

Supplementary Information

HEDDI-Net: Heterogeneous Network Embedding for Drug-Disease Association Prediction and Drug Repurposing, with Application to Alzheimer's Disease

Yin-Yuan Su¹, Hsuan-Cheng Huang¹, Yu-Tin Lin¹, Yi-Fang Chuang^{2,3,4}, Sheh-Yi Sheu^{1,5}, Chen-Ching Lin^{1*}

¹Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan.

²School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan.

³Institute of Public Health, National Yang Ming Chiao Tung University, Taipei, Taiwan.

⁴Department of Psychiatry, Far Eastern Memorial Hospital, New Taipei, Taiwan.

⁵Department of Life Science and Institute of Genome Science, National Yang-Ming University, Taipei, Taiwan.

Supplementary Information	1
SI Materials and Methods	2
Introduction of benchmarked methods	2
SI Figures	2
Figure S1: Distribution of zero similarity between the top 500 eigen-diseases and all CTD diseases	2
Figure S2: Distribution of eigenvector centrality in the MeSH tree	3
SI Tables	4
Table S1. Physicochemical properties used in this study.....	4
Table S2. Performance of comparison methods	5
Table S3. Performance of heterogeneous representative proteins	6
Table S4. Performance of different classifiers.....	7
References:.....	8

SI Materials and Methods

Introduction of benchmarked methods

- DRHGCN (1) employs an intra-domain feature extraction module to extract drug and disease embeddings based on similarity networks, and an inter-domain feature extraction module for message passing between drugs and diseases. The inter- and intra-domain embeddings are fused using a layer attention mechanism to obtain the final representation of drugs and diseases, improving prediction performance.
- DRWBNCF (2) uses a neural collaborative filtering approach based on neighborhood and neighborhood interactions. It constructs drug and disease relevant networks and integrates the information using a weighted bilinear graph convolution operation. Finally, a multi-layer perceptron optimized by the α -balanced focal loss function and graph regularization to model the drug-disease associations and infer potential new drugs for diseases.
- LAGCN (3) utilizes graph convolutional networks to capture structural information of drugs and diseases. It incorporates an attention mechanism to adaptively combine embeddings from multiple convolution layers, resulting in a more informative representation of drugs and diseases.
- HINGRL (4) is a novel drug repositioning model that learns the features of drugs and diseases from both topological and biological perspectives using different representation learning technique.

SI Figures

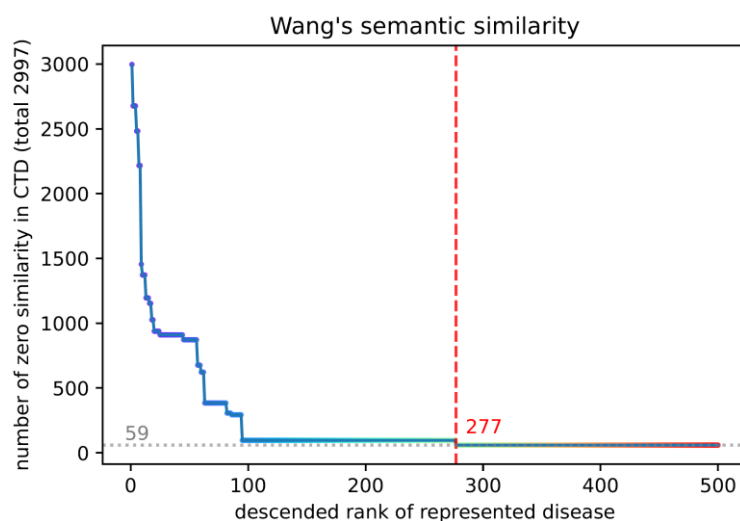


Figure S1: Distribution of zero similarity between the top 500 eigen-diseases and all CTD diseases

The x-axis represents the top 500 eigen-diseases ranked by descending eigenvector centrality. The y-axis represents the number of diseases in the CTD database with zero similarity based on the accumulated eigen-diseases, each having an eigenvector centrality no less than the corresponding disease on the x-axis. In the figure, it is evident that as the eigen-diseases accumulate up to 277, the number of diseases with zero similarity in the CTD dataset ceases to decrease. Therefore, we selected the diseases corresponding to the first 277 eigenvector centralities as our representative diseases.

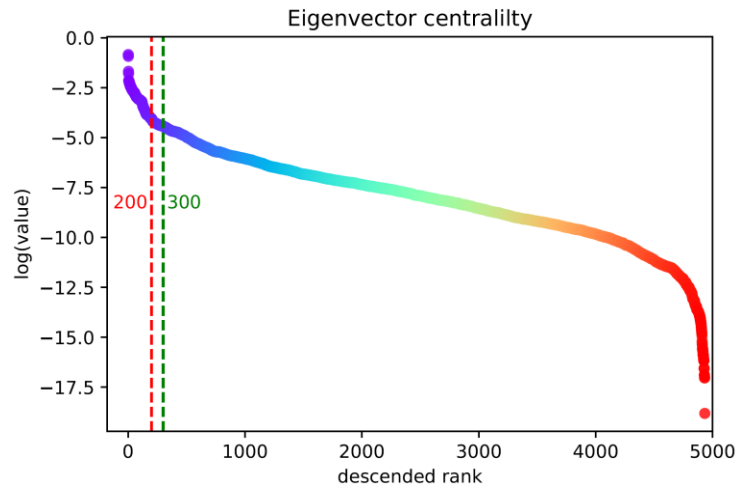


Figure S2: Distribution of eigenvector centrality in the MeSH tree

The x-axis represents diseases in the MeSH tree, ranked by descending eigenvector centrality. The y-axis represents the logarithmic value of the eigenvector centrality for each disease in the undirected MeSH tree. Notably, the figure illustrates that there is a turning point in the eigenvector centrality values for diseases ranging from the first 200 to 300 eigen-diseases.

SI Tables

Table S1. Physicochemical properties used in this study.

Molecule properties	Description	Molecule properties	Description
full_mwt	Molecular weight	mw_freebase	Molecular weight of the parent form of a molecule ^I .
mw_monoisotopic	The sum of the masses of the most abundant isotopes in the compound	alogP	Partition coefficient, is used to evaluate the lipophilicity or hydrophobic of a molecule ^{II} .
acd_logP	The calculated octanol/water partition coefficient using ChemAxon v17.29.0	acd_logD	The calculated octanol/water distribution coefficient at pH7.4 using ChemAxon v17.29.0.
aromatic_rings	The number of aromatic rings atoms in a molecule.	heavy_atoms	The number of non-hydrogen atoms in a molecule.
hba	The number of hydrogen bond acceptors.	hbd	The number of hydrogen bond donors.
hba_lipinski	The count of nitrogen and oxygen atoms in the molecule.	hbd_lipinski	The count of hydrogens attached to nitrogen or oxygen atoms.
num_ro5_violations	The number of properties in a molecule that fails to meet the definition in Lipinski's rule of 5 (RO5).	num_lipinski_ro5_violations	Replacing "hba" and "hbd" by "hba_lipinski" and "hbd_lipinski" in "num_ro5_violations".
rtb	The number of rotatable bonds in a molecule.	psa	The polar surface area ^{III} .
qed_weighted	A quantitative estimate of drug-likeness ^{IV} .		

^I A parent structure is a compound that consists of an unbranched chain of skeletal atoms, according to IUPAC nomenclature. In pharmacokinetics, logP is a significant indicator of ADME properties, since a drug needs to be lipophilic to pass through lipid bilayers in the intestinal epithelium after taken orally, but still hydrophilic enough to partition out again. As for pharmacodynamics, more lipophilicity means more toxic because a drug with high logP tends to be retained longer within the body.

^{II} AlogP is measured as the ratio of concentrations of molecular species in different aqueous media, usually octanol and water. The larger the logP value is, the higher the affinity of a molecule to the organic phase is.

^{III} PSA is defined as the sum of surface area over all polar atoms, chiefly oxygen and nitrogen and their attached hydrogen atoms.

^{IV} The value ranges from 0-1 where 1 is the most drug-like and 0 the least drug-like.

Table S2. Performance of comparison methods

all negative samples								
method / median (10 fold)	AUPR	AUC	F1	ACC	Recall	Specificity	Precision	Wilcoxon/t-test <i>p</i> -value*
HEDDI-Net	0.6273	0.9139	0.57	0.8941	0.596	0.9296	0.5284	0.078/0.1519 0.0078/0.0052 0.0078/0.0026 0.078/0.0292
HINGRL	0.588	0.8955	0.5648	0.8921	0.6116	0.9281	0.5279	
DRHGCN	0.5416	0.8703	0.5254	0.8793	0.5839	0.9178	0.4776	
DRWBNCF	0.4483	0.8379	0.4621	0.8573	0.5422	0.9003	0.4038	
LAGCN	0.27	0.8775	0.2278	0.9688	0.3633	0.9767	0.1661	
balanced dataset								
method / median (10 fold)	AUPR	AUC	F1	ACC	Recall	Specificity	Precision	Wilcoxon/t-test <i>p</i> -value*
HEDDI-Net	0.9334	0.9348	0.8574	0.8464	0.9188	0.7752	0.8071	0.0078/2.36 e-08
HINGRL	0.8816	0.8879	0.8143	0.7994	0.8765	0.7234	0.7595	

* The Wilcoxon signed-rank test and t-test were conducted to assess the right-tail significance of performance differences across all metrics between HEDDI-Net and the compared methods.

Table S3. Performance of heterogeneous representative proteins

model / median (10 fold)	AUPR	AUC	F1	ACC	Recall	Specificity	Precision	Variance
direct evidence ge5	0.9800	0.9802	0.9274	0.9259	0.9467	0.9054	0.9083	0.0969%
direct evidence ge4	0.9823	0.9823	0.9302	0.9294	0.9412	0.9125	0.9153	0.0855%
direct evidence ge3	0.9801	0.9801	0.9275	0.9260	0.9445	0.9061	0.9089	0.0949%
therapeutic ge5	0.9827	0.9834	0.9356	0.9345	0.9494	0.9164	0.9193	0.0764%
therapeutic ge4	0.9861	0.9862	0.9427	0.9423	0.9421	0.9360	0.9368	0.0515%
therapeutic ge3	0.9775	0.9785	0.9281	0.9272	0.9384	0.9158	0.9181	0.0710%

- ge5: 182 representative proteins with median z-score greater than or equal to 5
- ge4: 228 representative proteins with median z-score greater than or equal to 4
- ge3: 311 representative proteins with median z-score greater than or equal to 3

Table S4. Performance of different classifiers.

model / median (10 fold)	AUPR	AUC	F1	ACC	Recall	Specificity	Precision
direct evidence dataset							
DL	0.9823	0.9823	0.9302	0.9294	0.9412	0.9125	0.9153
LR	0.7773	0.7925	0.7437	0.7026	0.8636	0.5432	0.6541
linear SVC	0.7713	0.7874	0.7403	0.6995	0.8561	0.5418	0.6514
RBF SVC	0.8981	0.9003	0.8288	0.8233	0.8601	0.7866	0.8015
RF	0.9049	0.9071	0.8344	0.8272	0.8769	0.7749	0.7964
XGB	0.9335	0.9319	0.8630	0.8604	0.8829	0.8346	0.8423
therapeutic dataset							
DL	0.9861	0.9862	0.9427	0.9423	0.9421	0.9360	0.9368
LR	0.7575	0.7706	0.7261	0.6862	0.8406	0.5321	0.6413
linear SVC	0.7494	0.7637	0.7247	0.6793	0.8535	0.5059	0.6319
RBF SVC	0.8640	0.8627	0.7906	0.7790	0.8376	0.7210	0.7494
RF	0.8875	0.8857	0.8162	0.8072	0.8566	0.7708	0.7885
XGB	0.9153	0.9116	0.8393	0.8354	0.8595	0.8122	0.8209
benchmark dataset							
DL	0.9334	0.9348	0.8574	0.8464	0.9188	0.7752	0.8071
LR	0.7043	0.7245	0.7062	0.6371	0.8711	0.4038	0.5935
linear SVC	0.6963	0.7175	0.7044	0.6423	0.8610	0.4300	0.6013
RBF SVC	0.7731	0.7863	0.7406	0.7050	0.8511	0.5542	0.6560
RF	0.7804	0.7986	0.7509	0.7166	0.8597	0.5752	0.6685
XGB	0.8050	0.8148	0.7589	0.7265	0.8741	0.5686	0.6721

DL: Deep Learning; LR: Logistic Regression; SVC: Support Vector Classification; RBF: Radial Basis Function kernel; RF: Random Forest; XGB: eXtreme Gradient Boosting.

References:

1. Cai L, Lu C, Xu J, Meng Y, Wang P, Fu X, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Brief Bioinform.* 2021;22(6).
2. Meng Y, Lu C, Jin M, Xu J, Zeng X, Yang J. A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief Bioinform.* 2022;23(2).
3. Yu Z, Huang F, Zhao X, Xiao W, Zhang W. Predicting drug-disease associations through layer attention graph convolutional network. *Brief Bioinform.* 2021;22(4).
4. Zhao BW, Hu L, You ZH, Wang L, Su XR. HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. *Brief Bioinform.* 2022;23(1).