Graph neural network for protein-ligand binding affinity predictions

Rui Wang[1], Timothy Szocinski[1], Duc Nguyen[1], Guo-Wei Wei[1,*]
[1]Department of Mathematics, Michigan State University

Introduction

Learning tasks often require dealing with graph data which contains rich information among graph nodes. Graph Neural Network (GNN) has become one of the most popular models for learning from graph inputs in various fields such as physics, chemistry, biology and linguistics. Our work focused on protein-ligand binding affinity prediction by using flexibility-rigidity index (FRI) of protein-ligand complexes as graph inputs and training GNN hyper-parameters automatically. We employ datasets CASF-2007 to validate the Pearson correlation, robustness and reliability of our GNN model.

Flexibility-rigidity index (FRI)

Consider a biomolecule having N atoms with coordinates given as $\{r_i | r_i \in \mathbb{R}^3, i = 1,2,\dots,N\}$. Then commonly used FRI correlation functions include the generalized exponential functions will be

$$\Phi^E_{\kappa,\tau}(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|) = e^{-(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|/\eta_{ij})^{\kappa}}, \kappa > 0$$

and the generalized Lorentz functions

$$\Phi^L_{\upsilon,\tau}(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|) = \frac{1}{1 + (\|\boldsymbol{r}_i - \boldsymbol{r}_j\|/\eta_{ij})^{\upsilon}}, \upsilon > 0$$

where $\eta_{ij} = \tau(r_i + r_j)$ and r_i to be the van der Walls radius of ith atom.

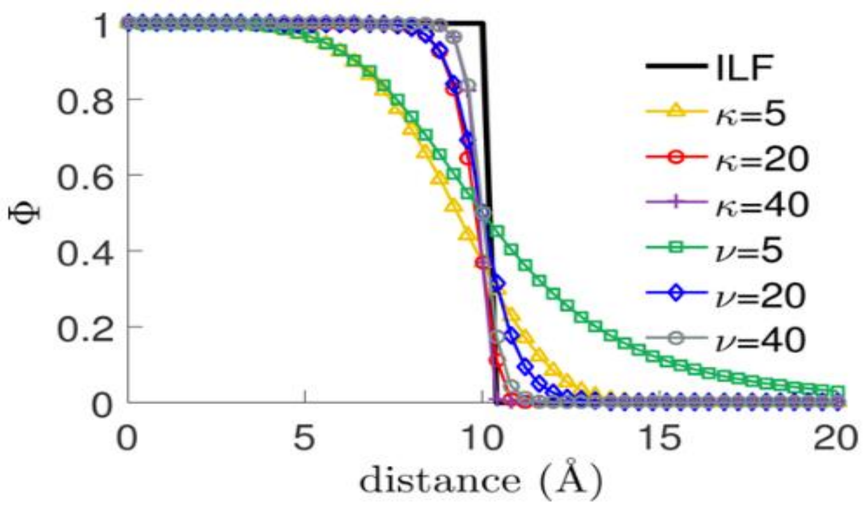


Figure1. FRI correlation functions, which behave like the ideal low filter (ILF) at large κ or υ values

RI-Score

- We define element–specific protein-ligand rigidity index by collecting cross correlations

$$RI^{\alpha}_{\beta,\tau,c}(X - Y) = \sum_{k \in X \in P} \sum_{l \in Y \in L} \Phi^{\alpha}_{\beta,\tau}(\|\boldsymbol{r}_k - \boldsymbol{r}_l\|),$$

with $\|\boldsymbol{r}_k - \boldsymbol{r}_l\| \le c$. Here, $\alpha = E, L$ is kernel index, c is cutoff distance to reduce computational complexity. X denotes heavy atoms $\{C, N, O, S\}$ in the protein and Y denotes heavy atoms $\{C, N, O, S, P, F, Cl, Br, I\}$ in the ligand.

- This representation allows the multiresolution analysis of protein-ligand binding interactions by varying hyper-parameter τ.

Why do we consider GNN?

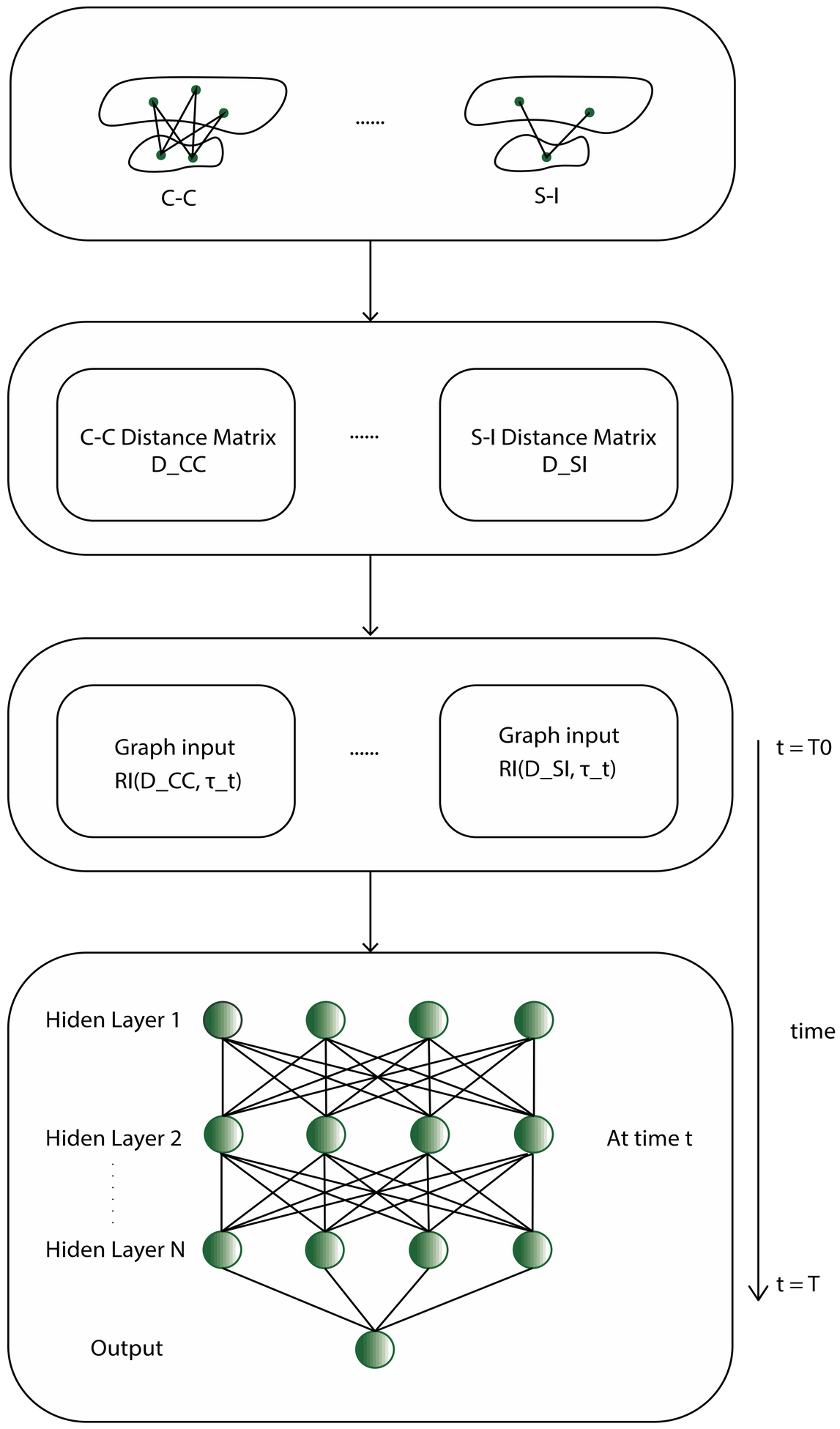
- Although ANN can be implement to predict the binding affinity for protein-ligand binding problems[1], it's time consuming to search all possible hyper-parameter τ.
- We will consider to use FRI of protein-ligand complexes as graph inputs and treat hyper-parameters τ as the parameter in GNN to search the best τ automatically.

Reference

[1] Nguyen, Duc, and Guo-Wei Wei. "AGL-Score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening." Journal of Chemical Information and Modeling (2019).

[2] Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein -ligand Binding Affinity with Applications to Molecular Docking. Bioinformatics 2010, 26, 1169−1175. (33)

Structure of GNN for PL Binding



Performance

- By choosing 3 exponential kernels with 1×36 different τ in (2.5,15) as initial parameters, we can search the best τ by neural network automatically. After 2000 epochs, the Pearson correlation coefficient on CASF-2007 is 0.781.
- We run our code on GPU and running time is about 6 hours.

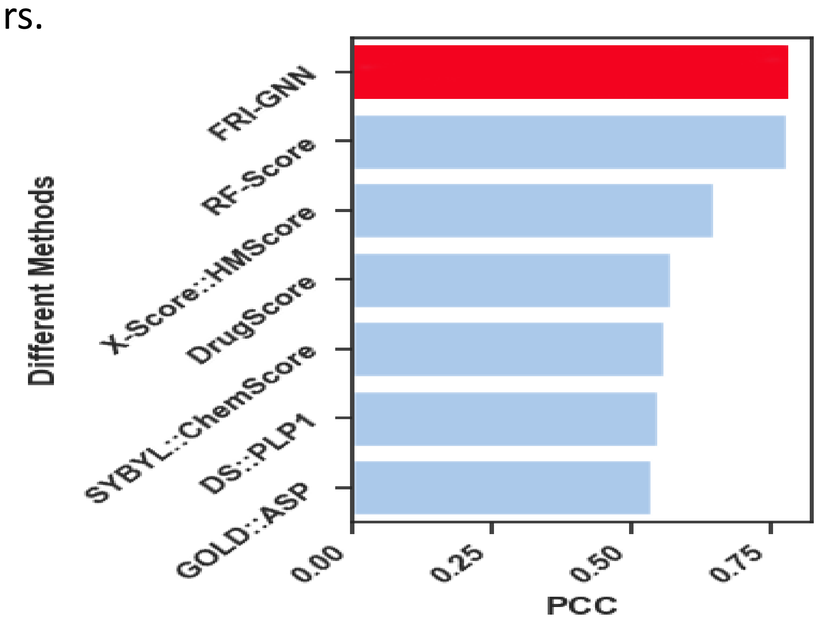


Figure2. Performance comparison between different methods on the PDBBind v2007 core set. The performance of other methods are adopted from ref [2].

Acknowledgement

This work was supported in part by NSF Grants DMS-1721024, DMS-1761320, and IIS1900473 and NIH grants GM126189 and GM129004. DDN and GWW are also funded by Bristol-Myers Squibb and Pfizer.