# Electronic Supplementary Information for: PIGNet: A physics-informed deep learning model toward generalized drug-target interaction predictions

Seokhyun Moon,‡[a] Wonho Zhung,‡[a] Soojung Yang,‡[a§]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]KI for Artificial Intelligence, KAIST, 291, Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea,
‡These authors contributed equally to this work.
*Corresponding author; E-mail: wooyoun@kaist.ac.kr.

January 25, 2022

## Contents

## List of Figures

## List of Tables

[§] Currently at Computational and Systems Biology, MIT, 77 Massachusetts Ave, Cambridge, MA

# 1. Training details

## (a) Input features of the neural network

### Atom features

Initial atom features in our model are summarized in Table 1. The X in the atom type corresponds to all other atom types except C, N, O, F, P, S, Cl, and Br. The final dimension of the node vector is 54.

| Feature | list |
|---|---|
| Atom type | C, N, O, F, P, S, Cl, Br, X (onehot) |
| Period | 1, 2, 3, 4, 5, 6 (onehot) |
| Group | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 (onehot) |
| Degree of atom | 0, 1, 2, 3, 4, 5 (onehot) |
| Hybridization | $s$, $sp$, $sp^2$, $sp^3$, $sp^3d$, $sp^3d^2$, unspecified (onehot) |
| Formal Charge | -2, -1, 0, 1, 2, 3, 4 (onehot) |
| Aromaticity | 0 or 1 |

Supplementary Table 1: The list of initial atom features

### Adjacency matrices

Our graph representation, $G(H, A)$, contains two adjacency matrices expressed as equations (1) and (2). $A^1$ and $A^2$ are constructed to account for the covalent bonds and intermolecular interactions in a protein-ligand complex, respectively.

$$A^1_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected by covalent bonds or } i{=}j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$A^2_{ij} = \begin{cases} 1 & \text{if } 0.5 \text{ Å} \leq d_{ij} \leq 5.0\text{Å} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

For $A^2$, we neglect the atom-atom pair interactions whose pairwise distance is smaller than 0.5 or larger than 5.0. By placing the upper threshold, we limit the effect of distant atoms and reduce the complexity of the graph representation. By setting the lower threshold, we avoid exceptional atom pairs within extremely short distances.

## (b) Model architecture

### Gated graph attention network (Gated GAT)

The $n^{th}$ unit of the gated GAT generates a set of the next node features from the set of the current node features, $H^n = \{h^n_1,\ h^n_2,\ \cdots,\ h^n_N\}$, and the adjacency matrix $A^1$, where $h^n_i \in R^F$. The scalar values $N$ and $F$ are the number of the atoms in a protein-ligand complex and the dimension of the node feature, respectively. The initial step of the gated GAT is the multiplication of a learnable weight, $W^n_1 \in R^{F\times F}$ and the node feature, $h^n_i$ to produce an embedded node feature, $m^n_i$, which has more information about protein-ligand complex. From the embedded node feature, $m^n_i$, the attention coefficient, $e^n_{ij}$ between the $i^{th}$ and the $j^{th}$ nodes is computed as follows:

$$e^n_{ij} = (m^n_i)^T W^n_2 m^n_j + (m^n_j)^T (W^n_2)^T m^n_i, \tag{3}$$

where $W^n_2 \in R^{F\times F}$ is also a learnable matrix. $e^n_{ij}$ implies the influence of the $i^{th}$ node to update the features of the $j^{th}$ node. The summation of $(m^n_i)^T W^n_2 m^n_j$ and $(m^n_j)^T (W^n_2)^T m^n_j$ forces $e^n_{ij}$ and $e^n_{ji}$ to be equal. We adopted the softmax activation function to normalize the attention coefficient, $e^n_{ij}$, across neighboring nodes. The normalized attention coefficient, $a^n_{ij}$, is given by

$$a_{ij}^n = \frac{exp(e_{ij}^n)}{\sum_{j \in N_i} exp(e_{ij}^n)}, \tag{4}$$

where $N_i$ is the set of the neighboring nodes of the $i^{th}$ node. Then, the current node feature, $\tilde{h}_i^n$ is calculated via the linear combination of the neighboring node features weighted by the attention coefficient, $a_{ij}$, with a ReLU activation function:

$$\tilde{h_i^n} = ReLU(\Sigma_j a_{ij}^n h_j^{n\prime}). \tag{5}$$

We also used the gate mechanism to effectively deliver the previous node features and the current node features to the next node features. The importance of the previous node features, $z_i$, is computed from $h_i^n$ and $\tilde{h}_i^n$ as follows:

$$z_i = \sigma(W_3^n((h_i^n \| \tilde{h_i^n}))), \tag{6}$$

where $\sigma$ is a sigmoid activation function which constrains $z_i$ between 0 and 1, $(\cdot||\cdot)$ is a concatenation operation, and $W_3^n \in R^{2F \times 1}$ is a learnable weight vector. Lastly, the next node feature, $h_i^{n+1}$, is a linearly interpolated value between $h_i^n$ and $\tilde{h}_i^n$:

$$h_i^{n+1} = z_i h_i^n + (1 - z_i)\tilde{h_i^n}. \tag{7}$$

We used three units of the gated GAT to incorporate intramolecular interactions into the node features.

### Interaction network

The interaction network takes the previous set of node features, $H^n = \{h_1^n,\ h_2^n,\ \cdots,\ h_N^n\}$, and the adjacency matrix $A^2$ to generate the next set of node features, $H^{n+1} = \{h_1^{n+1},\ h_2^{n+1},\ \cdots,\ h_N^{n+1}\}$. The interaction network first multiplies $h_i^n$ with a learnable weight, $W_1^n \in R^{F \times F}$ to get the set of embedded node features, $M^1 = \{m_1^1,\ m_2^1,\ \cdots,\ m_N^1\}$ as follows:

$$m_i^1 = W_1^n h_i^n, \tag{8}$$

where $i$ is the index of node features. The interaction network also makes a set of interaction embedded node features, $M^2 = \{m_1^2,\ m_2^2,\ \cdots,\ m_N^2\}$ with each previous node feature, $h_i^n$, a learnable weight, $W_2^n \in R^{F \times F}$ , and the adjacency matrix $A^2$. The interaction embedded feature of the $i^{th}$ node is represented by:

$$m_i^2 = \max_{j \in N_i}\{W_2^n h_j^n\}, \tag{9}$$

where $N_i$ is the set of nodes which have interactions with the $i^{th}$ node. By maximum aggregation for each node, the set of interaction embedded node features, $M^2$, becomes the most important node feature element within nodes with intermolecular interactions. From $M^1$ and $M^2$, we can get a set of total node features, $H'^n = \{h_1'^n, h_2'^n, \cdots, h_N'^n\}$ through the summation and a ReLU activation function:

$$h_i'^n = ReLU(m_i^1 + m_i^2). \tag{10}$$

The next set of node features, $H_i^{n+1}$, can be obtained from a gated recurrent unit (GRU)[1] by using a set of total node feature, $H'^n$, as a hidden state input and the previous set of node features, $H^n$.

$$h_i^{n+1} = GRU(h_i'^n,\ h_i^n) \tag{11}$$

Total node features, $H'^n$, updates the previous set of node features, $H^n$, recursively in the GRU cell. As a result, the next set of node features, $H^{n+1}$, is more likely to reflect only important features of the given protein-ligand complex when updating node features.

The interaction network makes a significant role in our model by transforming a set of node features of the protein-ligand complex to contain information about intermolecular interactions. To make a set of node features that sufficiently contains intermolecular interaction information, the interaction network consists of three units.

## (c) Physics-informed parameterized functions

PIGNet consists of several physics-informed parameterized functions: four energy components and a rotor penalty. Each energy component is computed with a set of pair-wise node features, $H^{concat}$, which represented as equation (12). Each pair-wise node feature consists of two node features, $h_i$ and $h_j$.

$$\begin{aligned} H^{concat} &= \{h_1^{concat},\ h_2^{concat},\ \cdots,\ h_{N^2}^{concat}\} \\ &= \{(h_1||h_1),\ (h_1||h_2),\ \cdots,\ (h_N||h_{N-1}),\ (h_N||h_N)\} \end{aligned} \tag{12}$$

Each energy component depends on $d_{ij}$, and $d'_{ij}$, which are distance, and corrected minimum distance between the $i^{th}$ node and the $j^{th}$ node. $d'_{ij}$ can be represented as follows:

$$d'_{ij} = r_i + r_j + c \cdot b_{ij}, \tag{13}$$

where $r_i$ and $r_j$ are van der Waals radii of the $i^{th}$ and the $j^{th}$ nodes respectively, and $b_{ij}$ is corresponding correction between two nodes. For the constant $c$ that scales $b_{ij}$, we used 0.2. The correction constant, $b_{ij}$, originates from a set of pair-wise node features, $H^{concat}$, by using learnable weights, $W^1 \in R^{2F\times F}$ and $W^2 \in R^{F\times 1}$, as the following:

$$b_{ij} = tanh(W^2(ReLU(W^1(h_{ij}^{concat})))). \tag{14}$$

### van der Waals interaction

We used the 12-6 Lennard-Jones potential to calculate a van der Waals interaction term, $e_{ij}^{vdw}$, between the $i^{th}$ and $j^{th}$ atoms. Equation (15) summarizes $e_{ij}^{vdw}$:

$$e_{ij}^{vdw} = c_{ij}\left[\left(\frac{d'_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{d'_{ij}}{d_{ij}}\right)^{6}\right], \tag{15}$$

where $c_{ij}$ indicates the minimum interaction energy which is also predicted from neural networks. We constrain the minimum and maximum values as 0.0178 and 0.0356, respectively, to render the predicted energy component similar to the true energy component. The maximum value of $c_{ij}$ is referred from the parameter of AutoDock Vina for steric interactions.[3] Equation (16) summarizes the calculation of $c_{ij}$:

$$c_{ij} = \sigma(W_2^{vdw}(ReLU(W_1^{vdw}h_{ij}^{concat}))) \times (0.0356 - 0.0178) + 0.0178, \tag{16}$$

where $W_1^{vdw} \in R^{2F\times F}$, and $W_2^{vdw} \in R^{F\times 1}$, are weight matrices. We consider all protein and ligand atom pairs except metal atoms whose van der Waals radii have high variance depending on atoms types. We obtain the total van der Waals energy, $E^{vdw}$, by summing $e_{ij}^{vdw}$ of all possible pairs, as follows.

$$E^{vdw} = \sum_{i,j} e_{ij}^{vdw} \tag{17}$$

The hydrogen bond, metal-ligand interaction, hydrophobic interaction components, and rotor penalty can be computed as described in the main article.

### Hydrogen bond, Metal-ligand interaction, Hydrophobic interaction

Table 2 shows SMARTS descriptors which are used to select the hydrogen bond donors and hydrogen bond acceptors.

| | |
|---|---|
| hydrogen bond acceptor | [$([!#6;+0]);!$([F,Cl,Br,I]);!$([o,s,nX3]);!$([Nv5,Pv5,Sv4,Sv6])] |
| hydrogen bond donor | [!#6;!H0] |

Supplementary Table 2: SMARTS descriptors for hydrogen bond acceptor and donor

### (d) Hyperparameter settings and computational resources

This section will show all the parameters used for the model training. The hidden dimension, $F$ used in the gated GAT and the interaction network were given as 128 for all implemented model architectures. During the training, the dropout ratio and learning rate were 0.1 and 0.0001, respectively. The minimum values and ratios for each loss term were used as mentioned in the previous section.

The batch size of every model trained with the data augmentation is fixed to 8 with the RTX 2080 Ti GPU. The 3D CNN-based model, the 3D GNN-based model and PIGNet are trained to 200, 1,100 and 2,300 epoch, respectively. Again, the PIGNet (single) result was obtained from the model without dropout, while the result of PIGNet (ensemble) was produced by using 30 ensemble models and the same dropout ratio as the training phase.

## 2. Benchmark methods

### (a) CASF-2016 benchmark metrics

In this supplementary section, we explain the CASF-2016 benchmark metrics we report as our results[2].

- **Scoring power**: This is defined as a linear correlation of predicted binding affinity and experimental binding data. The linear correlation is measured in the Pearson's correlation coefficient, R. In fact, the scoring functions in comparison with PIGNet produce log affinity ($logK_a$) values, while PIGNet computes binding free energy values [$kJ/mol$]. Nevertheless, since a constant multiplication converts $logK_a$ to the binding free energy, such a difference does not make a change in the R values.

- **Ranking power**: This refers to the ability of a model to correctly rank the binding affinities of the true binders for a certain target protein, given the binders' precise binding poses. The ranking power can be measured in terms of the average Spearman's rank correlation coefficient, $\rho$, the Kendall's rank correlation coefficient, $\tau$, and the Predictive Index ($PI$). We only report Spearman's rank correlation coefficients for our experiments, as all three metrics are well-correlated.

- **Docking power**: This is the ability of a model to find the native ligand binding pose among decoys with computer-generated poses. The metric for the docking power measurement is the overall success rate, which counts a complex identified by a model as a successful case if it has a high conformational similarity (RMSD < 2Å) with the native binding pose.

- **Screening power**: This is the ability of a model to identify the true binding ligands for a given target protein among a set of random molecules. We measure the screening power in terms of the enhancement factor (EF) and success rate, averaged across all 57 target proteins. The success rate is computed as a ratio of highest-affinity binder among the top $\alpha(\%)$ ligands. The enhancement factor for a target protein is defined as follows:

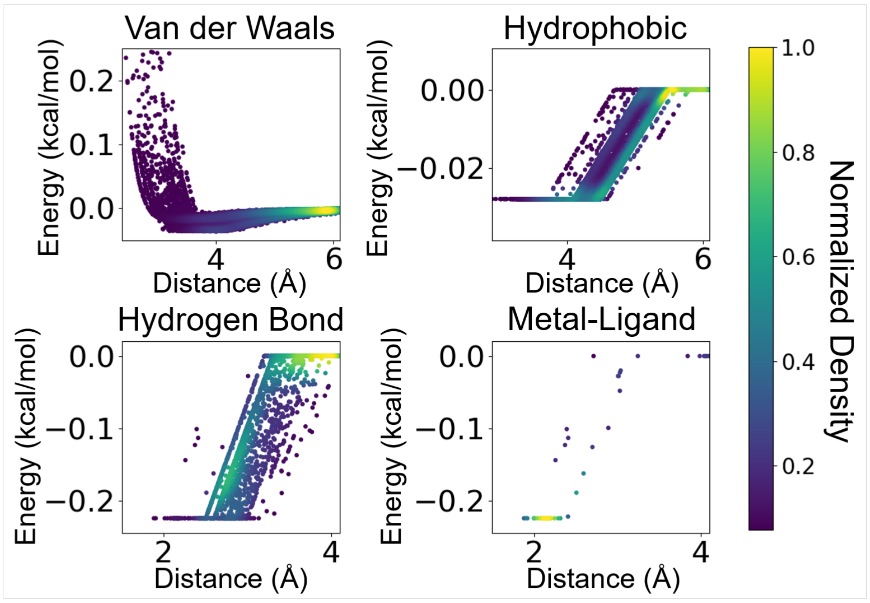
$$EF_\alpha = \frac{NTB_\alpha}{NTB_{total} \times \alpha}, \tag{18}$$

  where $NTB_\alpha$ is the number of the true binders among the top $\alpha(\%)$ candidates ranked by a model, and $NTB_{total}$ is the total number of the true binders for the given target protein. We cite our results in the average enhancement factor and success rate measured at $\alpha = 1\%$.

## 3. Interpretation of the physically modeled outputs

### (a) Distribution plot of atom-atom pairwise interaction in each energy component

By dissecting the predicted energies into individual energy components, we could observe that the model has learned the deviations within each energy component. Fig. 1 shows a distance-energy plot of each energy component, where the data points are the atom-atom pairs in the test set. Note that the pairwise energy plots are not a single solid line, but the multifariously deviated distributions. For the van der Waals component, while the plot generally complies with the form of the 12-6 Lennard-Jones potential, the deviations arise from the two learnable parameters, $b_{ij}$ and $c_{ij}$. $b_{ij}$ also contributes to the deviations in the hydrophobic, hydrogen bond, and metal energy components, as the parameter is used to calculate the corrected sum of van der Waals radii in equation (13).



Supplementary Figure 1: The distance-energy plot for each energy component in the test set. The closer to the yellow, the larger the number of pairs corresponding to the data point.

## References

[1] Junyoung Chung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *Preprint at arXiv:1412.3555* (2014).

[2] Minyi Su et al. "Comparative Assessment of Scoring Functions: The CASF-2016 Update". In: *Journal of Chemical Information and Modeling* 59.2 (2019), pp. 895–913.

[3] Oleg Trott and Arthur J Olson. "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of computational chemistry* 31.2 (2010), pp. 455–461.