Water position prediction with SE(3)-Graph Neural Network

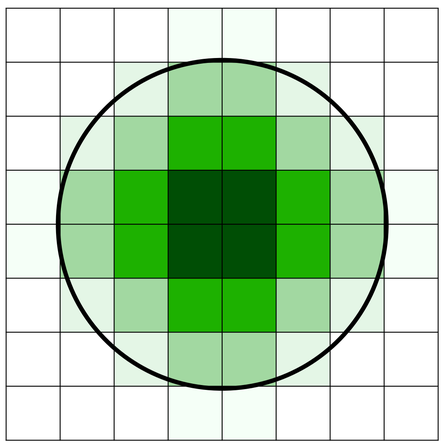
*Sangwoo Park**

Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea

ABSTRACT

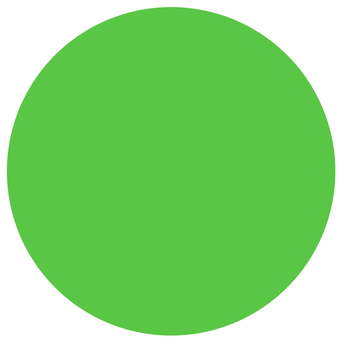
- Protein-bound **water molecules** affects the structure and function of the protein.
- To consider such water molecules, positions of bound water from given protein structure should be known. If only protein structure is given, **prediction of water molecule's position** could be done.
- For **fast and accurate prediction**, **WatGNN**, water position prediction method using **SE(3)-Graph Neural Network** (SE(3)-GNN) was developed in this research.

CNN (0.5Å spacing)



Multiple cells are needed for each atom

GNN



Single node is needed for each atom

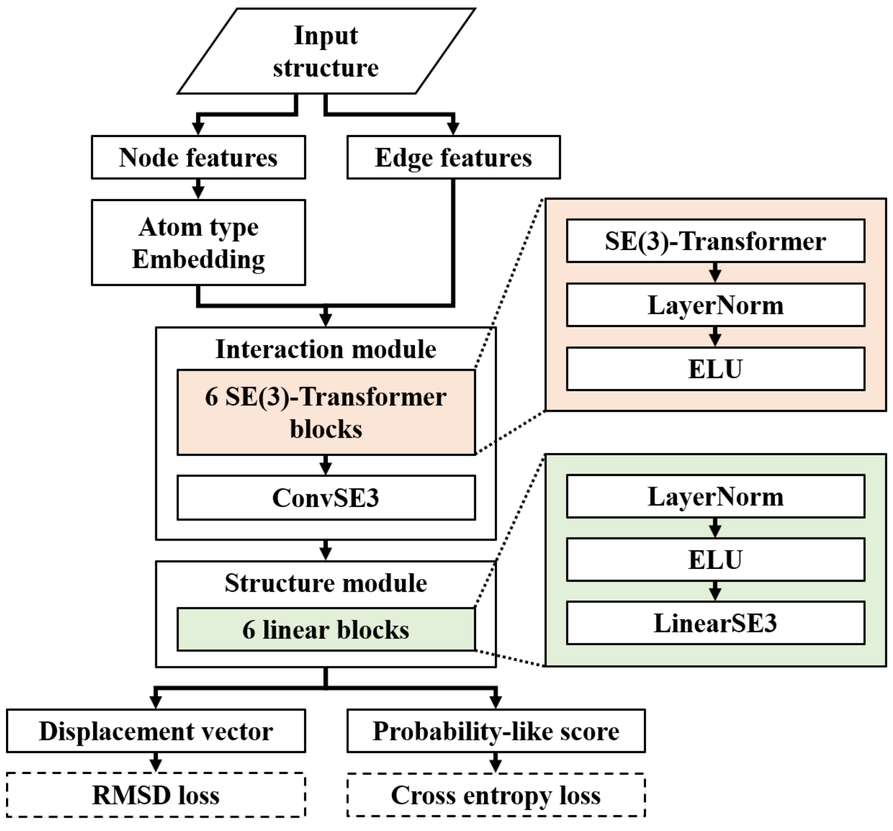
- With SE(3)-GNN based method, water position accuracy was better than state-of-art methods, while average computational cost was 1.86 seconds on i9-12900k CPU and RTX-4090 GPU, which is **12.8 times faster** than using our previous prediction method using Convolutional Neural Network.
- The preprint and prediction program is available on Biorxiv and Github, respectively.
  - Preprint: www.biorxiv.org/content/10.1101/2024.03.25.586555v1
  - Github: https://github.com/shadow1229/WatGNN (Github repository)

Preprint

Github

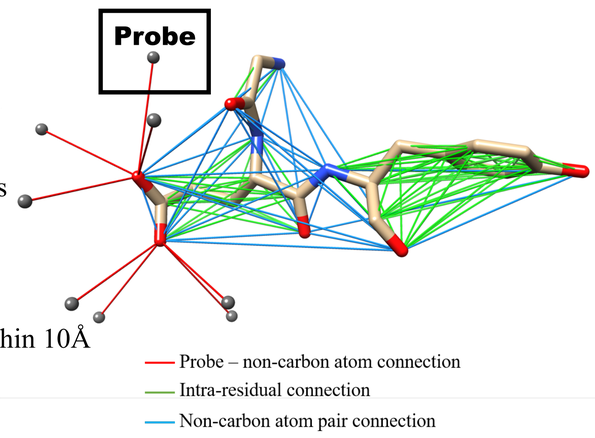
WatGNN Method

• Network Structure



• Features

- Node features: Atom types only
  - C,N,O,S,P, halogens, alkali metals, non-alkali metals, probes, other atoms
- Edge features
  - Probe-non carbon atom connections
  - Intra-residual connection
  - Non-carbon atom pair connection within 10Å



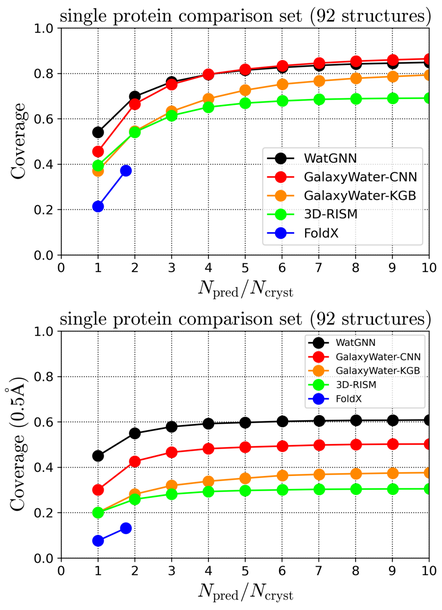
• Training set

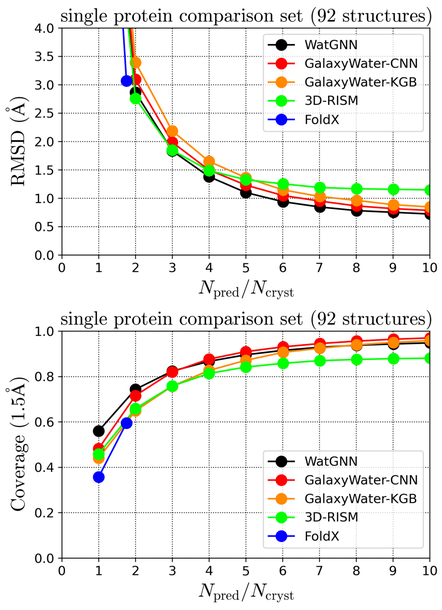
- High resolution X-ray crystallography protein structure set
  2.0Å or better resolution, sequence identity < 30%, < 5000 heavy atoms
- 7971 structures (curation date: Sep 21, 2023)
- training: 3971, validation: 300, test: 3700 structures

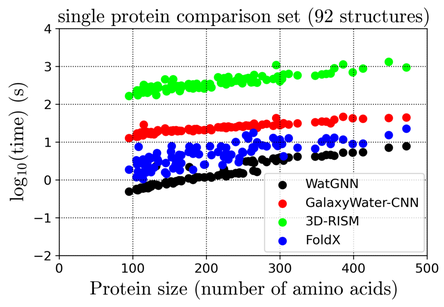
RESULT

• Prediction performance comparison

- Comparison set
  - Single protein structure set
    - 92 X-ray crystal structures, 1Å resolution or better, released before March 14th, 2015.
  - Protein-compound structure set
    - 397 structures curated from the PDBBind set, 2Å resolution or better, sequence identity < 30%, tanimoto similarity < 50%
- Performance for single protein structure set

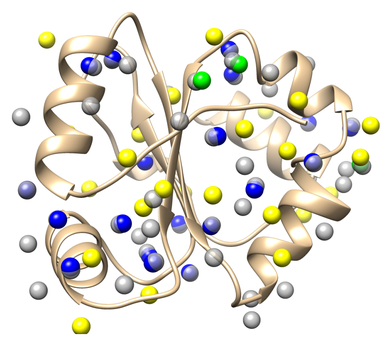






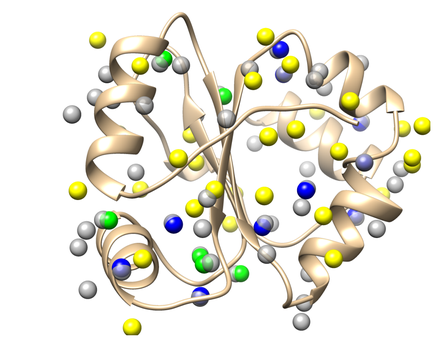
Average computation time
(i9-12900k + RTX4090)

WatGNN:	1.86s
GalaxyWater-CNN:	23.77s
3D-RISM:	398.77s
FoldX (CPU only):	5.72s



WatGNN

Predicted water sites



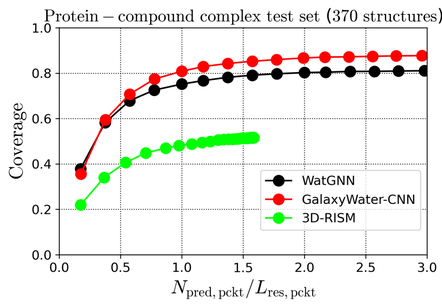
GalaxyWater-CNN

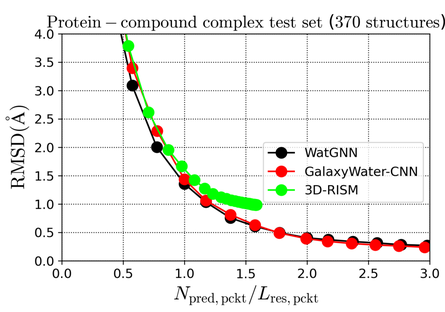
Crystallographic water w/ predicted sites within 0.5Å

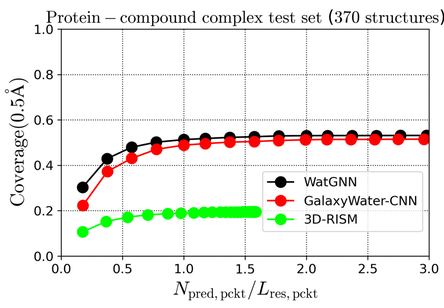
Crystallographic water w/ predicted sites within 1.0Å

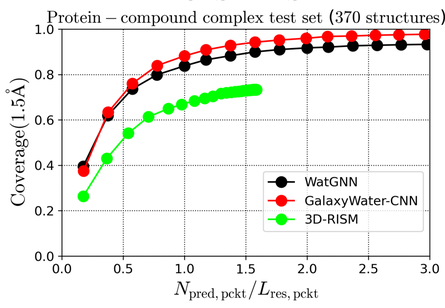
Crystallographic water w/o predicted sites within 1.0Å

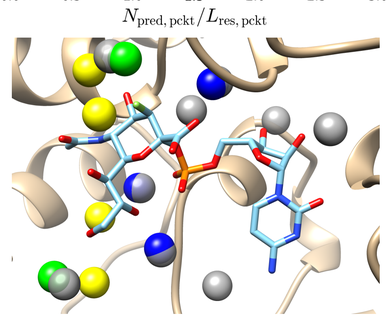
- Performance for protein-compound structure set



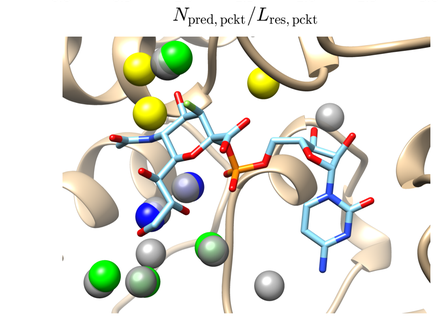








WatGNN



GalaxyWater-CNN

CONCLUSION

- Considering an atom as a single node for water position prediction via GNN greatly decreases computational cost compared with the previous method using CNN.
- Using GNN enables more accurate water position prediction without limited resolution due to CNN's grid representation.
- The limited amount of protein-ligand complex set for training limits the performance of WatGNN, which implies that the amount of training data impacts performance of WatGNN to what extent.