bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license.

NeoDTI: Neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions

Fangping Wan[1], Lixiang Hong[1], An Xiao[1], Tao Jiang[2,3,4,] and Jianyang Zeng[1,*]

Abstract

Accurately predicting drug-target interactions (DTIs) *in silico* can guide the drug discovery process and thus facilitate drug development. Computational approaches for DTI prediction that adopt the systems biology perspective generally exploit the rationale that the properties of drugs and targets can be characterized by their functional roles in biological networks. Inspired by recent advance of information passing and aggregation techniques that generalize the convolution neural networks (CNNs) to mine large-scale graph data and greatly improve the performance of many network-related prediction tasks, we develop a new nonlinear end-to-end learning model, called NeoDTI, that integrates diverse information from heterogeneous network data and automatically learns topology-preserving representations of drugs and targets to facilitate DTI prediction. The substantial prediction performance improvement over other state-of-the-art DTI prediction methods as well as several novel predicted DTIs with evidence supports from previous studies have demonstrated the superior predictive power of NeoDTI. In addition, NeoDTI is robust against a wide range of choices of hyperparameters and is ready to integrate more drug and target related information (e.g., compound-protein binding affinity data). All these results suggest that NeoDTI can offer a powerful and robust tool for drug development and drug repositioning.

1 Introduction

Identifying drug-target interactions (DTIs) through computational approaches can greatly narrow down the large search space of drug candidates for downstream experimental validation, and thus significantly reduce the high cost and the long period of developing a new drug [1]. Currently, the structure based [2], ligand-similarity based [3] and machine learning based methods [4, 5] are three main classes of prediction approaches in computational aided drug design. The structure based methods generally require the three-dimensional (3D) structures of proteins and have limited performance for those proteins with unknown structures, which unfortunately is the

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[3] MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST, Tsinghua University, Beijing, China.
[4] Department of Computer Science and Engineering, University of California, Riverside, CA, USA.
[*] To whom correspondence should be addressed. Email: zengjy321@tsinghua.edu.cn.

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

case for a majority of targets. The ligand-similarity based methods exploit the common knowledge of known interacting ligands to make prediction. Such approaches cannot lead to confident prediction results if the compound of interest is not indicated in the library of reference ligands. Recently, the machine learning based methods [6, 7, 5], which fully exploit the latent correlations among the related features of drugs and targets have become a highly promising strategy for DTI prediction. For instance, , the DTI network data have been integrated with the drug structure and protein sequence information into a network-based machine learning model (e.g., a regularized least square framework) for predicting new DTIs [8, 9, 10]. Inspired by the recent surge of deep learning techniques, models with higher predictive capacity have also been developed in various drug discovery settings (e.g., compound-protein interaction prediction, drug discovery with one-shot learning) [11, 12, 13, 14, 15, 16].

In addition to known DTI data, chemical structure and protein sequence information, the properties of drugs and targets can also be characterized by their various functional roles in biological systems (e.g., protein-protein interactions and drug-disease associations). Indeed, by integrating diverse information from heterogeneous data sources, methods like DTINet [5], MSCMF [17], HNM [18] can further improve the accuracy of DTI prediction. However, these methods still suffer from some limitations that need to be addressed. For example, in MSCMF [17], the employed matrix factorization operation of a given drug-target interaction network is regularized by the corresponding drug and protein similarity matrices, which are obtained by integrating multiple data sources through a weighted averaging scheme. Under such a data integration strategy, substantial loss of information may occur and thus result in a sub-optimal solution. DTINet [5] first uses an unsupervised manner to automatically learn the low-dimensional feature representations of drugs and targets from heterogeneous network data, and then applies inductive matrix completion [19] to predict new DTIs based on the learnt features. In such a framework, separating feature learning from the prediction task at hand may not yield the optimal solution, as the features learnt from the unsupervised learning procedure may not be the most suitable representations of drugs or targets for the final DTI prediction task. In addition, by constraining the learning models to only take relatively simple forms (e.g., bilinear or log-bilinear functions), these methods may not be sufficient enough to capture the complex hidden features behind the heterogeneous data. Recent advance of information passing and aggregation techniques that generalize the conventional convolution neural networks (CNNs) to large-scale graph data have shown substantial performance improvement on the network-related prediction tasks [20, 21]. This inspires us to incorporates deeper learning models to extract complex information from a highly heterogeneous network and discover new DTIs.

In this paper, we propose a new framework, called NeoDTI (**NE**ural integration of neighb**O**r information for **DTI** prediction) to predict new drug-target interactions from heterogeneous data. NeoDTI integrates neighborhood information of the heterogeneous network constructed from diverse data sources via a number of information passing and aggregation operations, which are achieved through the nonlinear feature extraction by neural networks. After that, NeoDTI applies a network topology-preserving learning procedure to enforce the extracted feature representations of drugs and targets to match the observed networks as possible. Comprehensive tests on several challenging and realistic scenarios in DTI prediction have demonstrated that our end-to-end prediction model can significantly outperform several state-of-the-art baseline prediction methods. Moreover, several novel DTIs predicted by NeoDTI with evidence supports from previous studies in the literature further indicate the strong predictive power of NeoDTI. In addition, the robustness of NeoDTI and its extendability to integrate more heterogeneous data (e.g., compound-protein binding affinity data) have been examined through various tests. All these results suggest that NeoDTI can provide a powerful and useful tool in predicting unknown DTIs,

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

and thus advance the drug discovery and repositioning fields.

In summary, our work includes the following contributions: (1) A *first* machine learning framework to integrate neural networks with neighborhood information passing and aggregation to predict new DTIs from heterogeneous data; (2) A new end-to-end model to integrate diverse information from multiple data sources for DTI prediction, which requires no hand engineered features; (3) Extensive validation tests demonstrating the superior prediction performance, the robustness and the extendability to integrate more heterogeneous data of our method; (4) A finding that several novel DTIs predicted by our method can be well supported by known evidence from the literature.

2 Method

2.1 Problem formulation

NeoDTI predicts unknown drug-target interactions (DTIs) from a drug and target related heterogeneous network, in which drugs, targets and other objects are represented as nodes, and DTIs and other interactions or associations are represented as edges. We first introduce the definition of a heterogeneous network.

DEFINITION 1. ***(Heterogeneous network)***. *A* ***heterogeneous network (HN)*** *is defined as a directed (or undirected) graph $G = (V, E)$, in which each node v in the node set V belongs to an object type from an object type set O, and each edge e in the edge set $E \subset V \times V \times R$ belongs to a relation type from a relation type set R.*

The datasets used in our framework to construct the HN (also see Section 3.1) include the object type set $O = \{drug, target, disease, side\text{-}effect\}$, and the relation type set $R = \{drug\text{-}drug\text{-}interaction, drug\text{-}protein\text{-}interaction, drug\text{-}disease\text{-}association, drug\text{-}structure\text{-}similarity, drug\text{-}side\text{-}effect\text{-}association, protein\text{-}protein\text{-}interaction, protein\text{-}drug\text{-}interaction, protein\text{-}sequence\text{-}similarity, protein\text{-}disease\text{-}association, disease\text{-}drug\text{-}association, disease\text{-}protein\text{-}association, side\text{-}effect\text{-}drug\text{-}association\}$. In our current framework, each node only belongs to a single object type although it can be relatively easily extended to a multi-object-type mapping scenario. In addition, all edges are undirected and non-negatively weighted. Also, the same two nodes can be linked by more than one edge, e.g., two drugs can be linked by a *drug-drug-interaction* edge and a *drug-structure-similarity* edge simultaneously.

Given an HN G, NeoDTI aims to automatically learn a network topology-preserving node-level embedding (i.e., a function that maps nodes to their corresponding feature representations that preserve the original topological characteristics as possible) from G that can be used to greatly facilitate the prediction of drug-target interactions. Most existing techniques for learning the embeddings of structured data mainly exploit the rationale that the elements of these structured data can be well characterized by their contextual information. For example, in natural language processing, the Word2vec technique [22] enforces the embedding of words to preserve the semantic relationships with their corresponding surrounding words. The graph embedding techniques, such as Deepwalk [23] and metapath2vec [24], have extended this embedding strategy to further learn the latent representations of network data. Recent advance in generalizing convolutional neural networks (CNNs) to analyze large-scale graph data [25, 26] and the integration of the information passing and aggregation techniques with different graph convolution operations into a unified framework [21, 20] have brought significant performance improvement for many network-related prediction tasks, such as predicting the biological activities of small molecules, graph signal processing and social network data analysis. Similar to GraphSAGE [20] and Mes-

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license.

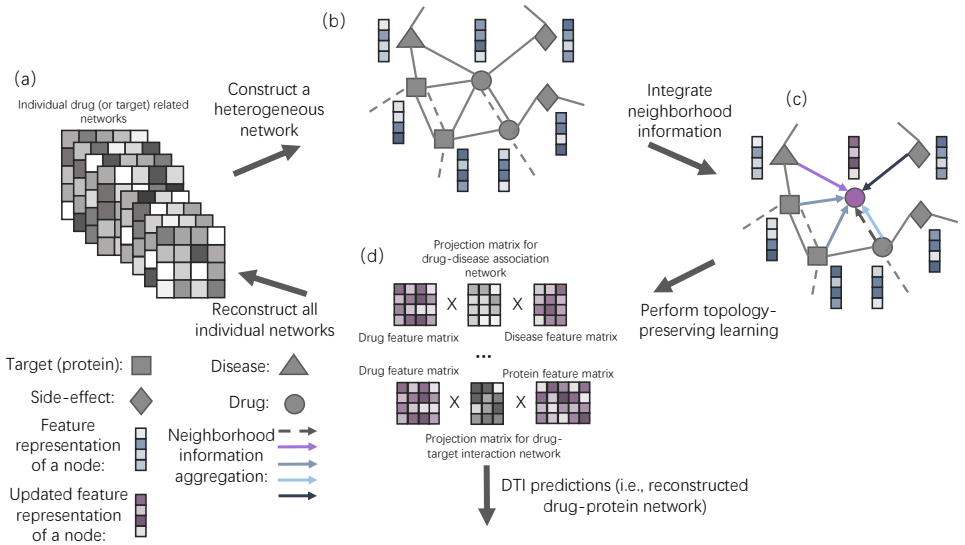

Figure 1: The schematic workflow of NeoDTI. (a) NeoDTI uses eight individual drug or target related networks (see Section 3.1 for more details of the used datasets). (b) NeoDTI first constructs a heterogeneous network from these eight networks. Different types of nodes are connected by distinct types of edges. Two nodes can be connected by more than one edge (e.g., the solid link between them representing drug-drug-interaction and the dashed link between two drugs representing drug-structure-similarity). In addition, NeoDTI associates each node with a feature representation. (c) To extract information from neighborhood, each node adopts a neighborhood information aggregation operation (see Definition 2 in the main text). Each colored arrow represents a specific aggregation function with respect to a specific edge type. Then each node updates its feature representation by integrating its current representation with the aggregated information (see Definition 3 in the main text). (d) By enforcing the node features to reconstruct the original individual networks as possible (see Definition 4 in the main text), NeoDTI effectively learns the topology-preserving node features that are useful for drug-target interaction prediction.

sage Passing Neural Networks (MPNNs) [21], our framework NeoDTI also applies neural networks to integrate neighborhood information from individual nodes. However, unlike GraphSAGE which mainly focuses on learning a node-level embedding from a homogeneous network or MPNNs which aim at learning a graph-level embedding from heterogeneous graphs for predicting molecular properties, NeoDTI focuses on learning a node-level embedding from a heterogeneous network. In addition, to the best of our knowledge, NeoDTI is the *first* framework to apply the neural information passing and aggregation techniques to extract the latent features of drugs and targets from a heterogeneous network to make DTI prediction.

2.2 The worflow of NeoDTI

NeoDTI consists of the following three main steps: (1) neighborhood information aggregation; (2) updating the node embedding; (3) topology-preserving learning of the node embedding. Through Steps (1) and (2), each node in the given HN generates a new feature representation by integrating its neighborhood information with its own features. Through Step (3), we enforce the embedding of nodes to be topology-preserving, which is useful for extracting the topological features of individual nodes for accurate DTI prediction. Next, we will introduce the mathematical formulations of these three steps.

Definition 2 (***Neighborhood information aggregation***). *Given an HN G, an initial node embedding*

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

function $f^0 : V \to \mathbb{R}^d$ that maps each node $v \in V$ to its d-dimensional vector representation $f^0(v)$ and an edge weight mapping function $s : E \to \mathbb{R}$ that maps each edge $e \in E$ to a non-negative real value $s(e)$, ***neighborhood information aggregation*** *for node v is defined as:*

$$\boldsymbol{a}_v = \sum_{r \in R} \underbrace{\sum_{\substack{u \in N_r(v), \\ e=(u,v,r) \in E}} \frac{s(e)}{M_{v,r}} \sigma(\boldsymbol{W}_r f^0(u) + \boldsymbol{b}_r)}_{\text{neighborhood information aggregation with respect to edge type } r}, \tag{1}$$

where $N_r(v) = \{u,\ u \in V,\ u \neq v,\ (u,v,r) \in E\}$ denotes the set of adjacent nodes connected to $v \in V$ through edges of type $r \in R$, $\sigma(\cdot)$ stands for a nonlinear activation function over a single-layer neural network parameterized by weights $\boldsymbol{W}_r \in \mathbb{R}^{d \times d}$ and a bias term $\boldsymbol{b}_r \in \mathbb{R}^d$, and $M_{v,r} = \sum_{u \in N_r(v),\ e=(u,v,r)} s(e)$ stands for a normalization term.

More specifically, for each edge type r, the neighborhood information aggregation operation for node v with respect to r can be obtained by first nonlinearly transforming the embedded feature representations of the corresponding adjacent nodes $f^0(u)$, $u \in N_r(v)$ through an edge-type specific single-layer neural network that is parameterized by weights $\boldsymbol{W}_r \in \mathbb{R}^{d \times d}$, a bias term $\boldsymbol{b}_r \in \mathbb{R}^d$ and a nonlinear activation function $\sigma(\cdot)$, and then averaging by the normalized edge weights, i.e., $\frac{s(e)}{M_{v,r}}$. Finally, the output of the neighborhood information aggregation operation $\boldsymbol{a}_v$ for node v is the summation of neighborhood information aggregation with respect to every edge type r. Here, the initial node embedding $f^0(u), \forall u \in V$ is obtained through a random mapping.

Definition 3 ***(Updating the node embedding)***. *Given the aggregated neighbor information $\boldsymbol{a}_v$'s for all nodes v's, the process of* ***updating the node embedding*** *is defined as:*

$$f^1(v) = \frac{\sigma(\boldsymbol{W}^1 concat(f^0(v), \boldsymbol{a}_v) + \boldsymbol{b}^1)}{||\sigma(\boldsymbol{W}^1 concat(f^0(v), \boldsymbol{a}_v) + \boldsymbol{b}^1)||_2}. \tag{2}$$

The above equation states that the new embedding of node $f^1(v)$ can be obtained using a single-layer neural network that is parameterized by weights $\boldsymbol{W}^1 \in \mathbb{R}^{(2d) \times d}$, a bias term $\boldsymbol{b}^1 \in \mathbb{R}^d$ and a nonlinear activation function $\sigma(\cdot)$ to nonlinearly transform the concatenation of the original embedding $f^0(v)$ and the neighborhood aggregation information $\boldsymbol{a}_v$, and then normalized by its l_2 norm.

Noted that in principle we could repeat the previous two steps alternately several times to produce more embeddings of nodes (e.g., $f^2(\cdot), f^3(\cdot), ...$). In practice, we find that we only need to conduct such a process once to obtain reasonably good prediction results, according to our validation tests (as described in Supplementary Information). In the rest part of this section, we will mainly use $f^1(\cdot)$ to demonstrate our algorithm for convenience. In addition, we choose to use $ReLU(x) = max(0, x)$ as the activation function $\sigma(\cdot)$.

Definition 4 ***(Topology-preserving learning of the node embedding)***. *Given the embedding of nodes $f^1(\cdot)$,* ***topology-preserving learning of the node embedding*** *is defined as:*

$$\min_{\{f^0(u),\ \boldsymbol{W}^1,\ \boldsymbol{b}^1,\ \boldsymbol{W}_r,\ \boldsymbol{b}_r,\ \boldsymbol{G}_r,\ \boldsymbol{H}_r,\ |u \in V,\ r \in R\}} \sum_{r \in R} \sum_{\substack{u,\ v \in V, \\ e=(u,v,r) \in E}} [s(e) - f^1(u)^\top \boldsymbol{G}_r \boldsymbol{H}_r^\top f^1(v)]^2, \tag{3}$$

where $\boldsymbol{G}_r,\ \boldsymbol{H}_r \in \mathbb{R}^{d \times k}$ are edge-type specific projection matrices.

The above equation states that, after edge-specific projections of $f^1(u)$ and $f^1(v)$ by $\boldsymbol{G}_r$ and $\boldsymbol{H}_r$, respectively, the inner product of the two projected vectors should reconstruct the original

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

edge weight $s(e)$ as possible. Note that a similar reconstruction strategy has also been used in [5, 19] to solve the link prediction problem. In addition, if the edge type r is symmetric, i.e., $r \in \{drug\text{-}structure\text{-}similarity,\ drug\text{-}drug\text{-}interaction,\ protein\text{-}sequence\text{-}similarity,\ protein\text{-}protein\text{-}interaction\}$, we use the tie weights (i.e., $\boldsymbol{G}_r = \boldsymbol{H}_r$) to enforce this symmetric property. Here, the summation of the squared reconstruction errors is minimized for all edges with respect to all unknown parameters. Since all mathematical operations in Equations 1, 2 and 3 are differentiable or subdifferentiable (e.g., for the ReLU activation function), all parameters can be trained through an end-to-end manner by performing gradient descent to minimize the final objective function described in Equation 3.

Finally, after Step (3), the predicted interaction confidence score between drug node u and protein node v can be obtained by

$$\begin{aligned} & f^1(u)^\top \boldsymbol{G}_r \boldsymbol{H}_r^\top f^1(v), \\ \text{subject to} \quad & \phi(u) = drug, \\ & \phi(v) = protein, \\ & r = drug\text{-}protein\text{-}interaction, \end{aligned} \tag{4}$$

where $\phi(u)$ and $\phi(u)$ stand for the node types of u and v, respectively.

The above operation is equivalent to reconstructing the *drug-protein* edge weight between nodes u and v. By collecting $f^1(u)$'s for all drugs and $f^1(v)$'s for all targets, we can form a drug feature matrix $\boldsymbol{F}_{drug}$ and a target feature matrix $\boldsymbol{F}_{target}$. Then, the reconstructed drug-target interaction matrix can be written as:

$$\boldsymbol{W}_{DTI_reconstruct} = \boldsymbol{F}_d \boldsymbol{G} \boldsymbol{H}_r^\top \boldsymbol{F}_t^\top \tag{5}$$

In this sense, we can consider our DTI prediction task as a matrix factorization or completion problem. However, unlike the conventional matrix factorization approaches [17, 19], NeoDTI incorporates a deeper learning model to construct the feature matrices $\boldsymbol{F}_d$ and $\boldsymbol{F}_t$ by explicitly defining the construction processes of $\boldsymbol{F}_d$ and $\boldsymbol{F}_t$ through Steps (1) and (2). In addition, through these two steps, NeoDTI incorporates the prior knowledge of network topology into $\boldsymbol{F}_d$ and $\boldsymbol{F}_t$ and specifies the forms of these two matrices. As a result, NeoDTI prevents the DTI network as well as other networks from being factorized arbitrarily in Step (3), which we believe can serve as a useful regularizer and thus lead to performance improvement for DTI prediction (as also demonstrated in our cross-validation tests; see the Results section).

3 Results

3.1 Datasets

We adopted the datasets that were curated in our previous study [5], which included six individual drug/protein related networks: **drug-protein interaction** and **drug-drug interaction** networks (interactions were extracted from Drugbank Version 3.0 [27]), the **protein–protein interaction** network (interactions were extracted from the HPRD database Release 9 [28]), **drug-disease association** and **protein-disease association** networks (associations were extracted from the Comparative Toxicogenomics Database [29]) and the **drug-side-effect association** network (associations were extracted from the SIDER database Version 2 [30]). The basic statistics of these datasets can be found in Table S1 in Supplementary Information. We also incorporated drug chemical structure information as well as protein sequence information by creating two extra networks: the

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

drug structure similarity network (i.e., a pair-wise chemical structure similarity network measured by the dice similarities of the Morgan fingerprints with radius 2 [31], which were computed by RDKit [32]) and the **protein sequence similarity** network (which was obtained based on the pair-wise Smith-Waterman scores [33]). All networks had binary edge weights (one represents a known interaction or association, and zero otherwise) except the drug structure similarity network and the protein sequence similarity network, which had non-negative real-valued edge weights. We combined all these eight networks to construct the heterogeneous network (Figure 1) for evaluating the prediction performance of NeoDTI.

3.2 NeoDTI yields superior performance in predicting new drug-target interactions

The DTI prediction can be considered as a binary classification problem, in which the known interacting drug-target pairs are regarded as positive examples, while the unknown interacting pairs are treated as negative examples. Several challenging and realistic scenarios were considered in our tests to evaluate the prediction performance of NeoDTI. The hyperparameters of NeoDTI were determined using an independent validation set (as described in Supplementary Information). We first ran a ten-fold cross validation test on all positive pairs and a set of randomly sampled negative pairs, whose number was ten times as many as that of positive samples. This scenario basically mimicked the practical situation in which the drug-target interactions are sparsely labeled. For each fold, a randomly chosen subset of 90% positive and negative pairs was used as training data to construct the heterogeneous network and then train the parameters of NeoDTI (i.e., during the topology-preserving learning process, we only calculated the reconstruction loss of the DTI network with respect to training data, while the reconstruction losses of other types of networks were computed as usual), and the remaining 10% positive and negative pairs were held out as the test set. We also compared the performance of NeoDTI with that of three baseline methods, including DTINet [5], HNM [18] and NetLapRLS [8]. The area under precision recall (AUPR) curve and the area under receiver operating characteristic (AUROC) curve were used to evaluate the prediction performances of all prediction methods. We observed that NeoDTI greatly outperformed other state-of-the-art baseline methods, with significant improvement (8.0% in terms of AUPR and 3.0% in terms of AUROC) over the second best method (Figures 2a and S1b). Next, we further increased the positive-negative ratio by including all negative examples (i.e., all unknown drug-target interacting pairs) in the ten-fold cross-validation procedure (the ratio between positive and negative samples was around 1.8×10^{-3}). We observed a larger AUPR improvement (29.4%) over the second best method (Figure 2b). Although the improvement of AUROC in this scenario was negligible (Figure S1c), as also stated in previous work [34], here, AUPR generally provides a more informative criterion than AUROC for the highly skewed datasets. Since drug discovery is generally a needle-in-a-haystack problem, the substantial improvement in AUPR truely demonstrated the superior prediction performance of NeoDTI over other methods.

Since the dataset may contain "redundant" DTIs (i.e., a same protein is connected to more than one similar drugs and vice versa), the prediction performance can be easily inflated by easy predictions in this case [5]. To consider this issue, we followed the same evaluation strategies as in [5] by conducting the following additional ten-fold cross validation tests: (1) removing DTIs with similar drugs (i.e., drug chemical structure similarities > 0.6) or similar proteins (i.e., protein sequence similarities $> 40\%$); (2) removing DTIs with drugs sharing similar drug interactions (i.e., Jaccard similarities > 0.6); (3) removing DTIs with drugs sharing similar side-effects (i.e., Jaccard similarities > 0.6); (4) removing DTIs with drugs or proteins sharing similar diseases (i.e., Jaccard

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license.

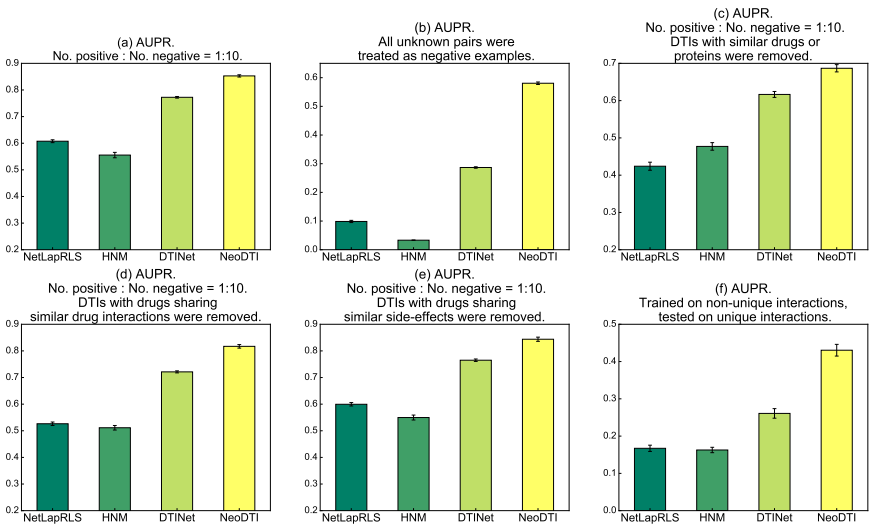

Figure 2: Performance evaluation of NeoDTI on several challenging scenarios in terms of the AUPR scores. (a) A ten-fold cross-validation test in which the ratio between positive and negative samples was 1 : 10. (b) A ten-fold cross-validation test in which all unknown drug-target interacting pairs were considered. (c-e) Ten-fold cross-validation with positive : negative ratio = 1 : 10 on several redundancy removed datasets: (c) DTIs with similar drugs and proteins were removed; (d) DTIs with drugs sharing similar drug interactions were removed; (e) DTIs with drugs sharing similar side-effects were removed. (f) NeoDTI was trained on non-unique drug-target interaction pairs and tested on unique drug-target interaction pairs. All results were summarized over 10 trials and expressed as mean $\pm$ standard deviation.

similarities > 0.6). In all these test scenarios, we kept the ratio between positive and negative samples to be 1 : 10. As expected, we observed a drop in prediction performance for all prediction methods after the removal of redundant DTIs (Figures 2c-e, S1a, S1d-g). However, NeoDTI still consistently outperformed other prediction methods in terms of both AUPR and AUORC, which also indicated the robustness of NeoDTI against the redundancy in data.

In dyadic prediction, if a dataset contains many drugs or targets with only one interaction, conventional cross-validation may not be a proper way to evaluate the prediction performance. Here, we call such drugs, proteins and interactions as "unique". In such a case, conventional training methods may lean to exploit the bias towards those unique drugs and targets to boost the performance [35]. To investigate this issue, we further evaluated the prediction performance of NeoDTI by separating unique DTIs from non-unique ones. That is, all methods were trained on non-unique DTIs and evaluated on unique DTIs. Note that in such a case, the negative examples in the test data were sampled by enforcing the corresponding drugs or targets (or both) to be unique. This scenario basically mimicked the situation in which the DTIs of new drugs or targets are predicted without much prior DTI knowledge. We found that NeoDTI significantly outperformed all the baseline methods at least by 17.0% in terms of AUPR and 6.7% in terms of AUROC, which suggested that NeoDTI can have a much better generalization capacity over other state-of-the-art methods, when predicting new DTIs for those drugs or targets with little prior DTI knowledge.

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

3.3 Robustness of NeoDTI

In this section, we further evaluated the robustness of NeoDTI by varying different types of data used in the heterogeneous network as well as the hyperparameters of NeoDTI. All computational experiments in this section were conducted using a ten-fold cross-validation procedure in which the ratio of positive versus negative samples was set to $1:10$.

To examine the effects of incorporating heterogeneous data, we first evaluated the performance of NeoDTI that was trained using only the drug-protein interaction network. We observed a substantial drop of prediction performance (11.1% in terms of AUPR and 9% in terms of AUROC), compared to that of the original NeoDTI model trained on all eight networks (Figure 3a). We then investigated the effects of incorporating individual networks by training NeoDTI again on a heterogeneous network constructed by each individual network and the drug-protein interaction network. As expected, we found that adding individual drug or target related networks can improve the prediction performance (Figures 3a, S2a-e). These results suggested that diverse information from multiple data sources can better characterize the latent properties of drugs and targets, and thus incorporating heterogeneous information is necessary to improve the accuracy of DTI prediction. In addition, to examine whether NeoDTI can also be easily extended to incorporate more drug or targeted related information beyond the previously used datasets, we further incorporated compound-protein binding affinity information into the heterogeneous network. Specifically, we collected all the binding affinity data between drug-like compounds and proteins that satisfied $Ki \leq 1nm$ from the ZINC15 database [36]. In total, we extracted 1,696 edges that connected $1,244$ compounds to the proteins used in our previous datasets. We set the negative logarithm of Ki (denoted by pKi) between a pair of compound and protein as the weight of their corresponding interaction. We also linked the compounds and drugs by drug-structure-similarity edges. In addition, if a compound and a drug were similar (i.e., chemical structure similarities > 0.6) and connected to the same protein in the test set, we removed this compound-protein pair from the training set to ease the inflation of prediction performance that may be resulted from the redundancy in data. We found that NeoDTI trained on this new heterogeneous network further improved AUPR from 85.3% to 86.3% and AUROC from 94.6% to 94.9% (Figure 3b), which demonstrated the easy extendability of NeoDTI to integrate more heterogeneous information.

In our topology-preserving learning of the node embedding, we enforce the feature representations of nodes to reconstruct all types of edges as possible. We further investigated the effect of this edge reconstruction strategy by conducting an additional test in which we constrained NeoDTI to only reconstruct the DTI edges. In this test, we observed a decrease in prediction performance with 5.5% in terms of AUPR and 2.7% in terms of AUROC (Figure S2f). Thus, reconstructing other types of network edges is useful for boosting the prediction performance. Such an operation probably serves as a beneficial regularizer to further overcome the potential overfitting problem.

In addition, we investigated the robustness of NeoDTI against different choices of hyperparameters: (1) For the dimension d of the node embedding, we tested $d = 256$, 512 and 1024; (2) For the dimension k of the projection matrices, we tested $k = 256$, 512 and 1024; (3) For the repetition time p of neighborhood information aggregation, we examined $p = 0$, 1, 2 and 3. We found that NeoDTI can produce relatively stable results over a wide range of choices for both d and k, although we observed that increasing the value of d can slightly improve the prediction results (Figure S3a-b). More importantly, we observed significant performance improvement when $p \geq 1$, demonstrating the necessity of integrating neighborhood information for the representation learning of node features (Figure S3c). However, we found that increasing the repetition time of neighborhood information aggregation from one to three did not improve the prediction

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license.

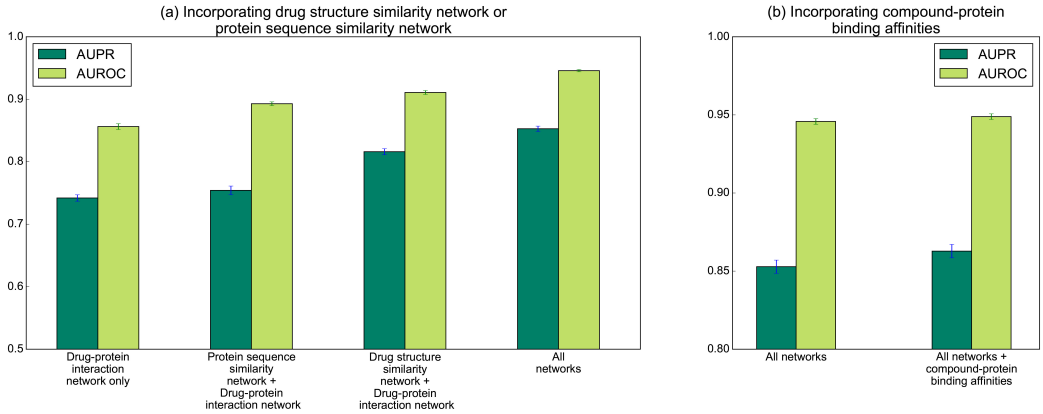

Figure 3: Incorporating more drug or target related information can improve the prediction performance of NeoDTI. (a) Incorporating drug structure similarity network or protein sequence similarity network. (b) Incorporating the compound-protein binding affinity data. All results were summarized over 10 trials and expressed as mean $\pm$ standard deviation.

performance (Figure S3c). Thus, in practice, we only need to run the operation of integrating neighborhood information once.

3.4 NeoDTI reveals novel DTIs with literature supports

We also predicted the novel DTIs by training NeoDTI using the whole heterogeneous network, including the aforementioned binding affinity data. We excluded those easy predictions by removing the predicted DTIs that were similar to the known DTIs (i.e., drug chemical structure similarities $>$ 0.6 and protein sequence similarities $>$ 40%). We then analyzed the predicted DTIs whose prediction confidence scores were significant (three-sigma rule) with respect to the corresponding drugs and targets. The network visualization of the top 100 novel drug-target interactions predicted by NeoDTI can be found in Figure 4.

Among the top twenty predicted DTIs ranked according to their confidence scores, eight DTIs can be supported by previous studies in the literature (Table S2). For instance, Sorafenib, a drug approved for the treatment of advanced renal cell carcinoma, was predicted by NeoDTI to interact with the colony stimulating factor 1 receptor (CSF1R), which plays an important role in the development of mammary gland and mammary gland carcinogenesis [37]. Such a prediction can be supported by a previous study showing that Sorafenib can block CSF1R and induce apoptosis in various classical Hodgkin lymphoma cell lines [38]. In addition, carbonic anhydrase 6 (CA6), an enzyme abundantly found in salivary glands, has been reported to be the target of three drugs, including zonisamide, ellagic acid and mafenide [27], was predicted by NeoDTI to also interact with acetazolamide. This prediction can be supported by the previous finding on the CA6 inhibitory activity of acetazolamide [39]. Overall, these novel DTIs predicted by NeoDTI with literature supports further demonstrated its strong predictive power.

4 Conclusion

In this paper, we develop a new framework, called NeoDTI, to integrate diverse information from a heterogeneous data to predict new drug-target interactions. NeoDTI extracts the complex hidden features of drugs and targets by applying neural networks to integrate neighborhood

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

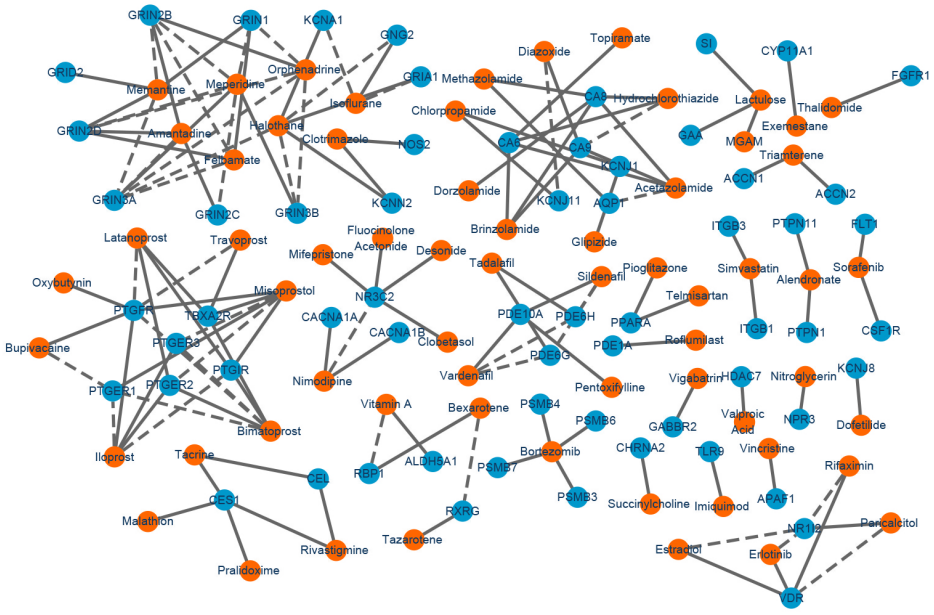

Figure 4: Network visualization of the top 100 novel drug-target interactions predicted by NeoDTI. Blue and orange nodes represent proteins and drugs, respectively. Dashed and solid lines represent the known and predicted drug-target interactions, respectively.

information in the input heterogeneous network. By simultaneously optimizeing the feature extraction process and the DTI prediction model through an end-to-end manner, NeoDTI eases the effort of designing handcrafted features. The effectiveness and robustness of NeoDTI have been extensively validated on several realistic prediction scenarios and supported by the finding that many of the novel predicted DTIs agree well with the previous studies in the literature. Moreover, NeoDTI can incorporate more drug and target related information readily (e.g., compound-protein binding affinity data). Therefore, we believe that NeoDTI can provide a powerful tool to facilitate the drug discovery or drug repositioning processes. In the future, we will further extend NeoDTI by integrating more heterogeneous information and validate some of the prediction results through wet-lab experiments.

5 Acknowledgement

The authors thank Yunan Luo for helpful discussions. This work was supported in part by the National Natural Science Foundation of China [61472205, 81630103], China's Youth 1000-Talent Program, Beijing Advanced Innovation Center for Structural Biology. We acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

References

[1] Gillian R Langley, Ian M Adcock, Francois Busquet, Kevin M Crofton, Elena Csernok, Christoph Giese, Tuula Heinonen, Kathrin Herrmann, Martin Hofmann-Apitius, Brigitte Landesmann, et al. Towards a 21st-century roadmap for biomedical research and drug discovery: consensus report and recommendations. *Drug discovery today*, 22(2):327–339, 2017.

[2] Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16):2785–2791, 2009.

[3] Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nature biotechnology*, 25(2):197–206, 2007.

bioRxiv preprint first posted online Feb. 7, 2018; doi: http://dx.doi.org/10.1101/261396. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[4] Qingjun Yuan, Junning Gao, Dongliang Wu, Shihua Zhang, Hiroshi Mamitsuka, and Shanfeng Zhu. Druge-rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics*, 32(12):i18–i27, 2016.

[5] Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8(1), 2017.

[6] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232, 2008.

[7] Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, 2009.

[8] Zheng Xia, Ling-Yun Wu, Xiaobo Zhou, and Stephen TC Wong. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In *BMC systems biology*, volume 4, page S6. BioMed Central Ltd, 2010.

[9] Twan van Laarhoven, Sander B Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21):3036–3043, 2011.

[10] Twan van Laarhoven and Elena Marchiori. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PloS one*, 8(6):e66952, 2013.

[11] Yuhao Wang and Jianyang Zeng. Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics*, 29(13):i126–i134, 2013.

[12] Fangping Wan and Jianyang Zeng. Deep learning with feature embedding for compound-protein interaction prediction. *bioRxiv*, page 086033, 2016.

[13] Masatoshi Hamanaka, Kei Taneishi, Hiroaki Iwata, Jun Ye, Jianguo Pei, Jinlong Hou, and Yasushi Okuno. Cgbvs-dnn: Prediction of compound-protein interactions based on deep learning. *Molecular informatics*, 36(1-2), 2017.

[14] Kai Tian, Mingyu Shao, Yang Wang, Jihong Guan, and Shuigeng Zhou. Boosting compound-protein interaction prediction by deep learning. *Methods*, 110:64–72, 2016.

[15] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

[16] Youjun Xu, Jianfeng Pei, and Luhua Lai. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *Journal of chemical information and modeling*, 57(11):2672–2685, 2017.

[17] Xiaodong Zheng, Hao Ding, Hiroshi Mamitsuka, and Shanfeng Zhu. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1025–1033. ACM, 2013.

[18] Wenhui Wang, Sen Yang, Xiang Zhang, and Jing Li. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, 30(20):2923–2930, 2014.

[19] Nagarajan Natarajan and Inderjit S Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

[20] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1025–1035. Curran Associates, Inc., 2017.

[21] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[23] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[24] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 135–144. ACM, 2017.

[25] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016.

[26] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[27] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, et al. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl_1):D1035–D1041, 2010.

[28] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl_1):D767–D772, 2008.

[29] Allan Peter Davis, Cynthia Grondin Murphy, Robin Johnson, Jean M Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Michael C Rosenstein, Thomas C Wiegers, et al. The comparative toxicogenomics database: update 2013. *Nucleic acids research*, 41(D1):D1104–D1114, 2012.

[30] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1):343, 2010.

[31] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[32] Greg Landrum. Rdkit: Open-source cheminformatics. *Online). http://www. rdkit. org. Accessed*, 3(04):2012, 2006.

[33] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[34] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[35] Twan van Laarhoven and Elena Marchiori. Biases of drug–target interaction network data. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 23–33. Springer, 2014.

[36] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *J. Chem. Inf. Model*, 55(11):2324–2337, 2015.

[37] Rulla M Tamimi, Joan S Brugge, Matthew L Freedman, Alexander Miron, J Dirk Iglehart, Graham A Colditz, and Susan E Hankinson. Circulating colony stimulating factor-1 and breast cancer risk. *Cancer research*, 68(1):18–21, 2008.

[38] Katrin Ullrich, Kathrin D Wurster, Björn Lamprecht, Karl Köchert, Andreas Engert, Bernd Dörken, Martin Janz, and Stephan Mathas. Bay 43-9006/sorafenib blocks csf1r activity and induces apoptosis in various classical hodgkin lymphoma cell lines. *British journal of haematology*, 155(3):398–402, 2011.

[39] Isao Nishimori, Tomoko Minakuchi, Saburo Onishi, Daniela Vullo, Andrea Scozzafava, and Claudiu T Supuran. Carbonic anhydrase inhibitors. dna cloning, characterization, and inhibition studies of the human secretory isoform vi, a new target for sulfonamide and sulfamate inhibitors. *Journal of medicinal chemistry*, 50(2):381–388, 2007.