Large-scale integration of heterogeneous pharmacogenomic data for identifying drug mechanism of action

Yunan Luo, Sheng Wang, Jinfeng Xiao and Jian Peng*

Department of Computer Science,
University of Illinois at Urbana-Champaign,
Urbana, Illinois 61801, USA
**Corresponding author:* `jianpeng@illinois.edu`

A variety of large-scale pharmacogenomic data, such as perturbation experiments and sensitivity profiles, enable the systematical identification of drug mechanism of actions (MoAs), which is a crucial task in the era of precision medicine. However, integrating these complementary pharmacogenomic datasets is inherently challenging due to the wild heterogeneity, high-dimensionality and noisy nature of these datasets. In this work, we develop Mania, a novel method for the scalable integration of large-scale pharmacogenomic data. Mania first constructs a drug-drug similarity network through integrating multiple heterogeneous data sources, including drug sensitivity, drug chemical structure, and perturbation assays. It then learns a compact vector representation for each drug to simultaneously encode its structural and pharmacogenomic properties. Extensive experiments demonstrate that Mania achieves substantially improved performance in both MoAs and targets prediction, compared to predictions based on individual data sources as well as a state-of-the-art integrative method. Moreover, Mania identifies drugs that target frequently mutated cancer genes, which provides novel insights into drug repurposing.

Keywords: data integration, drug mechanisms of action, drug target, drug similarity network, dimensionality reduction

1. Introduction

Accurate identification drug mechanism of actions (MoAs) and drug targets is of great importance for developing new drug as well as repurposing existing drugs. During the past decades, many computational approaches have been developed to identify drug MoAs and targets according to molecular docking analysis,[1] annotated target profiles,[2] adverse drug reactions,[3] and scientific literature.[4] However, these methods were limited to the prediction for drugs that are well-studied either in literature or existing biological experiment assays. Consequently, computational approaches that can be generalized to all drugs are a pressing need in the field.

Fortunately, with the recent advances in sequencing technology, large-scale pharmacogenomic data offers us exciting opportunities to systematically identify drug MoAs and targets. For example, chemical structure has been used to predict drug-target interaction.[5,6] The motivation behind this is that drugs that are structurally similar tend to interact with similar genes, thus sharing similar MoAs. Another notable dataset, drug perturbation data has also been widely used to identify MoAs.[7] Drug perturbation data, such as Connectivity Map (CMap) Library[8] and the L1000 dataset from the Library of Integrated Network-based Cellular Signatures (LINCS),[9] reveals drug-induced transcriptional profiles. It measures the gene expression change in the presence of a drug and these gene signatures enable the comparison between drugs. Moreover, high-throughput *in vitro* drug screening over large panels of tumor cell lines have been shown to be useful in identifying clinically relevant drugs. For example, the recent developed Cancer Therapeutics Response Portal (CTRP) project[10] contains the drug sensitivity profiles of 481 small-molecule compounds across 860 cancer cell lines, which provides additional insights into the MoA of small-molecule compounds and novel therapeutic hypotheses. Since drugs with the same MoAs tend to exhibit similar transcriptional and cellular responses, these more accessible pharmacogenomic collections can be used to systematically infer drug MoAs and targets.[7]

Intuitively, integrating these datasets can further improve the identification of drug MoAs and targets. However, the sheer amount and heterogeneity of these multi-omics data pose great challenges in the integration process: (i) the mixed formats, scales, and metrics, (ii) the complementary but high-dimensional information, and (iii) the incomplete and noisy nature of these datasets. As far as we know, Drug Network Fusion (DNF) [11] was the only previous attempt to simultaneously integrate the drug structure, perturbation and sensitivity data. Notably, DNF used a similarity network fusion approach,[12] in which a similarity network is constructed for each input data sources, and these similarity networks are then iteratively fused together until convergence to obtain a single similarity network. The major drawback of this approach is that the context-specific similarity measures were mixed together in the collapsed single network, where the context-specific information may be lost or obscured.

In this work, we introduce Mania (prediction of mechanism of action by network integration), a novel method for characterizing drug-drug relationships and predicting drug mechanism of actions (MoAs) and drug targets through integrating multiple large-scale pharmacogenomic data, including drug structure, sensitivity, and perturbation data. Mania takes full advantage of the fine-grained inherent structure in the individual data source and integrates heterogeneous information by learning low-dimensional vector representations for drugs, which best explain the relationships among drug across all pharmacogenomic data. We demonstrate that, unlike DNF which directly produces a drug-drug similarity matrix, Mania is a versatile method in that the low-dimensional vector representations of drugs not only capture more accurate similarity measure with any type of distance metric, but can also be used as plug-in feature vectors of many off-the-shelf machine learning algorithms for the prediction of drug MoAs and targets. Experiment results suggested that Mania outperforms DNF, the state-of-the-art method, with substantial improvements in MoAs/targets prediction. In addition, based on the low-dimensional vector representations of drugs, Mania consistently identified functionally-enriched drug clusters, in which drugs within the same cluster are interacting with same targets. Moreover, we show that Mania found new drugs that may target significantly mutated cancer genes, which provides potential insights into drug repurposing. Overall, our experiment results suggested the superior ability of Mania in integrating multiple pharmacogenomic data for drug MoAs and targets prediction, and also demonstrated its potential as a practical tool to support network pharmacology.

2. Materials and Methods

We first provide an overview of Mania (Fig. 1). Taking one or more types of drug-related data for the same set of drugs as input, Mania first constructs a similarity network for each type of data source separately. It then integrates these heterogeneous similarity networks by combining a network diffusion algorithm and a dimensionality reduction scheme to learn a low-dimensional vector representation for each drug. These vector representations of drugs simultaneously capture the complementary information from different data sources. Intuitively, the vector representations of two drugs will be co-localized in the low-dimensional space if they are structurally similar or functionally correlated, e.g., share common chemical structure features, have similar sensitivity profiles, and/or perturb the same set of genes. Finally, Mania constructs the integrated drug-drug similarity network and infers MoAs and targets based on the low-dimensional vector representations.

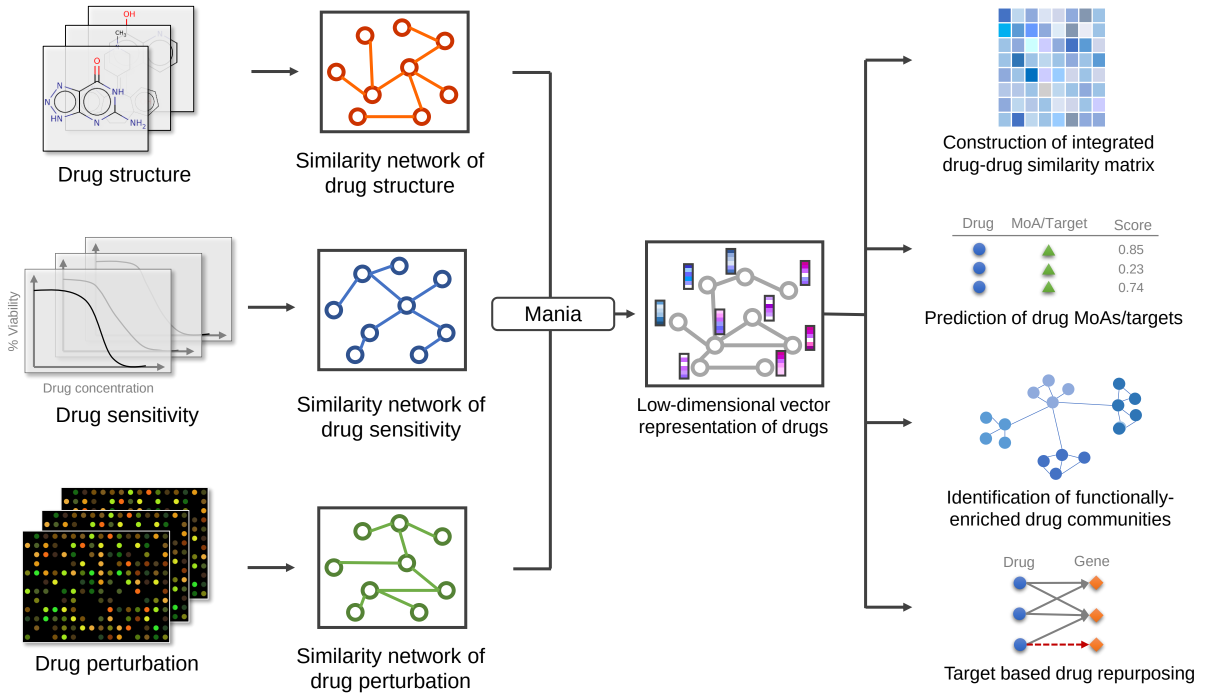
Drug structure
Similarity network of drug structure
% Viability
Drug concentration
Drug sensitivity
Similarity network of drug sensitivity
Drug perturbation
Similarity network of drug perturbation
Mania
Low-dimensional vector representation of drugs
Construction of integrated drug-drug similarity matrix
Drug
MoA/Target
Score
0.85
0.23
0.74
Prediction of drug MoAs/targets
Identification of functionally-enriched drug communities
Drug
Gene
Target based drug repurposing

Fig. 1. **Schematic illustration of Mania.** Mania integrates multiple pharmacogenomic data sources and represents each drug with a low-dimensional vector representation. Mania can be used in a variety of tasks, including constructing integrated drug-drug similarity network, predicting MoAs and targets for drugs, identifying drug communities that share common MoAs/targets, and target based drug repurposing.

2.1. *Construction of heterogeneous drug-drug similarity networks*

In this work, we integrate three different types of drug-related data, including drug structure, sensitivity, and perturbation data. Each type of data is presented in heterogeneous formats (e.g., sequences that represent the chemical structures of drugs, and matrices that represent gene expression in response to different concentrations of a drug), and characterizes the properties of drugs from various aspects. To extract the information of drug-drug relationships encoded in the heterogeneous data, we construct a similarity network for each type of the data sources.

Drug perturbation. We obtained the drug perturbation data from the L1000 dataset[9] from the Integrated Network-Based Cellular Signatures (LINCS) Program (`http://www.lincsproject.org/`). The L1000 dataset produced over one million gene expression profiles of 1,000 landmark genes in response to the treatment of 20,413 unique compounds across many cancer cell lines, subject to various perturbation conditions. We used the PhamacoGx package[13] to download the transcriptional profiles, and compute a "signature" for each drug that quantifies the effect of drug concentration on the gene expression with a linear regression model. To characterize the relationships between a pair of drugs based on whether they perturb the same set of genes with similar patterns, we compute the pairwise drug similarity using the Pearson correlation between their drug perturbation signatures.

Drug structure. We collected the canonical SMILES strings for the small molecules in the L1000 dataset from the PubChem database.[14] We used the RDKit[15] library to parse the SMILES strings, generate fingerprints, and compute structure similarity. We generated the Morgan fingerprint[16] (also known as circular fingerprints) with radius 2 for each drug, which takes into account both the atomic properties and the neighborhood information of each atom. Given the fingerprints of a pair of drugs, the structure similarity between them was then calculated using the Dice coefficient,[17] which is a real value in $[0, 1]$ that measures the extent to which pairs of drugs share similar structure features.

Drug sensitivity. We used the drug sensitivity data released in a recent work,[18] which is also available at the Cancer Cancer Therapeutics Response Portal (`http://www.broadinstitute.org/ctrp/`). The dataset contains sensitivity patterns for 481 compounds (including FDA-approved drugs and clinical candidates) spanning 842 different human cancer cell lines encompassing 25 lineages. We extracted the area under curve (AUC) of the concentration-response curve as the metric of sensitivity, which measures the cellular response to individual compound. To quantify the relationships between a pair of drugs based on whether they cause similar responses to same cancer cell lines, we calculate the pairwise drug similarity using the Pearson correlation between their drug sensitivity profiles.

Drug MoAs and targets. The recently released Drug Repurposing Hub[19] is a repository that contains a drug screening collection of 4,707 compounds with extensively curated annotations (e.g., mechanism of action, target, SMILES string, drug indication, and disease area) for each drug. Drug Mechanism of actions and targets were exported from the repository and used as the ground truth in the experiments of this work.

Intersection of drugs in multi-omic data. Overlapping the drugs in the drug perturbation data (20,413 drugs) and sensitivity data (481 drugs), we obtained a common set of 277 drugs that are shared in the two types of data. The drug structures of these 277 drugs were collected from the PubChem database. Each of the 277 drugs was then searched in the Drug Repurposing Hub, and 170 of them were found to have annotated MoA and target information. The MoAs and targets of these drugs were extracted for evaluation in our experiments.

2.2. *Integration of multi-omics data*

Multi-omics data provide drug-related information from diverse data sources and integration methods can shed light on the properties of less-characterized drugs. Here, we used our recently developed network integration algorithm Mashup[20,21] to integrate three types of multi-omics data, including drug structure, drug sensitivity and drug perturbation profiles. Mashup has been demonstrated to achieve significantly improved prediction for protein function prediction, gene ontology reconstruction, genetic interaction prediction, and drug-target interaction prediction.[20–23] It takes one or more networks as input, performs random walk with restart (RWR)[24] and extracts topological information from the diffusion distributions using informative but low-dimensional vector representations of drugs.

Formally, let $\mathbf{A}$ denote the weighted adjacency matrix of a certain type of similarity network of n drugs (for example, let $A_{i,j}$ be the chemical structure similarity between drugs i and j). The transition matrix of the RWR can then be calculated as $\mathbf{B}_{i,j} = \mathbf{A}_{i,j} / \sum_{j'} \mathbf{A}_{i,j'}$. Let $\mathbf{s}_i^t$ be an n-dimensional distribution vector in which each element stores the probability of a node being visited from node i after t iterations of the random walk, the RWR process is then defines as $\mathbf{s}_i^{t+1} = (1 - p_r)\mathbf{s}_i^t\mathbf{B} + p_r\mathbf{e}_i$, where $\mathbf{e}_i$ stands for an n-dimensional vector with $\mathbf{e}_i(i) = 1$ and $\mathbf{e}_i(j) = 0$, $\forall j \neq i$, and p_r is the restart probability controlling the relative influence between local and global topological information in the diffusion process. At this fixed point of the RWR, we can obtain the "diffusion state" $\mathbf{s}_i^\infty$ for drug i (i.e., $\mathbf{s}_i = \mathbf{s}_i^\infty$), in which the jth element $\mathbf{s}_{ij}$ of the diffusion state stores the probability of RWR starting node i and ending up at node j in equilibrium.

The diffusion states resulting from the aforementioned RWR process may not be entirely accurate, partially due to the low-quality and high-dimensionality of biological data. One of the strengths of Mashup is that it teases functionally relevant topological patterns apart from noise in the diffusion states and jointly integrates heterogeneous information from L similarity networks by learning low-dimensional vector representations of drugs. With the goal of denoise and dimensionality reduction, Mashup approximates each the diffusion state $\mathbf{s}_i^{(l)}$ of drug i in network l with a multinomial logistic model parameterized by low-dimensional feature vectors:

$$\hat{\mathbf{s}}_{ij}^{(l)} = \frac{\exp\left(\mathbf{x}_j^T \mathbf{w}_i^{(l)}\right)}{\sum_{j'} \exp\left(\mathbf{x}_{j'}^T \mathbf{w}_i^{(l)}\right)}, \tag{1}$$

where $\forall i$, $\mathbf{w}_i^{(l)}, \mathbf{x}_i \in \mathbb{R}^d$ for $d \ll n$. For drug i, We refer to $\mathbf{w}_i^{(l)}$ as the *context feature* which is network-specific for network l, and $\mathbf{x}_i$ as the *node feature* which is shared globally across all networks. Finally, Mashup uses the Kullback-Leibler (KL) divergence to guide the learning of the two low-dimensional vectors,

$$\min_{\mathbf{w},\mathbf{x}} C(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n} D_{KL}\left(\mathbf{s}_i^{(l)} \parallel \hat{\mathbf{s}}_i^{(l)}\right). \tag{2}$$

The vectors $\{\mathbf{x}_i\}$ are subsequently used as the low-dimensional vector representations of drugs. If two drugs have similar vector representations, it generally implies that they have similar positions with respect to other drugs in the network, and thus probably share similar functions.

2.3. *Prediction of MoAs and drug targets*

To predict the MoAs and drug targets, Mania first identifies similar drugs based on the low-dimensional vector representations of drugs. In the experiments throughout this work, we used the cosine distance between the feature vectors as the distance metric for a pair of drugs i and j, following the previous work:[21]

$$D_{cos}(i,j) = 1 - \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}, \tag{3}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors of drugs i and j, respectively.

After computing the distances, Mania is able to predict the MoAs and targets for drugs that are less well-characterized using a k-nearest neighbor approach, i.e., predicting the MoAs and targets for a drug by transferring the knowledge of its k most similar drugs based on the distances computed above. Specifically, Mania calculates the affinity score of drug i and MoA (or target) j as a weighted majority voting by the k most similar drugs of drug i:

$$\mathbf{s}_{i,j} = \sum_{d \in \mathcal{N}_i} \cos(\mathbf{x}_i, \mathbf{x}_j) \mathbb{I}[d \in M_j], \tag{4}$$

where $\mathcal{N}_i$ is the set of the k most similar drugs of drug i, $\mathbb{I}[\cdot]$ is the indicator function, and M_j is the set of drugs that are annotated with MoA j in the training data. We set $k = 10$ in our experiments.

3. Results

We evaluate the ability of our Mania framework on uncovering the drug-drug relationships and predicting drug MoAs and drug targets by integrating multi-omics data. The integrated drug-drug similarity network given by Mania achieved an AUPRC score of 0.892, which is a substantial improvement over DNF[11] (0.838), the state-of-the-art integration method for drug taxonomy. The low-dimensional vector representations of drugs learned by Mania can also be used as plug-in features for off-the-shelf machine learning algorithms. We show that by using its learned feature vectors as the input of a k-nearest neighbor (kNN) algorithm, Mania successfully recovering around 75% true MoAs associated with drugs when evaluated with five-fold cross-validation on the list of its top 10 predictions, which is remarkably 20% higher than the DNF method. The details of our experiments are described below.

3.1. *Mania improves the quantification of drug-drug similarity*

Accurate quantification of drug similarity can help elucidate the drug-drug relationships and predict new targets for existing drugs. To assess the ability of Mania on quantifying the drug similarity, we calculated the cosine similarity among the low-dimensional vector representations between pairs of drugs. The cosine similarity matrix was compared to a binary drug similarity matrix, where an entry was set to 1 if the pair of drugs shares at least one MoA, and 0 otherwise. We filtered the MoAs to retain only those MoAs that are associated with at least two drugs before computing the binary similarity matrix. We evaluate the performance by computing the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Note the whole process is unsupervised and no MoA information was available to Mania when learning the low-dimensional vector representations and computing the similarity metric. We implemented Mania based on Mashup[20] (`http://mashup.csail.mit.edu/`). We set the dimensionality of the low-dimensional vector representations of drugs as $d = 10$. The restart probability p_r of RWR was set to 0.8. We observed stable performances a wide range of values of d and p_r in our experiments.

We first compared the integrated similarity network by Mania's integration of multi-omics data with three similarity networks that were computed based on individual omics data, including the drug structure, drug sensitivity and drug perturbation (Fig. 2). We noticed that although the individual similarity networks of drug structure and drug sensitivity achieved roughly the same AUROC score (around 0.80), there was a 20% gap between their performances and that of the similarity network computed based on drug perturbation, which means there was a 20% fraction of drug-drug relationships (i.e., drug pairs that share same MoAs) that cannot be accurately predicted by drug perturbation data only. The similar effects were also observed for the AUPRC scores, where there were noticeable gaps between the individual networks of drug structure, drug sensitivity, and drug perturbation. These findings suggested that each omics data are not redundant. Instead, these multi-omics data are complementary and an integration of them would improve the quantification of drug-drug similarities. Even if sensitivity data and structure data have roughly the same AUROC score, the similar drug pairs identified by these two data sources were inherently different, thus motivating us to further integrate them. The performance of the integrated similarity network by Mania confirmed this hypothesis, where the AUROC score was substantially improved to 0.892, an 11% improvement over the best individual similarity network, and the AUPRC score was also significantly improved to 0.423, 43% higher than the best individual similarity network. We also observed similar results for the evaluation on the target data, in which the binary drug similarity matrix was computed based on whether two drugs share at least one common target. Notably, the performance of perturbation-based network was the worst among all three similarity networks, possibly due to the noisy and batch effect in large-scale perturbation experiments.

Furthermore, we compared Mania with DNF,[11] a state-of-the-art integration method for drug taxonomy. DNF was built upon the similarity fusion network (SNF) method, which takes individual networks as input, iteratively updates every network by message passing until convergence to a single network. Unlike our method that outputs low-dimensional vector representations that can be used to compute any kind of similarity of distance metric, the DNF method directly outputs the converged single network as a similarity network. We found that although both DNF and Mania improved the performance over individual networks, the performances of Mania were substantially higher than that of DNF when evaluated on the binary similar matrices based on both MoA and drug target data (one-sided Wilcoxon rank-sum test $P < 0.001$). For example, on the MoA data, Mania achieved a 25% improvement on AUPRC over DNF (AUPRC of 0.423 and 0.339 for Mania and DNF, respectively) and a 5% improvement on AUROC over DNF (AUPRC of 0.892 and 0.849 for Mania and DNF, respectively). Further comparisons suggested that Mania also outperformed a recently proposed matrix factorization-based integration framework, Collective-Matrix Factorization (CMF).[25]

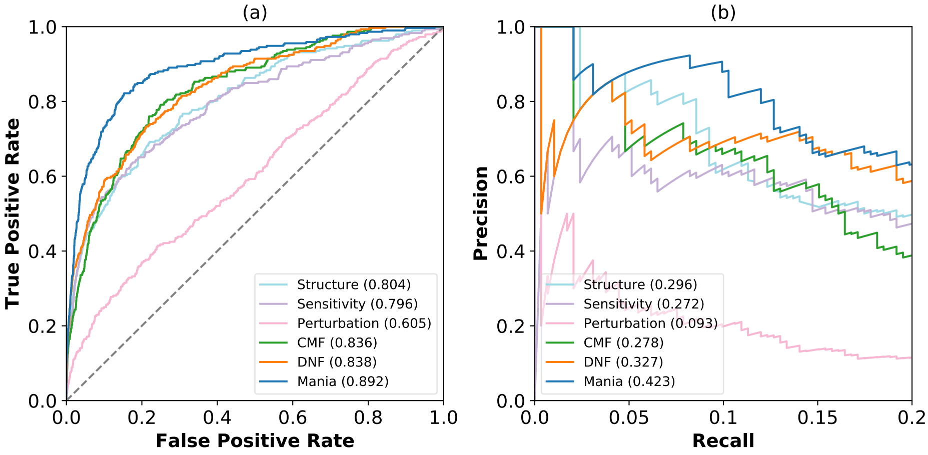

Fig. 2. **Accurate quantification of drug-drug similarity.** The similarity matrices based on individual data sources and the integration of DNF and Mania were compared to the binary similarity matrix derived from the ground truth data of drug MoA associations. Performance was evaluated using AUROC score (a) and AUPRC score (b).

These results suggested that Mania is capable of integrating drug properties from multi-omics data to provide a more comprehensive drug-drug similarity measure, which is helpful for elucidating the drug-drug relationships and potentially useful in MoA and target prediction of drugs, which we will demonstrate in the next section.

3.2. *Mania achieves accurate prediction of drug MoAs and targets*

In reality, one can collect and leverage the existing MoA (or target) information of well-studied drugs to infer the MoAs (or targets) of less well-characterized drugs. Therefore, it is also important to assess the ability of Mania on predicting drug MoAs and targets in a supervised way. As Mania outputs low-dimensional vector representations of drugs, we can use these vectors as plug-in features of drugs in off-the-shelf machine learning algorithms. In particular, we cast the prediction of drug MoAs and targets as a multi-label classification task and applied a k-nearest neighbor (kNN) method with the vector representations of drugs as input features. We assessed the performance of prediction using a five-fold cross-validation. For each test drug, we ran a weighted majority voting among its $k = 10$ most similar drugs based on the cosine distances imposed over the vector representations of drugs. Following the previous work,[26] we used the “recall@top-R” as the evaluation metric, which is defined as the fraction of true associated MoAs (or targets) that were retrieved in the list of top-R predictions for a drug. The motivation of using this metric was that a method that can recover the true MoAs (or targets) in the top-R predictions with high probability is desirable and useful in applications such as drug repurposing.

We compared the performance of Mania, CMF and DNF evaluated by recall@top-R for $R = 1, 5, 10$ on the MoA and target data (Fig. 3). We observed that Mania correctly recovered more MoAs and targets across all values of R. We would like to highlight the noticeable improvements of our method on smaller values of R (e.g., R =1 or 5), as evaluation under smaller

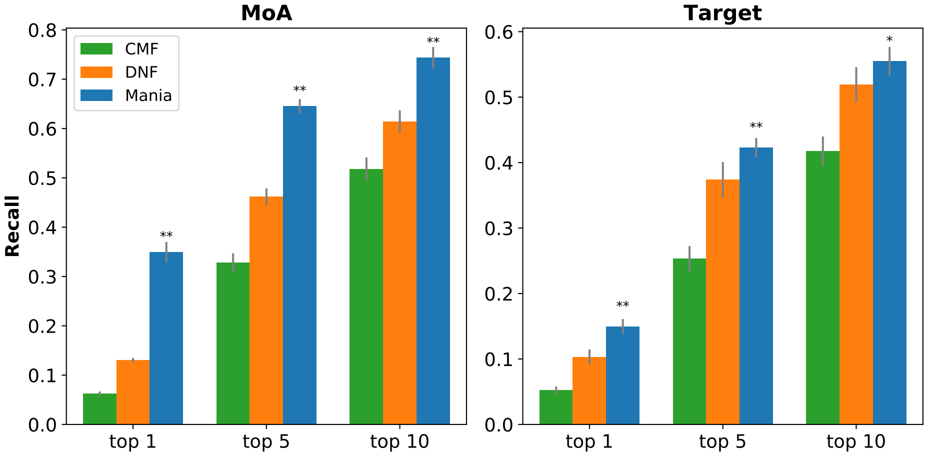

Fig. 3. **Comparison on the performance of drug MoAs and targets prediction.** We performed ten trials of five-fold cross-validation to compare the drug MoA/target prediction performance of Mania, CMF and DNF. The recall@top-R was used as the evaluation metric. *: $P < 0.01$ and **: $P < 0.001$, one-sided Wilcoxon rank-sum test.

values of R is a more challenging metric and reflects the precision of the very top predictions of a method. For instance, we observed Mania achieved a 0.350 recall@top-1 score, much higher than that of DNF (0.129) and CMF (0.062). We also found that the improvements of Mania over the other two methods were statistically significant (one-sided Wilcoxon rank-sum test $P < 0.01$). The superior performance of Mania demonstrates its potential on transferring the information of existing drugs to infer new MoAs or targets of drugs that are not well-studied, thus serving as a practical tool for drug repurposing.

3.3. *Identification of functionally-enriched drug communities*

The low-dimensional vector representations learned by Mania are able to capture the context-specific information and vectors of drugs that are functionally correlated will be co-localized in the low-dimensional space. To assess the pharmacological relevance among drugs exhibited by the vector representations of drugs learned by Mania, we applied an affinity propagation clustering algorithm on the low-dimensional vector representations of the common set of 277 drugs, with cosine distance as the distance metric. The affinity propagation algorithm requires an input "preference" parameter determines the likelihood of a particular drug to become an "exemplar" of a community (cluster). We set this preference parameter to be the 90% percentile of all pairwise similarities to encourage a relatively large number of communities to be produced, thus enabling each community to have more distinguishable pharmacological features. Note that the clustering was solely based on the low-dimensional vector representations that were learned by integrating multi-omics data, and no information of drug MoAs or targets was used to guide the clustering process.

We obtained 29 drug communities, with one drug selected as the representative (exemplar) drug in each community (Fig. 4). The size of the communities varies from 2 to 13, with a median size of 6 drugs. We observed that based on the low-dimensional vector representations

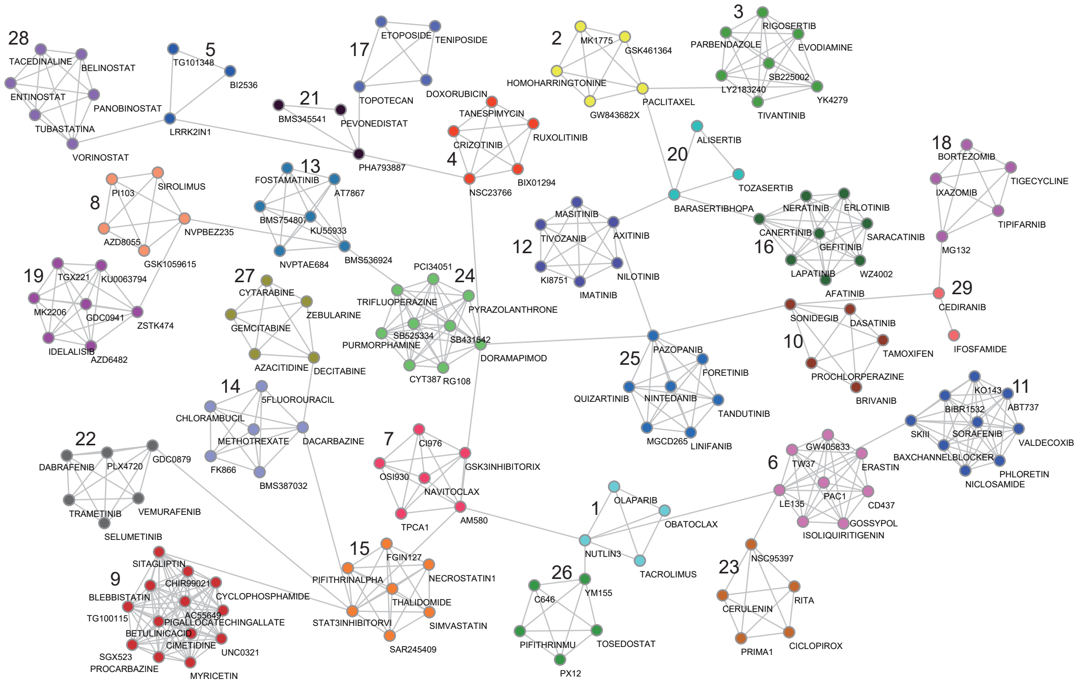

Fig. 4. **Network visualization of functionally-enriched drug communities identified by Mania.** Mania identified 29 drug communities by applying an affinity propagation clustering algorithm. One drug of each community was selected as the exemplar drug for that community. Inter-community edges were represented by exemplar-exemplar edges, which were obtained by building a minimal spanning tree among exemplar drugs.

of drugs, Mania produced various drug communities in which drugs with the same or similar functions were clustered together. For instance, Mania correctly identified the inhibitors for several targets, including the inhibitors of BRD4 (Cluster 5), PI3K/mTOR (Cluster 8), IGF-1R (Cluster 13), EGFR/ERBB (Cluster 16), TOP2 (Cluster 17), PSMB1 (Cluster 18), BRAD/MEK (Cluster 22), TGFBR1 (Cluster 24), and HDAC (Cluster 28). Among these, TOP2 (Topoisomerase II) has held great interest of researchers because of the discovery of active anti-cancer drugs that target TOP2,[27] and Mania identified four drugs (Teniposide, Etoposide, Topotecan, and Doxorubicin) targeting TOP2 (Cluster 17), which includes Etoposide and Doxorubicin, two clinically active agents.

We further conducted a Fisher's exact test between all the drugs grouped in a community and all drugs associated with a specific MoA or target, to assess whether the specific MoA or target is enriched in the community. We found that out of the 29 drug communities, 16 communities were significantly enriched ($P < 0.05$) for a direct target and 15 were significantly enriched for one MoA. For example, Cluster 28 were statistically enriched ($P < 10^{-5}$) for targets in the HDAC family and contained six known inhibitors for these targets, including Vorinostat and Belinostat, two FDA approved drugs. Another example is Cluster 16, where the 8 drugs were statistically enriched ($P < 10^{-7}$) for the ERBB and EGFR targets.

Taken together, the above results on clustering demonstrated the ability of Mania on illustrating the functional drug-drug relationships among drugs, by partitioning the drugs into disjoint function-related communities based on the low-dimensional vector representations.

3.4. *Predictions of drugs for significantly mutated genes*

We then proceeded with explosive analysis to test the ability of Mania for drug repurposing. To this end, we obtained a list of 224 significantly mutated cancer genes across 21 tumor types in a recent analysis of The Cancer Genome Atlas.[28] Mania first predicted targets for each drug through a weighted majority voting process, and the top 10 scored targets for each drug were recorded. Each significantly mutated gene was then searched against the list of top 10 predicted targets for each drug, and we found that 20 genes had been predicted by Mania to have new corresponding drugs that were not included in the Drug Repurposing Hub. Among these, EGFR, a significantly mutated gene in lung adenocarcinoma, was predicted by Mania as the top 1 gene that can be targeted by the drug Saracatinib but has not been predicted by DNF in its top 10 list. Our prediction of the use of Saracatinib to treat lung cancer through reducing the activation of EGFR is also supported by studies in the literature,[29] where Saracatinib was found to be able to efficiently reduce the activation of EGFR. Interestingly, Saracatinib was clustered into Cluster 16, which contains several well-known EGFR inhibitors, including two launched drugs, Afatinib and Erlotinib. This demonstrated the ability of Mania on transferring the knowledge of well-studied drugs to other similar but less characterized ones and providing additional potential insights into drug repurposing.

4. Discussion

We have presented Mania, a method for integrating heterogeneous pharmacogenomic data, which can be used to predict drug MoAs and targets, as well as to study the drug-drug relationships. Mania integrates multiple data sources and learns low-dimensional vector representations of drugs, which encode the structural and functional information for drugs. We have demonstrated Mania accurately quantifies the drug-drug similarity and substantially improves the performance of MoA/target prediction.Furthermore, Mania identifies functionally-enriched drug communities and new drugs that potentially target cancer mutated genes.

In the future, we plan to pursue further improvements of Mania. First, besides the Pearson correlation as the similarity measure for perturbation and sensitivity data, we plan to explore other approaches that can better capture the drug-drug relationships. In addition, we will test our method on non-redundant data (redundancy may arise, for example, when structures of some drugs were derived from others) and analyze the prediction ability of each data source. Furthermore, we plan to integrate more types of pharmacogenomic data (e.g., Cancer Cell Line Encyclopedia[30]) to provide a more complete view of the relationships among drugs.

Acknowledgements: This work was supported in part by the NSF CAREER Award, the Sloan Research Fellowship, and the PhRMA Foundation Award in Informatics.

References

1. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *Journal of computational chemistry* **30**, 2785 (2009).
2. K. Bleakley and Y. Yamanishi, *Bioinformatics* **25**, 2397 (2009).

3. M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, *Science* **321**, 263 (2008).
4. J. Li, X. Zhu and J. Y. Chen, *PLoS computational biology* **5**, p. e1000450 (2009).
5. F. Yang, J. Xu and J. Zeng, Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data, in *Pacific Symposium on Biocomputing*, 2014.
6. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics* **24**, i232 (2008).
7. F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi *et al.*, *Proceedings of the National Academy of Sciences* **107**, 14621 (2010).
8. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, *science* **313**, 1929 (2006).
9. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu *et al.*, *bioRxiv* , p. 136168 (2017).
10. B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould *et al.*, *Cancer discovery* **5**, 1210 (2015).
11. N. El-Hachem, D. M. Gendoo, L. S. Ghoraie, Z. Safikhani, P. Smirnov, C. Chung, K. Deng, A. Fang, E. Birkwood, C. Ho, R. Isserlin, G. D. Bader, A. Goldenberg and B. Haibe-Kains, *Cancer Research* **77**, 3057 (2017).
12. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, *Nature methods* **11**, 333 (2014).
13. P. Smirnov, Z. Safikhani, N. El-Hachem, D. Wang, A. She, C. Olsen, M. Freeman, H. Selby, D. M. Gendoo, P. Grossmann *et al.*, *Bioinformatics* **32**, 1244 (2015).
14. S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, *Nucleic acids research* **44**, D1202 (2015).
15. Rdkit: Open-source cheminformatics `http://www.rdkit.org`.
16. D. Rogers and M. Hahn, *Journal of chemical information and modeling* **50**, 742 (2010).
17. L. R. Dice, *Ecology* **26**, 297 (1945).
18. M. G. Rees, B. Seashore-Ludlow, J. H. Cheah, D. J. Adams, E. V. Price, S. Gill, S. Javaid, M. E. Coletti, V. L. Jones, N. E. Bodycombe *et al.*, *Nature chemical biology* **12**, p. 109 (2016).
19. S. M. Corsello, J. A. Bittker, Z. Liu, J. Gould, P. McCarren, J. E. Hirschman, S. E. Johnston, A. Vrcic, B. Wong, M. Khan *et al.*, *Nature Medicine* **23**, 405 (2017).
20. H. Cho, B. Berger and J. Peng, *Cell systems* **3**, 540 (2016).
21. H. Cho, B. Berger and J. Peng, Diffusion component analysis: Unraveling functional topology in biological networks. *RECOMB* 2015.
22. S. Wang, H. Cho, C. Zhai, B. Berger and J. Peng, *Bioinformatics* **31**, i357 (2015).
23. Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen and J. Zeng, *Nature Communications* **8** (2017).
24. H. Tong, C. Faloutsos and J.-y. Pan, *ICDM* , 613 (2006).
25. M. Žitnik and B. Zupan, *IEEE TPAMI* **37**, 41 (2015).
26. U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon and E. M. Marcotte, *PloS one* **8**, p. e58977 (2013).
27. J. L. Nitiss, *Nature reviews. Cancer* **9**, p. 338 (2009).
28. M. S. Lawrence, P. Stojanov, C. H. Mermel, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander and G. Getz, *Nature* **505**, p. 495 (2014).
29. L. Formisano, V. D'Amato, A. Servetto, S. Brillante, L. Raimondo, C. Di Mauro, R. Marciano, R. C. Orsini, S. Cosconati, A. Randazzo *et al.*, *Oncotarget* **6**, p. 26090 (2015).
30. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin *et al.*, *Nature* **483**, 603 (2012).