



## METHOD

# PIMD: An Integrative Approach for Drug Repositioning Using Multiple Characterization Fusion



Song He<sup>#</sup>, Yuqi Wen<sup>#</sup>, Xiaoxi Yang, Zhen Liu, Xinyu Song, Xin Huang, Xiaochen Bo<sup>\*</sup>

Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing 100850, China

Received 9 July 2018; revised 21 September 2018; accepted 10 October 2018

Available online 17 October 2020

Handled by Xin Gao

## KEYWORDS

Drug repositioning;  
Drug similarity network;  
Multiple characterization  
fusion;  
Network pharmacology;  
Drug discovery

**Abstract** The accumulation of various types of drug informatics data and computational approaches for **drug repositioning** can accelerate pharmaceutical research and development. However, the integration of multi-dimensional drug data for precision repositioning remains a pressing challenge. Here, we propose a systematic framework named PIMD to predict drug therapeutic properties by integrating multi-dimensional data for drug repositioning. In PIMD, drug similarity networks (DSNs) based on chemical, pharmacological, and clinical data are fused into an integrated DSN (iDSN) composed of many clusters. Rather than simple fusion, PIMD offers a systematic way to annotate clusters. Unexpected drugs within clusters and drug pairs with a high iDSN similarity score are therefore identified to predict novel therapeutic uses. PIMD provides new insights into the universality, individuality, and complementarity of different drug properties by evaluating the contribution of each property data. To test the performance of PIMD, we use chemical, pharmacological, and clinical properties to generate an iDSN. Analyses of the contributions of each drug property indicate that this iDSN was driven by all data types and performs better than other DSNs. Within the top 20 recommended drug pairs, 7 drugs have been reported to be repurposed. The source code for PIMD is available at <https://github.com/Sepstar/PIMD/>.

## Introduction

Despite the ever-increasing funding of pharmaceutical research and development (R&D), the number of new drugs approved has not increased significantly [1]. Traditional *de novo* drug development remains costly, risky, and time-consuming [2,3]. Drug repositioning, wherein an existing drug receives a new application, provides a new opportunity for pharmaceutical

\* Corresponding author.  
E-mail: [boxc@bmi.ac.cn](mailto:boxc@bmi.ac.cn) (Bo X).  
<sup>#</sup> Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

<https://doi.org/10.1016/j.gpb.2018.10.012>

1672-0229 © 2020 The Authors. Published by Elsevier B.V. and Science Press on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

R&D [4]. With the accumulation of drug informatics datasets, computational algorithms can be used for systematic identification of potential new indications for on-market drugs, thereby reducing the financial and time investment, as well as the risk, involved in pharmaceutical R&D [5–8].

Recent studies have shown that computational approaches based on drug similarity have the potential to reveal novel indications for on-market drugs. These approaches have been applied primarily in the following four ways. 1) The first group includes transcriptional response-based approaches. For instance, Iorio et al. [9] used drug-specific response profiles based on the Connectivity Map (CMap) database to find drugs with similar modes of action, whereas Xie et al. [10] used drug perturbation profiles from the Library of Integrated Network-based Cellular Signatures (LINCS) project to predict additional therapeutic properties of drugs. 2) The second group includes chemical structure-based approaches. For example, Keiser et al. [11] predicted potential drug targets based on combined drug–target structure. 3) The third group includes side effect- or other phenotype-based approaches. For example, Campillos et al. [12,13] constructed a side effect-driven drug similarity network (DSN) based on the assumption that drugs with similar side effects may share targets leading to the identification of novel drug targets. 4) The last group includes target property-based approaches. For example, Yildirim et al. [14] used known drug–target associations to assess ongoing trends and shifts in drug discovery and to quantify interrelationships between drug targets and disease-causing gene products. Although they shared the assumption that similar drugs tend to share therapeutic properties, each of these studies focused on a single drug feature in assessment of drug similarity, raising some doubts about the usefulness of these approaches. For example, Yildirim et al. [14] pointed out that most drugs with the same targets have different chemical structures, and Keiser et al. [15] demonstrated that a small change in drug structure could alter binding affinity dramatically. In addition, the transcriptional response to drug perturbation may differ across cell lines and drug dosages, thus introducing noise into drug repositioning strategies based on transcriptome data. Notwithstanding, our previous studies illustrated positive correlations between repositioning potential based on transcriptome data and that based on side effect profile or structure [10].

With the development of network pharmacology and systems biology, integrating multi-attribute data of drugs seems to be a feasible means of identifying new opportunities for drug repositioning [16]. One of the most common methods to integrate such data is to concatenate several measurements from various properties, such as side effects and chemical fragments, of each drug [17,18]. However, the already low signal-to-noise ratio in each data type could be diluted by concatenation [19]. To avoid this problem, many researchers have made some preliminary attempts to use DSNs based on different drug properties for data combination. For example, Napolitano et al. [20] constructed three DSNs, based on drug structures, distances between drug targets in protein–protein interaction (PPI) networks, and expression patterns of drug perturbations, separately. They then integrated these attribute datasets by averaging three drug similarity measurements to predict new therapeutic properties of drugs. Meanwhile, Wang et al. [21] proposed a new algorithm, called PreDR, which predicts as yet unidentified drug–disease associations by taking

the maximums of three drug similarity matrices derived from chemical structure, target protein sequence, and side effect profile similarities. Zhang et al. [22] proposed the Similarity-based LARge-margin learning of Multiple Sources (SLAMS) algorithm of drug similarity based on multiple sources of drug and disease property data. SLAMS outputs therapeutic scores for each drug–disease pair that correspond to multi-level drug properties and disease properties, and then averages the scores to predict the novel disease applications for drugs. Liu et al. [23] proposed the two-pass random walks with restart on a heterogeneous network (TP-NRWRH) to predict new indications for approved drugs. In the model, DSNs are integrated using the probability disjunction formula. Additionally, there are many articles on the drug–target predictions with data integration based on a linear combination of multiple attributes [24–42]. Although many studies take into account multiple drug sources, these integration strategies based on simple averaging or maximization are linear and cannot make full use of topology and non-linear information. Furthermore, most of the methods cannot evaluate the contribution and the relative importance of each property data. As for network-based analysis, to our knowledge, there is no integrated method offering a systematic way to annotate and evaluate the drug clusters. In recent years, non-linear multi-dimensional data fusion algorithms and tools have been widely applied in disease subtype identification, such as the Similarity Network Fusion (SNF) and Integrated Clustering of Multidimensional biomedical data (ICM) [19,43].

In this study, we propose a systematic and extensible paradigm of drug repositioning, namely prediction of drug therapeutic property by integrating multi-dimensional data (PIMD), and report the construction of an integrated DSN (iDSN) based on drug structure, side effect profile, and target protein sequence data. First, we integrated different types of drug information, including side effects, chemical structures, and molecular targets representing clinical, chemical, and pharmacological properties, separately, and constructed a DSN for each of these properties. Second, we used a non-linear fusion algorithm, namely SNF, to combine three DSNs into the iDSN iteratively [19]. Next, we evaluated the contributions of each dimension of data in the iDSN and the correlation among them.

Our study examined the types of data underlying the iDSN and how the iDSN performs relative to single-property networks. We used spectral clustering to divide the iDSN into clusters, and then conducted a systematic and comprehensive drug cluster analysis through five types of statistical analyses, including drug-based enrichment analysis, target-based enrichment analysis, drug property analysis, chemogenomic enrichment analysis (CGEA), and chemical ontology enrichment analysis. We hypothesize that if similar drugs have similar therapeutic properties, then the drugs that appear unexpectedly in a cluster based on their anatomical therapeutic chemical (ATC) label would represent repositioning candidates.

## Method

### Data source

We collected information on 7132 drugs with their corresponding target protein information from the DrugBank database, a

bioinformatics resource with detailed drug data and complete drug–target interaction data. We used version 5.0 of the DrugBank database to construct the DSN based on drug targets (DSN-T) [5]. Protein sequence data were extracted from the UniProt database, which provides high-quality, freely accessible protein sequence data [44].

We obtained 139,756 relationships between 1430 drugs and 5868 side effects from the Side Effect Resource (SIDER) database, which contains information about on-market drugs and their recorded adverse drug reactions or side effects, extracted from public documents and package inserts. We used version 4.1 of SIDER to construct the DSN based on drug side effects (DSN-S) [8].

PubChem Compound is a database containing more than 92 million unique structures of compounds. Similarly, We extracted chemical structures of drugs from PubChem Compound to construct the DSN based on drug chemical structure (DSN-C) [7]. We used the PubChem Compound Identifier (CID) as the only identifier of drugs to identify drugs shared across the databases. We used the identified drugs to construct three single DSNs.

## Drug similarity measurements

Drug similarity quantifies the degree of shared features between paired drugs. We restricted similarity scores to be between 0 (lowest) and 1 (highest). We defined the drug similarity measurements of the three properties examined separately.

### Drug similarity based on drug side effects

Side effects represent the clinical properties of a drug. We obtained side effect data from SIDER. We used drug side effect information with frequency data (as opposed to without) because such information was derived empirically and thus deemed more credible. Given there are risks of bias in observation and statistics, we filtered outsider effect terms if they occurred only once or with a frequency  $< 0.1\%$ . Finally, we characterized drug side effects according to the 2072-dimensional binary vector  $E(d)$ , known as side effect profile. Similarity based on side effects between two drugs  $d$  and  $d'$  was computed by the Tanimoto coefficient of their side effect profiles:

$$S_{d,d'}^{sideeffect} = \frac{E(d) \times E(d')}{|E(d)| + |E(d')| - E(d) \times E(d')} \quad (1)$$

where  $|E(d)|$  and  $|E(d')|$  are the number of side effect terms for drugs  $d$  and  $d'$ , respectively.  $E(d) \times E(d')$  represents the number of side effects shared by these two drugs.

### Drug similarity based on drug chemical structure

Drug chemical structure represents the chemical properties of a drug. We obtained chemical structure information from PubChem Compound and computed atom-pair descriptors of drugs using the R package “ChemmineR” [45]. The atom-pair descriptors used to quantify the chemical structure of

small molecule compounds encode all atom pairs in a drug. We computed similarity based on the chemical structure between drugs  $d$  and  $d'$  as the Tanimoto coefficient of the chemical atom-pair descriptors:

$$S_{d,d'}^{chem} = \frac{C(d) \times C(d')}{|C(d)| + |C(d')| - C(d) \times C(d')} \quad (2)$$

where  $|C(d)|$  and  $|C(d')|$  represent the number of atom pairs for drug  $d$  and drug  $d'$ , respectively.  $C(d) \times C(d')$  represents the number of the atom pairs shared by these two drugs.

### Drug similarity based on drug targets

Drug targets represent the pharmacological properties of a drug. We obtained drug–target interactions and target protein sequences from the DrugBank and Uniprot databases, respectively. Then, similarity based on drug targets between drugs  $d$  and  $d'$  was computed with a normalized Smith–Waterman score as follows:

$$S_{d,d'}^{target} = \frac{1}{|P(d)| \times |P(d')|} \sum_{i=1}^{|P(d)|} \sum_{j=1}^{|P(d')|} \frac{SW(P_i(d), P_j(d'))}{\sqrt{SW(P_i(d), P_i(d))} \sqrt{SW(P_j(d'), P_j(d'))}} \quad (3)$$

where  $P(d)$  represents a target protein set of drug  $d$ ,  $P_i(d)$  indicates the  $i$ th target of drug  $d$ , and  $|P(d)|$  represents the size of the target protein set of drug  $d$ .  $SW(P_i(d), P_j(d'))$  is the Smith–Waterman sequence alignment score of target proteins of drugs  $d$  and  $d'$  [46].

## SNF method

We used the “SNFtool” in R software to achieve SNF, a useful and popular computational method for data integration in the field of disease subtype identification [19]. It can deal with noise in different data types and make full use of common and complementary information across data types by integrating data in a non-linear way. We introduced SNF into multi-dimensional drug informatics data integration in this study for the first time, in the following three steps. 1) DSNs are built for each data type. 2) Multiple DSNs are integrated with SNF, and each of these DSNs is updated iteratively with information from other networks, making them more similar to each other than before. There are three main parameters in SNF: hyperparameter ( $\eta$ ), number of neighbors ( $K$ ), and number of iterations ( $T$ ). The integration is robust to these parameters as described previously [19]. Here we set  $\eta = 0.5$ ,  $K = 20$ , and  $T = 20$ , as recommended by Wang and colleagues [19]. 3) A final iDSN is obtained from SNF process convergence.

## Cluster validity index

We used two cluster validity indexes to determine the number of clusters.

### Dunn index

The Dunn index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn index has a value between zero and infinity and should be maximized. The Dunn index is calculated as follows:

$$DI = \min_{C_k \in C} \left( \min_{C_l \in C} \frac{\text{dist}(C_k, C_l)}{\max_{C_m \in C} \text{diam}(C_m)} \right) \quad (4)$$

where  $C$  is the collection of all clusters,  $\text{diam}(C_m)$  is the largest intra-cluster distance in Cluster  $C_m$ , whereas  $\text{dist}(C_k, C_l)$  is the distance between the nearest pair of samples in Cluster  $C_k$  and Cluster  $C_l$ .

### Silhouette index

The silhouette value is a measure of how well each object lies within its cluster. The silhouette value ranges from  $-1$  to  $1$ , and should be maximized. It is calculated as follows:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (5)$$

where  $a_i$  is the average distance between sample  $i$  and all other data points within the same cluster,  $b_i$  is the lowest average distance of sample  $i$  to all points in any other clusters.

### Evaluation measurements

Several evaluation measurements were used in the study, as introduced below.

#### Normalized mutual information

In probability theory and information theory, the mutual information (MI) of two random variables is a measure of their mutual dependence. In this study,  $X$  is the ATC label vector of all the 593 drugs in iDSN, and  $Y$  is the predicted label vector obtained by clustering these drugs. MI was calculated as follows:

$$MI(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \left( \frac{p(x, y)}{p(x)p(y)} \right) \quad (6)$$

where  $p(x, y)$  is the joint probability function of  $X$  and  $Y$ , whereas  $p(x)$  and  $p(y)$  are the marginal probability distribution functions of  $X$  and  $Y$ , respectively. Normalized MI (NMI) is calculated as follows:

$$NMI(X, Y) = \frac{MI(X, Y)}{\sqrt{H(X) \times H(Y)}} \quad (7)$$

where  $H(X)$  and  $H(Y)$  are the entropy of  $X$  and  $Y$ , respectively.

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \quad (8)$$

where  $p(x)$  is the marginal probability distribution function of  $X$ .

#### ATC overlap rate

In drug set  $D$ , the number of drug pairs is denoted as  $Sum$ . For each drug pair  $(d_x, d_y)$  in the set, we calculated the intersection of ATC first-level codes of drugs  $d_x$  and  $d_y$ . If the intersection

was not empty,  $AI(d_x, d_y) = 1$ , otherwise denoted  $AI(d_x, d_y) = 0$ . The ATC overlap rate (AOR) in drug set  $D$  is computed as follows:

$$AOR(D) = \sum_{(d_x, d_y) \in D} \frac{AI(d_x, d_y)}{Sum} \quad (9)$$

#### Superclass overlap rate

In drug set  $D$ , the number of drug pairs is denoted as  $Sum$ . For each drug pair  $(d_x, d_y)$  in the set, if the superclass label of drug  $d_x$  is the same as that of drug  $d_y$ ,  $SI(d_x, d_y) = 1$ , otherwise  $SI(d_x, d_y) = 0$ . The superclass overlap rate (SOR) in drug set  $D$  is computed as follows:

$$SOR(D) = \sum_{(d_x, d_y) \in D} \frac{SI(d_x, d_y)}{Sum} \quad (10)$$

#### Connectivity

The connectivity indicates the degree of connectedness of the clusters. Denote  $N$  as the number of observations and denote  $C$  as the collection of all clusters.  $L$  represents the number of nearest neighbors. Define  $mn_{i(j)}$  as the  $j^{\text{th}}$  nearest neighbor of observation  $i$ . Let  $x_{i, mn_{i(j)}}$  be zero if  $i$  and  $j$  are in the same cluster, and  $1/j$  otherwise. The connectivity is defined as:

$$Conn(C) = \sum_{i=1}^N \sum_{j=1}^L x_{i, mn_{i(j)}} \quad (11)$$

The connectivity has a value between zero and infinity and should be minimized.

#### Rogers–Tanimoto index

The Rogers–Tanimoto similarity rely on a  $2 \times 2$  contingency table, consisting of the following four cells:  $n_{11}$ ,  $n_{10}$ ,  $n_{01}$ , and  $n_{00}$ .  $n_{11}$  is the number of observation pairs, where the two observations belong to the same cluster according to both partition  $P_1$  and  $P_2$ .  $n_{10}$  is the number of observation pairs, where the two observations belong to the same cluster according to partition  $P_1$  but not to  $P_2$ .  $n_{01}$  is the number of observation pairs, where the two observations belong to the same cluster according to partition  $P_2$  but not to  $P_1$ .  $n_{00}$  is the number of observation pairs, where the two observations do not belong to the same cluster according to both partition  $P_1$  and  $P_2$ . The Rogers–Tanimoto similarity is defined as:

$$RT = \frac{n_{11} + n_{00}}{n_{11} + n_{00} + 2(n_{10} + n_{01})} \quad (12)$$

### Other integrative methods for comparison

We compared the network fusion performance of PIMD with three previous integrative methods: 1) the maximum method [21], 2) the weighted average method [20,22,24–32], and 3) the probability disjunction [23]. For the maximum method, we took the maximums of multiple drug similarity matrices. For the weighted average method, we averaged multiple drug similarity matrices by traversing weight. The weight of each drug similarity network is from 0 to 1 with step 0.1. For the probability disjunction, the formula is:



$$S_{d,d'} = 1 - (1 - S_{d,d'}^{\text{sideeffect}})(1 - S_{d,d'}^{\text{chem}})(1 - S_{d,d'}^{\text{target}}) \quad (13)$$

where  $S_{d,d'}$  is the integrative similarity measurement between drug  $d$  and  $d'$ .

### Data type contribution

For each edge in the iDSN, we used similarity scores from each single network to describe which data type was the primary contributor. First, we ranked three similarity scores of the edge in each single network as  $S_i \geq S_j \geq S_k$ , where  $i, j$ , and  $k$  refer to the three types of data. If  $S_i$  was  $\geq 10\%$  higher than  $S_j$ , the edge was attributed to the  $i$  data type. If  $S_i$  was  $< 10\%$  higher than  $S_j$  but  $S_j$  was  $\geq 10\%$  higher than  $S_k$ , the edge was attributed to both the  $i$  and  $j$  data types. If  $S_i$  is  $< 10\%$  higher than  $S_j$ , but  $S_j$  is  $< 10\%$  higher than  $S_k$ , the edge was attributed to all three data types.

### Statistical analysis

We performed five types of enrichment analyses to annotate the drug clusters for drug precision repositioning. For drug-based and target-based enrichment analyses, we calculated enrichment score (ES) as follows:

$$ES = \frac{kN}{nm} \quad (14)$$

where  $k$  is the number of drugs with a particular label (e.g., ATC code) in the cluster of interest,  $m$  is the number of drugs with the label in the overall dataset,  $n$  is the total number of drugs in the cluster of interest, and  $N$  is the total number of drugs in the overall dataset. Then we used the hypergeometric distribution to calculate the  $P$  value as follows:

$$P(X \geq k) = \sum_{x=k}^{\infty} \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}} \quad (15)$$

### Drug-based enrichment analysis

For each cluster, we performed drug class and absorption, distribution, metabolism, excretion, and toxicity (ADMET) property enrichment analyses based on the DrugBank database. We computed ESs and  $P$  values based on drug ATC code, superclass label, and ADMET properties.

### Target-based enrichment analysis

For target proteins of each cluster, we performed target class, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, and Gene Ontology (GO) enrichment analyses based on the International Union of Basic and Clinical Pharmacology/British Pharmacological Society (IUPHAR/BPS), KEGG, and GO databases, respectively [47–49]. KEGG and GO analyses were conducted with the ClusterProfiler tool in R software [50]. KEGG and GO enrichment results were selected by a  $P$  value threshold of 0.05.

### Drug property analysis

Physicochemical features were extracted from a drug property list in RepurposeDB [51]. Chemical descriptors were computed based on information from the Pybel, JOELib2, and Chemminer chemoinformatics libraries [45,52,53]. We used  $t$ -tests

to evaluate the significance of deviation of mean values for each cluster from that of all drugs in DrugBank. In each cluster, for property  $i$  in property set  $I$ , the deviation from the mean value of all drugs in DrugBank is calculated as follows:

$$D(i) = \frac{C(i) - A(i)}{|A(i)|}, i \in I \quad (16)$$

where  $A(i)$  is the mean property value for all drugs in DrugBank, and  $C(i)$  is the mean property value for drugs in the cluster.

### Chemogenomic enrichment analysis

We use the CGEA online tool to analyze drugs in each cluster (<http://server.dudleylab.org/index>). CGEA maps the drug list to various annotation resources, including drug-induced transcriptional modules, enzymes, and fragments. These potentially relevant features identified provide a rich chemogenomic context for drugs of interest.

### Chemical ontology enrichment analysis

BiNChE is a web tool for chemical enrichment analysis based on the Chemical Entities of Biological Interest (ChEBI) Ontology [54]. We performed enrichment analysis by mapping the drugs in each cluster to both ‘structure’ and ‘role’ subsets of ChEBI Ontology.

## Results

### An overview of PIMD framework

PIMD is a systematic framework that predicts new therapeutic properties of known drugs, calculates the contributions of various types of data, and annotates various aspects of drug grouping, by integrating multiple drug properties. In this study, we hypothesize that if drug  $d_A$  and drug  $d_B$  are similar with respect to a particular parameter, then the therapeutic property  $t$  of drug  $d_A$  may also be shared by drug  $d_B$ . The functions of a drug can be characterized by multiple drug informatics. Therefore, in PIMD, we investigated three representative drug properties: 1) chemical properties, based on chemical structure data derived from the PubChem Compound database [7]; 2) pharmacological properties, based on protein target sequence data derived from the DrugBank and UniProt databases [5,44]; and 3) clinical and phenotypic properties, based on side effect data derived from the SIDER database [8].

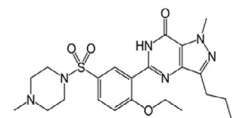
The PIMD consists of three parts (Figure 1). In the first part, we constructed three DSNs based on side effects, chemical structure, and drug targets, denoted as DSN-S, DSN-C, and DSN-T, respectively. Then, we applied the SNF method to fuse the DSN-S, DSN-C, and DSN-T into the iDSN in a manner that enables the relative contribution of each data type to be calculated. In the second part, we assessed drug repositioning using two approaches: finding drug pairs with a high similarity score, and identifying unexpected drugs in each cluster. In the iDSN, drug pairs are ranked according to their similarity scores, with highly ranked drug pairs being the most likely to achieve drug repositioning, wherein one drug in a pair may be repurposed for the therapeutic properties of the other drug. We used spectral clustering to distinguish clusters within the iDSN [55], wherein drugs within the same cluster have

## Similarity network fusion

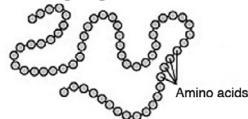
## • Side effects



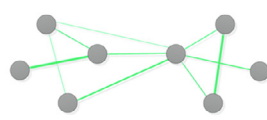
## • Chemical structure



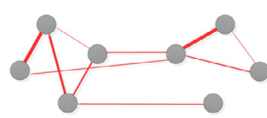
## • Drug targets



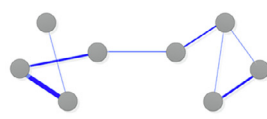
## • DSN-S



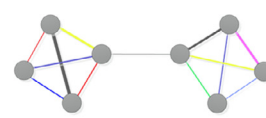
## • DSN-C



## • DSN-T



## • iDSN

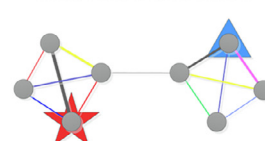


## Drug repositioning

## • Drug pairs with high similarity

Final ranking		
Rank	Drug pair	Similarity
1	●●	0.91
2	●●	0.84
3	●●	0.77
4	●●	0.75
5	●●	0.74
6	●●	0.73
...	...	...

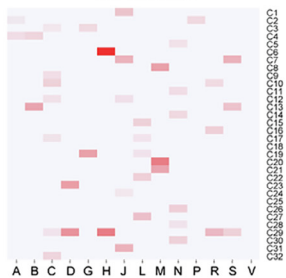
## • Unexpected drugs in clusters



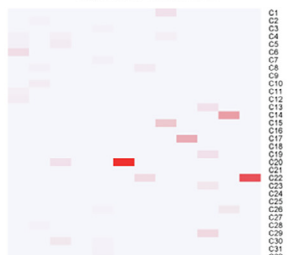
## Enrichment analysis

## • Drug-based enrichment analysis

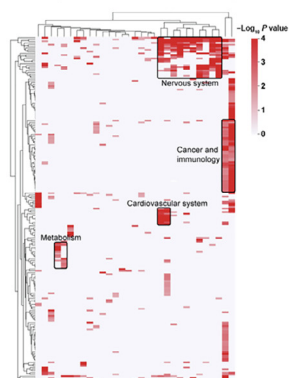
## ATC enrichment



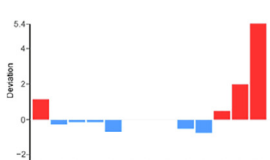
## Superclass enrichment



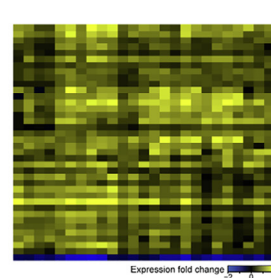
## • Target-based enrichment analysis



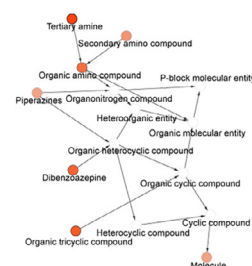
## • Drug property statistical analysis

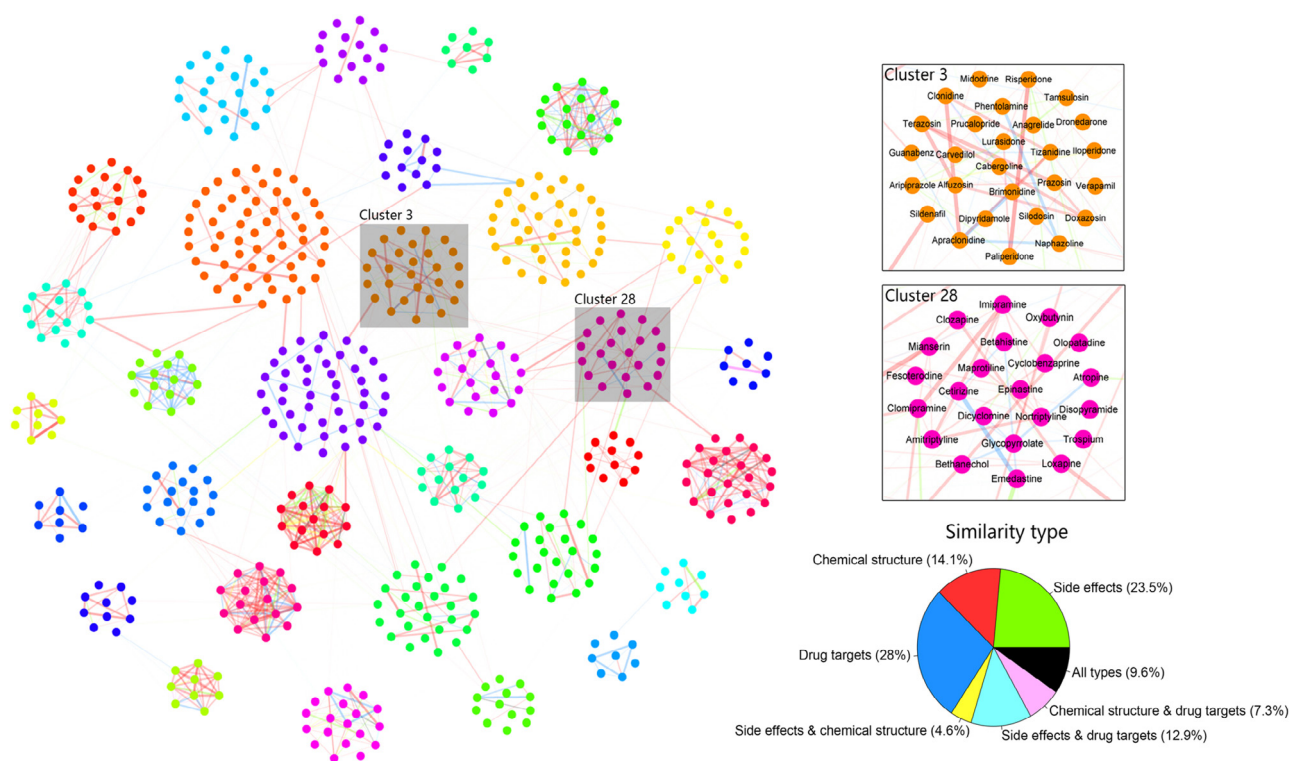


## • Chemogenomic enrichment analysis



## • Chemical ontology enrichment analysis





**Figure 2 The iDSN and contributions of various data types**

The 593 drugs in the iDSN are grouped into 32 clusters. Each node represents a drug, and node color signifies cluster assignment of the drug. Edges are weighted according to similarity score. Edges are color coded according to the major contributing data types, as defined in the pie chart on the right. Three types of data, including chemical structure, drug targets, and side effects, are used alone or in various combinations to evaluate their respective contributions. Clusters 3 and 28 are zoomed in as examples for better illustration.

similar therapeutic properties, as reflected by first-level ATC codes. Within each cluster, drugs with unexpected ATC labels are flagged as having the potential for a repositioning of therapeutic properties. In the third part, aimed at drug precision repositioning, we carried out the following series of enrichment analyses to label and annotate each cluster: 1) drug-based enrichment analysis; 2) target-based enrichment analysis; 3) drug property analysis; 4) CGEA; and 5) chemical ontology enrichment analysis. These analyses can provide guidance to researchers conducting drug repositioning studies from various perspectives. They may also help to elucidate differences in the mode of action of different drug clusters and reveal potential associations between drugs within a cluster.

### Global analysis of iDSN

We used PubChem CIDs as the sole drug identifiers and extracted chemical structures, target sequences, and side effect sources for 593 drugs from the PubChem Compound, DrugBank, UniProt, and SIDER databases. After constructing the DSN-C, DSN-T, and DSN-S, and combining these three networks into the iDSN with the SNF method, we applied spectral clustering to divide the iDSN into 32 subnetworks (labeled Cluster 1–32) based on two validity indexes (Figure 2, Figure S1) [55,56].

In the iDSN, each node represents a drug, and the edges connecting the nodes are thickness-weighted according to sim-

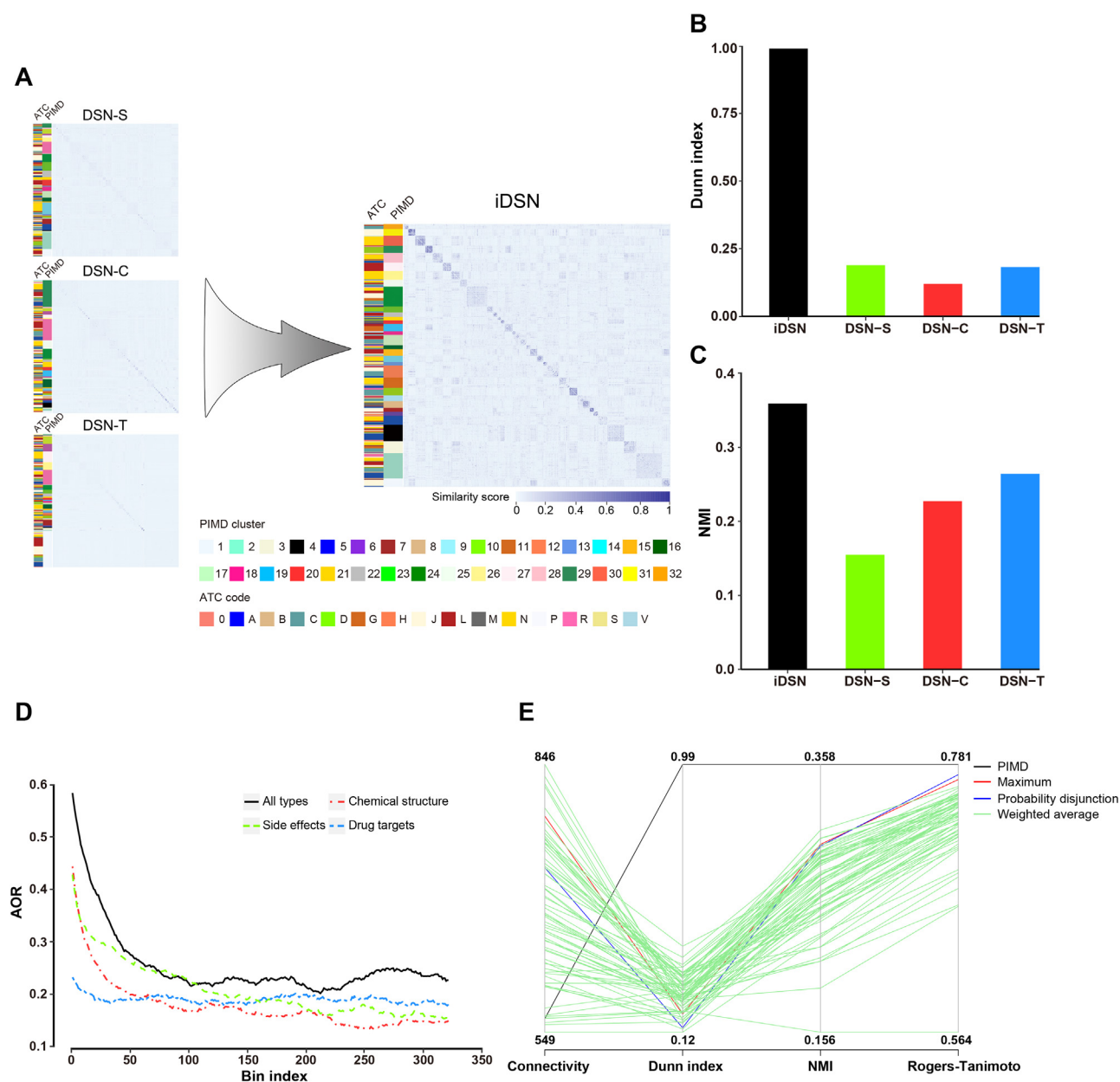
**Figure 1 Construction, drug repositioning, and annotation of iDSN using PIMD**

In the first part, we constructed three single-property DSNs. DSNs based on side effects, chemical structure, and drug targets were denoted as DSN-S, DSN-C, and DSN-T, respectively. Then we fused them into the iDSN. In the second part, possible drug repositioning applications were obtained based on high similarity or the occurrence of unexpected drugs within iDSN clusters. Edge color indicates the data type that is the main contributor to the similarity between the drugs. Pentagram and triangle represent unexpected drugs within iDSN clusters. In the third part, 593 drugs were included in the iDSN and grouped into 32 clusters (C1–C32) according to Dunn index and Silhouette values. We then performed a systematic analysis to annotate iDSN clusters from various perspectives to guide drug repositioning. PIMD, prediction of drug therapeutic property by integrating multi-dimensional data; DSN, drug similarity network; iDSN, integrated DSN; ATC, anatomical therapeutic chemical; A, alimentary tract and metabolism; B, blood and blood-forming organs; C, cardiovascular system; D, dermatological; G, genito-urinary system and sex hormones; H, systemic hormonal preparations, excluding sex hormones and insulins; J, anti-infectives for systemic use; L, antineoplastic and immunomodulating agents; M, musculoskeletal system; N, nervous system; P, anti-parasitic products, insecticides, and repellents; R, respiratory system; S, sensory organs; V, various.

ilarity of the connected drugs. Node color represents the cluster which the drug belongs to. Note that high-similarity connections are found predominantly within clusters. Edge color indicates the data type that is the main contributor to the similarity between the drugs, corresponding to the color scheme in the pie chart. Note that the iDSN model as a whole is supported by all analyzed types of data; in particular, drug pair similarities in the iDSN are supported by two or more types

of data, and among the three single drug properties, the highest relative contribution comes from the drug target-based data.

We divided the iDSN edges into two categories: within-cluster edges and between-cluster edges. Data contribution analysis for edges within the cluster (Figure S2) shows that side effect-based data, chemical structure-based data, and drug target-based data account for 11.0%, 22.2%, and 43.9%,



**Figure 3** Performance of the iDSN

**A.** Heatmaps of DSNs. DSN-S, DSN-C, and DSN-T were integrated into the iDSN. The two sidebars to the left of the networks correspond to the original ATC code (left) and PMID cluster label (right) separately. **B.** Dunn index for DSN-S, DSN-C, DSN-T, and iDSN. **C.** NMI for DSN-S, DSN-C, DSN-T, and iDSN. **D.** AOR for DSN-S, DSN-C, DSN-T, and iDSN. A bin composed of 3000 drug pairs was slid from the top to the bottom of the drug pair list with a step size of 100. AOR was calculated for each bin and that for the first 320 bins is plotted here. **E.** PMID outperforms other integrative methods for DSN clustering. For the weighted average method, multiple drug similarity matrices were averaged by traversing weight. The weight of each DSN ranges from 0 to 1 with the step of 0.1. NMI, normalized mutual information; AOR, ATC overlap rate.



respectively. Although the contributions of drug target-based data and side effect-based data are greater than that of chemical structure-based data for all edges in the iDSN (Figure 2), side effect-based data contribute the least to the edges inside the cluster and excessive contribution is from drug target-based data (Figure S2). These results indicate that drug target-based data play a more important role in drug clustering. Data contribution details for each cluster are reported in Table S1. For instance, the contribution of chemical structure-based data in Cluster 6 accounts for 94.4%. We found that Cluster 6 drugs are all peptide drugs, which have chemical structures significantly different from those of other drugs. Cluster 6 drugs also exhibit a significant deviation for physicochemical features (Figure S3A and B). The deviation is consistent with the nature of these peptide drugs.

Comparing the performances of the iDSN and each single-property network, we found that the iDSN has a clearer cluster structure than any of the three single-property DSNs (Figure 3A). To verify that, we calculated Dunn index, a metric for evaluating the quality of clustering results. The Dunn index for the iDSN is higher than those for the single-property DSNs (Figure 3B).

Furthermore, compared to the single-property DSNs, the iDSN has a larger overlap between the drug cluster and drug ATC labels. NMI, which reflects consistency across the original ATC and cluster labels, was examined, and a higher NMI for the iDSN was obtained than those for the DSN-C, DSN-T, or DSN-S (Figure 3C). In addition, we compared the AOR of each single-property DSN and the iDSN by first ranking similarity scores of all drug pairs in the iDSN or single-property DSNs and assigning overlapping consecutive drug pair bins with 3000 pairs per bin, such that bin 1 contains the top 3000 most similar drug–drug pairs, bin 2 contains pairs ranked 100th to 3100th in similarity, and so on (first 320 bins are shown in Figure 3D). Subsequent calculation of AORs for each bin shows clearly that the AORs for the iDSN are higher than those of the single-property DSNs (Figure 3D). These results suggest that PIMD makes full use of common and complementary information about drug properties and that the iDSN performs better than any single-property DSNs.

The network fusion of PIMD goes beyond a simple integration representing a maximum or an average of drug similarity measurements. It can capture potential links between drugs. On the one hand, if the similarity score in a single-property DSN is high, but the similarity scores in other single-property DSNs are low, PIMD does not dilute the original information. On the other hand, if the similarity scores for a drug pair in each of the three single-property DSNs are unremarkable, PIMD can still capture potential similarities. For example, the first-level ATC codes for fenoprofen and sulfasalazine are M and A, respectively. Fenoprofen (CID: 000003342) is used for symptomatic relief of rheumatoid arthritis, osteoarthritis, and mild to moderate pain, whereas sulfasalazine (CID: 005359476) is used to treat inflammatory bowel disease. The similarity scores for fenoprofen and sulfasalazine based on the three individual properties are relatively low (rank of 16,951 in DSN-C, 13,466 in DSN-T, and 146,145 in DSN-S), while the iDSN similarity score ranked much higher (rank of 2729) and both drugs were placed in Cluster 8. Although crossover in an application would not have been predicted by any of the single-property DSNs, it

has been reported that sulfasalazine can be used to treat rheumatoid arthritis [57,58].

### PIMD shows better performance than previous integrated methods

To better evaluate the network fusion performance of PIMD, we compared our results with three previous integrative methods: 1) the maximum method [21], 2) the weighted average method [20,22,24–32], and 3) the probability disjunction [23]. Here, we calculated two internal indices (connectivity and Dunn index) and two external comparison indices (NMI and Rogers–Tanimoto index). The internal indices are used to measure the goodness of a clustering structure without external information [59]. The external indices are a measure of agreement between two partitions where the first partition is the *a priori* known clustering structure, and the second partition results from the clustering procedure [60]. The connectivity indicates the degree of connectedness of the clusters. The Rogers–Tanimoto index reflects similarity between the original ATC and cluster labels. Among the four indices, only connectivity should be minimized. We found that PIMD performs the best compared with other methods (Figure 3E). The superior network fusion performance of PIMD results from the application of network-based approach, which is non-linear and utilizes topology information of the network.

### Drug repositioning from two aspects

There are two approaches for drug repositioning: the first is finding drug pairs with a high similarity score; the second is finding unexpected drugs in each cluster. Drug pairs with high similarity in the iDSN could provide us with clues for drug repositioning. The AORs of the top 10 (Table 1) and the top 100 (top 1000 drug pairs are listed in Table S2) drug pairs reached 80% and 94%, respectively. Drug pairs with high similarity scores but different ATC codes may have the potential for repositioning. For example, triptorelin (CID: 025074470) and nafarelin acetate (CID: 025077649) are both gonadotropin-releasing hormone receptor agonists despite having different first-level ATC codes. Interestingly, this drug pair has a high iDSN similarity score and is ranked second in the drug association list. Within the top 20 drug pairs, 7 drugs have been repurposed successfully according to RepurposeDB [51] and Repurposed Drug Database (<http://drugrepurposing-portal.com/>), which record all repurposed drugs thus far. We further checked whether the repositioning in the databases is relevant to our prediction and found that 6 out of these 7 drugs are repositioned for the same purpose as we predicted.

Unexpected drugs within an iDSN drug cluster can also signify the potential for repositioning based on distinct therapeutic properties. That is, a drug with an unusual ATC code within a certain cluster may be repositioned for alternative therapeutic properties. The unexpected drugs in each cluster are indicated in Table S3, and some cases are discussed in the section of drug repositioning case using PIMD.

We compared the two drug repositioning approaches together. In the top 100 drug pairs with a high similarity score, there are 6 drug pairs with totally different first-level ATC codes, among which 3 pairs can be discovered using the second

**Table 1** Top 10 drug pairs in the iDSN

Drug 1		Drug 2	
Name	ATC code	Name	ATC code
Atracurium besylate	M	Cisatracurium besylate	M
Triptorelin	L	Nafarelin acetate	H
Theophylline	R	Aminophylline	R
Gonadorelin	V, H	Nafarelin acetate	H
Gonadorelin	V, H	Triptorelin	L
Risedronate	M	Ibandronic acid	M
Citalopram	N	Escitalopram	N
Desonide	S, D	Desoximetasone	D
Lovastatin	C	Simvastatin	C, A
Ofloxacin	S, J	Lecofloxacin	S, J, A

*Note:* Similarity scores between drug pairs among the 593 drugs in the iDSN were calculated using SNF and the 10 drug pairs with highest similarity scores were selected. iDSN, integrated drug similarity network; SNF, Similarity Network Fusion; ATC, anatomical therapeutic chemical; A, alimentary tract and metabolism; C, cardiovascular system; D, dermatological; H, systemic hormonal preparations, excluding sex hormones and insulins; J, anti-infectives for systemic use; L, antineoplastic and immunomodulating agents; M, musculoskeletal system; N, nervous system; R, respiratory system; S, sensory organs; V, various.

approach. In the top 1000 drug pairs, there are 234 drug pairs with totally different first-level ATC codes, among which 75 pairs can be discovered using the second approach. These data indicate a certain overlap between the results predicted using these two approaches. The unexpected drugs in each cluster can achieve drug repositioning between different ATC codes. Compared to the second approach, top-ranking drug pairs with the same ATC codes are more inclined to achieve drug repositioning.

To explore the global repositioning association among ATC codes, we analyzed the top 5% of drug pairs in the iDSN. After dividing these drugs into 14 groups according to their ATC codes, we averaged the similarity scores between the different drug groups and used these averages as indices of the repositioning potential within each ATC code. As shown in Figure 4 (where edge thickness represents repositioning potential between two ATCs), there is a particularly high repositioning potential between dermatological drugs and respiratory system drugs. Among 78 dermatological drugs (D) and respiratory system drugs (R), 23 drugs have been repurposed successfully between D and R (Table S4).

### Drug cluster annotation in the iDSN

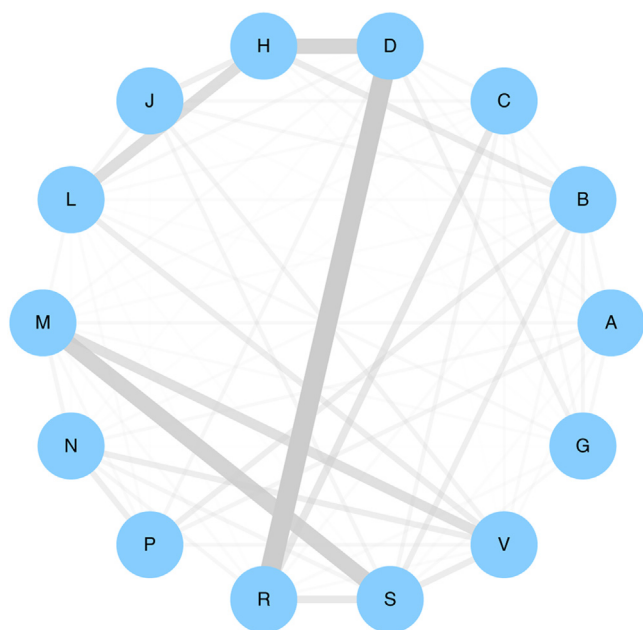
To label and annotate clusters for drug precision repositioning, we performed iDSN cluster analysis consisting of five statistical analysis methods from drug and target perspectives, thus providing multiple views to verify and select drug clusters for researchers in different fields. These analyses reveal potential links between drugs in the same clusters and differences in the modes of action between drugs in different clusters.

### ATC, superclass, and ADMET property enrichment analyses of drugs

We set to explore to what extent drugs within a cluster share a common ATC code. For each cluster, we computed an ATC code enrichment score and an accompanying *P* value. We found that 30 out of 32 clusters were significantly enriched for at least one ATC code ( $P < 0.05$ ; Figure 5A). Furthermore, 25 out of 32 clusters were significantly enriched for at least one superclass label ( $P < 0.05$ ; Figure 5B). The superclass label of the drug is extracted from the DrugBank database and focuses more on the chemical attributes of drugs. Combining the two enrichment results may lead to new discoveries. For example, Cluster 22 was found to be enriched in the lignan/norlignan superclass as well as in the antineoplastic and immunomodulating agent ATC code. Recent studies have shown that lignans/norlignans play an important role in anti-cancer therapies [61]. In addition, we extracted 18 ADMET property terms from the DrugBank database and examined whether drugs within the same cluster tend to have the same ADMET properties (Figure S3C). For example, 6 clusters are enriched in the nervous system ATC code, while four of them (Cluster 11, Cluster 26, Cluster 28, and Cluster 30) are also enriched in the blood brain barrier (+). This observation is in line with the knowledge that nervous system drugs usually need to penetrate the blood brain barrier.

### KEGG and GO enrichment analyses of targets

To explore whether drugs in the same cluster tend to target similar proteins and whether particular drug classes are associated with particular target classes, we performed KEGG pathway and GO enrichment analyses for targets of each drug



**Figure 4 Global repositioning association among ATC codes**

First, we divided drugs found in the top 5% of drug pairs in the iDSN into 14 groups according to their ATC codes. Then we averaged the similarity scores between the different drug groups and used these averages as indices of the repositioning potential within each ATC code. The 14 nodes represent ATC first levels. Edge thickness represents the repositioning potential of drugs between the two ATC codes connected.

cluster [48–50]. The most enriched pathways, biological processes, cellular components, and molecular functions (Figure 5C and D, Figure S3D and E; result matrices are listed in Table S5) differ substantially among the different clusters, highlighting differences in mode of action of drugs. Nonetheless, some drug clusters were found to contain common pathways and biological processes. We applied key biological themes for major clusters. For example, Cluster 27, which is enriched for antineoplastic and immunomodulating agents, was most enriched for the KEGG pathways of Rap1 signaling pathway, Ras signaling pathway, and central carbon metabolism in cancer. The most enriched GO biological process terms for Cluster 27 include protein autophosphorylation, positive regulation of MAPK cascade, and phosphatidylinositol-mediated signaling. These pathways and biological processes are indeed closely related to the mechanisms of antineoplastic drugs.

Another example is Cluster 16. Drugs in Cluster 16 are enriched for the ‘respiratory system’ ATC code. Chemical ontology enrichment analysis shows that drugs in this cluster are enriched for methylxanthine. The methylxanthines in Cluster 16, such as caffeine and theophylline, are used in therapy for respiratory diseases. However, the KEGG pathway enrichment analysis shows that drug targets in this cluster are also enriched for some cancer and immunology related pathways. These data suggest the association between methylxanthine and cancer, which is supported by recent studies [62,63].

Furthermore, we conducted target enrichment analysis for each cluster based on class information. The class information is collected from the IUPHAR/BPS database [47], which

indicates the type of drug targets, such as G-protein-coupled receptors (GPCR), catalytic receptor, and enzyme. We found that 24 of the 32 iDSN clusters were significantly enriched ( $P < 0.05$ ; Figure S3F). For example, Cluster 32 is enriched in the voltage-gated ion channel (VGIC) target class as well as in the ‘cardiovascular system’ ATC code. This observation is in line with the fact that most of Cluster 32 drugs are calcium channel blockers and used as antihypertensive drugs, whose targets mainly include calcium voltage-gated channel alpha1 (CACNA1) subunits.

#### *Physicochemical feature and chemical descriptor analyses of drug properties*

Given pharmacological profiling of small molecules may also affect drug repositioning, we analyzed the characteristic physicochemical features and chemical descriptors for each cluster. In total, we extracted 14 physicochemical features from the RepurposeDB drug property list [51] and 62 chemical (*i.e.*, atomic, compositional, and geometric) descriptors. These properties were quantitated using the Pybel, JOELib2, and Chemminer cheminformatics tools [45,52,53]. For each property, we calculated mean values of these drug properties for all drugs in DrugBank and in each cluster. We found that the mean value of drug properties for drugs in each cluster was deviated from that for all drugs in DrugBank (Table S6). The degrees of statistically significant deviation ( $P < 0.05$ ; see Method) are shown in Figure S3A and G. Considering the bias resulting from the incompleteness of drug set, we also compared the mean value of drugs in each cluster with those of all the 593 drugs in the iDSN (Figure S3B and H).

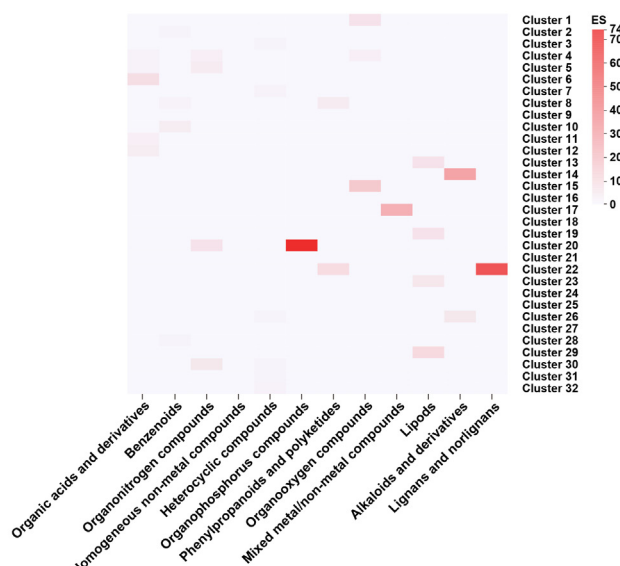
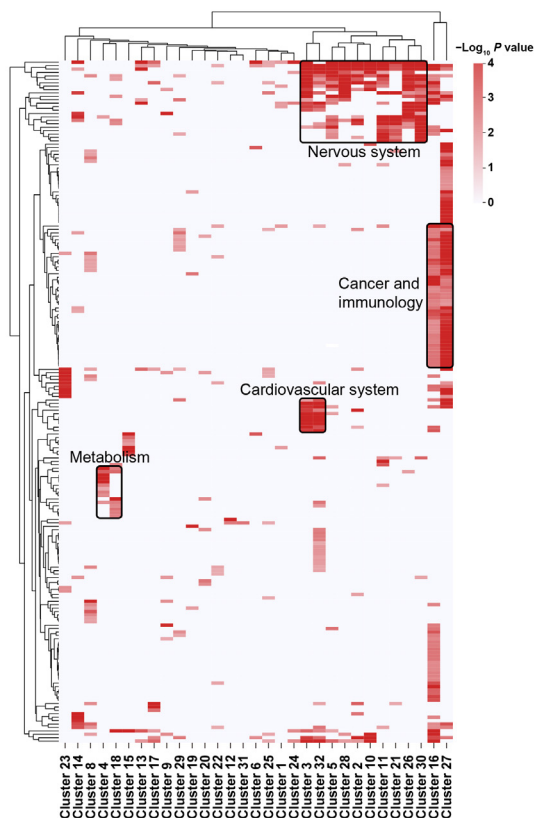
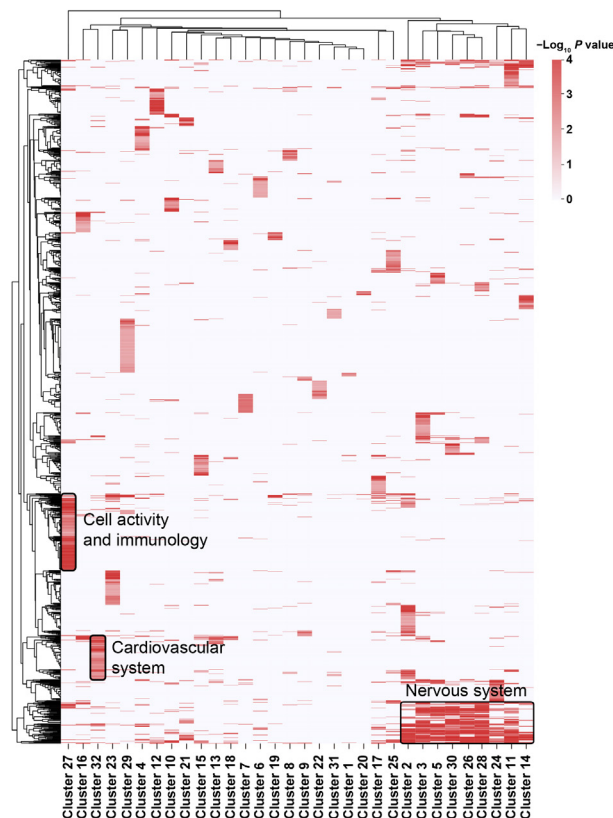
#### *Chemogenomic enrichment analysis*

CGEA, similar to gene set enrichment analysis, is a method that compares drugs with a range of biological and chemical annotations [64] (<http://server.dudleylab.org/index>). By identifying chemogenomic characteristics shared by sets of drugs such as enzymes, transporters, and structural fragments, we can obtain abundant chemogenomic context for a biological state of the drug set. We used CGEA to analyze drugs in each cluster for biological and chemical annotations. The results are listed in Table S7. The example of CGEA is described in detail in the section of ‘Case study: drug repositioning using 4 clusters of iDSN’.

#### *Enrichment analysis of chemical ontology*

Chemical ontology enrichment analysis is based on the ChEBI Ontology, which is a dictionary of chemical compounds with biological roles [65]. Mapping drugs to the ChEBI database can improve our understanding of biochemical nature of drugs. We performed a chemical ontology enrichment analysis of the drugs in each cluster using the BiChE tool [54]. The resultant ChEBI-based enriched structure terms and role terms for each cluster are provided in Table S8. The example of chemical ontology enrichment analysis is described in detail in the section of ‘Case study: drug repositioning using 4 clusters of iDSN’.

These analyses validate the rationality of the cluster division. Furthermore, by performing the 5 types of analyses described above, we have a comprehensive understanding of various properties of drugs in the cluster to conduct drug repositioning better. For example, Cluster 15 and Cluster 27 are

**A ATC enrichment****B Superclass enrichment****C KEGG pathway enrichment****D GO biological process enrichment****Figure 5 Drug-based and target-based enrichment analyses**

**A.** Enrichment landscape of ATC codes. **B.** Enrichment landscape of superclass codes from the DrugBank database. **C.** Enrichment landscape of KEGG pathways. **D.** Enrichment landscape of GO biological processes. Deeper color signifies greater ES in panels A and B, and signifies greater significance in panels C and D, respectively. Common biological themes shared by multiple clusters are boxed with names provided in the plot. ES, enrichment score; GO, Gene Ontology. Enrichment matrices of ATC codes and superclass codes were obtained using Fisher's exact test based on hypergeometric distribution. KEGG and GO enrichment results were obtained using the ClusterProfiler tool in R software.

both enriched for antineoplastic and immunomodulating agents, but their enriched KEGG pathways and GO terms

(Figure 5C and D, Figure S3D and E) differed from each other. Chemical ontology enrichment analysis shows that



Cluster 15 drugs have a role term of antimetabolite, while Cluster 27 drugs have a role term of protein kinase inhibitor. Most of Cluster 15 drugs are purine or pyrimidine analogs, whereas most of Cluster 27 drugs are tyrosine kinase inhibitors. These analyses highlight the differences in mode of action of the drugs. In Cluster 15, ribavirin (ATC code for ‘anti-infectives for systemic use’), which is a guanosine analog used for anti-virus, shows its therapeutic potential for cancers [66], and its mode of action is similar to other Cluster 15 drugs.

Additionally, these analyses provide us a chance to look at the same drug cluster in combination of different perspectives, which may bring us some new discoveries. For example, Cluster 5 drugs are enriched for the ‘nervous system’ ATC code. Chemical ontology enrichment analysis shows that Cluster 5 drugs have a role term of serotonergic drug, a type of nervous system-related drugs. KEGG pathway enrichment and GO enrichment analyses also reveal terms related to the nervous system. However, there are also terms related to cardiovascular system according to cellular components (CC) and molecular function (MF) of GO enrichment analysis results (Figure S3D and E). In ADMET property enrichment analysis, Cluster 5 drugs are enriched in human ether-a-go-go-related gene (hERG) inhibition (predictor I, strong inhibitor) and hERG inhibition (predictor II, inhibitor). hERG inhibition is related to QT prolongation. Moreover, CGEA results indicate that Cluster 5 drugs are enriched for the side effect electrocardiogram QT prolonged. These aforementioned analyses suggest that some serotonergic drugs (especially in Cluster 5) may cause QT prolongation. Indeed, some recent studies revealed the association between serotonergic drugs and QT prolongation [67,68].

#### Case study: drug repositioning using 4 clusters of iDSN

To illustrate the drug repositioning performance of PIMD, we examined some clusters in this section.

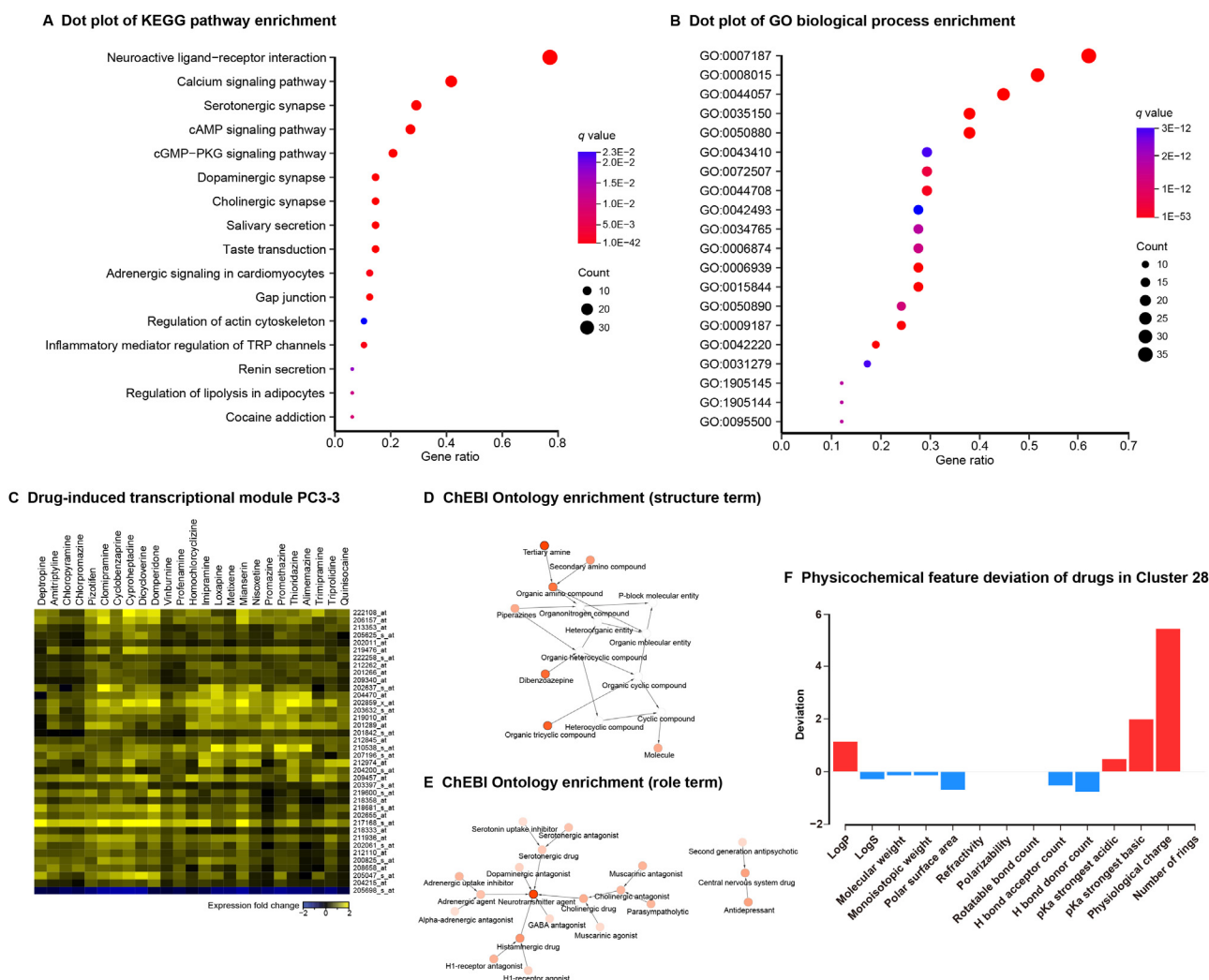
The Cluster 28 drugs (Figure 2) are enriched for the ‘nervous system’ ATC code, the ‘benzenoids’ superclass label, and the ‘G-protein coupled receptors’ target class. KEGG pathway enrichment and GO enrichment analysis results also include terms related to the nervous system (Figure 6A and B, Figure S4A and B; Table S9). Our CGEA analysis shows that Cluster 28 drugs are enriched for the molecular fragment CCCN(C)C and drug-induced transcriptional module PC3-3 (Figure 6C). The module contains expression profiles for 38 genes in response to 25 drugs, 6 drugs out of which are included in Cluster 28 [69]. The most prominent drug mode of action among Cluster 28 drugs is ‘Antihistamines for systemic use’, which can affect the central nervous system [70]. Chemical ontology enrichment analysis illustrates that Cluster 28 drugs are enriched with neurotransmitters, neurotransmitter derivatives, and central nervous system drugs (Figure 6D–F; drug property analysis results in Figure 6F). We found that cyclobenzaprine, a drug used to treat skeletal muscle spasms and fibromyalgia, is an unexpected drug in Cluster 28. Cyclobenzaprine, as well as another drug named the anti-depression drug amitriptyline in this cluster, exhibits antagonistic effects on the 5-hydroxytryptamine receptor 2A and possesses a tricyclic structure. Cyclobenzaprine is being studied for the post-traumatic stress disorder treatment according to ClinicalTrials.gov (<https://clinicaltrials.gov/>), a database of

clinical studies conducted around the world. Therefore, the novel therapeutic property for cyclobenzaprine might be N (nervous system). These results indicate that Cluster 28 drugs are related to nervous system and show the good compatibility. Unexpected drugs in Cluster 28 also have the potential to be repositioned to ‘nervous system’ ATC code.

Similarly, pentoxifylline, an unexpected drug in Cluster 16 (drugs in this cluster are enriched for the ‘respiratory system’ ATC code), is an interesting case. Pentoxifylline carries an ATC code for ‘cardiovascular system’ and is used in therapy for intermittent claudication [71]. Notably, it has a high similarity score with theophylline (ATC code for ‘respiratory system’), another drug in Cluster 16 used to treat respiratory diseases such as asthma. They are both members of the xanthine family, so they have a relatively high similarity score based on chemical structure (rank of 2898 in DSN-C). They also have a relatively high similarity score based on drug targets (rank of 7277 in DSN-T) because of the presence of 5 common targets. Pentoxifylline can increase red blood cell deformability and decrease blood viscosity [72]. But its similarity with theophylline has encouraged researchers to explore the potential of pentoxifylline to treat asthma [73].

Clusters simultaneously enriched in two therapeutic properties can also provide a unique perspective for drug repurposing. Cluster 3 drugs (Figure 2) are enriched for both the ‘cardiovascular system’ and ‘genitourinary system and reproductive hormones’ ATC codes (Figure S5A–G; Table S9). In this cluster, iloperidone, a drug used for schizophrenia, may be repositioned for applications related to the ‘cardiovascular system’ and the ‘genitourinary system and reproductive hormones’. Indeed, a previous study showed that repeat administration of iloperidone moderated hypotension [74]. Interestingly, we noted that sildenafil, a drug that was successfully repositioned from ‘cardiovascular system’ treatment to ‘genitourinary system and reproductive hormones’ treatment, is also present in Cluster 3, suggesting that drugs in this cluster may have the potential for both therapeutic properties. Likewise, prazosin and terazosin are an interesting pair. Prazosin (ATC code for ‘cardiovascular system’) is used to treat hypertension and also for urinary hesitancy associated with prostatic hyperplasia [75]. Terazosin (ATC code for ‘genitourinary system and reproductive hormones’) is used to treat enlarged prostate symptoms, which can also moderate hypertension [76,77]. Hence, the primary indications of these two drugs can be treated as secondary indications of each other. In terms of target proteins, both drugs have antagonistic effects on  $\alpha$ -1A, -1B, and -1D adrenergic receptors. In terms of chemical structures, their most common substructure is the molecular fragment O=CN1CCNCC1. Their common adverse reactions include dizziness, headache, drowsiness, lack of energy, and weakness.

Another interesting case is cyproterone acetate (ATC code for ‘genito-urinary system and sex hormones’) in Cluster 19. Cyproterone acetate can not only treat androgen-dependent conditions like excessive hair growth and acne but also treat prostate cancer [78]. Therefore, the novel therapeutic property for cyproterone acetate could be L (antineoplastic and immunomodulating agents). More interestingly, cyproterone acetate was originally developed as a progestin [79], but it was first marketed as an antiandrogen [80]. Cluster 19 drugs are enriched for both the ‘genito-urinary system and sex hormones’ and ‘antineoplastic and immunomodulating agents’



**Figure 6** Community analysis of Cluster 28

**A.** Dot plot of KEGG pathway enrichment results. **B.** Dot plot of GO biological process enrichment results. Dot size and color indicate the count of enriched genes in each of the categories and the corresponding significance of enrichment, respectively. Gene ratio represents the ratio of enriched genes to all genes in each of the categories. GO IDs are presented here for simplicity. The list of corresponding GO terms associated with these IDs can be found in Table S9. **C.** Drug-induced transcriptional module PC3-3. The transcriptional module data were obtained from the study of Iskar and colleagues [69]. Drug enrichment analysis was performed using CGEA. The horizontal axis represents the drug name and the vertical axis represents the gene probe ID, respectively. **D.** ChEBI Ontology structure term enrichment results. **E.** ChEBI Ontology role term enrichment results. Nodes indicate the enriched ChEBI terms. The lower the node transparency is, the more significantly the term is enriched. Arrows proceed from child to parent terms. **F.** Physicochemical feature deviation of drugs in Cluster 28. Red and blue bars represent the indicated drug feature of Cluster 28 that is higher or lower than the average of all drugs in the DrugBank database, respectively. CGEA, chemogenomic enrichment analysis; ChEBI, Chemical Entities of Biological Interest.

ATC codes. In this cluster, cyproterone acetate has high similarity scores with bicalutamide (ATC code for ‘antineoplastic and immunomodulating agents’) and medroxyprogesterone acetate (ATC code for ‘antineoplastic and immunomodulating agents’ and ‘genito-urinary system and sex hormones’). For the drug pair of cyproterone acetate and medroxyprogesterone acetate, their similarity score based on the chemical structure is relatively high (rank of 101 in DSN-C), but the similarity score based on drug targets is relatively low (rank of 35,974 in DSN-T). On the contrary, cyproterone acetate and bicalutamide have relatively high similarity score based on drug targets (rank of 178 in DSN-T) but relatively low similarity score based on chemical structures (rank of 81,257 in DSN-C). This

is because cyproterone acetate is more structurally similar to medroxyprogesterone acetate (a type of progestin) but has antiandrogenic effects on androgen receptor like bicalutamide (a kind of antiandrogen). This explains the uniqueness of cyproterone acetate. These results also show that PIMD takes advantage of the universality, individuality, and complementarity of different drug properties.

## Discussion

Here, we proposed PIMD, an integrative, systematic, and extensible framework for discovering novel therapeutic

properties of drugs from heterogeneous data sources. PIMD characterizes the iDSN by integrating chemical structure data, target protein sequence data, and side effect data; and performing spectral clustering of the iDSN identified 32 drug clusters. Additionally, PIMD facilitates drug repositioning from two aspects: drug pairs with high iDSN similarity score and unexpected drugs in each cluster. Finally, via a series of enrichment analyses, PIMD annotates and evaluates all clusters from chemical, pharmacological, and genomic views. Thus, PIMD screens suitable clusters for drug precision repositioning.

By integrating multi-dimensional drug informatics data, PIMD can capture potential similarities between drugs with sensitivity. The iDSN is superior to single-property DSNs in multiple evaluation measurements, including AOR, NMI, and cluster compactness. In this study, we primarily used ATC label as a golden standard to annotate drugs and evaluate the performance of PIMD and other compared methods. Interestingly, if the superclass label is used to calculate the evaluation measurements instead of ATC code, the SOR and NMI values of chemical structure-based DSN are higher than those of other DSNs, even the iDSN (Figure S6A and B). This suggests that chemical property has a large positive effect on superclass, whereas clinical and pharmacological properties could affect drug superclass label negatively. Indeed, ATC label and superclass label are two different drug catalogs. ATC label can comprehensively reflect therapeutic properties of drugs, thus commonly used to classify and label drugs, whereas superclass label primarily represents the chemical properties of the drugs. This result can improve our understanding of the difference between ATC and superclass labels of the drugs.

PIMD offers new insights into the universality, individuality, and complementarity of three drug properties, including chemical, pharmacological, and clinical properties. Our calculations of the relative contribution of each data type indicate that the iDSN is driven by all data types. PIMD makes full use of the information on each property. Examination of NMI among single-property, dual-property, and three-integrated-property DSNs (Figure S7A–D) shows that networks based on single property alone overlap marginally but are complementary to each other instead. Furthermore, PIMD provides the drug property contribution to each cluster, improving our understanding of the cluster characteristics (Table S1).

Rather than simple fusion, PIMD provides a systematic way to evaluate the drug cluster and drug repositioning. Pharmacology and pharmacogenomics researchers can use PIMD to screen drug clusters based on their own requirements (Figure 5A–D, Figure S3A–F; Tables S7 and S8).

As a highly extensible framework, PIMD can fuse various properties beyond the three properties examined here. For example, combining drug expression profiles would allow us to elucidate similarities between drugs at the transcriptional level. Drugs have multi-dimensional properties, and drug effects on disease processes are an interdisciplinary issue. The complementarity among multiple properties allows us to assess drug similarities more accurately and thus to provide a more comprehensive and clearer direction for drug repositioning.

There are two reasons why we chose these three data sources as an example. Firstly, these three data sources are representative of a variety of drug informatics data. Side effect,

chemical structure, and molecular target data represent the clinical, chemical, and pharmacological properties of drugs, respectively. The three properties comprehensively summarize the drug characteristics. Secondly, these three data sources have sufficient data to extract. We can collect plenty of these data from the public databases. Moreover, despite insufficient data available, we also compared iDSN with DSNs based on three other data types: 1) drug 3D structure, 2) drug expression profiles, and 3) PPI network (similarity measurements can be found in File S1). The results show that iDSN performs best compared with other DSNs (Figure S8).

The current study is limited by the set of drugs available when using the intersection of drug sets with multi-dimensional properties. If a certain property of a drug is not available, it would be excluded from the construction of iDSN. Therefore, with the accumulation of drug informatics data in the future, we expect that the scale of iDSN would be expanded, and the performance of PIMD would be further improved accordingly.

In summary, PIMD provides a new perspective for drug repositioning through multi-property fusion and an analysis package. It facilitates to understand the integration of drug properties at a deeper level, and its high expansibility and modularity would allow users to explore drugs from a wider range of fields.

## Code availability

Source code of PIMD is available at <https://github.com/Sepstar/PIMD/>.

## CRedit author statement

**Song He:** Conceptualization, Methodology, Investigation, Software, Visualization, Writing - original draft, Writing - review & editing. **Yuqi Wen:** Conceptualization, Methodology, Investigation, Software, Visualization, Writing - original draft, Writing - review & editing. **Xiaoxi Yang:** Investigation, Visualization. **Zhen Liu:** Investigation, Visualization. **Xinyu Song:** Investigation, Visualization. **Xin Huang:** Investigation, Visualization. **Xiaochen Bo:** Conceptualization, Methodology, Supervision, Software, Writing - review & editing. All authors read and approved the final manuscript.

## Competing interests

The authors have declared no competing interests.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. U1435222) and the Program of International Sci-Tech Cooperation, China (Grant No. 2014DFB30020).

## Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gpb.2018.10.012>.



## ORCID

0000-0002-4136-6151 (Song He)  
 0000-0002-2391-3297 (Yuqi Wen)  
 0000-0002-2674-1902 (Xiaoxi Yang)  
 0000-0002-3327-4118 (Zhen Liu)  
 0000-0002-1905-3503 (Xinyu Song)  
 0000-0002-6569-0095 (Xin Huang)  
 0000-0003-1911-7922 (Xiaochen Bo)

## References

- [1] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;9:203–14.
- [2] Adams CP, Brantner VV. Estimating the cost of new drug development: is it really \$802 million dollars?. *Health Aff* 2006;25:420–8.
- [3] DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ* 2003;22:151–85.
- [4] Boguski MS, Mandl KD, Sukhatme VP. Repurposing with a difference. *Science* 2009;324:1394–5.
- [5] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;42:D1091–7.
- [6] Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, et al. SuperTarget and Matador: resources for exploring drug–target relationships. *Nucleic Acids Res* 2008;36:D919–22.
- [7] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. *Nucleic Acids Res* 2016;44:D1202–13.
- [8] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;44:D1075–9.
- [9] Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A* 2010;107:14621–6.
- [10] Xie L, He S, Wen Y, Bo X, Zhang Z. Discovery of novel therapeutic properties of drugs from transcriptional responses based on multi-label classification. *Sci Rep* 2017;7:7136.
- [11] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. *Nature* 2009;462:175–81.
- [12] Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science* 2008;321:263–6.
- [13] Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. *PLoS One* 2011;6:e28025.
- [14] Yıldırım MA, Goh KI, Cusick ME, Barabási AL, Vidal M. Drug–target network. *Nat Biotechnol* 2007;25:1119–26.
- [15] Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;25:197–206.
- [16] Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* 2008;4:682–90.
- [17] Coelho ED, Arrais JP, Oliveira JL. Computational discovery of putative leads for drug repositioning through drug–target interaction prediction. *PLoS Comput Biol* 2016;12:e1005219.
- [18] Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 2016;32:2338–45.
- [19] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;11:333–7.
- [20] Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. *J Cheminform* 2013;5:30.
- [21] Wang Y, Chen S, Deng N, Wang Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One* 2013;8:e78518.
- [22] Zhang P, Agarwal P, Obradovic Z. Computational drug repositioning by ranking and integrating multiple data sources. *Mach Learn Knowl Discov Databases* 2013:579–94.
- [23] Liu H, Song Y, Guan J, Luo L, Zhuang Z. Inferring new indications for approved drugs via random walk on drug-disease heterogeneous networks. *BMC Bioinformatics* 2016;17:539.
- [24] Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-learning-based drug–target interaction prediction. *J Proteome Res* 2017;16:1401–9.
- [25] Nascimento AC, Prudêncio RB, Costa IG. A multiple kernel learning algorithm for drug–target interaction prediction. *BMC Bioinformatics* 2016;17:46.
- [26] Shi JY, Yiu SM, Li Y, Leung HC, Chin FY. Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods* 2015;83:98–104.
- [27] Naveed H, Hameed US, Harrus D, Bourguet W, Arold ST, Gao X. An integrated structure- and system-based framework to identify new targets of metabolites and known drugs. *Bioinformatics* 2015;31:3922–9.
- [28] Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. *AMIA Annu Symp Proc* 2014;2014:1258–67.
- [29] Yamanishi Y, Kotera M, Moriya Y, Sawada R, Kanehisa M, Goto S. DINIES: drug–target interaction network inference engine based on supervised analysis. *Nucleic Acids Res* 2014;42:W39–45.
- [30] Fakhraei S, Huang B, Raschid L, Getoor L. Network-based drug–target interaction prediction with probabilistic soft logic. *IEEE/ACM Trans Comput Biol Bioinform* 2014;11:775–87.
- [31] Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug–target interactions. *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2013;1025–33.
- [32] Mei JP, Kwok CK, Yang P, Li XL, Zheng J. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 2013;29:238–45.
- [33] Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, et al. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017;8:573.
- [34] Wang W, Yang S, Zhang X, Li J. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 2014;30:2923–30.
- [35] Xia Z, Wu LY, Zhou X, Wong ST. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst Biol* 2010;4:S6.
- [36] Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009;25:2397–403.
- [37] van Laarhoven T, Marchiori E. Predicting drug–target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* 2013;8:e66952.
- [38] van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 2011;27:3036–43.
- [39] Wang Y, Zeng J. Predicting drug–target interactions using restricted Boltzmann machines. *Bioinformatics* 2013;29:i126–34.
- [40] Yuan Q, Gao J, Wu D, Zhang S, Mamitsuka H, Zhu S. DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 2016;32:i18–27.



- [41] Wan F, Hong L, Xiao A, Jiang T, Zeng J. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* 2019;35:104–11.
- [42] Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;26: i246–54.
- [43] He S, He H, Xu W, Huang X, Jiang S, Li F, et al. ICM: a web server for integrated clustering of multi-dimensional biomedical data. *Nucleic Acids Res* 2016;44:W154–9.
- [44] UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;43:D204–12.
- [45] Cao Y, Charisi A, Cheng LC, Jiang T, Girke T. ChemmineR: a compound mining framework for R. *Bioinformatics* 2008;24:1733–4.
- [46] Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;147:195–7.
- [47] Pawson AJ, Sharman JL, Benson HE, Faccenda E, Alexander SP, Buneman OP, et al. The IUPHAR/BPS guide to pharmacology: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res* 2014;42:D1098–106.
- [48] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353–61.
- [49] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;32: D258–61.
- [50] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;16:284–7.
- [51] Shameer K, Glicksberg BS, Hodos R, Johnson KW, Badgeley MA, Readhead B, et al. Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. *Brief Bioinform* 2018;19:656–78.
- [52] O’Boyle NM, Morley C, Hutchison GR. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem Cent J* 2008;2:5.
- [53] Wegner J. JOELib/JOELib2. Germany: Department of Computer Science, University of Tübingen; 2005.
- [54] Moreno P, Beisken S, Harsha B, Muthukrishnan V, Tudose I, Dekker A, et al. BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics* 2015;16:56.
- [55] Luxburg UV. A tutorial on spectral clustering. *Stat Comput* 2007;17:395–416.
- [56] Brock G, Pihur V, Datta S, Datta S. clValid, an R package for cluster validation. *J Stat Softw* 2008;25:1–22.
- [57] O’Dell JR, Haire CE, Erikson N, Drymalski W, Palmer W, Eckhoff PJ, et al. Treatment of rheumatoid arthritis with methotrexate alone, sulfasalazine and hydroxychloroquine, or a combination of all three medications. *N Engl J Med* 1996;334:1287–91.
- [58] Hannonen P, Möttönen T, Hakola M, Oka M. Sulfasalazine in early rheumatoid arthritis. a 48-week double-blind, prospective, placebo-controlled study. *Arthritis Rheumatol* 1993;36:1501–9.
- [59] Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* 2006;22:2405–12.
- [60] Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 2002;3:research0036.
- [61] Wrobel A, Eklund P, Bobrowska-Hägerstrand M, Hägerstrand H. Lignans and norlignans inhibit multidrug resistance protein 1 (MRP1/ABCC1)-mediated transport. *Anticancer Res* 2010;30:4423–8.
- [62] Ohta A, Sitkovsky M. Methylxanthines, inflammation, and cancer: fundamental mechanisms. *Handb Exp Pharmacol* 2011;469–81.
- [63] Uccella S, Mariani A, Wang A, Vierkant R, Cliby WA, Robien K, et al. Intake of coffee, caffeine and other methylxanthines and risk of Type I vs Type II endometrial cancer. *Br J Cancer* 2013;109:1908–13.
- [64] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545–50.
- [65] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 2012;41:D456–63.
- [66] Borden KL, Culjkovic-Kraljacic B. Ribavirin as an anti-cancer therapy: acute myeloid leukemia and beyond? *Leuk Lymphoma* 2010;51:1805–15.
- [67] Ballesteros A, Saiz H, Rosero Á, Portilla A, Montes L, Elorza R. SSRIs and QT interval prolongation management. A review. *Eur Psychiatry* 2017;41:S750.
- [68] Ojero-Senard A, Benevent J, Bondon-Guitton E, Durrieu G, Chebane L, Araujo M, et al. A comparative study of QT prolongation with serotonin reuptake inhibitors. *Psychopharmacology* 2017;234:3075–81.
- [69] Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, Kaminska KH, et al. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol Syst Biol* 2013;9:662.
- [70] Lieberman P. Histamine, antihistamines, and the central nervous system. *Allergy Asthma Proc* 2009;30:482–6.
- [71] Salhiyyah K, Forster R, Senanayake E, Abdel-Hadi M, Booth A, Michaels JA. Pentoxifylline for intermittent claudication. *Cochrane Database Syst Rev* 2015;9:CD005262.
- [72] Ward A, Clissold SP. Pentoxifylline. *Drugs* 1987;34:50–97.
- [73] Entzian P, Bitter-Suermann S, Burdon D, Ernst M, Schlaak M, Zabel P. Differences in the anti-inflammatory effects of theophylline and pentoxifylline: important for the development of asthma therapy?. *Allergy* 1998;53:749–54.
- [74] Sainati SM, Hubbard JW, Chi E, Grasing K, Brecher MB. Safety, tolerability, and effect of food on the pharmacokinetics of iloperidone (HP 873), a potential atypical antipsychotic. *J Clin Pharmacol* 1995;35:713–20.
- [75] Tsujii T. Comparison of prazosin, terazosin and tamsulosin in the treatment of symptomatic benign prostatic hyperplasia: a short-term open, randomized multicenter study. *Int J Urol* 2000;7:199–205.
- [76] Mersey JH. Long-term experience with terazosin for treatment of mild to moderate hypertension. *Am J Med* 1986;80:68–72.
- [77] Itskovitz HD. Alpha 1-blockade for the treatment of hypertension: a megastudy of terazosin in 2214 clinical practice settings. *Clin Ther* 1994;16:490–504.
- [78] Torri V, Floriani I. Cyproterone acetate in the therapy of prostate carcinoma. *Arch Ital Urol Androl* 2005;77:157–63.
- [79] Pucci E, Petraglia F. Treatment of androgen excess in females: yesterday, today and tomorrow. *Gynecol Endocrinol* 1997;11:411–33.
- [80] Wakelin SH, Maibach HI, Archer CB. Systemic drug treatment in dermatology: a handbook. Boca Raton: CRC Press; 2002.