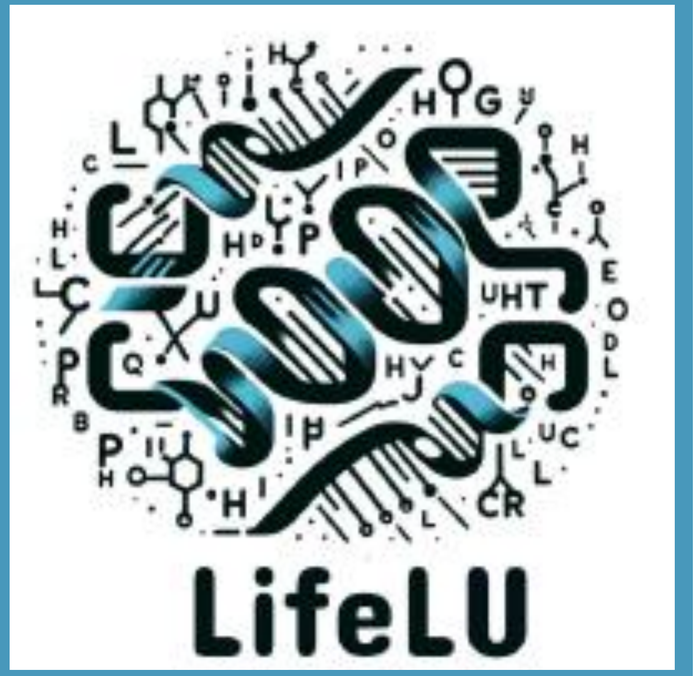
LifeLU - UNDERSTANDING THE LANGUAGE OF LIFE

READING GROUP



Presenter: Özlem Şimşek

Paper: HEAL: Hierarchical graph transformer with contrastive learning for protein function prediction

Link: https://academic.oup.com/bioinformatics/article/39/7/btad410/7208864

Date: 05/12/2024

Hierarchical graph transformer with contrastive learning for protein function prediction

Gu, Z., Luo, X., Chen, J., Deng, M., & Lai, L. (2023). Hierarchical graph transformer with contrastive learning for protein function prediction. Bioinformatics, 39.

Bioinformatics, 2023, **39(7)**, btad410
https://doi.org/10.1093/bioinformatics/btad410
Advance access publication 27 June 2023
Original Paper

OXFORD

Structural bioinformatics

Hierarchical graph transformer with contrastive learning for protein function prediction

Zhonghui Gu [1,†], Xiao Luo[2,†], Jiaxiao Chen[3], Minghua Deng[3,4,*], Luhua Lai [1,3,5,*]

[1]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90024, United States
[3]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[4]School of Mathematics Sciences, Peking University, Beijing 100871, China
[5]BNLMS, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

*Corresponding authors. Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China. E-mail: dengmh@math.pku.edu.cn (M.D.); Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China. E-mail: lhlai@pku.edu.cn (L.L.)

†Equal contribution.

Associate Editor: Peter Robinson

Abstract

Motivation: In recent years, high-throughput sequencing technologies have made large-scale protein sequences accessible. However, their functional annotations usually rely on low-throughput and pricey experimental studies. Computational prediction models offer a promising alternative to accelerate this process. Graph neural networks have shown significant progress in protein research, but capturing long-distance structural correlations and identifying key residues in protein graphs remains challenging.

Results: In the present study, we propose a novel deep learning model named Hierarchical graph transformEr with contrAstive Learning (HEAL) for protein function prediction. The core feature of HEAL is its ability to capture structural semantics using a hierarchical graph Transformer, which introduces a range of super-nodes mimicking functional motifs to interact with nodes in the protein graph. These semantic-aware super-node embeddings are then aggregated with varying emphasis to produce a graph representation. To optimize the network, we utilized graph contrastive learning as a regularization technique to maximize the similarity between different views of the graph representation. Evaluation of the PDBch test set shows that HEAL-PDB, trained on fewer data, achieves comparable performance to the recent state-of-the-art methods, such as DeepFRI. Moreover, HEAL, with the added benefit of unresolved protein structures predicted by AlphaFold2, outperforms DeepFRI by a significant margin on Fmax, AUPR, and Smin metrics on PDBch test set. Additionally, when there are no experimentally resolved structures available for the proteins of interest, HEAL can still achieve better performance on AFch test set than DeepFRI and DeepGOPlus by taking advantage of AlphaFold2 predicted structures. Finally, HEAL is capable of finding functional sites through class activation mapping.

Availability and implementation: Implementations of our HEAL can be found at https://github.com/ZhonghuiGu/HEAL.

1 Introduction

Recent development in high-throughput sequencing has resulted in a great increase in the number of protein sequences in benchmark databases such as (Apweiler *et al.* 2004, UniProt Consortium 2019). However, the bulk of protein sequences lack functional annotation owing to the exorbitant expense and low-throughput experimental studies (Radivojac *et al.* 2013, Zhou *et al.* 2019). Therefore, computational approaches that can automatically and precisely deduce protein functions are much wanted. Commonly used methods for inferring functions for a new protein sequence include sequence-alignment that identify similar domains (FunFam) (Das *et al.* 2015) or local alignments (Blast) (Altschul *et al.* 1990, Buchfink *et al.* 2015), to transfer the functions of proteins that have been experimentally confirmed before. With the advance of machine learning, a variety of computational approaches for protein function prediction have been developed (Yang *et al.* 2015, Fa *et al.* 2018, Kulmanov *et al.* 2018, Gelman *et al.* 2021). In the Critical Assessment of Functional Annotation (CAFA), a blind prediction challenge, machine learning methods have demonstrated superior performance compared to traditional sequence alignment-based methods (Radivojac *et al.* 2013). These machine learning methods can be broadly categorized into knowledge-based, sequence-based, and structure-based approaches. Knowledge-based approaches typically incorporate information from external sources such as protein–protein interaction (PPI) networks (Mostafavi *et al.* 2008, Cho *et al.* 2016, You *et al.* 2021). However, the absence of prior knowledge may limit their practical analysis of newly discovered protein sequences (Gligorijević *et al.* 2021). Sequence-based approaches often use primary sequence as well as some other hand-crafted features to predict protein functions (Fa *et al.* 2018, Kulmanov *et al.* 2018, Zhang *et al.* 2019, Cao and Shen 2021, Kulmanov and Hoehndorf 2021, Yao *et al.* 2021, Zhu *et al.* 2022). Additionally, since structural information has a direct connection with protein functions, structure-based methods have become increasingly popular (Gligorijević *et al.* 2021,

Received: January 12, 2023. **Revised:** May 25, 2023. **Editorial Decision:** June 19, 2023. **Accepted:** June 26, 2023

SUMMARY

Aim: predict GO terms from protein structure and sequence

Problem type: multi-label classification

Method: combination of message passing neural network and hierarchical graph transformer

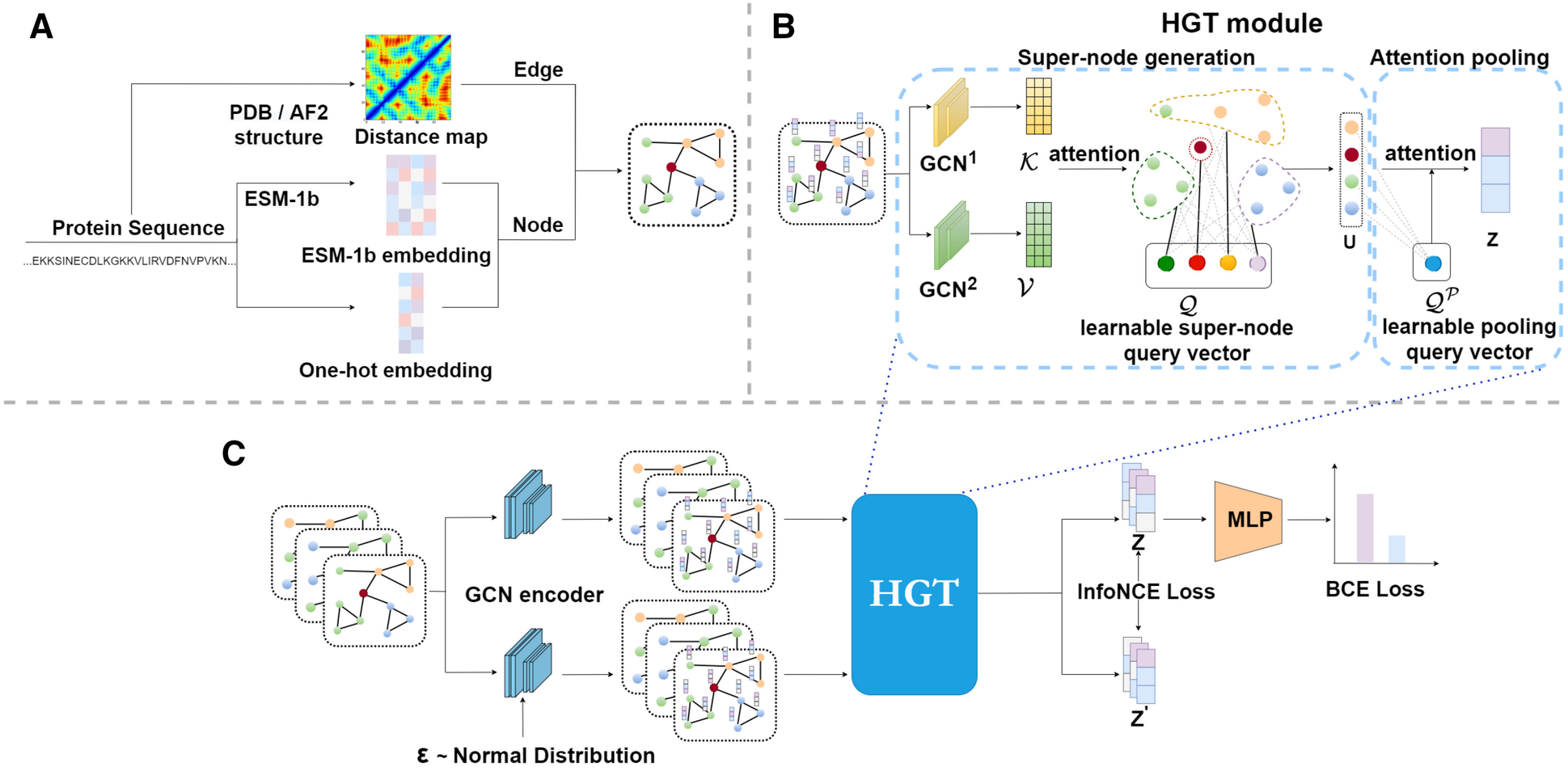
Motivation: high number of protein sequences but lack of functional annotation

RELATED WORK

- sequence-alignment (FunFam)
- local alignments (Blast)

- Machine learning
  - knowledge-based approaches
    - use additional information like PPI networks
  - sequence-based approaches
    - use primary sequence and hand crafted features
  - structure-based methods
    - use protein structure information besides sequence

METHOD


A
PDB / AF2 structure
Distance map
Edge
Protein Sequence
...EKKSINECDLKGKKVLIRVDFNVPVKN...
ESM-1b
ESM-1b embedding
Node
One-hot embedding
B
HGT module
Super-node generation
Attention pooling
GCN1
GCN2
attention
learnable super-node query vector
attention
U
Z
learnable pooling query vector
C
GCN encoder
ε ~ Normal Distribution
HGT
Z
Z'
InfoNCE Loss
MLP
BCE Loss

DATASET

- **dataset of DeepFRI**
  - 36 641 protein structures from PDB database (PDBch)
  - 244 775 protein structures from SWISS-MODEL repository (SMch)

  - PDBch:
    - 95% sequence identity
    - training, validation, and test sets => 8:1:1 ratio
    - The GO-term annotations were retrieved from SIFTS (Dana et al. 2019) and UniProtKB
    - A PDB model needs to
      - share at least 90% sequence identity and
      - cover at least 70% of the UniProtKB sequence to transfer the annotations

DATASET

- SMch:
  - homology models of the PDBch dataset with at least one annotation from the SWISS-MODEL repository
  - 95% sequence identity
  - training and validation => 9:1 ratio

- **dataset with AlphaFold (AFch)**
  - 25% sequence identity
  - training set with 43 072 sequences and test set with 567 sequences
  - random 10% of training set as validation set

METHOD

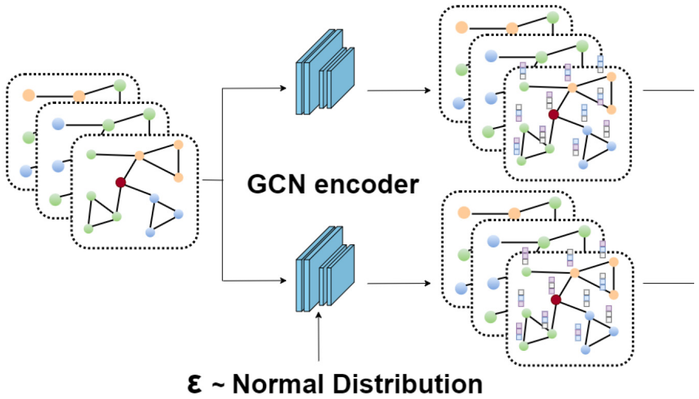
Features:

- Graph input
  - Each protein is represented as a graph input G=(V,E)
    - Each node is an aa
    - Each edge is defined by contact map
  - Node features: ESM-1b embeddings for each aa
  - Edge criteria: There is an edge between two aa, if the distance between their αC atoms is less than 10 A°.

METHOD

Message passing neural network:

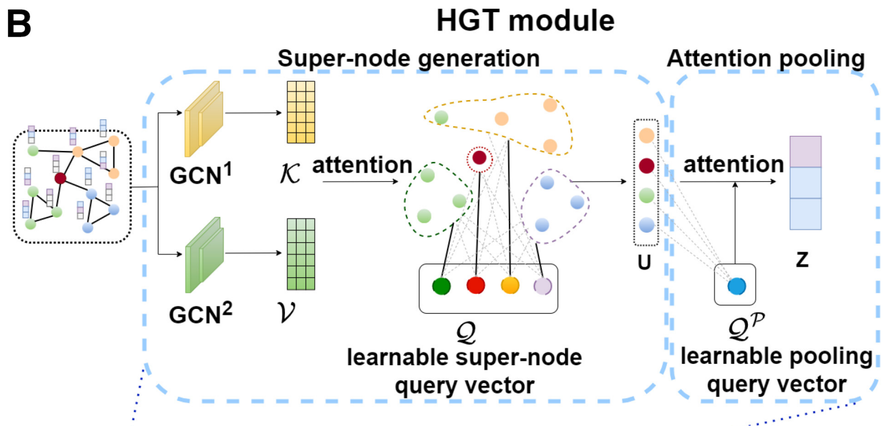
- To collect local information in the protein graph
- GCN encoder



METHOD

Hierarchical graph transformer:

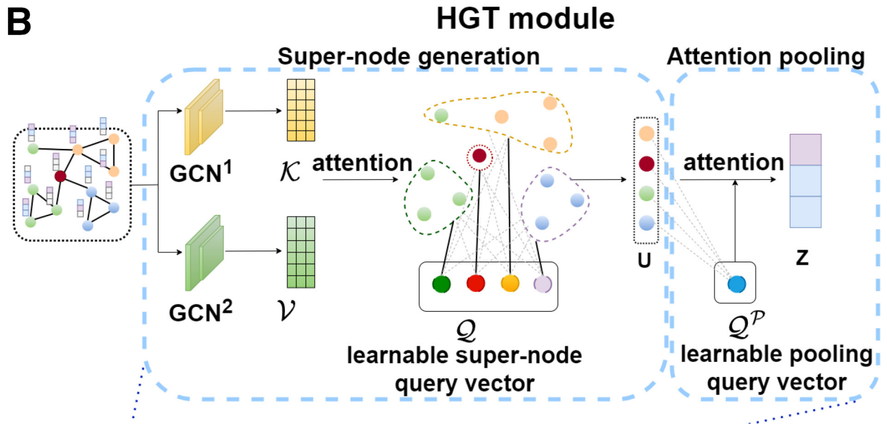
- Instead of global pooling on hidden embeddings from message passing
  - incapable of recognizing important nodes
  - cannot infer long-distance structural relationships
- They introduce hierarchical graph transformer
  - learnable super-nodes to explore long-distance correlations
  - attention module to generate graph-level representations



METHOD

Hierarchical graph transformer:

- Super-node Generation:
  - K super-nodes with learnable features, q1,…,qk
    - interact with nodes in the protein graph for global structure information
  - super-nodes as query vectors in transformers
  - semantics-aware super-node embedding matrix, U



$$\mathbf{\Gamma} = \mathrm{softmax}\left(\frac{\mathcal{Q}\cdot\mathcal{K}^{\top}}{\sqrt{D}}\right)\cdot\mathcal{V}, \tag{2}$$

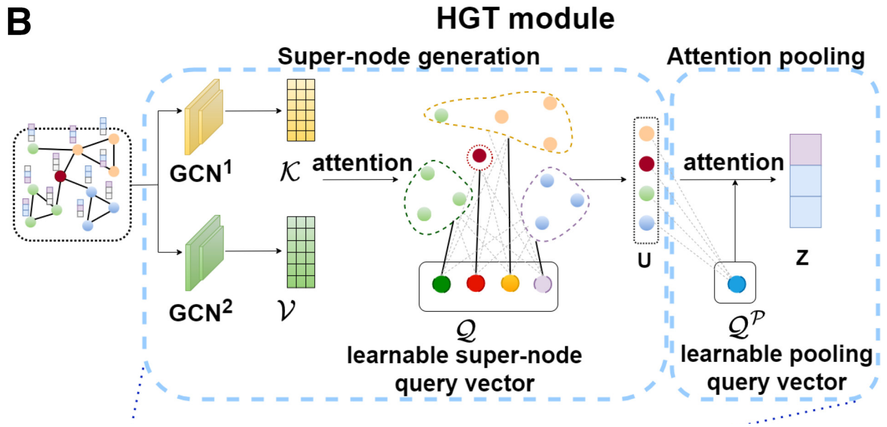
$$\mathcal{K} = \mathrm{GCN}^{1}(H,A), \quad \mathcal{V} = \mathrm{GCN}^{2}(H,A) \tag{3}$$

$$\boldsymbol{U} = FC^{1}([\mathbf{\Gamma}_1,\ldots,\mathbf{\Gamma}_H]) \tag{4}$$

METHOD

Hierarchical graph transformer:

- Attention Pooling:
  - summarizes semantics-aware super-node representations into graph representations



$$z = \mathrm{softmax}\left(\frac{\mathcal{Q}^P \cdot (\boldsymbol{U} \cdot \mathcal{K}^P)^\top}{\sqrt{D}}\right) \cdot \boldsymbol{U} \cdot \mathcal{V}^P, \qquad (5)$$

METHOD

Graph contrastive learning:

- for regularization in the model
- for each graph G, inject noise to every node in the graph to provide a different view
- maximize the similarity between graph representations of different views based on InfoNCE loss

METHOD

Supervised Loss:

- binary cross-entropy (BCE) loss for multilabel classification

METHOD

Summary:

- Training parameters:
  - Adam optimizer
  - learning rate = 0.0001
  - batch size = 64
  - 100 epochs
  - early-stopping = 5 epochs

METHOD

Baseline methods:

- Blast
- FunFam
- DeepGO: sequence based
- DeepFRI: both sequences and structures
- DeepGOPlus

RESULTS

Table 1. AUPR, Fmax, and Smin of different methods on PDBch test set.[a]

Method	Training set	AUPR (↑)			Fmax (↑)			Smin (↓)		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
Blast	–	0.136	0.067	0.097	0.328	0.336	0.448	0.632	0.651	0.628
FunFams	–	0.367	0.260	0.288	0.572	0.500	0.627	0.531	0.579	0.503
DeepGO	PDBch+SMch training set	0.391	0.182	0.263	0.577	0.493	0.594	0.472	0.577	0.550
DeepFRI	PDBch+SMch training set	0.495	0.261	0.274	0.625	0.540	0.613	0.437	0.543	0.527
HEAL-PDB	PDBch training set	0.571	0.259	0.342	0.691	0.565	0.655	0.401	0.540	0.501
HEAL-SW	PDBch+SMch training set	0.653	0.308	0.432	0.711	0.581	0.654	0.366	**0.509**	0.489
HEAL	PDBch+AFch training set	**0.691**	**0.337**	**0.467**	**0.747**	**0.595**	**0.687**	**0.342**	**0.509**	**0.458**

[a] Best performance in bold. Fmax and AUPR, highest; Smin, lowest.

- Performance on PDBch test set
- Compared to DeepFRI, HEAL-PDB:
  - performs significantly better on the MF and CC tasks
  - has comparable results on the BP task,
  - despite being trained on much less data

RESULTS

Table 2. Ablation study of HEAL on PDBch test set.[a]

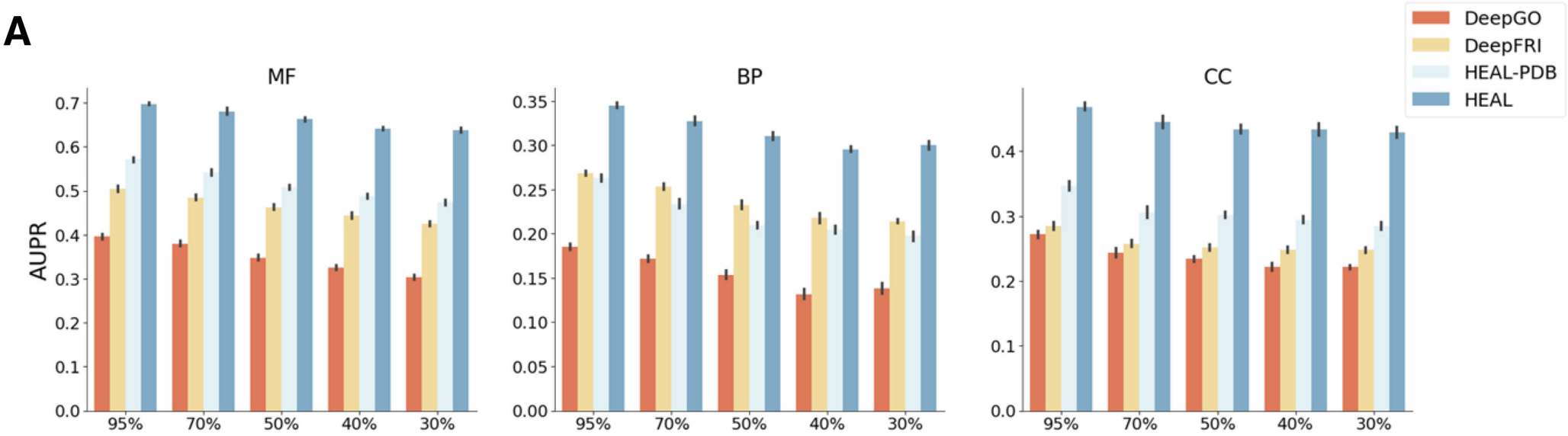
Method	AUPR (↑)			Fmax (↑)			Smin (↓)		
	MF	BP	CC	MF	BP	CC	MF	BP	CC
HEAL	**0.691**	**0.337**	**0.467**	**0.747**	**0.595**	**0.687**	**0.342**	**0.509**	**0.458**
HEAL w/o CL	0.635	0.304	0.410	0.708	0.586	0.672	0.375	0.521	0.478
HEAL w/o MP	0.588	0.252	0.378	0.666	0.552	0.665	0.416	0.547	0.486
HEAL w/o EE	0.284	0.130	0.222	0.478	0.447	0.579	0.554	0.607	0.553

[a] Best performance in bold. Fmax and AUPR, highest; Smin, lowest. The three variants of HEAL are: (i) HEAL w/o CL (contrastive learning): it removes the contrastive learning objective. (ii) HEAL w MP (max pooling): it utilizes the max pooling to replace HGT (iii) HEAL w/o EE (ESM-1b embeddings): it removes ESM-1b embeddings from the node attributes.

- Ablation study on how different components of HEAL contribute to its performance using PDBch test set

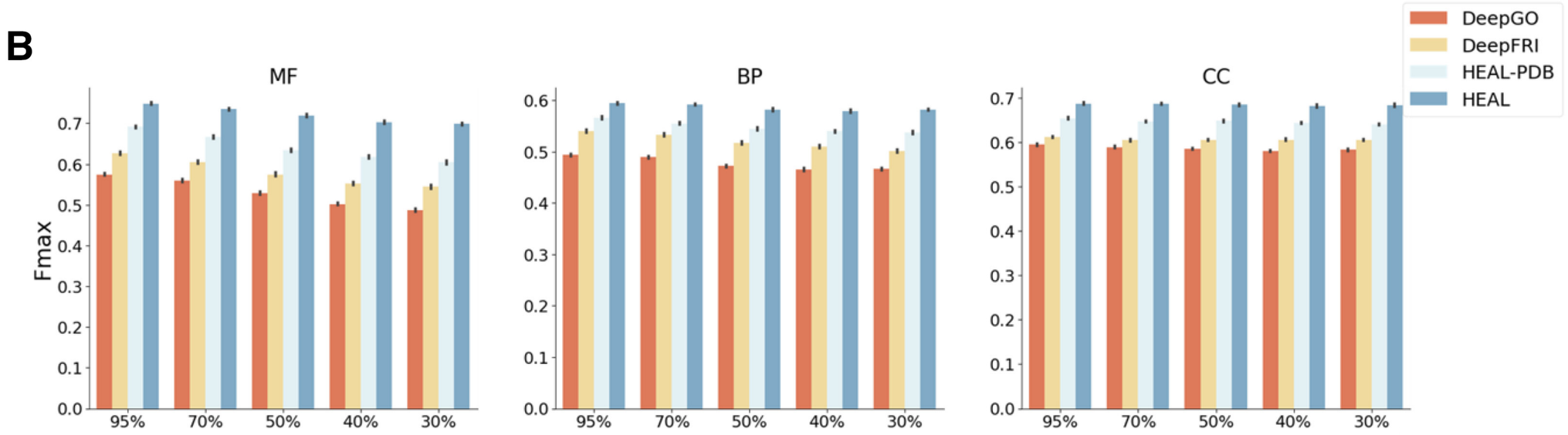
RESULTS

A



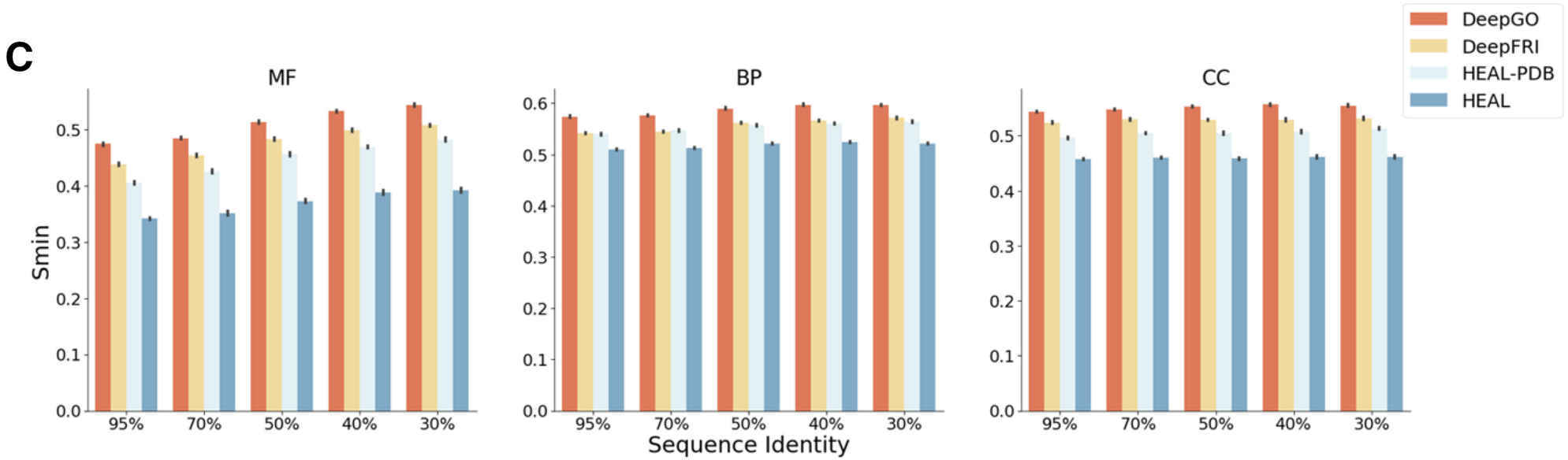
- Evaluate the generalizability of HEAL
- AUPR for different sequence identity thresholds
- DeepGO (sequence based) has min AUPR values
- HEAL-PDB > DeepFRI for MF and CC but DeepFRI > HEAL-PDB for BP
- HEAL > DeepFRI for all

RESULTS



- Evaluate the generalizability of HEAL
- Fmax for different sequence identity thresholds

RESULTS



- Evaluate the generalizability of HEAL
- Smin for different sequence identity thresholds

RESULTS

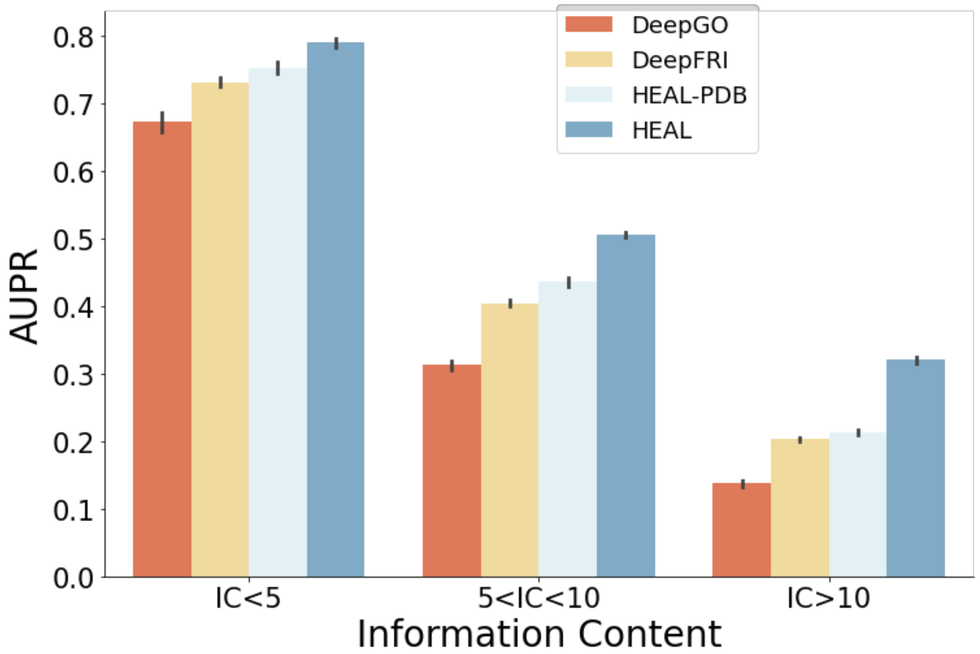


Fig. S3. AUPR of different methods on PDBch test set over different IC (information content) thresholds.

- AUPR on different IC thresholds
- HEAL-PDB outperforms DeepGO and achieves similar results with DeepFRI

RESULTS

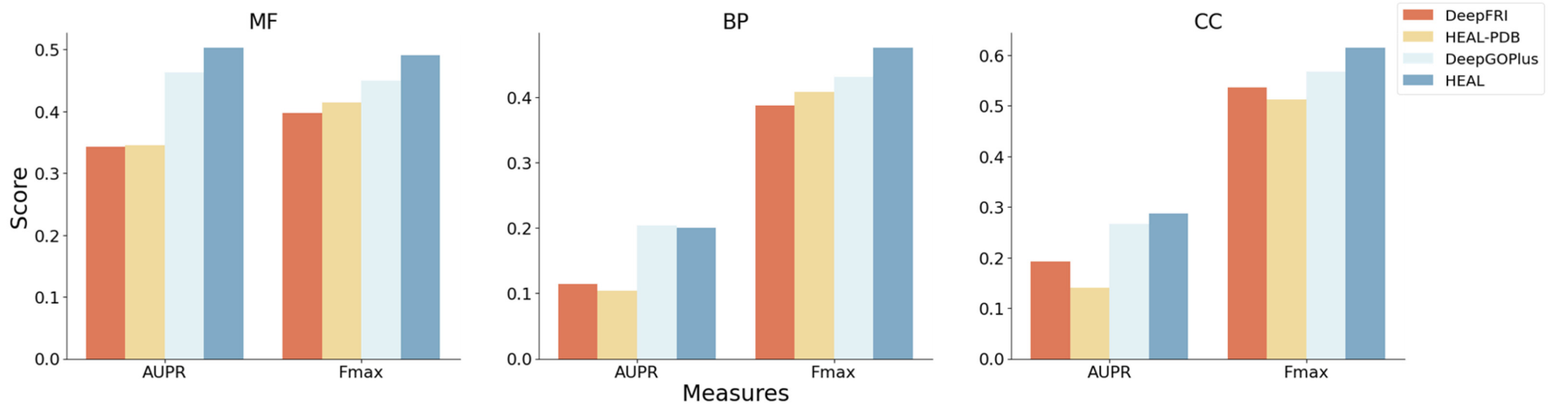
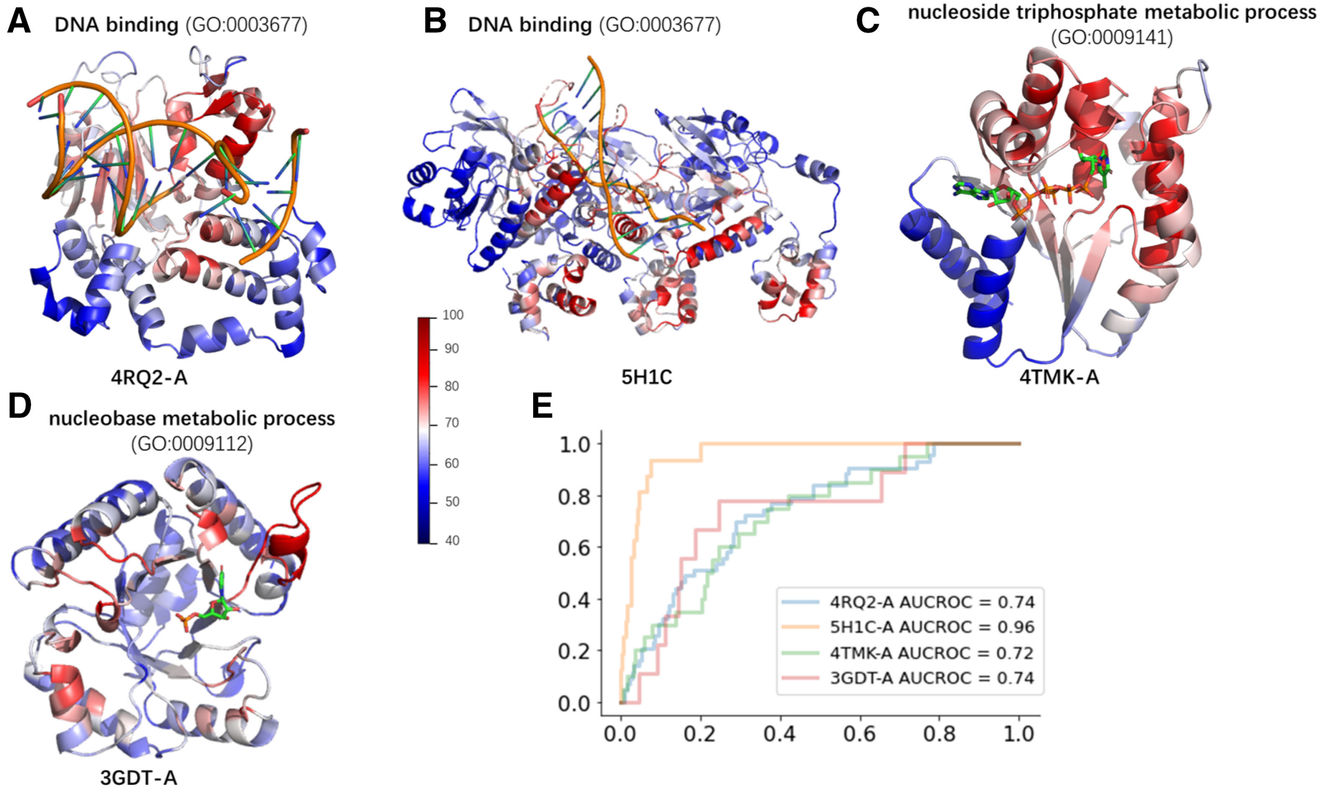


Figure 3. AUPR and Fmax of different methods on AFch test set.

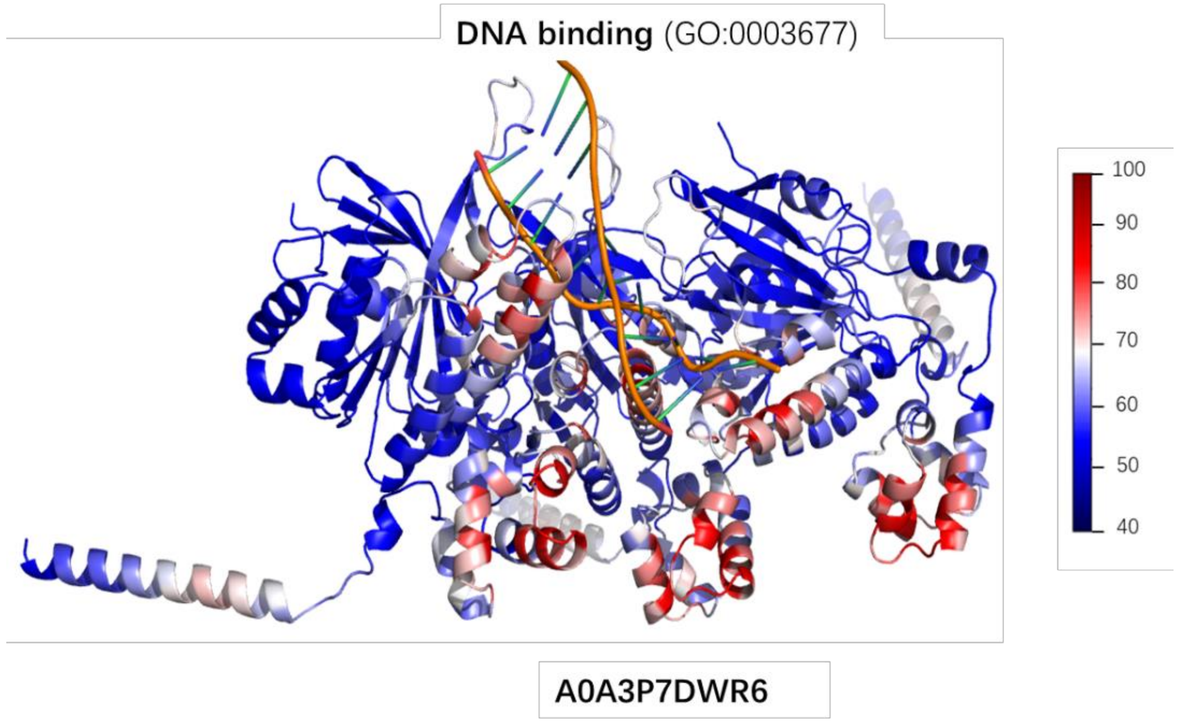
- Task: predict biological functions for proteins with neither experimentally resolved structure nor annotated similar sequences
- Experiment on AFch test set

RESULTS



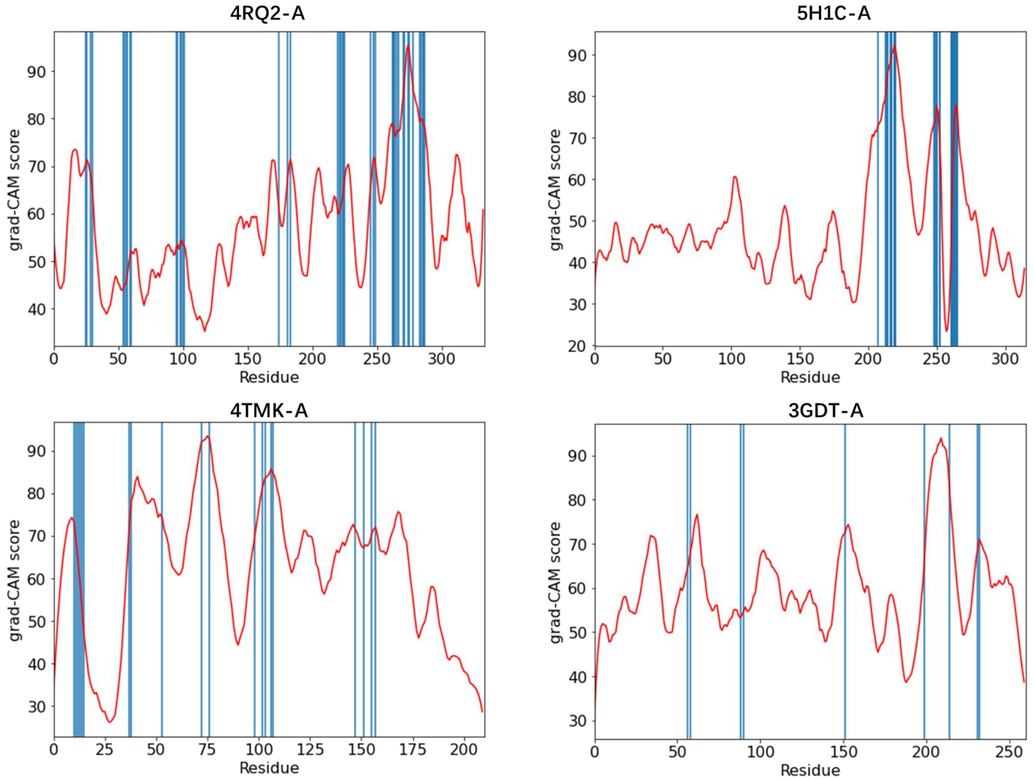
- gradient-weighted Class Activation Map (grad-CAM)
- Projected the heat-map onto the protein structure
- observed strong signals in regions where DNA binds for A and B
- observed strong signals on residues around the inhibitor for C
- observed strong signals on UP6-binding site for D

RESULTS



- gradient-weighted Class Activation Map (grad-CAM)
- Projected the heat-map onto the protein structure
- identify the core binding sites
- A0A3P7DWR6 doesn’t have structure information in PDB

RESULTS



- gradient-weighted Class Activation Map (grad-CAM)
- Red line : grad-CAM score
- Blue line : binding sites stored by the BioLiP database

CONCLUSION

- HEAL predicts protein functions based on protein structures and sequences.
- HEAL uses:
  - hierarchical graph Transformer
  - contrastive learning
- HEAL outperformed the state-of-art model DeepFRI.

- On AFch test set, which includes AF2 predicted structures with low sequence similarity to the training set and no experimentally resolved structures, HEAL outperforms other state-of-the-art methods.
- Therefore, HEAL has great potential for application in real-world scenarios.

FUTURE WORK

- In the future, we aim to modify single sequence structure prediction models so that the learned evolutionary and structural information can be leveraged to annotate more sequences in larger datasets such as CAFA.