Supplementary Information for: HybridDTA: Hybrid Data Fusion through Pairwise Training for Drug-Target Affinity Prediction

Hongyu Luo[1*], Yingfei Xiang[1*], Xiaomin Fang[1†], Wei Lin[2], Fan Wang[1†], Hua Wu[3], Haifeng Wang[3]

[1] Baidu Inc., Shenzhen, China
[2] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
[3] Baidu Inc., Beijing, China
[*] Equal contributions
[†] Corresponding authors

S.1 Data pre-processing

S.1.1 Davis and KIBA

We follow the previous work[4] to pre-process Davis and KIBA datasets. The details can be found at https://github.com/hkmztrk/DeepDTA/tree/master/data.

S.1.2 BindingDB

The BindingDB dataset with the version of May 2021 contains 2,221,487 compound-protein pairs, including 7,965 protein and 963,425 compounds.

We pre-process the raw BindingDB data following steps below: (1) We keep compound-protein pairs with at least one of the measurements (K_D, K_I, IC_{50} and EC_{50}). (2) We remove the affinity values with '>' or '<'. (3) We modify the extreme affinity values by replacing the values more than 10,000 with 10,000. (4) We drop the duplicates. (5) We define a ranking task by considering the proteins, pH degrees ('pH'), temperatures ('Temp (C)'), and data sources ('Curation/DataSource', 'Article DOI', 'PMID', 'PubChem AID', 'Patent Number', 'Authors', 'Institution). (6) For a protein-compound pair in a ranking task, we keep the median affinity value of that pair. (7) We remove the ranking tasks with no less than ten candidate compounds.

For the backbone model GraphDTA, we conduct additional data cleaning. We remove the illegal SMILES sequence which can not be converted by the Cheminformatics software RDKit (https://rdkit.org) and the single atom sequence (['F', '[SH-]', '[I-]', 'S', 'I', '[F-]']) that can not be converted to a molecular graph. We also remove the protein sequences which do not conform to FASTA format (https://zhanggroup.org/FASTA/).

Besides, instead of using original K_D score as the binding affinity value to make the prediction, we normalize and transform it into pK_D (shown in Equation 1), which is similar to the previous works[1–4]:

$$pK_D = -log10(\frac{K_D}{10^9}). \tag{S1}$$

Note that, K_I, IC_{50} and EC_{50} are normalized by the same way.

S.2 Hyper-parameters of HybridDTA

For Davis and KIBA datasets, we use grid search to search the best hyper-parameters for each DTA backbone model on each dataset. The candidate settings of the hyper-parameters for searching are shown in Table S1. We use Adam optimizer and train 100, 30, 50 epochs for GraphDTA, DeepDTA, and MolTrans, respectively.

Table S1: Candidate settings of hyper-parameters for Davis and KIBA.

Hyper-parameter	Candidate settings
deviation ϵ	0.2
sample times (original dataset)	10
sample times (fused dataset)	$\{0.5,1,3,5\}$
learning rate	$\{1\times10^{-3}, 5\times10^{-4}, 1\times10^{-5}\}$
batch size	$\{32, 256, 512\}$

For BindingDB, the candidate settings of the hyper-parameters for searching are shown in Table S2. We use Adam optimizer and train 200, 200, 50 epochs for GraphDTA, DeepDTA, and MolTrans, respectively.

Table S2: Candidate settings of hyper-parameters for BindingDB.

Hyper-parameter	Candidate settings
deviation ϵ	0.2
sample times (K_I)	10
sample times (K_D)	$\{1,3,5\}$
sample times (IC_{50})	$\{0.2, 0.5, 1\}$
learning rate	$\{1\times10^{-3}, 5\times10^{-4}, 1\times10^{-5}\}$
batch size	$\{32, 256, 512\}$

References

[S1] He, T. *et al.* (2017). Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, **9**(1), 1–14.

[S2] Jiang, M. *et al.* (2020). Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, **10**(35), 20701–20712.

[S3] Nguyen, T. *et al.* (2020). GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, **37**(8), 1140–1147.

[S4] Öztürk, H. *et al.* (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, **34**(17), i821–i829.