Predicting Drug-Target Interaction for New Drugs Using Enhanced Similarity Measures and Super-Target Clustering[1]

Jian-Yu Shi[†], Siu-Ming Yiu[‡], Yiming Li[§], Henry C. M. Leung[‡] and Francis Y. L. Chin[‡]*
[†]School of Life Sciences, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[‡]Department of Computer Science, The University of Hong Kong, Hong Kong
[§]Department of Psychiatry, The University of Hong Kong, Hong Kong
Email addresses: JYS: jianyushi@nwpu.edu.cn, SMY: smyiu@cs.hku.hk, YL: liym1018@hku.hk,
HCML: cmleung2@cs.hku.hk, FYLC: chin@cs.hku.hk
* To whom correspondence should be addressed.

***Abstract*—Predicting drug-target interaction using computational approaches is an important step in drug discovery and repositioning. To predict whether there will be an interaction between a drug and a target, most existing methods identify similar drugs and targets in the database. The prediction is then made based on the known interactions of these drugs and targets. This idea is promising. However, there are two shortcomings that have not yet been addressed appropriately. Firstly, most of the methods only use 2D chemical structures and protein sequences to measure the similarity of drugs and targets respectively. However, this information may not fully capture the characteristics determining whether a drug will interact with a target. Secondly, there are very few known interactions, i.e. many interactions are "missing" in the database. Existing approaches are biased towards known interactions and have no good solutions to handle possibly missing interactions which affect the accuracy of the prediction. In this paper, we enhance the similarity measures to include non-structural (and non-sequence-based) information and introduce the concept of a "super-target" to handle the problem of possibly missing interactions. Based on evaluations on real data, we show that our similarity measure is better than the existing measures and our approach is able to achieve higher accuracy than the two best existing algorithms, WNN-GIP and KBMF2K.**

I. Introduction

Drug targets (or simply targets) are proteins that are related to diseases. If a drug interacts with a target, that drug can possibly be used to treat the corresponding disease. The number of (approved) drugs ($< 7,000$) having known interactions with targets is extremely small compared to the number of all available chemical compounds (35 million) that could be potential drug candidates [1]. Testing these candidates against each possible target using laboratory experiments would require a huge amount of money and a very long time. In the past, there has been quite a number of successful cases where an approved drug was found to be useful to treat another disease that is not the original target the drug was designed for (drug repositioning [2]). Identifying new drug-target interactions (DTI) for either approved drugs or new drug candidates is a crucial step in drug discovery and repositioning.

To speed up the process, a possible direction is to predict new interactions for new (or approved) drugs based on known drug-target interactions using computational approaches before we conduct laboratory experiments. Existing computational approaches could be divided into two main categories – docking simulation and machine learning. Docking simulation usually requires the three-dimensional (3D) structures of targets (traditional docking) or a large set of drugs (inverse docking). This requirement is always difficult to meet due to the small size of known 3D structures or available drug set. Besides, the docking approach could be rather time-consuming. In contrast, machine learning is becoming more and more popular in DTI prediction because it is much more efficient when dealing with a large number of drug or target candidates.

We classify the DTI prediction problem into 4 types. A known (approved) drug is a drug with at least one known interaction with a target. A known target is a target known to interact with at least one drug. A potential drug candidate ("new drug") is one without any known interaction. Similarly, a "new target" is a potential target with no known interaction with any drugs. Given a set of known DTI interactions, we can predict (S1) *new* DTI between known drugs and known targets; (S2) DTI between new drugs and known targets; (S3) DTI between known drugs and new targets; and (S4) DTI between new drugs and new targets. It is obvious that S1 has the most information about the concerned drug and target pair, S2 and S3 have less information, whereas S4 has the least information. S2 and S3 are sort of symmetric (i.e., methods developed for S2 can usually be used to solve S3 and vice versa). In this paper, we mainly focus on new drugs, i.e. S2 and S4.

Existing algorithms predicting new drug-target interaction based on known drug-target interaction usually represent the known drug-target interaction as a bipartite graph [3] in which drugs and targets are nodes and the interactions are undirected edges between these nodes. Thus, predicting new drug-target interaction is equivalent to predicting new edges in the bipartite graph. Some algorithms (e.g. [4]) try to predict new edges based on the topology of this graph. To predict if a drug d interacts with a target t, a general idea is to consider the edges involving d and t, e.g. (1) the shortest path between d and t in the graph; (2) the number of length-2 paths connecting d and another drug d' interacting with t; and (3) the degree of t. However, these algorithms do not work well for S2, S3, and S4 since these problems involve either a new drug or a new target which is not connected to the rest of the graph. Thus, this approach could only handle S1 and may not be very useful

[1] This work was supported by Hong Kong Scholars Program (No. XJ2011028), China Postdoctoral Science Foundation (No. 2012M521803) and HK GRF – HKU (7111/12E), as well as partially supported by NSFC (11171086) and Shenzhen basic research project (NO.JCYJ20120618143038947).

for other cases.

Another popular approach for predicting DTI is to build a classifier based on known interactions (machine learning approach). For example, [5] considers each drug-target pair as one sample. Each drug-target pair is labeled as positive if and only if they are known to interact and can be represented by a feature vector. The features of a certain drug are based on its two-dimensional (2D) chemical structure (see Fig. 1 for an example) while those of a target are derived from its protein sequence. The classifier, e.g. support vector machines (SVMs), is built on the structural similarity of the drugs and the sequence similarity of the targets. By using kernel tricks, [6] follows a similar approach which allows SVMs to learn a nonlinear classifier effectively. Instead of building a general classifier for all targets, [7], [8] tries to build a separate classifier for each target based on the bipartite local model (BLM) approach. For each target, each drug is considered as one sample and the classifier is built based on the structural similarity of the drugs. However, the BLM approach and all the above-mentioned classifier based methods suffer the same problem of biased training data since there are a lot more negative samples than positive samples (known interactions). In fact, quite a number of those negative samples should be positive, just that we do not know that they interact ("missing interactions").

To solve this problem of biased training, Laarhoven et al. introduced a variant of the BLM, named Gaussian interaction profile (GIP) [9], by only using positive samples to build the classifiers. However, this variant has another bias. The targets interacting with more drugs are more likely to be predicted to interact with a new drug since the classifier is built on the sum of similarities between the new drug and the drugs interacting with the concerned target. Its extension by incorporating a weighted nearest neighbor (WNN) algorithm, WNN-GIP [10] tries to rectify this bias by only considering the nearest neighbors of the new drug (i.e., the drugs most similar to the new drug). Nevertheless, GIP and WNN-GIP could not handle S4 well because no positive samples are available for the new drugs and targets in S4.

Recently, matrix factorization techniques such as kernelized Bayesian matrix factorization (KBMF2K) [11], which were originally employed in recommendation systems [12], have been applied to DTI prediction. It tries to predict a set of common features to represent each drug and target such that the drug similarity, target similarity and known interactions can be described based on these features. Then it uses these common feature to predict the drug-target interactions. The problem of this approach is that the features have no explicit chemical or biological meaning. It is difficult to justify if the prediction is reasonable and could practically guide drug design. Again, since the factorization depends on the known interactions, the missing interactions will still affect the correctness of both D and T.

To summarize, two issues are not yet solved.

(1) *Missing interactions*: No existing approach is able to handle the huge number of missing interactions in the databases. For example, the drug D02356 has interactions with 121 targets, however, most of the drugs similar to D02356 only have a few interactions and these drugs also share no common targets. It is likely that there are missing interactions among these drugs and targets.

(2) *The similarity measure*: Many approaches rely on the similarity measures of drugs and targets. Existing measures only consider the chemical 2D structures of the drugs and the protein sequence of the targets. There are quite a few examples showing that these measures cannot reflect the true similarity with respect to the interaction with targets. One example is the drugs D00316 and D01132. They share many targets but their 2D chemical structural similarity is only 0.275 which ranks 10 out of the most similar 25 drugs to D01132 (Fig. 1 shows their dissimilar 2D structures). This implies that structural similarity alone cannot fully capture the characteristics of drugs affecting DTI. The same argument applies to the sequence similarity of targets.

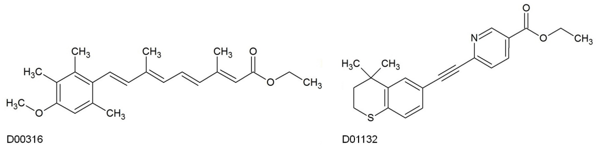


Fig. 1. The 2D chemical structures of the drugs D00316 and D01132, as extracted from KEGG [13].

Our contributions: To handle the large amount of missing interactions, we propose the concept of "super-target" to cluster similar targets. A drug, if known to interact with a target, is assumed to interact with the corresponding super-target group. For example, in S2, given a new drug and a known target, we consider the super-target group containing the known target. Our prediction is based on two levels of probabilities: how likely the new drug will interact with the members in the super-target group and how likely the new drug will interact with the concerned target. By combining the probabilities, we compute a confidence score[1]. The idea can be extended to handle S4 since even for a new target, we can still group it into a target group based on the target similarity (Section II-B).

In order to further reduce the bias on positive samples, we extend WNN-GIP's idea of using the top K nearest neighbors to negative samples as well. Given a new drug and a known target (S2), in addition to considering the top K nearest neighbors of the known drugs interacting with the concerned target, we also consider the top K nearest neighbors of the known drugs that *do not* have interaction with the concerned target. Thus, we have a more balanced set of positive and negative samples. Considering that under most cases, only a few drugs interact with the concerned target, we ran experiments with K ranging from 1 to 5, and finally selected 3 as K's value since beyond $K = 3$, both the area under the curve (AUC) and the area under precision-recall curve (AUPR) drop or do not increase much.

We also developed new similarity measures for drugs and targets. For drugs, in addition to 2D chemical structures, we incorporate a score based on the Anatomical Therapeutic Chemical (ATC) Classification System. The ATC Classification System divides drugs into different groups according to the organ or system on which they act as well as their chemical,

[1] We can use a similar concept of a "super-drug" to handle S3.

pharmacological and therapeutic properties [14]. Hence, the ATC codes provide non-structural information related to target interaction for drugs. For targets, besides the sequence similarity score, we also make use of their functional categories (FCs), which measure the target similarity according to the classified chemical reactions catalyzed by targets or the annotating functions of protein-coding genes.

Based on the real data, we show that our similarity measure is superior to the existing similarity measures, and with the concept of a "super-target", our approach is able to handle some of the missing interactions and achieve a higher accuracy than both WNN-GIP and KBMF2K, which are the most popular existing performing tools.

II. Method

A. *DTI prediction as probabilistic events*

The interactions between m known drugs and n known targets are represented by an interaction matrix $A_{m\times n}$, in which $A(i,j)=1$ when there is a known interaction between drug d_i and target t_j, and $A(i,j)=0$ otherwise. Given a new drug d_x (i.e. $d_x \notin$ the set of known drugs $\{d_i\}$), we aim to predict the interaction between drug d_x and target t_j (DTI prediction). In addition to the interaction matrix $A_{m\times n}$, both the pairwise similarities between drugs and the pairwise similarities between targets would be used for DTI prediction (Section II-C).

Inspired by the multi-label K nearest neighbors algorithm (ML-KNN) [15], we treat the DTI prediction between d_x and t_j as probabilistic events. Consider a new drug d_x with 3 similar drugs that interacts with a target t_j. Assume among all known drugs with exactly 3 similar drugs interacting with t_j, 90% of them also interact with t_j. Then d_x would have a 90% probability to interact with t_j. The calculation of this probability depends on the drug similarities and known drug-target interactions (Section II-B).

Let $p^{x,j}$ be the probability that d_x interacts with t_j, i.e. $A(x,j)=1$. When $d_x \in \{d_i\}$, if it interacts with t_j, $p^{x,j}=1$; otherwise, $p^{x,j}=0$. When $d_x \notin \{d_i\}$, $p^{x,j}\in[0,1]$ is the confidence score of the potential interaction between d_x and t_j. Based on the pairwise drug similarities, the K *neighbors* of a drug are defined to be the top K most similar drugs to it. Let $N(x,K)$ be the set of K neighbors of d_x, and $n(x,j,K)=\sum_{d_i\in N(x,K)} A(i,j)=c$ be the number of neighbors that interact with t_j. We attempt to calculate $p^{x,j}$ for a new drug $d_x \notin \{d_i\}$ based on the observed $n(x,j,K)$ as in (1).

$$\frac{\Pr\left[a(x,j)=1\right]\cdot\Pr\left[n(x,j,K)=c|a(x,j)=1\right]}{\sum_{b=0,1}\Pr\left[a(x,j)=b\right]\cdot\Pr\left[n(x,j,K)=c|a(x,j)=b\right]} \tag{1}$$

where $a(x,j)$ is the binary indication of whether d_x interacts with t_j. We adopt the same way as that of ML-KNN [15] to estimate the two probabilistic components $\Pr\left[a(x,j)=b\right]$ and $\Pr\left[n(x,j,K)=c|a(x,j)=b\right]$ in (1) from the m known drugs. $\Pr\left[a(x,j)=b\right]$ is the prior probability which could be estimated from $A_{m\times n}$ by (2).

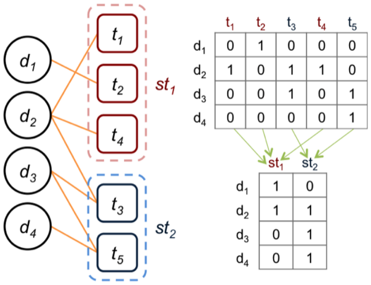
$$\begin{aligned}\Pr\left[a(x,j)=1\right] &\approx \left[1+\sum\nolimits_{i=1}^{m} A(i,j)\right]/(m+2);\\ \Pr\left[a(x,j)=0\right] &= 1-\Pr\left[a(x,j)=1\right]\end{aligned} \tag{2}$$

$\Pr\left[n(x,j,K)=c|a(x,j)=b\right]$ can be estimated similarly, for each known drug d_i, its K neighbors $N(i,K)$ are first obtained via drug similarities and the number $n(i,j,K)$ of these drugs interacting with each known target t_j can be counted. After collecting a set of $n(i,j,K)$ for different known d_i's and t_j's, the value of $\Pr\left[n(x,j,K)=c|a(x,j)=b\right]$ can be estimated by (3).

$$\frac{1+\sum_i Ind[A(i,j)=b\,\&\,n(i,j,K)=c]}{(K+1)+\sum_{c'=0}^{K}\sum_i Ind[A(i,j)=b\,\&\,n(i,j,K)=c']} \tag{3}$$

where $Ind[P]$ is the binary indication of whether statement P is correct or not. Notably, the 1's in the numerators of formulas (2) and (3) guarantee non-zero probabilities for the interaction between a new drug and a new target (S4). Finally, for each value of b, a table is built for the drugs interacting with ($b=1$) and not interacting with ($b=0$) target t_j respectively, which contains $K+1$ above probability entries corresponding to the $K+1$ possible values of $c=0,1,\ldots,K$.

B. *Super-targets for DTI prediction*



	t_1	t_2	t_3	t_4	t_5
d_1	0	1	0	0	0
d_2	1	0	1	1	0
d_3	0	0	1	0	1
d_4	0	0	0	0	1

	st_1	st_2
d_1	1	0
d_2	1	1
d_3	0	1
d_4	0	1

Fig. 2. Illustration of the concept of a super-target.

When considering the missing but potential interaction between t_j and d_x in $A_{m\times n}$ (e.g. similar drugs share no common target), we first cluster all the targets which are similar to t_j into one group. If d_x interacts with this group of targets, d_x would very likely interact with t_j as well. Let the n targets be $T=\{t_1,\ldots,t_n\}$ which can be partitioned into p groups $\{st_1,\ldots,st_p\}$ by a clustering algorithm according to the pairwise similarities between targets such that each group $st_q (q=1,\ldots,p)$ contains non-overlapped targets and $\bigcup_{q=1}^{p} st_q = T$. Each group st_q is defined as a *super-target*. All drugs interacting with at least one target in the super-target form the set of drugs interacting with the super-target. Thus, the profiles of targets belonging to the same group represented in $A_{m\times n}$ are combined into the profile of a super-target as represented in a new collapsed matrix $SA_{m\times p}$ which indicates the interactions between known drugs and super-targets. Fig. 2. illustrates the definition of a super-target.

We propose a two-stage approach to calculate the confidence score of the potential interaction between a new drug d_x and a known target t_j. Using the same procedure in Section II-A, we construct a new estimation of $\Pr[a(x, j) = b]$ and $\Pr[n(x, j, K) = c|a(x, j) = b]$ for each super-target by replacing $A_{m\times n}$ and target t_j's with $SA_{m\times p}$ and super-targets st_q's respectively. Given a new drug d_x and a known target t_j, we first calculate the confidence score s_1 between d_x and st_q where $t_j \in st_q$. Then we calculate the confidence score s_2 between d_x and t_j within super-target st_q. The final confidence score between d_x and t_j is defined to be the product of s_1 and s_2, which will be directly used to calculate the AUC and AUPR when assessing the performance of our proposed method.

So far, we have focused on S2. S4 can also be handled in a similar manner. Although there is no known interaction between a new drug and a new target, if the new target can be grouped with other targets (according to the similarity measure) to form a super-target which interacts with some known drugs that are similar to the new drug, we may predict there is an interaction between the new drug and new target.

C. Similarity used in this paper

The drug similarities and target similarities widely used in former publications are chemical-structure-based and sequence-based respectively [8], [9], [10], [11], [16]. The pairwise similarity between drugs is measured by aligning their chemical structures [17], whereas the Smith-Waterman alignment score [18] between protein sequences is used to measure that between the targets.

Since drugs with different structures can interact with the same target and proteins with different sequences may have similar 3D structures corresponding to similar functions, we propose to incorporate also ATC-based semantic similarity for drugs and FC-based semantic similarity for targets. Both ATC codes and FC codes are hierarchically semantic schemes represented by a sequence of symbols and numbers, e.g. the ATC code A10BA02 and the FC code 3.4.11.4, organized from high to low level. We calculate the semantic similarity between two drugs according to their ATC codes ($S^d_{ATC}(d_u, d_v)$) and that between targets according to their FC codes ($S^t_{FC}(t_u, t_v)$) by counting the common sub-codes. Consider a ATC code or FC code represented by a vector consisting of N entries where each entry denotes the sub-code in its corresponding level of hierarchical scheme. If the first f entries of the two vectors are the same, the semantic similarity between the two drugs (targets) is $\frac{f}{N}$. As we expect that drugs interacting with the same target are similar, we only add the new semantic similarities if combining them can make the drugs more similar, otherwise we do not. Besides, when calculating the sequence similarity score between two nuclear receptors (NRs) which contain a highly conserved DNA-binding domain (DBD) and a ligand-binding domain (LBD) [19], we calculate the sequence similarity by performing Smith-Waterman alignment on the sub-sequences in the LBDs of the proteins rather than their whole sequences. The sub-sequences in the LBDs can be extracted by the annotated boundary in SMART [20].

III. Results

In this section, we first illustrate the effectiveness of the new similarity measures for drugs and targets. Then we shall compare our approach with two state-of-the-art approaches, KBMF2K [11] and WNN-GIP [10], for predicting potential interactions between new drugs and known targets (S2) using four DTI datasets involving targets belonging to categories (1) enzyme, (2) ion channel (IC), (3) G protein-coupled receptor (GPCR) and (4) nuclear receptor (NR) respectively, which were benchmark datasets for comparing the performance of DTI prediction algorithms originally provided in [7]. These datasets were also used by KBMF2K [11] and WNN-GIP [10]. Last, we shall exhibit the performance of our approach in handling the missing interactions and S4.

A. Assessment

To evaluate the performance of the prediction algorithms, based on all known DTI, we adopted the same datasets, the same procedure and the same assessment as those used by both KBMF2K [11] and WNN-GIP [10]. Note that the targets are classified into four types: enzymes, IC, GPCR, and NR. We adopted 5-fold cross validation (5-CV) as the testing strategy in which we split randomly drugs in each dataset into 5 subsets of roughly equal sizes, and used each subset as the testing set and the remaining 4 subsets as the training set in turn. We then repeated the whole procedure 5 times and evaluated the performance of methods on four datasets. Finally, we adopted Receiver Operator Characteristic (ROC) curve and Precision-Recall (PR) curve to assess the performance by calculating the areas under them (AUC and AUPR). Since AUPR punishes incorrect top ranking predictions more than AUC [16], [21], we paid more attention to AUPR in our experiments. The detailed information of AUC and AUPR can be found in [21].

B. The effectiveness of new similarity

Recall that a good similarity measure is crucial in interaction prediction since similar drugs tend to interact with similar targets. We validated the effectiveness of our new measures by comparing their performance for new drugs with that using the old measures. Utilizing the new similarity measures, we improved the performance for all datasets (Table I) in terms of AUC and AUPR in cross-validation. This provides further evidence that simply using the drug 2D chemical structures and the target protein sequences may not be good enough to capture their characteristics for prediction. Take the drug D00066 as an example which interacts with a protein HSA:2099. Consider the top-five similar drugs to D00066 interacting with HSA:2099. If we use 2D structural similarity to rank the top-five similar drugs, they are at ranks 7, 9, 10, 14 and 20 out of the 54 drugs (AUC = 0.938 and AUPR = 0.333). If we integrate ATC-based semantic and 2D structural similarity measures, they will be at ranks 4, 5, 7, 9 and 10 (AUC = 0.979 and AUPR = 0.792). Thus, using the ATC-based semantic similarity measures can improve the performance of predicting DTI. Since KBMF2K (AUC = 0.625 and AUPR = 0.093) and WNN-GIP (AUC = 0.438 and AUPR = 0.066) consider 2D structural similarity only, their performances on drug D00066 are not good.

Similarly, there is also a need to combine other information derived from non-sequence based properties (e.g. FC) of the target in order to assess target similarity more accurately. In particular, if the sequences are remotely homologous, sequence similarity may not be a good similarity measure.

TABLE I. COMPARING THE PERFORMANCES OF NEW AND OLD SIMILARITY MEASURES

	Enzyme		IC		GPCR		NR	
	AUC|std	AUPR|std	AUC|std	AUPR|std	AUC|std	AUPR|std	AUC|std	AUPR|std
Old	0.805|0.005	0.332|0.009	0.776|0.009	0.296|0.027	0.854|0.008	0.304|0.017	0.860|0.016	0.476|0.028
New	**0.812**|0.010	**0.385**|0.011	**0.811**|0.012	**0.367**|0.018	**0.875**|0.002	**0.414**|0.021	**0.871**|0.018	**0.533**|0.043

C. Comparison with other methods for new drugs

We compared our method with two recent approaches, KBMF2K [11] and WNN-GIP [10] (Table II). From the results, we can see that our method achieved better results in terms of both AUC and AUPR. Though the performance of our approach on Enzyme dataset shows a $\sim 5\%$ decrease in the AUC value, it has a $\sim 10\%$ increase in the AUPR value. We emphasis more on AUPR since we want to penalize highly-ranked false positive predictions more when the number of pairs without known interaction greatly exceeds the number of pairs with known interaction. Regarding the running time, our method has the lowest time complexity of $O(kmn)$, where k is the number of nearest neighbors, whereas the complexity of KBMF2K is $O(rm^3 + rn^3 + r^3)$ where r is the dimension of subspace in the method, and the complexity of WNN-GIP is $O(m^3 + n^3)$ (we exclude the time for computing the similarity matrices for drugs and targets). Therefore, our approach outperforms both KBMF2K and WNN-GIP in accuracy and time complexity (see the actual running times in Table II which show that our approach is over 20 times faster in practice than the other two approaches).

The drug D00548 is an example of explaining why we out-performed other approaches in addition of using a better similarity measurement. D00548 connects with 15 similar targets of which 9 connect with other drugs dissimilar to D00548. After wrapping those targets into a super-target, we found that the top ranked drugs (such as D00711) similar to D00548 connect with the super-target and the known interaction between D00711 and the proteins in the super-target can be predicted with a high confidence level (AUC = 1.000 and AUPR = 1.000) which is better than our approach without considering super-targets (AUC = 0.978 and AUPR = 0.743). For the other two approaches KBMF2K (AUC = 0.714 and AUPR = 0.115) and WNN-GIP (AUC = 0.832 and AUPR = 0.176) which do not introduce the notion of super-targets, the performances are even worse.

D. Possible missing interactions

Missing interactions usually occur when a known drug interacts with many targets but the drugs similar to it only interact with a few targets and share with no or very few common targets. The other case is for S4 (new drugs and new targets) since we have little knowledge about both the drugs and the targets. In this section, we investigate the performance of our approach in two scenarios.

Defining a drug's *degree* as the number of targets interacting with it, we firstly investigated the algorithms' performances with high-degree drugs. We selected those drugs with $\frac{1}{3}$ maximum drug degree assuming that drug degrees follow the power law distribution [22]. The prediction results confirm that our approach outperforms the other two approaches with high-degree drugs (Table III).

For S4, since the drug and target are new, there should not be any known interaction between the new drug (target) and known targets (drugs) and it is likely that there is only one interaction between a new drug and a new target which is unknown to us. Thus, the performance of a DTI prediction algorithm cannot be appropriately evaluated using AUC and AUPR as there is only one possible point on the curve. To evaluate the performance of solving S4, we focus on the top-ranked (i.e. most confidently) predicted interaction for each new drug. We apply the one-error [23] metric to evaluate the performance of the DTI prediction algorithms. For each drug, we determine whether it has known interaction between its top-ranked predicted target. If there is no known interaction between the drug and its top-ranked predicted target, the prediction is considered wrong. The one-error measures the fraction of drugs with a wrong prediction – note that when there is only one target interacting with the drug (the usual case in S4), one-error is identical to an ordinary classification error [15]. The smaller the value of one-error, the better the performance. The prediction results for the drugs under S4 are listed in Table IV. Since our approach have smaller one-errors than the others (except in the NR dataset where all approaches have the same performance), our approach outperforms other approaches in predicting interactions between new drugs and new targets.

IV. CONCLUSIONS

In this paper, we have addressed the following two important issues that were not solved in previous approaches. All the previous methods only compute the similarity of drugs based on the 2D chemical structures of the drugs and the similarity of targets based on the corresponding protein sequences. However, cases have shown that the 2D structures and sequence similarity may not fully capture the characteristics of the interaction between drugs and targets. To resolve this issue, we introduced a non-structural-based similarity metric for drugs and a functional-category-based similarity metric for targets, integrating them with structure-based similarity of drugs and sequence-based similarity of targets respectively.

Another critical issue is that many interactions are missing in the database, and none of the previous approaches has a good solution to handle missing interactions. In our approach, we have proposed the concept of a "super-target" to group similar individual targets together as well as drugs interacting with them so that the drugs interacting with the same super-target are as similar as possible when no or only a few similar drugs interact with the same individual target. This "super-target" idea is particularly useful in two scenarios in which missing interactions may play an important role in the prediction: (1) it could predict the interactions for a drug which potentially interacts with a lot of similar targets even if other drugs have few interactions with these targets; (2) it could predict the interactions between new drugs and new targets if

TABLE II. COMPARING THE PERFORMANCES OF KBMF2K, WNN-GIP AND OUR METHOD

	Enzyme		IC		GPCR		NR		Total
	AUC|std	AUPR|std	AUC|std	AUPR|std	AUC|std	AUPR|std	AUC|std	AUPR|std	running time
KBMF2K	0.812|0.004	0.287|0.021	0.802|0.006	0.245|0.023	0.840|0.009	0.347|0.028	0.810|0.025	0.354|0.063	115.4 min
WNN-GIP	**0.861**|**0.004**	0.280|0.014	0.775|0.006	0.233|0.024	0.872|0.008	0.311|0.021	0.839|0.023	0.456|0.065	190.9 min
Ours	0.812|0.010	**0.385**|0.011	**0.811**|0.012	**0.367**|0.018	**0.875**|0.002	**0.414**|0.021	**0.871**|0.018	**0.533**|0.043	5.5 min

TABLE III. PERFORMANCE OF OUR METHOD: HIGH-DEGREE DRUGS

	Enzyme		IC		GPCR		NR	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
KBMF2K	0.759	0.388	0.559	0.418	0.839	0.662	0.727	0.543
WNN-GIP	**0.859**	0.496	0.546	0.400	0.893	0.647	0.757	0.581
Ours	0.832	**0.531**	**0.724**	**0.631**	**0.933**	**0.737**	**0.878**	**0.691**

TABLE IV. ONE-ERROR OF OUR METHOD: NEW DRUGS, NEW TARGETS

	Enzyme	IC	GPCR	NR
KBMF2K	0.774	0.600	0.654	0.600
WNN-GIP	0.931	0.600	0.692	0.600
Ours	**0.657**	**0.500**	**0.500**	0.600

a new drug tends to interact with the super-target containing the new target and some other known targets.

Besides solving the above issues, our approach also has another advantage in further reducing the bias of positive samples (in all the datasets, due to the limitation of knowledge, we always have a lot more negative samples than positive samples). We have modeled the interactions between a new drug and a known target by two probabilistic tables, which are built via the same number of nearest neighbors for both positive and negative samples. Based on four real benchmark datasets, we have shown that integrating new similarities provides better measurements of drug similarities and target similarities, and our approach can handle some of the missing interactions with the "super-target" and performs superior to the other best performing tools in both S2 and S4 problems.

REFERENCES

[1] E. E. Bolton, Y. Wang, P. A. Thiessen *et al.*, "PubChem: integrated platform of small molecules and biological activities," *Annual Reports in Computational Chemistry*, vol. 4, pp. 217–241, 2008.

[2] M. Hurle, L. Yang, Q. Xie *et al.*, "Computational drug repositioning: from data to therapeutics," *Clin. Pharmacol. Ther.*, vol. 93, no. 4, pp. 335–341, 2013.

[3] M. A. Yıldırım, K.-I. Goh, M. E. Cusick *et al.*, "Drug-target network," *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1119–1126, 2007.

[4] F. Cheng, C. Liu, J. Jiang *et al.*, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput. Biol.*, vol. 8, no. 5, p. e1002503, 2012.

[5] D.-S. Cao, S. Liu, Q.-S. Xu *et al.*, "Large-scale prediction of drug–target interactions using protein sequences and drug topological structures," *Anal. Chim. Acta.*, vol. 752, pp. 1–10, 2012.

[6] L. Jacob and J.-P. Vert, "Protein-ligand interaction prediction: an improved chemogenomics approach," *Bioinformatics*, vol. 24, no. 19, pp. 2149–2156, 2008.

[7] Y. Yamanishi, M. Araki, A. Gutteridge *et al.*, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[8] J.-P. Mei, C.-K. Kwoh, P. Yang *et al.*, "Drug–target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2013.

[9] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[10] T. van Laarhoven and E. Marchiori, "Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile," *PloS One*, vol. 8, no. 6, p. e66952, 2013.

[11] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[12] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[13] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.

[14] "ATC structure and principles," http://www.whocc.no/atc/structure_and_principles/, accessed: 2014-06-30.

[15] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recogn.*, vol. 40, no. 7, pp. 2038–2048, 2007.

[16] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug–target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[17] M. Hattori, Y. Okuno, S. Goto *et al.*, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways," *J. Am. Chem. Soc.*, vol. 125, no. 39, pp. 11 853–11 865, 2003.

[18] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, 1981.

[19] M. Tsuji, "Local motifs involved in the canonical structure of the ligand-binding domain in the nuclear receptor superfamily," *J. Struct. Biol.*, vol. 185, no. 3, pp. 355–365, 2014.

[20] I. Letunic, T. Doerks, and P. Bork, "SMART 7: recent updates to the protein domain annotation resource," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D302–D305, 2012.

[21] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[22] F. J. Azuaje, L. Zhang, Y. Devaux *et al.*, "Drug-target network in myocardial infarction reveals multiple side effects of unrelated drugs," *Scientific Reports*, vol. 1, 2011.

[23] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2010, pp. 667–685.