SUPPLEMENTARY INFORMATION

Title: Network-based Prediction of Drug-Target Interactions using an Arbitrary-Order Proximity Embedded Deep Forest

Zeng et al., *Bioinformatics* 2020

*To whom correspondence should be addressed:

Feixiong Cheng, PhD

Genomic Medicine Institute

Lerner Research Institute, Cleveland Clinic

9500 Euclid Avenue, Cleveland, Ohio 44195

Email: chengf@ccf.org

Phone: +1-216-4447654

Fax: +1-216-6361609

Supplementary Information files contain Supplemental Methods, 8 Supplementary Tables, 2 Supplementary Figures, and Supplementary References.

Details for reconstruction of the drug-target interaction network

We assembled high-quality physical drug-target interactions on FDA-approved drugs from 6 commonly used data sources, and defined a physical drug-target interaction using reported binding affinity data: inhibition constant/potency (K_i), dissociation constant (K_d), median effective concentration (EC_{50}), or median inhibitory concentration (IC_{50}) ≤ 10 μM. Drug-target interactions were acquired from the DrugBank database (v4.3)(Law, et al., 2014), the Therapeutic Target Database (TTD, v4.3.02)(Zhu, et al., 2012), and the PharmGKB database(Hernandez-Boussard, et al., 2008). Specifically, bioactivity data of drug-target pairs were collected from three widely used databases: ChEMBL (v20)(Gaulton, et al., 2012), BindingDB(Liu, et al., 2007), and IUPHAR/BPS Guide to PHARMACOLOGY(Pawson, et al., 2014). In total, 4,978 drug-target interactions connecting 732 FDA-approved drugs and 1,915 unique human targets (proteins) were used.

Details for building the human protein-protein interactome

To build a comprehensive human protein-protein interactome, we assembled data from a total of 15 bioinformatics and systems biology databases with multiple experimental evidences. Specifically, we focused on high-quality protein-protein interactions (PPIs) with five types of experimental evidences: (i) Binary PPIs tested by high-throughput yeast-two-hybrid (Y2H) systems from two public available high-

quality Y2H datasets(Rolland, et al., 2014; Rual, et al., 2005) and one unpublished dataset, publicly available at: http://ccsb.dana-farber.org/ interactome-data.html; (ii) Kinase-substrate interactions by literature-derived low-throughput or high-throughput experiments from KinomeNetworkX (Cheng, et al., 2014), Human Protein Resource Database (HPRD) (Peri, et al., 2004), PhosphoNetworks(Hu, et al., 2014; Newman, et al., 2013), PhosphositePlus (Hornbeck, et al., 2015), dbPTM 3.0 (Lu, et al., 2013), and Phospho. ELM (Dinkel, et al., 2011); (iii) Literature-curated PPIs identified by affinity purification followed by mass spectrometry (AP-MS), Y2H, or by literature-derived low-throughput experiments from BioGRID (Chatr-Aryamontri, et al., 2015), PINA (Cowley, et al., 2012), HPRD (Peri, et al., 2004), MINT (Licata, et al., 2012), IntAct (Orchard, et al., 2014), and InnateDB (Breuer, et al., 2013); (iv) Signaling network by literature-derived low-throughput experiments as annotated in SignaLink2.0(Fazekas, et al., 2013); and (v) Binary, physical PPIs from protein three-dimensional (3D) structures reported in Instruct (Meyer, et al., 2013). In this study, all inferred data, including evolutionary analysis, gene expression data, and metabolic associations, were excluded. The resulting human protein-protein interactome used in this study includes 16,133 PPIs connecting 1,915 unique drug-target coding gene products.

Details for collecting drug-drug interactions

Drug-drug interactions (DDIs). We collected clinically reported DDI data from the DrugBank database (v4.3) (Law, et al., 2014). Chemical name, generic name or commercial name of each drug were standardized by Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS) vocabularies (Bodenreider, 2004) and further transferred to DrugBank ID. In total, 132,768 clinically reported DDIs connecting 732 unique FDA-approved drugs were used.

Description of re-constructing drug-disease network

We collected the known drug-disease associations from several public resources, including repoDB (Brown and Patel, 2017), DrugBank (v4.3) (Law, et al., 2014), and Drug Central (Ursu, et al., 2017) databases. Compound name, generic name or commercial name of each drugs and disease names were standardized by MeSH and UMLS vocabularies (Bodenreider, 2004). In total, 1,208 drug-disease pairs connecting 732 drugs and 440 diseases were used.

Description of re-constructing drug-side effect network

We collected the clinically reported drug side effects or adverse drug event (ADE) information by assembling data from MetaADEDB (Cheng, et al., 2013), CTD (Davis, et al., 2011), SIDER (version 2) (Tatonetti, et al., 2012), and OFFSIDES

(Kuhn, et al., 2010). Only ADE data with clinically reported evidence were used. In total, 263,805 drug−ADE associations collecting 732 approved drugs and 12,904 ADEs were used.

Chemical similarity analysis of drug pairs

We downloaded chemical structure information (SMILES format) from the DrugBank database and computed MACCS fingerprints of each drug using Open Babel v2.3.1(O'Boyle, et al., 2011). If two drug molecules have *a* and *b* bits set in their MACCS fragment bit-strings, with *c* of these bits being set in the fingerprints of both drugs, the Tanimoto coefficient (*T*) (Willett, 2006) of a drug-drug pair is defined as:

$$T = \frac{c}{a+b-c} \qquad \text{(S1)}$$

Protein sequence similarity (identity) analysis

We downloaded the canonical protein sequences of drug targets (proteins) in *Homo sapiens* from Uniprot database (http://www.uniprot.org/).

Similarity of drug targets. We calculated the protein sequence similarity $S_p(a, b)$ of two drug targets *a* and *b* using the Smith-Waterman algorithm (Smith and Waterman, 1981).

Similarity of drug pairs. The overall sequence similarity of the drug targets

binding two drugs *A* and *B* is determined by equation (S2) by averaging all pairs of proteins *a* and *b* with $a \in A$ and $b \in B$ under the condition $a \neq b$. This condition ensures that for drugs with common targets we do not take pairs into account where a target would be compared to itself.

$$< \mathrm{S}_p >= \frac{1}{n_{pairs}} \sum_{\{a,b\}} \mathrm{S}_p(a,b) \qquad \text{(S2)}$$

Gene co-expression analysis for drug targets

We downloaded the RNA-seq data (RPKM value) across 32 tissues from GTEx V6 release (https://gtexportal.org/home/). For each tissue, we regarded those genes with RPKM ≥ 1 in more than 80% samples as tissue-expressed genes.

Co-expression analysis of drug targets. To measure the extent to which drug target-coding genes (*a* and *b*) associated with the drug-treated diseases are co-expressed, we calculated the Pearson's correlation coefficient ($\mathrm{PCC}(a,b)$) and the corresponding p-value via F-statistics for each pair of drug target-coding genes *a* and *b* across 32 human tissues. In order to reduce the noise of co-expression analysis, we mapped $\mathrm{PCC}(a,b)$ into the human protein-protein interactome network to build a co-expressed protein-protein interactome network as described previously (Cheng, et al., 2014).

Co-expression analysis of drug pairs. The co-expression similarity of the drug target-coding genes associated with two drugs *A* and *B* is computed by averaging

$\mathrm{PCC}(a,b)$ over all pairs of targets *a* and *b* with $a \in A$ and $b \in B$ as below:

$$< \mathrm{S}_{co} >= \frac{1}{n_{pairs}} \sum_{\{a,b\}} |\mathrm{PCC}(a,b)| \qquad \text{(S3)}$$

Gene Ontology (GO) similarity analysis for drug targets

We downloaded the Gene Ontology (GO) annotation for all drug target-coding genes from website: http://www.geneontology.org/. We used three types of the experimentally validated or literature-derived evidences: *cellular component (CC)*, *biological processes (BP)*, and *molecular function (MF)*.

Similarity of drug targets. We computed GO similarity $\mathrm{S}_{GO}(a,b)$ for each pair of drug target-coding genes *a* and *b* using a graph-based semantic similarity measure algorithm (Wang, et al., 2007) and GOSemSim (Yu, et al., 2010).

Similarity of drug pairs. The overall GO similarity of the drug target-coding genes binding to two drugs *A* and *B* is determined by equation (S4), averaging all pairs of drug target-coding genes *a* and *b* with $a \in A$ and $b \in B$.

$$< \mathrm{S}_{GO} >= \frac{1}{n_{pairs}} \sum_{\{a,b\}} \mathrm{S}_{GO}(a,b) \qquad \text{(S4)}$$

Clinical similarity analysis for drug pairs

We computed clinical similarities of drug pairs derived from the drug Anatomical Therapeutic Chemical (ATC) classification systems (Cheng, et al., 2013). We downloaded all ATC codes from the DrugBank database (v4.3) (Law, et al., 2014).

The *k*th level drug clinical similarity (S_k) of drugs *A* and *B* is defined via the ATC codes as below.

$$S_k(A,B) = \frac{\mathrm{ATC}_k(A) \cap \mathrm{ATC}_k(B)}{\mathrm{ATC}_k(A) \cup \mathrm{ATC}_k(B)} \quad \text{(S5)}$$

where ATC_k represents all ATC codes at the *k*th level. A score $S_{atc}(A,B)$ is used to define the clinical similarity between drugs *A* and *B*:

$$S_{atc}(A,B) = \frac{\sum_{k=1}^{n} S_k(A,B)}{n} \quad \text{(S6)}$$

Where n represents the five levels of ATC codes (ranging from 1 to 5).

Description of collecting disease-gene network

We assembled disease-gene annotation data from three commonly used database: 1) The OMIM database (http://www.omim.org/) (Amberger, et al., 2015); 2) The Comparative Toxicogenomics Database (http://ctdbase.org/) (Davis, et al., 2015); 3) HuGE Navigator (Yu, et al., 2008). We annotated all protein-coding genes using gene Entrez ID, chromosomal location, and the official gene symbols from the NCBI database(Coordinators, 2016). In total, 23,080 disease-genes pairs connecting 440 diseases and 1,915 drug targets-coding genes were used.

Preprocess of association networks

For the homogeneous interaction networks (e.g., drug-drug interaction network) and similarity networks (e.g., drug chemical similarity network), we use AROPE

to extract features from these networks directly. For the association networks, i.e., drug-disease, drug-side-effect, and protein-disease networks, we construct the corresponding similarity networks based on the Jaccard similarity coefficient first, and then run the AROPE model on these similarity networks. Jaccard similarity is a common statistic used for characterizing the similarity and diversity between two sets of samples. Taking the drug-disease association network as an example, we use the following formula to measure the similarity between drug i and drug j:

$$\mathrm{Sim}(i,j) = \frac{|Disease_i \cap Disease_j|}{|Disease_i \cup Disease_j|} \qquad \text{(S7)}$$

Where $Disease_i$ denotes the set of diseases of drug i. Then we run the AROPE algorithm on this similarity network to obtain the feature representation of drugs. In the same manner, we can construct the similarity networks of proteins.

In summary, we construct 8 types of similarity networks for drugs, based on (1) drug-disease associations, (2) drug-side-effect associations, (3) chemical structures, (4) therapeutic similarity, (5) primary protein sequence-derived drug-drug similarity, (6) biological process, (7) cellular component, (8) molecular function. Similarly, we construct 5 types of similarity networks for proteins, based on (1) gene/protein-disease associations, (2) primary protein sequence, (3) biological process, (4) cellular component, (5) molecular function. With 13 similarity networks and another two interaction networks (drug-drug interactions and protein-protein interactions), we can learn the low-dimensional feature vector

representations of drugs and proteins through network embedding scheme.

Baseline Methods

We compare our method against five previously-proposed methods, including NeoDTI (Wan, et al., 2018), deepDTnet (Zeng), RLSWNN (van Laarhoven and Marchiori, 2013), KBMF2K (Gonen, 2012) and NetLapRLS (Xia, et al., 2010). Among these methods, NeoDTI, and deepDTnet can integrate multiple heterogeneous information to predict new DTIs, while RLSWNN, KBMF2K and NetLapRLS are not particularly designed to exploit multiple drug or protein network data for DTI prediction. To make a fair comparison, we followed the same strategy as NeoDTI (Wan, et al., 2018) to integrate multiple networks into a single network for RLSWNN, KBMF2K and NetLapRLS in our comparison tests. In particular, we combined multiple networks into a single network by assigning the edge weight $p_{i,j} = 1 - \prod_k \left(1 - p_{i,j}^{(k)}\right)$, where $p_{i,j}^{(k)} \epsilon [0,1]$ is the interaction probability or similarity between node i and node j in network $k \epsilon \{1,2, \dots, K\}$, where K stands for the total number of networks. For the hyperparameters used in the baseline methods, we tuned them to get the best performance in the cross validation. We will briefly describe these methods below.

1. **NeoDTI:** Neural integration of neighbor information for DTI prediction is a nonlinear end-to-end learning model that integrates diverse information from heterogeneous network data via a number of information passing and

aggregation operations, and automatically learns topology-preserving representations of drugs and targets to make predictions. We chose node embedding dimension d from $\{256,512,1024\}$, the dimension of the edge-type specific projection matrices k from $\{256,512,1024\}$. We used the Adam optimizer (Kingma and Ba, 2014) with the learning rate 0.001 to perform gradient descent.

2. **deepDTnet:** A network-based, deep learning methodology (Zeng, et al., 2019) for drug repositioning, that integrates a deep neural network algorithm for network embedding, which embeds each vertex in a network into a low-dimensional vector space, and a Positive-Unlabeled (PU)-matrix completion algorithm for prediction, which is a vector space projection scheme for predict drug-target interactions. We designed different network architecture with different number of layers and different number of hidden nodes. We chose embedding dimension d of each network from $\{50,100,200\}$, according to the prediction performance. The biased value α and regulation parameter λ in PU-matrix completion are selected over the grid search. Specifically, we chose α from $\{0.1,0.2,\dots,1\}$, we chose λ from $\{0.005,0.01,0.05,0.1,0.2\}$.
3. **RLSWNN:** Regularized Least Squares with Weighted Nearest Neighbors (van Laarhoven and Marchiori, 2013), which uses a weighted nearest neighbor procedure for inferring a profile for a drug compound by using interaction profiles of the compounds in the training data. The regularization parameter

was selected from $\{0.1,0.2,\ldots,1\}$.

4. **KBMF2K:** Kernelized Bayesian matrix Factorization method (Gonen, 2012), which uses a kernelized Bayesian matrix factorization with twin kernels to predict drug-target interactions. KBMF2K combines dimensionality reduction, matrix factorization and binary classification for predicting drug-target interaction networks using only chemical similarity between drugs compounds and genomic similarity between target proteins. This approach proposed a joint Bayesian formulation of projecting drug compounds and target proteins into a unified subspace using the similarities and estimating the interaction network in that subspace. The subspace dimensionality parameter R was chosen from {5,10, …,40}.
5. **NetLapRLS:** An algorithm that is based on the bipartite local model concept (Xia, et al., 2010), which perform two sets of predictions, one from the drug side and one from the target side, and then aggregates these predictions to give the final prediction scores for the potential interaction candidates. The ratios $\lambda_{d2}/\lambda_{d1}$ and $\lambda_{p2}/\lambda_{p1}$ were chosen from $\{10^{-5},10^{-4},\ldots,10\}$ and the parameters β_d and β_p were selected from $\{3\cdot10^{-4},3\cdot10^{-3},\ldots,3\cdot100\}$

Supplementary Tables

Supplementary Table S1. The number of nodes of individual types in the constructed heterogeneous drug-target-disease network.

Type of node	Count
Drug	732
Protein	1,915
Disease	440
Side-effect	12,904
Total	15,991

Supplementary Table S2. The size of individual networks or association matrices in the constructed heterogeneous network.

Type of edge	Count
Drug-Protein	4,978
Drug-Drug	132,768
Drug-Disease	1,208
Drug-Side-effect	263,805
Protein-Protein	16,133
Protein-Disease	23,080
Total	441,972

Supplementary Table S3. Overlap analysis of two external validations collected from the DrugCentral database (Ursu, et al., 2018) and ChEMBL database (Mendez, et al., 2018) respectively.

	# of drugs	#of proteins	# of interactions
DrugCentral	446	483	1507
ChEMBL	559	826	3034
overlap	371	409	589

Supplementary Table S4. The area under the receiver operating characteristic curve (AUROC) and the area under precision-recall curve (AUPR) during cross-validation on the gold standard drug-target network. We performed 10 times random 5-fold cross-validation and standard derivation was provided.

Methods	AUROC	AUPR
AOPEDF	**0.985±0.0009**	**0.985±0.0009**
NeoDTI	0.971±0.0017	0.970±0.0019
deepDTnet	0.965±0.0011	0.969±0.0013
RLSWNN	0.949±0.0024	0.955±0.0029
KBMF2K	0.936±0.0011	0.947±0.0012
NetLapRLS	0.923±0.0018	0.936±0.0013

Supplementary Table S5. Performance of ablation analysis of different components implemented in AOPEDF.

	AUROC	AUPR
$LINE_{1st}$+deep forest	0.969	0.969
$LINE_{2st}$+deep forest	0.971	0.973
AROPE+SVM	0.957	0.943
AROPE+RF	0.977	0.977
AROPE+DNN	0.976	0.975
$LINE_{1st}$+SVM	0.929	0.922
$LINE_{1st}$+RF	0.964	0.964
$LINE_{1st}$+DNN	0.955	0.953
$LINE_{2st}$+SVM	0.941	0.926
$LINE_{2st}$+RF	0.965	0.967
$LINE_{2st}$+DNN	0.961	0.958
AOPEDF	**0.985**	**0.985**

Note: AUROC: the area under ROC curve; AUPR; the area under PR; RF: standard random forest; SVM: support vector machine; DNN: deep neural network.

Supplementary Table S6. Performance of the AOPEDF models built from 15 single network separately and the total 15 networks.

Network	AUROC	AUPR
Drug-drug	0.8020±0.0050	0.7960±0.0099
Drug-disease	0.8054±0.0030	0.8067±0.0034
Drug-side-effect	0.8095±0.0040	0.8091±0.0032
Drugsim1	0.8083±0.0025	0.8011±0.0044
Drugsim2	0.8058±0.0062	0.8011±0.0108
Drugsim3	0.8099±0.0098	0.8034±0.0136
Drugsim4	0.8118±0.0016	0.8031±0.0023
Drugsim5	0.8121±0.0036	0.8113±0.0088
Drugsim6	0.8046±0.0019	0.7910±0.0047
Drugs	0.8204±0.0017	0.8187±0.0015
Protein-protein	0.8877±0.0021	0.8816±0.0034
Protein-disease	0.8889±0.0031	0.8869±0.0032
Proteinsim1	0.8982±0.0046	0.8858±0.0060
Proteinsim2	0.8920±0.0015	0.8817±0.0040
Proteinsim3	0.8654±0.0023	0.8580±0.0044
Proteinsim4	0.8949±0.0023	0.8872±0.0024
Proteins	0.9020±0.0021	0.8947±0.0005
Total	**0.9831±0.0008**	**0.9840±0.0007**

Note:
Drugsim1: using drug chemical similarity network
Drugsim2: using drug therapeutic similarity network
Drugsim3: using drug target sequence similarity network
Drugsim4: using drug Gene Ontology (GO) biological process similarity network
Drugsim5: using drug GO cellular component similarity network
Drugsim6: using drug GO molecular function similarity network
Drugs: using all drug-related networks
Proteinsim1: using protein sequence similarity network
Proteinsim2: using protein Gene Ontology (GO) biological process similarity network
Proteinsim3: using protein GO cellular component similarity network
Proteinsim4: using protein GO molecular function similarity network
Proteins: using all protein-related networks
Total: using all 15 networks.

We used each network separately to validate the contribution of integrating 15 networks, and observe the influence of each single network. The deep forest we

use contains two random forests, two completely-random tree forests and two gradient boosting tree forests, each forest contains 100 trees. From the experiment results we can find that integrating multiple networks performs better than using single network. Besides, using both drug-related network and protein-related network performs better than using only one kind of network.

Supplementary Table S7. Performance of the AOPEDF models built from 14 network separately by leaving each network out separately.

Remove	AUROC	AUPR
Drug-drug	0.8621±0.0062	0.8649±0.0083
Drug-disease	0.8646±0.0050	0.8686±0.0056
Drug-side effect	0.8622±0.0041	0.8673±0.0078
Drugsim1	0.8608±0.0066	0.8652±0.0084
Drugsim2	0.8640±0.0070	0.8683±0.0092
Drugsim3	0.8680±0.0052	0.8686±0.0093
Drugsim4	0.8567±0.0112	0.8544±0.0119
Drugsim5	0.8643±0.0043	0.8640±0.0074
Drugsim6	0.8563±0.0092	0.8576±0.0096
Protein-protein	0.8517±0.0114	0.8563±0.0098
Protein-disease	0.8475±0.0099	0.8582±0.0106
Proteinsim1	0.8579±0.0070	0.8568±0.0034
Proteinsim2	0.8643±0.0125	0.8669±0.0113
Proteinsim3	0.8577±0.0060	0.8583±0.0101
Proteinsim4	0.8540±0.0056	0.8610±0.0094
All 15 networks	**0.8682±0.0066**	**0.8698±0.0050**

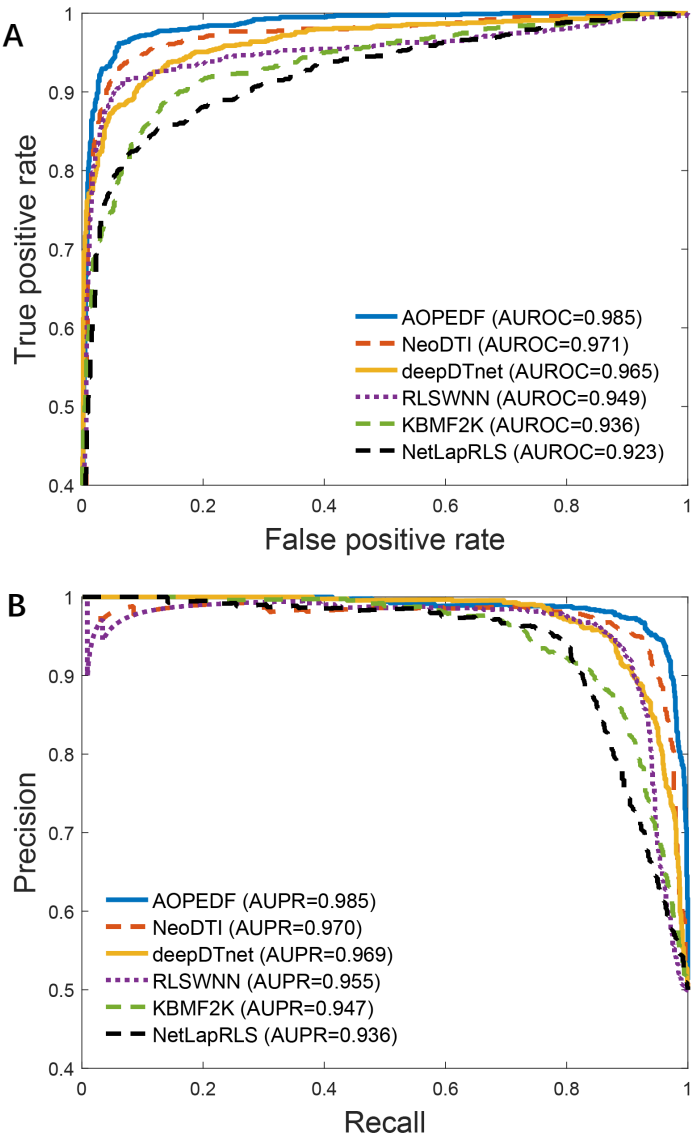
Note:
Drugsim1: remove drug chemical similarity network
Drugsim2: remove drug therapeutic similarity network
Drugsim3: remove drug sequence-derived drug-drug similarity network
Drugsim4: remove drug Gene Ontology (GO) biological process similarity network
Drugsim5: remove drug GO cellular component similarity network
Drugsim6: remove drug GO molecular function similarity network
Proteinsim1: remove protein sequence similarity network
Proteinsim2: remove protein Gene Ontology (GO) biological process similarity network
Proteinsim3: remove protein GO cellular component similarity network
Proteinsim4: remove protein GO molecular function similarity network

Note: We left each single network out separately and using the reminding 14 networks to build models. We repeated each experiment in 5 times and standard deviation was shown.

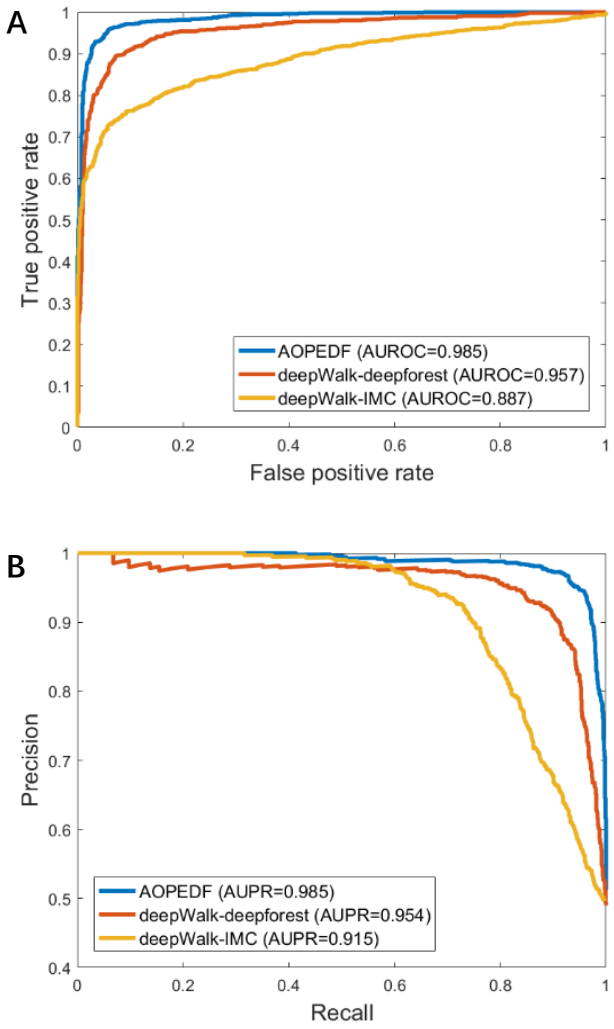
Supplementary Table S8. The robustness to the hyper-parameter settings. We use two random forests, two completely-random tree forests and two gradient boosting tree forests. We vary the number of trees in each forest, and observe the performance. From the results, we find that the prediction performance remains stable under different tree number settings.

Tree numbers	AUROC	AUPR
50	0.9824+0.0013	0.9832+0.0017
100	0.9831+0.0008	0.9840+0.0007
200	0.9837+0.0007	0.9843+0.0010
500	0.9851+0.0009	0.9852+0.0009
600	0.9840+0.0009	0.9849+0.0008

Supplementary Figures



Supplementary Figure S1. Performance of different methods on the experimentally validated drug-target network (Supplementary Tables 1 and 2). (**A**) Receiver operating characteristic (ROC) curves of prediction results obtained by applying AOPEDF and five previously reported methods in 5-fold cross-validation. (**B**) Precision-recall (PR) curves for AOPEDF and other methods in 5-fold cross-validation. AUROC: the area under ROC curve; AUPR: the area under PR curve.



Supplementary Figure S2. Performance of AOPEDF and deepWalk (Zong, et al., 2017) on the experimentally validated drug-target network. (**A**) Receiver operating characteristic (ROC) curves of prediction results obtained by applying AOPEDF and deepWalk in 5-fold cross-validation. (**B**) Precision-recall (PR) curves for AOPEDF and deepWalk in 5-fold cross-validation. AUROC: the area under ROC curve; AUPR: the area under PR curve. We performed both deepWalk+IMC (Inductive matrix completion) and deepWalk + deepforest.

Supplementary References

Amberger, J.S.*, et al.* OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic acids research* 2015;43(Database issue):D789-798.

Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research* 2004;32(Database issue):D267-270.

Breuer, K.*, et al.* InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic acids research* 2013;41(Database issue):D1228-1233.

Brown, A.S. and Patel, C.J. A standard database for drug repositioning. *Sci Data* 2017;4:170029.

Chatr-Aryamontri, A.*, et al.* The BioGRID interaction database: 2015 update. *Nucleic acids research* 2015;43(Database issue):D470-478.

Cheng, F.*, et al.* Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol Biol Evol* 2014;31(8):2156-2169.

Cheng, F.*, et al.* Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget* 2014;5(11):3697-3710.

Cheng, F.*, et al.* Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J Chem Inf Model* 2013;53(4):753-762.

Coordinators, N.R. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 2016;44(D1):D7-19.

Cowley, M.J.*, et al.* PINA v2.0: mining interactome modules. *Nucleic acids research* 2012;40(Database issue):D862-865.

Davis, A.P.*, et al.* The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic acids research* 2015;43(Database issue):D914-920.

Davis, A.P.*, et al.* The Comparative Toxicogenomics Database: update 2011. *Nucleic acids research* 2011;39(Database issue):D1067-1072.

Dinkel, H.*, et al.* Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic acids research* 2011;39(Database issue):D261-267.

Fazekas, D.*, et al.* SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC systems biology* 2013;7:7.
Gaulton, A.*, et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 2012;40(D1):D1100-D1107.
Gonen, M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 2012;28(18):2304-2310.
Hernandez-Boussard, T.*, et al.* The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic acids research* 2008;36(Database issue):D913-918.
Hornbeck, P.V.*, et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research* 2015;43(Database issue):D512-520.
Hu, J.*, et al.* PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics* 2014;30(1):141-142.
Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* 2014.
Kuhn, M.*, et al.* A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology* 2010;6:343.
Law, V.*, et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;42(Database issue):D1091-1097.
Licata, L.*, et al.* MINT, the molecular interaction database: 2012 update. *Nucleic acids research* 2012;40(Database issue):D857-861.
Liu, T.Q.*, et al.* BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic acids research* 2007;35:D198-D201.
Lu, C.T.*, et al.* DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic acids research* 2013;41(Database issue):D295-305.
Mendez, D.*, et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* 2018;47(D1):D930-D940.
Meyer, M.J.*, et al.* INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 2013;29(12):1577-1579.
Newman, R.H.*, et al.* Construction of human activity-based phosphorylation networks. *Molecular systems biology* 2013;9:655.
O'Boyle, N.M.*, et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* 2011;3:33.

Orchard, S.*, et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research* 2014;42(Database issue):D358-363.
Pawson, A.J.*, et al.* The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic acids research* 2014;42(D1):D1098-D1106.
Peri, S.*, et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic acids research* 2004;32(Database issue):D497-501.
Rolland, T.*, et al.* A proteome-scale map of the human interactome network. *Cell* 2014;159(5):1212-1226.
Rual, J.F.*, et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;437(7062):1173-1178.
Smith, T.F. and Waterman, M.S. Identification of common molecular subsequences. *Journal of molecular biology* 1981;147(1):195-197.
Tatonetti, N.P.*, et al.* Data-driven prediction of drug effects and interactions. *Science translational medicine* 2012;4(125):125ra131.
Ursu, O.*, et al.* DrugCentral 2018: an update. *Nucleic Acids Research* 2018;47(D1):D963-D970.
Ursu, O.*, et al.* DrugCentral: online drug compendium. *Nucleic Acids Res* 2017;45(D1):D932-D939.
van Laarhoven, T. and Marchiori, E. Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. *Plos One* 2013;8(6):e66952.
Wan, F.*, et al.* NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* 2018;35(1):104-111.
Wang, J.Z.*, et al.* A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;23(10):1274-1281.
Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* 2006;11(23-24):1046-1053.
Xia, Z.*, et al.* Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst Biol* 2010;4 Suppl 2:S6.
Yu, G.*, et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010;26(7):976-978.
Yu, W.*, et al.* A navigator for human genome epidemiology. *Nature genetics* 2008;40(2):124-125.

Zeng, X.*, et al.* Target identification among known drugs by deep learning from heterogeneous networks. *Chemical Science, 2020*.
Zhu, F.*, et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research* 2012;40(Database issue):D1128-1136.
Zong, N.*, et al.* Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics (Oxford, England)* 2017;33(15):2337-2344.