Supplementary Materials for

Multi-task Joint Strategies of Self-Supervised Representation Learning on Biomedical Networks for Drug Discovery

Xiaoqi Wang[1†], Yingjie Cheng[1†], Yaning Yang[1], Fei Li[2*], and Shaoliang Peng[1,3*]

[†]These authors contributed equally to this work.

*Corresponding author. Email:

pittacus@gmail.com (Fei Li);

and slpeng@hnu.edu.cn (Shaoliang Peng).

The Supplementary Materials file includes:

S1. Biomedical heterogeneous network constructions

To construct biomedical heterogeneous networks (BioHNs), we extract three types of nodes and five types of edges from public databases according to deepDTnet [1]. More specifically, the drug-drug interactions are extracted from the DrugBank database (v4.3) [2], where we only select drugs that have experimentally validated target information. The chemical name of each drug is transferred to a DrugBank ID. The drug-protein interaction networks are collected from the DrugBank database (v4.3), PharmGKB [3], and the Therapeutic Target database (TTD) [4]. We extract human protein-protein interactions with multiple pieces of evidences from the HPRD database (Release 9) [5], HuRI [6] and BioGRID [7]. Each protein name is transferred into an Entrez ID (https://www.ncbi.nlm.nih.gov/gene) via the NCBI (https://www.ncbi.nlm.nih.gov/). Drug-disease associations are attained via the fusion of the drug indications in the repoDB [8], DrugBank (v4.3), and DrugCentral databases [9]. Disease-protein associations are collected from two bioinformatic databases, including the Online Mendelian Inheritance in Man (OMIM) database [10] and the Comparative Toxicogenomics database (CTD) [11]. The disease names are standardized according to Unified Medical Language System (UMLS) vocabularies [12], and mapped to the MedGen ID (https://www.ncbi.nlm.nih.gov/medgen/) based NCBI database. In total, the numbers of nodes and edges in the BioHNs are shown in Table S1.

Table S1. The numbers of nodes and edges in BioHNs

Type of node	Count	Type of edge	Count
Drug	721	Drug-Drug interactions	66,384
Protein	1,894	Drug-protein interactions	4,978
Disease	431	Drug-disease associations	1,201
/	/	Protein-protein interactions	16,133
/	/	Disease-protein associations	23,080
Total	3046	Total	111,776

S2. Basic self-supervised tasks on BioHNs

S2.1 Clustering coefficient regression (ClusterPre): In this work, we firstly design ClusterPre to develop the self-supervised representation learning (SSL) for drug discovery. In this pretext task, we aim to predict the clustering coefficient of each vertex to capture the local structure information in a BioHN. Formally, we adopt the mean squared error as the loss function of ClusterPre:

$$L_c(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta\left(f_\theta(i)\right) - Y_{c_i}\right)^2 \quad (1)$$

where θ is the parameter of a graph neural network model $f_\theta(\cdot)$, n represents the number of nodes, $f_\theta(i)$ denotes the representations of node i, $\delta(\cdot)$ is a Sigmoid function, and Y_{c_i}, which is the clustering coefficient for a given node i, can be calculated as follows:

$$Y_{c_i} = \frac{2l_i}{deg_i(deg_i - 1)} \tag{2}$$

where deg_i is the degree of node i, and l_i is the number of links between the deg_i neighbors of node (i.e., the number of triangles that go through node i).

Generally, the clustering coefficients of nodes are larger when they have denser connections to other nodes. The closeness centrality can reflect the local structures in BioHNs to a large extent. The goal of ClusterPre is to ultimately learn the low-dimension representations that preserve the local structure information in BioHNs.

S2.2 Pairwise distance classification (PairDistance): The PairDistance self-supervised task is not limited to a node itself and its local neighborhoods; it also takes a global view of BioHNs. Three key steps as follows form the PairDistance task.

Step1: Randomly select a certain number of node pairs (i, j) for which there is a path between nodes i and j, and calculate the shortest path length $d_{i,j}$ for each node pair. This is done mainly because calculating the shortest path lengths of all node pairs would be computationally expensive, and might be full of challenges for large-scale networks.

Step2: Divide all path lengths $d_{i,j}$ into four categories, that is, $d_{i,j} = 1, d_{i,j} = 2,$ $d_{i,j} = 3$ and $d_{i,j} \geq 4$. Formally, we let $Y_{d_{i,j}} = \{d_{i,j} \mid d_{i,j} = 1, 2, 3, \text{and } d_{i,j} \geq 4\}$ denote the distance categories of node pairs.

Step3: Utilize graph neural networks (GNNs) to predict the distance category of each node pair.

As described in *Step3*, PairDistance can be treated as a multiclass classification problem in which the objective function is formulated as follows:

$$L_{CD}(\theta) = -\frac{1}{|S|}\sum_{(i,j)\in S} \ell\left(\sigma\left(\langle f_\theta(i), f_\theta(j)\rangle\right), Y_{d_{i,j}}\right) \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is an operation that concatenates two vectors, $\ell(\cdot, \cdot)$ represents the cross entropy loss function, and $\sigma(\cdot)$ represents the Softmax function. S and $|S|$ denote the selected set of node pairs (i, j) and the number of node pairs (i, j), respectively.

S2.3 Edge type masked prediction (EdgeMask): In this task, the edge representations, which are obtained by concatenating the representations of its two end-nodes, are fed into the Softmax function to predict the type of the masked edges. EdgeMask can be treated as a five classification problem since there are five types of edges in the constructed BioHN. Similar to PairDistance, we also adopt the cross entropy loss function in EdgeMask.

S2.4 Bio-meta path classification (PathClass): A meta path is defined as a composite sequence that integrates the semantic relationships in BioHNs. For example, "protein $\rightarrow$ disease $\rightarrow$ drug" describes the situation in which a protein causes a disease that is treated by a drug. Given a meta path, we can sample many path instances that have the same semantics, and belong to the same path type. Inspired by the multi-hub characteristics [13] within BioHNs, we design 16 types of meta paths as shown in Table S2, where the first and last objects are drugs and proteins, respectively. Note that all the meta paths include only four nodes mainly because meta paths longer than four nodes may reduce the quality of the associated semantic meanings.

Table S2. The types of meta paths

NO.	Meta path
1	drug-drug-drug-protein
2	drug-drug-protein-protein
3	drug-drug-disease-protein
4	drug-protein-drug-protein
5	drug-protein-protein-protein
6	drug-protein-disease-protein
7	drug-disease-drug-protein
8	drug-disease-protein-protein
9	protein-drug-drug-drug
10	protein-protein-drug-drug
11	protein-disease-drug-drug
12	protein-drug-protein-drug
13	protein-protein-protein-drug
14	protein-disease-protein-drug
15	protein-drug-disease-drug
16	protein-protein-disease-drug

S2.5 Node similarity regression (SimReg): During the GNNs learning process, we perform node message propagation by aggregating local neighborhoods. Concurrently, we also wish to somewhat maintain the similarity attributes in the original feature space. Therefore, we develop SimReg, which requires GNNs to fit the similarity distributions of node pairs in the original feature space. Formally, the objective function employs the mean squared error and is given as follows:

$$L_{simReg}(\theta) = -\frac{1}{|S|}\sum_{(i,j)\in S}\left(\cos\left(f_\theta(i), f_\theta(j)\right) - Y_{sim_{i,j}}\right)^2 \tag{4}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity value between the embedding vectors of two nodes. $Y_{sim_{i,j}}$ is the similarity value between two nodes in the original feature space.

In this work, we use different property similarity measurements according to various types of node. **Chemical similarities among drug pairs:** The simplified molecular input line entry system (SMILES) of each drug is extracted from DrugBank. For a given drug, we transform its SMILES sequence into an MACCS fingerprint by using Open Babel v2.3.1 (http://openbabel.org/wiki/Main_Page). Based on these MACCS fingerprints, we calculate the Tanimoto coefficient [14] of each drug-drug pair as its chemical similarity score. The Tanimoto coefficient offers a value in the range of zero to one and is widely used for drug discovery.

Protein sequence similarity: We download the protein sequences from the Uniprot database (http://www.uniprot.org/). We leverage the Smith-Waterman model [15] to calculate the sequence similarity scores of protein pairs. The Smith-Waterman algorithm performs local sequence alignment by comparing segments of all possible lengths and optimizing the similarity measure for determining similar regions between two strings of protein sequences.

Disease similarity based on protein-protein interaction (PPI) networks: The disease module theory [16] suggests that diseases with overlapping modules in gene-gene networks show significant symptom similarity and comorbidity. We calculate the disease similarity scores by using the ModuleSim algorithm [17-18], which is an extension of disease module theory.

$$sim(dis_1, dis_2) = \frac{2 * SIM(G_1, G_2)}{SIM(G_1, G_1) + SIM(G_2, G_2)} \tag{5}$$

where $G_1 = \{g_{11}, g_{12}, \ldots, g_{1m}\}$ denotes a disease module, which contains m genes for disease dis_1. G_2 is another disease module with a similar definition. $SIM\left(G_1, G_2\right)$ is calculated as follows:

$$SIM\left(G_1, G_2\right) = \frac{\sum_{1\le z\le m_1} F_{G_2}\left(g_{1z}\right) + \sum_{1\le r\le m_1} F_{G_1}\left(g_{2r}\right)}{m_1 + m_2} \tag{6}$$

where $F_{G_2}(g_{1z}) = \text{avg}\left(\sum_{g \in G_2} sp(g_{1z}, g)\right)$. $sp(g_{1z}, g)$ is calculated as follows:

$$sp(g_{1z}, g) = \begin{cases} 1, & \text{if } g_{1z} = g \\ e^{-d_{g_{1z},g}}, & \text{otherwise} \end{cases} \quad (7)$$

where $d_{g_{1z},g}$ is the length of the shortest path between g_{1z} and g in a PPI network.

$F_{G_1}(g_{2r})$ is also calculated according to similar definitions.

S3. The multi-task SSL results obtained in two drug discovery scenarios

Various combinations of multi-task SSL models are used for drug-drug interaction (DDI) and drug-target interaction (DTI) predictions. We design the following two experimental scenarios. **Warm start predictions:** The training set may include the drugs and targets that appear the test set. The results of multi-task combinations are listed in Table S3 for old drug predictions. **Cold start for drugs:** we predict the potential targets and drugs that may interact with newly discovered chemical compounds. In other words, the test set only contains drugs that are unseen in the training set. The results of multi-task combinations are listed in Table S4 for new drug predictions.

Table S3 The results obtained with fifteen SSL combinations in warm start predictions

No.	Self-supervised tasks	Modal size	DDI		DTI	
			AUROC±Std	AUPR±Std	AUROC±Std	AUPR±Std
1	EdgeMask-PairDistance	2	0.917±0.003	0.912±0.005	0.958±0.007	0.951±0.009
2	ClusterPre-PathClass	2	0.915±0.001	0.910±0.003	0.956±0.007	0.956±0.007
3	ClusterPre-PairDistance	1	0.895±0.002	0.882±0.004	0.945±0.006	0.936±0.011
4	EdgeMask-PathClass	1	0.882±0.009	0.869±0.011	0.946±0.005	0.933±0.007
5	PairDistance-PathClass	2	0.887±0.004	0.871±0.005	0.943±0.006	0.937±0.007
6	PathClass-SimCon	2	0.883±0.004	0.867±0.006	0.947±0.009	0.945±0.010
7	PairDistance-SimCon	2	0.880±0.006	0.860±0.012	0.942±0.007	0.935±0.008
8	EdgeMask-SimReg	2	0.875±0.004	0.856±0.009	0.946±0.006	0.932±0.011
9	PairDistance-SimReg	2	0.873±0.004	0.847±0.010	0.936±0.005	0.924±0.008
10	ClusterPre-EdgeMask	2	0.863±0.003	0.839±0.007	0.938±0.007	0.928±0.012
11	SimReg-SimCon	1	0.858±0.002	0.836±0.003	0.916±0.010	0.915±0.013
12	ClusterPre-PairDistance-PathClass	2	0.914±0.003	0.909±0.003	0.958±0.008	0.955±0.013
13	ClusterPre-PathClass-SimReg	3	0.926±0.004	0.923±0.005	0.968±0.016	0.966±0.018
14	PairDistance-SimReg-SimCon	2	0.879±0.008	0.858±0.016	0.944±0.007	0.938±0.010
15	PairDistance-EdgeMask-SimCon	3	0.926±0.004	0.923±0.005	0.968±0.016	0.966±0.018

Table S4 The results obtained with fifteen SSL combinations in cold start predictions

No.	Self-supervised tasks	Modal size	DDI		DTI	
			AUROC±Std	AUPR±Std	AUROC±Std	AUPR±Std
1	ClusterPre-PathClass	2	0.890±0.017	0.873±0.024	0.923±0.028	0.902±0.054
2	EdgeMask-PairDistance	2	0.889±0.014	0.852±0.017	0.927±0.030	0.890±0.077
3	ClusterPre-PairDistance	1	0.871±0.018	0.845±0.026	0.911±0.027	0.864±0.074
4	EdgeMask-PathClass	1	0.862±0.018	0.833±0.025	0.912±0.024	0.861±0.057
5	PairDistance-PathClass	2	0.862±0.021	0.825±0.028	0.912±0.028	0.864±0.067
6	PathClass-SimCon	2	0.858±0.016	0.837±0.020	0.903±0.037	0.868±0.065
7	PairDistance-SimCon	2	0.848±0.022	0.814±0.026	0.913±0.029	0.867±0.072
8	EdgeMask-SimReg	2	0.85±0.020	0.801±0.035	0.909±0.032	0.866±0.079
9	PairDistance-SimReg	2	0.837±0.021	0.792±0.029	0.910±0.026	0.875±0.058
10	ClusterPre-EdgeMask	2	0.822±0.054	0.774±0.070	0.912±0.029	0.860±0.073
11	SimReg-SimCon	1	0.843±0.019	0.813±0.026	0.905±0.028	0.876±0.085
12	ClusterPre-PairDistance-PathClass	2	0.894±0.017	0.878±0.027	0.924±0.027	0.897±0.061
13	ClusterPre-PathClass-SimReg	3	0.909±0.013	0.898±0.016	0.948±0.022	0.939±0.052
14	PairDistance-SimReg-SimCon	2	0.863±0.021	0.825±0.024	0.918±0.024	0.880±0.060
15	PairDistance-EdgeMask-SimCon	3	0.909±0.008	0.895±0.011	0.940±0.020	0.915±0.048

References

[1] Zeng, X. at al. Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* **11**, 1775-1797 (2020).

[2] Law, V., *et al*. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*. **42**, D1091-7 (2014).

[3] Hernandez, T., *et al*. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res*. **36**, D913-8 (2008).

[4] Zhu, F., *et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res*. **40**, D1128-36 (2012).

[5] Keshava Prasad,T. *et al*. Human protein reference database 2009 update. *Nucleic Acids Res*. **37**, D767-D772 (2009).

[6] Figeys D. Mapping the human protein interactome. *Cell Res.* **18**, 716-24 (2008).

[7] Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529-D541 (2019).

[8] Brown, A. S., & Patel, C. J. A standard database for drug repositioning. *Sci. Data*, **4**, 170029 (2017).

[9] Ursu, Oleg. *et al*. DrugCentral: online drug compendium. *Nucleic Acids Res*. **45**, D932-D939 (2017).

[10] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. **33**, D514-D517 (2005)

[11] Davis, A.P. *et al*. The comparative toxicogenomics database: update 2013. *Nucleic Acids Res*. **41**, D1104-D1114 (2013).

[12] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-D270 (2004).

[13] Wang, X. et al. DeepR2cov: deep representation learning on heterogeneous drug networks to discover anti-inflammatory agents for COVID-19. *Brief. Bioinformatics* **22**, 1-14 (2021).

[14] Vilar, S. et al. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat. Protoc.* **9**, 2147-2163 (2014).

[15] Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708 (1982).

[16] Menche, Jörg, et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 6224 (2015).

[17] Luo, P., Li, Y., Tian, L. P., & Wu, F. X. Enhancing the prediction of disease-gene associations with multimodal deep learning. *Bioinformatics* **35**, 3735-3742 (2019).

[18] Ni, P. at al. Constructing disease similarity networks based on disease module theory. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 906-915 (2018).