# The Supplementary Materials file includes：

## S1. Result analysis of single task-driven SSL on drug warm start predictions

PathClass attains approximately 10-15% improvements over EdgeMask in terms of AUROC and AUPR scores for DDI and DTI predictions. Another aspect to note is that SimCon is superior to SimReg, with approximately 12.5% average improvements for DDI predictions. To further investigate the difference among various methods, we provide a Student's t-test on the DTI and DDI results, respectively. Here, we assume that there is significant difference between two methods when the *p*-value is below 0.05. In Table S1, we summarize the *p*-value among single task-driven self-supervised representation learning (SSL) models for DDI and DTI predictions. We find that there is significant difference among most of methods. In particularly, all of the *p*-value between local information- and global information-based SSL models are below 0.05, as shown in the yellow shaded areas. Analogously, there is significant difference between attribute strong constraint- and weak constraint-based SSL models, as shown in the green shaded areas. These results further suggest that the global information (or attribute weak constraint)-driven SSL approaches significantly outperform the local information (or attribute strong constraint)-based SSL tasks.

## S2. Result analysis of two-task combinations in warm start scenarios

The results obtained by SSL on warm start drug predictions are shown in Table S2. Although PairDistance and SimCon generate great results, we find that PairDistance-SimCon shows the unsatisfactory performance (DDI-AUROC=0.880, and DTI-AUROC=0.942). In contrast, EdgeMask-PairDistance (DDI-AUROC=0.917, DTI-AUROC=0.958) and ClusterPre-PathClass (DDI-AUROC=0.915, DTI-AUROC= 0.956) produce relatively high results, with 2.0-5.9% higher AUROC and 2.8%-7.6% higher AUPR than other task combinations for DDI predictions. Concurrently, ClusterPre-PairDistance and EdgeMask-PathClass also produce promising results on DTI and DDI predictions. More interestingly, we find that EdgeMask-PairDistance, ClusterPre-PathClass, ClusterPre-PairDistance, and EdgeMask-PathClass are the combinations of global and local SSL tasks. In addition, PairDistance-SimCon (DDI-AUPR=0.860, DTI-AUPR=0.935) is superior to PairDistance-SimReg (DDI-AUPR= 0.847, DTI-AUPR=0.924). Similar situations are observed in comparison between SimCon and SimReg. Correspondingly, in Table S3, we summarize that the *p*-value among 11 double-modality combination models for DTI predictions. We observe that 9 out of 55 t-test experiments obtain *p*-value > 0.05. However, as shown in blue shaded area, there is significant difference among local-global combination models (i.e., EdgeMask-PairDistance, ClusterPre-PathClass, ClusterPre-PairDistance, and EdgeMask-PathClass) and other random combination models when there are the same number of modalities. We find a similar phenomenon in t-test experiments based on DDI results, as shown in Table S4. These results further suggest that the joint training of local and global SSL tasks tends to obtain higher performance than random two-task combinations when there are the same number of modalities.

## S3. Performance analysis of multi-task combinations in warm start scenarios

Although ClusterPre-PairDistance and EdgeMask-PathClass are local-global combination SSL models, as shown in Table S3 and Table S4, there is no significant difference ($p$-value>0.05) between them and other four combination models (i.e., PairDistance-PathClass, PathClass-SimCon, PairDistance-SimCon, and EdgeMask-SimReg). This may be attributed to the fact that EdgeMask-PathClass includes only single type of modality information (i.e., structures). However, PairDistance-PathClass and PathClass-SimCon capture two modalities of information (i.e., the structures and semantics of BioHNs). These results seem to indicate that we should consider the effect of different modalities. Therefore, we further design four combination models of multimodal tasks that refer to the structures, semantics and attributes of BioHNs.

We find that the top two combination models are ClusterPre-PathClass-SimReg and PairDistance-EdgeMask-SimCon in Table S2. Interestingly, they include three modalities of information, i.e., structure, semantic and attribute knowledge. Although EdgeMask-PathClass and EdgeMask-PairDistance belong to the local-global task combinations, EdgeMask-PairDistance is superior to EdgeMask-PathClass, with DDI-AUROC and DDI-AUPR improvements of approximately 3.5% and 4.3%, respectively. Similar phenomena is observed in comparison between ClusterPre-PathClass (DDI-AUROC=0.915, DDI-AUPR=0.910) and ClusterPre-PairDistance (DDI-AUROC= 0.895, DDI-AUPR=0.882). In other words, the combinations of double-modality tasks (e.g., EdgeMask-PairDistance and ClusterPre-PathClass) generate better results than the combinations of single-modality tasks (e.g., EdgeMask-PathClass and ClusterPre-PairDistance). More interestingly, we notice that PairDistance-SimReg-SimCon has one more task than PairDistance-SimReg. However, its DTI prediction performance exhibits no significant improvement. In contrast, for DDI predictions, PairDistance-SimReg-SimCon leads to a slight reduction compared to PairDistance-SimReg. This may be because PairDistance-SimReg-SimCon, in which the three tasks have only double-modality views (i.e., structure and semantic information), fails to increase the number of modalities over that used by PairDistance-SimReg and generates some noise. Similarly, ClusterPre-PairDistance-PathClass and ClusterPre-PathClass exhibit the same trend and phenomena. To further investigate the difference among 10 models, we conduct Student's t-test on the mixed results of DDI and DTI predictions. As shown in Table S5, there is significant difference ($p$-value <0.05) across various methods which include different modality information. The multi-task SSL models with multimodal information (e.g., PairDistance-EdgeMask-SimCon and ClusterPre-PathClass-SimReg) achieve greater results than other combination models. These results further suggest that combinations of multimodal tasks can achieve best performance for drug discovery.

## S4. Performance verification of MSSL2drug in cold start scenarios

For cold start scenarios, the results of single-task SSL models are shown in Fig. S1. PairDistance and PathClass yield better results than ClusterPre and EdgeMask. In particular, PathClass outperforms EdgeMask with 8.4% and 11.2% improvements in

terms of AUROC and AUPR for DDI predictions, respectively. These results are straightforward and effective demonstrations that global information-based SSL can achieve better performance than local information-based SSL. Similarly, SimCon is superior to SimReg further suggesting that the attribute weak constraint-based SSL models outperform the strong constraint-based models.

In the two-task combination scenarios, as shown in Fig. S2 and Table S6, we observe the same phenomena as those exhibited in the warm start drug scenarios: the top three models are the local-global SSL tasks, that is, ClusterPre-PathClass (DDI-AUROC=0.890, DTI-AUROC=0.923), EdgeMask-PairDistance (DDI-AUROC=0.889, DTI-AUROC=0.927), and ClusterPre-PairDistance (DDI-AUROC=0.871, DTI-AUROC=0.911). These results further certify that the local and global SSL tasks jointly guide GNNs to generate superior drug discovery predictions when there are the same number of modalities.

In addition, the results of the multimodal tasks are shown in Fig. S3. We find that the SSL task combinations containing three modalities of information, such as ClusterPre-PathClass-SimReg (DDI-AUROC=0.909, DTI-AUROC=0.948) and PairDistance-EdgeMask-SimCon (DDI-AUROC=0.909, DTI-AUROC=0.940), are superior to the task combinations capturing double-modality knowledge, such as ClusterPre-PairDistance-PathClass (DDI-AUROC=0.894, DTI-AUROC=0.924), and PairDistance-SimReg-SimCon (DDI-AUROC=0.863, DTI-AUROC=0.918). Concurrently, the double-modality SSL task combinations outperform the one-modality SSL task combinations. In other words, as the number of modalities increases, the performance of cold start predictions is improved. These results further verify the multimodal combination strategy, that is, combinations of multimodal SSL tasks can achieve state-of-the-art the prediction performance of drug discovery.

## S5. Performance validation on the external dataset

MSSL2drug is used for Luo's dataset [1] and evaluated by warm and cold start predictions with different splitting ratios. The Luo's dataset is a biomedical heterogeneous network integrating four types of nodes and six types of edges. However, in this experiment, we only choose two types of nodes (i.e., drugs and proteins) and three types of edges (i.e., drug-drug interactions, drug-target interactions, protein-protein interactions) because of the limitations of our hardware (NVIDIA Tesla V100 GPU, Memory16G). Subsequently, we filter out isolated nodes. The information of final biomedical heterogeneous network is shown in Table S7.

In warm start predictions, all the known interactions are positive samples, and an equal number of negative samples are randomly selected from unknown interactions. To evaluate the robustness of MSSL2drug, the positive and negative samples are divided into training and test set by two ratios that include 9:1 and 5:5, respectively. These splitting ratios can respectively simulate the situations that there are large and small sample labels. For warm start predictions, as shown in Fig. S4, we observe that the local-global combination models (i.e., EdgeMask-PairDistance, ClusterPre-PathClass, ClusterPre-PairDistance, and EdgeMask-PathClass) achieve better

prediction performance than other two-task combinations when there are the same number of modalities. Nevertheless, we find that PairDistance-EdgeMask-SimCon (DDI-AUROC=0.951, and DTI-AUROC=0.956) generates the best performance, as shown in Table S8. This could probably be attributed to that PairDistance-EdgeMask-SimCon integrates multimodal information including structures, semantics, and attributes in BioHN. Surprisingly, EdgeMask-PairDistance (DDI-AUROC=0.938, DDI-AUPR= 0.931) seems to be slightly higher than ClusterPre-PathClass-SimReg (DDI-AUROC= 0.932, DDI-AUPR=0.925). A possible explanation for this result is that ClusterPre-PathClass achieves a large margin improvement compared to EdgeMask-PairDistance. Even if SimReg is added into EdgeMask-PairDistance (i.e., ClusterPre-PathClass-SimReg) cannot outperform ClusterPre-PathClass. As shown in Fig. S5, these results suggest that the multimodal and local-global combination strategies still conducive to improving the performance of drug discovery on Luo's dataset.

The performance of all models on small training data (training set:test set = 5:5) is reduced when compared to warm start predictions with 9:1 ratios, as shown in Table S9. For EdgeMask-PairDistance, the performance is reduced by 1.8% and 2.3% in term of DDI-AUROC and DDI-AUPR, respectively. All of methods do not perform as well because the volume of training set is reduced. However, we find the same performance distribution, that is, multimodal combination models consistently achieve the best performance compared to other multi-task joint strategies. Furthermore, the combinations of local and global SSL tasks outperform random task combinations when there are the same number of modalities.

For cold start predictions, we randomly select 5% (or 10%) drugs and their interactions (i.e., DDIs and DTIs) to construct test sets. The rest of DDI and DTI samples are treated as training set. For 5% drugs cold start predictions, as shown in Table S10, we observe that the top three models are still PairDistance-EdgeMask-SimCon (DDI-AUROC=0.939, DTI-AUROC=0.957), EdgeMask-PairDistance (DDI-AUROC=0.927, DTI-AUROC= 0.941), and ClusterPre-PathClass-SimReg (DDI-AUROC=0.913, DTI-AUROC= 0.933). A similar phenomenon can also be observed in cold start predictions of 10% drugs, as shown in Table S11. These results further suggest that the multimodal and local-global combination strategies contribute to generating better representations on Luo's dataset, thus improving the performance of DDI and DTI predictions. In addition, these results on different splitting ratios demonstrate that MSSL2drug has the great robustness and generalization.

## S6. Performance comparisons and analysis

MSSL2drug is compared with six state-of-the-art methods, including DTINet [1], deepDTnet [2], MoleculeNet [3], KGE_NFM [4], DDIMDL [5], and DeepR2cov [6]. This is mainly attributed to two reasons. (1) Similar to MSSL2drug, these methods use multiple biomedicine networks to predict DDIs or DTIs. (2) These methods have achieved great results on various drug discovery datasets. In this experiment, we only select PairDistance-EdgeMask-SimCon to compare with baseline methods, because it

achieves best performance.

For warm start predictions on the constructed biomedical network data, PairDistance-EdgeMask-SimCon outperforms other existing methods. In particularly, PairDistance-EdgeMask-SimCon generates 2.5%~6.8% AUROC and 1.9~7.9% AUPR improvements than baselines for DDI predictions. For the cold start scenarios, we observe that MSSL2drug consistently achieves higher performance. Interestingly, KGE_NFM, that is the third best method in warm start scenarios, is reduced by 18% and 19.3% in term of AUROC and AUPR for DDI predictions in cold start scenarios, respectively. However, results of MSSL2drug slightly drop compared to warm start scenarios. Interestingly, although MoleculeNet generates poor results for warm start prediction scenarios, it becomes the third best method for cold start predictions. It is worth noting that both MoleculeNet and PairDistance-EdgeMask-SimCon are based on graph neural network. However, PairDistance-EdgeMask-SimCon attains over 5% and 3% improvement compared to MoleculeNet in term of DDI-AUROC and DTI-AUROC, respectively. This is mainly because PairDistance-EdgeMask-SimCon integrates multimodal and local-global features in biomedical heterogeneous networks.

The results of all methods on Luo dataset are shown in Table S12. For warm start prediction scenarios, we find that PairDistance-EdgeMask-SimCon still outperforms other existing methods, with 4.03% higher DDI-AUROC and 3.92% higher DTI-AUROC than the average performance of six baselines. For the cold start scenarios, we observe that MSSL2drug consistently achieves higher performance than baselines, with 7.79% higher DDI-AUROC and 7.03% higher DTI-AUROC than the average performance of six baselines. In particular, in DDI cold start predictions, KGE_NFM is reduced by 17.3% and 19.8% in term of AUROC and AUPR compared to warm start scenarios, respectively. However, for cold start predictions, results of PairDistance-EdgeMask-SimCon slightly drop compared to warm start scenarios. These results suggest that MSSL2drug can achieve higher performance on different datasets and scenarios, and is an effective self-supervised representation learning method for drug discovery predictions.

MSSL2drug and six baselines are evaluated under different splitting ratios between training and test sets, as shown in Fig. S6. We observe that the performance of all methods are reduced when there are only few training samples. However, the performance of MSSL2drug is without much fluctuation, and superior to baselines for different volume of training sets. In particularly, when the ratio of training:test sets is 3:7, KGE_NFM, DDIMDL, and DeepR2cov achieve poor results below 0.9 in terms of AUROC and AUPR for DDI and DTI predictions whereas PairDistance-EdgeMask-SimCon shows consistently the excellent performance with results close to 0.94. An interesting finding is that there are margin changes of PairDistance-EdgeMask-SimCon, ClusterPre-PathClass-SimReg, DTINet and DeepR2cov for DDI predictions when the splitting ratios is changed from 30:70 to 90:10. A possible explanation for this phenomenon is that PairDistance-EdgeMask-SimCon, ClusterPre-PathClass-SimReg, DeepR2cov, and DTINet integrate self-supervised representation learning technologies, thus relieving the dependence on the size of DDI and DTI training samples. In other words, these approaches may reach convergence point under small-scale DDI and DTI

training samples. However, there are relatively large-scale DDI samples that include 66,384 positive samples and the same number of negative samples. Therefore, even training set:test set=3:7 (i.e., the number of training samples is approximately 66,384*2*0.3=39,830), it is enough to make models reach saturation and convergence points. Surprisingly, although MoleculeNet achieves relatively lower performance, it shows high robustness. The main reasons behind this are that MoleculeNet has the small-scale parameters, thus it is easy to saturate. These results suggest that most existing methods are prone to be influenced when applying to a small dataset, while MSSL2drug can partly overcome this limitation.

In addition, we conduct the comparison between MSSL2drug with traditional graph embedding or matrix factorization, including Laplacian Eigenmaps (LE) [7], Graph Factorization (GF) [8] and DeepWalk [9]. The LE and GF use matrix factorization technique to learning representation of node. DeepWalk is a classic graph embedding approach and achieve great performance. Their hyperparameters is set as the default value. The results of MSSL2drug, LE, GF and DeepWalk are shown in Table S13. We find that PairDistance-EdgeMask-SimCon from MSSL2drug achieves the best performance compared to LE, GF and DeepWalk. Especially, MSSL2drug significantly improves the performance of DTI predictions which include more less label data, with 3~12.4% in term of AUROC and 3.1~11.8% in term of AUPR compared to traditional graph embedding or matrix factorization models. We also notice that DeepWalk is superior to LE and GF. These results suggest that MSSL2drug outperforms the traditional graph embedding and matrix factorization for DDI and DTI predictions in sparse networks.

We compare the performance of MF2A [10] and MSSL2drug on two datasets in this work. As shown in Table S14, we observe that PairDistance-EdgeMask-SimCon from MSSL2drug outperforms MF2A. More interestingly, MSSL2drug achieves 62.1% and 50.6% improvement in term of AUPR on different datasets. This phenomenon is also consistent with the results in original paper of MF2A. However, there is no the significant difference between AUROC and AUPR values of MSSL2drug. In addition, we find that the results (AUROC=0.915, AUPR=0.441) of MF2A is less than results (AUROC=0.981, AUPR=0.653) in original paper. This is mainly because the Luo's dataset in MSSL2drug is smaller and include less biomedical networks than the original dataset in [R40]. These results suggest that MSSL2drug achieves better performance on all evaluation criterias.

Finally, we test the performance of MIRACLE [11] on our dataset and Luo's small dataset, as shown in Table S15. We find that PairDistance-EdgeMask-SimCon from MSSL2drug outperforms MIRACLE on two datasets. In particularly, on Luo's dataset, PairDistance-EdgeMask-SimCon achieves 9.3% and 13.1% over MIRACLE in terms of AUROC and AUPR scores, respectively. Similar to the original study of MIRACLE, we observed that its AUPR score is less than AUROC values. However, there is no the significant difference between AUROC and AUPR values of MSSL2drug. The previous works [12-13] have suggested that AUPR can provide a better assessment compared to AUROC that is likely to be an overoptimistic metric. These results suggest that MSSL2drug achieves better performance for DDI predictions.

## S7. Drug repositioning for COVID-19

Recently, the coronavirus disease 2019 (COVID-19) has posed a global health threat. The COVID-19 patients tend to be accompanied by an excessive inflammatory response that is a main factor of death and indicates a poor prognosis in COVID-19 [14]. Subsequently, a large number of clinical data has suggested that interleukin (IL)-6 plays key roles in the inflammatory storms. Based on this intuition, some studies found that the inhibitors targeting IL-6 are likely to become promising agents for the treatment of COVID-19 patients [15-16]. Therefore, PairDistance-EdgeMask-SimCon, which achieves the best performance in MSSL2drug, is applied to drug repositioning for COVID-19, which aims to discover agents related to IL-6 for blocking the excessive inflammatory response in patients. To be specific, we predict the confidence score of interaction between each drug and IL-6 by multilayer perceptron, and then top-10 drugs with the highest scores are selected as potential anti-inflammatory agents for COVID-19 patients. Finally, we utilize the knowledge from PubMed publications and clinical reports to explain the anti-inflammatory effects of candidate drugs. It is noted that in the construed BioHN, there is no interaction between IL-6 and all drugs. In other words, there is no the known drugs interacting with IL-6. Therefore, top-10 drugs are novel predictions for IL-6. Concurrently, the top-10 drug-IL-6 interactions are not used in training the PairDistance-EdgeMask-SimCon representation.

Based on PubMed publications and clinical studies, we find that nine out of ten drugs predicted by MSSL2drug can inhibit the release of IL-6 and reduce inflammatory response, as shown in column 4 in Table S16. Triclosan inhibits the release of IL-6 by decreasing mRNA levels. Tazarotene decreases the expression of IL-6 to exert the anti-inflammatory effects. Bosutinib inhibits the production of IL-6 and tumor necrosis factor (TNF)-α induced by Lipopolysaccharide stimulation. Vandetanib significantly inhibits the levels of IL-6, IL-10, and TNF-α, and reduces inflammatory cell infiltrates in the lungs of a COVID-19 infection mouses. Pazopanib can relieve the negative prognostic effect of high IL-6. More importantly, we find that ruxolitinib (ClinicalTrials.gov ID: NCT04414098, and NCT04377620) and chlorhexidine (NCT04941131) have been determined in clinical test against COVID-19. Therefore, these drugs can inhibit inflammatory responses and should be taken into consideration in clinical studies against COVID-19.

On the other hand, we conduct molecular docking [17], molecular dynamics (MD) simulations [18-19], and surface plasmon resonance (SPR) [20] to explore physical interactions between the predicted drugs and IL-6.

- **Molecular docking:** The docking program AutoDock4.2 [17] was used to model the molecular interactions between IL-6 and each drug. The three-dimensional structure of IL-6 are extracted from the Protein Data Bank (PDB ID: 4CNI). The grid box is a cubed with 12 Å sides centered on selected amino acid residues. Others parameters are set as default values. During docking process, the final docking conformations are selected by using cluster algorithms based on energy and the root mean square deviation (RMSD) values. To be specific, for a given

drug-IL-6 pair, at least 50 docking poses are obtained for position clustering. These docking poses are clustered by position differentiation at 0.5Å RMSD. First, all conformations are sorted by docked energy and the conformation with lowest energy is treated as the center of the first cluster. Second, the second conformation is compared to the first. If its RMSD value is below 0.5Å, it is added to the first cluster. Otherwise, it becomes the first member of a new cluster. This process is repeated until each conformation are divided into a group in which members are most similar to it. Finally, the center conformation within the biggest position cluster is selected as the final conformation.

The docking affinity in Fig. S7(a) suggest that there is the good binding ability between 10 small molecules and IL-6. In addition, we find that the protein binding pocket are located at hydrophobic regions that contain more hydrophobic amino acids, as shown in Fig. S7(b). These amino acids containing relatively many functional groups tend to form the hydrophobic interaction with small molecules. In particular, vandetanib, pazopanib, and chlorhexidine may have stronger binding interactions compared to other drugs, because of corresponding to relatively lower energies. However, chlorhexidine corresponds to lowest confidence score predicted by multilayer perceptron. Therefore, vandetanib and pazopanib are conducted molecular dynamics simulations to further explore physical interactions between them and IL-6.

- **Molecular dynamics simulations:** In this experiment, we use AMBER16 [18-19] to conduct molecular dynamics simulations. The results from molecular docking are set as the initial structure. Each complex is placed into a box that is delimited by at least 10 Å from any heavy atom of the protein. The protein-ligand complex is then filled with TIP3P water molecules, and $Na^+$ ions are added to neutralize net charge of systems. Finally, one 100ns MD simulation is run for each complex, the binding free energy is computed for each snapshot and averaged using the MM-GBSA module.

To evaluate the fluctuation equilibrium and structural stability of the protein, we monitor the root mean square deviation (RMSD) of the backbone atoms relative to the initial structure during MD simulations, as shown in Fig. S8. We find that the pazopanib-IL-6 system and vandetanib-IL-6 system reach stability after 20ns and 40ns, respectively. Interestingly, RMSD of pazopanib-IL-6 system and vandetanib-IL-6 system stabilize at 4.6~4.8Å and 4.8~5Å, respectively. These results indicate that two complex systems are stable and their fluctuations is similar.

For the pazopanib-IL-6 complex model in Fig. S9(a), the binding is mediated by direct hydrogen bonds from Leu151 to pazopanib. The same atoms in pazopanib form hydrogen bond with Leu147 of IL-6. The corresponding binding free energy values are -9.61 kcal/mol for pazopanib. Fig. S9(b) shows that vandetanib mainly binds to Asn63 and Tyr97 in IL-6 through two hydrogen bonds. The corresponding binding free energy values are -11.81 kcal/mol for vandetanib. These results suggest that pazopanib and vandetanib may be able to form physical interactions with IL-6.

- **Surface plasmon resonance (SPR):** We further validate physical interactions between these two molecules (i.e., vandetanib and pazopanib) and interleukin (IL)-6 through surface plasmon resonance (SPR) [20] which has been used for detecting

protein-ligand interactions. SPR is conducted via Nicoya system. The mixture of 400 mM EDC and 100 mM NHS are used for activating COOH chip. The recombinant human IL-6 protein are diluted to 40 μg/mL through immobilization buffer (Sodium Acetate, pH5.0) and then these diluted proteins are injected to the chip at a flow rate of 20 μL/min. The chip is deactivated by injecting 1M Ethanolamine hydrochloride at a flow rate of 20 μL/min for 240s. Analogously, the micromolecule is diluted to the different concentrations by the same running buffer, as shown in Fig. S10. The fixed concentrations of micromolecule is injected into a channel flow cell at a flow rate of 20 μL/min and keep an association of 240s, followed by 360s dissociation. Note that the association and dissociation process are performed in the Running Buffer PBS (pH7.4). Glycine-HCl as regeneration buffer is injected on sensor chip surface at a flow rate of 150 μL/min to remove any bound analyte. We repeat perform the above dilution, association, dissociation and regeneration procedures according to analyte concentrations in ascending order.

The different concentrations of vandetanib and pazopanib are injected on the chip surface, resulting in five response curves as shown in Fig. S10. According to the obtained association and dissociation rates, we achieve the equilibrium dissociation constants ($K_D$) of the selected drugs and recombinant human IL-6. We find that vandetanib ($K_D$=28.6μM) and pazopanib ($K_D$=20.7μM) can bind to Recombinant Human IL-6 with high affinity. In comparison, the affinity of pazopanib and recombinant human IL-6 is stronger than that of vandetanib. These results demonstrate that MSSL2drug can identify new physical drug-target interactions.

In this work, it is noted that the drug-protein interaction (DTI) networks are collected from the DrugBank [21], PharmGKB [22], and Therapeutic Target database (TTD) [23] where there may be DTI samples in which drugs indirectly interact with proteins by regulate the related pathways. Unfortunately, it is different to completely eliminate these noises. In other words, there may be the indirect relationships between drugs and proteins in training samples. These samples with indirect relationships are treated as drug-target physical interactions by deep learning models. Naturally, drug-target interactions predicted by deep learning may include some indirect relationships between drugs and proteins. Therefore, these predicted drugs may exert inhibitory action on IL-6 by regulating down-stream pathway. However, it is important and interesting to validate whether the predicted drugs physically bind to IL-6. These results and analyses further emphasis that it is important and interesting to validate whether there are the indirect relationship or physical interactions between these drugs and IL-6 by standard and systematic experiments. In addition, all predicted drugs must be validated in preclinical models experiments and randomized clinical trials before being used in COVID-19 patients.

## S8. Impact of key components on SSL tasks

### S8.1 Selection of centrality measurements in ClusterPre

It is interesting and important to discuss the impact of centrality measurements on MSSL2drug. There are popular centrality measurements, including degree centrality,

eigenvector centrality [24], clustering coefficient [25], closeness centrality [26], and betweenness centrality [27]. Nevertheless, the closeness and betweenness centrality are calculated by shortest paths between target node and all of the other nodes in networks. Time complexity of the closeness and betweenness centrality is very high, and it is not easy to obtain the closeness and betweenness centrality for the large real networks. Therefore, we use graph attention network (GAT) models to predict the each node degree centrality (DegreePre) and eigenvector centrality (EigenvectorPre) for extracting representations that preserve the local structure information in BioHNs and are applied to drug discovery.

As shown in Table S17, we observe that DegreePre (AUPR=0.651) and EigenvectorPre (AUPR=0.596) generate the poor prediction results for DDI predictions. For DTI predictions, ClusterPre outperforms DegreePre and EigenvectorPre, with 8.35% higher AUROC and 12% higher AUPR average values. A possible explanation for these results is that the degree and eigenvector centralities only consider the importance and distribution of neighboring nodes. However, the degree and eigenvector centralities fail to capture the triangle (loops of order 3) structures in networks, i.e., if node *i* is connected to nodes *j* and *k*, there is a high probability of nodes *j* and *k* being connected [28]. Nevertheless, the clustering coefficients are not only extract the distribution of neighboring nodes, but also the triangle (loops of order 3) structures in networks.

**S8.2 Division of "major" class in PairDistance**

Recently, $S^2$GRL [29] suggests that the distinctions between node pairs with long distance ($d_{ij}$>4) are relatively vague, and divided into one "major" class. In $S^2$GRL, it can be found that clearly distinguishing 1-hop, 2-hop, and 3-hop nodes into three classes is beneficial to improving the quality of representations, while further differentiating 4-hop and higher-hop node pairs would degrade the performance. Therefore, in pairwise distance classification (PairDistance), we also divide 4-hop and higher-hop node pairs into a "major" class. In addition, according to your opinions, we have added the experiments to investigate the effect of "major" class selection on the prediction performance of PairDistance. As shown in Table S18, we find a similar phenomenon in $S^2$GRL, that is, dividing 4-hop and higher-hop node pairs into a "major" class achieves better performance compared to 3-hop and 5-hop. These results indicate that the distance of "major" class should be set to 4 to optimize the perfection performance of drug discovery.

**S8.3 Lengths of meta path in PathClass**

The length selection of meta path is an interesting and important question. Fortunately, previous studies have suggested that the long meta-paths may reduce the quality, while short meta paths contributes to link predictions [30-31]. Therefore, we further investigate the effect of meta path length on the performance of PathClass. As shown in Table S19, we summarize the predictive results of PairDistance with meta path of different lengths for warm start drug discovery. We find that meta path with lengths 4 achieve best performance, with approximately 1% higher the average DDI-AUPR and 6% higher the average DTI-AUPR than other length meta path for warm-start

predictions. Meta paths with lengths 5 increase the time complexity, while it generates lower performance. These results suggest that selecting meta paths with lengths 4 is contribute to the performance of PathClass compared to other length paths.

**S8.4 Selection of similarity measurement in SimCon**

In the task of similarity contrast **(**SimCon**)**, cosine similarity is used to measure the similarity between two embedding representations. Therefore, we replace cosine similarity by Euclidean Distance (termed SimCon-ED) to verify the influence of different similarity measurements on SimCon. Table S20 summarizes the results of SimCon and SimCon-ED for warm start predictions. We observed that SimCon-ED (DDI-AUPR=0.805) achieves higher performance than SimCon (DDI-AUPR=0.783). In contrast, SimCon obtains 1% higher AUROC and 1.6% AUPR than SimCon-ED for DTI predictions. These results indicate that the different similarity measurements bring the marginal improvements or reductions to SimCon. A possible explanation for this result is that cosine function and Euclidean Distance have the same level of ability to measure the similarity distributions among different nodes. Concurrently, SimCon only requires to distinguish the similarity distributions between node pairs, and reduces the dependence on similarity measurements.

**S8.5 Ablation result analysis of PairDistance-EdgeMask-SimCon**

We conduct ablation analysis for PairDistance-EdgeMask-SimCon. To be specific, PairDistance-EdgeMask-SimCon is transformed into EdgeMask-PairDistance, PairDistance-SimCon, and EdgeMask-SimCon by removing different tasks, respectively. The results of ablation experiment related to PairDistance-EdgeMask-SimCon are shown in Table S21. We find that EdgeMask-PairDistance achieves higher results than PairDistance-SimCon and EdgeMask-SimCon for both warm start predictions and cold start predictions. A possible explanation for these results is that EdgeMask-PairDistance captures local and global feature in BioHNs. Interestingly, EdgeMask-SimCon (AUROC=0.878, AUPR=0.871) outperforms PairDistance-SimCon (AUROC=0.858, AUPR=0.837) for cold start DDI predictions. In contrast, EdgeMask-SimCon is lower than PairDistance-SimCon for cold start DTI predictions. More importantly, we find that PairDistance-EdgeMask-SimCon achieves best performance compared to EdgeMask-PairDistance, PairDistance-SimCon, and EdgeMask-SimCon. These results suggest that PairDistance-EdgeMask-SimCon integrating multimodal and local-global task is beneficial to improve performance of drug discovery, in which the contribution of SimCon is relatively lower than EdgeMask and PairDistance to some extent.

## S9. Contributions of key components in multi-task learning

MSSL2drug integrates adversarial training- and orthogonality constraint-based multi-task learning mechanism for improving representation quality. Therefore, we have added the experiments to evaluate the contribution of the adversarial training strategy and orthogonality constraint mechanism to MSSL2drug, respectively. In this

experiment, MSSL2drug is transformed into ADL and ORC patterns. To be specific, ADL pattern denotes that MSSL2drug only uses adversarial training-based multi-task framework to learning representation vectors of nodes, as shown in Fig. S11. In contrast, ORC pattern denotes that MSSL2drug only uses an orthogonality constraint-based multi-task learning strategy, as shown in Fig. S12.

In this experiment, we select ClusterPre-PathClass and PairDistance-EdgeMask-SimCon to conduct this experiment because they achieve best performance in two-task and multi-task combinations, respectively. Therefore, ClusterPre-PathClass is transformed into ADL and ORC patterns, termed CP-ADL and CP-ORC. Analogously, PairDistance-EdgeMask-SimCon is transformed into PES-ADL and PES-ORC. The results of all models for warm start predictions are shown in Table S22. An interesting finding is that CP-ADL obtains the slight improvement compared to CP-ORC for DTI predictions, while CP-ORC (DDI-AUROC=0.915) is superior to CP-ADL (DDI-AUROC=0.879) by a large margin. A similar phenomenon can also be observed in PES-ADL and PES-ORC. More importantly, we find that MSS2drug achieves best performance compared to ADL and ORC models. These results suggest that MSS2drug integrating ADL and ORC is beneficial to improve performance, in which the contribution of ORC is higher than ADL to some extent.

## S10. High-quality representations

If the representations perfectly keep the characteristic of networks, the traditional machine learning models can generate great results. To further demonstrate the performance of MSSL2drug, the representations are fed into Random forest (RF) [32] and support vector machine (SVM) [33] for DDI and DTI predictions. In this experiment, we only select PairDistance-EdgeMask-SimCon, because it achieves best performance in 15 multi-task combinations. The representations from PairDistance-EdgeMask-SimCon are fed into SVM and RF for drug discovery, termed PES-SVM and PES-RF, respectively. As shown in Table S23, we observe that there is no significant difference between PES-SVM and PairDistance-EdgeMask-SimCon for DDI and DTI predictions. However, an interesting finding is that PES-RF (AUROC= 0.987) achieves the higher results than PairDistance-EdgeMask-SimCon (AUROC= 0.939) for DDI predictions. These results suggest that MSSL2drug can generate the high-quality representations that can keep the inherent nature of biomedical heterogeneous networks, thus improving the performance of drug discovery.

## S11. Test set contamination studies

In this experiment, PairDistance-EdgeMask-SimCon (PES) is transformed into PESReM through four key steps. (1) We extract 10% of DTIs as the positive samples of test set and remove them from BioHNs while ensuring no node is isolated. Concurrently, all DTIs in the residual network are treated as positive samples of training set. (2) An equal number of negative samples are selected randomly and added into training set and test set, respectively. (3) The residual network is used for self-

supervised learning to obtain low-dimension representations. (4) All drug representations as features are fed into multilayer perceptron for DTI predictions.

As shown in Fig. S13, we find that the performance of PESReM is not changed much compared to PairDistance-EdgeMask-SimCon (PES). Interestingly, although the AUROC score of PESReM is 1% worse than that of PairDistance-EdgeMask-SimCon, PESReM achieves approximately 1% improvement in term of AUPR for cold start predictions. More importantly, recent studies also made a similar finding [34-35]. These results would suggest that the data contamination in SSL does not cause significant change for the performance of MSSL2drug. This is mainly attributed to two reasons. On the one hand, the DTIs in test set are only used for GNN message passing rather than the label of self-supervised learning. For example, in reading comprehension task, GNNs only see the passage during self-supervised training but not see the questions and answers, this does not constitute cheating and bring an advantage. In contrast, previous studies have suggested that removing all test data in downstream tasks maybe lead to overly punishment for false positive samples [34]. On the other hand, a large amount of GNN models uniformly sample a part of neighbors for each node during training instead of using full neighborhood sets, in order to keep the high precision and computational efficiency [36-38]. These studies further indicate that there is no significant influence on drug discovery when removing a certain number of edges. In summary, MSSL2drug is relatively insensitive to data contamination. However, in the future, we will more synthetically analyze the data contamination on the performance of downstream tasks.

## S12. Comparisons of run-time and parameter sizes

We compared the run-time and parameter sizes of PairDistance-EdgeMask-SimCon with baselines. All methods are run on NVIDIA Tesla V100 GPU (Memory16G). As shown in Table S24, the parameter sizes of MSSL2drug (i.e., PairDistance-EdgeMask-SimCon) is smallest compared to most of baselines. However, the run-time of MSSL2drug is relatively higher than DDIMDL, DTINet, MoleculeNet, and deepDTnet. This is mainly because MSSL2drug integrates multiple GATs that are trained by large-scale samples with pretext labels. A similar issue can also be observed in DeepR2cov that is a SSL technique.

## S13. Small world property of biological networks

Here, we use the statistical analysis to evidence that the constructed biological network is a small-world network. In Table S25, we list the shortest path length and the number of node pairs. We find that there are only 170 node pairs in which the shortest path length is over 5. In addition, we observe the fat-tailed degree distributions (https://en.wikipedia.org/wiki/Fat-tailed_distribution) in Fig. S14, and average degree with 73. This phenomenon is consistent with the other biological networks in [39-40]. These results suggest that the biological network has the small world property.

## S14. Basic self-supervised tasks on BioHNs

### S14.1 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 node to capture the local structure information in BioHNs. Formally, we adopt the mean squared error as the loss function of ClusterPre:

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

where $\theta$ 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.

### S14.2 Pairwise distance classifications (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 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 GATs 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:

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

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.

**S14.3 Edge type masked predictions (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 the four classification problems. Similar to PairDistance, we also adopt the cross entropy loss function in EdgeMask. In the construed BioHN, there are five types of edges (e.g., drug-drug interactions, drug-protein interactions, drug-disease associations, protein-protein interactions, and protein-disease associations). However, previous studies [31, 41-42] have suggested the schema of BioHNs as shown in Fig.S15, where both drug-drug relationships and protein-protein relationships are treated as 'interaction'. Concurrently, drug-drug networks and protein-protein networks are homogeneous networks. Inspired by these works, in MSSL2drug, drug-drug relationships and protein-protein relationships are treated as the same types of semantic.

**S14.4 Bio-meta path classifications (PathClass)**

For a given heterogeneous network, a meta path is defined as a sequence in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, which describes a composite relations between $A_1$ and $A_{l+1}$, where $A_l$ is node types, and $R_l$ represents edge type between nodes. In this work, the length of a meta path is defined as the total number of nodes in current meta path. A meta path integrates the semantic relationships in BioHNs. For example, "$\text{protein} \rightarrow \text{disease} \rightarrow \text{drug}$" 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 [41, 43] within BioHNs, we design 16 types of meta paths as shown in Table S26, where the first and last objects are drugs and proteins, respectively. This is mainly because drugs and proteins are interconnected with other entities by more edges, as shown in Table S28. 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.

In addition, we generate false path instances by randomly replacing some nodes in true path instances. There are three key points in the false path generation, as shown in Fig. S16, (1) For a given true path instance, it has 6.25% (i.e., 1/16) chance being generated a false path instance to avoid the label imbalance questions. (2) There is no relationship between the permutation nodes and the context nodes in current paths. (3) The number of replaced nodes is less than four that is the length of meta paths. In other words, there are at most two true edges in a false path. Therefore, these false paths can improve the generalization and robustness of PathClass when compared to the false paths with two true edges or one true edge. Nevertheless, it is interesting and important to verify what happens if the false paths have two true edges or one true edge.

**S14.5 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 values of node pairs in the original feature space. Formally, the objective function employs the mean squared error and is given as follows:

$$L_{simReg}(\theta) = \frac{1}{|S|} \sum_{(i,j)\in S} \ell'\left(\delta\left(\left\langle f_\theta(i), f_\theta(j)\right\rangle\right), Y_{sim_{i,j}}\right) \tag{4}$$

where $\theta$ is the parameters of a graph neural network $f_\theta(\cdot)$, $f_\theta(i)$ and $f_\theta(j)$ denote the embeddings of node $i$ and $j$. $\langle\cdot,\cdot\rangle$ is an operation that concatenates two nodes, $\ell'(\cdot,\cdot)$ represents the mean square error (MSE) loss function, and $\delta(\cdot)$ represents the Sigmoid function. $Y_{sim_{i,j}}$ is the similarity value between two nodes in the original feature space. $S$ and $|S|$ denote the selected set of node pairs $(i, j)$ and the total number of node pairs $(i, j)$, respectively.

In this work, we use different property similarity measurements according to various types of nodes. **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 [44] 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

[45] 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 [46] 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 [47-48], 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. $SIM(G_1, G_2)$ is calculated as follows:

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

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:

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

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.

## S15. Hyperparameter setting

In SSL stage, we empirically consider the selections of optimization algorithm, weight initialization, and activation functions. The number of hidden layers, hidden units, head attentions, and batch size is selected according to the limitations of hardware (NVIDIA Tesla V100 GPU, Memory16G). The selections L2 regularization, learning rate, and epoch size are slightly tuning according to the performance of single tasks. Finally, we adopt the Glorot initialization [49], the Adam optimizer [50] with a learning rate [1e-5,1e-2], L2 regularization 5e-4, 8 hidden units and 8 head attentions. The number of epoch is set to 30. In supervised drug discovery, an MLP with three fully connected layers (including an input layer, a hidden layer and an output layer) is used to decode the embedding vectors. The size of the input layer depends on the dimensionality of the

input feature, and the size of the hidden layer is set to 64. We also use the Adam optimizer to train the MLP for 30 epochs with batch size 128. For the learning rate, we select 10 points that are equidistant from the interval [5e-4, 5e-1].

The representation dimension can directly affect the performance and time efficiency of self-supervised representation learning approaches. Therefore, several researchers have investigated the influence of different embedding dimensions in various SSL methods [51-55]. These studies found that the performance first increases when the embedding dimensionality increases. However, the performance tends to saturate or reduce when the dimension reaches to a threshold that is often close to 100. This is intuitive since higher dimensionality can encode more useful information, while too large value may lead to over-fitting phenomenon and excessive time complexity. In addition, the algorithm models with too large-scale parameters cannot run due to the limitations of hardware implementation. Therefore, the representation dimension of each private or shared GAT models is set to 64 (i.e., 8 hidden units $\times$8 head attentions) that is largest values under our hardware.

## S16. Baselines

In this work, we compare MSSL2drug with six state-of-the-art methods, including deepDTnet [2], MoleculeNet [3], KGE_NFM [4], DTINet [1], DDIMDL [5], and DeepR2cov [6], Laplacian Eigenmaps (LE) [7], Graph Factorization (GF) [8] and DeepWalk [9], MF2A [10], and MIRACLE [11]. We only select PairDistance-EdgeMask-SimCon to compare with baseline methods, because it achieves best performance in 15 multi-task combinations. The hyperparameters of all baseline methods are set according to the guidelines in [1-6], in which use Adam as the optimization algorithm and cross entropy as loss function to train the deep learning networks.

- deepDTnet: it proposes a deep neural network to learn low-dimensional graph representations for both drugs and targets. In addition, these representation vectors are fed into Positive-Unlabeled-matrix completion models for target identification among known drugs. In deepDTnet, we set the number of random walk steps T= 3, the biased value $\alpha$=0.5 and regulation parameter $\lambda$=0.1.
- MoleculeNet: this is a benchmark for molecular machine learning which includes multiple methods. Here, we select a graph convolutional model to compare with MSSL2drug, which is specifically designed for network (graph) structure data and is widely applied to drug discovery. The graph convolutional models consist of a graph convolutional layer, a batch normalization layer, followed by a fully-connected dense layer to predict DDIs or DTIs. The number of units in graph convolutional layers and fully-connected layers are set to 64 and 128, respectively.
- KGE_NFM: This model firstly learns a low-dimensional representations based on the knowledge graphs, and then integrates the multimodal information via a neural factorization machine for DTI predictions. In KGE_NFM, we set weight of regularization loss as 1e-5, regularizer norm as 3, the layer number and units in each layer of deep net as 128, L2 Regularizer strength applied to DNN as 1e-5.

- DTINet: It applies a diffusion component analysis algorithm to learning the low-dimensional vectors, which capture the topological properties in biomedical heterogeneous networks. Based on these representations, DTINet makes DTI predictions via an inductive matrix completion model [56]. In DTINet, we set restart probability to 0.8, dimensionality of drug and protein representations to 100 and 400, respectively.
- DDIMDL: It is a multimodal deep learning framework for DDI predictions. DDIMDL firstly constructs deep neural network sub-models based on diverse drug features, and then combines all sub-models to learn cross-modality representations for predicting DDIs. In sub-models of DDIMDL, we adopt three hidden layers, numbers of whose nodes are 512, 256 and 1, and set the dropout rate to 0.3.
- DeepR2cov: This study proposes a meta path-based deep representation model to learn low-dimensional embedding vectors. DeepR2cov also uses inductive matrix completion model [43] for bio-link predictions including disease-gene associations, DTIs and drug-side effect associations. More importantly, DeepR2cov predicts 22 agents to accelerate treatment of the inflammatory responses in COVID-19 patients. In DeepR2cov, the number of Transformer blocks, the hidden sizes, the number of self-attention heads, and batch size are set to 12, 768, 12, and 256, respectively.
- Laplacian Eigenmaps: This is a graph Laplacian-based geometrically motivated algorithm for representing the high-dimensional graph data.
- Graph Factorization: It is a streaming graph embedding that factorizes a graph so as to minimize the number of neighboring vertices rather than all possible edges.
- DeepWalk: The node sequences from random walks in graphs are treated as sentences and fed into SkipGram model [57] to learning node representations. The length of random walk is set to 32. The window size is 10.
- MF2A: This model use matrix factorization to optimize AUPR and AUROC for drug-target prediction. MF2A adopt a local interaction consistency to incorporate drug and target similarity information. The hyperparameters of MF2A is set as the default value in [10].
- MIRACLE: this is a multi-view graph contrastive representation learning for drug-drug interaction prediction. MIRACLE is able to capture inter-view molecule structure intra-view interactions between molecules. The number of the hidden states and GCN layers are set to 256 and 3, respectively. The ratio of dropout is 0.3.

## Supplementary Figures

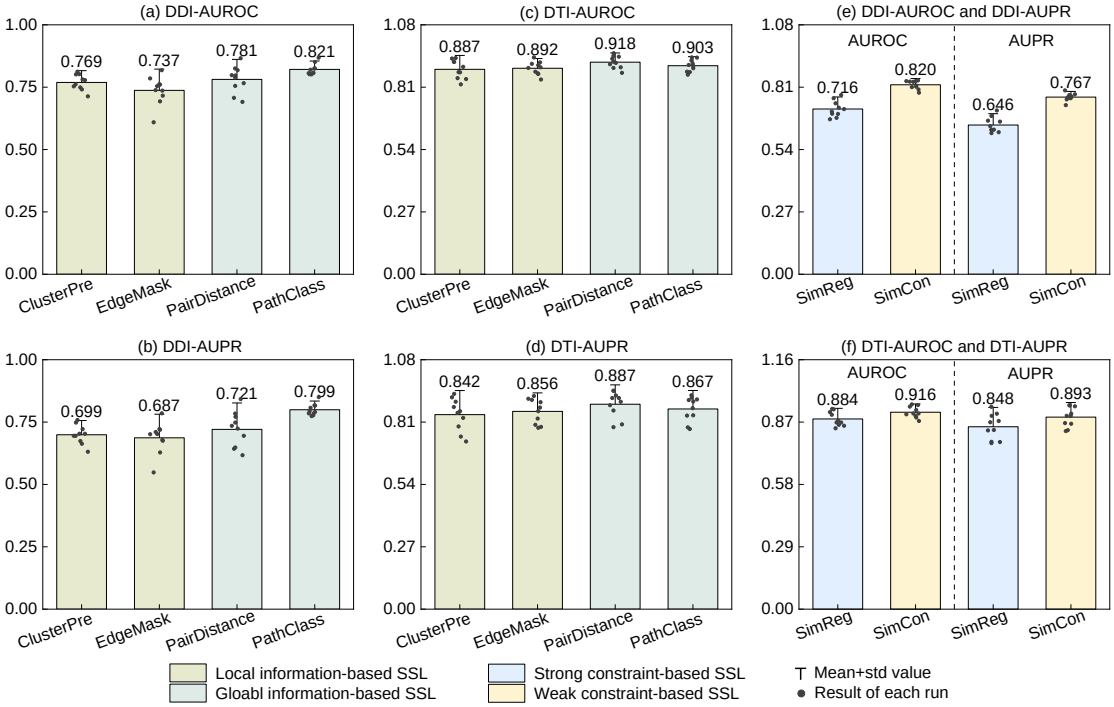



**Figure S1.** The results of single SSL tasks for cold start predictions where mean and std values denote average and standard deviation values that are calculated across ten results.

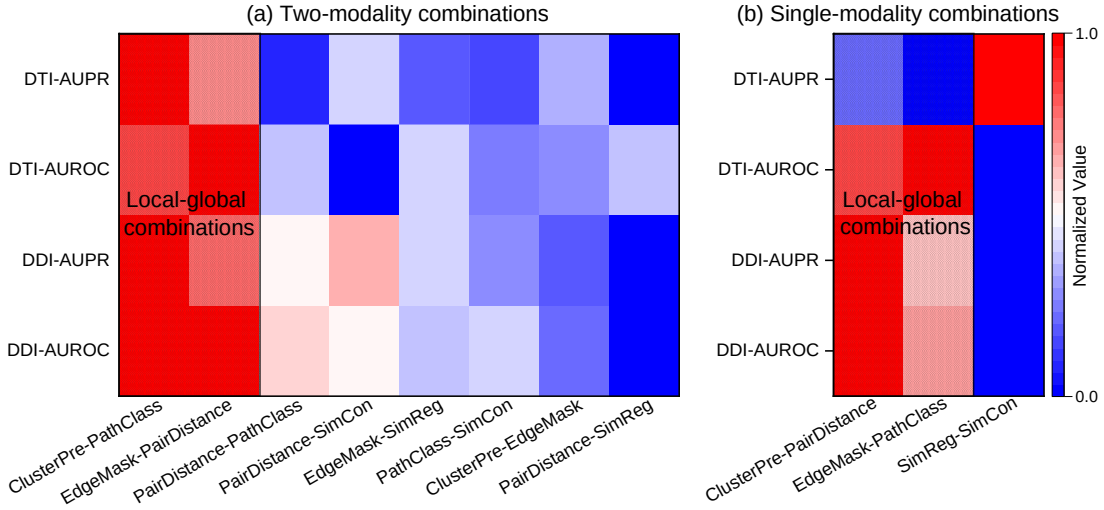



**Figure S2.** Heatmap of two-task combinations for cold start predictions where the results are normalized to [0,1] along the x-axis by Min-Max normalization technique. The redder (bluer) squares denote the greater (smaller) value. The shaded area denotes the combinations of global and local SSL tasks.

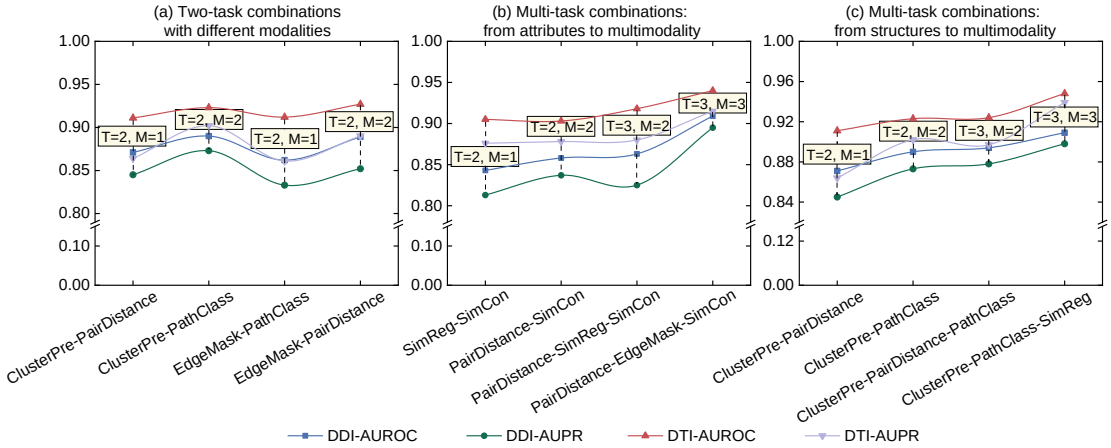



**Figure S3.** The results obtained with multimodal SSL tasks for cold start drug discovery, where 'T' and 'M' denote the total number of tasks and modalities in each multi-task combination, respectively.

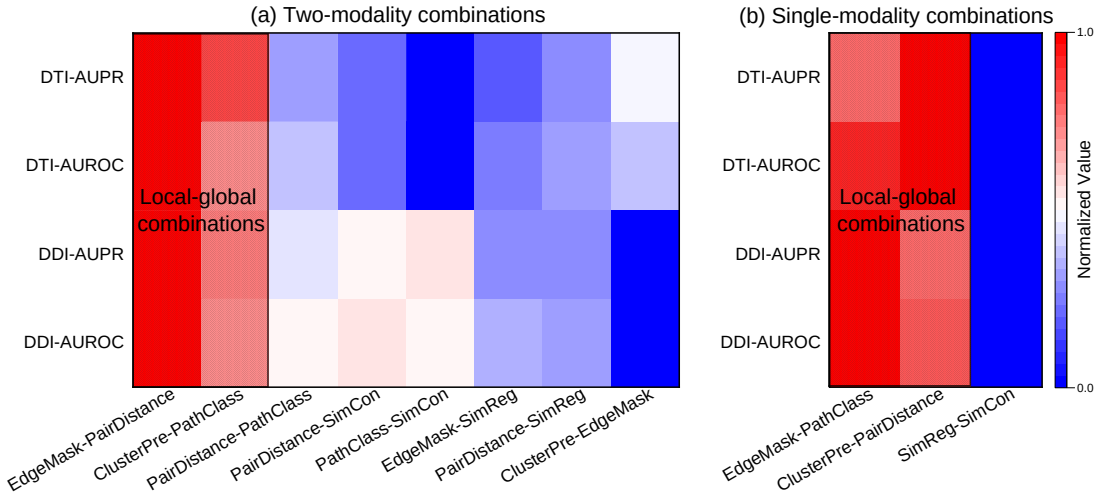



**Figure S4.** Heatmap of two-task combinations on Luo's dataset for warm start predictions where the results are normalized to [0,1] along the x-axis by Min-Max normalization technique. The redder (bluer) squares denote the greater (smaller) value. The shaded area denotes the combinations of global and local SSL tasks.

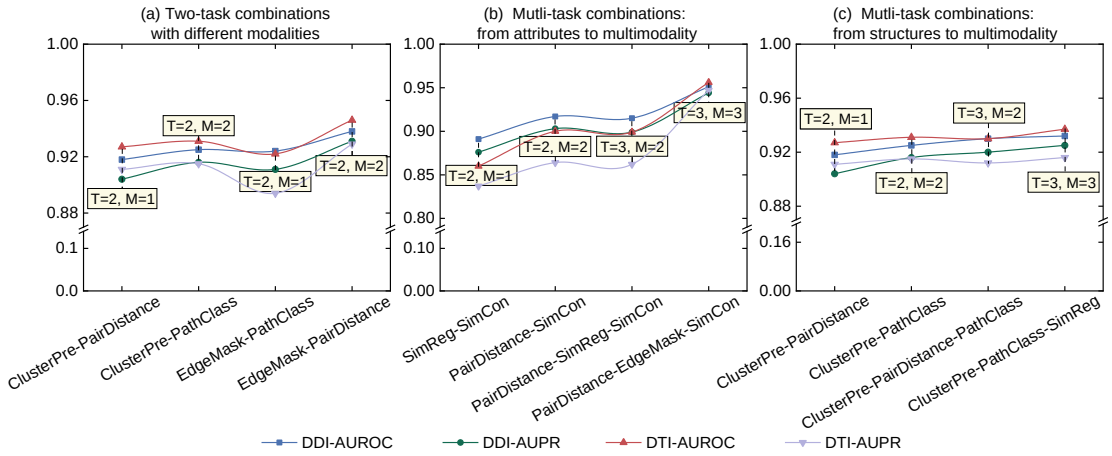



**Figure S5.** The results obtained with multimodal SSL tasks on Luo's dataset for warm start drug discovery, where 'T' and 'M' denote the total number of tasks and modalities in each multi-task combination, respectively.

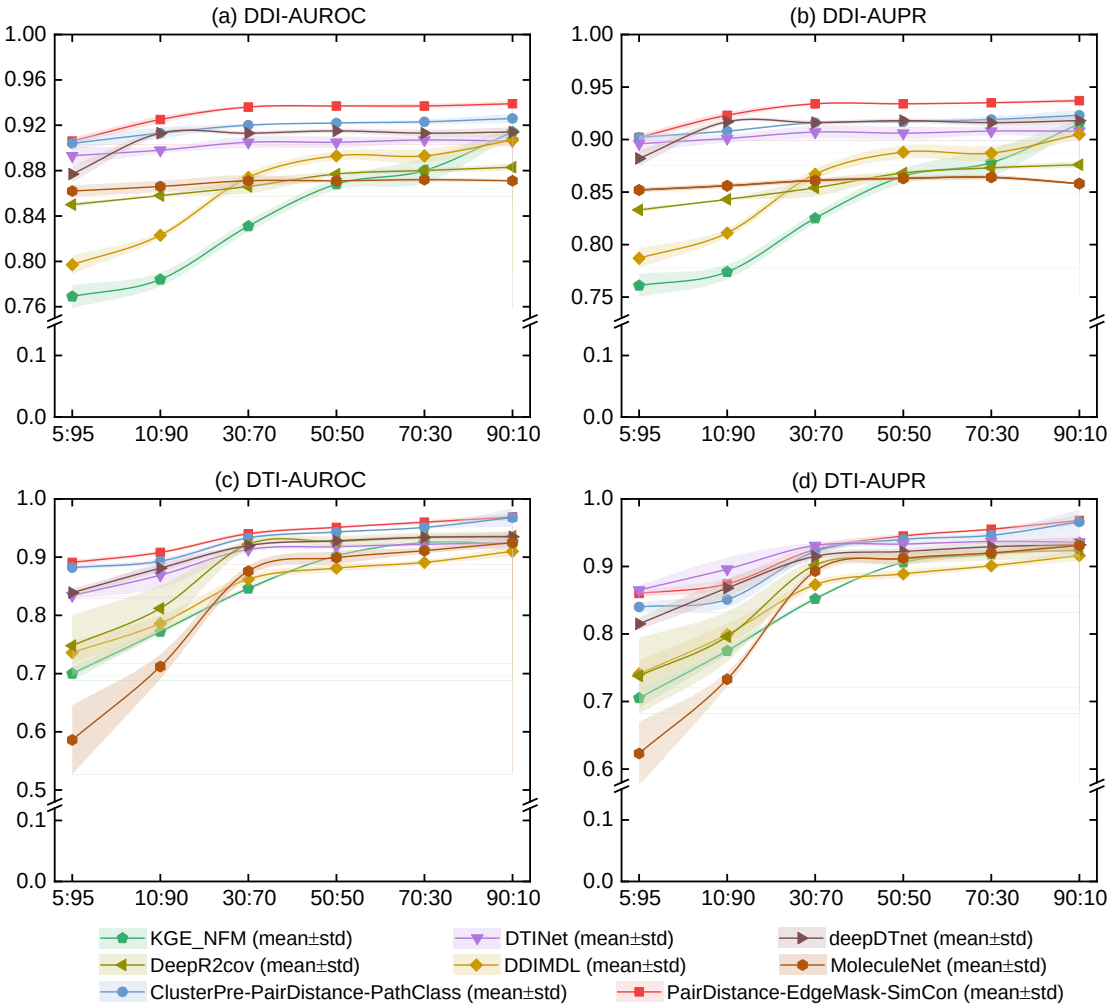
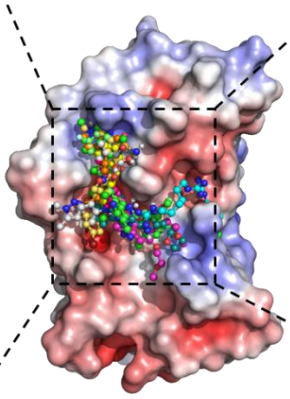


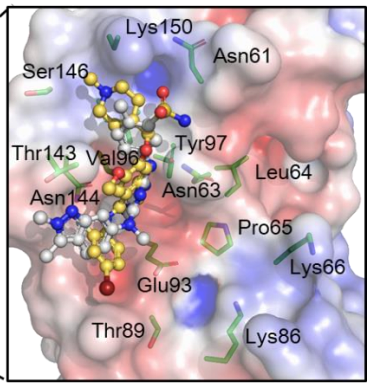

**Figure S6.** Performance of all methods on different splitting ratios, where the x-axis represents the ratios between training sets and test sets. The mean and std values denote average and standard deviation values that are calculated across ten results.

| Drug name | Energy (kcal/mol) |
|---|---|
| Triclosan | -5.9 |
| Hexachlorophene | -5.7 |
| Ruxolitinib | -6.3 |
| Tazarotene | -6.4 |
| Bosutinib | -6.7 |
| **Vandetanib** | **-7.0** |
| Sunitinib | -6.7 |
| Erlotinib | -6.1 |
| **Pazopanib** | **-7.4** |
| Chlorhexidine | -7.8 |

(a) Docking affinity



(b) Protein surface




(c) Two small molecules in protein binding pocket

**Figure S7.** Molecular docking results for small molecules binding to IL-6. (a) Docking affinity between 10 small molecules and IL-6. (b) Protein surface colored according to the interpolated atomic charge. The hydrophilic and hydrophobic regions are denoted by blue and red color, respectively. The different colors ball-and-sticks denote the different small molecules. (c) Two small molecules with in protein binding pocket surrounding some residues where the yellow and light grey ball-and-sticks denote vandetanib and pazopanib, respectively.

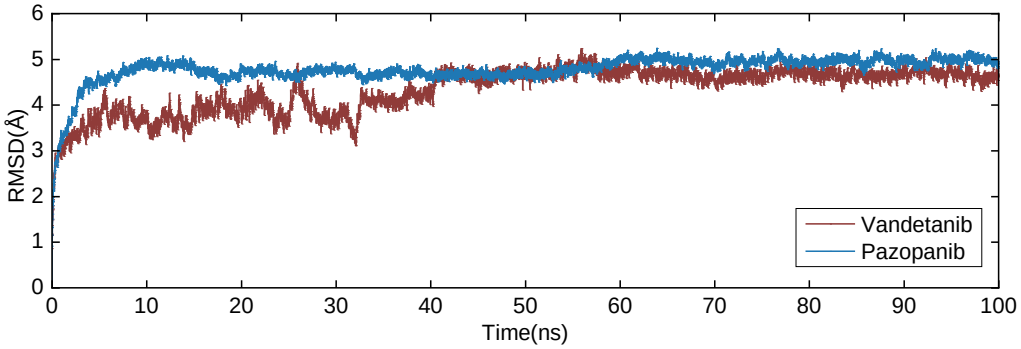
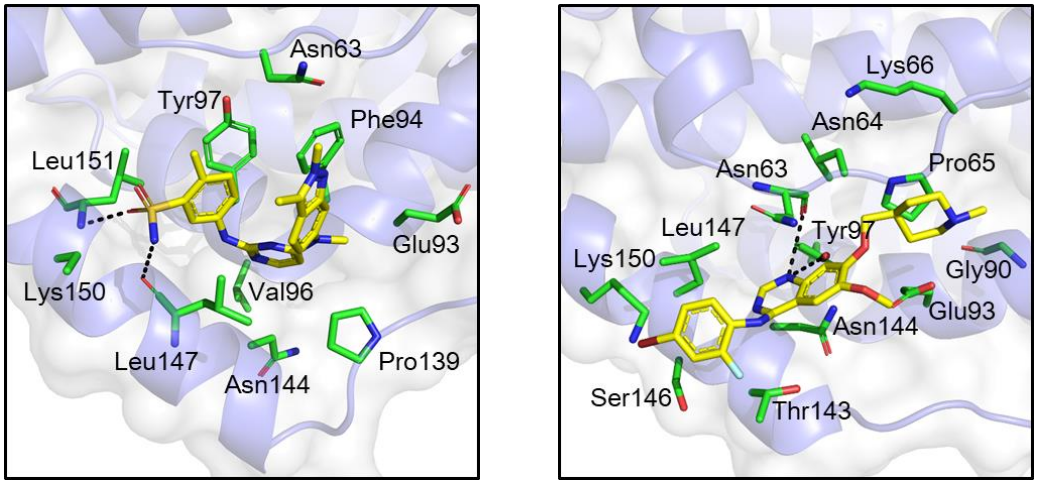



**Figure S8.** Root mean square deviation (RMSD) of the protein in two complex system as the 100 function of simulation time.




(a) Pazopanib-IL-6 (b) Vandetanib-IL-6

**Figure S9.** Molecular dynamics simulation results for small molecules binding to IL-6 where the black dotted line represent hydrogen bonds.

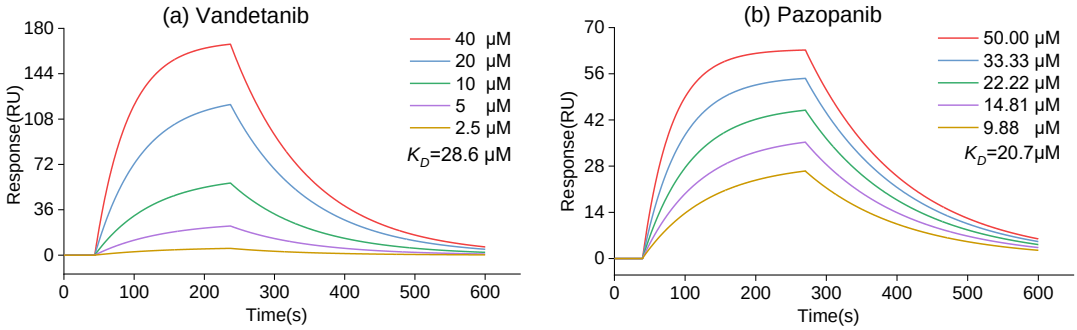



**Figure S10.** SPR sensorgrams and binding affinity between two molecules and IL-6.

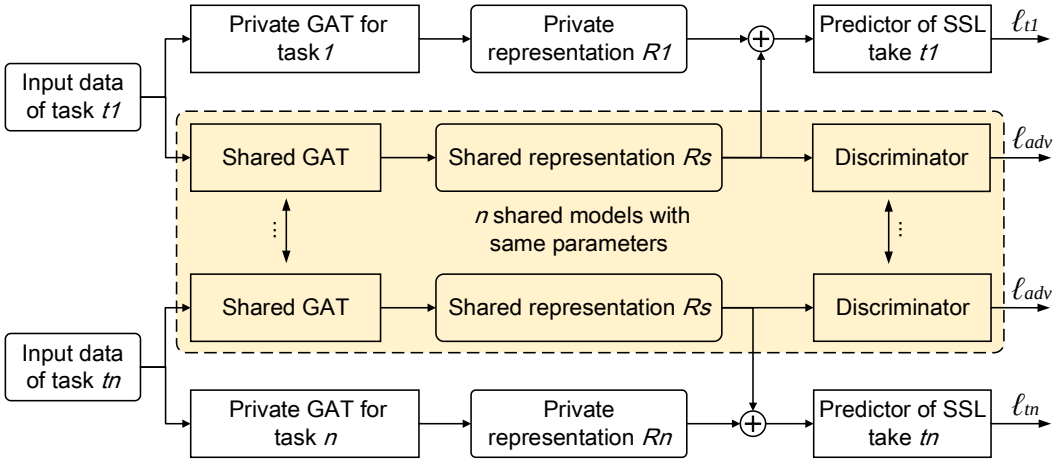



**Figure S11.** Adversarial training-based multi-task learning framework.

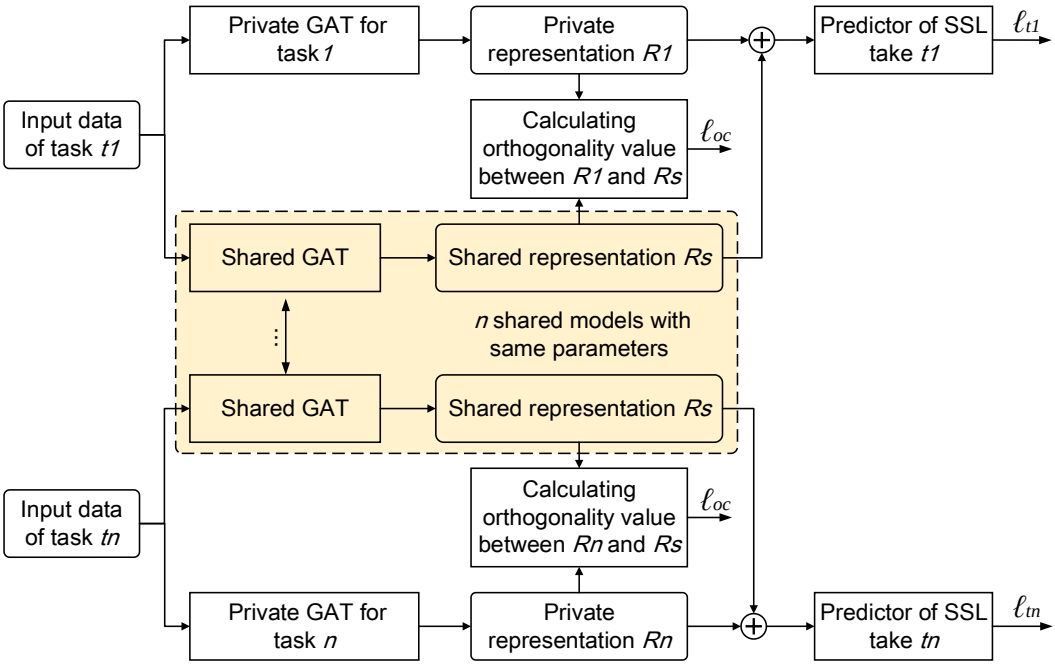



**Figure S12.** Orthogonality constraint-based multi-task training framework.

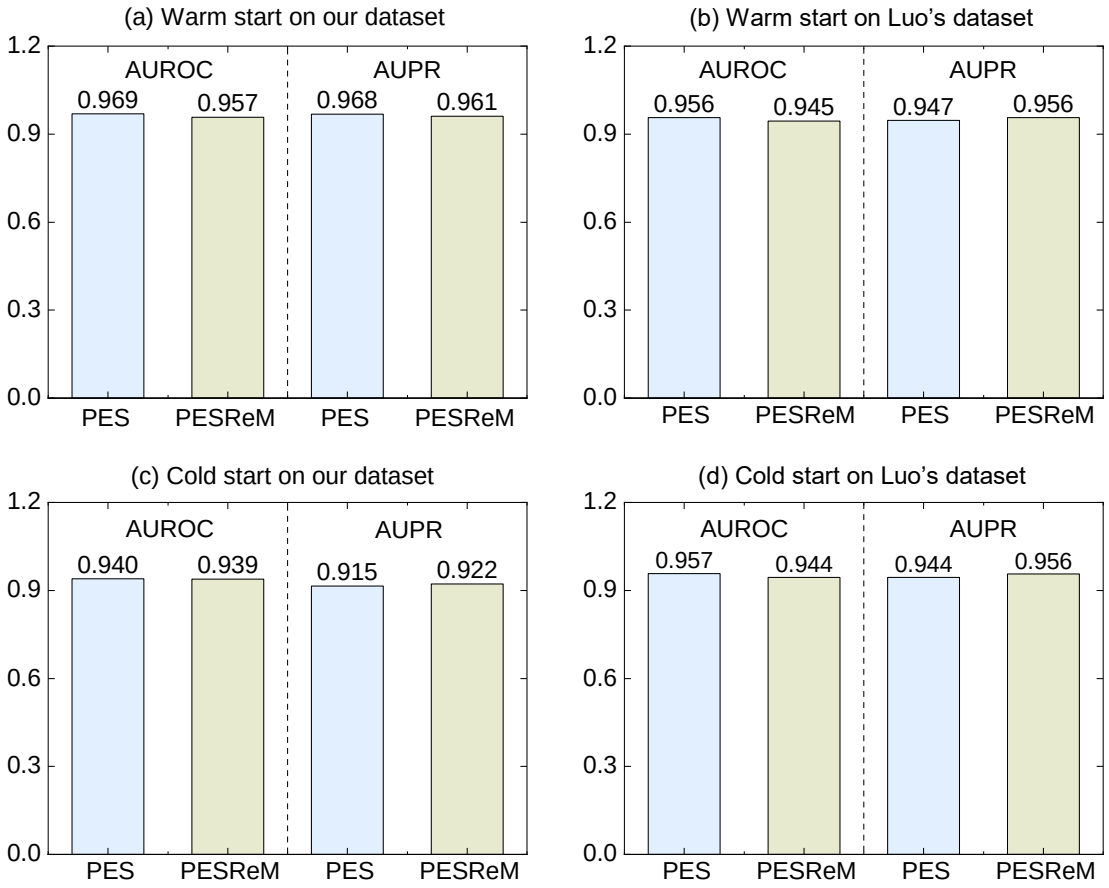



**Figure S13.** Performance comparison of PESReM and PairDistance-EdgeMask-SimCon for DTI predictions where PSE denotes PairDistance-EdgeMask-SimCon; PESReM denotes the variant of PairDistance-EdgeMask-SimCon when test data is removed from SSL.

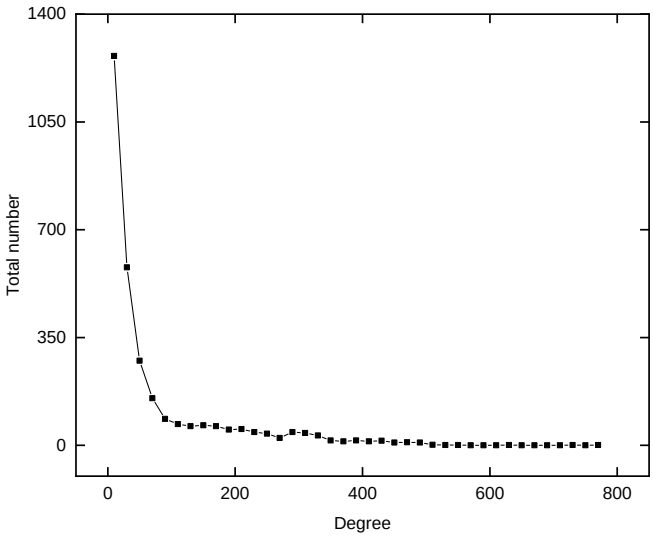



**Figure S14.** The distribution of node degree.

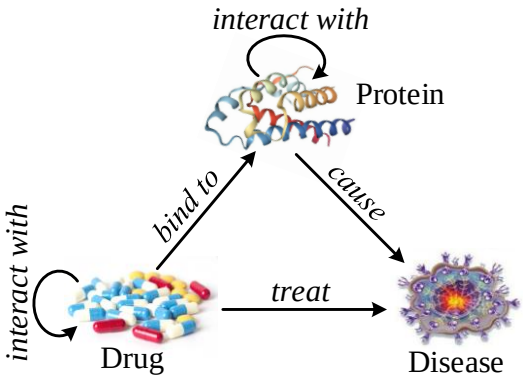



**Figure S15.** Schema of the BioHN.

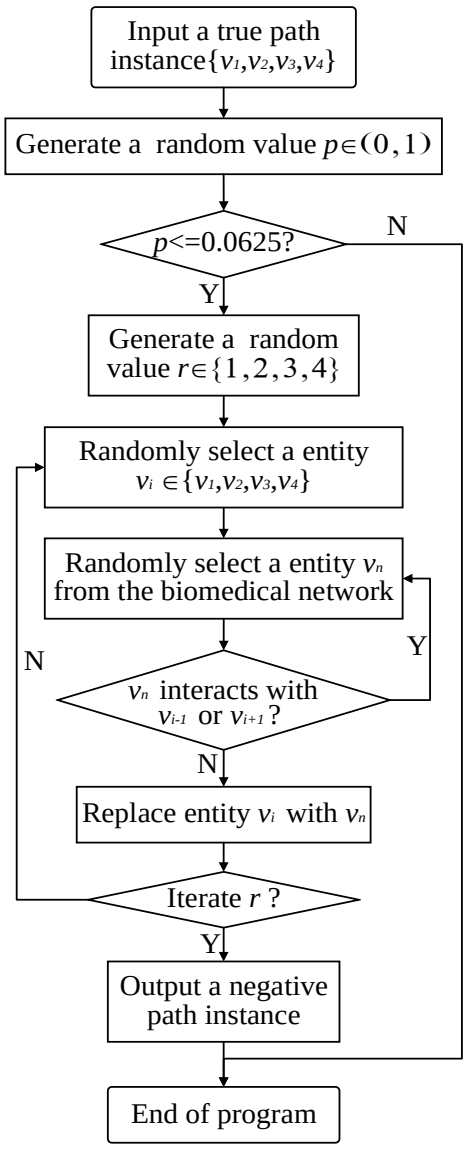


**Figure S16.** The procedure of negative path generation.

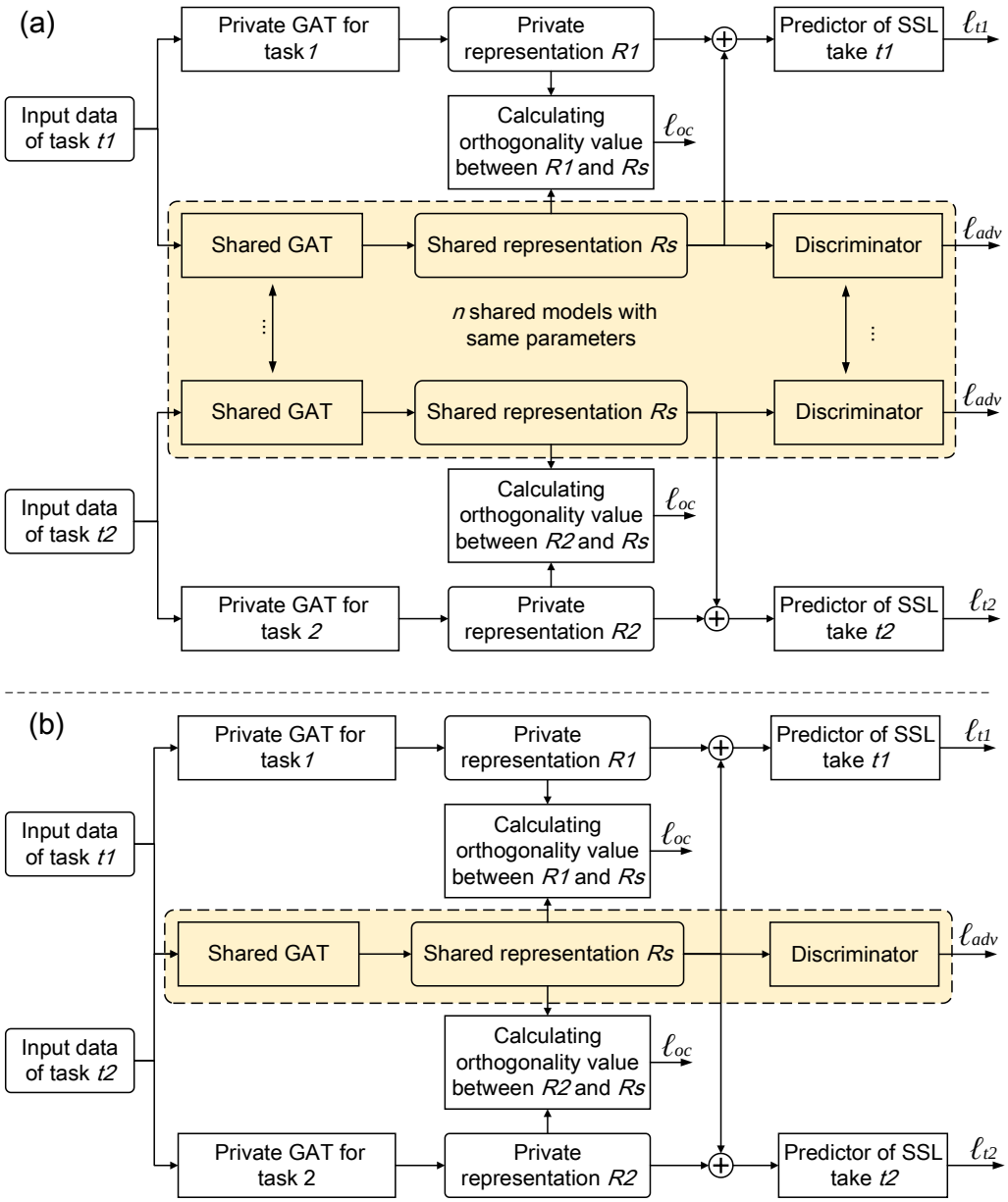
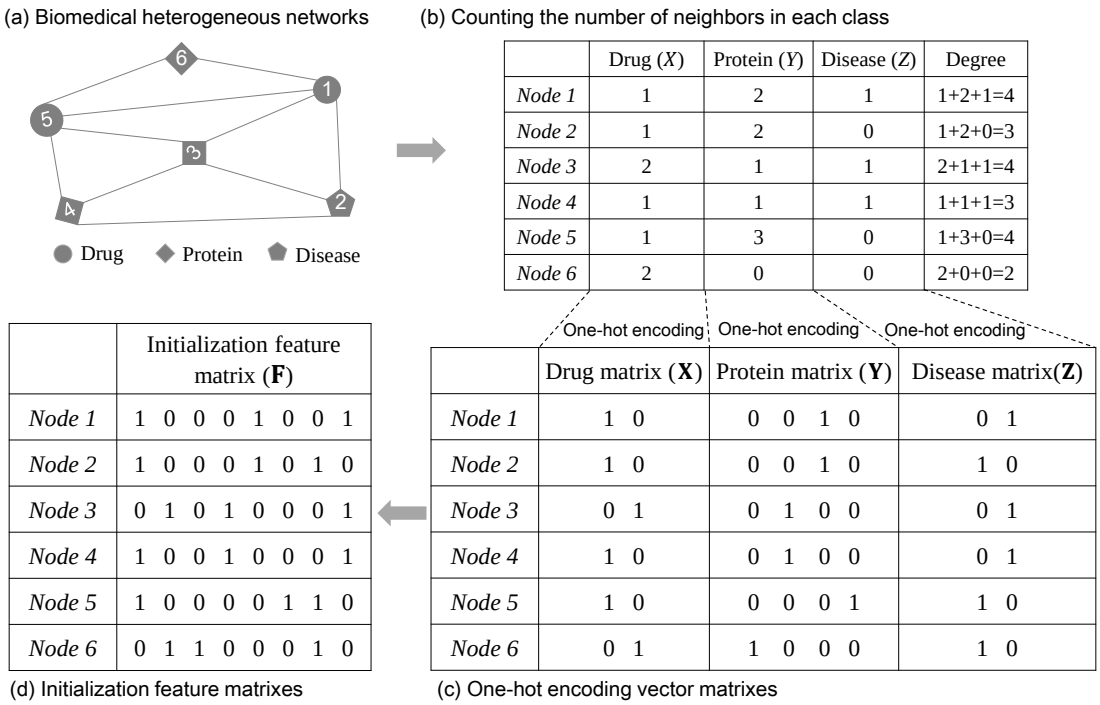



**Figure S17.** The frameworks of graph attention-based two-task learning.




| | Drug ($X$) | Protein ($Y$) | Disease ($Z$) | Degree |
|---|---|---|---|---|
| *Node 1* | 1 | 2 | 1 | 1+2+1=4 |
| *Node 2* | 1 | 2 | 0 | 1+2+0=3 |
| *Node 3* | 2 | 1 | 1 | 2+1+1=4 |
| *Node 4* | 1 | 1 | 1 | 1+1+1=3 |
| *Node 5* | 1 | 3 | 0 | 1+3+0=4 |
| *Node 6* | 2 | 0 | 0 | 2+0+0=2 |

| | Initialization feature matrix (**F**) |
|---|---|
| *Node 1* | 1 0 0 0 1 0 0 1 |
| *Node 2* | 1 0 0 0 1 0 1 0 |
| *Node 3* | 0 1 0 1 0 0 0 1 |
| *Node 4* | 1 0 0 1 0 0 0 1 |
| *Node 5* | 1 0 0 0 0 1 1 0 |
| *Node 6* | 0 1 1 0 0 0 1 0 |

| | Drug matrix (**X**) | Protein matrix (**Y**) | Disease matrix(**Z**) |
|---|---|---|---|
| *Node 1* | 1 0 | 0 0 1 0 | 0 1 |
| *Node 2* | 1 0 | 0 0 1 0 | 1 0 |
| *Node 3* | 0 1 | 0 1 0 0 | 0 1 |
| *Node 4* | 1 0 | 0 1 0 0 | 0 1 |
| *Node 5* | 1 0 | 0 0 0 1 | 1 0 |
| *Node 6* | 0 1 | 1 0 0 0 | 1 0 |

**Figure S18.** An example of features initialization process.

## Supplementary Tables

**Table S1.** The *p*-value among single task-driven SSL models for DTI and DDI predictions where we used the two-sided Student's t-test with a significance threshold of 0.05. No adjustments were made for multiple comparisons.

| | Method | ClusterPre | EdgeMask | PairDistance | PathClass | SimReg | SimCon |
|---|---|---|---|---|---|---|---|
| DTI | ClusterPre | / | 0.112 | 2.38E-8 | 7.23E-5 | 5.41E-4 | 1.63E-5 |
| | EdgeMask | 0.112 | / | 1.84E-3 | 4.66E-3 | 6.08E-7 | 3.30E-6 |
| | PairDistance | 2.38E-8 | 1.84E-3 | / | 2.62E-3 | 0.815 | 5.17E-3 |
| | PathClass | 7.23E-5 | 4.66 E-3 | 2.62E-3 | / | 1.59E-2 | 1.80E-5 |
| | SimReg | 5.41E-4 | 6.08E-7 | 0.815 | 1.59E-2 | / | 3.58E-3 |
| | SimCon | 1.63E-5 | 3.30E-6 | 5.17E-3 | 1.80E-5 | 3.58E-3 | / |
| DDI | ClusterPre | / | 1.50E-6 | 3.3E-4 | 2.68E-8 | 1.20E-5 | 1.54E-11 |
| | EdgeMask | 1.50E-6 | / | 7.49E-9 | 3.50E-14 | 2.87E-12 | 3.75E-16 |
| | PairDistance | 3.3E-4 | 7.49E-9 | / | 1.29E-6 | 0.121 | 9.35E-12 |
| | PathClass | 2.68E-8 | 3.50E-14 | 1.29E-6 | / | 3.86E-8 | 2.84E-12 |
| | SimReg | 1.20E-5 | 2.87E-12 | 0.121 | 3.86E-8 | / | 1.35E-15 |
| | SimCon | 1.54E-11 | 3.75E-16 | 9.35E-12 | 2.84E-12 | 1.35E-15 | / |

a. The yellow shaded areas denote the *p*-value between local information- and global information-based SSL models.

b. The green shaded areas denote the *p*-value between attribute strong and weak constraint-based SSL models.

c. The *p*-value among SSL models higher than 0.05 are marked in red.

**Table S2.** 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 | **0.939**±0.002 | **0.937±0.002** | **0.969**±0.006 | **0.968±0.007** |

a. ‘std’ denotes the standard deviation value calculated across ten results.

b. The best results are marked in **boldface.**

**Table S3.** The *p*-value among 11 two-task combination models for DTI predictions where we used the two-sided Student's t-test with a significance threshold of 0.05. No adjustments were made for multiple comparisons.

| | EdgeMask-PairDistance | ClusterPre-PathClass | ClusterPre-PairDistance | EdgeMask-PathClass | PairDistance-PathClass | PathClass-SimCon | PairDistance-SimCon | EdgeMask-SimReg | PairDistance-SimReg | ClusterPre-EdgeMask | SimReg-SimCon |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EdgeMask-PairDistance | / | 0.258 | 1.05 E-5 | 1.17 E-5 | 2.38 E-8 | 1.37 E-4 | 8.56 E-8 | 2.79 E-5 | 3.12 E-8 | 1.55 E-8 | 3.57 E-8 |
| ClusterPre-PathClass | 0.258 | / | 1.56 E-4 | 6.01 E-4 | 3.22 E-5 | 4.11 E-4 | 3.06 E-5 | 2.17 E-3 | 1.53 E-6 | 5.63 E-6 | 1.30 E-8 |
| ClusterPre-PairDistance | 1.05 E-5 | 1.56 E-4 | / | 0.935 | 0.025 | 0.582 | 0.029 | 0.661 | 7.51 E-5 | 3.43 E-4 | 2.39 E-6 |
| EdgeMask-PathClass | 1.17 E-5 | 6.01 E-4 | 0.935 | / | 0.015 | 0.700 | 0.011 | 0.466 | 8.62 E-6 | 5.87 E-4 | 5.25 E-6 |
| PairDistance-PathClass | 2.38 E-8 | 3.22 E-5 | 0.025 | 0.015 | / | 0.043 | 0.244 | 0.011 | 1.75 E-4 | 3.07 E-3 | 3.14 E-6 |
| PathClass-SimCon | 1.37 E-4 | 4.11 E-4 | 0.582 | 0.700 | 0.043 | / | 0.030 | 0.915 | 8.98 E-4 | 4.24 E-3 | 8.39 E-8 |
| PairDistance-SimCon | 8.56 E-8 | 3.06 E-5 | 0.029 | 0.011 | 0.244 | 0.030 | / | 4.41 E-3 | 1.20 E-3 | 0.013 | 4.74 E-6 |
| EdgeMask-SimReg | 2.79 E-5 | 2.17 E-3 | 0.661 | 0.466 | 0.011 | 0.915 | 4.41 E-3 | / | 3.66 E-5 | 1.97 E-4 | 6.05 E-6 |
| PairDistance-SimReg | 3.12 E-8 | 1.53 E-6 | 7.51 E-5 | 8.62 E-6 | 1.75 E-4 | 8.98 E-4 | 1.20 E-3 | 3.66 E-5 | / | 0.257 | 2.00 E-5 |
| ClusterPre-EdgeMask | 1.55 E-8 | 5.63 E-6 | 3.43 E-4 | 5.87 E-4 | 3.07 E-3 | 4.24 E-3 | 0.013 | 1.97 E-4 | 0.257 | / | 3.29 E-5 |
| SimReg-SimCon | 3.57 E-8 | 1.30 E-8 | 2.39 E-6 | 5.25 E-6 | 3.14 E-6 | 8.39 E-8 | 4.74 E-6 | 6.05 E-6 | 2.00 E-5 | 3.29 E-5 | / |

a. The blue shaded areas represent the *p*-value between local-global combination models and other models.

b. The *p*-value among SSL models higher than 0.05 are marked in red.

**Table S4.** The *p*-value among 11 two-task combination SSL models for DDI predictions where we used the two-sided Student's t-test with a significance threshold of 0.05. No adjustments were made for multiple comparisons.

| | EdgeMask-PairDistance | ClusterPre-PathClass | ClusterPre-PairDistance | EdgeMask-PathClass | PairDistance-PathClass | PathClass-SimCon | PairDistance-SimCon | EdgeMask-SimReg | PairDistance-SimReg | ClusterPre-EdgeMask | SimReg-SimCon |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EdgeMask-PairDistance | / | 0.145 | 1.53 E-9 | 3.80 E-7 | 7.03 E-9 | 1.70 E-9 | 1.46 E-10 | 3.71 E-12 | 4.26 E-12 | 7.77 E-14 | 7.79 E-14 |
| ClusterPre-PathClass | 0.145 | / | 2.67 E-10 | 1.43 E-6 | 5.92 E-10 | 2.90 E-9 | 4.22 E-12 | 4.72 E-10 | 3.85 E-10 | 9.97 E-15 | 4.27 E-13 |
| ClusterPre-PairDistance | 1.53 E-9 | 2.67 E-10 | / | 2.03 E-3 | 3.55 E-4 | 5.96 E-05 | 5.33 E-8 | 1.90 E-7 | 9.92 E-8 | 6.40 E-12 | 5.18 E-12 |
| EdgeMask-PathClass | 3.80 E-7 | 1.43 E-6 | 2.03 E-3 | / | 0.304 | 0.813 | 0.418 | 0.025 | 0.011 | 5.59 E-5 | 5.86 E-6 |
| PairDistance-PathClass | 7.03 E-9 | 5.92 E-10 | 3.55 E-4 | 0.304 | / | 0.056 | 4.12 E-3 | 2.87 E-5 | 7.13 E-6 | 2.53 E-8 | 8.86 E-9 |
| PathClass-SimCon | 1.70 E-9 | 2.90 E-9 | 5.96 E-5 | 0.813 | 0.056 | / | 0.062 | 2.38 E-4 | 1.64 E-4 | 2.19 E-7 | 3.03 E-8 |
| PairDistance-SimCon | 1.46 E-10 | 4.22 E-12 | 5.33 E-8 | 0.418 | 4.12 E-3 | 0.062 | / | 9.02 E-3 | 2.03 E-3 | 9.54 E-9 | 2.57 E-9 |
| EdgeMask-SimReg | 3.71 E-12 | 4.72 E-10 | 1.90 E-7 | 0.025 | 2.87 E-5 | 2.38 E-4 | 9.02 E-3 | / | 0.014 | 4.19 E-6 | 4.33 E-8 |
| PairDistance-SimReg | 4.26 E-12 | 3.85 E-10 | 9.92 E-8 | 0.011 | 7.13 E-6 | 1.64 E-4 | 2.03 E-3 | 0.014 | / | 1.98 E-5 | 1.05 E-7 |
| ClusterPre-EdgeMask | 7.77 E-14 | 9.97 E-15 | 6.40 E-12 | 5.59 E-5 | 2.53 E-8 | 2.19 E-7 | 9.54 E-9 | 4.19 E-6 | 1.98 E-5 | / | 1.21 E-3 |
| SimReg-SimCon | 7.79 E-14 | 4.27 E-13 | 5.18 E-12 | 5.86 E-6 | 8.86 E-9 | 3.03 E-8 | 2.57 E-9 | 4.33 E-8 | 1.05 E-7 | 1.21 E-3 | / |

a. The blue shaded areas represent the *p*-value between local-global combination models and other models.

b. The *p*-value among SSL models higher than 0.05 are marked in red.

**Table S5.** The *p*-value among 10 SSL models based on the mixed results of DDI and DTI predictions where we used the two-sided Student's t-test with a significance threshold of 0.05. No adjustments were made for multiple comparisons.

| | ClusterPre-PathClass | EdgeMask-PairDistance | ClusterPre-PairDistance | EdgeMask-PathClass | PairDistance-SimCon | SimReg-SimCon | PairDistance-SimReg-SimCon | PairDistance-EdgeMask-SimCon | ClusterPre-PairDistance-PathClass | ClusterPre-PathClass-SimReg |
|---|---|---|---|---|---|---|---|---|---|---|
| ClusterPre-PathClass | / | 0.145 | 2.67 E-10 | 1.43 E-6 | 4.22 E-12 | 4.27 E-13 | 1.26 E-7 | 1.99 E-12 | 0.061 | 8.62 E-7 |
| EdgeMask-PairDistance | 0.145 | / | 1.53 E-9 | 3.80 E-7 | 1.46 E-10 | 7.79 E-14 | 4.94 E-8 | 2.42 E-9 | 0.017 | 5.87 E-5 |
| ClusterPre-PairDistance | 2.67 E-10 | 1.53 E-9 | / | 0.002 | 5.33 E-8 | 5.18 E-12 | 2.15 E-4 | 1.66 E-12 | 1.67 E-7 | 3.22 E-11 |
| EdgeMask-PathClass | 1.43 E-6 | 3.80 E-7 | 0.002 | / | 0.418 | 5.86 E-6 | 0.263 | 8.79 E-9 | 3.04 E-6 | 4.18 E-7 |
| PairDistance-SimCon | 4.22 E-12 | 1.46 E-10 | 5.33 E-8 | 0.418 | / | 2.57 E-9 | 0.451 | 2.63 E-14 | 1.39 E-10 | 1.20 E-10 |
| SimReg-SimCon | 4.27 E-13 | 7.79 E-14 | 5.18 E-12 | 5.86 E-6 | 2.57 E-9 | / | 7.45 E-6 | 4.34 E-15 | 2.33 E-12 | 1.28 E-12 |
| PairDistance-SimReg-SimCon | 1.26 E-7 | 4.94 E-8 | 2.15 E-4 | 0.263 | 0.451 | 7.45 E-6 | / | 1.00 E-9 | 5.33 E-8 | 1.19 E-7 |
| PairDistance-EdgeMask-SimCon | 1.99 E-12 | 2.42 E-9 | 1.66 E-12 | 8.79 E-9 | 2.63 E-14 | 4.34 E-15 | 1.00 E-9 | / | 2.16 E-10 | 7.00 E-7 |
| ClusterPre-PairDistance-PathClass | 0.061 | 0.017 | 1.67 E-7 | 3.04 E-6 | 1.39 E-10 | 2.33 E-12 | 5.33 E-8 | 2.16 E-10 | / | 1.55 E-5 |
| ClusterPre-PathClass-SimReg | 8.62 E-7 | 5.87 E-5 | 3.22 E-11 | 4.18 E-7 | 1.20 E-10 | 1.28 E-12 | 1.19 E-7 | 7.00 E-7 | 1.55 E-5 | / |

a. The gray shaded areas represent the *p*-value between multimodal combinations and other models.

b. The *p*-value among SSL models higher than 0.05 are marked in red.

**Table S6.** 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 | PairDistance-SimCon | 2 | 0.858±0.016 | 0.837±0.020 | 0.903±0.037 | 0.878±0.065 |
| 7 | EdgeMask-SimReg | 2 | 0.848±0.022 | 0.814±0.026 | 0.913±0.029 | 0.867±0.072 |
| 8 | PathClass-SimCon | 2 | 0.850±0.020 | 0.801±0.035 | 0.909±0.032 | 0.866±0.079 |
| 9 | ClusterPre-EdgeMask | 2 | 0.837±0.021 | 0.792±0.029 | 0.910±0.026 | 0.875±0.058 |
| 10 | PairDistance-SimReg | 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** |

The best results are marked in **boldface**.

**Table S7.** The numbers of nodes and edges in Luo's dataset

| Type of node | Count | Type of edge | Count |
|---|---|---|---|
| Drug | 660 | Drug-drug interactions | 10,036 |
| Protein | 1,324 | Drug-protein interactions | 1,923 |
| / | / | Protein-protein interactions | 7,363 |
| Total | 1,984 | Total | 19,322 |

**Table S8.** The results of multi-task models on Luo's dataset for warm start predictions (training set:test set = 9:1)

| No. | Self-supervised tasks | Modal size | DDI | | DTI | |
|---|---|---|---|---|---|---|
| | | | AUROC±std | AUPR±std | AUROC±Std | AUPR±Std |
| 1 | EdgeMask-PairDistance | 2 | 0.938±0.005 | 0.931±0.011 | 0.946±**0.008** | 0.929±0.019 |
| 2 | ClusterPre-PathClass | 2 | 0.925±0.008 | 0.916±0.015 | 0.931±0.011 | 0.915±0.016 |
| 3 | ClusterPre-PairDistance | 1 | 0.918±0.005 | 0.904±0.012 | 0.927±0.008 | 0.911±**0.012** |
| 4 | EdgeMask-PathClass | 1 | 0.924±**0.003** | 0.911±0.008 | 0.922±0.013 | 0.894±0.022 |
| 5 | PairDistance-PathClass | 2 | 0.916±0.004 | 0.899±0.009 | 0.910±0.016 | 0.873±0.027 |
| 6 | PathClass-SimCon | 2 | 0.916±0.008 | 0.905±0.014 | 0.887±0.021 | 0.845±0.037 |
| 7 | PairDistance-SimCon | 2 | 0.917±0.006 | 0.903±0.012 | 0.900±0.017 | 0.864±0.030 |
| 8 | EdgeMask-SimReg | 2 | 0.908±0.009 | 0.888±0.020 | 0.901±0.016 | 0.861±0.030 |
| 9 | PairDistance-SimReg | 2 | 0.907±0.009 | 0.888±0.020 | 0.905±0.017 | 0.870±0.031 |
| 10 | ClusterPre-EdgeMask | 2 | 0.893±0.009 | 0.872±0.019 | 0.909±0.016 | 0.885±0.022 |
| 11 | SimReg-SimCon | 1 | 0.891±0.011 | 0.876±0.020 | 0.860±0.012 | 0.837±0.029 |
| 12 | ClusterPre-PairDistance-PathClass | 2 | 0.930±0.006 | 0.920±0.011 | 0.930±0.017 | 0.912±0.029 |
| 13 | ClusterPre-PathClass-SimReg | 3 | 0.932±0.004 | 0.925±**0.007** | 0.937±0.010 | 0.916±0.017 |
| 14 | PairDistance-SimReg-SimCon | 2 | 0.915±0.009 | 0.899±0.015 | 0.899±0.017 | 0.862±0.032 |
| 15 | PairDistance-EdgeMask-SimCon | 3 | **0.951**±0.004 | **0.944**±0.011 | **0.956±0.008** | **0.947**±0.014 |

The best results are marked in **boldface**.

**Table S9.** The results of multi-task models on Luo's dataset for warm start predictions (training set:test set = 5:5)

| No. | Self-supervised tasks | Modal size | DDI | | DTI | |
|---|---|---|---|---|---|---|
| | | | AUROC±std | AUPR±std | AUROC±Std | AUPR±Std |
| 1 | EdgeMask-PairDistance | 2 | 0.920±0.006 | 0.908±0.009 | 0.931±0.005 | 0.908±0.008 |
| 2 | ClusterPre-PathClass | 2 | 0.908±**0.004** | 0.896±0.008 | 0.913±0.005 | 0.892±0.008 |
| 3 | ClusterPre-PairDistance | 1 | 0.900±0.005 | 0.884±0.006 | 0.910±0.006 | 0.885±0.011 |
| 4 | EdgeMask-PathClass | 1 | 0.909±**0.004** | 0.894±**0.004** | 0.908±0.005 | 0.863±0.011 |
| 5 | PairDistance-PathClass | 2 | 0.896±0.006 | 0.881±0.007 | 0.899±0.006 | 0.855±0.013 |
| 6 | PathClass-SimCon | 2 | 0.884±0.007 | 0.873±0.011 | 0.871±0.009 | 0.817±0.018 |
| 7 | PairDistance-SimCon | 2 | 0.895±0.009 | 0.877±0.012 | 0.886±0.006 | 0.844±0.011 |
| 8 | EdgeMask-SimReg | 2 | 0.881±0.010 | 0.8600.016 | 0.887±0.006 | 0.843±0.011 |
| 9 | PairDistance-SimReg | 2 | 0.894±0.006 | 0.877±0.009 | 0.892±0.006 | 0.854±0.010 |
| 10 | ClusterPre-EdgeMask | 2 | 0.872±0.006 | 0.846±0.011 | 0.894±0.005 | 0.864±0.008 |
| 11 | SimReg-SimCon | 1 | 0.841±0.011 | 0.813±0.017 | 0.854±0.009 | 0.828±0.015 |
| 12 | ClusterPre-PairDistance-PathClass | 2 | 0.900±0.005 | 0.886±0.005 | 0.918±0.007 | 0.896±0.011 |
| 13 | ClusterPre-PathClass-SimReg | 3 | 0.912±0.004 | 0.902±0.007 | 0.910±**0.004** | 0.886±0.010 |
| 14 | PairDistance-SimReg-SimCon | 2 | 0.900±0.006 | 0.883±0.008 | 0.885±0.007 | 0.842±0.012 |
| 15 | PairDistance-EdgeMask-SimCon | 3 | **0.932±0.004** | **0.922**±0.006 | **0.946±0.004** | **0.933±0.005** |

The best results are marked in **boldface**.

**Table S10.** The results of multi-task models on Luo's dataset for 5% drugs cold start predictions

| No. | Self-supervised tasks | Modal size | DDI | | DTI | |
|---|---|---|---|---|---|---|
| | | | AUROC±std | AUPR±std | AUROC±Std | AUPR±Std |
| 1 | EdgeMask-PairDistance | 2 | 0.927±0.013 | 0.912±0.026 | 0.941±0.019 | 0.923±0.024 |
| 2 | ClusterPre-PathClass | 2 | 0.904±0.015 | 0.884±0.026 | 0.928±0.023 | 0.908±0.033 |
| 3 | ClusterPre-PairDistance | 1 | 0.902±0.012 | 0.878±0.023 | 0.924±0.023 | 0.897±0.046 |
| 4 | EdgeMask-PathClass | 1 | 0.918±**0.012** | 0.898±0.022 | 0.921±0.033 | 0.877±0.039 |
| 5 | PairDistance-PathClass | 2 | 0.905±0.019 | 0.883±0.035 | 0.909±0.029 | 0.856±0.047 |
| 6 | PathClass-SimCon | 2 | 0.905±**0.012** | 0.887±0.028 | 0.891±0.034 | 0.832±0.062 |
| 7 | PairDistance-SimCon | 2 | 0.907±0.014 | 0.884±0.029 | 0.899±0.030 | 0.843±0.051 |
| 8 | EdgeMask-SimReg | 2 | 0.898±0.015 | 0.875±0.038 | 0.902±0.027 | 0.845±0.049 |
| 9 | PairDistance-SimReg | 2 | 0.897±0.012 | 0.868±0.024 | 0.904±0.029 | 0.853±0.050 |
| 10 | ClusterPre-EdgeMask | 2 | 0.872±0.017 | 0.842±0.046 | 0.904±0.028 | 0.865±0.042 |
| 11 | SimReg-SimCon | 1 | 0.865±0.016 | 0.833±0.045 | 0.863±0.037 | 0.824±0.052 |
| 12 | ClusterPre-PairDistance-PathClass | 2 | 0.906±0.013 | 0.888±0.029 | 0.930±0.021 | 0.909±0.041 |
| 13 | ClusterPre-PathClass-SimReg | 3 | 0.913±0.013 | 0.898±0.024 | 0.933±0.026 | 0.918±0.038 |
| 14 | PairDistance-SimReg-SimCon | 2 | 0.906±0.015 | 0.886±0.031 | 0.899±0.029 | 0.843±0.051 |
| 15 | PairDistance-EdgeMask-SimCon | 3 | **0.939±0.012** | **0.930±0.017** | **0.957±0.015** | **0.944±0.020** |

The best results are marked in **boldface**.

**Table S11.** The results of multi-task models on Luo's dataset for 10% drugs cold start predictions

| No. | Self-supervised tasks | Modal size | DDI | | DTI | |
|---|---|---|---|---|---|---|
| | | | AUROC±std | AUPR±std | AUROC±Std | AUPR±Std |
| 1 | EdgeMask-PairDistance | 2 | 0.928±**0.008** | 0.912±0.017 | 0.940±0.014 | 0.922±**0.016** |
| 2 | ClusterPre-PathClass | 2 | 0.907±0.013 | 0.890±0.020 | 0.922±0.015 | 0.900±0.025 |
| 3 | ClusterPre-PairDistance | 1 | 0.901±0.012 | 0.877±0.023 | 0.921±0.015 | 0.899±0.025 |
| 4 | EdgeMask-PathClass | 1 | 0.916±0.009 | 0.895±0.019 | 0.915±0.013 | 0.872±0.024 |
| 5 | PairDistance-PathClass | 2 | 0.905±0.013 | 0.881±0.028 | 0.903±0.012 | 0.850±0.031 |
| 6 | PathClass-SimCon | 2 | 0.902±0.012 | 0.885±0.019 | 0.883±0.013 | 0.823±0.032 |
| 7 | PairDistance-SimCon | 2 | 0.904±0.011 | 0.879±0.024 | 0.892±0.012 | 0.840±0.029 |
| 8 | EdgeMask-SimReg | 2 | 0.894±0.010 | 0.868±0.021 | 0.896±0.014 | 0.844±0.025 |
| 9 | PairDistance-SimReg | 2 | 0.895±0.009 | 0.866±0.025 | 0.898±**0.011** | 0.845±0.028 |
| 10 | ClusterPre-EdgeMask | 2 | 0.870±0.015 | 0.834±0.030 | 0.901±0.013 | 0.862±0.027 |
| 11 | SimReg-SimCon | 1 | 0.858±0.016 | 0.824±0.022 | 0.853±0.020 | 0.815±0.028 |
| 12 | ClusterPre-PairDistance-PathClass | 2 | 0.904±0.013 | 0.884±0.023 | 0.926±0.017 | 0.902±0.026 |
| 13 | ClusterPre-PathClass-SimReg | 3 | 0.912±0.012 | 0.897±0.018 | 0.923±0.018 | 0.900±0.021 |
| 14 | PairDistance-SimReg-SimCon | 2 | 0.904±0.012 | 0.877±0.025 | 0.891±0.012 | 0.836±0.029 |
| 15 | PairDistance-EdgeMask-SimCon | 3 | **0.939±0.008** | **0.928±0.014** | **0.953**±0.013 | **0.941**±0.017 |

The best results are marked in **boldface**.

**Table S12.** Results of all methods on Luo dataset

| Scenarios | Methods | DDI | | DTI | |
|---|---|---|---|---|---|
| | | AUROC ±std | AUPR ±std | AUROC ±std | AUPR ±std |
| Warm start | DeepR2cov | 0.931 ±0.009 | 0.912 ±0.012 | 0.922 ±0.011 | 0.910 ±0.016 |
| | DDIMDL | 0.913 ±0.009 | 0.905 ±0.014 | 0.914 ±0.009 | 0.915 ±0.009 |
| | DTINet | 0.929 ±0.006 | 0.927 ±**0.009** | 0.922 ±0.007 | 0.924 ±0.008 |
| | KGE_NFM | 0.916 ±0.008 | 0.907 ±0.010 | 0.879 ±0.008 | 0.892 ±**0.007** |
| | MoleculeNet | 0.912 ±0.007 | 0.899 ±0.013 | 0.905 ±0.016 | 0.905 ±0.021 |
| | deepDTnet | 0.923 ±0.008 | 0.921 ±0.010 | 0.911 ±**0.006** | 0.901 ±0.012 |
| | PairDistance-EdgeMask-SimCon | **0.951** ±**0.004** | **0.944** ±0.011 | **0.956** ±0.008 | **0.947** ±0.014 |
| Cold start | DeepR2cov | 0.909 ±0.012 | 0.875 ±0.031 | 0.920 ±0.035 | 0.901 ±0.044 |
| | DDIMDL | 0.852 ±**0.007** | 0.855 ±**0.007** | 0.898 ±0.026 | 0.908 ±0.025 |
| | DTINet | 0.918 ±0.017 | 0.913 ±0.029 | 0.910 ±0.023 | 0.911 ±0.026 |
| | KGE_NFM | 0.743 ±0.034 | 0.709 ±0.034 | 0.782 ±**0.003** | 0.793 ±**0.005** |
| | MoleculeNet | 0.905 ±0.019 | 0.878 ±0.039 | 0.898 ±0.045 | 0.905 ±0.031 |
| | deepDTnet | 0.913 ±0.014 | 0.904 ±0.025 | 0.888 ±0.028 | 0.892 ±0.044 |
| | PairDistance-EdgeMask-SimCon | **0.939** ±0.008 | **0.928** ±0.014 | **0.953** ±0.013 | **0.941** ±0.017 |

The best results are marked in **boldface**.

**Table S13.** Performance of MSSL2drug, LE, GF and DeepWalk for drug discovery

| Methods | DDI | | DTI | |
|---|---|---|---|---|
| | AUROC ±std | AUPR ±std | AUROC ±std | AUPR ±std |
| LE | 0.795 ±0.001 | 0.763 ±**0.001** | 0.845 ±**0.001** | 0.850 ±**0.001** |
| GF | 0.921 ±0.005 | 0.919 ±0.006 | 0.858 ±0.017 | 0.867 ±0.014 |
| DeepWalk | 0.922 ±0.004 | 0.917 ±0.005 | 0.939 ±0.003 | 0.937 ±0.003 |
| PairDistance-EdgeMask-SimCon | **0.939** ±**0.002** | **0.937** ±0.002 | **0.969** ±0.006 | **0.968** ±0.007 |

The best results are marked in **boldface**.

**Table S14.** Performance of MSSL2drug and MF2A for DTI predictions

| Methods | Our dataset | | Luo's dataset | |
|---|---|---|---|---|
| | AUROC ±std | AUPR ±std | AUROC ±std | AUPR ±std |
| MF2A | 0.942 ±0.008 | 0.347 ±0.008 | 0.915 ±0.001 | 0.441 ±0.007 |
| PairDistance-EdgeMask-SimCon | **0.969** ±0.006 | **0.968** ±0.007 | **0.956** ±0.008 | **0.947** ±0.014 |

The best results are marked in **boldface**.

**Table S15.** Performance of MSSL2drug and MIRACLE for DDI predictions

| Methods | Our dataset | | Luo's dataset | |
|---|---|---|---|---|
| | AUROC±Std | AUPR±Std | AUROC±Std | AUPR±Std |
| MIRACLE | 0.898±0.048 | 0.871±0.048 | 0.858±0.009 | 0.813±**0.009** |
| PairDistance-EdgeMask-SimCon | **0.939±0.002** | **0.937±0.002** | **0.951±0.004** | **0.944**±0.011 |

The best results are marked in **boldface**.

**Table S16.** Anti-inflammatory candidate agents for COVID-19 patients

| No. | DrugBank ID: Name | Confidence score | PMID |
|---|---|---|---|
| 1 | DB08604:Triclosan | 0.800 | 29568771, 29067681 |
| 2 | DB00756:Hexachlorophene | 0.772 | NA |
| 3 | DB08877:Ruxolitinib* | 0.737 | 32789663, 32679107 |
| 4 | DB00799:Tazarotene | 0.728 | 7512583, 30134735 |
| 5 | DB06616:Bosutinib | 0.713 | 25351958, 33685634 |
| 6 | DB05294:Vandetanib | 0.709 | 34981062 |
| 7 | DB01268:Sunitinib | 0.651 | 31039345, 23867310, 33191180 |
| 8 | DB00530:Erlotinib | 0.650 | 32566018 |
| 9 | DB06589:Pazopanib | 0.643 | 22759480, 28683470 |
| 10 | DB00878:Chlorhexidine* | 0.594 | 32581176 |

a. 'NA' represents that there has been no study proving that the drug can inhibit IL-6 release.

b. Drugs with '*' have been determined in clinical studies against COVID-19.

**Table S17.** Results of SSL based on different centralities for warm start predictions

| Methods | DDI | | DTI | |
|---|---|---|---|---|
| | AUROC±std | AUPR±std | AUROC±std | AUPR±std |
| DegreePre | 0.721±0.045 | 0.651±0.062 | 0.838±0.022 | 0.792±0.038 |
| EigenvectorPre | 0.673±0.052 | 0.596±**0.051** | 0.837±0.023 | 0.793±0.041 |
| ClusterPre | **0.793±0.009** | **0.745**±0.014 | **0.922±0.017** | **0.913±0.020** |

The best results are marked in **boldface**.

**Table S18.** Results of PairDistance with different "major" classes for warm start predictions

| Classes | DDI | | DTI | |
|---|---|---|---|---|
| | AUROC±std | AUPR±std | AUROC±std | AUPR±std |
| 3 | 0.816±0.008 | 0.769±0.016 | 0.899±0.014 | 0.866±0.027 |
| 4 | 0.818±0.007 | **0.779**±0.017 | **0.945**±0.015 | **0.933±0.016** |
| 5 | **0.820±0.003** | **0.779±0.008** | 0.911±**0.009** | 0.894±0.019 |

The best results are marked in **boldface**.

**Table S19.** Results of PathClass with meta path of different lengths for warm start predictions

| Path lengths | DDI | | DTI | |
|---|---|---|---|---|
| | AUROC±std | AUPR±std | AUROC±std | AUPR±std |
| 3 | 0.845±0.009 | 0.831±0.013 | 0.893±0.010 | 0.865±0.023 |
| 4 | **0.850±0.004** | **0.841±0.006** | **0.938**±0.019 | **0.931±0.013** |
| 5 | 0.845±0.006 | 0.832±0.010 | 0.900±**0.009** | 0.869±0.022 |

The best results are marked in **boldface**.

**Table S20.** Results of SimCon with different similarity measurements for warm start predictions

| Path lengths | DDI | | DTI | |
|---|---|---|---|---|
| | AUROC±std | AUPR±std | AUROC±std | AUPR±std |
| SimCon-ED | 0.821±0.008 | **0.805**±0.011 | 0.936±**0.006** | 0.923±0.013 |
| SimCon | **0.821±0.003** | 0.783±**0.007** | **0.946**±0.008 | **0.939±0.010** |

The best results are marked in **boldface**.

**Table S21.** Results of different combinations in PairDistance-EdgeMask-SimCon

| Scenarios | Methods | DDI | | DTI | |
|---|---|---|---|---|---|
| | | AUROC ±std | AUPR ±std | AUROC ±std | AUPR ±std |
| Warm start | EdgeMask-PairDistance | 0.917 ±0.003 | 0.912 ±0.005 | 0.958 ±0.007 | 0.951 ±0.009 |
| | PairDistance-SimCon | 0.880 ±0.006 | 0.860 ±0.012 | 0.942 ±0.007 | 0.935 ±0.008 |
| | EdgeMask-SimCon | 0.885 ±0.005 | 0.881 ±0.006 | 0.946 ±0.006 | 0.936 ±0.012 |
| | PairDistance-EdgeMask-SimCon | **0.939 ±0.002** | **0.937 ±0.002** | **0.969 ±0.006** | **0.968 ±0.007** |
| Cold start | EdgeMask-PairDistance | 0.889 ±0.014 | 0.852 ±0.017 | 0.927 ±0.030 | 0.890 ±0.077 |
| | PairDistance-SimCon | 0.858 ±0.016 | 0.837 ±0.020 | 0.903 ±0.037 | 0.878 ±0.065 |
| | EdgeMask-SimCon | 0.878 ±0.019 | 0.871 ±0.022 | 0.901 ±0.034 | 0.855 ±0.080 |
| | PairDistance-EdgeMask-SimCon | **0.909 ±0.008** | **0.895 ±0.011** | **0.940 ±0.020** | **0.915 ±0.048** |

The best results are marked in **boldface**.

**Table S22.** Results of ClusterPre-PathClass and PairDistance-EdgeMask-SimCon under ADL and ORC learning patterns for warm start predictions

| Methods | DDI | | DTI | |
|---|---|---|---|---|
| | AUROC ±std | AUPR ±std | AUROC ±std | AUPR ±std |
| CP-ADL | 0.879 ±0.006 | 0.877 ±0.006 | 0.927 ±0.008 | 0.897 ±0.018 |
| CP-ORC | **0.915** ±0.003 | **0.913** ±0.004 | 0.915 ±0.008 | 0.895 ±0.017 |
| ClusterPre-PathClass | **0.915 ±0.001** | 0.910 ±**0.003** | **0.956 ±0.007** | **0.956 ±0.007** |
| PES-ADL | 0.901 ±0.005 | 0.899 ±0.005 | 0.940 ±0.007 | 0.921 ±0.016 |
| PES-ORC | 0.932 ±**0.002** | 0.930 ±0.003 | 0.930 ±0.007 | 0.912 ±0.016 |
| PairDistance-EdgeMask-SimCon | **0.939 ±0.002** | **0.937 ±0.002** | **0.969 ±0.006** | **0.968 ±0.007** |

The best results are marked in **boldface**.

**Table S23.** Results of PairDistance-EdgeMask-SimCon under SVM and RF for warm start predictions

| Methods | DDI | | DTI | |
|---|---|---|---|---|
| | AUROC ±std | AUPR ±std | AUROC ±std | AUPR ±std |
| PES-SVM | 0.933 ±0.003 | 0.926 ±0.004 | 0.962 ±**0.004** | 0.963 ±**0.005** |
| PES-RF | **0.987 ±0.001** | **0.987** ±0.005 | 0.963 ±0.007 | 0.961 ±0.006 |
| PairDistance-EdgeMask-SimCon | 0.939 ±0.002 | 0.937 ±**0.002** | **0.969** ±0.006 | **0.968** ±0.007 |

The best results are marked in **boldface**.

**Table S24.** The comparisons of run-time and parameter sizes

| Methods | Run-time (s) | Parameters (M) |
|---|---|---|
| DeepR2cov | 11,556 | 354.421 |
| DDIMDL | **176** | 11.234 |
| DTINet | 382 | 1.523 |
| MoleculeNet | 135 | **0.172** |
| deepDTnet | 569 | 9.676 |
| PairDistance-EdgeMask-SimCon | 8,115 | 0.628 |

The best results are marked in **boldface**.

**Table S25.** The shortest path length and number of node pairs

| The shortest path length between node pairs | The number of node pairs |
|---|---|
| 1 | 221,140 |
| 2 | 3,359,298 |
| 3 | 5,042,652 |
| 4 | 582,024 |
| 5 | 15,048 |
| >=6 | 170 |
| Total | 4,177,680 |

**Table S26.** The numbers of nodes and edges in the constructed BioHN

| 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 |

**Table S27.** 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 |

**Table S28.** The total number of edges connected to drugs, proteins, and diseases, respectively.

| Node types | Drugs | Proteins | Diseases |
|---|---|---|---|
| The total number of edges | 72,563 | 44,191 | 24,281 |

**Table S29** The examples of multi-task combinations with different modalities

| Modal size | Multi-task combinations with different modalities | |
|---|---|---|
| 1 | SimReg-SimCon | ClusterPre-PairDistance |
| 2 | PairDistance-SimCon | ClusterPre-PathClass |
| 2 | PairDistance-SimReg-SimCon | ClusterPre-PairDistance-PathClass |
| 3 | PairDistance-EdgeMask-SimCon | ClusterPre-PathClass-SimReg |

## References


[1] Luo, Y. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8**, 1-13 (2017).

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

[3] Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. 9, 513-530 (2018).

[4] Ye, Q. et al. A unified drug-target interaction prediction framework based on knowledge graph and recommendation system. *Nat. Commun.* **12**, 1-12 (2021).

[5] Deng, Y. et al. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* **36**, 4316-4322 (2020).

[6] 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).

[7] Belkin,M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15**, 1373-1396 (2003).

[8] Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., and Smola, A. J. Distributed large-scale natural graph factorization. *In Proceedings of the 22nd International World Wide Web Conference* 37-48 (ACM, 2013).

[9] Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. *In Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining* 701-710 (ACM, 2014).

[10] Liu, B., & Tsoumakas, G. Optimizing Area Under the Curve Measures via Matrix Factorization for Predicting Drug-Target Interaction with Multiple Similarities. Preprint at https://arxiv.org/abs/2105.01545 (2021).

[11] Wang, Y., Min, Y., Chen, X., & Wu, J. Multi-view graph contrastive representation learning for drug-drug interaction prediction. *In Proceedings of the 30th Web Conference* 2921-2933 (ACM, 2021).

[12] Van Lahoven, T., Nabuurs, S. & Marchiori, E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics.* **27**, 3036-3043 (2011).

[13] Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. *In Proceedings of the 23rd International Conference on Machine learning.* 233-240 (ACM, 2006).

[14] Huang, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497-506 (2020).

[15] Mehta, P., et al. Covid-19: consider cytokine storm syndromes and immune-suppression. *Lancet* **395**, 1033-1034 (2020).

[16] Moore, J. B., & June, C. H. Cytokine release syndrome in severe COVID-19. *Science* **368**, 473-474 (2020).

[17] Morris, G. M., et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem.* 30, 2785-2791 (2009).

[18] Case, D. A. et al. The Amber biomolecular simulation programs. *J Chem Theory Comput.* 26, 1668-1688 (2005).

[19] Salomon-Ferrer, R., Gotz, A. W., Poole, D., Le Grand, S., & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J Chem Theory Comput* 9, 3878-3888 (2013).
[20] Hou, W., & Cronin, S. B. A review of surface plasmon resonance-enhanced photocatalysis. *Adv. Funct. Mater.* **23**, 1612-1619 (2013).
[21] Law, V., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091-7 (2014).
[22] Hernandez, T., et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* **36**, D913-8 (2008).
[23] Zhu, F., et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.* **40**, D1128-36 (2012).
[24] Sigman, M., & Cecchi, G. A. Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences* **99**, 1742-1747 (2002).
[25] Watts, D. J., & Strogatz, S. H. Collective dynamics of 'small-world'networks. *Nature* **393**, 440-442 (1998).
[26] Newman, M. E. J. A measure of betweenness centrality based on random walks. *Social networks* **27**, 39-54 (2005).
[27] Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry*, 35-41 (1977).
[28] Costa, L. D. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. Characterization of complex networks: A survey of measurements. *Advances in physics* **56**, 167-242 (2007).
[29] Peng, Z., Dong, Y., Luo, M., Wu, X. M., & Zheng, Q. Self-supervised graph representation learning via global context prediction. Preprint at https://arxiv.org/abs/2003.01604 (2020).
[30] Fu, G. et al. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC bioinformatics* **17**, 1-10 (2016).
[31] Wu, G, Liu, J. & Yue, X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC bioinformatics* **20**, 1-13 (2019).
[32] Breiman L. Random forests. *Machine learning* **45**, 5-32 (2001).
[33] Cortes, C., & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273-297 (1995).
[34] Brown, T., et al. Language models are few-shot learners. *In Proceedings of the 33rd International Conference on Neural Information Processing Systems.* **33**, 1877-1901 (2020).
[35] Chowdhery, A., et al. Palm: Scaling language modeling with pathways. Preprint at https://doi.org/10.48550/arXiv.2204.02311 (2022).
[36] Hamilton, W. L., Ying, R., & Leskovec, J. Inductive representation learning on large graphs. *In Proceedings of the 31st International Conference on Neural Information Processing Systems* 1025-1035 (MIT Press, 2017).
[37] Chen, J., Zhu, J., & Song, L. Stochastic training of graph convolutional networks with variance reduction. *In Proceedings of the 35th International Conference on Machine Learning* 80 941-949 (2018).

[38] Ying, R., et al. Graph convolutional neural networks for web-scale recommender systems. *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* 974-983 (2018).
[39] Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41-42 (2001).
[40] Van Noort, V., Snel, B., & Huynen, M. A. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, **5**, 280-284 (2004).
[41] Wang, X. et al. BioERP: biomedical heterogeneous network-based self-supervised representation learning approach for entity relationship predictions. *Bioinformatics* **37**, 4793-4800 (2021).
[42] Huang, J. et al. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Comput Biol* **9**, e1002998 (2013).
[43] Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, **29**, 17-37 (2016).
[44] Vilar, S. et al. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat. Protoc.* **9**, 2147-2163 (2014).
[45] Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708 (1982).
[46] Menche, Jörg, et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 6224 (2015).
[47] 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).
[48] Ni, P. et al. Constructing disease similarity networks based on disease module theory. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 906-915 (2018).
[49] Glorot, X., & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* 249-256 (JMLR, 2010).
[50] Kingma, D. P., & Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations* (OpenReview.net, 2015).
[51] Yue, X., et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* **36**, 1241-1251 (2020).
[52] Lin, X., et al. KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction. *In Proceedings of the 29th International Joint Conference on Artificial Intelligence* **380**, 2739-2745 (Morgan Kaufmann, 2020).
[53] Pan, S., Wu, J., Zhu, X., Zhang, C., & Wang, Y. Tri-party deep network representation. *In Proceedings of the 25th International Joint Conference on Artificial Intelligence* 1895-1901 (Morgan Kaufmann, 2016).
[54] Huang, X., Li, J., & Hu, X. Label informed attributed network embedding. *In proceedings of the 10th ACM international conference on web search and data mining* 731-739 (ACM, 2017).

[55] Li, C., et al. PPNE: property preserving network embedding. *In proceedings of the 22nd International Conference on Database Systems for Advanced Applications* 163-179 (Springer, 2017).
[56] Natarajan N, & Dhillon I S. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, **30**, i60-i68 (2014).
[57] Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient estimation of word representations in vector space. *In Proceedings of the* 1st International Conference on Learning Representations. (OpenReview.net, 2013).