SCIENCE CHINA
Information Sciences

• Supplementary File •

Integrating Sequence and Graph Information for Enhanced Drug-Target Affinity Prediction

Haohuai He[†] , Guanxing Chen[†] & Calvin Yu-Chian Chen[*]

Artificial I ntelligence M edical R esearch C enter, S chool o f I ntelligent S ystems Engineering,
Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107*, China*

Appendix A Methods

To address the limitations of both sequence-based and graph-based methods, we propose a method that leverages the strengths of both approaches while mitigating their shortcomings. Specifically, we utilize the sequence-based DTA prediction method for feature extraction and represent the internal structure using a graph-based method.

Like the sequence-based approach, we initially represent the DP pair with the SMILES of the drug and the amino acid sequence of the protein. These are then tokenized into separate tokens. Assuming the number of drug tokens is l_1 and the number of protein tokens is l_2, we can obtain the following formula:

$$\begin{aligned} token_d &= \left[d_1, d_2, \ldots, d_{l_1}\right], \\ token_p &= \left[p_1, p_2, \ldots, p_{l_2}\right]. \end{aligned} \tag{A1}$$

Subsequently, we use a feature extractor (FE) to extract features from them. In this study, the LSTM model serves as the FE, as illustrated in the following formula:

$$\begin{aligned} \left[f_{d_1}, f_{d_2}, \ldots, f_{d_{l_1}}\right] &= FE\left(token_d\right), \\ \left[f_{p_1}, f_{p_2}, \ldots, f_{p_{l_2}}\right] &= FE\left(token_p\right). \end{aligned} \tag{A2}$$

As indicated by the formula, each token $token_i$ in the corresponding position can obtain its own feature representation f_i. Then, we construct graph structures for the DP pairs separately, i.e., g_d and g_p. The nodes in the drug graph g_d represent atoms, and the edges represent chemical bonds. As for the protein graph g_p, the nodes represent amino acids, and the edges are formed by contact maps, i.e., there is an edge between two amino acids when they are close in space.

The connectivity of edges in the DP graphs can be represented using the following matrix:

$$g_d = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,l_1} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,l_1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{l_1,1} & m_{l_1,2} & \cdots & m_{l_1,l_1} \end{pmatrix}, \tag{A3}$$

$$g_p = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,l_2} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,l_2} \\ \vdots & \vdots & \ddots & \vdots \\ m_{l_2,1} & m_{l_2,2} & \cdots & m_{l_2,l_2} \end{pmatrix}, \tag{A4}$$

where $m_{i,j}$ represents the connectivity between the i-th and j-th nodes in the graph. The value of $m_{i,j}$ is 1 if there is an edge between the i-th and j-th nodes, and 0 otherwise:

$$m_{i,j} = \begin{cases} 1, & \text{There is an edge between node i and node j} \\ 0, & \text{There is no edge between node i and node j} \end{cases} \tag{A5}$$

Next, we connect the matrices using a virtual node to form a hybrid graph. Virtual nodes facilitate information exchange between two graphs, hence they must connect edges with other nodes in the mixed graph. The connectivity of the virtual node is represented by the following matrix:

$$g_h = \begin{pmatrix} 1 & 1 & 1 \\ 1 & g_d & 0 \\ 1 & 0 & g_p \end{pmatrix}. \tag{A6}$$

[*] Corresponding author (email: chenyuchian@mail.sysu.edu.cn)

[†] Haohuai He and Guanxing Chen have the same contribution to this work.

Node features in the graph are derived from the feature values extracted by the sequence-based method. The hybrid graph is then passed into a graph neural network to obtain the DTA prediction of the DP pair.

The sequence-based feature extraction method extracts features from the nodes in the hybrid graph, specifically from the SMILES strings and amino acid sequences of drugs and proteins. These features are then incorporated into the hybrid graph to enhance the representation of the DP pairs. This method efficiently and effectively encodes the structural information of drugs and proteins, vital for accurate DTA prediction.

During the training process, features from nodes in the hybrid graph are also extracted using the sequence-based feature extraction method. This improves the quality of the node feature representation and promotes interaction between the sequence-based and graph-based methods.

Our approach combines the strengths of both sequence-based and graph-based methods to extract features from DP pairs and predict DTAs. By integrating these two methods, the approach achieves a high level of accuracy and interpretability in DTA prediction.

Appendix A.1 Graph neural network

Graph neural networks (GNNs) are a class of neural networks that operate on graphs, used to process data with graph structures, such as social networks, citation networks, and biological networks. GNNs have been widely used in many fields, such as node classification, link prediction, and graph classification. In this study, we use GNNs to learn the features of the DP pair and predict the binding affinity, specifically employing the Graph Isomorphism Network (GIN) model. The GIN model is a generalization of the Weisfeiler-Lehman (WL) graph isomorphism test. It is a simple and efficient GNN model that can be applied to a variety of graph-based tasks. The GIN model is based on the following formula:

$$h_v^{(k)} = MLP^{(k)}\left(\left(1+\epsilon^{(k)}\right)h_v^{(k-1)} + \sum_{u\in N(v)} h_u^{(k-1)}\right), \tag{A7}$$

where $h_v^{(k)}$ represents the representation of node v in the k-th layer, $N(v)$ denotes the set of neighbors of node v, and $MLP^{(k)}$ signifies a multi-layer perceptron. The $\epsilon^{(k)}$ is a trainable parameter. The GIN model can be viewed as a generalization of the GCN model, but it uses a sum aggregator instead of a mean aggregator.

Appendix A.2 Long short-term memory (LSTM)

LSTM is a recurrent neural network (RNN) architecture capable of learning long-term dependencies. It is extensively used in natural language processing (NLP) and speech recognition. In this study, we employ LSTM to learn the features of the DP pair and predict the binding affinity. The LSTM model is based on the following formula:

$$\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right),\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right),\\
c_t &= f_t c_{t-1} + i_t\sigma\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right),\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right),\\
h_t &= o_t\sigma\left(c_t\right),
\end{aligned} \tag{A8}$$

where x_t is the input vector, h_t is the hidden state vector, c_t is the cell state vector, i_t is the input gate, f_t is the forget gate, o_t is the output gate, and σ is the sigmoid function. The W and b are the weight matrix and bias vector, respectively. The LSTM model can be considered a generalization of the RNN model. Its unique feature is the use of a memory cell to store information, which effectively addresses the problem of vanishing and exploding gradients inherent in the RNN model.

Table A1 Description of symbols mentioned in the text

Symbols	Description
d_i	Token of the i-th drug sequence
p_i	Token of the i-th protein sequence
f_{d_i}	The feature extracted from token di
g_d	Drug molecular graph
g_p	Protein molecular graph
g_h	Hyper graph of drugs and proteins
$h_v^{(k)}$	Representation of node v in the k-th layer
$N(v)$	The set of neighbors of node v
x_t	Input vector
h_t	The hidden state vector
c_t	The cell state vector
i_t	The input gate
f_t	The forget gate
o_t	The output gate
σ	The sigmoid function
W	The weight matrix
b	Bias vector

Appendix B Innovations

We propose a hybrid graph that integrates both the DP interaction network and the DP sequence similarity network. This hybrid graph effectively amalgamates information from both networks, thereby enhancing the performance of the DTA prediction model. Furthermore, the hybrid graph can be readily extended to other graph-based DTA prediction models.

During the training process, the sequence-based feature extraction method is employed to extract initial features from the nodes in the hybrid graph. However, these initial features may not be optimal for the task of DTA prediction as they do not take into account the structure of the graph. Consequently, these features are adapted through training, using the graph structure to refine their representation.

By adaptively updating the node features in the hybrid graph, our approach is capable of learning a more accurate and informative representation of the DP pair. This not only enhances the accuracy of DTA prediction, but also boosts the interpretability of the model by providing insights into the crucial features for prediction.

The innovative adaptive feature extraction method incorporated in our approach facilitates a more effective integration of sequence-based and graph-based methods, leading to a more potent and versatile approach for predicting DTAs.

Moreover, by amalgamating the strengths of both sequence-based and graph-based methods, our approach attains a high degree of accuracy and interpretability in DTA prediction. The graph-based method offers a comprehensive representation of the interactions between drugs and proteins at a molecular level, while the sequence-based method captures the structural information of drugs and proteins at a local level. This combination of global and local information guarantees a more holistic and accurate prediction of DTAs. Additionally, the use of graph-based methods permits the interpretation of the prediction results, which is vital for understanding the mechanism of drug action and designing innovative DP pairs.

Appendix C Datasets

We evaluate our method on two datasets: Davis and KIBA. The two datasets are widely used in the field of DTA prediction.

Appendix C.1 Davis

Davis et al. [1] conducted a comprehensive analysis of kinase inhibitors by testing 72 of them for their interaction with 442 kinases. This study covered more than 80% of the human catalytic protein kinome. The resulting dataset comprises 30,056 DP pairs and includes the SMILES representation of each drug, the protein sequence, and the IC50 activity value as a label. To standardize the activity value, we transformed it to the negative logarithm of pIC50.

Appendix C.2 KIBA

Tang et al. [2] proposed a model-based ensemble approach called KIBA to leverage the complementary information captured by various bioactivity types, such as IC50, Ki, and Kd, for an integrated drug-target bioactivity matrix. The resulting matrix includes 118,083 DP pairs, encompassing 2,068 drugs and 229 proteins. Each DP pair includes a KIBA score as the binding activity value label. The KIBA score is a weighted sum of the binding affinity values of the three bioactivity types.

Appendix D Evaluation Metrics

Appendix D.1 Mean Squared Error

Mean Squared Error (MSE) is a common metric used to evaluate the performance of regression models. It measures the average squared difference between the predicted values and the actual values. The formula for calculating MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{D1}$$

where y_i is the actual value, $\hat{y_i}$ is the predicted value, and n is the number of samples.

Appendix D.2 Concordance Index (CI)

The concordance index (CI) is a metric commonly used in survival analysis to evaluate the predictive accuracy of a model [3]. It measures the proportion of all possible pairs of patients for which the predicted survival time is correctly ordered. The CI ranges from 0.5 (no better than random prediction) to 1 (perfect prediction accuracy). The formula for calculating the CI is:

$$CI = \frac{1}{Z}\sum_{i=1}^{n}\sum_{j=1}^{n} I(\hat{y_i} > \hat{y_j}) I(y_i > y_j) \tag{D2}$$

where $I(\hat{y_i} > \hat{y_j})$ is an indicator function that returns 1 if $\hat{y_i} > \hat{y_j}$ and 0 otherwise, and $I(y_i > y_j)$ is an indicator function that returns 1 if $y_i > y_j$ and 0 otherwise. Z is a normalization factor, which is equal to the number of pairs of samples that can be formed from the dataset.

Appendix D.3 Regression Toward the Mean Index (r_m^2)

The r_m^2 metric is a measure used to evaluate the external prediction performance of a regression model. It is based on the standard coefficient of determination (r^2) but incorporates information about the model's intercept. The formula for calculating r_m^2 is:

$$r_m^2 = (1 - sqrt(r^2 - r_0^2)) * r^2 \tag{D3}$$

where r^2 is the standard coefficient of determination, and r_0^2 is the coefficient of determination of the model's intercept. The r_0^2 value is calculated as follows:

$$r_0^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \tag{D4}$$

where y_i is the actual value, $\hat{y_i}$ is the predicted value, $\bar{y_i}$ is the mean of the actual values, and n is the number of samples. Intuitively, r_m^2 can be thought of as a way to penalize models that overfit the data by adding an additional term that depends on the model's intercept. The term (1 - sqrt(r^2 - r_0^2)) represents the degree of overfitting, with larger values indicating more overfitting. By multiplying this term by the standard r^2 value, we obtain a modified coefficient of determination that reflects both the model's accuracy and its degree of overfitting.

Appendix E Results

Appendix E.1 Compare with 3D graph methods

Intuitively, incorporating the 3D structures of drugs and proteins can improve the accuracy of DTA prediction. Therefore, building on intuitive reasoning and previous research [4, 5], we developed a 3D graph model inspired by GraphDTA. Specifically, we substituted the drug encoding component of GraphDTA with a 3D graph neural network (GNN) and evaluated its performance using the Davis dataset. Our findings, as presented in Table E1, reveal that the DTI network employing a 3D GNN does not provide a significant advantage over GraphDTA, which uses a 2D GNN. Furthermore, the performance of the 3D GNN-based approach is notably inferior to that of NHGNN, which capitalizes on the strengths of both sequence-based and graph-based methods. This discrepancy could potentially be attributed to the 3D GNN's inability to effectively simulate the DTA binding pocket and model chemical bonds, despite its utilization of the drug's 3D structure. Although 3D GNNs may prove advantageous in other tasks, they may not be the optimal choice for DTI prediction.

Table E1 Compared 3D GNN on DAVIS with other DTA methods.

Methods	MSE ↓	CI↑
GraphDTA	0.229	0.893
GraphDTA & 3D GNN	0.228	0.891
Ours	**0.196**	**0.914**

Appendix E.2 Error analysis

To further examine the factors influencing the accuracy of our model, we conducted an error analysis on its predictions across subsets of different sizes of drugs and proteins within the Davis and KIBA datasets. We divided the test sets into multiple subsets based on the length of drugs and proteins and evaluated NHGNN's performance on each subset. The SMILES length distribution in the Davis dataset is relatively uniform, with the exception of the 40-45 interval, which has a higher sample count. Conversely, the KIBA dataset exhibits an imbalanced distribution, and for ease of visualization, we excluded 39 outliers with SMILES lengths exceeding 100. Excluding intervals with insufficient samples, an increasing trend in MSE with longer SMILES lengths is noticeable in both datasets, suggesting that extended SMILES lengths may negatively impact prediction accuracy. The protein sequence length distribution in both datasets provided adequate samples for protein lengths under 1500, while the sample count for lengths exceeding 1500 was minimal. An upward trend in MSE was observed for intervals shorter than 1500. However, intervals exceeding 1500 exhibited erratic MSE, likely due to limited data availability. As the size of the protein or drug increases, the DTA model must handle larger inputs, which can necessitate more message passing for feature interactions within GNNs. Our error analysis substantiates this observation.

References

1 Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nat biotechnol*, 29(11):1046–1051, 2011.

2 Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inf Model*, 54(3):735–743, 2014.

3 Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, an Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. *Adv Neural Inf Process Syst*, 20, 2007.

4 Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

5 Haohuai He, Guanxing Chen, and Calvin Yu-Chian Chen. 3GDT-DDI: 3d graph and text based neural network for drug–drug interaction prediction. *Brief in Bioinformatics*, 23(3):bbac134, 2022.