Toward reliability evaluation of computational models of protein molecules and their interactions

Md Hossain Shuvo, Ph.D.
Assistant Professor
Department of Computer Science
Prairie View A&M University

CCSB@PVAMU

February 19th,2025

Proteins

A fundamental molecule and workhorse of cells

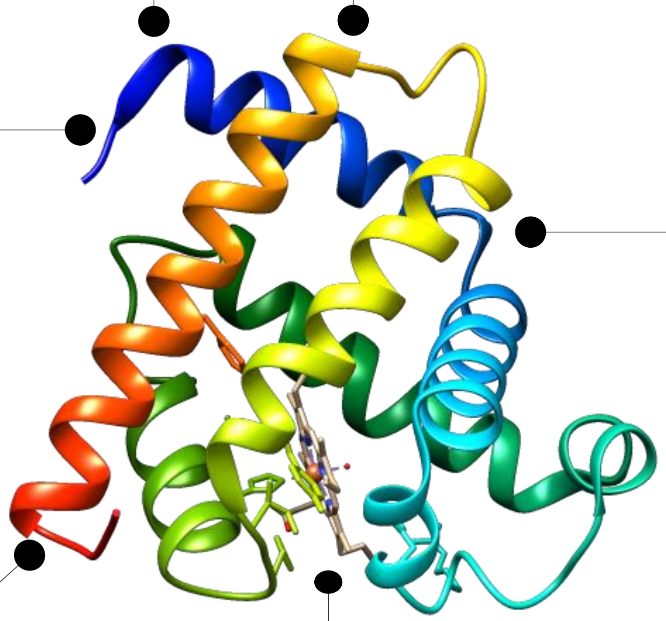
Participates in most biological processes

Diving mammals

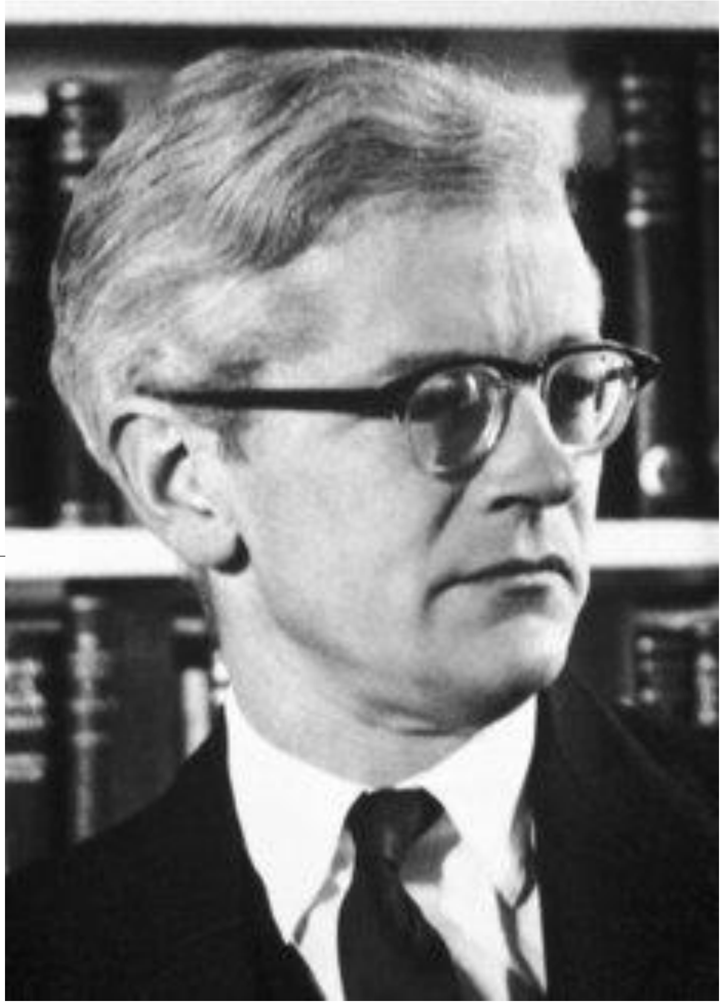
Skeletal muscle

Myoglobin

PDB ID: 1MBN



Oxygen storage



Dr. John Cowdery Kendrew

Solved the structure in 1957

Noble prize in Chemistry 1962

How to obtain protein 3D structure?

Experimental approaches

- X-ray crystallography
- Nucleic Magnetic Resonance
- Cryo-Electron Microscopy

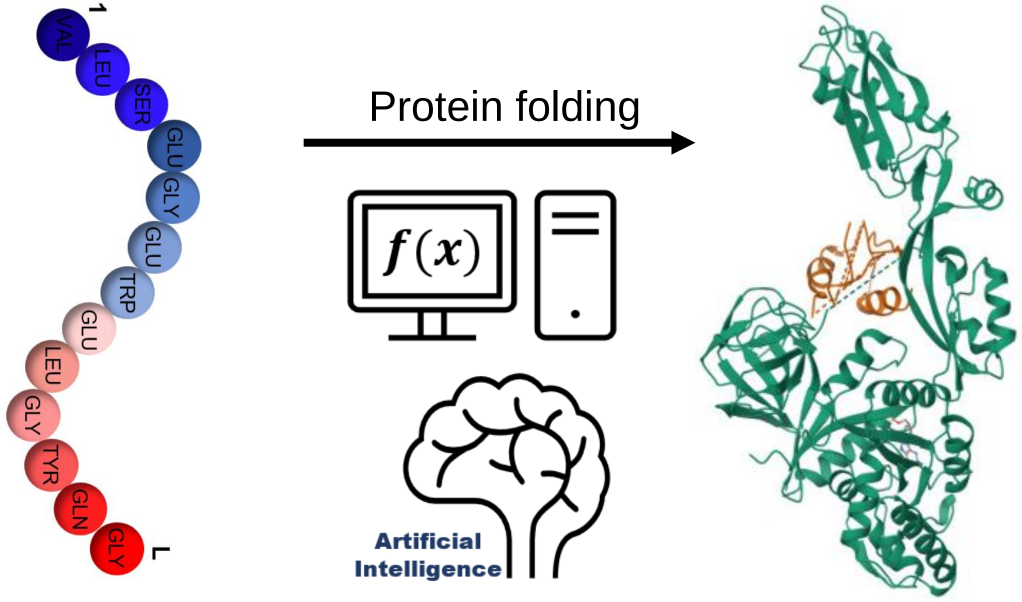
Drawbacks:

- Expensive
- Extensive
- Leads to gap between sequence and structure

Dr. Christian Anfinsen

Noble prize in 1972

“demonstrated that the amino acid sequence of a protein contained all of the information needed for the protein to reach the native conformation”



Computational protein structure prediction can help

Clark, 2008

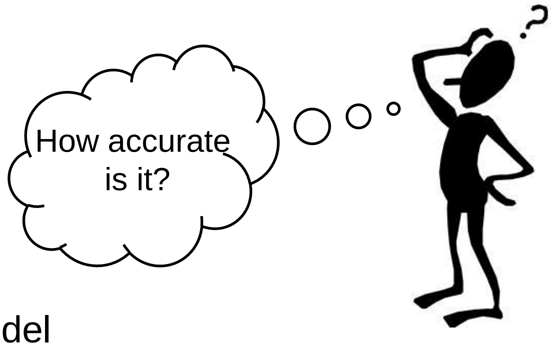
Computationally predicted protein models may have error..

Protein model quality estimation

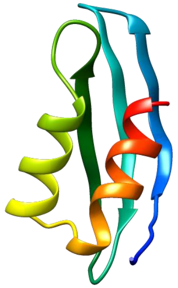
Protein model quality estimation

Model quality estimation

Computationally
predicted protein model



Model



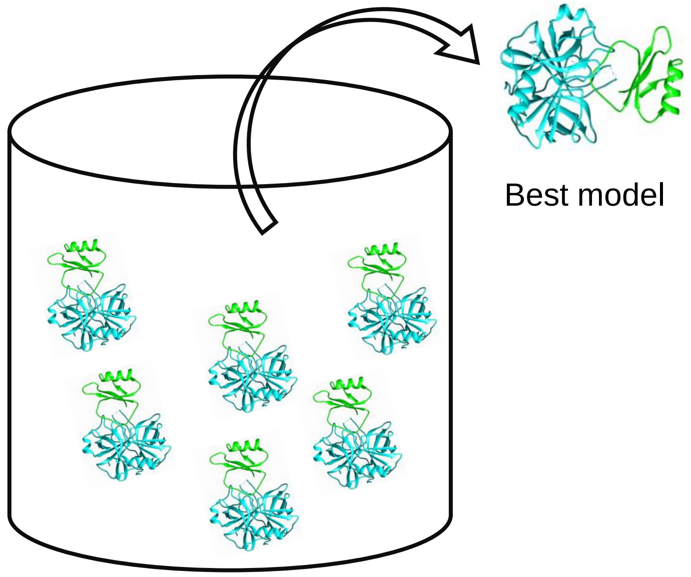
Experimental/Ground truth

Estimation of protein model quality

- When the experimental structure is not known/absent

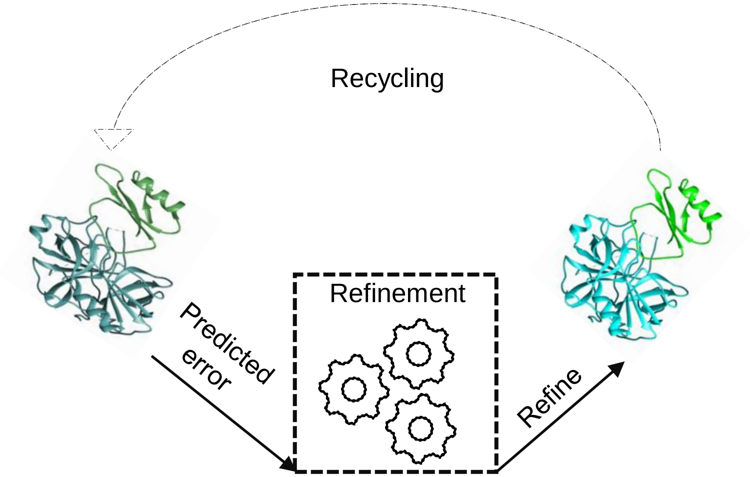
Application: protein structure prediction

Model selection



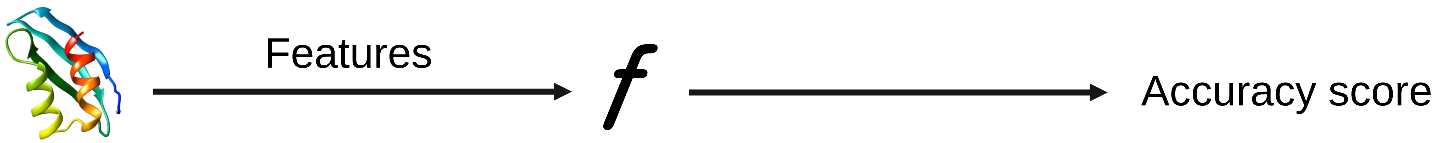
Pool of candidate structures

Model refinement



Helps in accurately guide the process of protein structure prediction

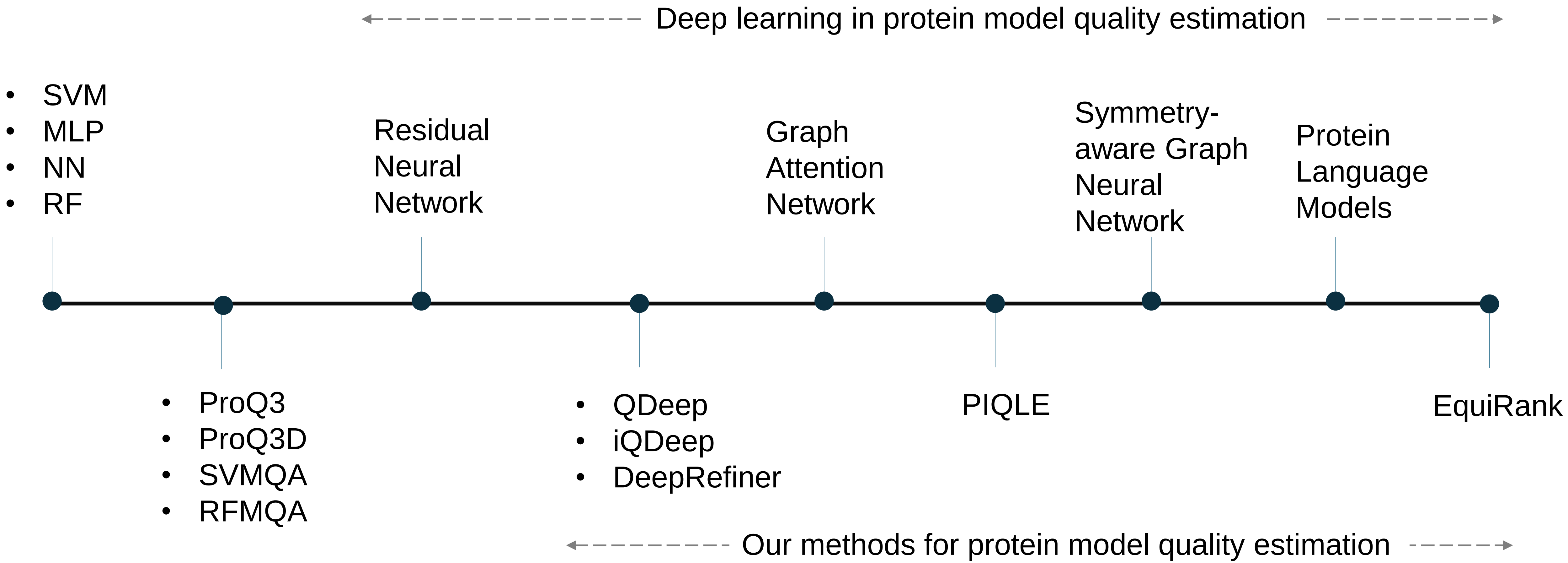
Application of deep learning in model quality estimation



Features

f

Accuracy score



Deep learning in protein model quality estimation

- SVM
- MLP
- NN
- RF

Residual Neural Network

Graph Attention Network

Symmetry-aware Graph Neural Network

Protein Language Models

- ProQ3
- ProQ3D
- SVMQA
- RFMQA

- QDeep
- iQDeep
- DeepRefiner

PIQLE

EquiRank

Our methods for protein model quality estimation

Uziela *et al.*, 2016
Uziela *et al.*, 2017

Manavalan and Lee, 2017
Manavalan *et al.*, 2014

He *et al.*, 2016)
Satorras *et al.*, 2022

Lin *et al.*, 2023
Shuvo *et al.*, 2020, 2021, 2023, 2025

Fundamental research questions

1

How to estimate the quality of predicted protein models?

1. **CSBJ**, 2025
2. **ISMB** 2020
   also, in **Bioinformatics**
   Oxford Press 2020
3. **JMB** 2023
4. **Bioinformatics advances** 2023
5. **PLOS ONE** 2020
6. **Proteins** 2021

2

How to apply quality estimation method to improve quality of predicted protein models?

Improving the quality of less accurate protein models

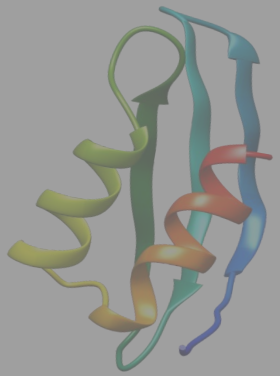
7. **Nucleic Acids Research** 2021

JMB: Journal of Molecular Biology
ISMB: International Society for Computational Biology
CSBJ: Computational and Structural Biotechnology Journal

Key research objectives

1

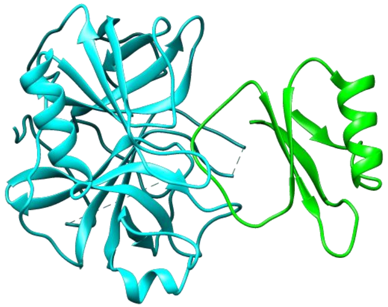
Estimation of **monomeric** protein model quality



Protein model with single chain

2

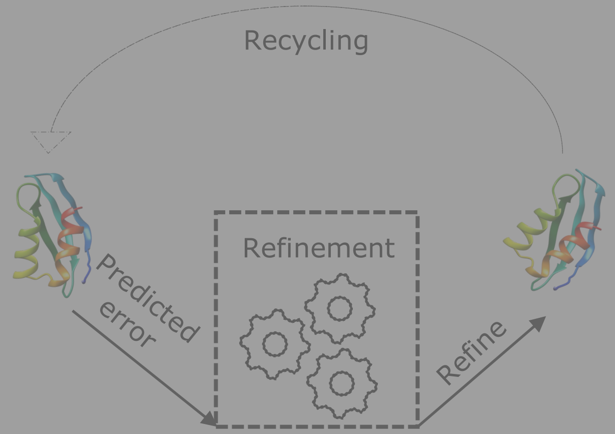
Estimation of **multimeric** protein model quality



Protein model with multiple chains/subunits

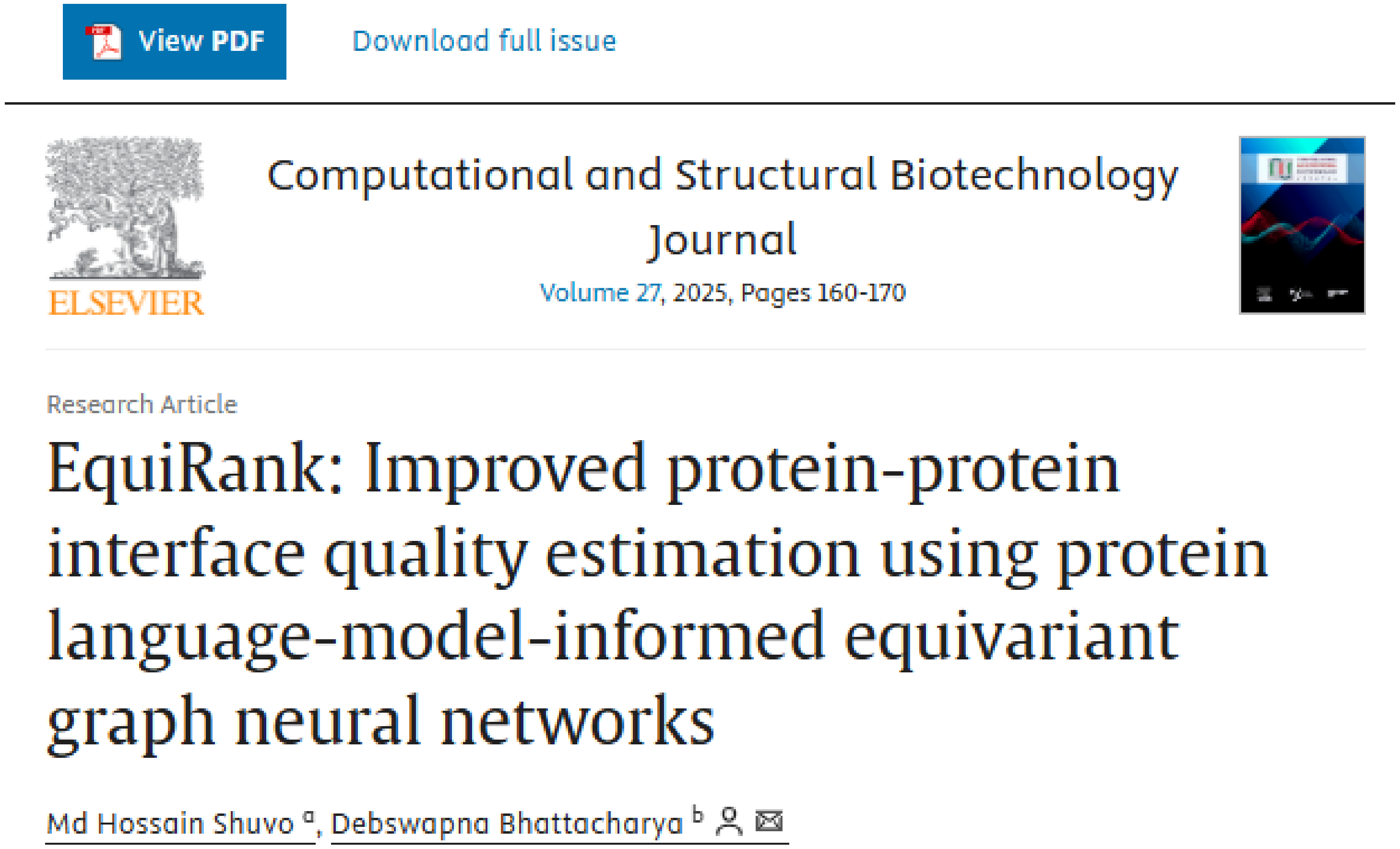
3

Application of quality estimation to improve protein model quality



Iterative refinement of predicted protein models

In this talk…



Computational and Structural Biotechnology Journal

Volume 27, 2025, Pages 160-170

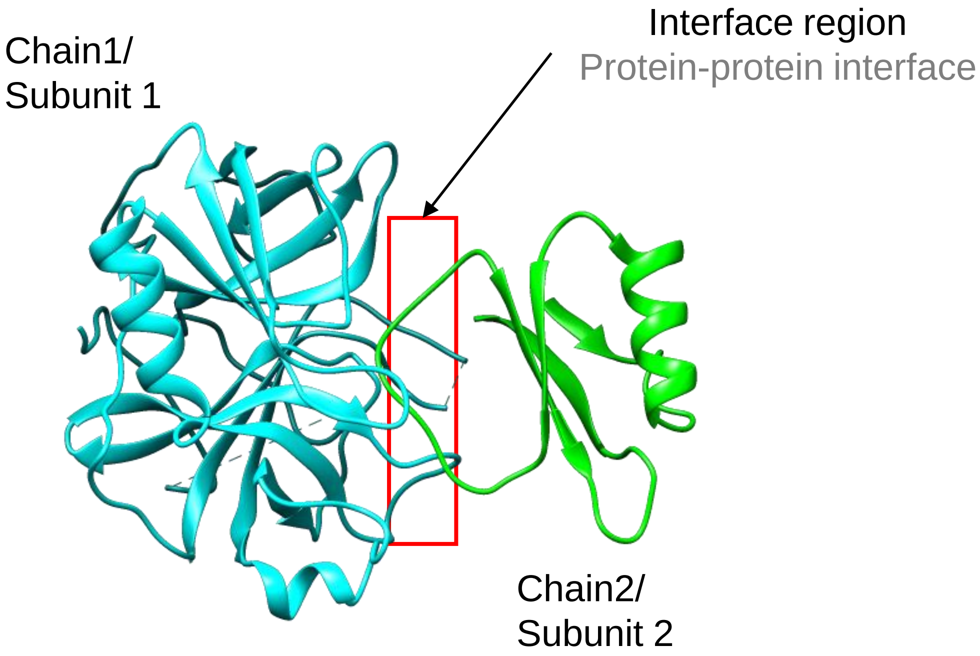
Research Article

EquiRank: Improved protein-protein interface quality estimation using protein language-model-informed equivariant graph neural networks

Md Hossain Shuvo [a], Debswapna Bhattacharya [b]

https://github.com/mhshuvo1/EquiRank

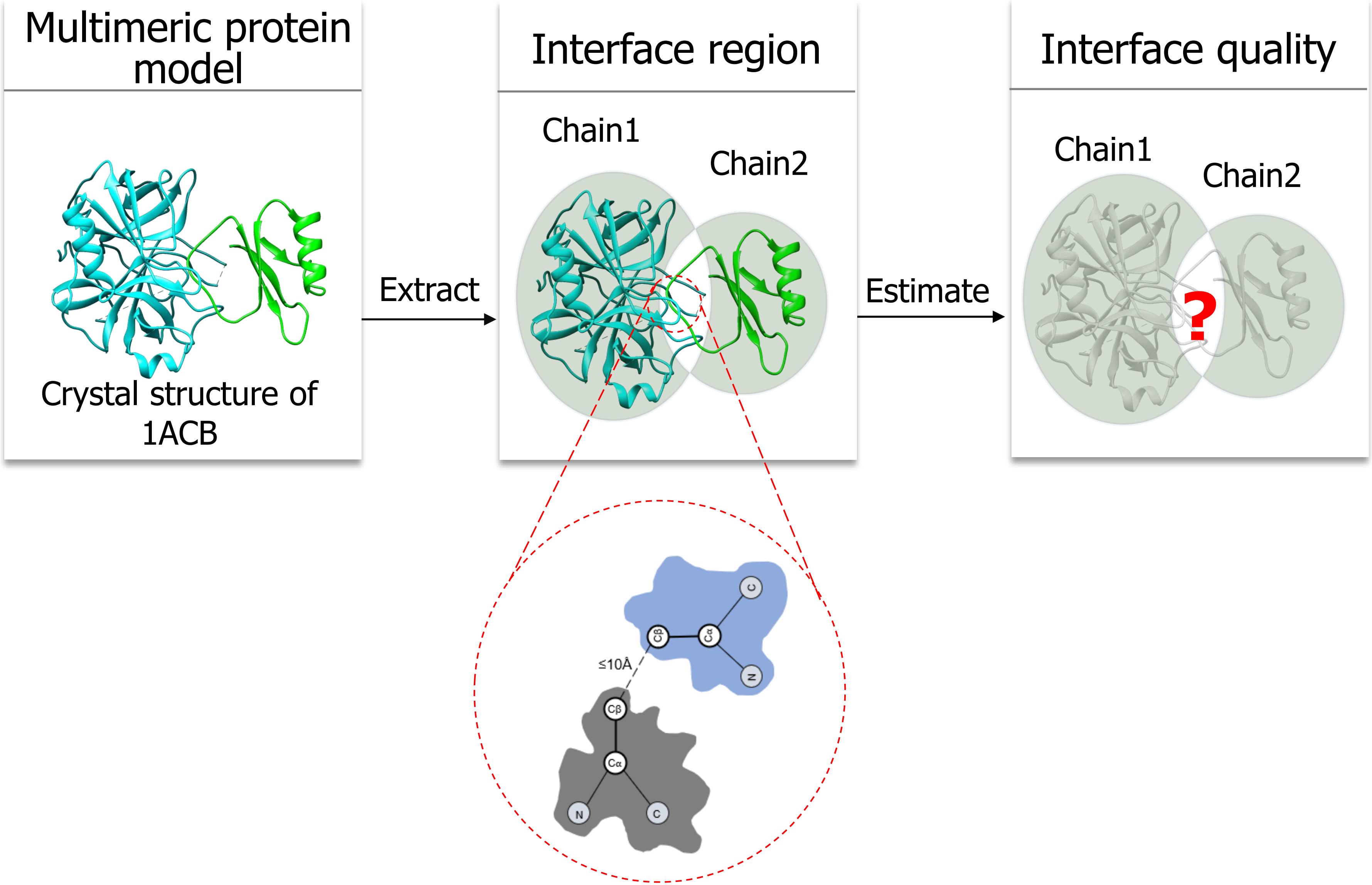
Protein-protein interface of multimeric model



- # of chains/subunits > 1
- A.K.A. protein complex
- Protein-protein interaction
- Catalyzes biological processes
- Interface quality → multimeric protein quality

Protein-protein interface quality estimation

- ➢ Estimation of protein-protein interface quality of computationally predicted protein multimer/protein complexes
- ➢ When the experimental structure is not available/absent



Multimeric protein model

Crystal structure of 1ACB

Extract

Interface region

Chain1

Chain2

Estimate

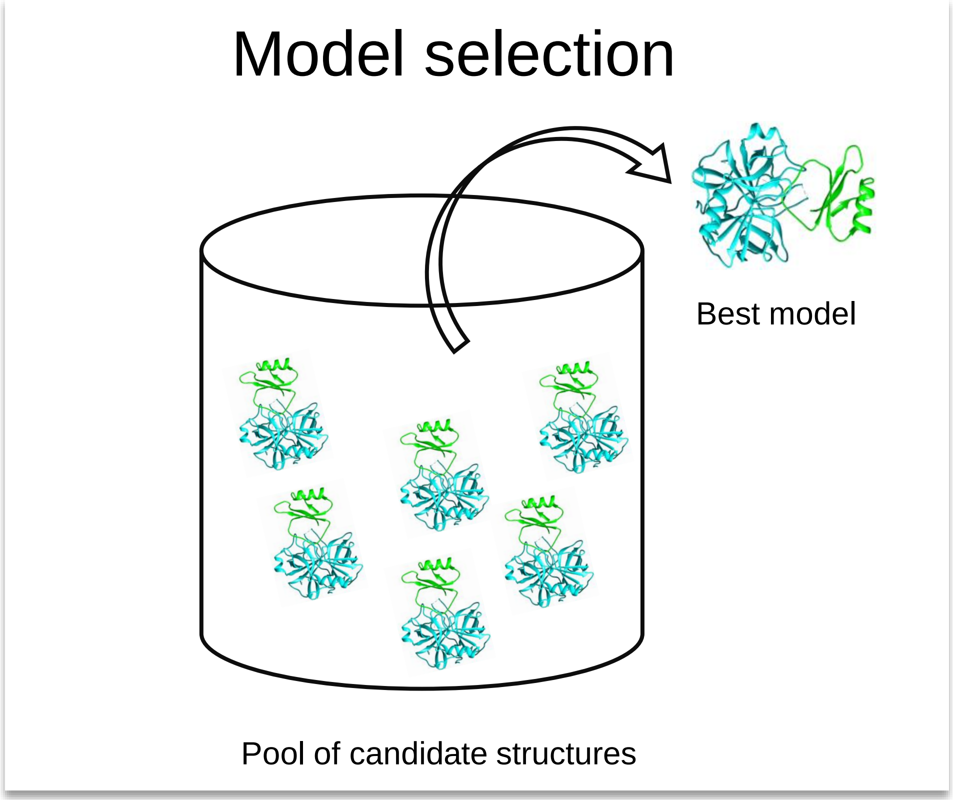
Interface quality

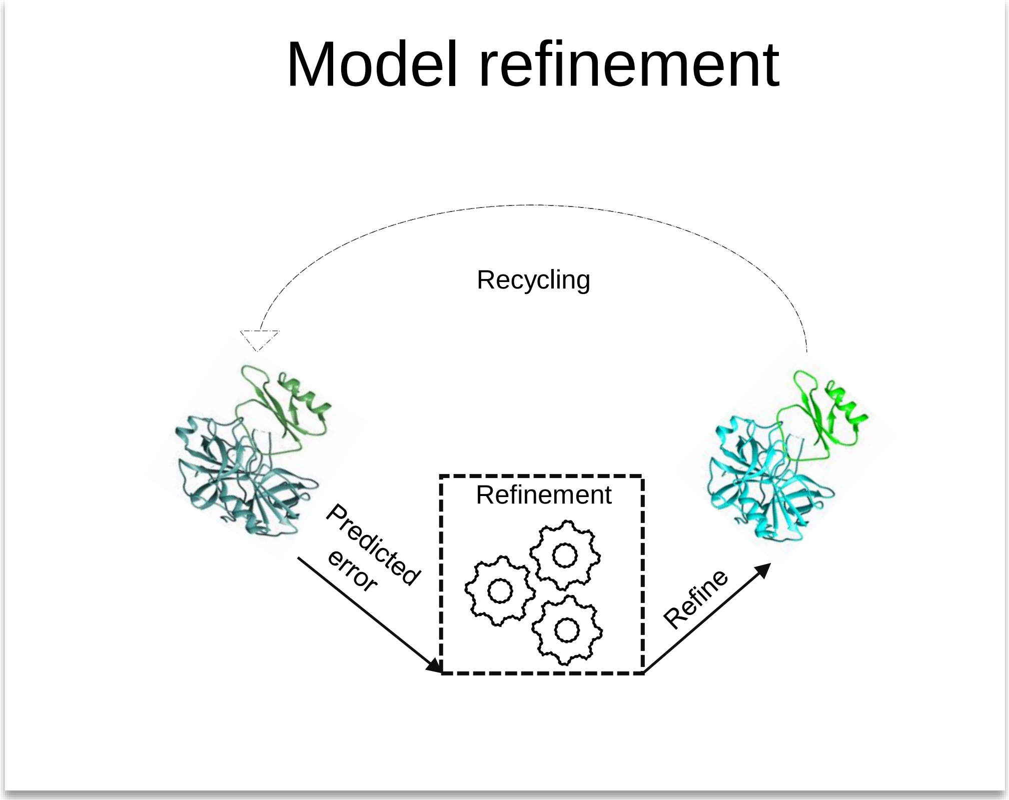
Chain1

Chain2

Marze *et al.*, 2018

Protein complex structure prediction





Helps in accurately guide the process of protein complex structure prediction

Research questions

Representation

1

How to better represent the interface of a protein complex?

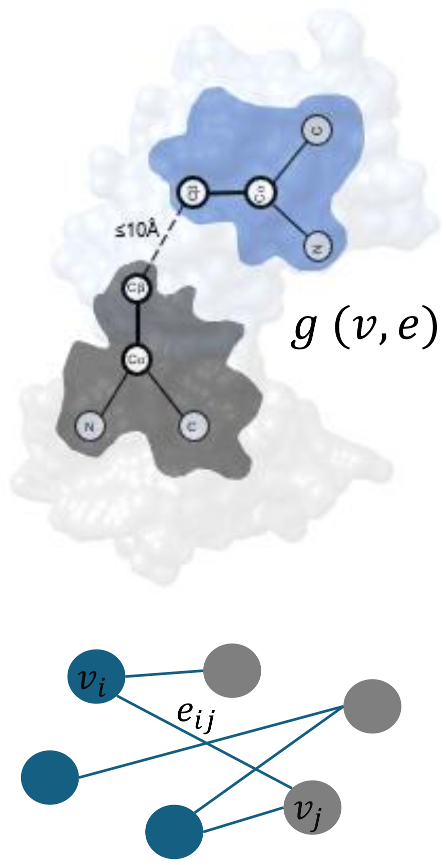
Learning the representation

2

How to better learn the interface representation?

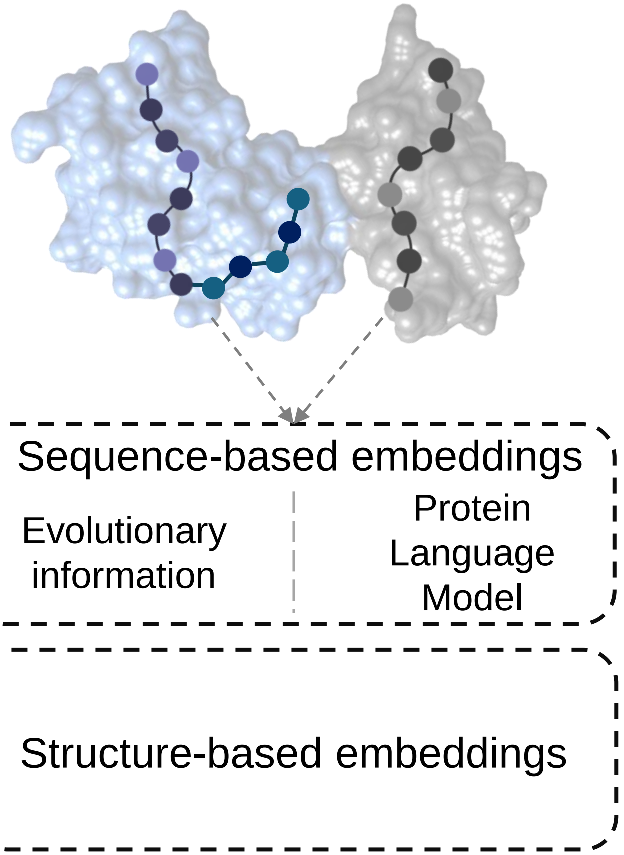
1. Representation: protein-protein interface

Graph representation



Input representations:

Sequence-and structure- based embeddings



Sequence-based embeddings

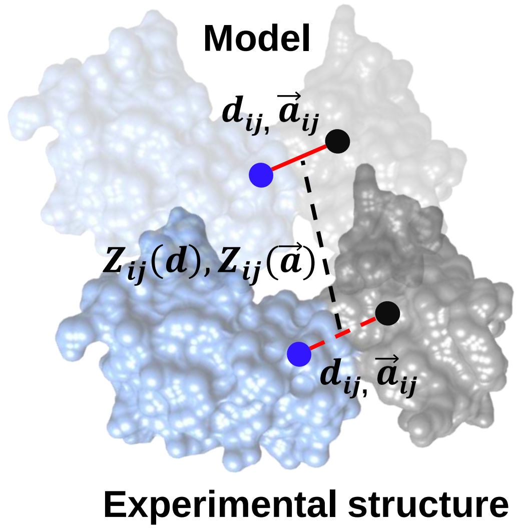
Evolutionary information

Protein Language Model

Structure-based embeddings

Output representations:

Multimeric geometry error

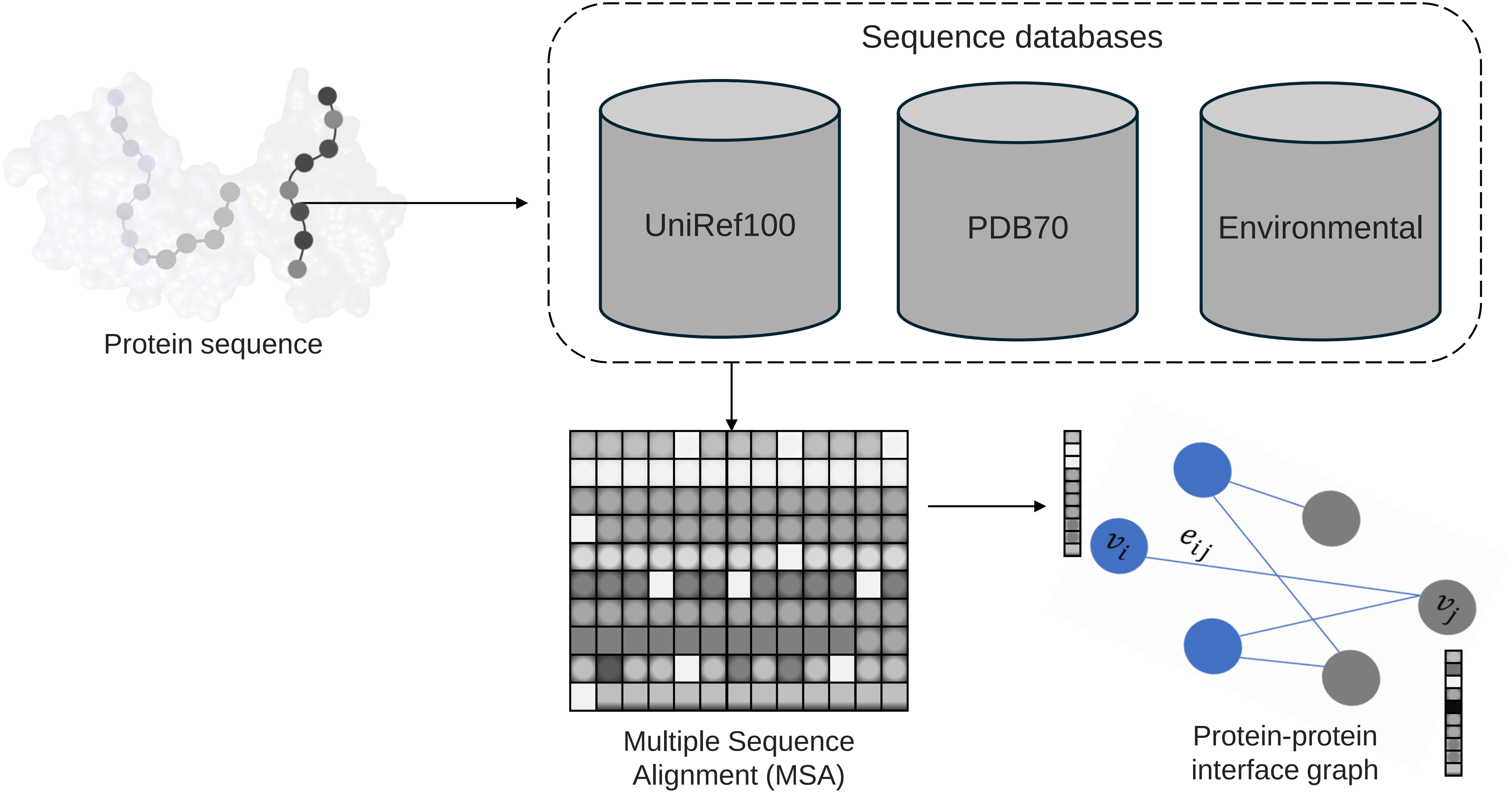


Multimeric distance error: $Z_{ij}(d)$

Multimeric orientation errors:

$Z_{ij}(\vec{a})$, where $\vec{a} \in \Omega, \tau_{ij}, \tau_{ji}, \lambda_{ij}, \lambda_{ji}$

1. Representation: sequence-based embeddings using evolutionary information



Sequence databases

UniRef100

PDB70

Environmental

Protein sequence

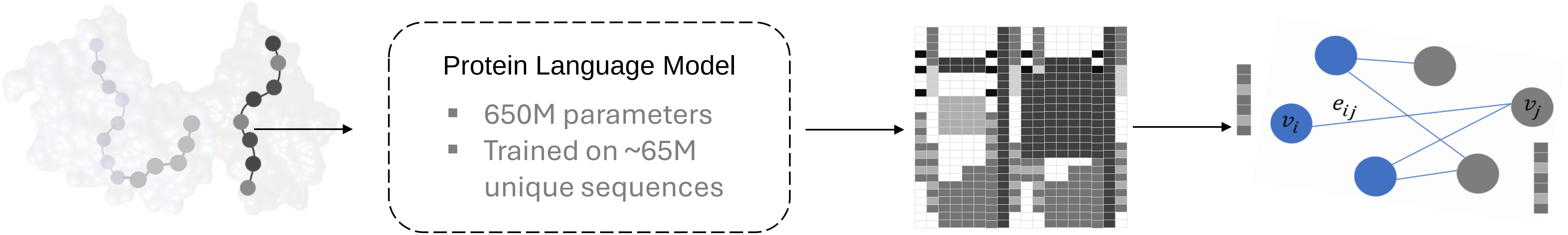
Multiple Sequence Alignment (MSA)

Protein-protein interface graph

UniRef: UniProt Reference Clusters

Mirdita *et al.*, 2022

1. Representation: sequence-based embeddings using Protein Language Model



Protein Language Model

- 650M parameters
- Trained on ~65M unique sequences

Protein sequence

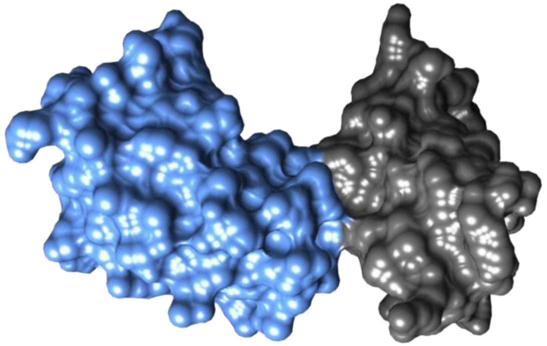
Sequence-based embeddings

Protein-protein interface graph

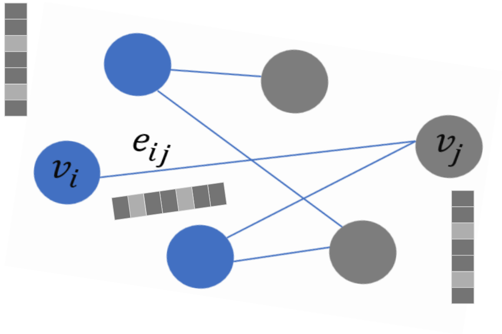
Captures the evolutionary patterns at scale

Lin et al., 2022

1. Representation: structure-based representations



- Physiochemical properties
- Structural topology
- Neighborhood information
- Structural geometry

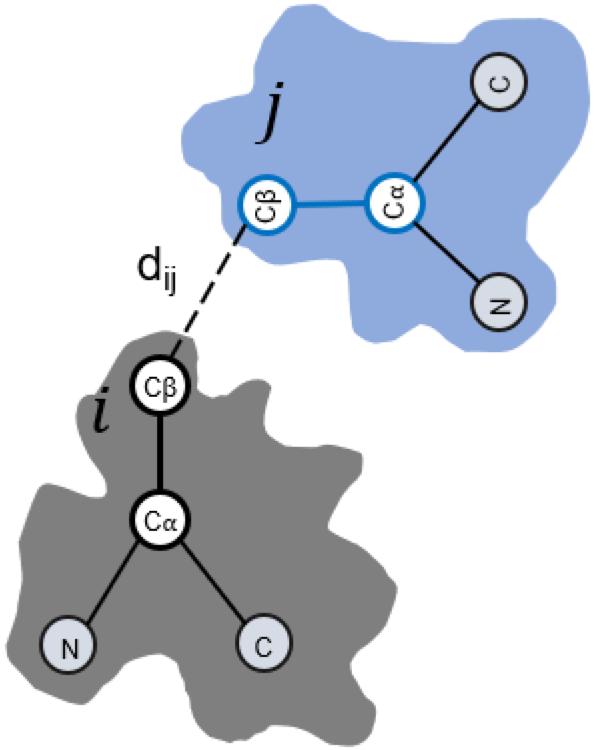


Protein-protein interface graph

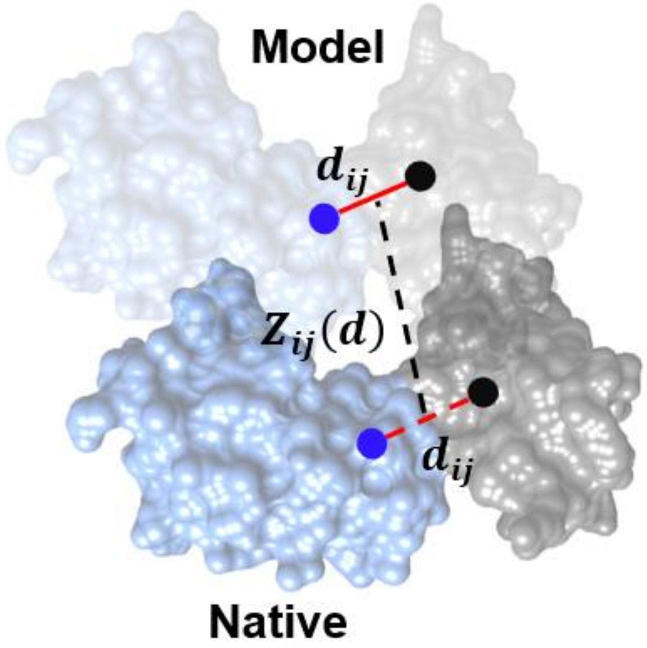
Expressive set of structural embeddings for protein-protein interface

1. Representation: output representation of multimeric **distance errors**

Multimeric distance: d_{ij}



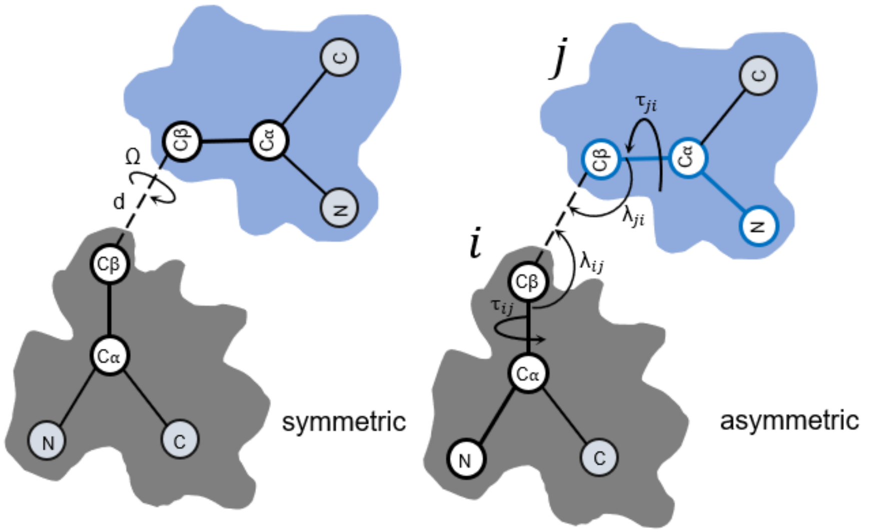
Multimeric distance error: z_{ij}



$$z_{ij}(d) = \begin{cases} 1 & \text{if } d_{ij}^{\text{model}} < 10\,\text{Å and } d_{ij}^{\text{native}} < 10\,\text{Å} \\ \frac{1}{1+\left(\frac{\left|d_{ij}^{\text{model}} - d_{ij}^{\text{native}}\right|}{d_0}\right)^2} & \text{otherwise} \end{cases}$$

1. Representation: output representation of **multimeric orientation errors**

Multimeric orientations: $\vec{a}_{ij}$



Torsion angles: Ω, τ_{ij}, τ_{ji}

Planar angles: λ_{ij}, λ_{ji}

Symmetric $\Omega_{ij} = c\alpha_i - c\beta_i - c\beta_j - c\alpha_j$

Asymmetric

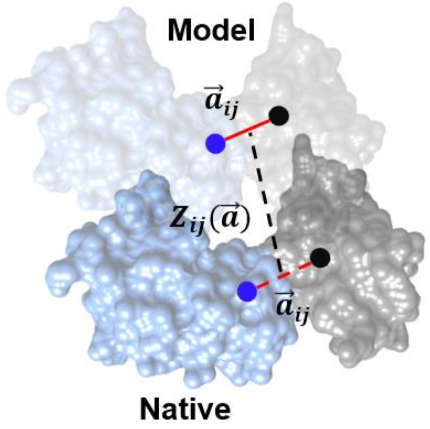
$$\tau_{ij} = N_i - c\alpha_i - c\beta_i - c\beta_j$$

$$\tau_{ji} = N_j - c\alpha_j - c\beta_j - c\beta_i$$

$$\lambda_{ij} = c\alpha_i - c\beta_i - c\beta_j$$

$$\lambda_{ji} = c\alpha_j - c\beta_j - c\beta_i$$

Multimeric orientations errors



Angular RMSD

$$z_{ij}(\vec{a}) = \sqrt{\left(\min\left(\left|a_{ij}^{\text{native}} - a_{ij}^{\text{model}}\right|, 2\pi - \left|a_{ij}^{\text{native}} - a_{ij}^{\text{model}}\right|\right)\right)^2}$$

Where, $\vec{a} \in \Omega, \tau_{ij}, \tau_{ji}, \lambda_{ij}, \lambda_{ji}$

Yang *et al.*, 2020

Representation

1

How to better represent the interface of a protein complex?

Learning the representation

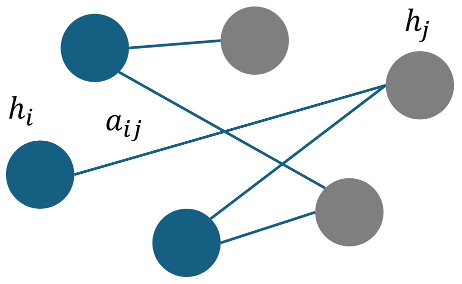
2

How to better learn the interface representation?

How to learn a better mapping between the input and the output representations?

2. Learning the representation

Graph neural network



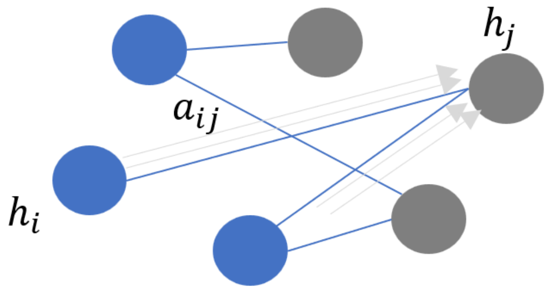
$$\mathbf{m}_{ij} = \phi_e(\mathbf{h}_i^l, \mathbf{h}_j^l, a_{ij})$$

$$\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}$$

$$\mathbf{h}_i^{l+1} = \phi_h(\mathbf{h}_i^l, \mathbf{m}_i)$$

Permutation invariant

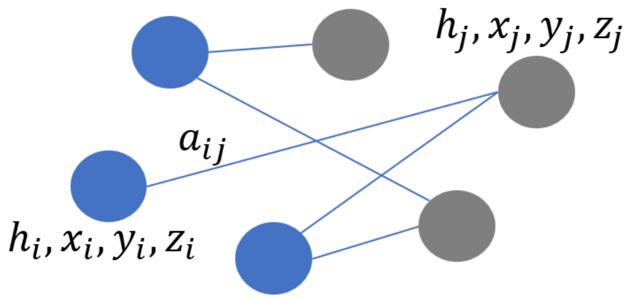
Graph attention network



PIQLE, published in Bioinformatics advances

Shuvo *et al.*, 2023

Symmetry-aware graph neural network



New output representations

- Multimeric distance error
- Multimeric orientation errors

Equivariance

Symmetry-aware equivariant neural network

Zhou et al., 2020
Veličković et al., 2018

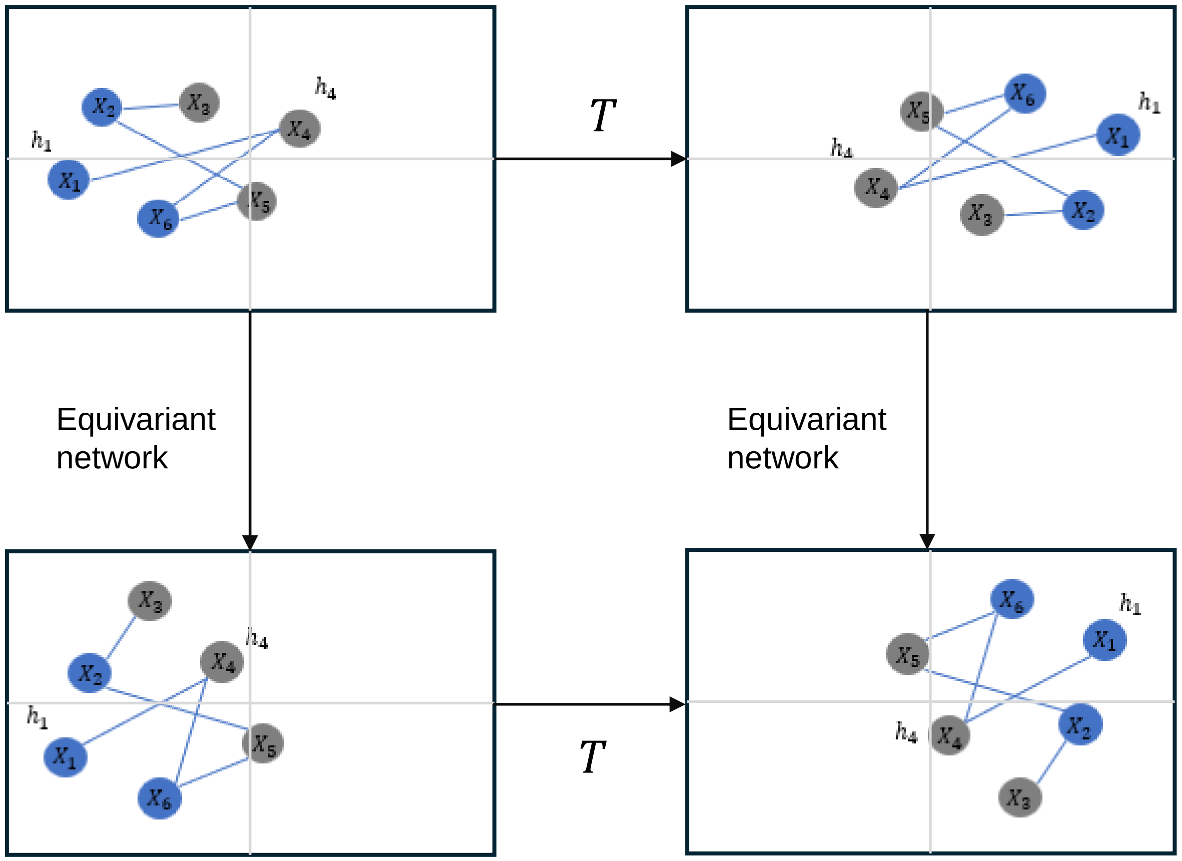
2. Learning the representation

Equivariance

- Let $f\colon X \rightarrow Y$ be a neural network function
- $\emptyset^x$ and $\emptyset^y$ are the transformations on X and Y, respectively
- f is equivariant iff

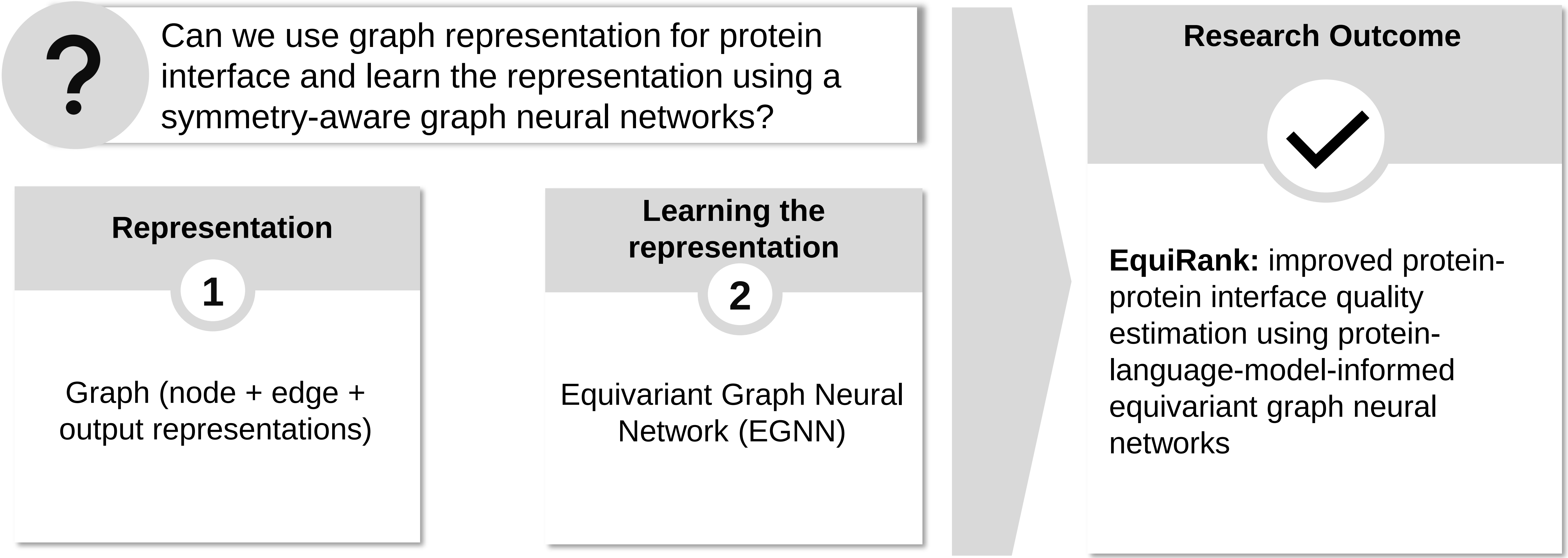
$$f \circ \emptyset_g^x = \emptyset_g^y \circ f$$

$$\begin{tikzcd}
X \arrow[r, "\emptyset_g^x"] \arrow[d, "f"'] & X \arrow[d, "f"] \\
Y \arrow[r, "\emptyset_g^y"'] & Y
\end{tikzcd}$$



Satorras *et al.*, 2022

Research outcome



? Can we use graph representation for protein interface and learn the representation using a symmetry-aware graph neural networks?

Representation

1

Graph (node + edge + output representations)

Learning the representation

2

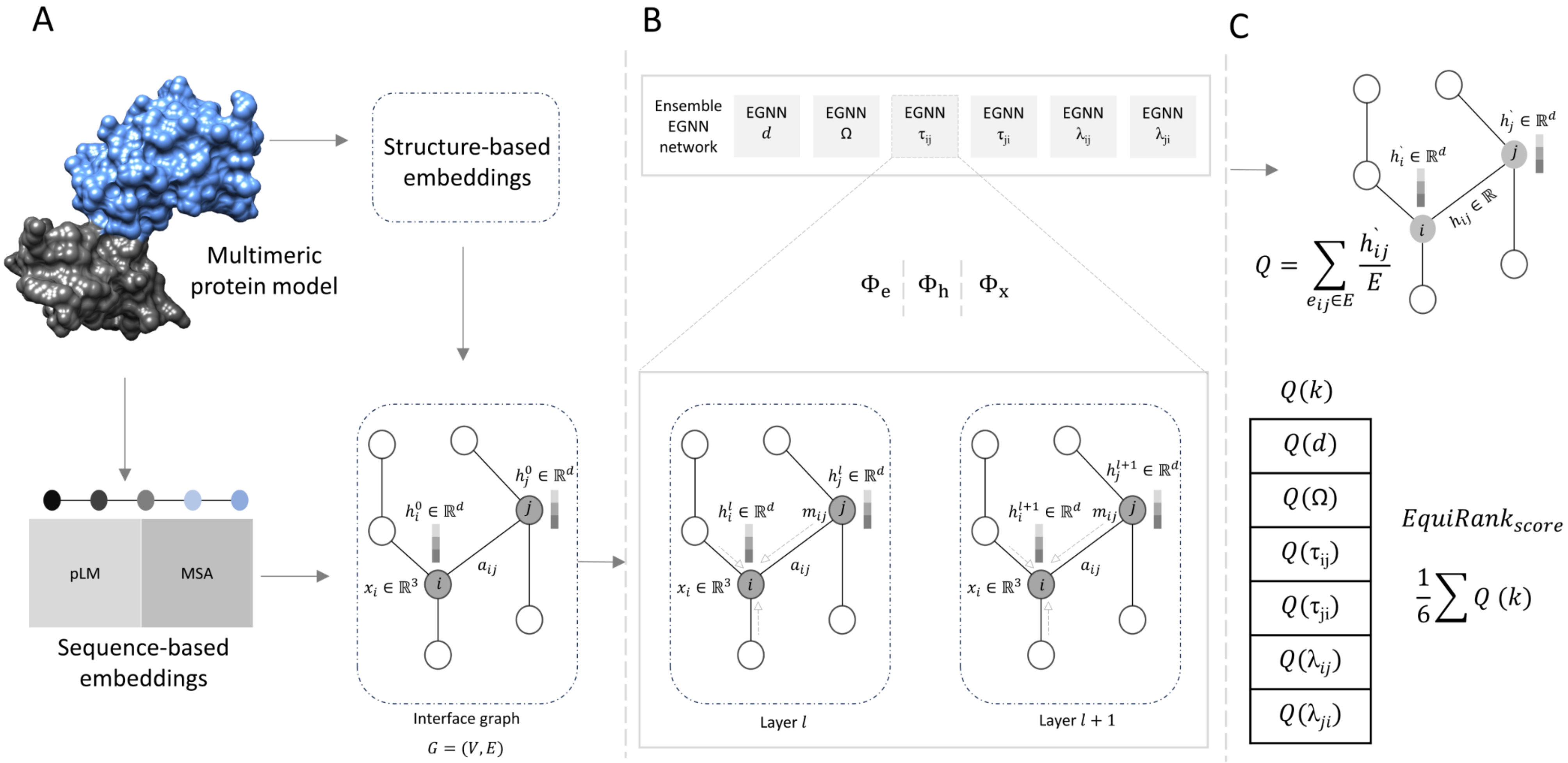
Equivariant Graph Neural Network (EGNN)

Research Outcome

✓

EquiRank: improved protein-protein interface quality estimation using protein-language-model-informed equivariant graph neural networks

Flowchart of EquiRank



$EquiRank_{score}$ is the probabilistic combination of estimated multimeric distance and orientation error

Available at: https://github.com/mhshuvo1/EquiRank

Datasets

	Dataset	Num. targets	Num. decoys	Correct (DockQ ≥ 0.23)	Incorrect (DockQ < 0.23)
Training	VoroIF_GNN_train	1,097	14,400	42%	52%
	CASP13	20	2,386	17.24%	82.76%
	CASP14	10	1,329	15.95%	84.05%
Testing	VoroIF_GNN_test	235	2,845	42%	58%
	CASP15	26	6,850	45.37%	54.63%
	Dockground v1	23	2,500	10.72%	89.28%
Validation	VoroIF_GNN_validation	235	2,814	40.96%	59.14%

CASP: Critical Assessment of Protein Structure Prediction

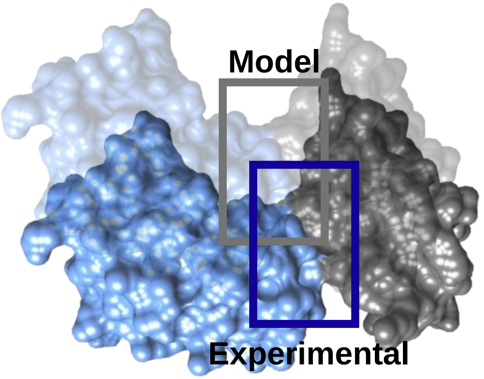
Dockground v1: Dockground version 1

Olechnovič & Venclovas, 2023
Kundrotas et al., 2018

Evaluation metrices

Ground truth

- ➢ DockQ score



- ➢ DockQ = [0, 1]

1 Ability to Rank

- ➢ Reproducibility of ranking w.r.t. ground truth DockQ scores
- ❑ **Spearman Correlation** between the DockQ and predicted interface quality scores.
- ❑ Higher correlation indicates better reproducibility
- ❑ **Top-N Success Rate (N = 1, 5, 10, 15, 20, 25, 30)**

$$SR(N) = \frac{S(N)}{K} \times 100$$

- ❑ **Top-N Hit Rate (N = 1, 5, 10, 15, 20, 25, 30)**

$$HR(N) = \frac{H(N)}{M} \times 100$$

- ❑ Higher Rates indicates better performance

2 Ability to Distinguish

- ➢ High quality protein complex models
- ❑ DockQ cutoff = 0.80
- ❑ Receiver Operating Characteristics Area Under the Curve
- ❑ Higher AUC is better

Basu and Wallner, 2016

Competing methods

- **Graph Neural Network-based methods**
    1. PIQLE
    2. EuDockScore
    3. VoroIF_GNN
    4. DProQA
    5. GDockScore
    6. DeepRank-GNN-esm
    7. GNN-DOVE
- **Transformer-based method**
    8. AlphaFold-Multimer
- **Convolutional Neural Network-based methods**
    9. TRScore
    10. DOVE_ATOM20
    11. DOVE_ATOM40
    12. DOVE_GOAP
    13. DOVE_ATOM_GOAP

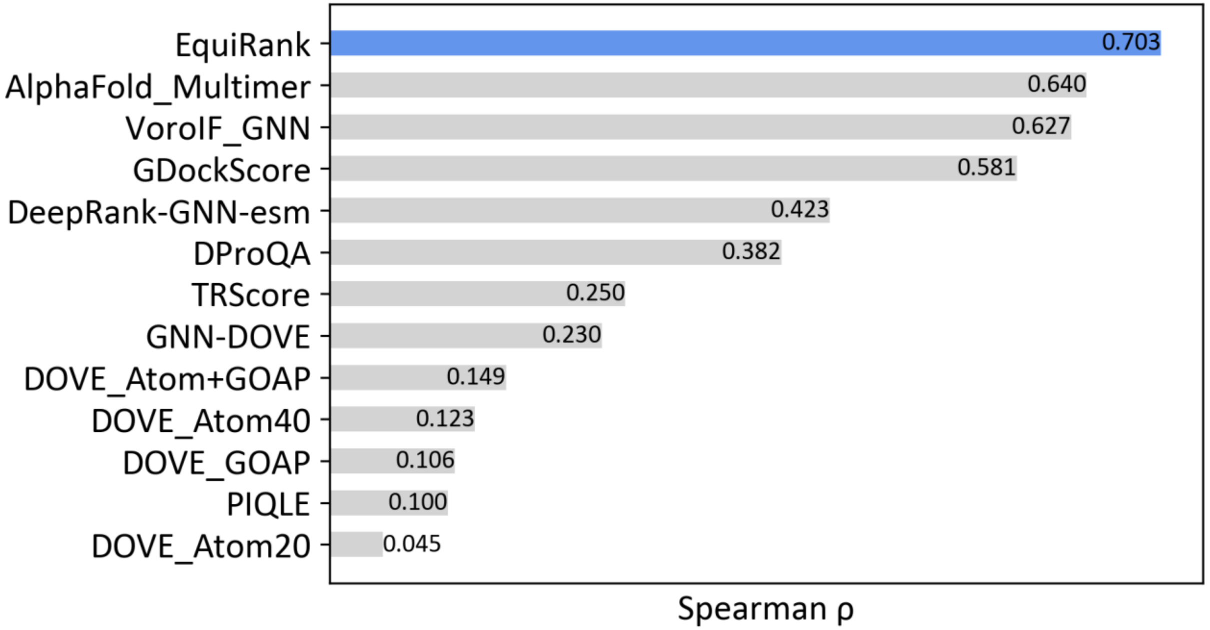
Matthew et al., 2024
Shuvo *et al.*, 2023
Olechnovič and Venclovas, 2023
Chen *et al.*, 2023
McFee and Kim, 2023
Xu and Bonvin, 2023
Wang *et al.*, 2021
Evans *et al.*, 2022
Guo *et al.*, 2022
Wang *et al.*, 2020

Results

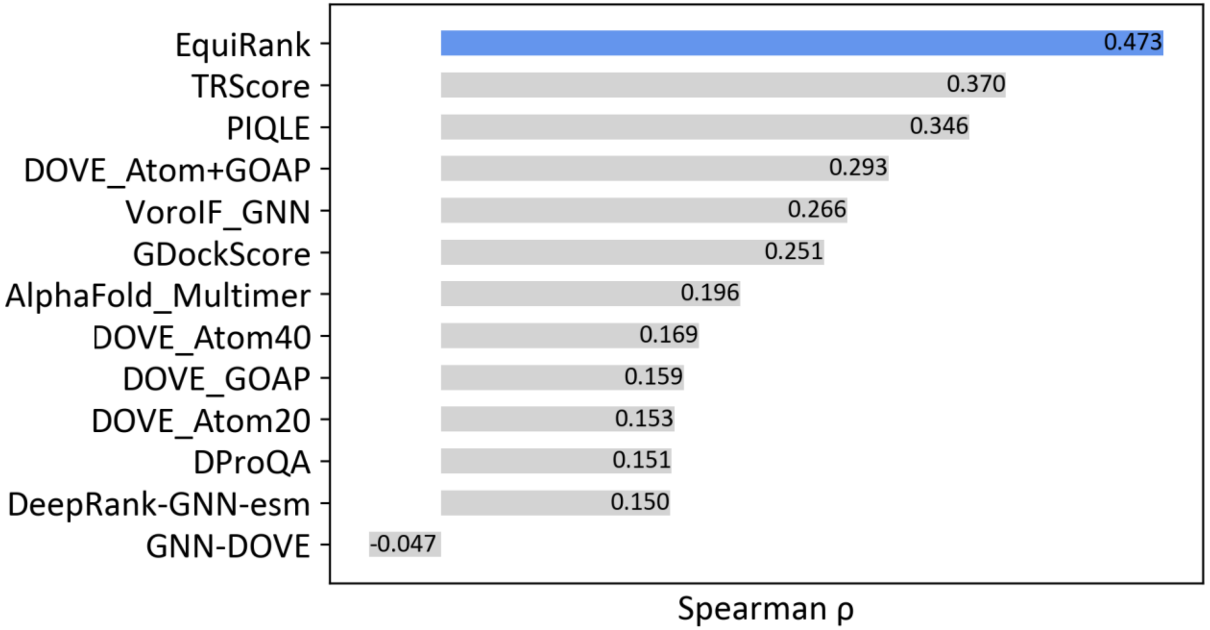
Ability to Rank: Reproducibility of DockQ ranking

Spearman correlation coefficient between predicted and DockQ scores

VoroIF_GNN_test → total decoys: 2,845 Correct: 42% Incorrect: 58%)



Dockground v1→ total decoys: 2,500 Correct: 10.72% Incorrect: 89.28%

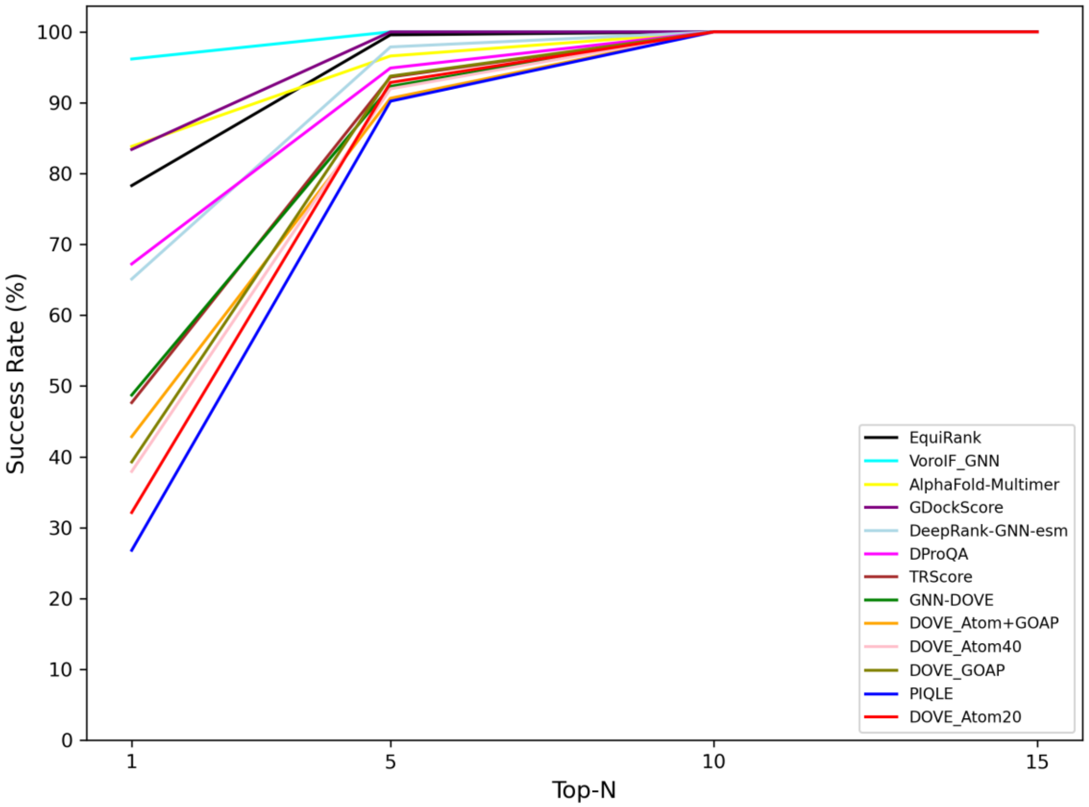


EquiRank is better in reproducing ground truth ranking

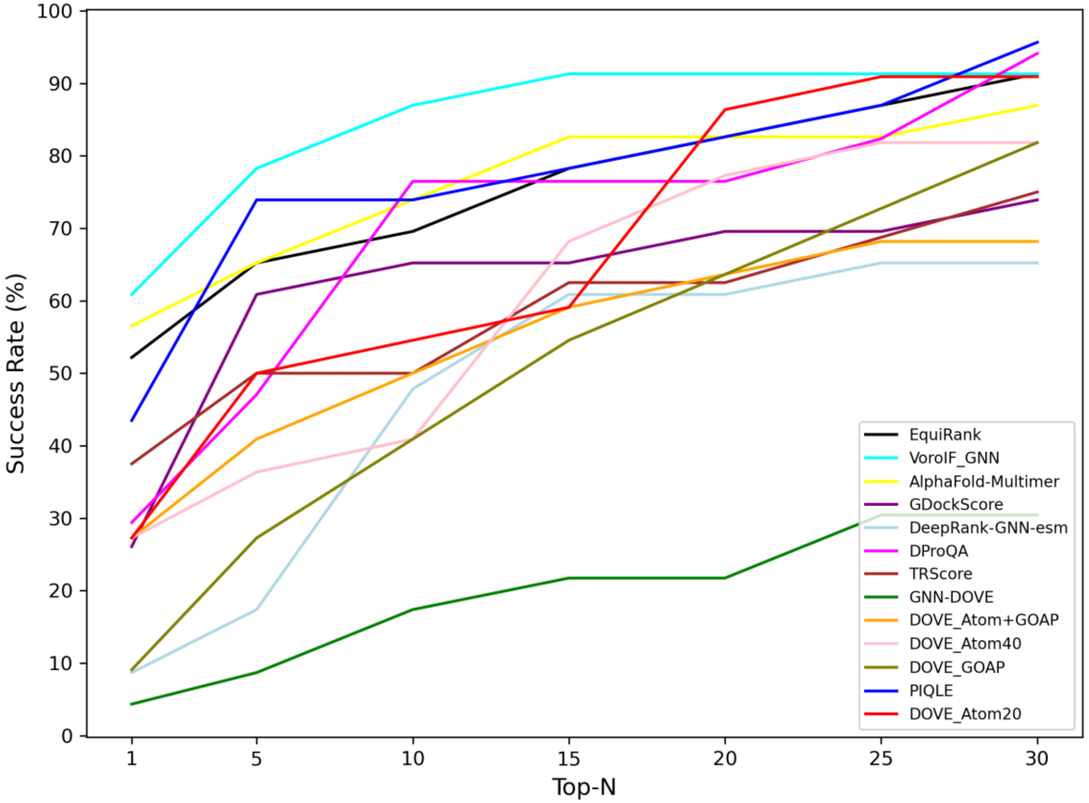
Ability to Rank: Top-N Success Rate

Percentage of targets with at least one acceptable model among top-N ranked models

VoroIF_GNN_test → total decoys: 2,845 Correct: 42% Incorrect: 58%)



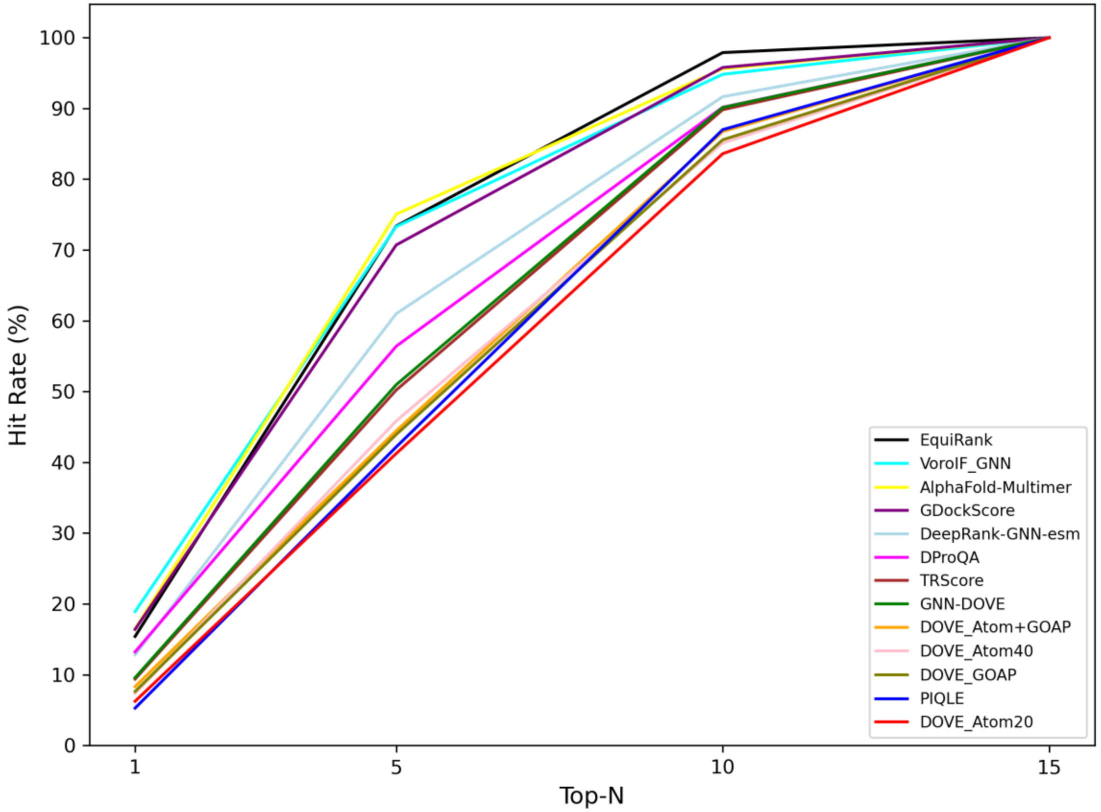
Dockground v1→ total decoys: 2,500 Correct: 10.72% Incorrect: 89.28%



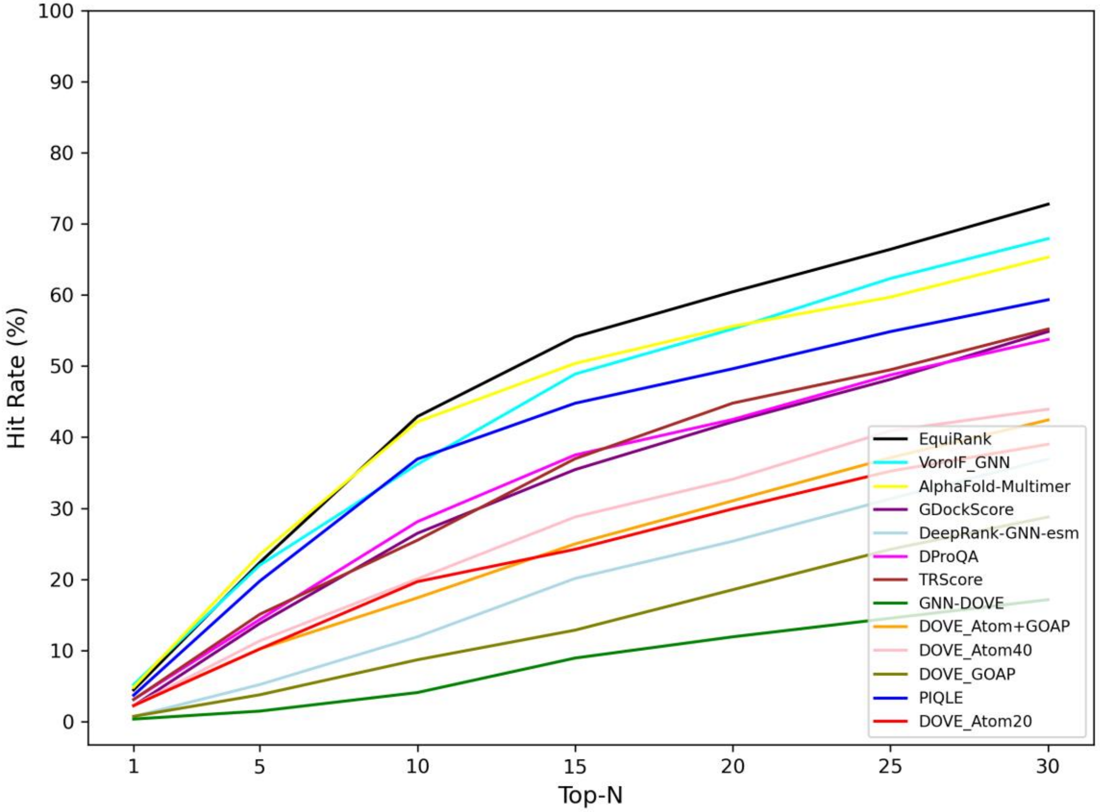
Ability to Rank: Top-N Hit Rate

Fraction of acceptable models among top-ranked models relative to all acceptable models in the dataset

VoroIF_GNN_test → total decoys: 2,845 Correct: 42% Incorrect: 58%)



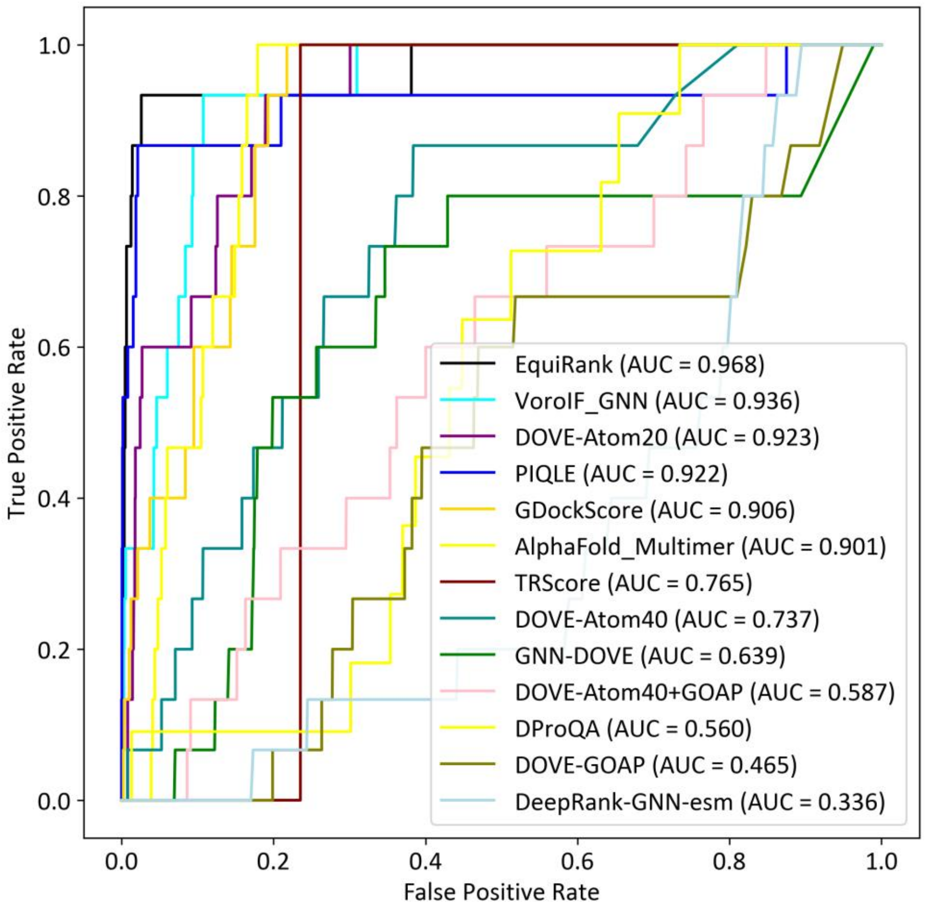
Dockground v1→ total decoys: 2,500 Correct: 10.72% Incorrect: 89.28%



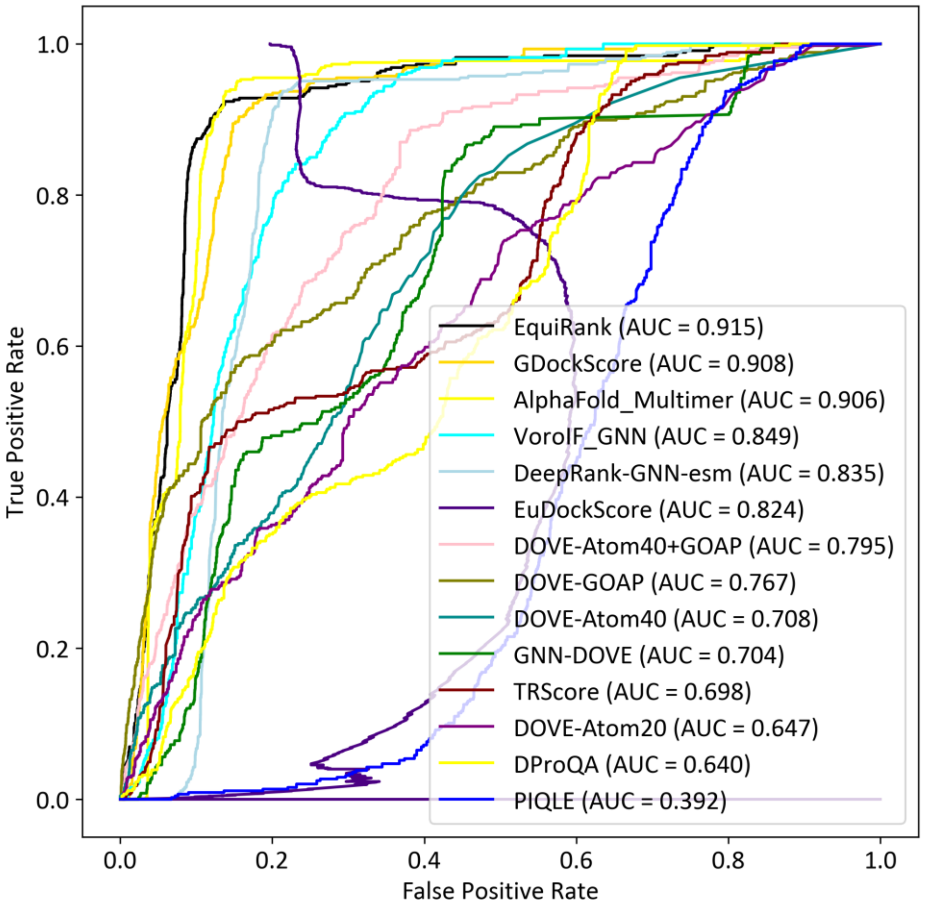
Ability to Distinguish: Distinguishing high-quality complex models

Area Under the ROC curve with DockQ threshold = 0.80

Dockground v1

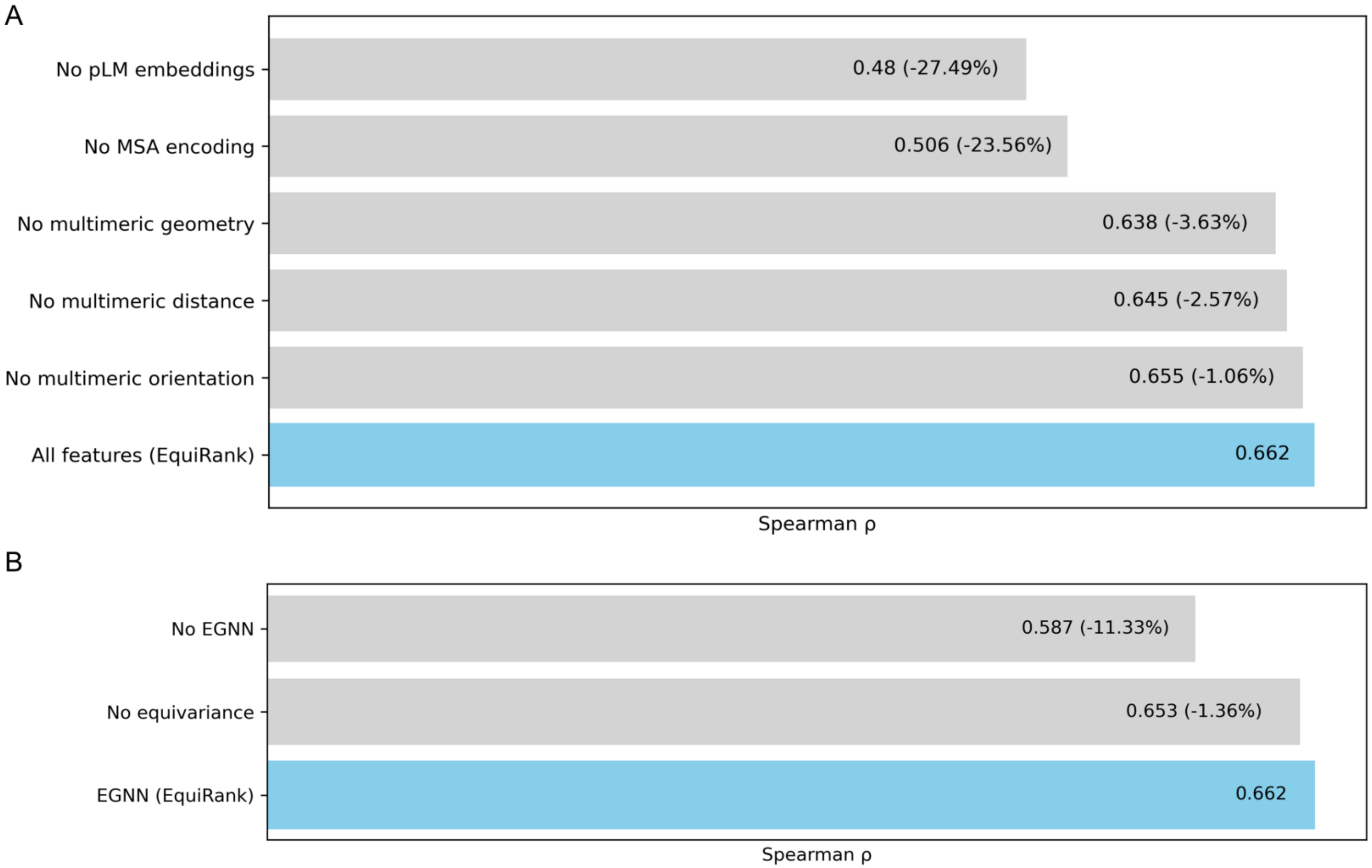


CASP15



EquiRank is better in distinguishing high-quality protein complex models

Ablation studies (on VoroIF_GNN_validation)



Protein Language Model embeddings and EGNN contributes to the improved model quality estimation performance

Conclusion and future works

- Application of an Ensemble 6 Equivariant Graph Neural Networks
- EquiRank is better than other competing methods in terms of reproducibility and distinguishability
- EquiRank demonstrates consistent performance on datasets having diverse quality
- Protein-language-model-informed equivariant neural network contributes to improved performance

In the future…

- Improve the generalizability of the multimeric quality estimation method
    - Hyperparameter optimization of the underlying model
- Improve the reproducibility of ground truth scores
- Application of multimeric model quality estimation to improve the predicted multimeric protein complex models
- Development of an integrated framework for multimeric protein model quality estimation

Thank You!

Toward reliability evaluation of computational models of protein molecules and their interactions

https://github.com/mhshuvo1/EquiRank

Md Hossain Shuvo, Ph.D.

mhshuvo@pvamu.edu