Supplementary Materials for DeepDrug: A general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction

Qijin Yin[1], Rui Fan[2], Xusheng Cao[2], Qiao Liu[3,*], Rui Jiang[1,*] and Wanwen Zeng[2,3,*]

[1] Ministry of Education Key Laboratory of Bioinformatics, Research Department of Bioinformatics at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China;

[2] College of Software, Nankai University, Tianjin, 300350, China;

[3] Department of Statistics, Stanford University, Stanford, CA 94305, USA;

* To whom correspondence should be addressed.

Supplementary Notes

Supplementary Note1. Drug embedding interpretation

Drug Embedding of DrugBank DDI dataset. We firstly evaluated the drug embeddings on the DrugBank DDI dataset. For each drug in the dataset, the 128-dimension latent feature from drug feature extraction module of pretrained DeepDrug model is extracted. And then t-SNE was performed on these features for dimension reduction and Leiden algorithm[1] is used for clustering.

The clustering results were analyzed by silhouette score and Drug Category Enrichment Score (DCES). The silhouette score is used for evaluating the clustering results. To calculate the Drug Category enrichment, 4149 drug categories were firstly collected from the DrugBank website. We noted that a drug may have multiple categories. In a certain cluster, we calculated the precision, recall and thus F1-score for each drug category. We then regarded the drug category with highest F1-score as the assigned category of this cluster. We defined the DCES as the average of the F1-scores of all clusters. The silhouette score and DCES can be used as a criterion to judge whether the drug embedding is good or not.

Drug Embedding of unseen drugs from DrugBank website. We collected a total of 11,172 drugs from the DrugBank website and removed drugs without any drug category, which results in 6587 drugs in total and only 1701 drugs were used for pre-training DeepDrug. Similar to above, the 128-dimension features of these 6587 drugs are extracted and dimension reduction and Leiden clustering are performed.

To calculate whether the DCES for this dataset is significant or not, we shuffled the drug categories for 1000 times and then calculated the DCES for each time, as disrupting the drug categories results in disrupting the embeddings of the drugs. We found that the mean and the standard deviation of DCES is 0.096.

Supplementary Note2. SARS-CoV-2 applications

Dataset. We collected two drug-target datasets for SARS-CoV-2 from a recent paper[2], which provide a literature-based and an expert-confirmed list of drugs and target proteins for SARS-CoV-2 with 42 and 34 pairs respectively. To construct the corresponding negative samples, we randomly paired the proteins of SARS-CoV-2 and 11164 drugs from the DrugBank website. Two approaches are used for the graph features construction of SARS-Cov-2 proteins. Firstly, the simulation structures of SARS-Cov-2 proteins are provided in SARS-CoV-2 3D database. Secondly, for each protein in the SARS-CoV-2, the most similar templates in the RCSB database are used as the crystal structure of the protein, which are also provided in the SARS-CoV-2 3D database[3].

Drug similarity and protein similarity. To measure the similarity of two drugs, topological fingerprints of two drugs are calculated respectively by rdkit.Chem.Fingerprints.FingerprintMols.FingerprintMol function with default settings, then drug similarity is determined by rdkit.DataStructs.FingerprintSimilarity function. To measure the similarity of SARS-CoV-2 and proteins in BindingDB dataset, we first calculated the pairwise sequence similarity via EMBOSS Needle package with default setting. The drug similarity for each drug in BindingDB dataset is defined as the maximum topological similarity between this drug to each SARS-Cov-2 interacting drugs. Similar to definition of drug similarity, the protein similarity for each protein in BindingDB dataset is defined as the maximum sequence similarity between this protein to each protein in SARS-CoV-2. The drug similarity threshold is set to 0.6, which results in the removal of 64,980 drugs (15.6%). The protein similarity threshold is set to 0.3, resulting in the removal of 124 proteins (6.0%). To sum up, 128045 pairs (17.0%) are removed from BindingDB dataset to build a stringent training dataset for SARS-CoV-2 potential drug prediction.

Affinity prediction. We used the DeepDrug DTI models pretrained on the BindingDB dataset to predict the binding affinity of the drug-target pairs constructed above. Five-fold cross-validation pretrained model are used to make binding affinity predictions and final prediction for each pair is obtained by taking the mean and maximum values of the predictions for these 5 models.

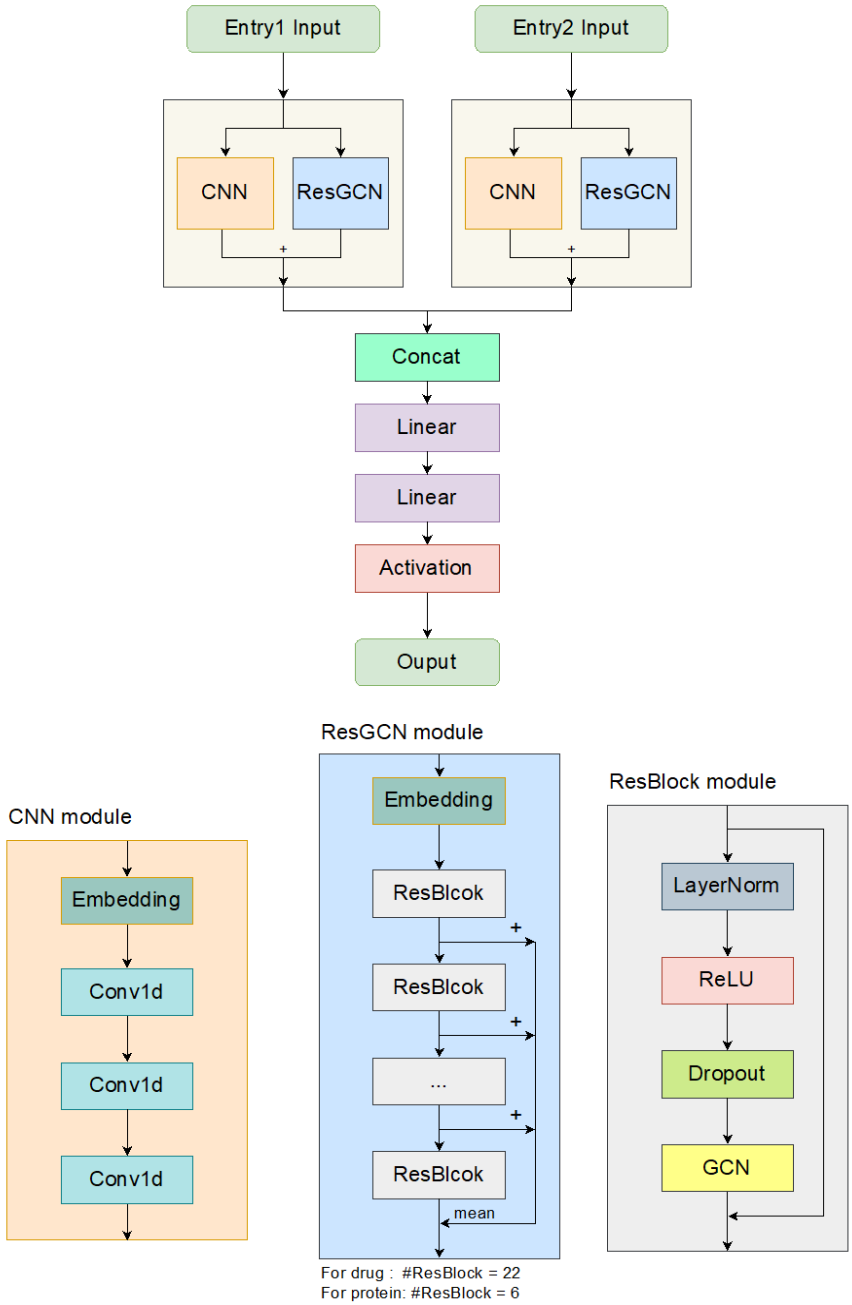
Supplementary Note3. Hyperparameter tuning

DeepDrug was composed of RES-GCN modules, CNN modules, and a combined prediction module. The structural information and sequencing of each drug or protein were fed to a RES-GCN module and a CNN module for feature extraction, respectively. We used Ray-project[4] for hyper-parameters searching. Specifically, we used DrugBank and KIBA datasets for hyperparameter optimization in DDI and DTI tasks, respectively, including the number of the graph convolutional residual blocks (3,6,12,18,22), the size of channels in the graph convolutional layer (32,64,128,256), learning rate (0.01, 0.001, 0.0001) and dropout rate (0.1, 0.3).

References

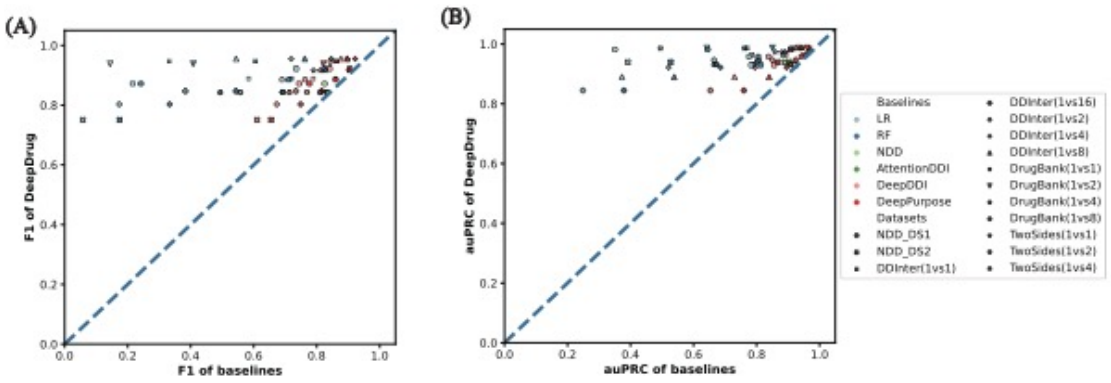
1. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports 9, 1-12 (2019).
2. Gordon DE, et al. A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. bioRxiv, (2020).
3. Alsulami AF, et al. SARS-CoV-2 3D database: understanding the coronavirus proteome and evaluating possible drug targets. Briefings in Bioinformatics 22, 769-780 (2021).
4. Moritz P*, et al.* Ray: A distributed framework for emerging {AI} applications. In: *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*) (2018).

Supplementary Figures

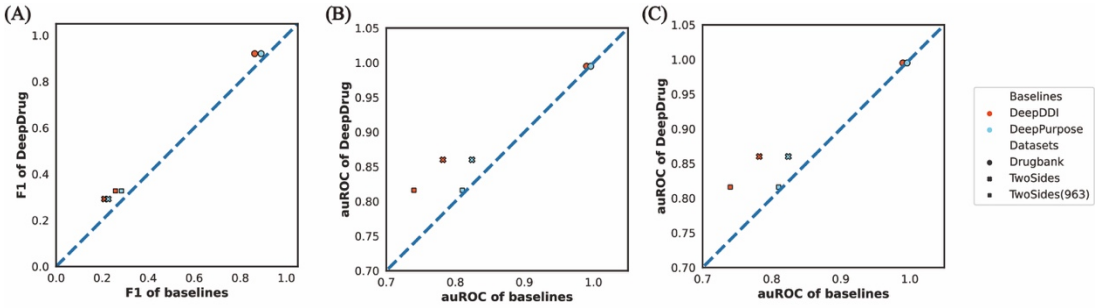


Supplementary Figure 1. Model detailed architecture of DeepDrug. DeepDrug was composed of RES-GCN modules, CNN modules, and a combined prediction module. The structural information and sequencing of each drug or protein were fed to a RES-GCN module and a CNN module for feature extraction, respectively. The CNN module consisted of three one-dimensional convolutional layers with 32,64 and 96 kernels respectively, followed by a one-dimensional adaptive max-pooling layer and a linear layer to extract the sequence embeddings. The kernel sizes of the convolutional layers were 4, 6 and 8 for drugs and 4,8 and 12 for proteins, respectively. Note that the adaptive max-pooling layer is aimed at reducing each kernel features across the sequence dimension by maximum function. The two RES-GCN or CNN modules had shared weights during DDI tasks and are independent for DTI tasks. The features

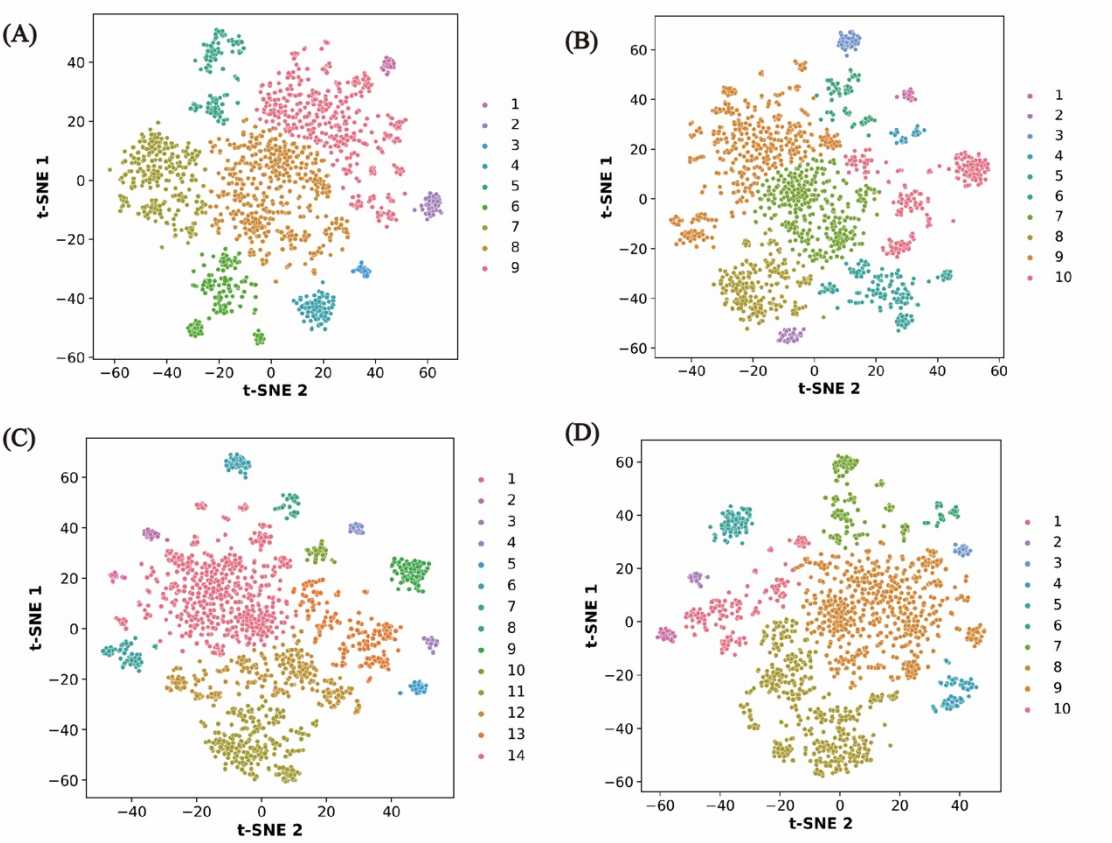
extracted by these modules are concatenated together and fed to the combined prediction module, which consisted of two linear layers and a final prediction layer. The two linear layers had 128 and 32 nodes respectively, and each was followed by a batch normalization layer, a dropout layer and a ReLU nonlinear layer.



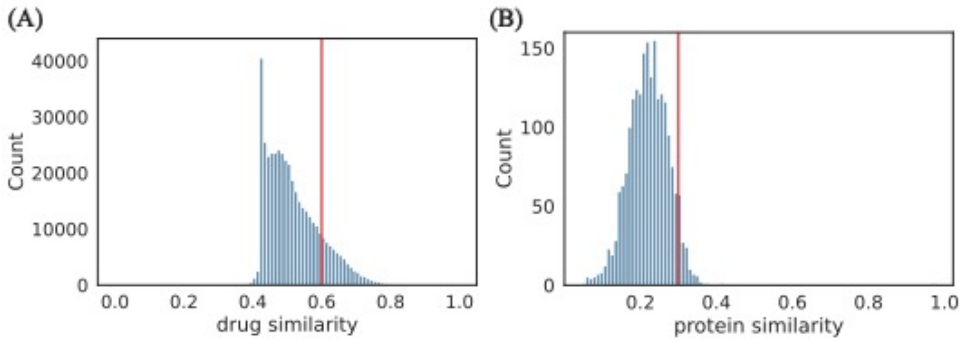
Supplementary Figure 2. DeepDrug is compared with 6 baselines on 3 datasets in DDI binary prediction tasks in terms of F1 (A) and auPRC (B) with different positive-to-negative ratios. The x axis and the y axis of each dot indicate the performance of a certain baseline (indicated by dot color) and DeepDrug on a certain dataset (indicated by the dot shape).



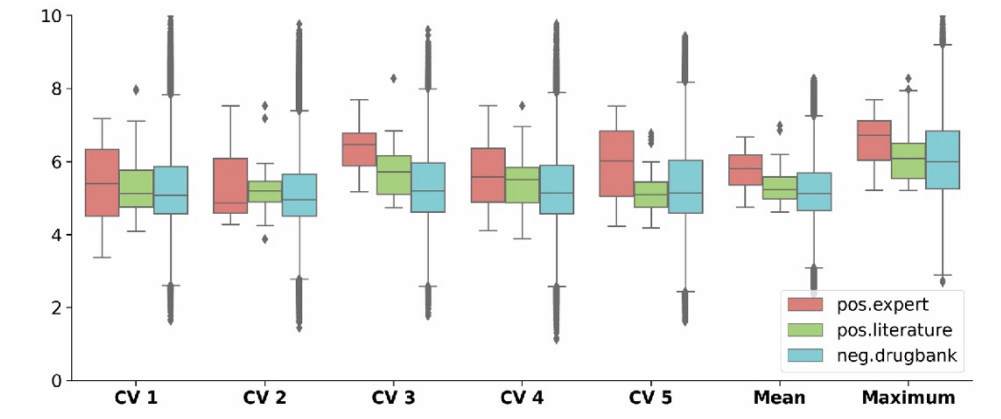
Supplementary Figure 3. Benchmark for DeepDrug on the multi-class/multi-label DDI tasks. DeepDrug are benchmarked with 3 baselines on the 2 datasets in terms of F1 score (A), auROC (B) and auPRC (C). “TwoSides(963)” indicates the Twosides dataset with 963 types of interactions which filters out classes with less than 500 samples.



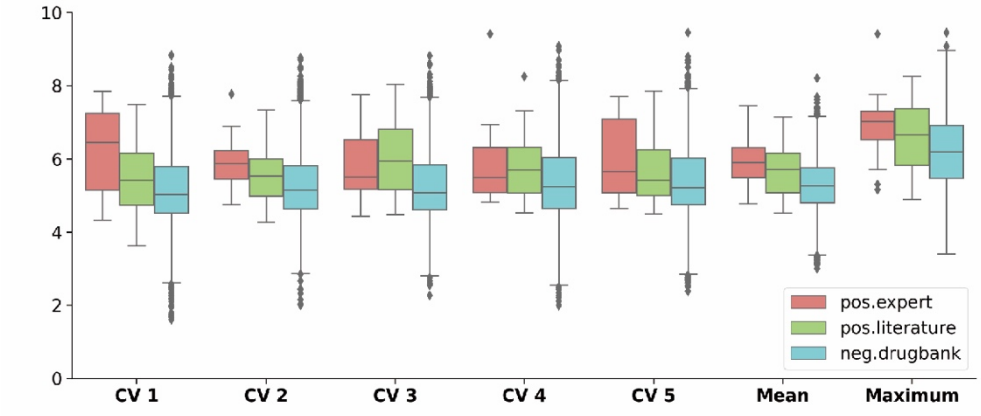
Supplementary Figure 4. Drug embeddings for DeepDrug. The t-SNE visualization for drug embeddings on the DrugBank dataset by DeepDrug in cross-validation (CV) fold 2 (A), CV fold 3 (B), CV fold 4 (C) and CV fold 5 (D).



Supplementary Figure 5. Drug similarity and protein similarity. (A) histogram of drug similarity between SAS-CoV-2 interacting drugs and drugs in the BindingDB dataset. (B) histogram of protein similarity between SAS-CoV-2 proteins and proteins in the BindingDB dataset.



Supplementary Figure 6. Potential drug prediction for SARS-CoV-2 based on SARS-CoV-2 protein fragments. Similar to Figure 5, the performance of DeepDrug to discriminate the potential drug- SARS-CoV-2 pairs and random pairs of the same drugs and SARS-CoV-2 proteins. Unlike Figure 5, the graph features for SARS-CoV-2 proteins are constructed from the similar templates in the RCSB database, rather than the simulation structures.



Supplementary Figure 7. Potential drug prediction for SARS-CoV-2. Similar to Figure 5, the performance of DeepDrug to discriminate the potential drug- SARS-CoV-2 pairs and random pairs of the same drugs and SARS-CoV-2 proteins. Unlike Figure 5, DeepDrug was trained on the original BindingDB dataset, rather than a stringent dataset.