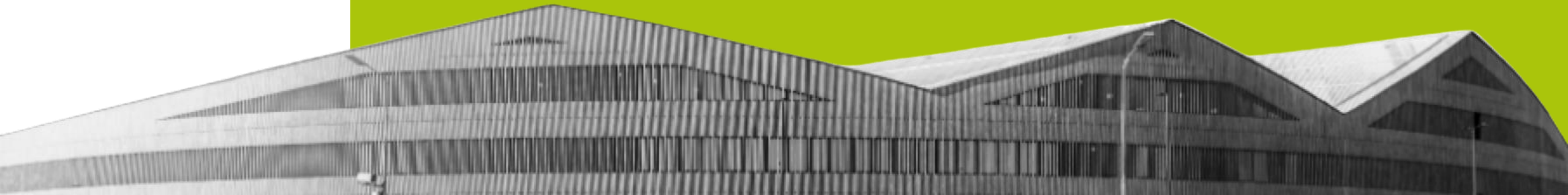
- Training set size and splitting strategy do not influence the performance much, while dataset quality is the key factor.

Skoltech

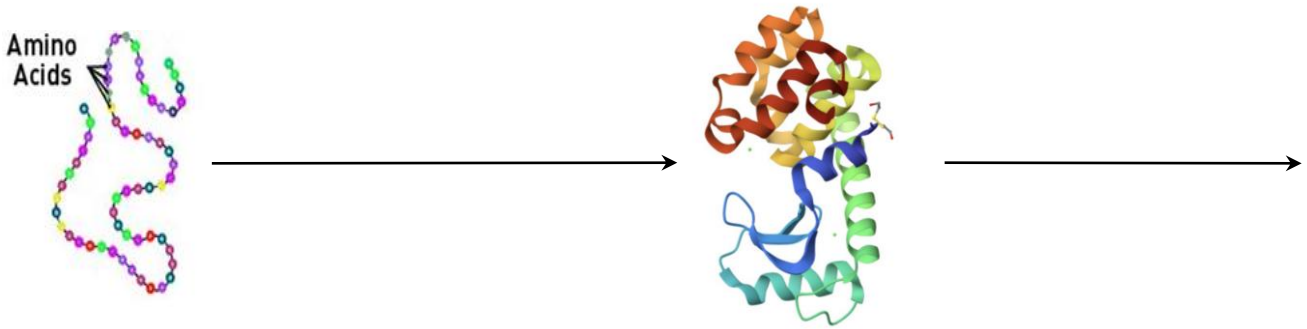
https://github.com/ivankovlab/abyssal

Structure-based ΔΔG prediction

Skoltech



Protein representation learning task



Protein properties
Protein interactions

Protein sequence
[Available]

Protein structure
[Available with AlphaFold]

Protein language models (PLM), eg. ESM-2 [1]

- Graph Neural Networks [3]
- Structure-aware PLMs
  - trained from scratch (SaProt [4])
  - add large number of parameters

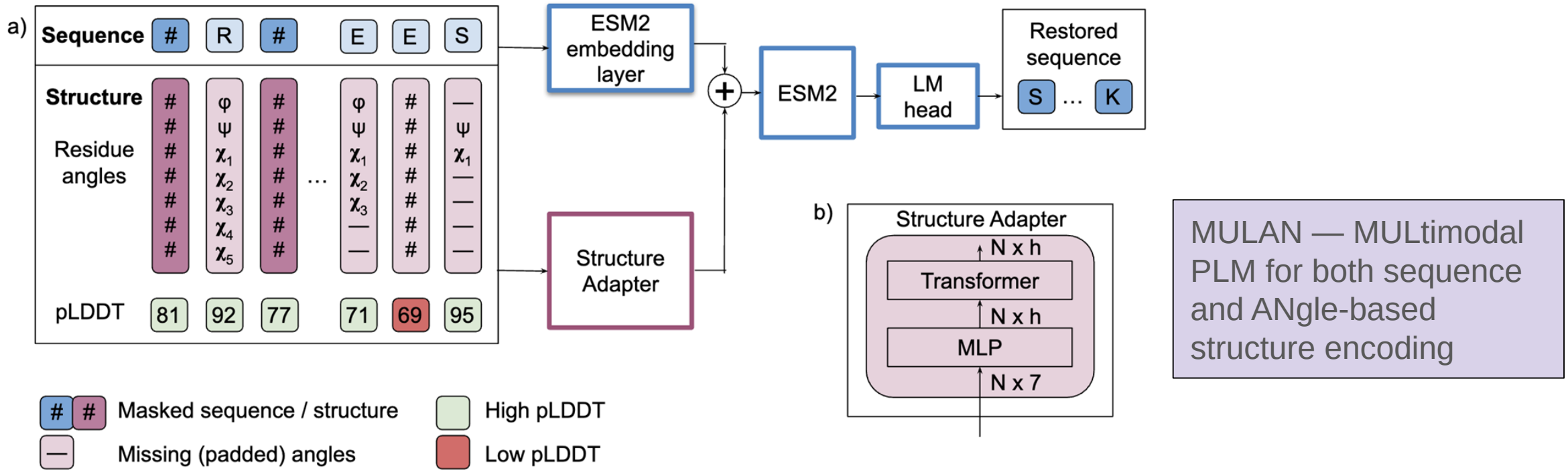
Goals:

- improve protein representation quality with the use of the novel structure-aware PLM;
- add protein structural information to a transformer model with a lightweight finetuning of existing PLMs.

Image credits: https://byjus.com/biology/proteins-structure-and-functions/

[1] Language models of protein sequences at the scale of evolution enable accurate structure prediction, Lin Z. et al., 2022
[2] Highly accurate protein structure prediction with AlphaFold, Jumper J. et al., 2021
[3] Diffdock: Diffusion steps, twists, and turns for molecular docking, Corso G. et al., 2022
[4] Saprot: Protein language modeling with structure-aware vocabulary, Su J. et al., 2023

MULAN architecture



MULAN — MULtimodal PLM for both sequence and ANgle-based structure encoding

Figure 2: The architecture of MULAN. a) MULAN processes sequence inputs with the ESM2 embeddings module, while structure inputs are passed to the Structure Adapter. Both sequence and structure embeddings are summed up and passed to the ESM2 model, which is then finetuned. Sequence-only ESM2 modules (blue) are initialized from the pre-trained ESM2 checkpoint. Structure processing modules are shown in pink. b) The architecture of the Structure Adapter.

Skoltech

Experimental setup

- Train on top of existing PLMs:
  - sequence-only ESM-2 8M, 35M, 650M
  - structure-aware SaProt 35M, 650M
- Only finetune base PLM together with the Structure Adapter
- Use dataset with 17M AlphaFold structures for training
- Evaluate protein embeddings on 7 downstream tasks
  - eg. protein property prediction and protein interaction prediction
  - protein embedding = average of all residue embeddings
  - train small downstream model on protein embeddings for each downstream task independently

Skoltech

Results

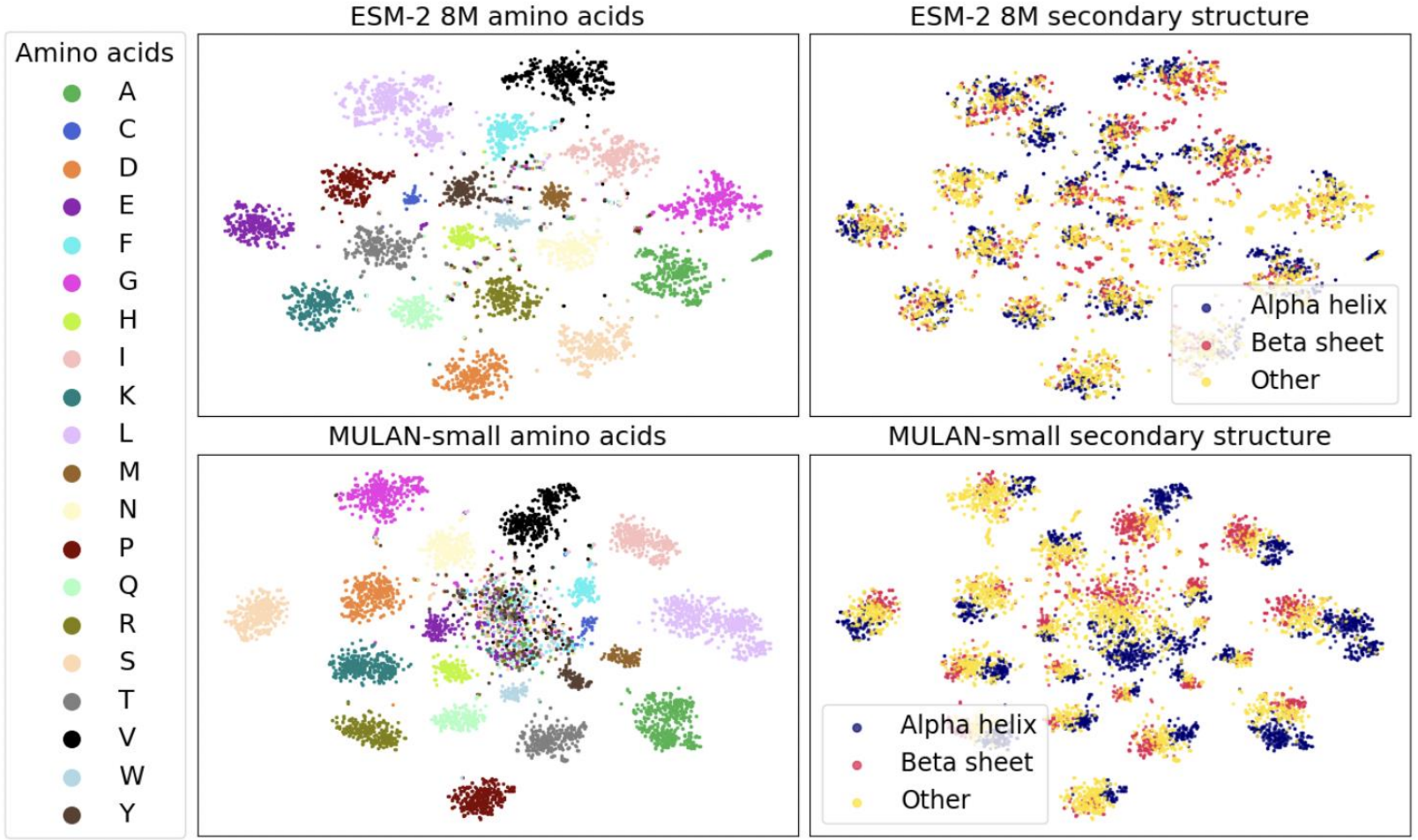
Table 6: The improvement shown by adding MULAN to various PLMs and SPLMs on all downstream tasks. The best results for each base model are shown in bold

Model name	Thermo-stability	Fluore-scence	Metal Ion Binding	Human PPI	GO		
					CC	MF	BP
	SCC ↑	SCC ↑	AUC ↑	AUC ↑	F_{max} ↑	F_{max} ↑	F_{max} ↑
Small models							
ESM-2 8M	.666	.579	.731	.698	.490	.529	.400
Δ MULAN-small 8M	**.006**	**.017**	**.047**	**.055**	**.002**	**.058**	**.026**
Medium models							
ESM-2 35M	.689	.592	.793	.751	.489	.621	.443
Δ MULAN-ESM2 35M	**.012**	**.017**	**.001**	**.031**	**.027**	**.015**	**.004**
Saprot AF 35M	.699	.639	.783	.731	.501	**.632**	.440
Δ MULAN-SaProt 35M	**.005**	**.003**	**.017**	**.048**	**.004**	−.001	**.002**
Large models							
ESM-2 650M	.694	.601	.781	.754	**.523**	**.678**	**.479**
Δ MULAN-ESM 650M	**.009**	**.007**	**.013**	**.117**	−.004	−.001	−.004
SaProt AF 650M	**.711**	.668	.776	.720	.540	.658	.464
Δ MULAN-SaProt 650M	−.008	**.001**	**.026**	**.048**	**.005**	**.005**	**.006**

MULAN generally improves the quality of base PLMs (and even structure-aware PLMs) of various sizes

Skoltech

Visualization of structural awareness



MULAN produces structure-aware protein representations

T-SNE visualization of residue embeddings of MULAN-small and ESM-2 8M on CASP12 dataset.

We use different colors for amino acid residue types (left) and for the 3 states of secondary structure (right)

Skoltech

Conclusion #2

- Proposed **MULAN** — MULtimodal PLM for both sequence and ANgle-based structure encoding.
- Evaluated the obtained structure-aware protein representations on a wide range of downstream tasks. We show that **MULAN improves over any base PLM it is applied to**.
- MULAN requires finetuning of the underlying base PLM together with the Structure Adapter → MULAN offers a **cheap increase in performance**.
- Demonstrated the **structural awareness of MULAN** embeddings.

Skoltech

https://github.com/DFrolova/MULAN

Acknowledgements

Marina Pak

Nikita Dovidchenko

Darya Frolova

Marina Pak

Ilya Sharov

Satyarth Mishra Sharma

Anna Litvin

Ivan Oseledets

thx.

Skoltech