Estimation of interfacial quality of protein complex models

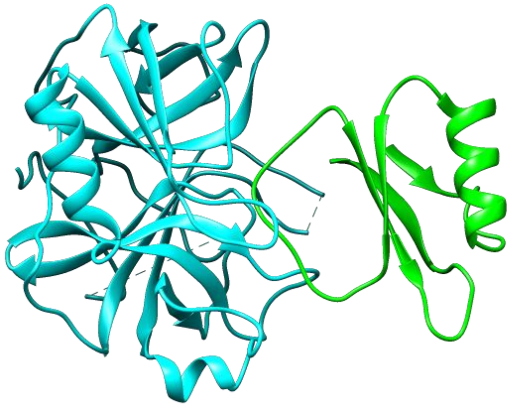
Md Hossain Shuvo

Ph.D. student

Virginia Tech

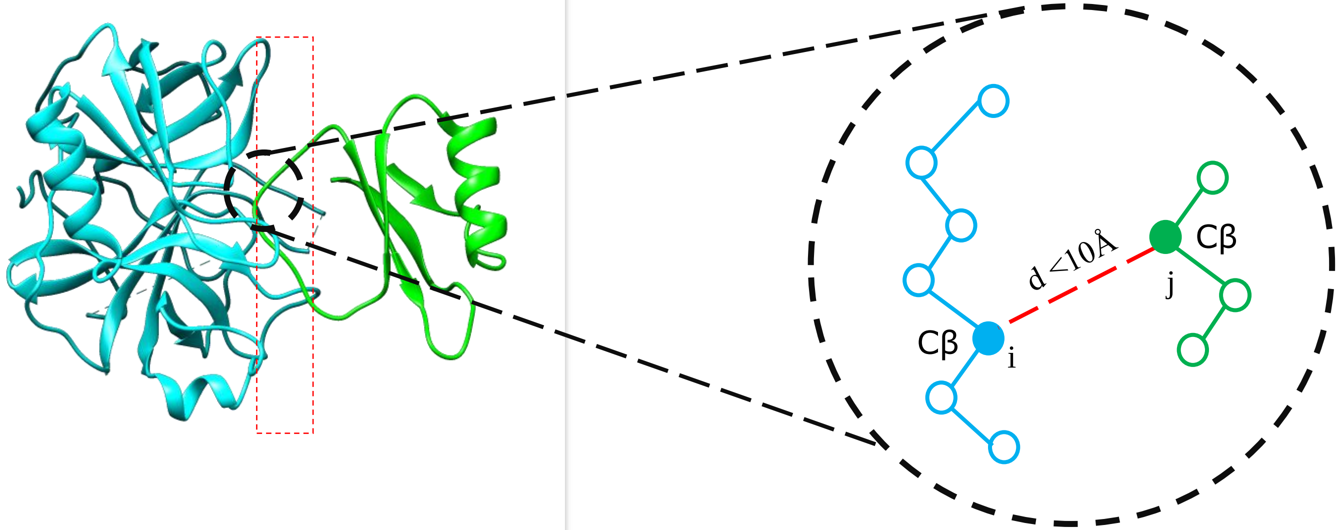
Background

Protein Complex



Crystal structure of 1ACB

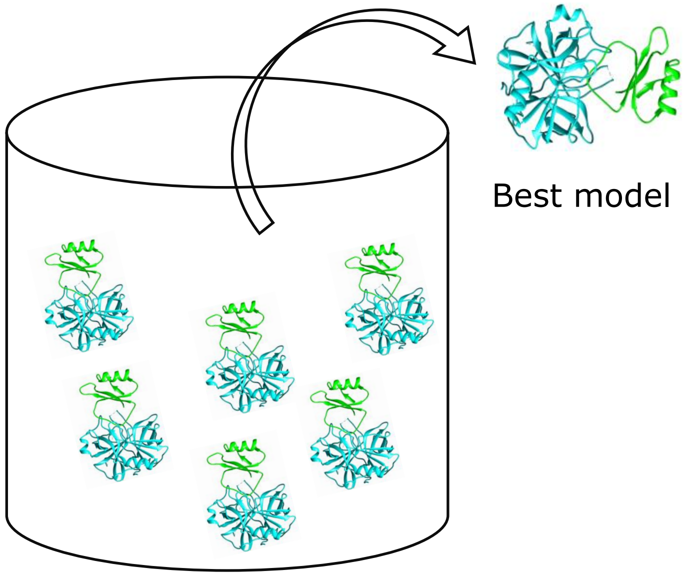
Interfacial region



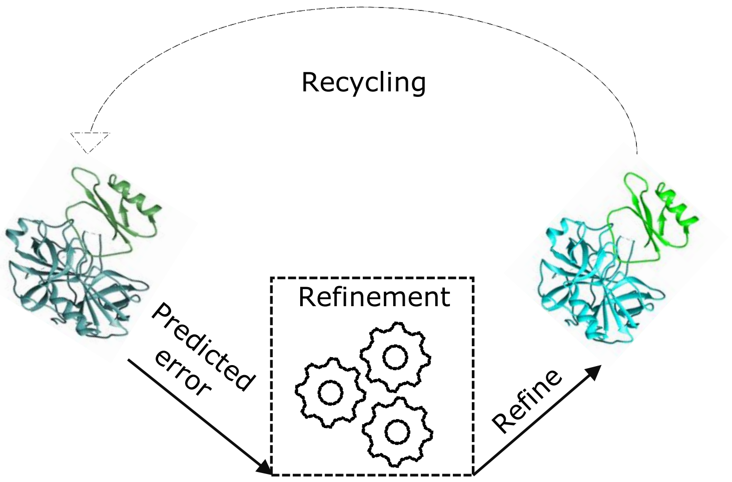
https://www.rcsb.org/structure/1ACB

Motivation

Model section



Model refinement



Helps in accurately guiding the process of protein complex prediction

Approach

- Dataset curation
- Feature extraction
- Model training
- Quality estimation

Dataset

Training

- Dockground docking decoy set v2
- 180 complex targets
- ~18000 docking decoys

Testing

- Dockground docking decoy set v1
- 23 complex targets
- ~2600 docking decoys

Feature extraction

- Node features (30)
  - Amino acids encoding (10)
  - Secondary structure (6)
  - solvent accessibility encoding (4)
  - Relative residue positioning (2)
  - MSA-based features (NEFF) (4)
  - Dihedral angles (4)
- Edge features (23)
  - Orientations between connecting nodes (theta, omega, phi) (6)
  - Edge distance encoding from 2 – 10 Å (17)

Learning algorithm

- Graph neural network
- Ideal for learning for graph representation
- Regression problem

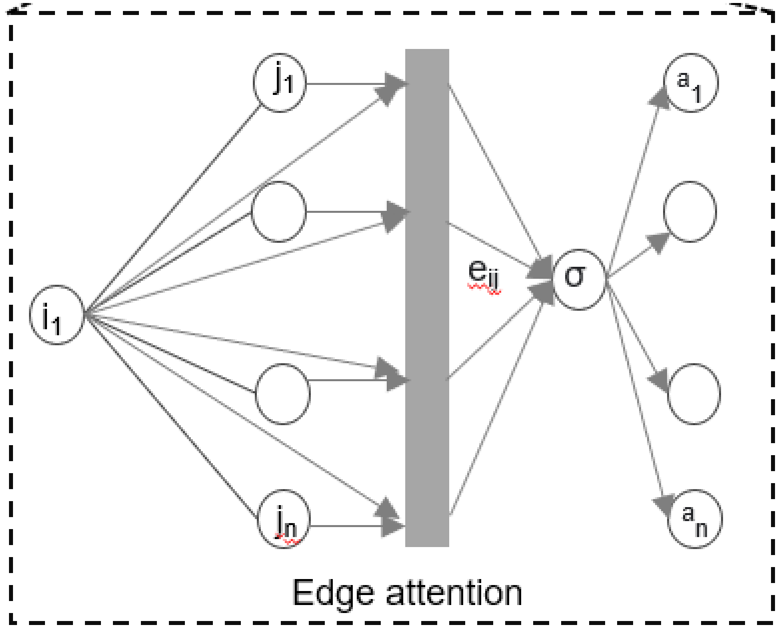
Graph attention network

GCN embedding

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)}\right)$$

$$c_{ij} = \sqrt{|\mathcal{N}(i)|}\sqrt{|\mathcal{N}(j)|}$$

GAT embedding



$$z_i^{(l)} = W^{(l)} h_i^{(l)}, \tag{1}$$

$$e_{ij}^{(l)} = \mathrm{LeakyReLU}(\vec{a}^{(l)^T}(z_i^{(l)} \| z_j^{(l)})), \tag{2}$$

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}(i)} \exp(e_{ik}^{(l)})}, \tag{3}$$

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} z_j^{(l)}\right), \tag{4}$$

Veličković,P. *et al., arXiv* (2018)

Multi-head attention

$$\text{concatenation}: h_i^{(l+1)} = ||_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k W^k h_j^{(l)} \right)$$

$$\text{average}: h_i^{(l+1)} = \sigma \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k W^k h_j^{(l)} \right)$$

Quality estimation

Target label

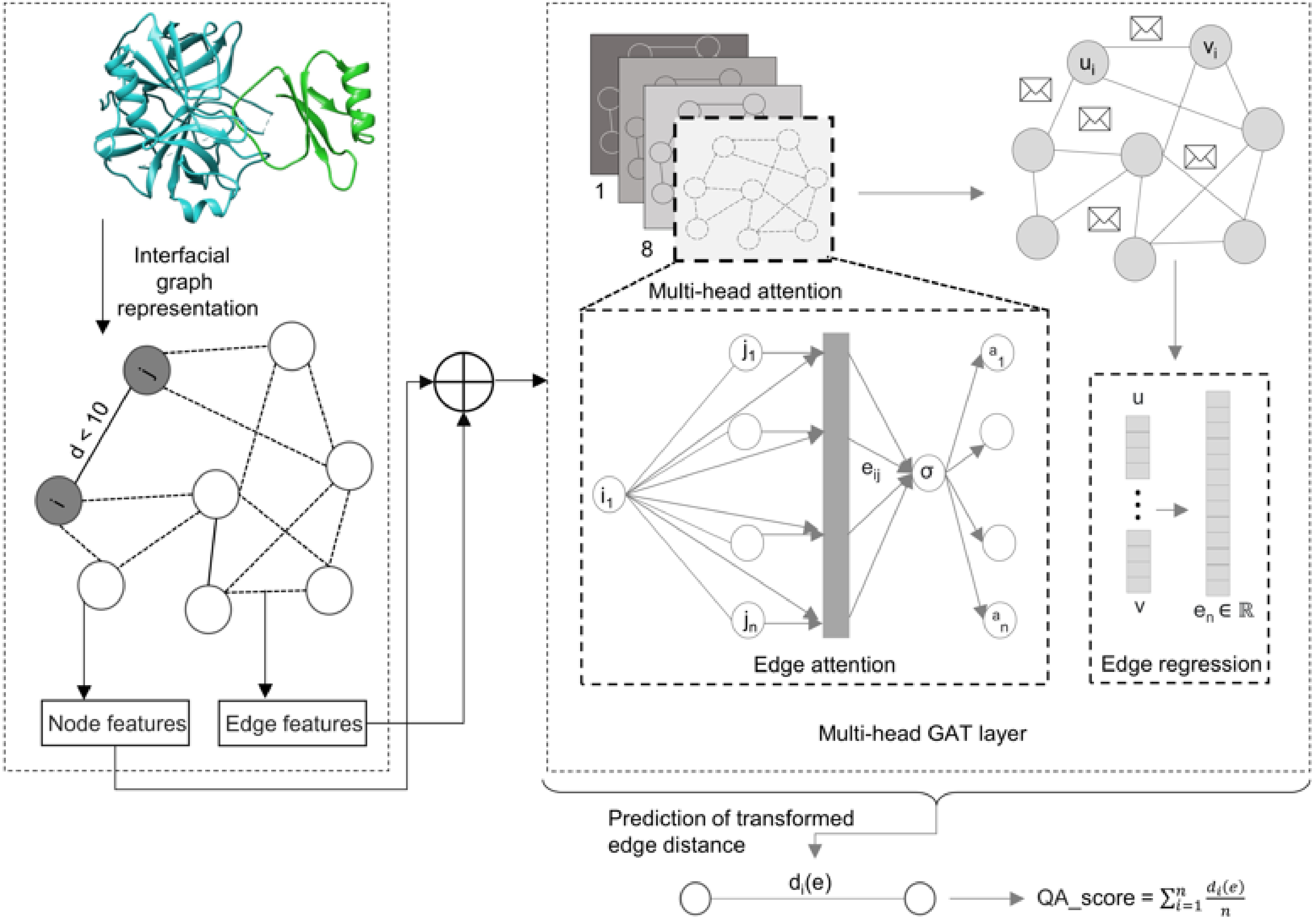
- For each edge (local quality)
- d_i = 10

$$s\_score = \frac{1}{1 + \left(\frac{d}{d_i}\right)^2}$$

- Global quality

$$global_{quality} = \frac{\sum_1^e s\_score_e}{n}$$

Flowchart



Interfacial
graph
representation
d < 10
Node features
Edge features
1
8
Multi-head attention
Edge attention
Edge regression
Multi-head GAT layer
Prediction of transformed
edge distance
QA_score = $\sum_{i=1}^{n} \frac{d_i(e)}{n}$

Model training

- Number of multi-headed GAT layers: 2
- Number of heads: 8
- Hidden dimension: 32
- Learning rate: 0.001
- Weight decay: 0.0005
- Loss: Mean Squared Error (MSE) with sum reduction
- Optimizer: Adam
- Number of batch: ~80
- Number of epochs: 500
- Patience: 40

Evaluation metrices

- Ground truth:
  - Observed s-score w.r.t iRMSD

$$s\_score = \frac{1}{1 + \left(\frac{d}{d_i}\right)^2}$$

- Pearson correlation between $\text{global}_{\text{quality}}$ and the s-score
- Spearman correlation between $\text{global}_{\text{quality}}$ and the s-score
- Kendall’s Tau correlation between $\text{global}_{\text{quality}}$ and the s-score

Competing methods

- DOVE_ATOM20
- DOVE_ATOM40
- DOVE_GOAP
- DOVE_ATOM_GOAP

Results

Dataset	Method	Avg. r	Avg. ρ	Avg. τ	Global r	Global ρ	Global τ
Dockground v1	**This work**	**0.441**	**0.314**	**0.224**	**0.531**	**0.593**	**0.421**
	DOVE_ATOM20	0.195	0.130	0.089	0.360	0.274	0.185
	DOVE_ATOM40	0.181	0.157	0.111	0.244	0.130	0.087
	DOVE_GOAP	0.084	0.140	0.094	-0.059	-0.085	-0.056
	DOVE_ATOM_GOAP	0.263	0.258	0.180	0.227	0.101	0.067

Contribution of GAT

Dataset	Method	Avg. r	Avg. ρ	Avg. τ	Global r	Global ρ	Global τ
Dockground v1	**GAT (This work)**	**0.441**	**0.314**	**0.224**	**0.531**	**0.593**	**0.421**
	GCN	0.284	0.223	0.156	0.412	0.451	0.311

Discussion and future plan

- Variable length graph
- Global and local quality

- Hyperparameter tuning
- Additional similar network
- Additional dataset
- Competing methods
- Additional accuracy metrics and case study

Challenges

- Variable length graph
- Regression problem

Reviewers' comments

- "It is representing only the interfacial region as a graph. But in decoys, there will be some orientations, where interface regions would be completely different compared to that of the corresponding native. I am wondering, if considering the interfacial region would cause some form of information loss. Therefore, considering the whole complex as a graph could provide more information during the learning process." (Computationally demanding, Pre-trained model, learning method, QA)
- "A visualization of the problem/dataset would be helpful to show the reader what exactly you'll be focusing on within the dataset." (Interfacial region, case study)
- "Can some node features be directly extracted from the interface coordinates themselves?" (Edge features, agreement)

Acknowledgement

Debswapna Bhattacharya, Ph.D.
Associate Professor
Virginia Tech

References

1. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. Int J Proteomics 2014;2014:147648.
2. Kundrotas PJ, Anishchenko I, Dauzhenka T, Kotthoff I, Mnevets D, Copeland MM, Vakser IA. Dockground: A comprehensive data resource for modeling of protein complexes. Protein Sci 2018;27(1):172–181.
3. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The Graph Neural Network Model. IEEE Transactions on Neural Networks 2009;20(1):61–80.
4. Lensink MF, Velankar S, Baek M, Heo L, Seok C, Wodak SJ. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. Proteins 2018;86 Suppl 1:257–273.
5. Wang X, Terashi G, Christoffer CW, Zhu M, Kihara D. Protein docking model evaluation by 3D deep convolutional neural networks. Bioinformatics 2020;36(7):2113–2118.
6. Wang X, Flannery ST, Kihara D. Protein Docking Model Evaluation by Graph Neural Networks. Frontiers in Molecular Biosciences 2021;8.
7. Renaud N, Geng C, Georgievska S, Ambrosetti F, Ridder L, Marzella DF, Réau MF, Bonvin AMJJ, Xue LC. DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. Nat Commun 2021;12(1):7068.