Supplementary Information for *Deep Learning Can Identify Explainable Reasoning Paths of Mechanism of Drug Action for Drug Repurposing from Multilayer Biological Network*

Jiannan Yang, Zhen Li, William Ka Kei Wu, Shi Yu, Qian Chu, Qingpeng Zhang

1 The metrics used in this study and experiment setup.

In this study, the drug repurposing problem can be viewed as a binary classification task and recommendation task. For the binary classification task, we introduce commonly used metrics, including accuracy, recall, AUROC, AUPRC, whose values are larger, the better the prediction performance of the models. We introduce two commonly used metrics in the recommendation system for the recommendation task: $NDCG@K$ and $Hit@K$. Here, K denotes the top K drugs to be predicted in this study, in other words, for the top K drugs predicted by the model, these metrics will compare each model's performance by checking how well the prediction is in agreement with the ideal ranking.

iDPath and other baselines are implemented with Python 3.7, PyTorch 1.6.0, NumPy 1.19.1 and scikit-learn 0.23.2 with a Rtx 3090. All the models are trained by the Adam optimizer. The ratio of training, validation, and test set is 7: 1: 2. The hyper-parameters are determined by optimizing AUROC on the validation set. The hyper-parameters of iDPath is set as follows: the batch size, learning rate, the number of shortest path L of one drug-disease pair, the fixed-length l_{max} of all the shortest path are set to 512, 0.001, 256, and 8, respectively; the dropout is set to 0.5, the dimension of the embedding and the size of the GCN layers are 64 and 64-128-64, and the size of the LSTM layers are 64-64. For DeepWalk, the walk length $t = 10$, the window size $w = 5$, and the embedding size $d = 128$. The hyper-parameters of GCN, LSTM, and KPRN are the same as that of iDPath.

2 The performance of iDPath compared with state-of-the-art baselines.

Models	Accuracy	Recall	AUROC	AUPRC	F1-Score
iDPath	**0.9416**	**0.9501**	**0.9738**	**0.9684**	**0.9433**
GCN	0.8184	0.8174	0.8681	0.8710	0.8220
LSTM	0.7834	0.8062	0.8668	0.8572	0.7915
KPRN	0.8890	0.8752	0.9551	0.9541	0.8893
DeepWalk	0.6766	0.7063	0.7281	0.6607	0.6826
Random	0.4922	0.4969	0.4913	0.5021	0.5005

Table 1: The performance of iDPath and baselines on the binary classification task in view of the drug repurposing problem. The bold number denotes the best performance.

3 The number of shortest paths has an impact on the performance of iDPath.

In our study, we select a fixed number of shortest paths (L) and fixed number of maximum length of paths (l_{max}) for parallel computing for deep learning. To investigate the impact of L and l_{max}, we report the performance of iDPath on both the binary classification and recommendation tasks under different settings of L and l_{max}. As shown in the bottom figures of Figure 1a and b, we divide all the drug-disease pairs into ten categories based on the logarithm of the number of shortest paths ($\log L$) and the average length of shortest paths l_{max}, respectively. For $\log L$, in our setting, we randomly select 256 shortest paths with replacement (red line in Figure 1) for all the drug-disease pairs. We notice that the performance of iDPath is basically unchanged for the pairs whose L are smaller than 256 while decreased when L is larger than 256, indicating that with more coverage of the whole shortest paths between drug-disease pairs, the better prediction. On the other hand, since the proportion of the pairs whose L are larger than 256 is small (6.65%), even we chose 512 as a setting, the performance of iDPath just slightly increases ($NDCG$@10 increases 0.76% in Figure 1c). Thus, the setting of $L = 256$ is a reasonable choice to balance the performance and training expenses.

We set the maximum length of shortest paths (l_{max}) to 8 to cover almost all the pairs, which means for the paths whose lengths are smaller than 8, we use the padding method to add it to 8, and for the paths whose lengths are longer than 8, we cut these paths from the tail to a length of 8. As shown in Figure 1b, the average length of shortest paths for all the pairs are distributed between 3 and 5, and there is an increasing trend of the performance with the increasing of the average length of shortest paths under the setting of 8. Furthermore, the first category that the average length of shortest paths is 3.65 performs worst compared with other categories. Note that the shortest path in this study includes the head (drug) and the tail (disease), thus the minimal length of these paths is 3. Under this setting, 3.65 indicates that the path info between drug-disease pairs is extremely insufficient, where iDPath cannot work well on. In Figure 1d, the

performance of iDPath cannot improve once l_{max} is larger than 5, which further proves that with comprehensive coverage of the nodes in the shortest paths, the performance of iDPath will not lose.

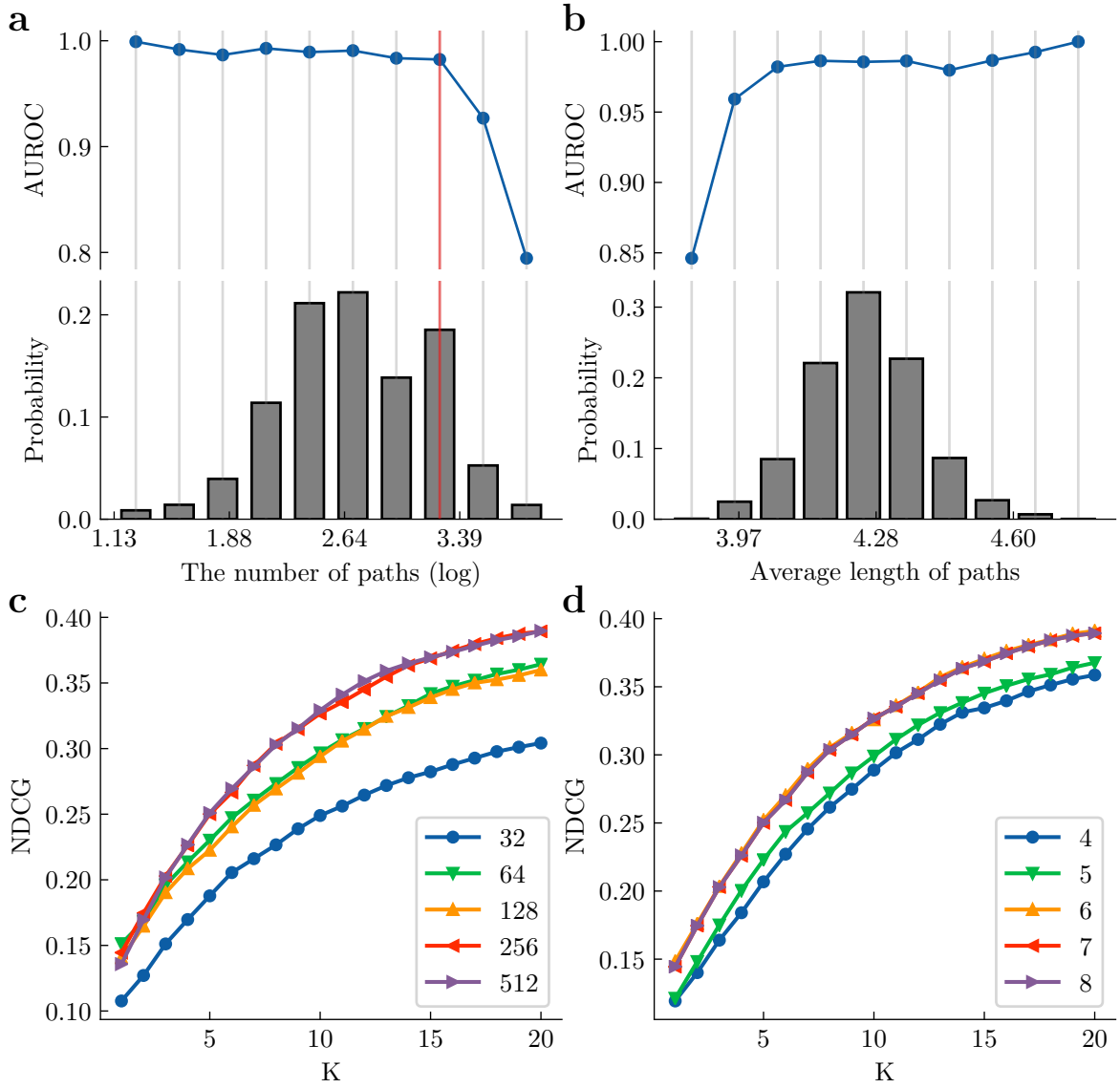

Figure 1: The impact of the setting of the number of paths and the maximum length of paths on the performance of iDPath. **a** Under the setting of the number of the shortest paths (L) as 256, the binary classification performance (AUROC) of iDPath (up) on the drug-disease pairs with different L (distribution of $\log L$ shown in the bottom figure). **b** Under the setting of the maximum length of shortest paths as 8, the binary classification performance (AUROC) of iDPath (up) on the drug-disease pairs with a different average length of shortest paths (distribution shown in the bottom figure). **c** The statistical analysis of the setting of the number of shortest paths (L) of iDPath with other hyper-parameters fixed. **d** The statistical analysis of the setting of the maximum length of shortest paths (l_{max}) of iDPath with other hyper-parameters fixed.

4 The performance of iDPath with different biological network layers.

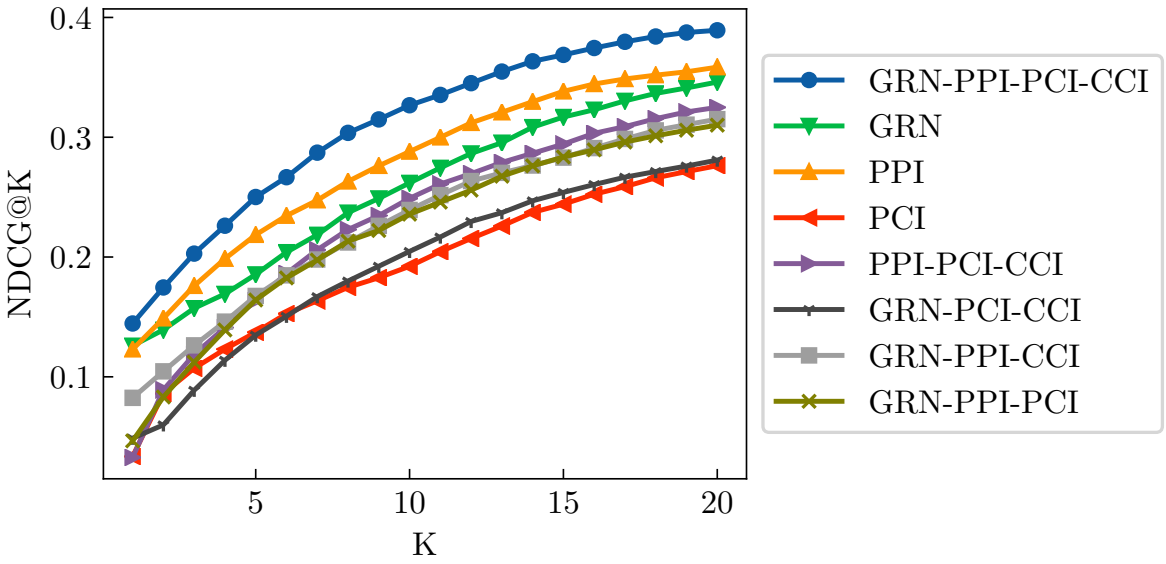

Figure 2: The performance of iDPath with different biological network layers. Here GRN-PPI-PCI-CCI denotes the multilayer biological network generated by these four networks, GRN denotes using gene regulatory network (GRN) alone, and the same goes for PPI, PCI. PPI-PCI-CCI denotes that using PPI, PCI, and CCI together to train iDPath for prediction, and the same goes for GRN-PCI-CCI, GRN-PPI-CCI, and GRN-PPI-PCI. The K values denote the top K drugs used to compute for $NDCG$.