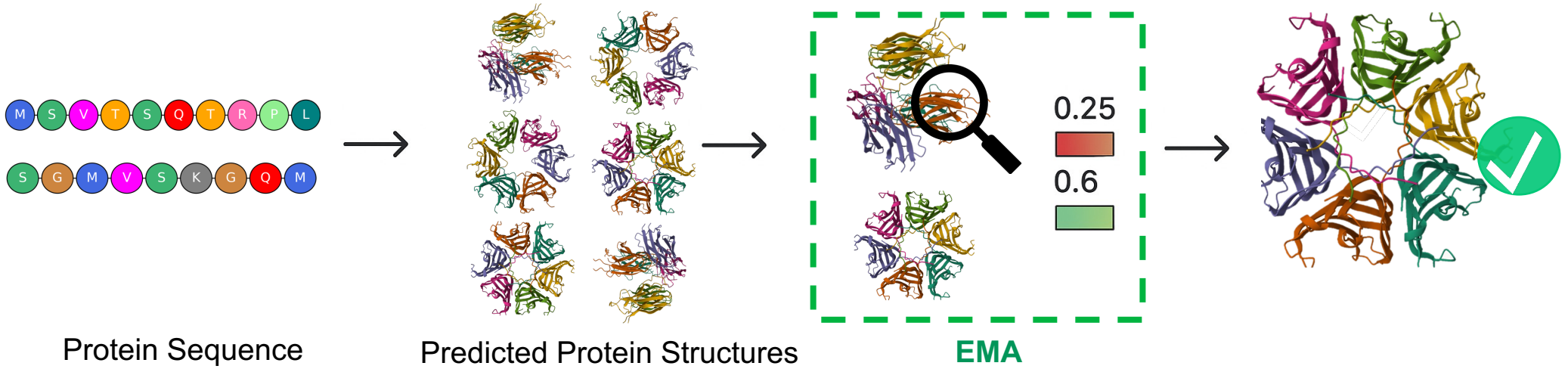
PSBench: a large-scale benchmark for estimating the accuracy of protein complex structural models

Pawan Neupane, Jian Liu, Jianlin Cheng

Bioinformatics and Machine Learning Laboratory (BML)

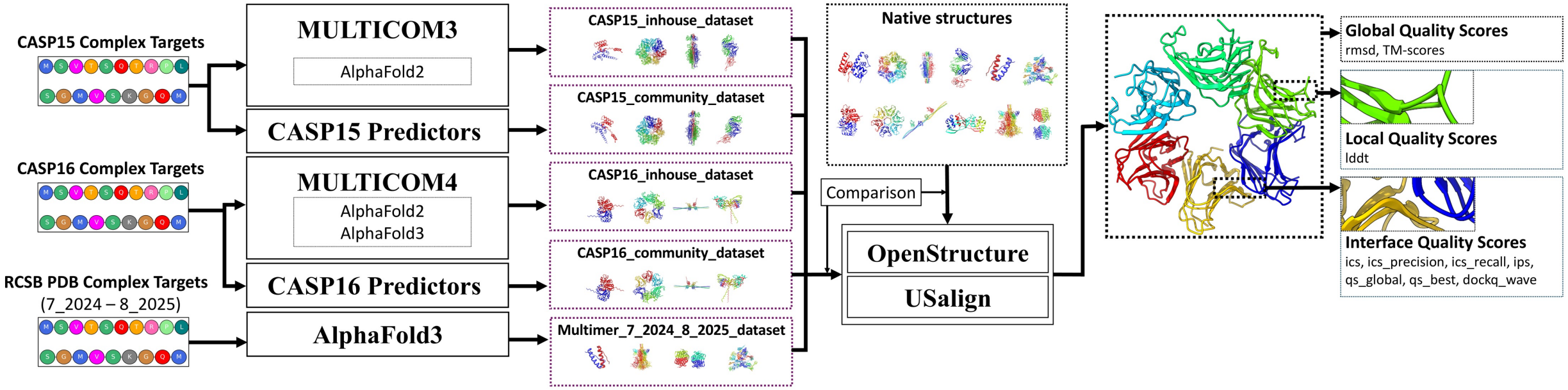
Dept. of Electrical Eng. & Computer Sci.
University of Missouri, Columbia, MO, USA

Motivation



- Protein complex structures are central to understanding biological function, disease, and drug design.
- AlphaFold2 and AlphaFold3 generate accurate structural models, but their self-reported confidence scores are not always reliable for model selection.
- Selecting high-quality models is often harder than generating them.
- **E**stimation of **M**odel **A**ccuracy (**EMA**) is therefore essential — but limited by lack of large, diverse, well-annotated datasets to train and test EMA methods.

Model Generation & Annotation Pipeline



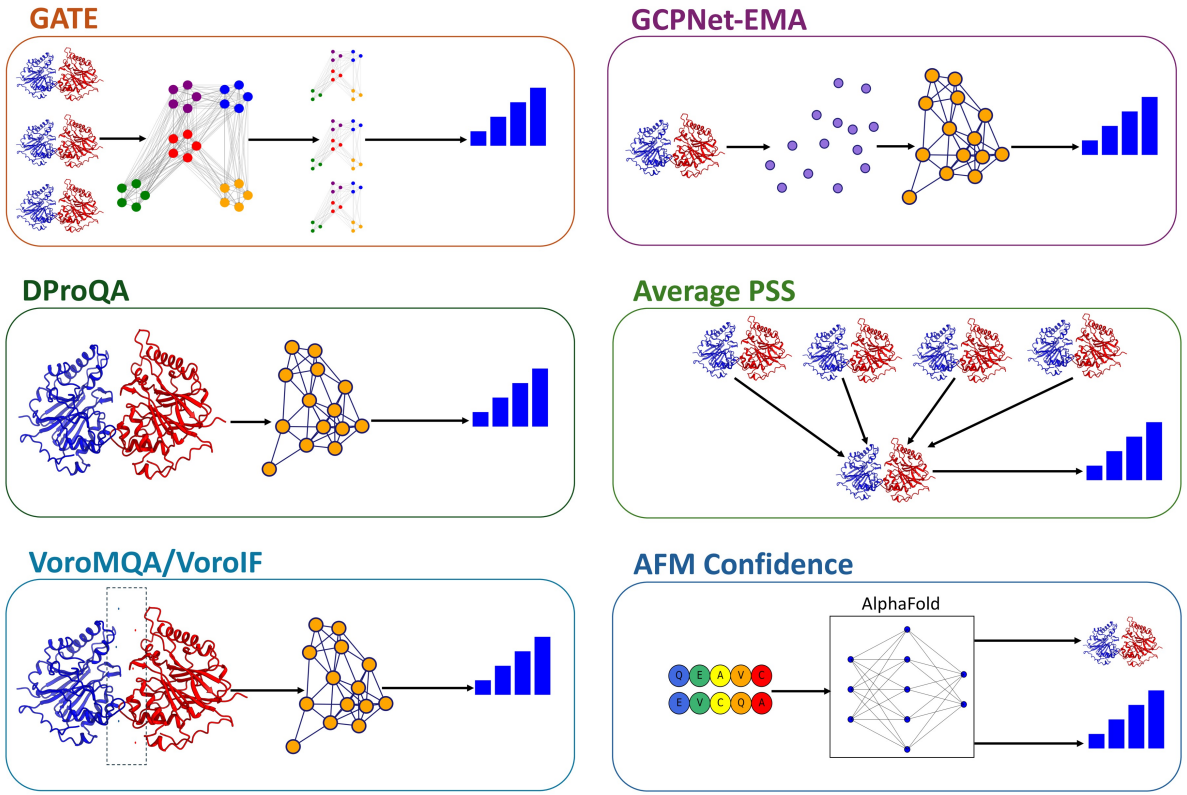
May 1st, 2022

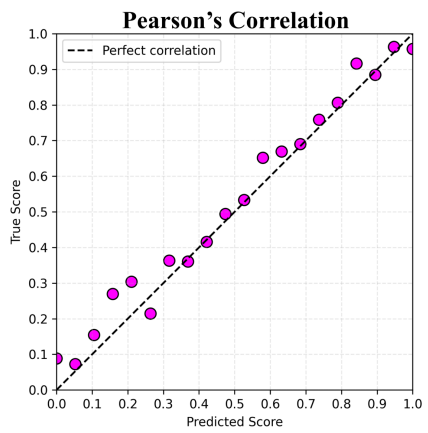
May 1st, 2024

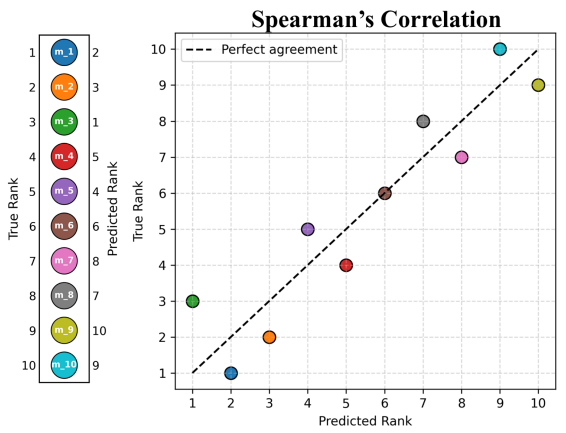
July, 2024 – August 2025

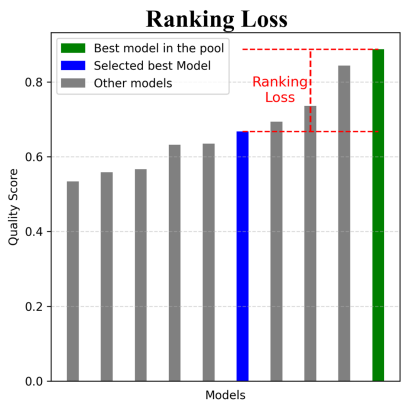
- **CASP15_inhouse_dataset: 7,885 models (31 targets)**
- **CASP15_community_dataset: 10,942 models (40 targets)**
- **CASP16_inhouse_dataset: ~1 M models (36 targets)**
- **CASP16_community_dataset: 12,904 models (39 targets)**
- **Multimer_7_2024_8_2025_dataset: 400,400 models (2,002 targets)**

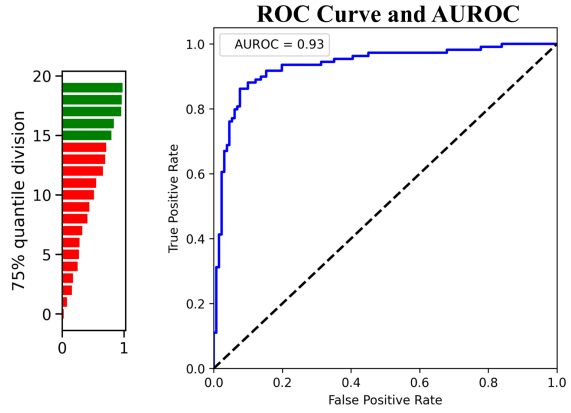
Methods & Metrics



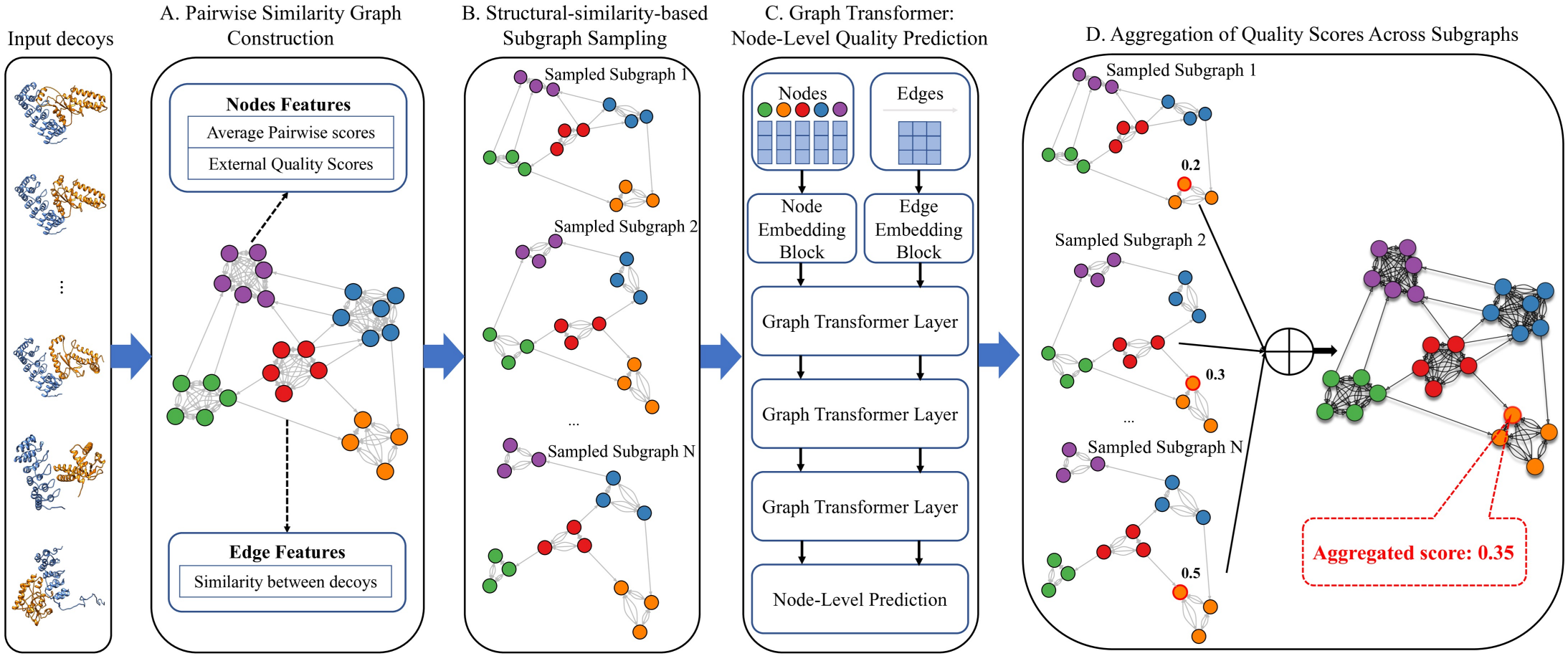








GATE (Trained and Tested on PSBench)



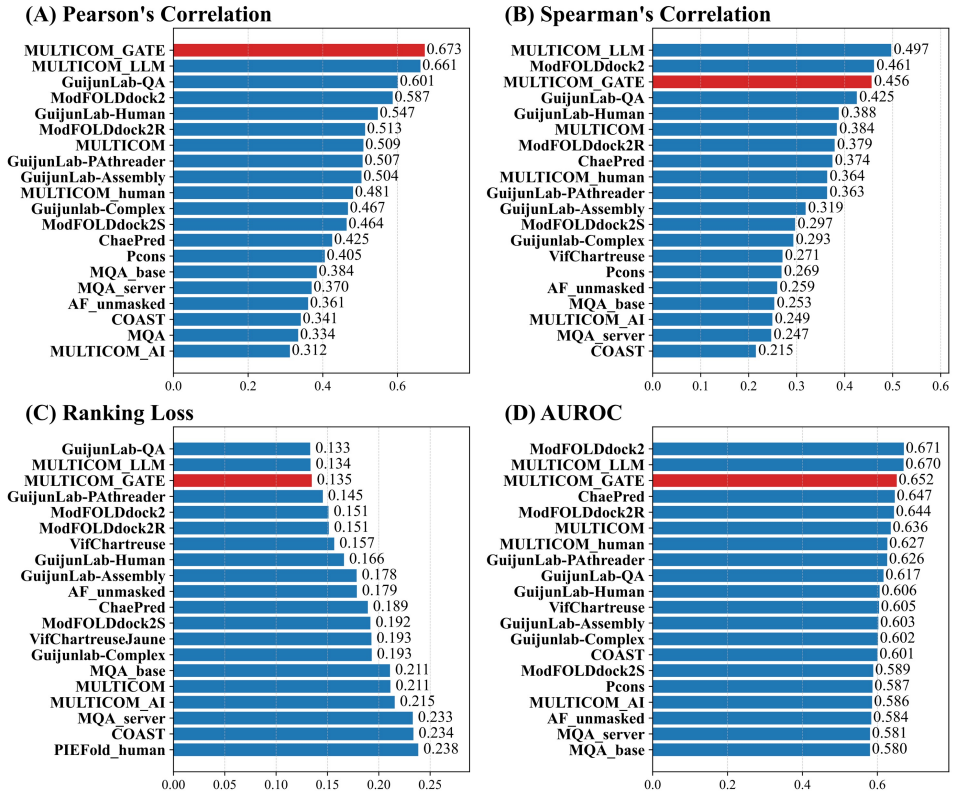
Evaluation (Blind CASP16 experiment)

CASP16 Inhouse models

Method	TM-score				DockQ_wave			
	$Corr^P$ ↑	$Corr^S$ ↑	Loss ↓	AUROC ↑	$Corr^P$ ↑	$Corr^S$ ↑	Loss ↓	AUROC ↑
GATE-AFM	0.372	0.283	0.102	0.658	0.431	0.322	0.138	0.662
AFM Confidence	0.259*	0.143*	0.106	0.597*	0.252*	0.114*	0.151	0.593*
PSS	0.394	0.261	0.114	0.647	0.369	0.284	0.154	0.645
GCPNet-EMA	0.360	0.249	0.135	0.643	0.355	0.264	0.169	0.648
VoroMQA-dark	0.039*	0.144	0.129	0.609	-0.013*	0.146*	0.163	0.622
VoroIF-GNN-pCAD-score	0.073*	0.105*	0.167*	0.589*	0.074*	0.137*	0.204	0.615
VoroIF-GNN-score	0.065*	0.116*	0.193*	0.599*	0.114*	0.170*	0.207*	0.622
DProQA	-0.051*	0.011*	0.194*	0.569*	0.032*	0.071*	0.223*	0.587*

Green: Best
Yellow: Second-best

CASP16 EMA competition



Datasets and codes

- Codes: https://github.com/BioinfoMachineLearning/PSBench
- Datasets: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/75SZ1U

Acknowledgements:

Tools and data sources:

Funding:

PSBench (Paper)

PSBench (Github)