Electronic Supplementary Information for:
PIGNet2: A Versatile Deep Learning-based Protein–Ligand Interaction Prediction Model for Binding Affinity Scoring and Virtual Screening†

Seokhyun Moon,[a] Sang-Yeon Hwang,[b] Jaechang Lim,[b] and Woo Youn Kim*[abc]

[a]Department of Chemistry, KAIST, 291, Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea,
[b]HITS Incorporation, 124, Teheran-ro, Gangnam-gu, Seoul, 06234, Republic of Korea,
[c]AI Institute, KAIST, 291, Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea,
*Corresponding author; E-mail: wooyoun@kaist.ac.kr.

July 17, 2023

Contents

List of Tables

1 Training details

1.1 Mathematical formulation

PIGNet2 predicts the binding affinity E^{pred}, utilizing both the protein graph, $\mathcal{G}^p = (\mathcal{V}^p, \mathcal{E}^p)$ and the ligand graph, $\mathcal{G}^l = (\mathcal{V}^l, \mathcal{E}^l)$. Here, we only considered heavy atoms as nodes of the graph for both proteins and ligands. For the initial node features, refer to Table 1. We adopted two types of adjacency matrices: intramolecular and intermolecular. The former is used to learn internal information about either the ligand or protein, considering edges only between nodes connected by a covalent bond. The latter, on the other hand, is designed to update the ligand and protein node features with additional information about their counterparts. For this, we consider an edge to exist only when the distance between a ligand (protein) node and protein (ligand) node is greater than 0.5 Å and less than 5 Å.

PIGNet2 shares the same model architecture as PIGNet.[1] The initial node features of protein and ligand are embedded in a node feature, h, using the same feedforward network. Following this, the node features incorporate both intramolecular and intermolecular information through two networks: the gated graph attention network and the interaction network. All ligand and protein node features are then concatenated pairwise to predict the binding affinity of a given complex.

Feature	List of available elements
Atom type	C, N, O, F, P, S, Cl, Br, X
Degree of atom	0, 1, 2, 3, 4, 5
Hybridization	s, sp, sp^2, sp^3, sp^3d, sp^3d^2, unspecified
Period	1, 2, 3, 4, 5, 6
Group	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
Aromaticity	0, 1

Supplementary Table 1: The list of initial atom features.

1.2 Loss functions

During the training process of PIGNet2, we consider the structures from two major data augmentation strategies: near-native structures, which include crystal structures from positive data augmentation (PDA) and non-binding structures from negative data augmentation (NDA). To design a versatile deep learning-based PLI prediction, it is essential to explicitly predict the binding affinity of near-native structures and non-binding structures. In order to accurately predict the binding affinity of the near-native structures, we utilized the regression loss ($L_{\text{regression}}$), which is the mean squared error loss between the experimental binding affinity, E^{true}, and the predicted binding affinity, E^{pred}, as shown in Equation 1.

$$L_{\text{regression}} = \frac{1}{N}\sum_i (E_i^{\text{pred}} - E_i^{\text{true}})^2, \tag{1}$$

Meanwhile, similar to the previous PIGNet, for non-binding structures, we employed a hinge function denoted as $L_{\text{augmentation}}$, as shown in Equation 2, to better distinguish them from the near-native structures.

$$L_{\text{augmentation}} = \frac{1}{N}\sum_i \max(E_i^{\text{target}} - E_i^{\text{pred}}, \epsilon_{\text{criterion}}) \tag{2}$$

The values of E^{target} and $\epsilon_{\text{criterion}}$ can vary depending on the type of data augmentation. As a result, $L_{\text{augmentation}}$ encourages training to predict E^{pred} to be greater than $E^{\text{target}}+\epsilon_{\text{criterion}}$. Specifically, $L_{\text{augmentation}}$ divides into three components depending on the type of data augmentation: $L_{\text{re-docking}}$, $L_{\text{cross-docking}}$, and $L_{\text{random-docking}}$. $L_{\text{re-docking}}$ is a loss function for structures generated by re-docking data augmentation. To predict the binding affinity to be a little more unstable than the experimental binding affinity corresponding to the native structure, we set E^{target} to be E^{true} and $\epsilon_{\text{criterion}}$ to be -1. Both $L_{\text{cross-docking}}$ and $L_{\text{random-docking}}$ are considered non-binding structures. We set E^{target} to -6.8 kcal/mol, so that the predicted binding affinity is greater than -6.8 kcal/mol, which reflects a generally accepted criterion for ineffective binding, equivalent to the condition that pIC_{50} exceeds 10 μM. Correspondingly, $\epsilon_{\text{criterion}}$ has been set to 0.

In summary, the total loss function is a linear combination of the aforementioned loss functions and can be expressed as follows:

$$\begin{aligned} L_{\text{total}} = & L_{\text{regression}} \\ & + C_{\text{re-docking}} L_{\text{re-docking}} \\ & + C_{\text{random-docking}} L_{\text{random-docking}} \\ & + C_{\text{cross-docking}} L_{\text{cross-docking}}, \end{aligned} \tag{3}$$

where $C_{\text{re-docking}}$, $C_{\text{cross-docking}}$, and $C_{\text{random-docking}}$ are constant hyperparameters, each set as 10.0, 5.0, and 5.0, respectively.

References

[1] S. Moon, W. Zhung, S. Yang, J. Lim and W. Y. Kim, *Chem. Sci.*, 2022, **13**, 3661–3673.