Learning to design protein–protein interactions with enhanced generalization

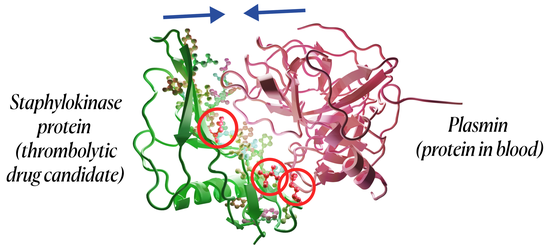
ICLR 2024

1 CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE
2 LOSCHMIDT LABORATORIES
3 FNUSA ICRC ST. ANNE'S UNIVERSITY HOSPITAL BRNO INTERNATIONAL CLINICAL RESEARCH CENTER
4 ÚOCHB AV ČR IOCB PRAGUE

Anton Bushuiev[*1], **Roman Bushuiev**[*1,4], Petr Kouba[1,2], Anatolii Filkin[1], Marketa Gabrielova[1], Michal Gabriel[1], Jiri Sedlar[1], Tomas Pluskal[4], Jiri Damborsky[2,3], Stanislav Mazurenko[2,3], Josef Sivic[1]

Paper | PPIRef package | PPIformer package | Hugging Face Space

Overview



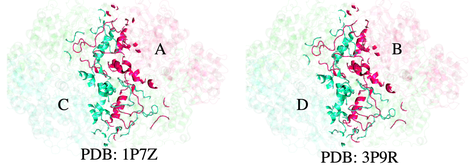
- **Problem:** Increasing binding affinity of a protein–protein interaction (PPI) by changing amino acids
- **Challenge:** Computational methods are often slow while machine learning methods struggle to generalize due to the scarcity of annotated mutations (ΔΔG data)
- **Our approach:** Large-scale self-supervised pre-training from coarse-grained PPI structures to enhance the generalization of PPI design

New dataset: PPIRef

- **PPIRef:** New largest and non-redundant dataset of PPI structures mined from the PDB database

Dataset	PPI structures	Unique interfaces
MaSIF-search	6K	5K
DIPS / DIPS-Plus	40K	9K
PPIRef (ours)	**322K**	**46K**

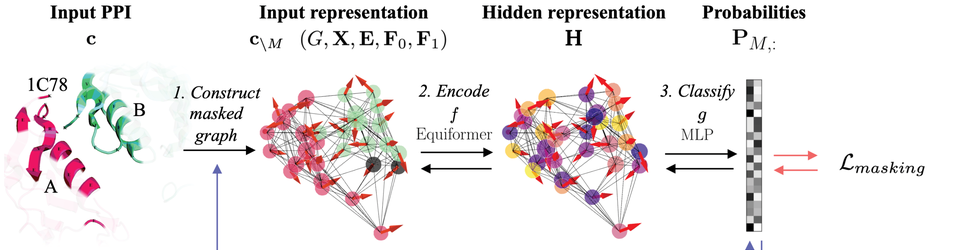
- **iDist:** New efficient approximation of 3D PPI alignment (iAlign, USalign), used to construct PPIRef
- **iDist** is **>100x faster** than iAlign and USalign and finds same near duplicates with **99% precision** and **97% recall**



PDB: 1P7Z PDB: 3P9R

New model: PPIformer

Architecture: Coarse-grained PPI structures modelled with the Equiformer graph neural network



$$\widehat{\Delta\Delta G} = \sum_{i \in M} \log p(\hat{c}_i = c_i | \mathbf{c}_{\setminus M}) - \sum_{i \in M} \log p(\hat{c}_i = m_i | \mathbf{c}_{\setminus M})$$

$\mathcal{L}_{mutation}$ (MSE)

Input mutant
m

$$M = \{i | c_i \neq m_i\}$$

Masking of mutated amino acids

Fine-tuning via log odds ratio

Pre-training: Regularized structural masked modeling

$$\mathcal{L}_{masking} = -\sum_{i \in M} w_{c_i} \left[(1-\epsilon) \cdot \log p(\hat{c}_i = c_i | \mathbf{c}_{\setminus M}) + \epsilon \sum_{j \in \mathcal{A} \setminus c_i} \log p(\hat{c}_i = c_j | \mathbf{c}_{\setminus M}) \cdot \frac{1}{|\mathcal{A}|} \right]$$

Sum over all masked residues

Weight of native amino acid type c_i

Probability of native amino acid type c_i

Smoothing hyper-parameter

Label smoothing regularization (non-native amino acid types)

Fine-tuning: Thermodynamically motivated log odds ratio

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt} = RT(\log(K_{wt}) - \log(K_{mut}))$$

Difference in binding energy of mutated and wild-type (native amino acid) complexes

Environment constants

Ratios of the concentraion of interacting and non-interacting complexes

Results: Standard data

- **State of the art** machine learning method on the non-leaking **SKEMPI v2.0** test set

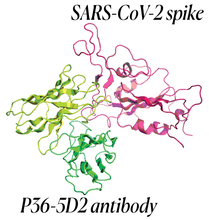
		ΔΔG correlation		*Negative ΔΔG detection*			*Absolute ΔΔG value*	
Category	Method	Spearman ↑	Pearson ↑	Precision ↑	Recall ↑	ROC AUC ↑	MAE ↓	RMSE ↓
Force field simulations	FLEX ddG *	0.55	0.57	0.63	0.62	0.84	1.59	2.00
Machine learning	GEMME	0.38	0.41	**0.60**	0.49	0.74	2.16	2.81
	MSA TRANSFORMER	0.31	0.36	0.51	0.38	0.70	6.13	6.93
	ESM-IF	0.18	0.18	0.33	0.41	0.68	1.87	2.15
	RDE-NET.	0.24	0.30	0.54	**0.65**	0.67	1.70	2.02
	PPIFORMER (OURS) **	**0.44** ± 0.03	**0.46** ± 0.03	**0.60** ± 0.014	0.61 ± 0.012	**0.78** ± 0.019	**1.64** ± 0.011	**1.94** ± 0.006

** A single prediction requires ~1 CPU hour (5 orders of magnitude slower than the other methods)*

*** Average over three fine-tuning seeds*

Results: Independent drug design case studies

- **Effective generalization** to a **coronavirus antibody design** case study



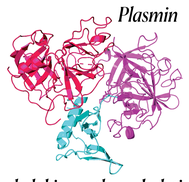
Ranks of known favorable mutations (5 out of 494)

Method	TH31W↓	AH53F↓	NH57L↓	RH103M↓	LH104F↓
MSA TRANSFORMER	56.88	42.11	63.56	49.19	18.83
ESM-IF	49.39	17.61	17.00	51.42	48.58
ROSETTA	10.73	76.72	93.93	11.34	27.94
FOLDX	13.56	**6.88** ***	**5.67**	16.60	66.19
DDGPRED	68.22	**2.63**	12.35	**8.30**	**8.50**
END-TO-END	29.96	**2.02**	14.17	52.43	17.21
MIF-NET.	24.49	**4.05**	**6.48**	80.36	36.23
RDE-NET.	**1.62**	**2.02**	20.65	61.54	**5.47**
PPIFORMER (OURS)	18.02	**0.20**	**7.69**	21.46	10.9

Lowest overall average rank for favorable mutations among all the methods

Accurate top-1 hit

- **Effective generalization** to a **thrombolytic design** case study



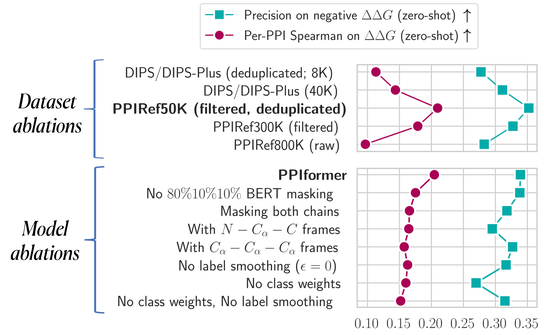
Ranks of known favorable mutations (6 out of 80)

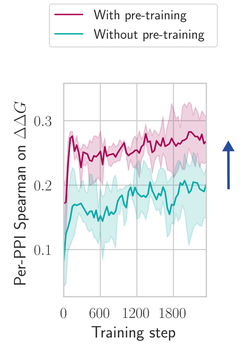
Method	Mutations with ≥ 2× activity enhancement					
	KC130A↓	KC130T↓	KC135R KC130T↓	KC135A↓	KC74Q KC130E KC135R↓	KC74R KC130T KC135R↓
MSA TRANSFORMER	52.50	32.50	55.00	40.00	70.00	78.75
ESM-IF	45.00	33.75	46.25	25.00	42.50	58.75
RDE-NET.	51.25	33.75	22.50	15.00	27.50	**5.00** ***
PPIFORMER (OURS)	66.25	15.00	**2.50**	52.50	33.75	**1.25**

Accurate top-2 hits

**** Mutations that are in top 10% of predictions are in bold*

Ablations





- Non-redundant **PPIRef** dataset is crucial for pre-training
- As well as our virtual C_β-based **coarse-grained representation** and **regularizations**
- **Pre-training** is crucial for ΔΔG fine-tuning