Designing protein-protein interactions with self-supervised geometric deep learning

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

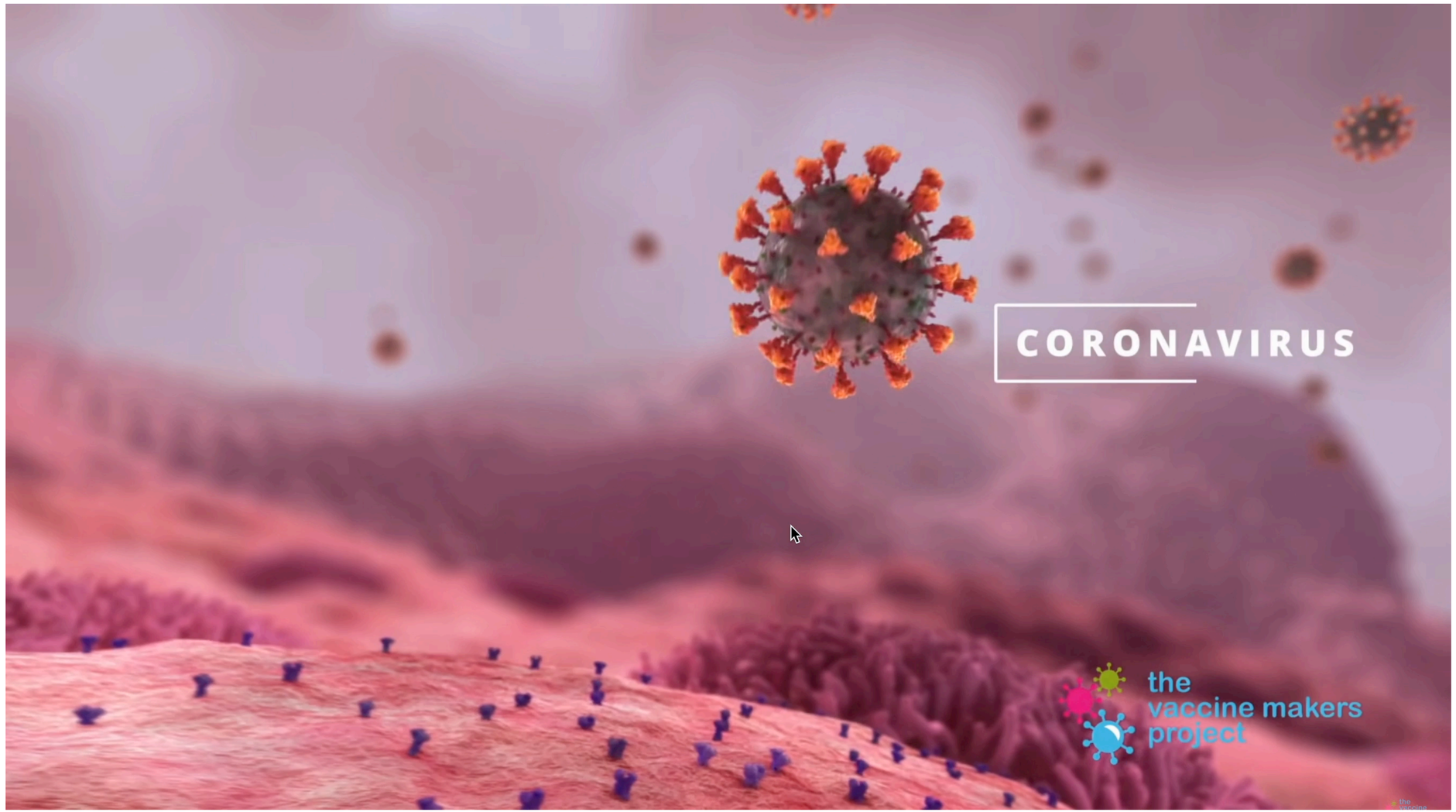
1 CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE

2 FNUSA ICRC ST. ANNE'S UNIVERSITY HOSPITAL BRNO INTERNATIONAL CLINICAL RESEARCH CENTER

3 LOSCHMIDT LABORATORIES

4 ÚOCHB AV ČR IOCB PRAGUE

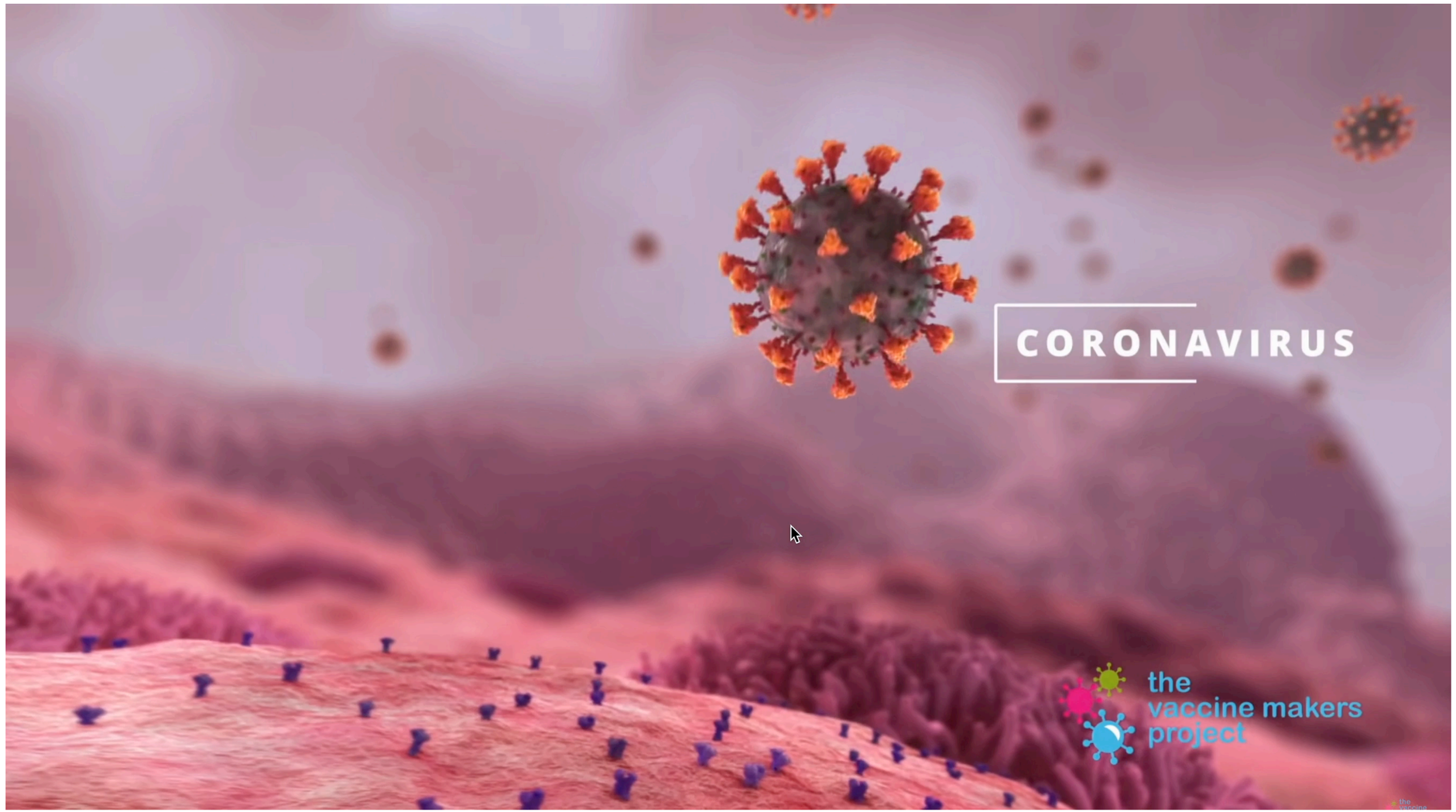
Protein—protein interactions



Directly linked to the development and treatment of **viruses**, **stroke**, **cancer**, **Alzheimer**, ...

Video source: YouTube channel Vaccine Makers Project

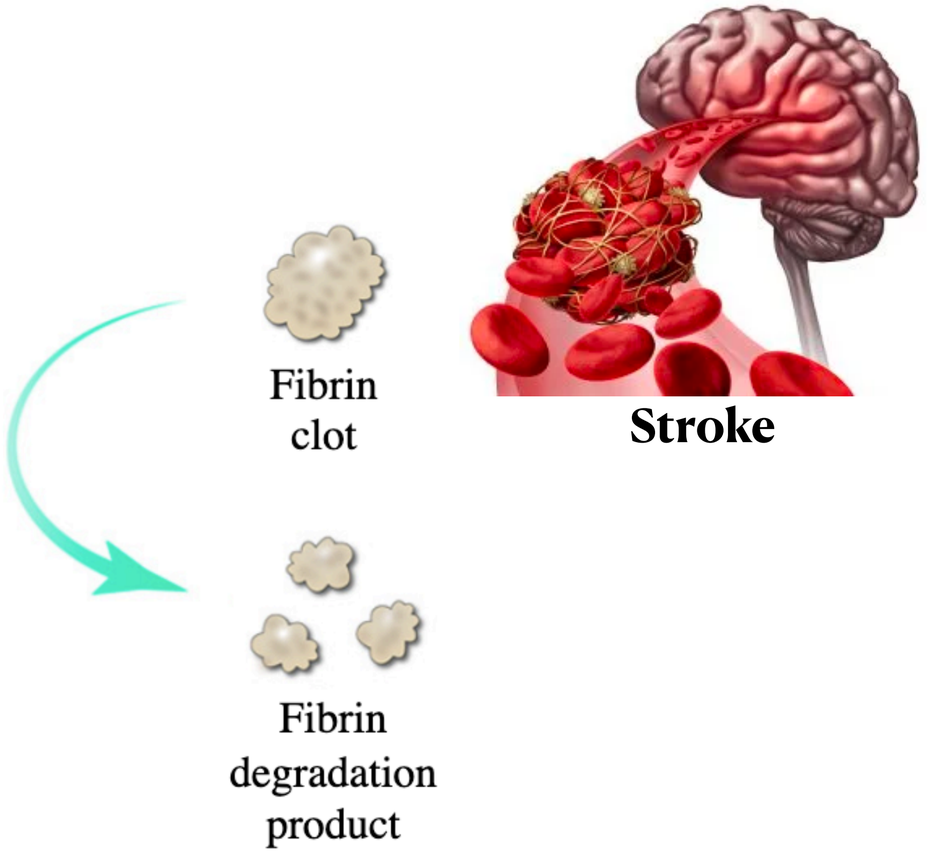
Protein—protein interactions



Directly linked to the development and treatment of **viruses**, **stroke**, **cancer**, **Alzheimer**, ...

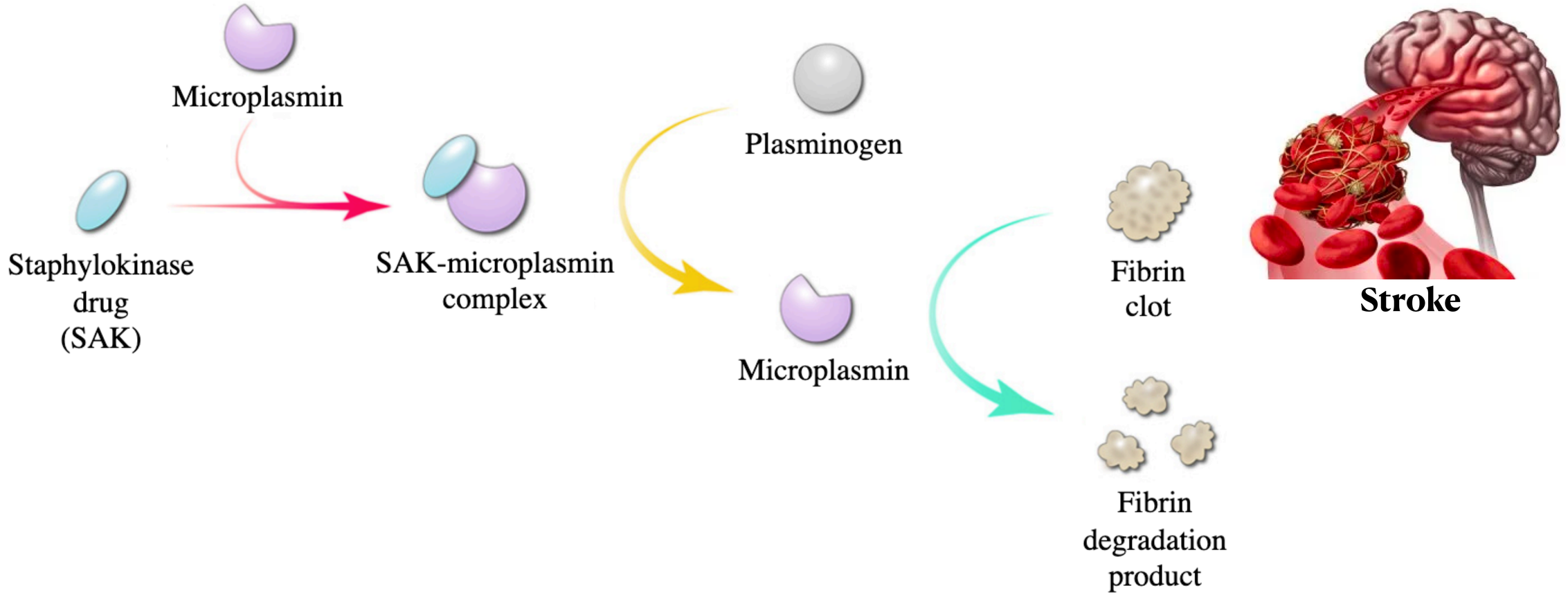
Video source: YouTube channel Vaccine Makers Project

Staphylokinase: thrombolytic drug candidate



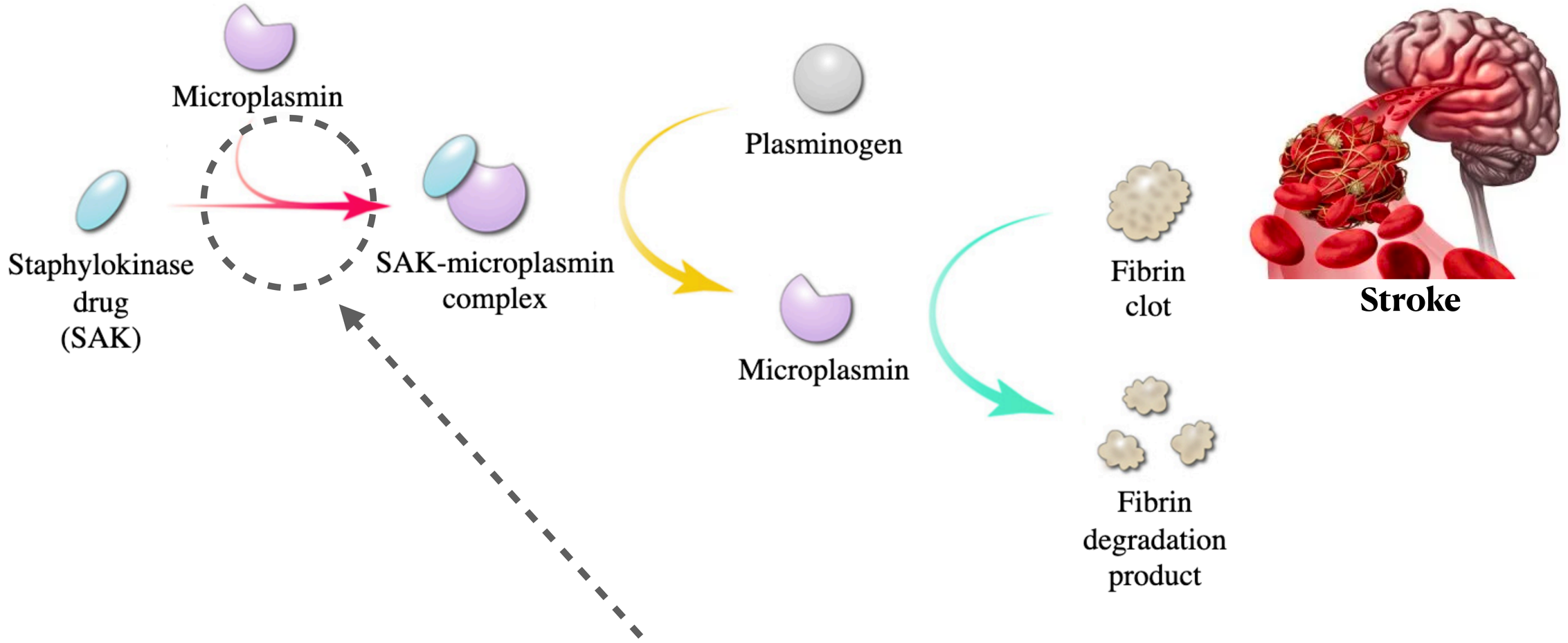
Brain image source: Bel Marra Health website

Staphylokinase: thrombolytic drug candidate



Brain image source: Bel Marra Health website

Staphylokinase: thrombolytic drug candidate

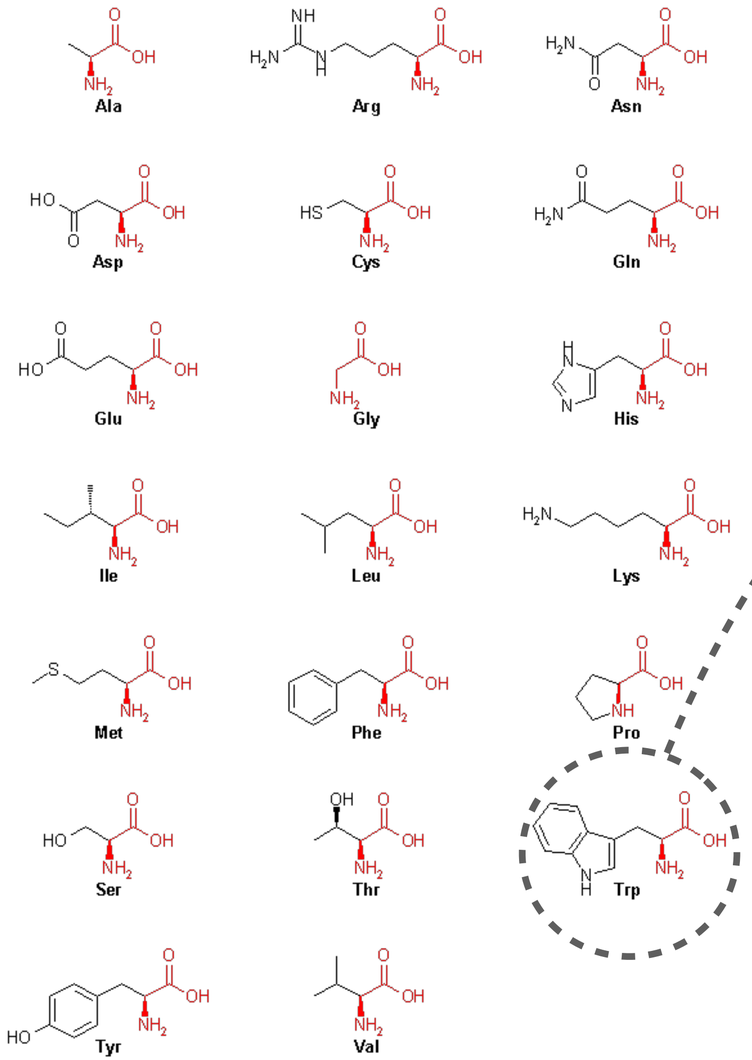


How to enhance the *binding affinity* of the interaction for effective thrombolysis?

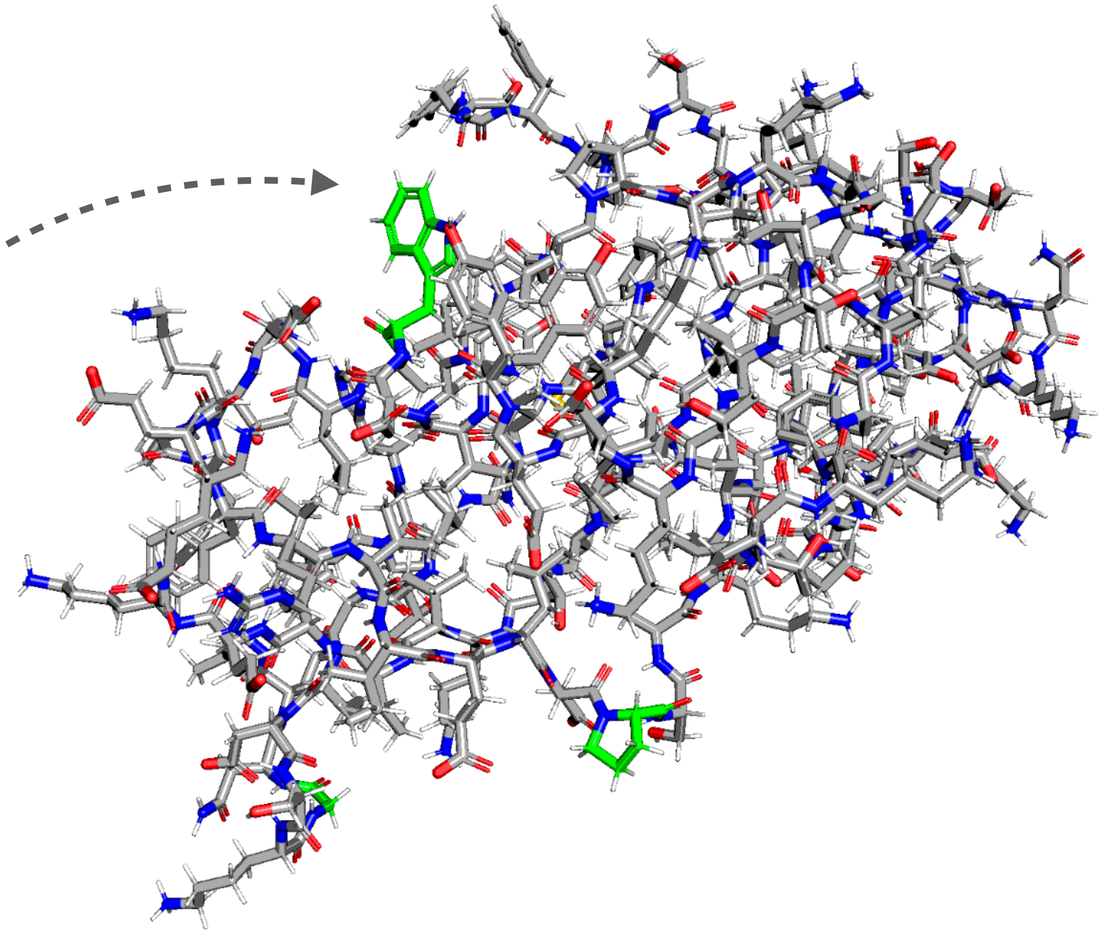
Brain image source: Bel Marra Health website

What is a protein?

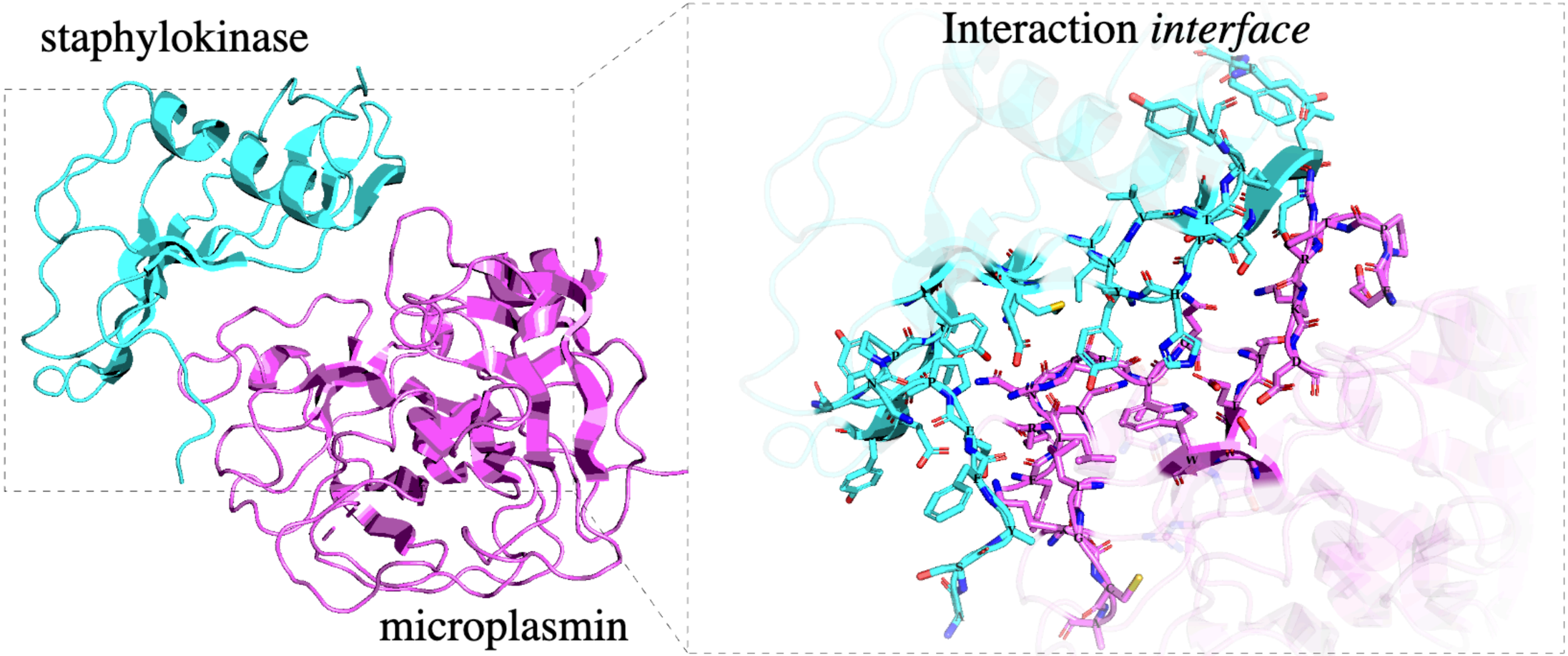
20 amino acids (building block types)



Protein: a folded chain of amino acids

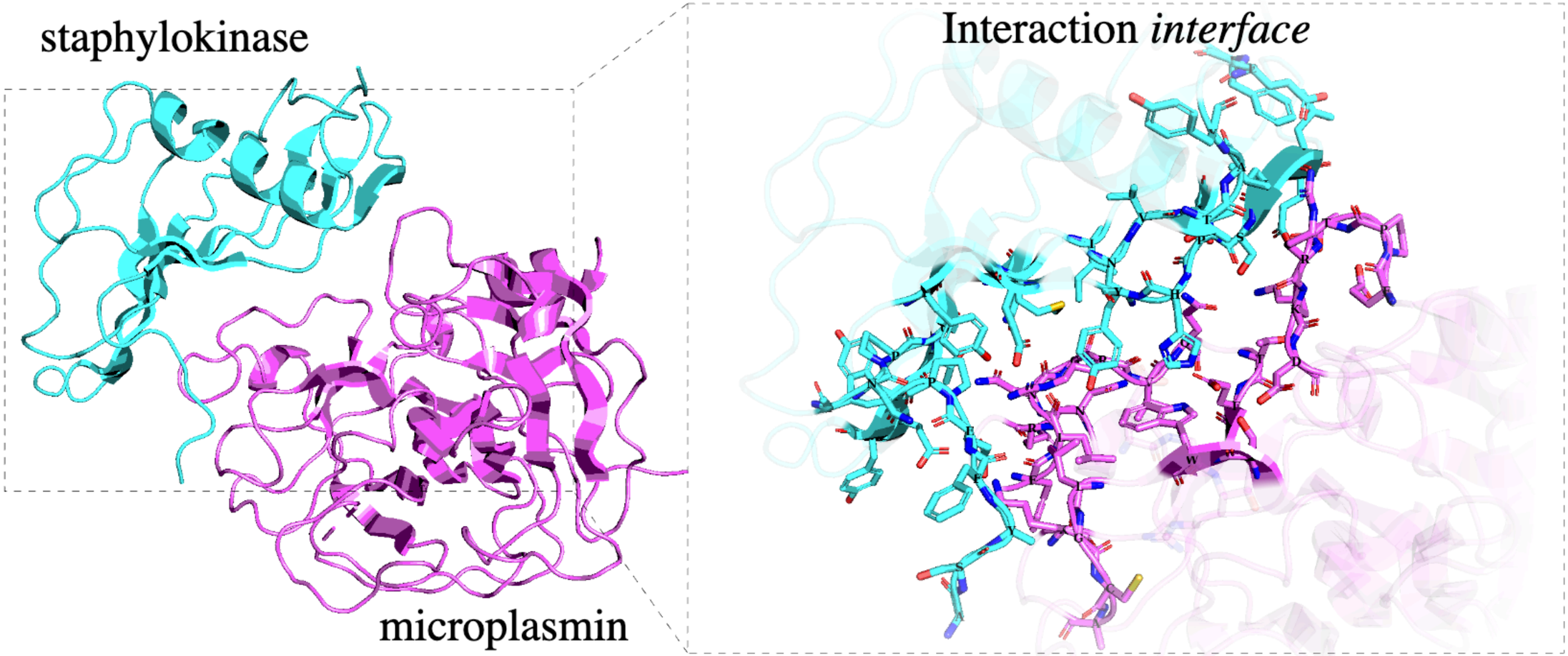


Staphylokinase—microplasmin interaction



How to enhance the *binding affinity* of the interaction?

Staphylokinase—microplasmin interaction



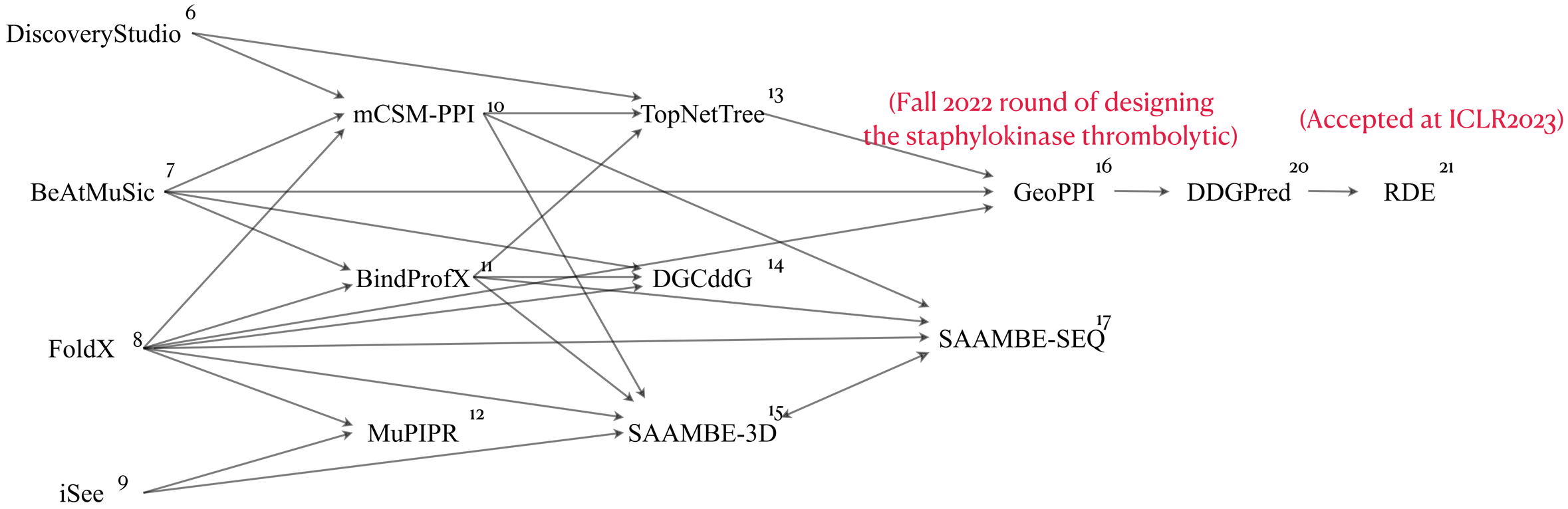
What amino acids of staphylokinase to mutate and how? 20^n combinations

Standard approach: ΔΔG screening

1. Screen thousands or millions of mutations according to ΔΔG — binding energy change upon mutation ranging roughly in [-12, 12]
2. Select several best candidates (with lowest ΔΔG) and test in a lab

Staphylokinase mutants

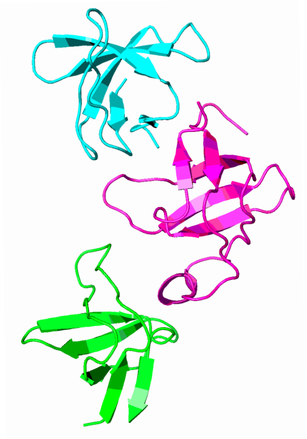
State of the art for predicting ΔΔG



- Rely on **small data** (7K annotated mutations from SKEMPI2) → unstable, weak generalization
- Often require **mutant 3D structure** → slow
- **Weak evaluation protocol** → poor generalization

Labeled data (SKEMPI2)

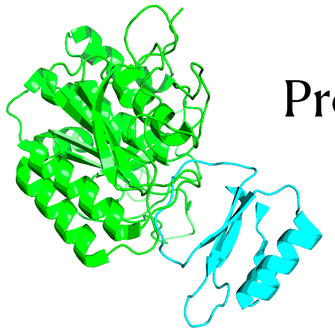
300 interactions, 7K mutations



Protein C, amino acid 21

Cys → Val

ΔΔG = -0.025



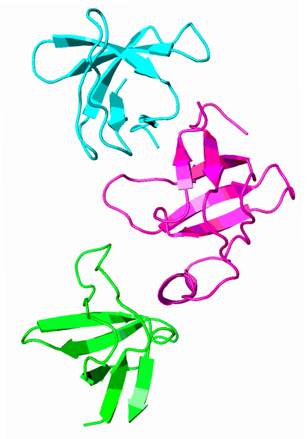
Protein I, amino acid 45

Leu → Ser

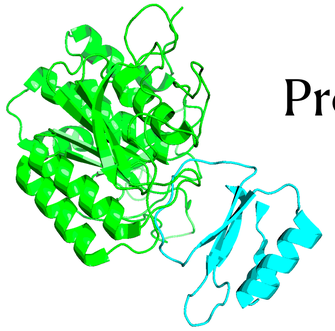
ΔΔG = 1.17

PPIRef: New large dataset of PPIs

Labeled data (SKEMPI2)
300 interactions, 7K mutations

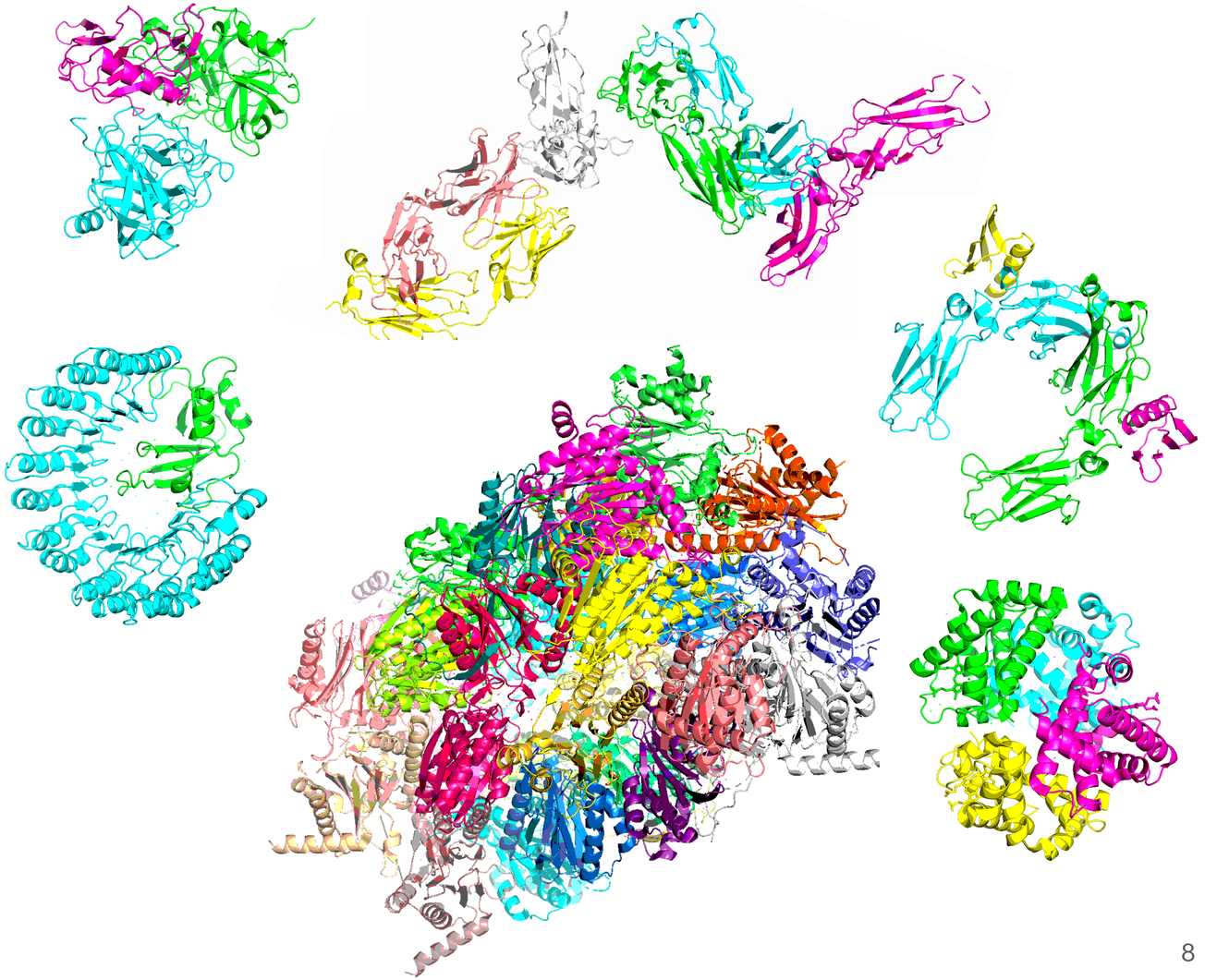


Protein C, amino acid 21
Cys → Val
ΔΔG = -0.025



Protein I, amino acid 45
Leu → Ser
ΔΔG = 1.17

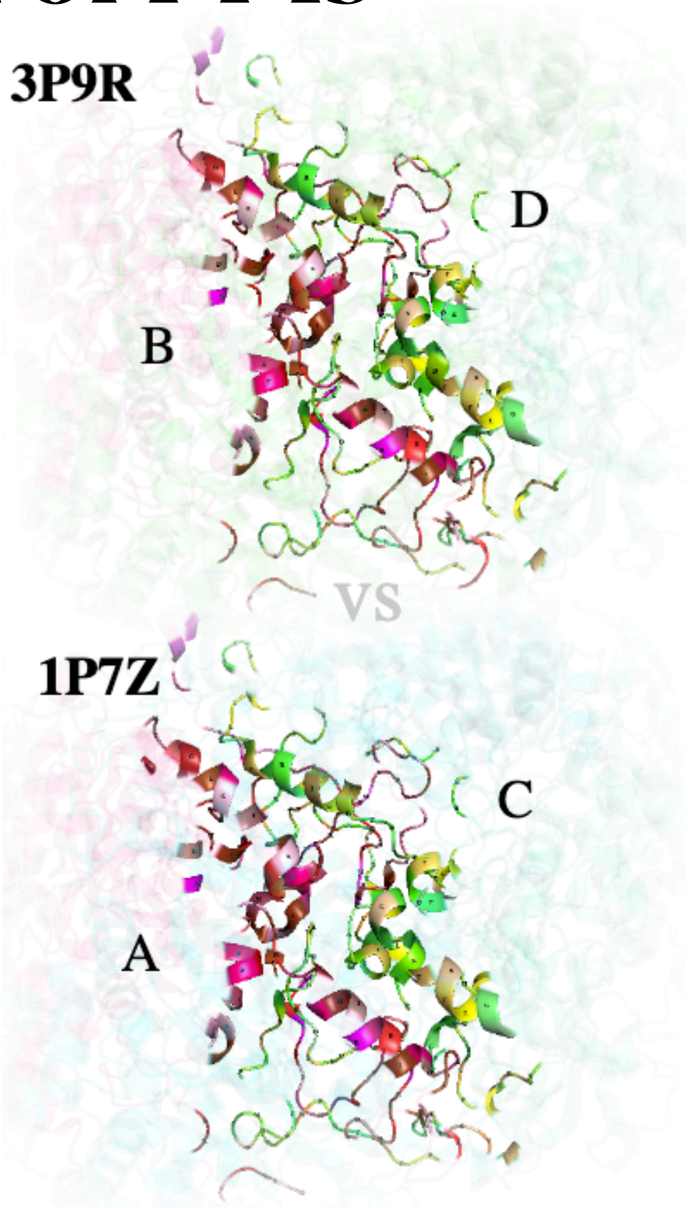
Unlabeled data (Protein Data Bank)
322K interactions in our PPIRef, **41K** in DIPS



iDist: Scalable comparison of PPIs

- **iDist** accurately approximates iAlign[18] (TM-score for PPIs)

  (near-duplicate detection with 99% precision and 97% recall)
- **iDist** is ~500 times faster than iAlign



iDist = 0.0035

iDist: Scalable comparison of PPIs

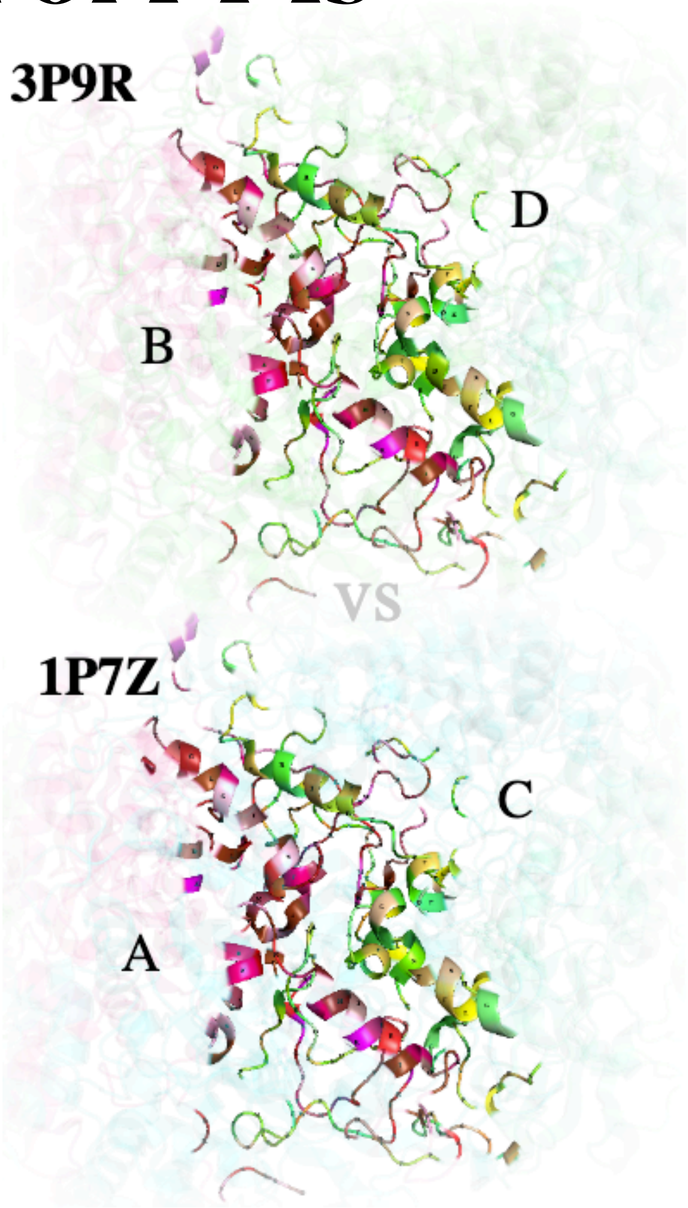
- **iDist** accurately approximates iAlign[18] (TM-score for PPIs)

  (near-duplicate detection with 99% precision and 97% recall)
- **iDist** is ~500 times faster than iAlign

- Available datasets are redundant and incomplete

  (Some PPIs represented > 500 times, many missing)
- Existing train-test splits suffer from data leakage

  (>53% of test sets have near-duplicates in train sets)



iDist = 0.0035

iDist: Scalable comparison of PPIs

- **iDist** accurately approximates iAlign[18] (TM-score for PPIs)

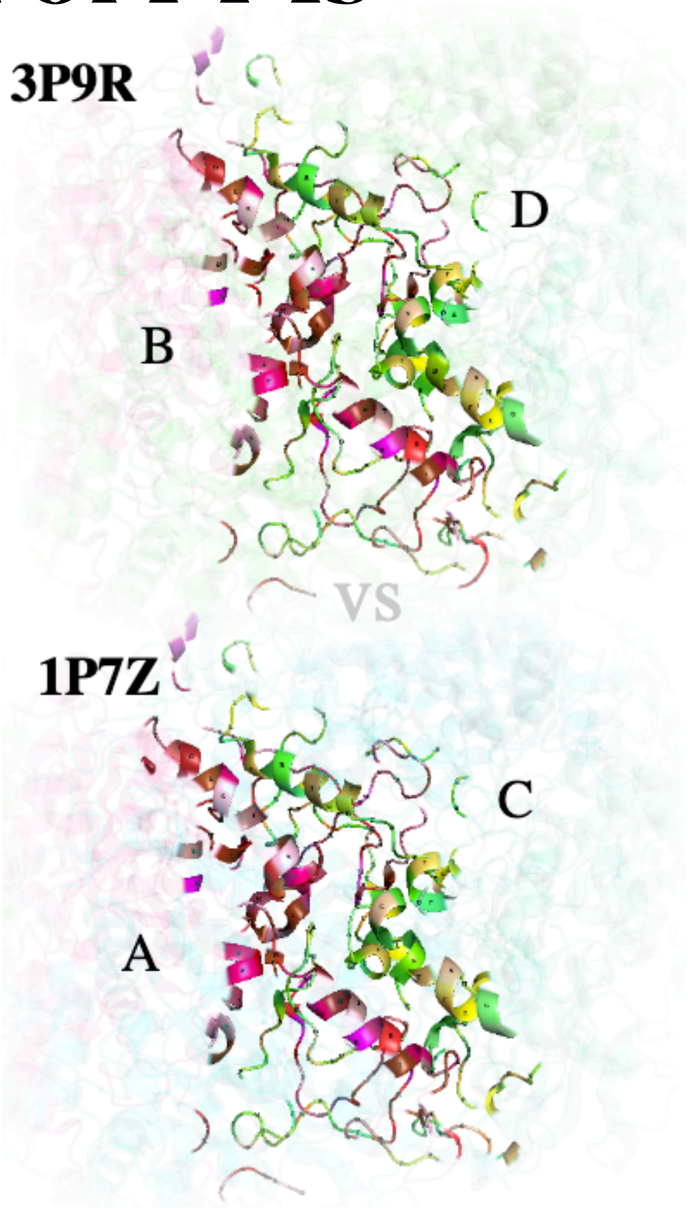
  (near-duplicate detection with 99% precision and 97% recall)
- **iDist** is ~500 times faster than iAlign

- Available datasets are redundant and incomplete

  (Some PPIs represented > 500 times, many missing)
- Existing train-test splits suffer from data leakage

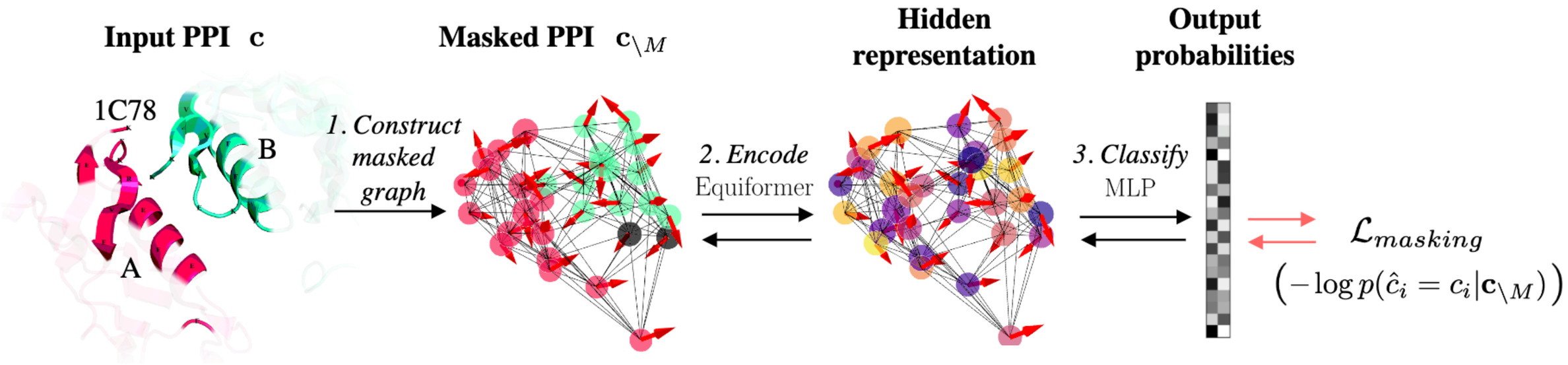
  (>53% of test sets have near-duplicates in train sets)

- **PPIRef**: non-redundant complete set of PPIs from PDB

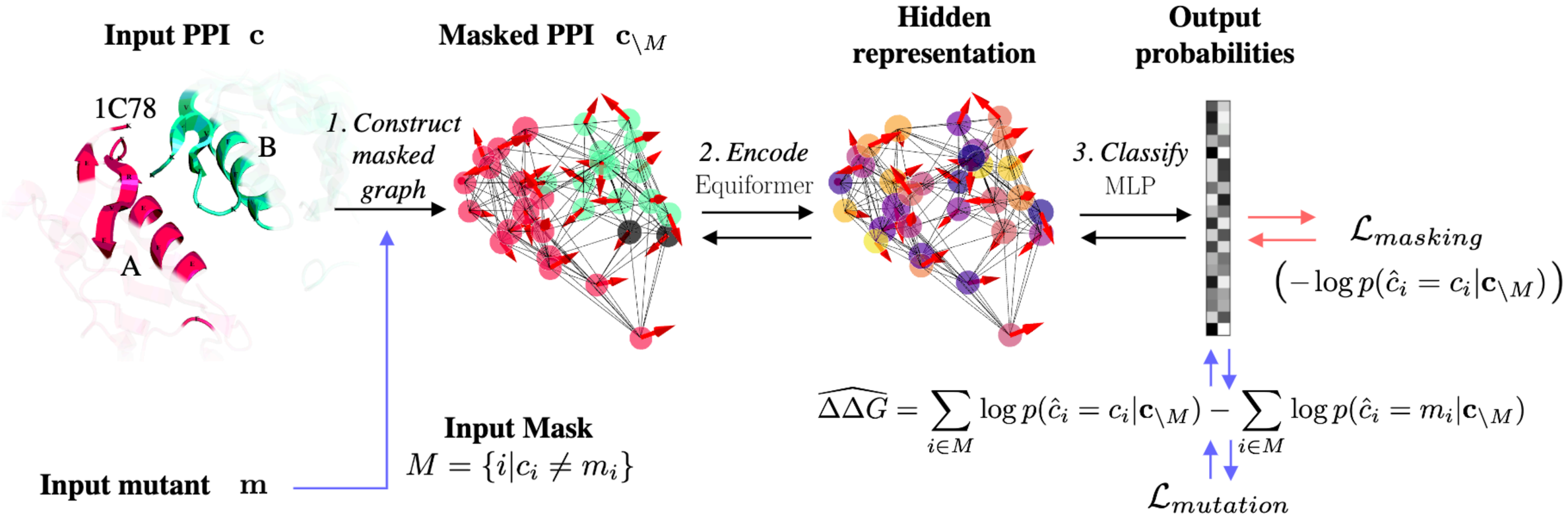


iDist = 0.0035

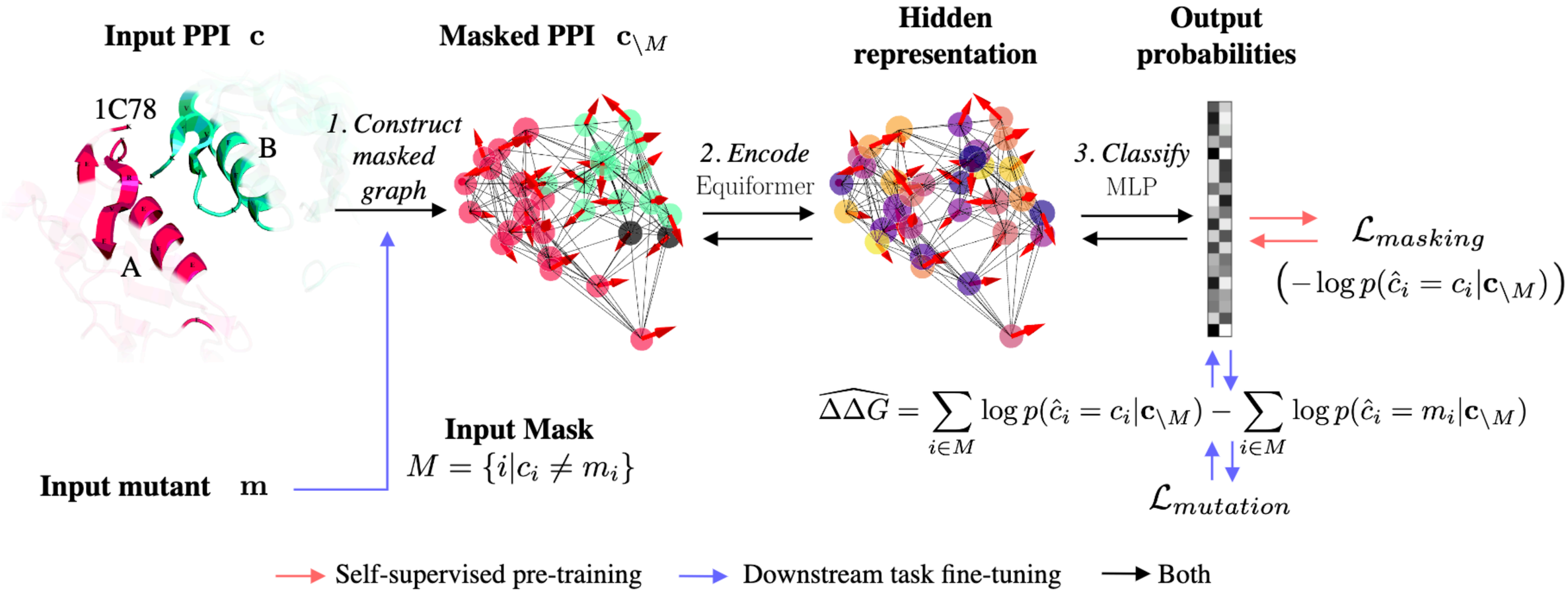
PPIformer


Input PPI c
Masked PPI $\mathbf{c}_{\setminus M}$
Hidden representation
Output probabilities
1C78
B
A
1. Construct masked graph
2. Encode
Equiformer
3. Classify
MLP
$\mathcal{L}_{masking}$
$\left(-\log p(\hat{c}_i = c_i | \mathbf{c}_{\setminus M})\right)$

PPIformer

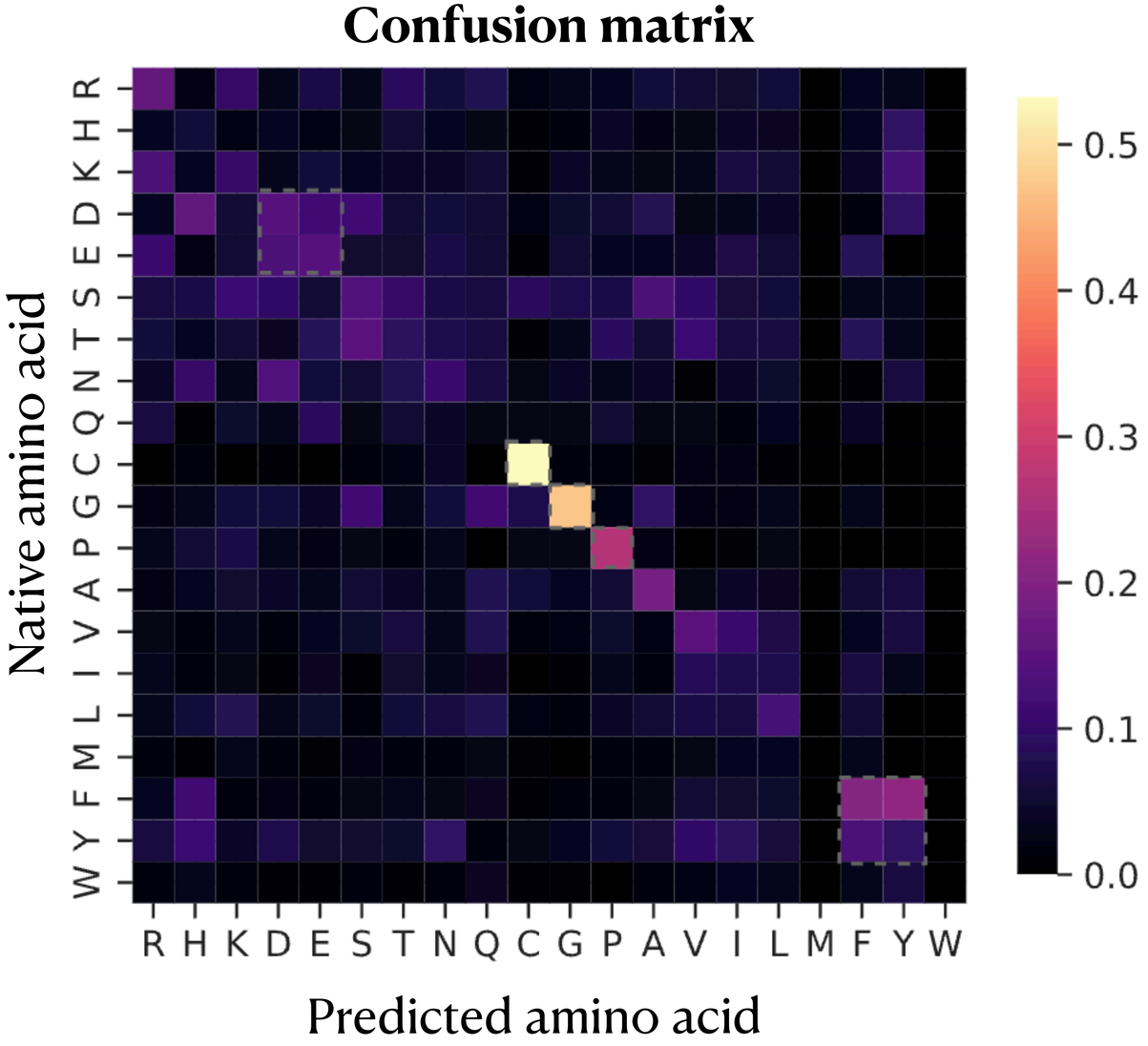

Input PPI c
Masked PPI $\mathbf{c}_{\setminus M}$
Hidden representation
Output probabilities
1C78
B
A
1. Construct masked graph
2. Encode
Equiformer
3. Classify
MLP
$\mathcal{L}_{masking}$
$\left(-\log p(\hat{c}_i = c_i|\mathbf{c}_{\setminus M})\right)$
$\widehat{\Delta\Delta G} = \sum_{i\in M} \log p(\hat{c}_i = c_i|\mathbf{c}_{\setminus M}) - \sum_{i\in M} \log p(\hat{c}_i = m_i|\mathbf{c}_{\setminus M})$
$\mathcal{L}_{mutation}$
Input Mask
$M = \{i|c_i \neq m_i\}$
Input mutant m
Self-supervised pre-training
Downstream task fine-tuning
Both

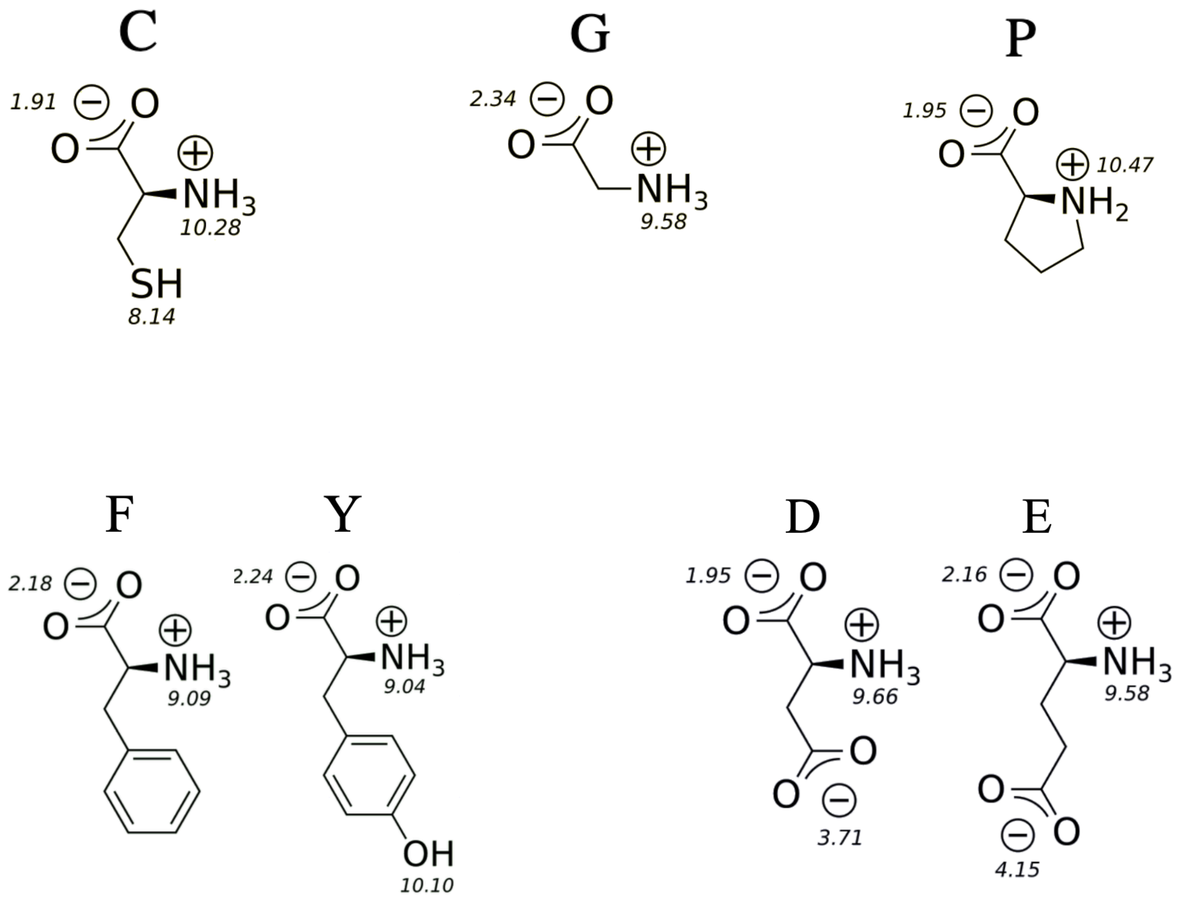
PPIformer



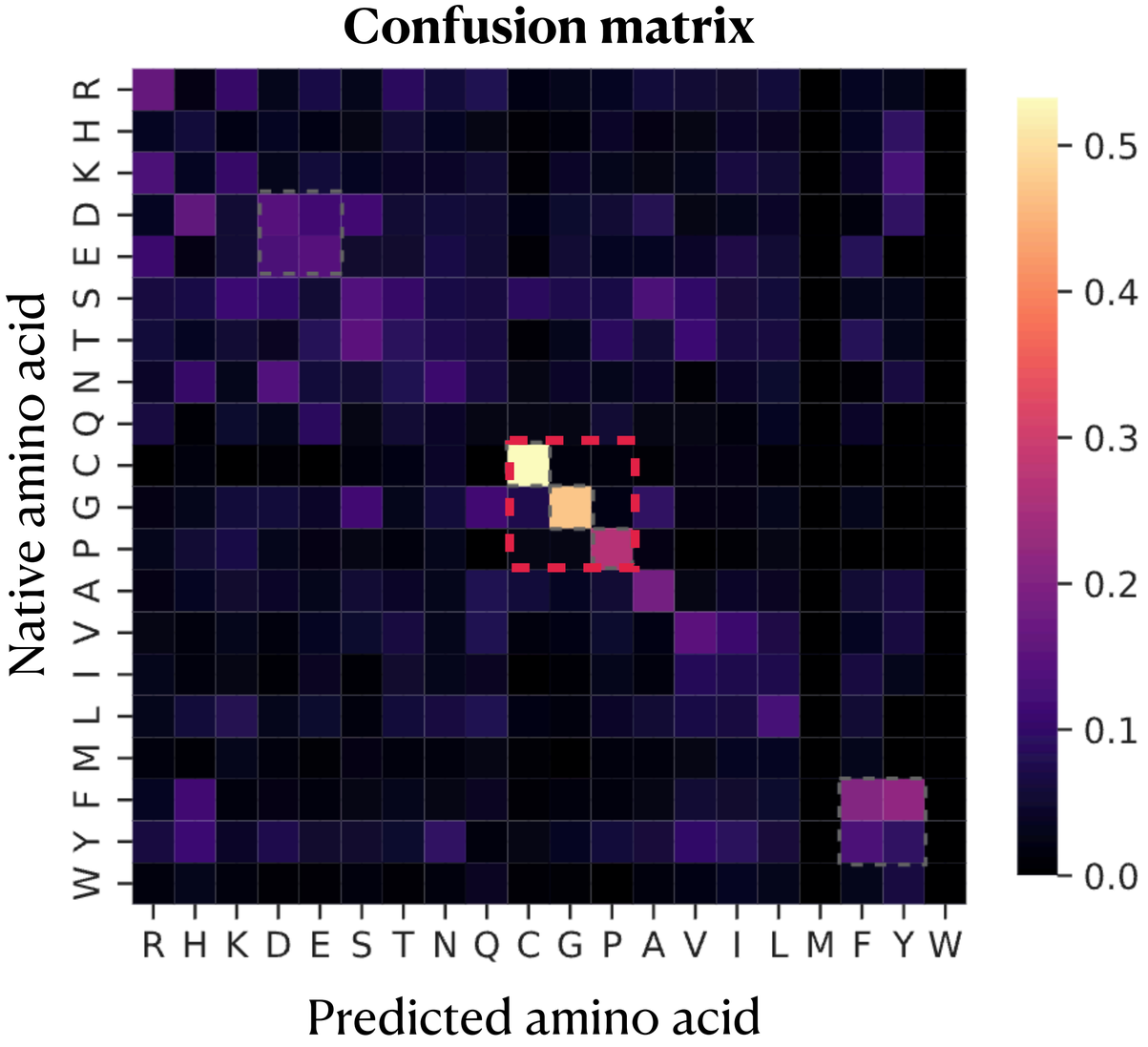
- Leverages **big data** (millions of masked examples from PPIRef during pre-training)
- Very fast, requires **a single forward pass on the native 3D structure**
- Fine-tuned and evaluated on **non-leaking ΔΔG data** using **practically-important metrics**

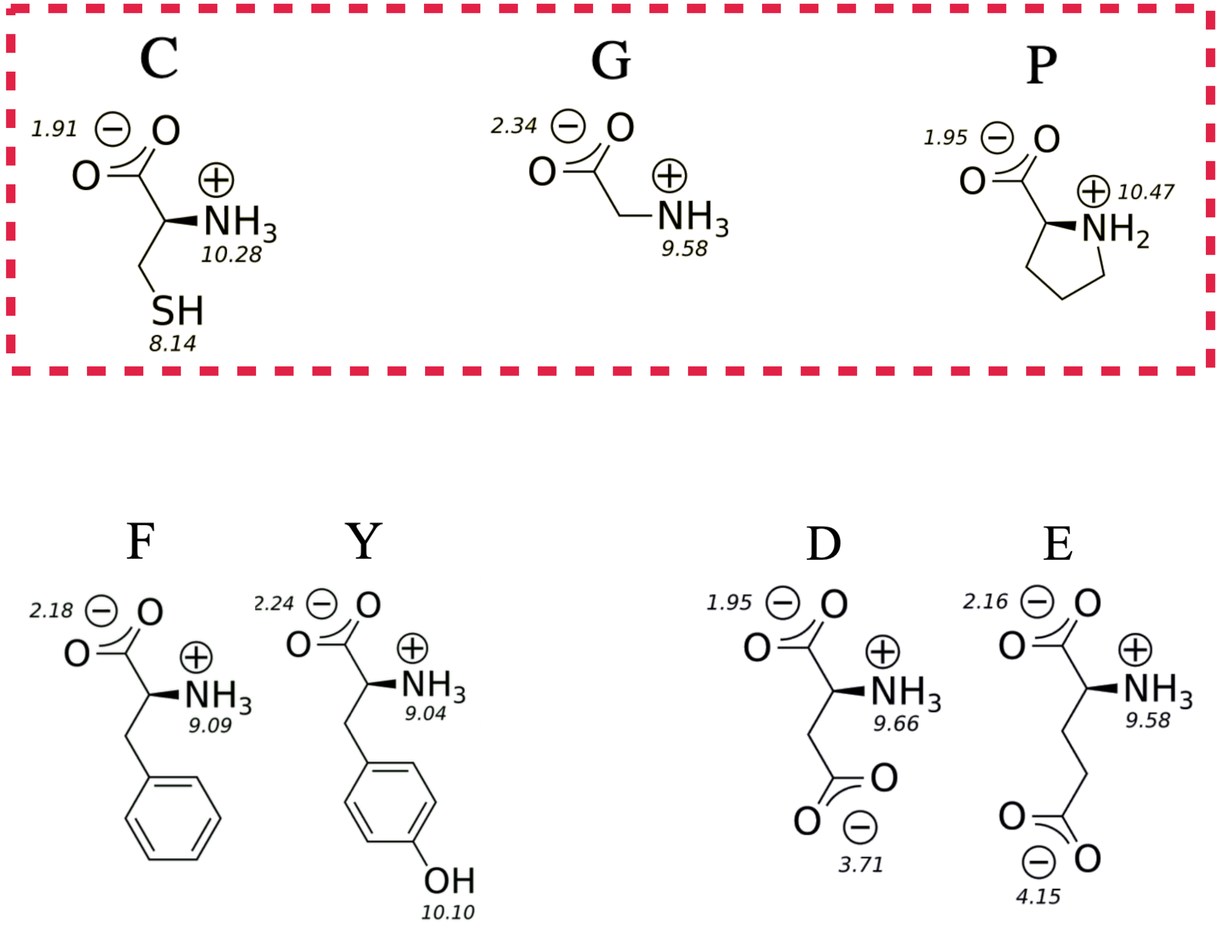
PPIformer captures biochemical principles



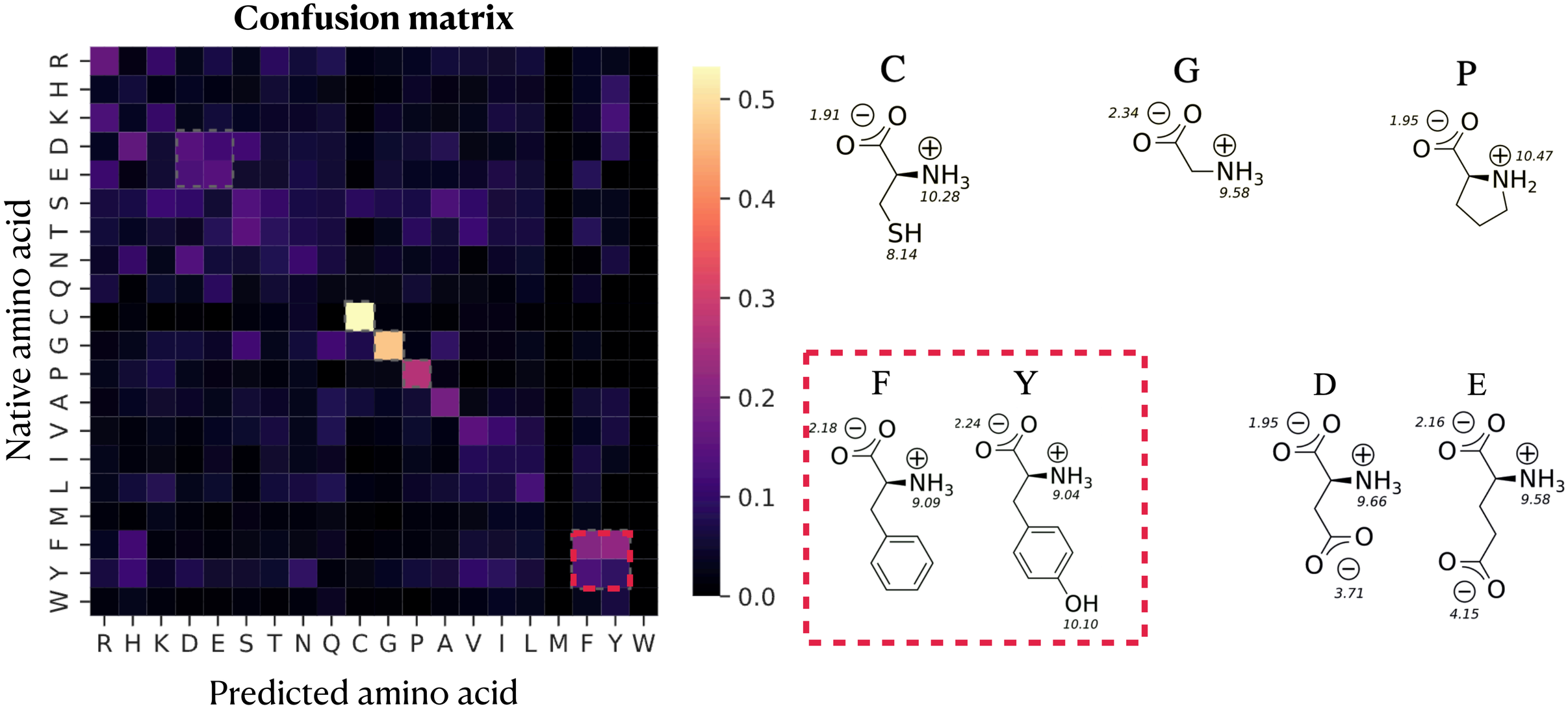


PPIformer captures biochemical principles

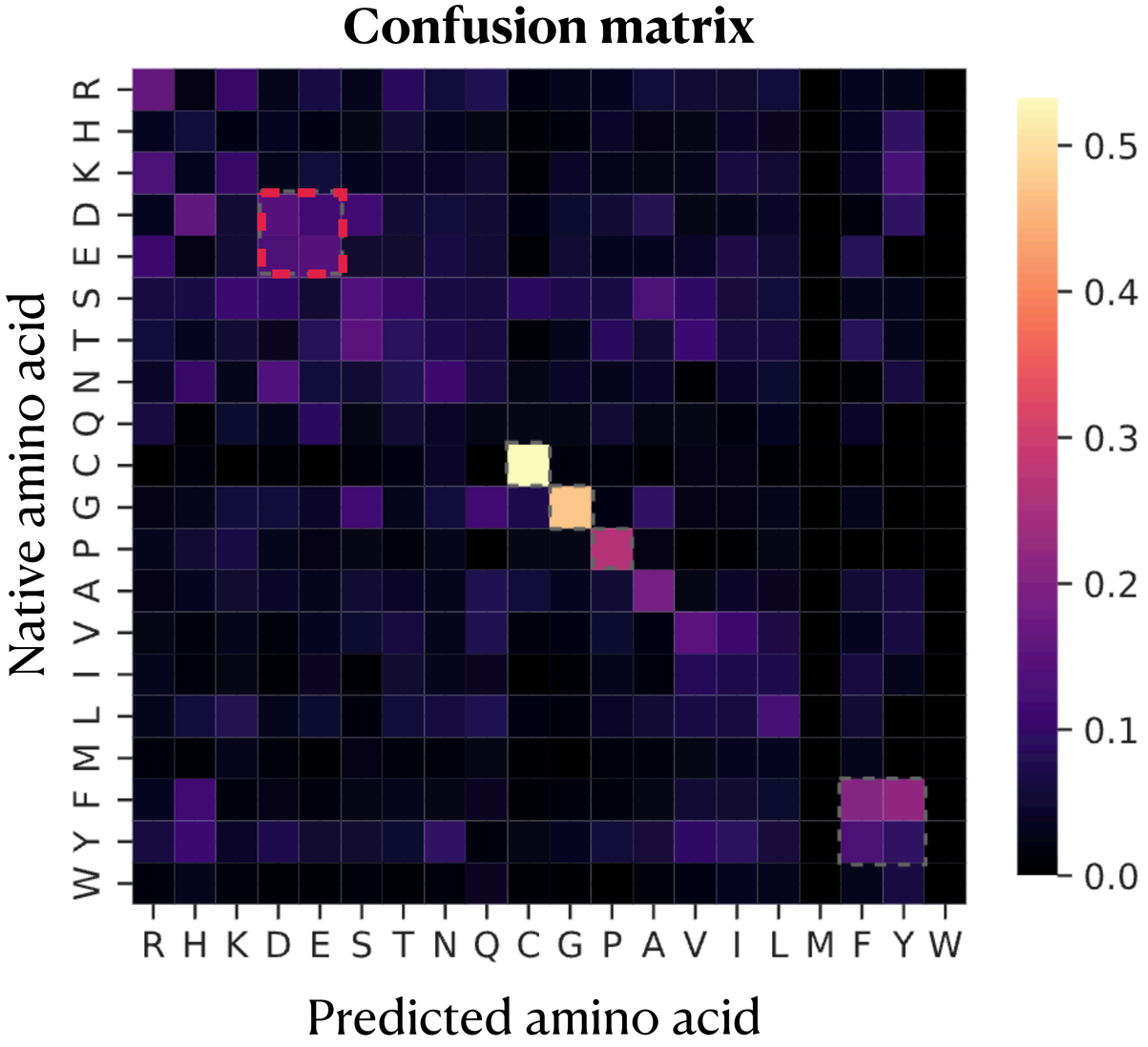


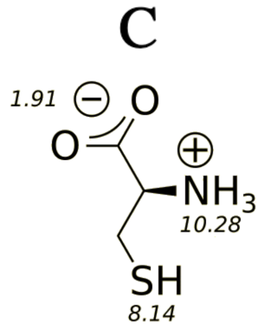


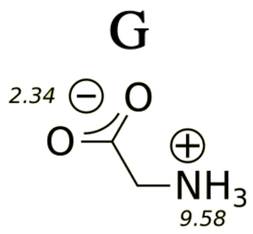
PPIformer captures biochemical principles

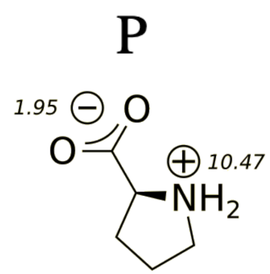


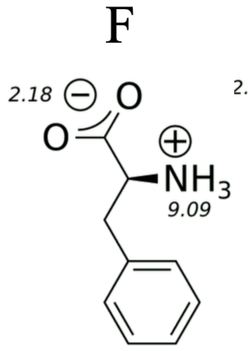
PPIformer captures biochemical principles

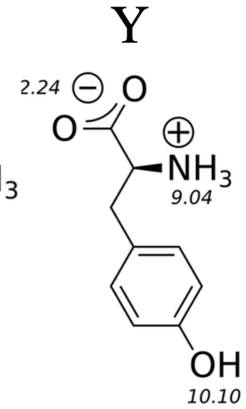


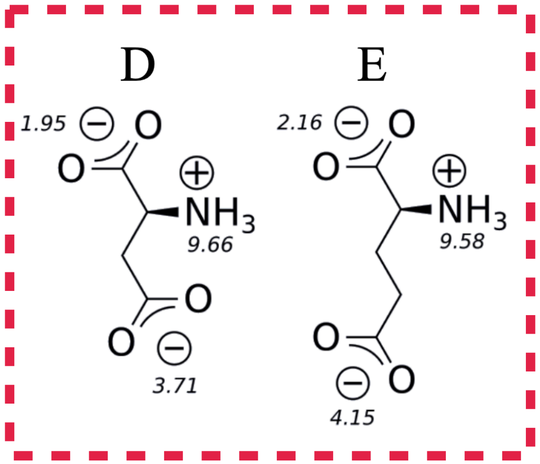




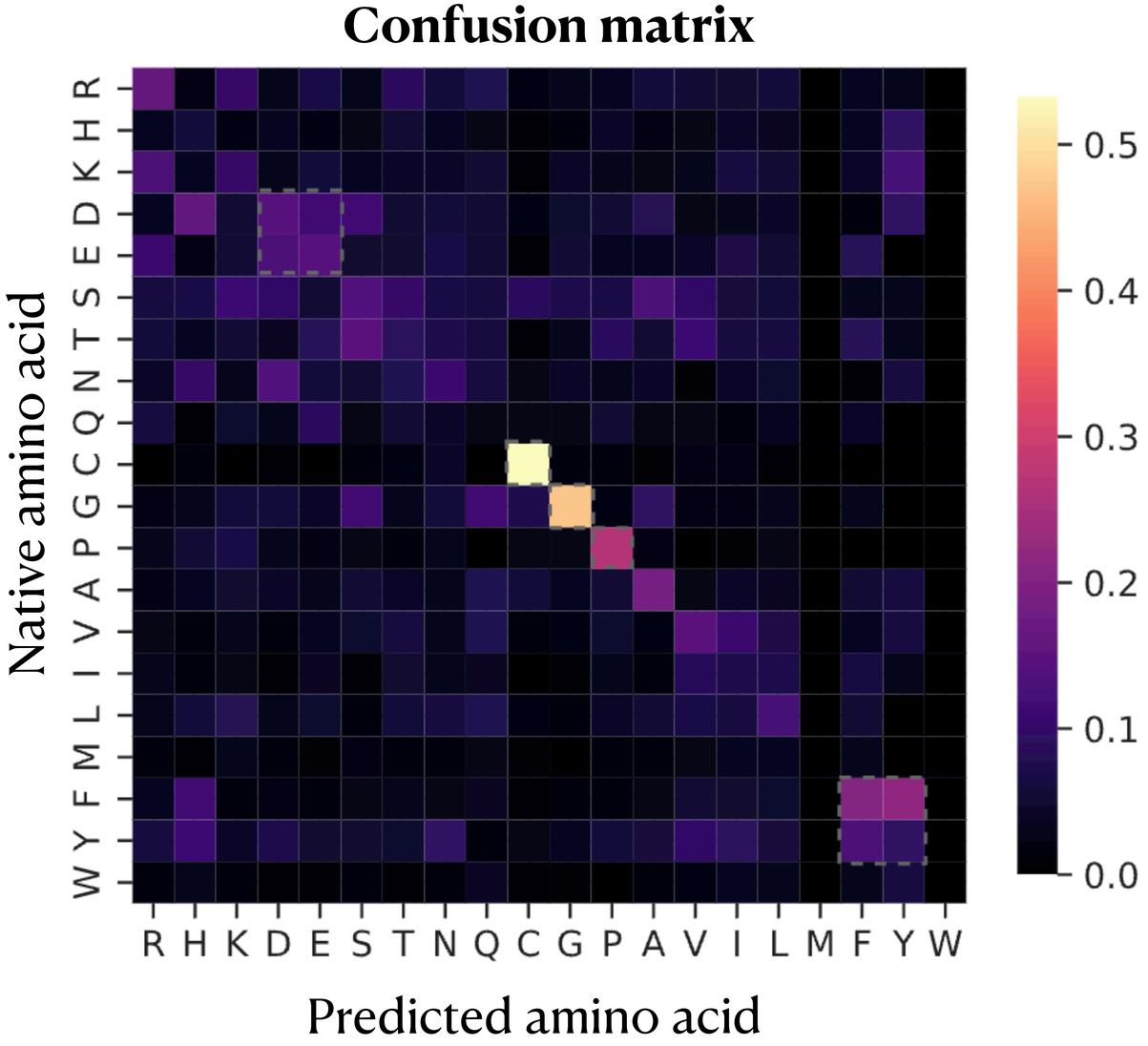


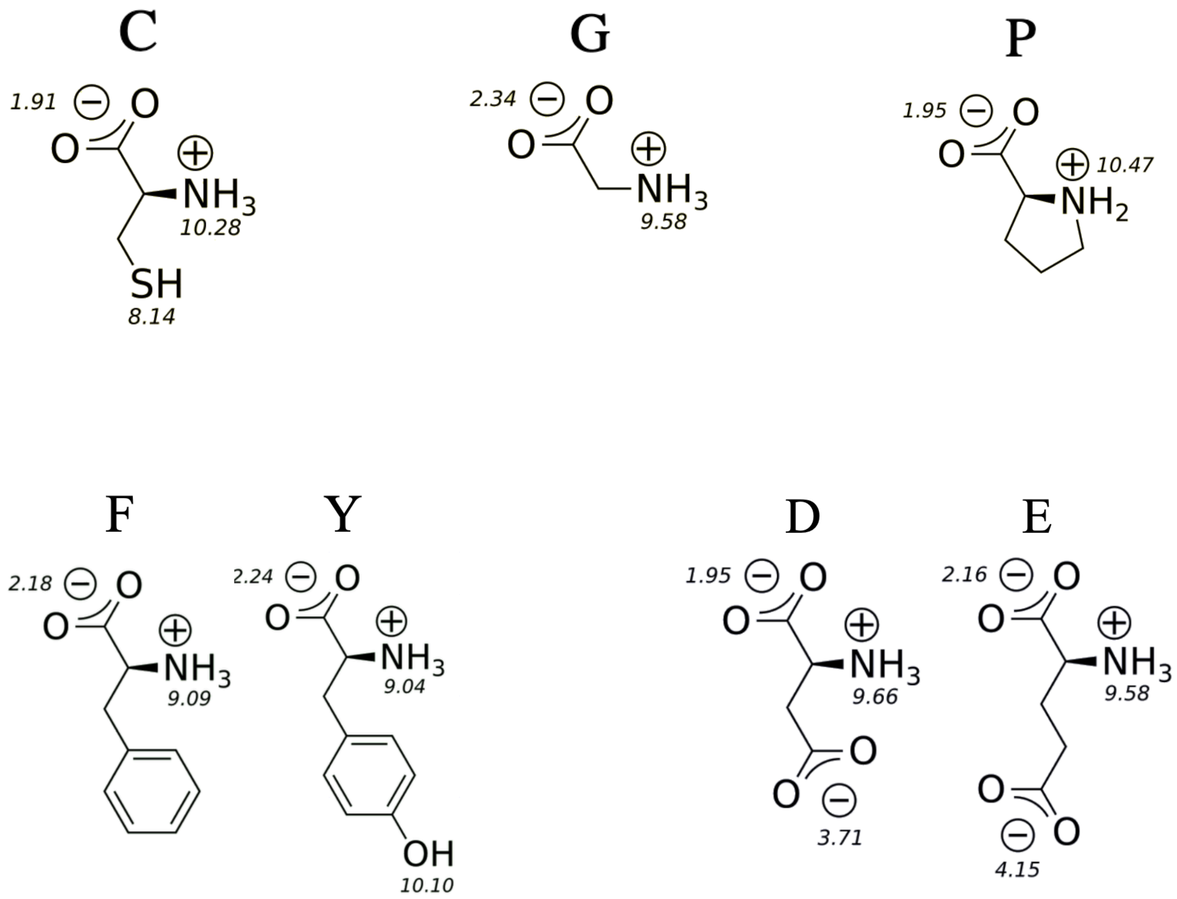




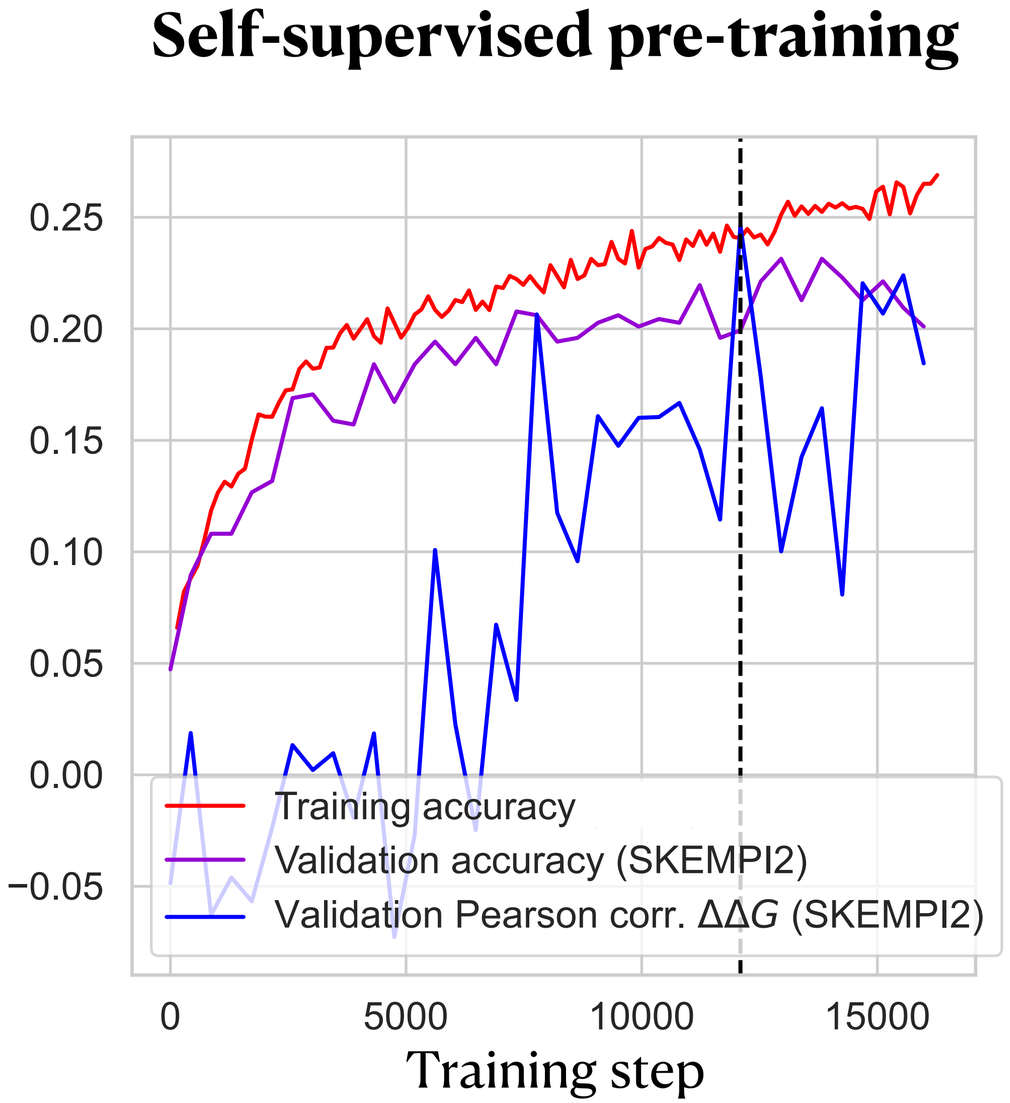


PPIformer captures biochemical principles



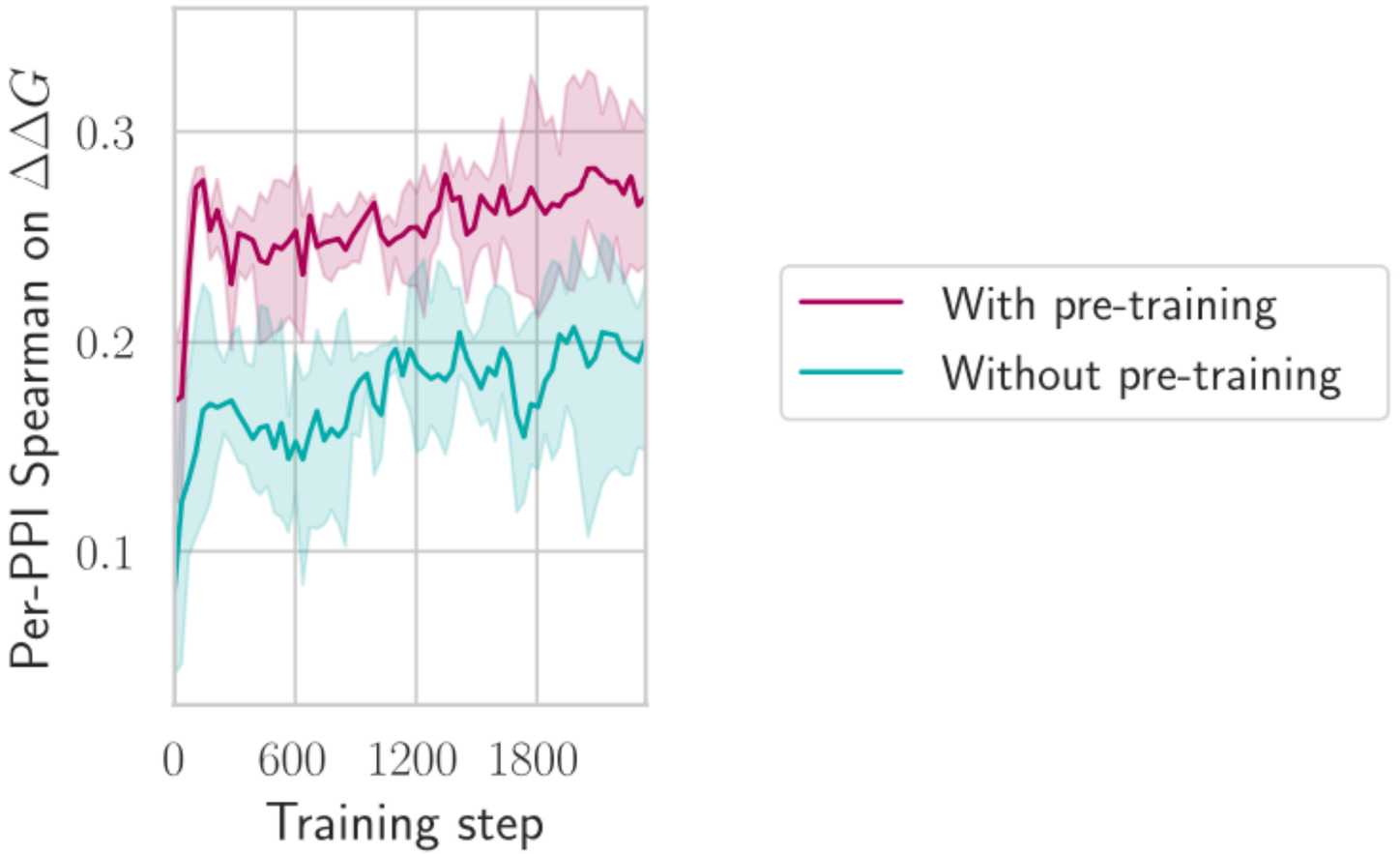


Emergence of mutation scoring capabilities



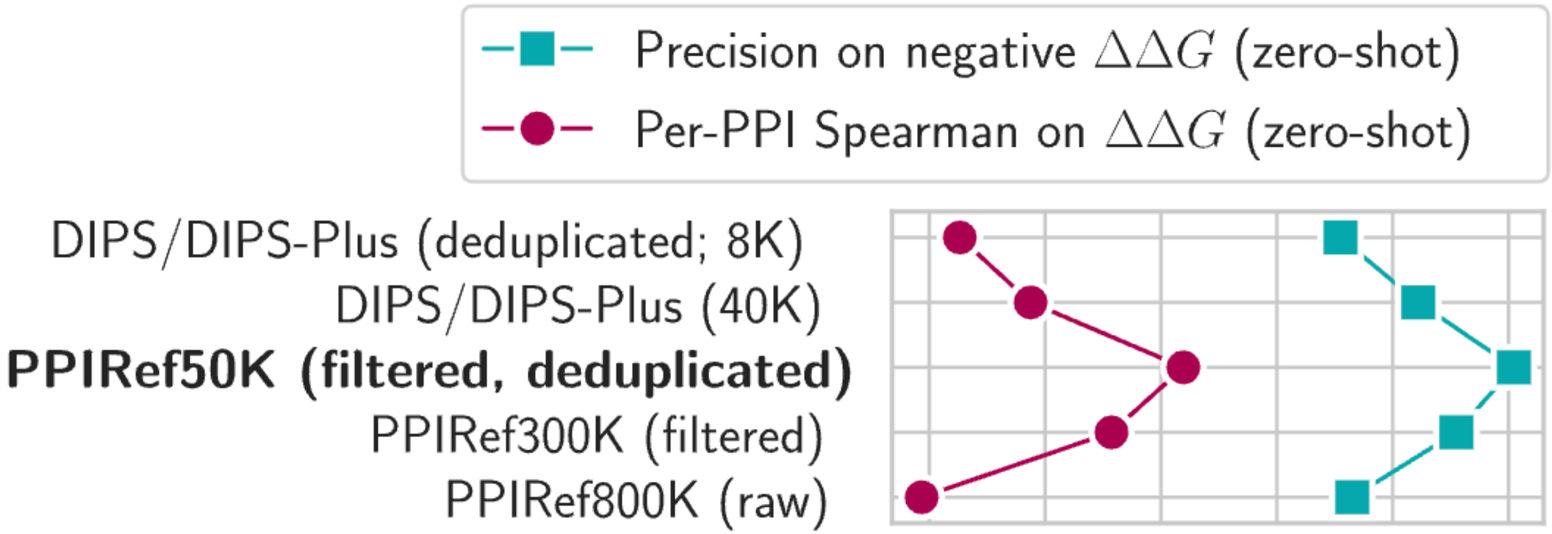
Pre-training is crucial for fine-tuning

Supervised ΔΔG fine-tuning



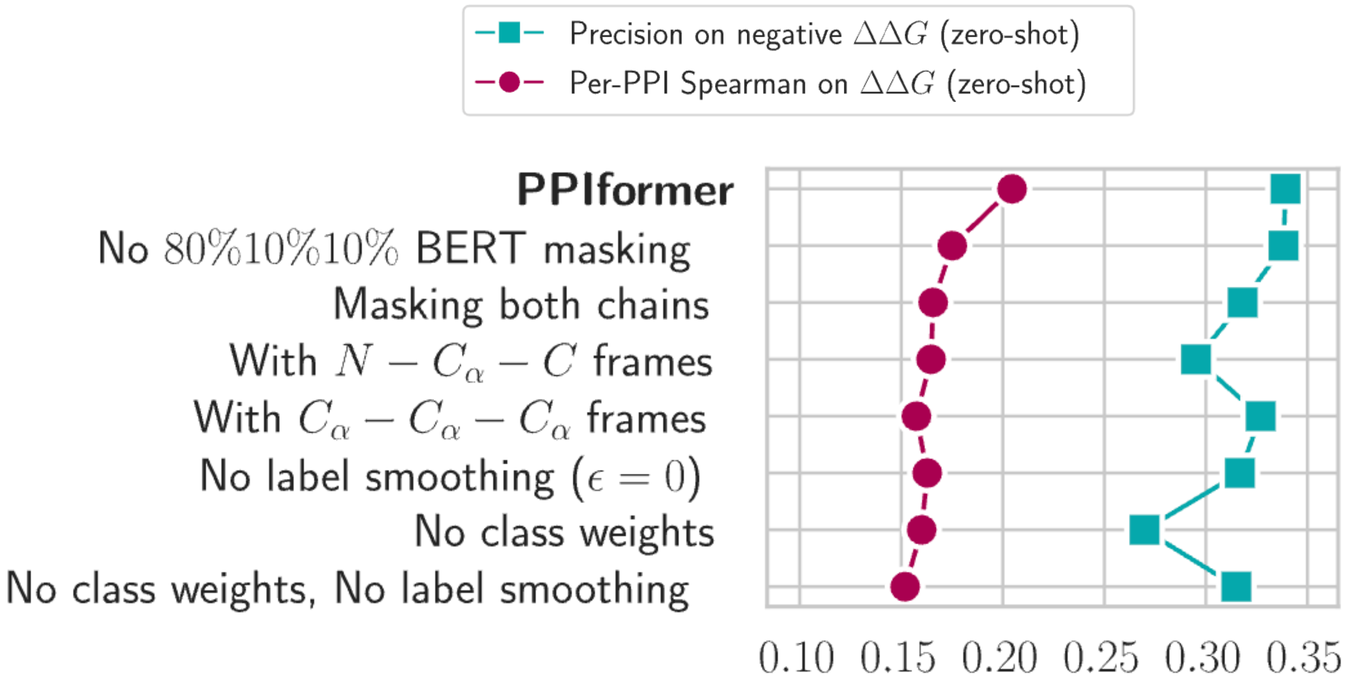
Pre-training from PPIRef is crucial

Pre-training



Other key ingredients

Pre-training



Comparison with the state of the art: 5 independent PPIs from SKEMPI

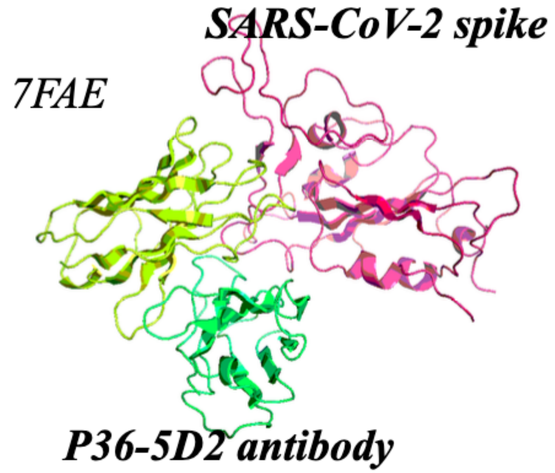
Category	Method	Spearman ↑	Pearson ↑	Precision ↑	Recall ↑	ROC AUC ↑	MAE ↓	RMSE ↓
Force field simulations	FLEX DDG* [22]	0.54	0.57	0.63	0.62	0.84	1.60	2.00
Machine learning	MSA TRANSFORMER [23]	<u>0.37</u>	<u>0.45</u>	0.51	0.38	<u>0.76</u>	5.99	6.77
	ESM-IF [24]	0.32	0.31	0.36	0.28	0.69	1.84	2.11
	RDE-NET. [21]	0.24	0.30	<u>0.54</u>	**0.65**	0.67	<u>1.70</u>	<u>2.02</u>
	PPIFORMER (OURS)	**0.42**	**0.46**	**0.58**	<u>0.61</u>	**0.77**	**1.64**	**1.94**

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

Comparison with the state of the art: COVID

	known favorable mutations					precision		
Method	TH31W↓	AH53F↓	NH57L↓	RH103M↓	LH104F↓	P@1↑	P@5%↑	P@10%↑
MSA TRANSFORMER	56.88	42.11	63.56	49.19	18.83	0.00	0.00	0.00
ROSETTA	10.73	76.72	93.93	13.56	**6.88**	0.00	0.00	2.04
FOLDX	**5.67****	68.22	**2.63**	12.35	29.96	0.00	4.00	4.08
DDGPRED	**2.02**	14.17	24.49	**4.05**	**6.48**	0.00	**8.00**	**6.12**
END-TO-END	11.34	16.60	**8.30**	52.43	80.36	0.00	0.00	2.04
MIF-NET.	27.94	66.19	**8.50**	17.21	36.23	0.00	0.00	2.04
ESM-IF	49.39	17.61	17.00	51.42	48.58	0.00	0.00	0.00
RDE-NET.	**1.62**	**2.02**	20.65	61.54	**5.47**	0.00	**8.00**	**6.12**
PPIFORMER (OURS)	18.02	**0.20**	**7.69**	21.46	10.93	**100**	4.00	4.08

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



Comparison with the state of the art: stroke

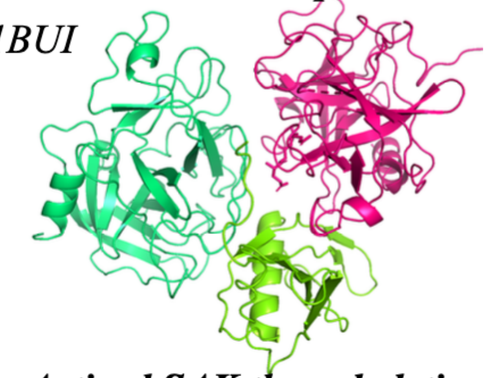
known favorable mutations

precision

Method	Mutations with ≥ 2× activity enhancement						Activity enhancement		
	KC130A ↓	KC130T ↓	KC135R KC130T ↓	KC135A ↓	KC74Q KC130E KC135R ↓	KC74R KC130T KC135R ↓	P@1 ↑	P@5% ↑	P@10% ↑
MSA TRANSFORMER	52.50	32.50	55.00	40.0	70.00	78.75	**100**	50.00	37.50
ESM-IF	45.00	33.75	46.25	25.0	42.50	58.75	0.00	0.00	25.00
RDE-NET.	51.25	33.75	22.50	15.00	27.50	**5.00**	0.00	50.00	62.50
PPIFORMER (OURS)	66.25	15.00	**2.50**	52.50	33.75	**1.25**	**100**	**75.00**	**87.50**

Microplasmin

1BUI



Actived SAK thrombolytic

LEARNING TO DESIGN PROTEIN–PROTEIN INTERACTIONS WITH ENHANCED GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Discovering mutations enhancing protein–protein interactions (PPIs) is critical for advancing biomedical research and developing improved therapeutics. While machine learning approaches have substantially advanced the field, they often struggle to generalize beyond training data in practical scenarios. The contributions of this work are three-fold. First, we construct PPIRef, the largest and non-redundant dataset of 3D protein–protein interactions, enabling effective large-scale learning. Second, we leverage PPIRef to pre-train PPIformer, a new SE(3)-equivariant model generalizing across diverse protein-binder variants. We fine-tune PPIformer to predict effects of mutations on protein–protein interactions via a thermodynamically motivated adjustment of the pre-training loss function. Finally, we demonstrate the enhanced generalization of our new PPIformer approach by outperforming other state-of-the-art methods on the new non-leaking splits of the standard labeled PPI mutational data and independent case studies optimizing a human antibody against SARS-CoV-2 and increasing staphylokinase thrombolytic activity.

PPIformer Github

anton.bushuiev@cvut.cz

References

1. Beckstette et al., Fast index based algorithms and software for matching position specific scoring matrices, BMC bioinformatics, 2006
2. Sumbalova et al., Hotspot wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information, Nucleic acids research, 2018
3. Meier et al., Language models enable zero-shot prediction of the effects of mutations on protein function, Advances in Neural Information Processing Systems, 2021
4. Shroff et al., Discovery of novel gain-of-function mutations guided by structure-based deep learning, ACS synthetic biology, 2020
5. Dauparas et al., Robust deep learning–based protein sequence design using ProteinMPNN, Science, 2022
6. Biota et al., Discovery studio modeling environment, Dassault Systemes, 2017
7. Dehouck et al., Beatmusic: prediction of changes in protein–protein binding affinity on mutations, Nucleic acids research, 2013
8. Delgado et al., Foldx 5.0: working with rna, small molecules and a new graphical interface, Bioinformatics, 2019
9. Geng et al., isee: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations, Proteins: Structure, Function and Bioinformatics, 2019
10. Rodrigues et al., mmcsm-ppi: predicting the effects of multiple point mutations on protein–protein interactions, Nucleic Acids Research, 2021
11. Xiong et al., Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts, Journal of molecular biology, 2017
12. Zhou et al., Mutation effect estimation on protein–protein interactions using deep contextualized representation learning 2020, NAR genomics and bioinformatics, 2020
13. Wang et al., A topology-based network tree for the predic- tion of protein–protein binding affinity changes following mutation, Nature Machine Intelligence, 2020
14. Jiang et al., Dgcddg: Deep graph convolution for predicting protein-protein binding affinity changes upon mutations, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023
15. Pahari et al., Saambe-3d: predicting effect of mutations on protein–protein interactions, International journal of molecular sciences, 2020
16. Liu et al., Deep geometric representations for modeling effects of mutations on protein-protein binding affinity, PLoS computational biology, 2021
17. Li et al., Saambe-seq: a sequence-based method for predicting mutation effect on protein–protein binding affinity, Bioinformatics, 2021
18. Gao and Skolnick, ialign: a method for the structural comparison of protein– protein interfaces, Bioinformatics, 2010
19. Townshend et al., End-to-end learning on 3d protein structure for interface prediction, Advances in Neural Information Processing Systems, 2019
20. Shan, Sisi, et al. "Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization." Proceedings of the National Academy of Sciences 119.11 (2022): e2122954119.
21. Luo, Shitong, et al. "Rotamer Density Estimator is an Unsupervised Learner of the Effect of Mutations on Protein-Protein Interaction." bioRxiv (2023): 2023-02.
22. Barlow, Kyle A., et al. "Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation." The Journal of Physical Chemistry B 122.21 (2018): 5389-5399.
23. Rao, Roshan M., et al. "MSA transformer." International Conference on Machine Learning. PMLR, 2021.
24. Hsu, Chloe, et al. "Learning inverse folding from millions of predicted structures." International Conference on Machine Learning. PMLR, 2022.