1 Evaluation metrics

The mean squared error (MSE) is used to evaluate the prediction accuracy of the model, that is, the deviation between the predicted values and the label values:

$$\mathrm{MSE} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} (\widehat{y_\mathrm{i}} - \mathrm{y_i})^2 \tag{1}$$

where $\widehat{y_\mathrm{i}}$ is the predicted value of the i-th sample and $\mathrm{y_i}$ is the label value of the i-th sample. And m is the number of samples in a group. The smaller MSE values indicate smaller error between the predicted values and the label values.

The concordance index (CI) is used to evaluate the consistency in ranking the predicted values and the corresponding label values for two samples:

$$\mathrm{CI} = \frac{1}{\mathrm{Z}}\left(\sum_{\mathrm{y_i > y_j}} \mathrm{S}(\widehat{y_\mathrm{i}} - \widehat{y_\mathrm{j}})\right) \tag{2}$$

where Z is a normalization constant, $\widehat{y_\mathrm{j}}$ is the predicted value of the j-th sample and $\mathrm{y_j}$ is the label value of the j-th sample. The step function S(h) is defined as follows:

$$\mathrm{S(h)} = \begin{cases} 1, & \text{if } \mathrm{h} > 0 \\ 0.5, & \text{if } \mathrm{h} = 0 \\ 0, & \text{if } \mathrm{h} < 0 \end{cases} \tag{3}$$

The higher CI values show that the consistency of sorting is higher between the predicted values and the label values.

The adjusted coefficient of determination (R_m^2) is used to evaluate the goodness of fit of nonlinear regression model, and the definition is as follows:

$$R_m^2 = R^2 * \left(1 - \sqrt{\left|(R^2)^2 - (R_0^2)^2\right|}\right) \tag{4}$$

The R^2 is the traditional coefficient of determination, reflecting the degree of correlation between the predicted values and the label values, defined as follows:

$$R^2 = \frac{\left(\sum_{i=1}^{m} (y_i - \bar{y}) * (\widehat{y_\mathrm{i}} - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{m} (\mathrm{y_i} - \bar{y})^2\right) * \left(\sum_{i=1}^{m} (\widehat{y_\mathrm{i}} - \bar{\hat{y}})^2\right)} \tag{5}$$

where $\bar{y}$ is the average value of all label values in a group of samples and $\bar{\hat{y}}$ is the average value of all predicted values in a group of samples.

The R_0^2 is a special determining coefficient, which scales the predicted value by adjusting the factor k to eliminate the influence caused by different scales. It is defined as follows:

$$R_0^2 = 1 - \frac{\sum_{i=1}^{m} (\mathrm{y_i} - k * \widehat{y_\mathrm{i}})^2}{\sum_{i=1}^{m} (\mathrm{y_i} - \bar{y})^2} \tag{6}$$

where k is defined as follows:

$$k = \frac{\sum_{i=1}^{m} \mathrm{y_i} * \hat{y}_i}{\sum_{i=1}^{m} \hat{y}_i * \hat{y}_i} \tag{7}$$

The $\mathrm{R_m^2}$ combines both the R^2 and $\mathrm{R_0^2}$ to evaluate the model performance more comprehensively. And the higher R_m^2 values indicate the better prediction performance of the model.

2 Baseline methods

- **MgraphDTA:** it uses multi-scale GNN (MGNN) to learn the local and global feature of molecular graph and uses multi-scale CNN (MCCN) to extract protein sequence features at various scales. The obtained feature vectors of compounds and proteins are finally concatenated into FCNN to predict DTA values.
- **BACPI:** it uses a graph attention network (GAT) to capture graph feature representations of compounds and uses 1D-CNNs to encode protein sequences. And an end-to-end bidirectional attention module to explore the mutual relationship between compounds and proteins and then learned features are entered into a FCNN.
- **TEFDTA:** it uses a transformer to extract molecular fingerprints and 1D-CNNs to encode protein sequences. Finally, these mixed features are entered into a FCNN.
- **DeepGLSTM:** it captures molecular characteristics by stacking multiple GCN blocks and uses a bidirectional LSTM module to learn protein features. FCNN is also used to the integrate interactive features.
- **PerceiverCPI:** it utilizes GAT and CNN to extract molecular features of compounds and sequence features of protein respectively. And it innovatively uses nested cross-attention mechanisms to grasp the relationships between proteins and compounds.

Here we downloaded the original codes of these methods and ran them according to the hyper-parameter settings provided by authors.

S1. Supplemental Tables

Table S1. The number of compounds with different length for three datasets

Dataset	Number of compounds	Mean_length	Max_length	Length covering 90%	Length covering 95%	Selected length
Davis (public)	68	54	72	66	68	70
DB_K (ours)	18764	78	729	125	196	200
DB_E (ours)	80404	68	819	88	114	115

Table S2. The number of protein sequences with different length for three datasets

Dataset	Number of proteins	Mean_length	Max_length	Length covering 80%	Length covering 85%	Selected length
Davis (public)	414	720	1400	1030	1072	1050
DB_K (ours)	1536	636	7096	826	976	900
DB_E(ours)	1242	699	7096	883	1011	900

Table S3. Detailed information about the vertex features.

Name	Description	Dim
Atomic Num.	The atomic number (integer)	1
Atom type	[H,C,N,O,F,Cl,s,Br,I] (one-hot)	9
Hydrogens	Number of connected hydrogens (integer)	1
Hybridization	[sp,sp2,sp3] (one-hot)	3
Radical electrons	Number of radical electrons for the atom(integer)	1

Acceptor	Accepts electrons [0/1] (binary)	1
Donor	Donates electrons [0/1] (binary)	1
Aromatic	In an aromatic system [0/1] (binary)	1
Formal charge	Formal charge of the atom (integer)	1
Implicit valence	Implicit valence of the atom(integer)	1
Explicit Hs.	Number of implicit Hs the atom is bound to (integer)	1
Explicit valence	Explicit valence of the atom (integer)	1