Applied Intelligence
https://doi.org/10.1007/s10489-021-02495-z

Drug-target interaction prediction via an ensemble of weighted nearest neighbors with interaction recovery

Bin Liu[1] · Konstantinos Pliakos[2,3] · Celine Vens[2,3] · Grigorios Tsoumakas[1]

Accepted: 1 May 2021

Abstract

Predicting drug-target interactions (DTI) via reliable computational methods is an effective and efficient way to mitigate the enormous costs and time of the drug discovery process. Structure-based drug similarities and sequence-based target protein similarities are the commonly used information for DTI prediction. Among numerous computational methods, neighborhood-based chemogenomic approaches that leverage drug and target similarities to perform predictions directly are simple but promising ones. However, existing similarity-based methods need to be re-trained to predict interactions for any new drugs or targets and cannot directly perform predictions for both new drugs, new targets, and new drug-target pairs. Furthermore, a large amount of missing (undetected) interactions in current DTI datasets hinders most DTI prediction methods. To address these issues, we propose a new method denoted as Weighted k-Nearest Neighbor with Interaction Recovery (WkNNIR). Not only can WkNNIR estimate interactions of any new drugs and/or new targets without any need of re-training, but it can also recover missing interactions (false negatives). In addition, WkNNIR exploits local imbalance to promote the influence of more reliable similarities on the interaction recovery and prediction processes. We also propose a series of ensemble methods that employ diverse sampling strategies and could be coupled with WkNNIR as well as any other DTI prediction method to improve performance. Experimental results over five benchmark datasets demonstrate the effectiveness of our approaches in predicting drug-target interactions. Lastly, we confirm the practical prediction ability of proposed methods to discover reliable interactions that were not reported in the original benchmark datasets.

Keywords Drug-target interactions · Nearest neighbor · Interaction recovery · Local imbalance · Ensemble learning

1 Introduction

Prediction of drug-target interactions (DTIs) is fundamental to the drug discovery process [1, 2]. However, the identification of interactions between drugs and specific targets via wet-lab (*in vitro*) experiments is extremely costly, time-consuming, and challenging [3, 4]. Computational (*in silico*) methods can efficiently complement existing *in vitro* activity detection strategies, leading to the identification of interacting drug-target pairs and accelerating the drug discovery process.

In the past, computational approaches identified DTIs mainly based on known ligands for targets [5] or 3D protein structures [6]. However, these methods suffer from two limitations. Firstly, they would collapse when the required information (ligand-related information or 3D protein structures) is unavailable. Secondly, they are built on only one kind of information. Recently, chemogenomic approaches have attracted extensive interest, because they integrate both drug and target information into a unified framework [1] (e.g. chemical structure information related to drugs and genomic information related to target proteins). Such information is often obtained via publicly available databases.

✉ Bin Liu
binliu@csd.auth.gr

Konstantinos Pliakos
konstantinos.pliakos@kuleuven.be

Celine Vens
celine.vens@kuleuven.be

Grigorios Tsoumakas
greg@csd.auth.gr

[1] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

[2] Campus KULAK, Department of Public Health and Primary Care, KU Leuven, Oude Markt 13, 3000 Leuven, Belgium

[3] ITEC, IMEC Research Group at KU Leuven, Kortrijk, Belgium

Published online: 09 July 2021

Popular chemogenomic approaches rely on machine learning algorithms, which have been widely employed for DTI prediction tasks due to their verified effectiveness [7]. There are several machine learning strategies to handle DTI prediction [1]. Here, we focus on neighborhood approaches. These are typically straightforward and effective methods that are based on similarity functions [8]. Albeit rather simple, they are very promising. For instance, neighborhood information is often used to regularize matrix factorization tasks, leading to powerful DTI prediction methods [9]. Drug-drug similarities based on chemical structure and target-target similarities based on protein sequence are the most common types of information employed in DTI prediction tasks [1].

Many of the existing similarity-based methods follow the transductive learning scheme [9–14], where test pairs are presented and used during the training process. Every time new pairs arrive, the learning model has to be re-trained in order to perform predictions for them. This is computationally inefficient and substantially affects the scalability of such models, especially in cases where not all test pairs of drugs and targets can be gathered in advance. On the other hand, an inductive model is built upon a training dataset consisting of a drug set, a target set and their interactivity, and can predict any unseen drug-target pairs without re-training. Therefore, we employ similarity-based approaches following the inductive learning scheme, which is more flexible and effective to perform predictions for newly arrived test pairs.

Furthermore, many false negative drug-target pairs are typically included in DTI training sets. These drug-target pairs are actually interacting, but the interactions have not yet been reported (detected) due to the complex and costly experimental verification process [15–17]. Hereafter, these interactions shall be called *missing interactions*. When we treat the unverified DTIs as non-interacting, we inevitably lose valuable information and introduce noise to the data. Therefore, exploiting possible missing interactions is crucial for the DTI prediction approach [15, 18].

In DTI prediction, there are four main prediction settings. These include predictions for: unseen pairs of training (known) drugs and targets (S1), pairs of test (new) drugs and training targets (S2), pairs of test (new) targets and training drugs (S3), and pairs of test (new) drugs and test (new) targets (S4). Compared to S1, which only considers pairs consisting of known drugs and targets, the other three settings that focus on predicting interactions for new drugs and/or new targets are more difficult, because they relate to the cold-start problem where interacting information of new drugs (targets) is unavailable [18]. This paper focuses on these three (S2, S3, S4) more challenging settings.

S4 is substantially more arduous than S2 and S3, as the bipartite components of test pairs are new in S4. Many methods, especially most nearest neighbor based ones [11, 16, 19], either cannot be applied or show a major drop in prediction performance when it comes to S4. An existing neighborhood-based approach designed for S4 specifically [20], can not perform predictions in S2 and S3. Therefore, there is a lack of neighborhood-based DTI prediction methods that can successfully handle each and every one of S2, S3, and S4. Such methods are useful when the prediction setting of interest is unknown at training time.

We address DTI prediction with an emphasis on new drugs (S2), new targets (S3), and pairs of new drugs and new targets (S4). First, the formulation of the inductive DTI prediction task that aims to perform predictions for any unseen drug-target pair is defined and its differences with the transductive one are clarified. Next, we propose a neighborhood-based DTI prediction method called Weighted k-Nearest Neighbor with Interaction Recovery (WkNNIR). The proposed method can deal with all prediction settings, as well as effectively handle missing interactions. Specifically, WkNNIR detects neighbors of test drugs, targets, and drug-target pairs to estimate interactions in S2, S3, and S4, respectively. It updates the original interaction matrix based on the neighborhood information to mitigate the impact of missing interactions. In addition, WkNNIR exploits the concept of local class imbalance [21] to weigh drug and target similarities, which boosts interaction recovery and prediction.

Furthermore, we propose three ensemble methods to further improve the performance of WkNNIR and other DTI prediction methods. These methods follow a common framework that aggregates multiple DTI prediction models built upon diverse training sets deriving from the original training set by sampling drugs and targets. They employ three different sampling strategies, namely Ensemble with Random Sampling (ERS), Ensemble with Global imbalance based Sampling (EGS), and Ensemble with Local imbalance based Sampling (ELS). A short preliminary account of these three ensemble methods is given in [22].

The performance of the proposed methods is evaluated on five benchmark datasets. The obtained results show that WkNNIR outperforms other state-of-the-art DTI prediction methods. We also show that ELS, EGS and ERS are able to promote the performance of six different base models including WkNNIR in all prediction settings, and ELS is the most effective one. Last, we demonstrate cases where our methods succeed in discovering interactions that had not been reported in the original benchmark datasets. The latter highlights the potential of the proposed computational methods in finding new interactions in the real world.

The rest of this paper is organized as follows. In Section 2, existing DTI prediction approaches are briefly presented. In Section 3, we define the formulation of inductive DTI prediction and compare it with the transductive one. The proposed neighborhood-based and ensemble methods are presented in Sections 4 and 5, respectively. Experimental results and analysis are reported in Section 6. Finally, we conclude our paper in Section 7.

2 Related work

The fundamental assumption of similarity-based methods is that similar drugs tend to interact with similar targets and vice versa [23]. Similarity-based methods can be divided into four types according to the learning method they employ, namely Nearest Neighborhood (NN), Bipartite Local Model (BLM), Matrix Factorization (MF) and Network Diffusion (ND).

Nearest neighborhood based methods predict the interactions based on the information of neighbors. Nearest Profile [19] infers interactions for a test drug (target) from only its nearest neighbor in the training set. Weighted Profile (WP) [19] integrates all interactions of training drugs (targets) by weighted average to make predictions. Weighted Nearest Neighbor (WNN) [11] sorts all training drugs (targets) based on their similarities to the test drug (target) in descending order and assigns the weight of each training drug (target) according to its rank. In WNN, interactions of training drugs (targets) whose weights are larger than a predefined threshold are considered to make predictions. The Weighted k-Nearest Known Neighbor (WKNKN) [16] was initially proposed as a pre-processing step that transforms the binary interaction matrix into an interaction likelihood matrix, while it can estimate interactions for test drug (target) based on k nearest neighbors of training drugs (targets) as well. The Similarity Rank-based Predictor [24] predicts interactions for test drug (target) based on the likelihood of interaction and non-interaction obtained by similarity ranking. All the above NN methods are restricted to the S2 and S3 settings. In [20], three neighborhood-based methods, namely individual-to-individual, individual-to-group, and nearest-neighbor-zone, designed specifically for predicting interactions between test drug and test target (S4) are proposed. Furthermore, the well-known neighborhood-based multi-label learning method MLkNN [25] is employed for drug side effect prediction [26]. In [15], MLkNN with Super-target Clustering (MLkNNSC) takes advantage of super-targets that are constructed by clustering the interaction profiles of targets.

BLM methods build two independent local models for drugs and targets, respectively. They integrate the predictions of the two models to obtain the final scores of test interactions. The first BLM method is proposed in [27], where a support vector machine is employed as the base local classifier. Furthermore, regularized least squares (RLS) formalized as kernel ridge regression (RLS-avg) and RLS using Kronecker product kernel (RLS-kron) are two other representative BLM approaches that process drug and target similarities as kernels [28]. One weakness of the above BLM approaches is that they cannot train local models for unseen drugs or targets. To address this issue, BLM-NII [10], which is based on RLS-avg, introduces a step to infer interactions for test drugs (targets) based on training data. Analogously, GIP-WNN [11] extends RLS-kron by adding the WNN process to estimate the interactions for test drugs (targets). In Advanced Local Drug-Target Interaction Prediction (ALADIN) [17], the local model is a k-Nearest Neighbor Regressor with Error Correction (ECkNN) [29], which corrects the error caused by bad hubs. Such hubs are often located in the neighborhood of other instances but have different labels from those instances. Generally, BLM can predict interactions for test drug-target pairs via a two-step learning process [30, 31], i.e. the interactions between test drugs (targets) and training targets (drugs) are estimated first, then two models based on those predictions are built to estimate interactions between test drugs and test targets.

MF methods deal with DTI prediction by factorizing the interaction matrix into two low-rank matrices which correspond to latent features of drugs and targets, respectively. Kernelized Bayesian Matrix Factorization with Twin Kernels (KBMF2K) [32] and Probabilistic Matrix Factorization (PMF) [33] conduct matrix factorization based on probability theory. KBMF2K follows a Bayesian probabilistic formulation and PMF leverages the probabilistic linear model with Gaussian noise. Collaborative Matrix Factorization (CMF) [34] is applied to DTI prediction via adding low-rank decomposition regularization on similarity matrices to ensure that latent features of similar drugs (targets) are similar as well. In [12], soft weighting based self-paced learning is integrated into CMF to avoid bad local minima caused by the non-convex objective function and achieve better generalization. Weighted Graph Regularized Matrix Factorization (WGRMF) [16] performs graph regularization on latent features of drugs and targets to learn a manifold for label propagation. Neighborhood Regularized Logistic Matrix Factorization (NRLMF) [9] combines the typical matrix factorization with neighborhood regularization in a unified framework to model the probability of DTIs within the logistic function. Dual-Network Integrated Logistic Matrix Factorization (DNILMF) [35], is an extension of NRLMF that utilizes diffused drug and target kernels instead of similarity matrices for the logistic function. MF approaches usually comply with the transductive learning scheme.

Network-based inference (NBI) [36] applies network diffusion on the DTI bipartite network leveraging graph-based techniques to predict new DTIs. This approach uses only the interactions between drugs and targets. Domain Tuned-Hybrid (DT-Hybrid) [37] and Heterogeneous Graph Based Inference (HGBI) [38] extends NBI via incorporating drug-target interactions, drug similarities and target similarities to the diffusion of the heterogeneous network. Apart from network inference, random walk [39], probabilistic soft logic [40], and finding simple path [41] are three other approaches that can be applied to the heterogeneous DTI network to predict interactions. All of the ND methods are transductive, as the network diffusion step should be recomputed if new drugs or new targets are added in the heterogeneous network.

Apart from using similarities, there are methods that treat interaction prediction as a classification task, building binary or multi-label classifiers over input feature sets. In [42], traditional ensemble tree strategies, such as random forests (RF) [43] and extremely randomized trees (ERT) [44], are extended to the bi-clustering tree setting [45]. Another example is AGHEL [46], which is a heterogeneous ensemble approach integrating two different classifiers, namely RF and XGBoost [47]. Moreover, three tree-based multi-label classifiers which incorporate various label partition strategies to effectively capture the correlations among drugs and targets are proposed for DTI prediction in [48]. Furthermore, delivering low dimensional drug and target embeddings from a DTI network using graph embedding [14, 49] or path category based extraction techniques [13, 50, 51] has been shown very effective. However, these methods are utilized in a transductive setting.

Finally, we briefly present approaches dealing with the issue of missing interactions. The DTI prediction task with missing interactions can be treated as a Positive-Unlabeled (PU) learning problem [52], where the training data consists of positive and unlabeled instances and only a part of positive instances are labeled. Several methods mentioned above discover missing interactions by using matrix completion techniques [9, 34]. Others construct super-targets containing more interacting information [15], correct possible missing interactions [17], and recover interactions as a pre-processing step [16]. Similar to the idea of interaction recovery, Bi-Clustering Trees with output space Reconstruction (BICTR) [18] utilizes NRLMF to restore the training interactions on which an ensemble of Bi-clustering trees [42] is built. In [53, 54], traditional PU learning algorithms, such as Spy and Rocchio, are employed to extract reliable non-interacting drug-target pairs. Based on a more informative PU learning assumption that interactions are not missing at random, a probabilistic model without bias to labeled data is proposed in [55].

3 Inductive DTI prediction

In this section, the formulation of similarity based DTI prediction in an inductive learning scheme is presented. Then, the comparison between the inductive and transductive settings for the DTI prediction task is given.

3.1 Formulation

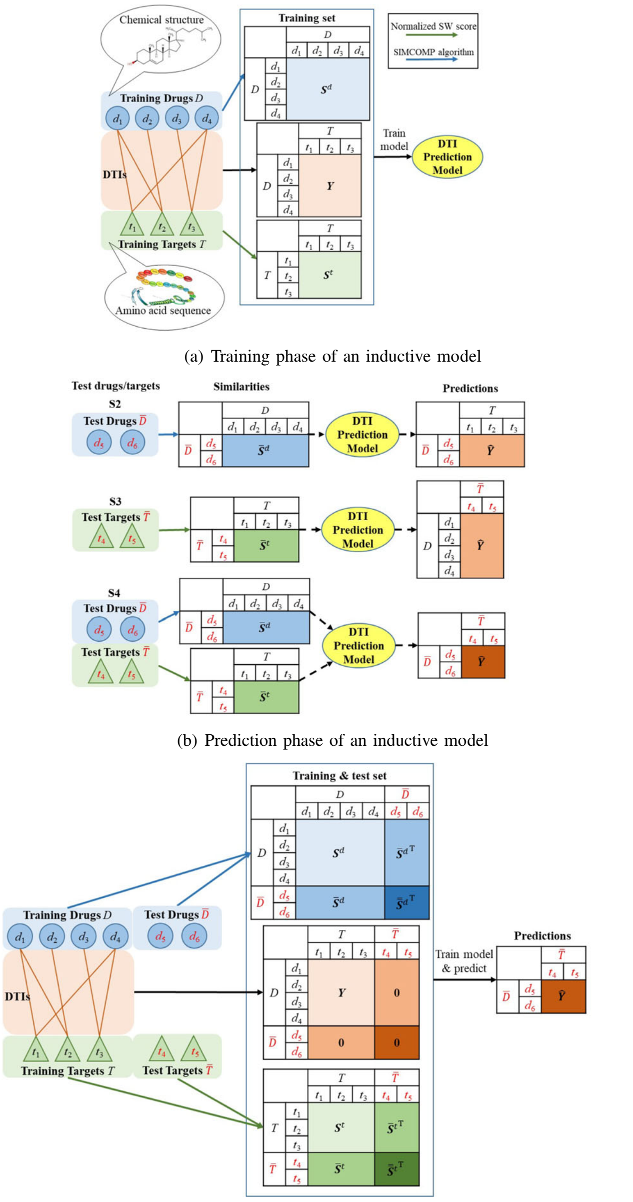
In this part, we define the inductive DTI prediction problem, as well as the training and estimation procedures of a DTI prediction model in inductive learning, where the drug and target similarities are utilized as input information.

Let $D = \{d_i\}_{i=1}^{n}$ be the training drug set containing n drugs, where each drug is a compound described by its chemical structure. Let $T = \{t_i\}_{i=1}^{m}$ be the training target set consisting of m targets, where each target is a protein represented by its amino acid sequence. There is a set of known interactions between drugs in D and targets in T. Figure 1(a) illustrates the process of training an inductive model. Initially, to cater for a similarity-based DTI prediction model, the drug similarity matrix $\boldsymbol{S}^d \in \mathbb{R}^{n\times n}$ and the target similarity matrix $\boldsymbol{S}^t \in \mathbb{R}^{m\times m}$ are computed, where S^d_{ij} is the similarity between d_i and d_j, and S^t_{ij} is the similarity between t_i and t_j. In this paper, drug similarities are computed by SIMCOMP algorithm [56], which assesses the common chemical structure of two drugs, and target similarities are calculated by using the normalized Smith-Waterman (SW) score [57], which evaluates the shared amino acid sub-sequence of two targets. In addition, DTIs are represented with an interaction matrix $\boldsymbol{Y} \in \{0, 1\}^{n\times m}$, where $Y_{ij} = 1$ if d_i and t_j are known to interact with each other and $Y_{ij} = 0$ indicates that d_i and t_j either actually interact with each other but their interaction is undetected, or d_i and t_j do not interact. An inductive DTI prediction model is built based on a training set consisting of D, T, $\boldsymbol{S}^d$, $\boldsymbol{S}^t$ and $\boldsymbol{Y}$.

In the prediction phase, as discussed in the introduction, we distinguish three settings of DTI prediction, according to whether the drug and target involved in the test pair are included in the training set or not. In particular:

- S2: predict the interactions between test drugs $\bar{D}$ and training targets T.
- S3: predict the interactions between training drugs D and test targets $\bar{T}$.
- S4: predict the interactions between test drugs $\bar{D}$ and test targets $\bar{T}$.

where $\bar{D} = \{d_u\}_{u=1}^{\bar{n}}$ is a set of test drugs disjoint from the training drug set (i.e. $\bar{D} \cap D = \emptyset$), and $\bar{T} = \{t_v\}_{v=1}^{\bar{m}}$ is a set of test targets disjoint from T.



(a) Training phase of an inductive model

(b) Prediction phase of an inductive model

(c) The training and prediction procedure of a transductive model dealing with S4

Fig. 1 Comparison between inductive and transductive settings

The prediction procedure is shown in Fig. 1(b). In all prediction settings, the similarities between test drugs (targets) and training drugs (targets), required by the similarity-based model, are firstly computed. Next, the learned DTI prediction model receives these similarities to perform predictions for the corresponding test drug-target pairs. In S2, given a set of test drugs $\bar{D}$, the similarities between $\bar{D}$ and D ($\bar{\boldsymbol{S}}^d \in \mathbb{R}^{\bar{n}\times n}$) computed by the SIMCOMP algorithm are input to the model, and the predictions are a real-valued matrix $\hat{\boldsymbol{Y}} \in \mathbb{R}^{\bar{n}\times m}$ indicating the confidence of the affinities between test drugs and training targets. Similarly in S3, the normalized SW score is employed to calculate the similarities between $\bar{T}$ and T ($\bar{\boldsymbol{S}}^t \in \mathbb{R}^{\bar{m}\times m}$), upon which the model outputs a prediction matrix $\hat{\boldsymbol{Y}} \in \mathbb{R}^{n\times \bar{m}}$. In S4, both $\bar{\boldsymbol{S}}^d$ and $\bar{\boldsymbol{S}}^t$ are computed, and the prediction matrix is $\hat{\boldsymbol{Y}} \in \mathbb{R}^{\bar{n}\times \bar{m}}$.

3.2 Comparison between inductive and transductive settings

In the transductive learning scheme, the model is learned for the purpose of predicting specific test pairs. Figure 1(c) describes the training and prediction processes of a model in the transductive setting dealing with S4. Both training and test drugs (targets) are available in the learning phase of a transductive model. Given that $\tilde{D} = D \cup \bar{D}$ and $\tilde{T} = T \cup \bar{T}$, one would first compute the extended drug and target similarity matrices for $\tilde{D}$ and $\tilde{T}$, respectively:

$$\tilde{\boldsymbol{S}}^d = \begin{bmatrix} \boldsymbol{S}^d & \bar{\boldsymbol{S}}^{d\top} \\ \bar{\boldsymbol{S}}^d & \bar{\bar{\boldsymbol{S}}}^d \end{bmatrix} \quad \tilde{\boldsymbol{S}}^t = \begin{bmatrix} \boldsymbol{S}^t & \bar{\boldsymbol{S}}^{t\top} \\ \bar{\boldsymbol{S}}^t & \bar{\bar{\boldsymbol{S}}}^t \end{bmatrix} \tag{1}$$

where $\bar{\bar{\boldsymbol{S}}}^d \in \mathbb{R}^{\bar{n}\times\bar{n}}$ ($\bar{\bar{\boldsymbol{S}}}^t \in \mathbb{R}^{\bar{m}\times\bar{m}}$) stores similarities between test drugs (targets). In addition, the interaction matrix is extended to $\tilde{\boldsymbol{Y}} \in \{0,1\}^{(n+\bar{n})\times(m+\bar{m})}$, where the rows and columns of $\tilde{\boldsymbol{Y}}$ corresponding to test drugs and targets contain “0”s. The transductive model is trained upon the input consisting of $\tilde{D}$, $\tilde{T}$, $\tilde{\boldsymbol{S}}^d$, $\tilde{\boldsymbol{S}}^t$ and $\tilde{\boldsymbol{Y}}$, and could perform predictions for the specific test pairs included in the input once it has been built. The processes of the transductive model handling S2 and S3 are similar with S4, except for that no test target and drug are used in S2 and S3, respectively (e.g in S3, $\bar{D} = \emptyset$, the drug similarity matrix is $\boldsymbol{S}^t$ and the interaction matrix is $[\boldsymbol{Y}, \boldsymbol{0}]$).

The difference between the inductive and transductive settings relies on their input information and the predicting ability of the model. On one hand, an inductive model is built in the training phase, using similarity and interaction matrices referring to training drugs and targets. After the completion of this process, it can perform predictions for any setting (S2, S3, S4) and any unseen pairs. On the other hand, a transductive model is built upon both the information of the training set and the similarities of the test drugs and/or targets. Hence, it can provide predictions only for these specific test pairs and cannot generalize to unseen test data. If a transductive model needs to predict the interaction of another unseen drug-target pair different to the one included in the training process, it should be re-trained, incorporating the information corresponding to the unseen test pair. This is extremely demanding, especially when it comes to large scale data. Therefore, in this paper, we focus on the inductive setting.

4 WKNNIR

In this section, we first propose the Weighted k-Nearest Neighbor (WkNN) as a comprehensive nearest neighbor based approach that could handle all prediction settings. Next, the local imbalance is introduced to measure the reliability of similarity matrices. Lastly, WkNNIR, which incorporates interaction recovery and local imbalance driven similarity weighing into WkNN, is presented.

4.1 WkNN

Neighborhood-based methods predict the interactions of test drugs (S2) or test targets (S3) by aggregating the interaction profiles of their neighbors [11, 16, 19]. The major limitation of these methods is that they cannot directly predict interactions between test drugs and test targets (S4), as interactions between test drugs (targets) and their neighbors are unavailable in the training set. To overcome this drawback of existing methods, we propose WkNN.

WkNN employs the prediction function of WKNKN [16] to deal with S2 and S3 because of its simplicity and efficacy. WkNN is a lazy learning method that does not have any specific training phase. Given a test drug-target pair (d_u, t_v) belonging to the prediction setting S2 or S3, WkNN predicts their interaction profiles based on the interactions of either k nearest training drugs of d_u or k nearest training targets of t_v as follows:

$$\hat{Y}_{uv} = \begin{cases} \frac{1}{z_{d_u}} \sum\limits_{d_i \in \mathcal{N}^k_{d_u}} \eta^{i'-1} \bar{S}^d_{ui} Y_{iv}, & \text{if } d_u \notin D \text{ and } t_v \in T \\ \frac{1}{z_{t_v}} \sum\limits_{t_j \in \mathcal{N}^k_{t_v}} \eta^{j'-1} \bar{S}^t_{vj} Y_{uj}, & \text{if } d_u \in D \text{ and } t_v \notin T \end{cases} \tag{2}$$

In (2), $\mathcal{N}^k_{d_u}$ $(\mathcal{N}^k_{t_v})$ corresponds to k nearest neighbors of d_u (t_v) which are retrieved by choosing k training drugs (targets) having k largest values in u-th row of $\bar{\boldsymbol{S}}^d$ (v-th row of $\bar{\boldsymbol{S}}^t$) and sorting the selected k drugs (targets) in the descending order according to their similarities to d_u (t_v). Moreover, d_i (t_j) is the i'-th (j'-th) nearest neighbor of d_u (t_v), i.e. i' (j') is the index of d_i (t_j) in $\mathcal{N}^k_{d_u}$ $(\mathcal{N}^k_{t_v})$, $\eta \in [0, 1]$ is the decay coefficient shrinking the weight of further neighbors, and $z_{d_u} = \sum_{d_i \in \mathcal{N}^k_{d_u}} \bar{S}^d_{ui}$ and $z_{t_v} = \sum_{t_j \in \mathcal{N}^k_{t_v}} S^t_{vj}$ are normalization terms.

To make predictions for pairs of test drugs and test targets (S4), WkNN follows the tetrad rule: when a drug interacts with a target, another similar drug probably interacts with another similar target [40]. For the sake of adapting to the tetrad rule, WkNN considers similar drug-target pairs instead of similar drugs or targets. We define the neighbors of a test drug-target pair (d_u, t_v) as $\mathcal{N}^k_{d_u t_v} = \{(d_i, t_j) | d_i \in \mathcal{N}^k_{d_u}, t_j \in \mathcal{N}^k_{t_v}\}$. Although the direct way to find neighbors of a drug-target pair is to search all nm drug-target combinations, our definition shrinks the search space to $n + m$ leading to increased efficiency. As in [11, 28], we define the similarity between two drug-target pairs as the product of the similarity of two drugs and the similarity of two targets, i.e. the similarity of (d_u, t_v) and (d_i, t_j) is $\bar{S}^d_{ui} \bar{S}^t_{vj}$. Thus, the prediction function of WkNN for S4 is:

$$\hat{Y}_{uv} = \frac{1}{z_{uv}} \sum_{(d_i, t_j) \in \mathcal{N}^k_{d_u t_v}} \eta^{i'+j'-2} \bar{S}^d_{ui} \bar{S}^t_{vj} Y_{ij} \tag{3}$$

where the normalization term z_{uv} is equal to $\sum_{(d_i, t_j) \in \mathcal{N}^k_{d_u t_v}} \bar{S}^d_{ui} \bar{S}^t_{vj}$. The weight in (3) consists of two parts: the first one corresponds to the decay coefficient, where the exponent is determined by the index of d_i in $\mathcal{N}^k_{d_u}$ (i') and the index of t_j in $\mathcal{N}^k_{t_v}$ (j') simultaneously, and the second one is the similarity between the test pair and its neighbor.

4.2 Local imbalance

The concept of local imbalance concerns the label distribution of an instance within the local region, playing a key role in determining the difficulty of a dataset to be learned [21]. Concerning DTI data that contain two kinds of similarities, the local imbalance can be assessed in both drug space and target space.

Firstly, we define the drug-based local imbalance. The local imbalance of a drug d_i for target t_j is measured as the proportion of $\mathcal{N}^k_{d_i}$ having the opposite interactivity to t_j as d_i:

$$C^d_{ij} = \frac{1}{k} \sum_{d_h \in \mathcal{N}^k_{d_i}} [\![Y_{hj} \neq Y_{ij}]\!] \tag{4}$$

where $C^d_{ij} \in [0, 1]$ and $[\![x]\!]$ is the indicator function that returns 1 if x is true and 0 otherwise. The larger the value of C^d_{ij}, the fewer the drugs in the local region of d_i having the same interactivity to t_j, and the higher the local imbalance of d_i for t_j based on drug similarities. In the ideal case, similar drugs share the same interaction profiles (i.e. interact with the same targets), rendering DTI prediction a very simple task. However, there are several cases, where similar drugs interact with different targets, which makes DTI prediction more challenging. Therefore, we employ local imbalance C^d_{ij} to assess the reliability of similarities between d_i and other training drugs. Specifically, lower C^d_{ij} indicates more reliable similarity information.

Likewise, we calculate the local imbalance of t_j for d_i based on the targets as:

$$C^t_{ij} = \frac{1}{k} \sum_{t_h \in \mathcal{N}^k_{t_j}} [\![Y_{ih} \neq Y_{ij}]\!] \tag{5}$$

4.3 WkNNIR

The existence of missing interactions (not yet reported) in the training set can lead to biased DTI prediction models and an inevitable accuracy drop. To address this issue, we propose WkNNIR, which couples WkNN with interaction recovery to perform predictions upon the completed interaction matrix. Moreover, in S4, where both drug and target similarities are used to perform predictions, WkNNIR has the advantage that the importance (weight) of drug and target similarities is differentiated depending on their local imbalance.

Firstly, WkNNIR computes recovered interactions, which replace the original interactions in the prediction phase. Based on the assumption that similar drugs interact with similar targets and vice versa, missing interactions can be completed by considering the interactions for neighbor drugs or targets. There are two ways to recover the interaction matrix: one on the drug side and another on the target side. The drug side recovery is conducted row-wise, where each row of $\boldsymbol{Y}$ is recovered by the weighted average of the neighbor rows identified by drug similarities. The drug-based recovery interaction matrix $\boldsymbol{Y}^d$ is:

$$\boldsymbol{Y}^d_{i\cdot} = \frac{1}{z_{d_i}} \sum_{d_h \in \mathcal{N}^k_{d_i}} \eta^{h'-1} S^d_{ih} \boldsymbol{Y}_{i\cdot},\ i = 1, \cdots, n \tag{6}$$

where h' is the index of d_h in $\mathcal{N}^k_{d_i}$, and $z_{d_i} = \sum\limits_{d_h \in \mathcal{N}^k_{d_i}} S^d_{ih}$.

The target-based recovery interaction matrix $\boldsymbol{Y}^t$ is obtained by reconstructing $\boldsymbol{Y}$ column-wise:

$$\boldsymbol{Y}^t_{\cdot j} = \frac{1}{z_{t_j}} \sum_{t_l \in \mathcal{N}^k_{t_j}} \eta^{l'-1} S^t_{lj} \boldsymbol{Y}_{\cdot j},\ j = 1, \cdots, m \tag{7}$$

where l' is the index of t_l in $\mathcal{N}^k_{t_j}$ and $z_{t_j} = \sum_{t_l \in \mathcal{N}^k_{t_j}} S^t_{lj}$.

However, $\boldsymbol{Y}^d$ ($\boldsymbol{Y}^t$) exploits only one kind of similarity and neglects the other one. To address this issue, we combine these two recovered interaction matrices into a complementary one that incorporates the recovery information from both drug and target views. Besides, interactions restored via more reliable similarity measures tend to be more credible. Therefore, instead of treating $\boldsymbol{Y}^d$ and $\boldsymbol{Y}^t$ equally, we distinguish the effectiveness of different recovered interaction matrices according to the local imbalance of the similarity used in the recovery process.

The interacting pair (d_i, t_j) with lower drug-based local imbalance indicates that d_i is close to other drugs interacting with t_j in the drug space, and therefore the recovered interactions inferred from the pair are more reliable in the drug view. Hence, we define the weight of recovered interaction Y^d_{ij} according to the average local imbalance of interacting pairs that are used to estimate Y^d_{ij}:

$$W^d_{ij} = \exp\left(- \frac{\sum_{d_h \in \mathcal{N}^k_{d_i}} C^d_{hj} Y_{hj}}{\sum_{d_h \in \mathcal{N}^k_{d_i}} Y_{hj}} \right) \tag{8}$$

Similarly, the weight of recovered interaction Y^t_{ij} is computed as:

$$W^t_{ij} = \exp\left(- \frac{\sum_{t_l \in \mathcal{N}^k_{t_j}} C^t_{lj} Y_{il}}{\sum_{t_l \in \mathcal{N}^k_{t_j}} Y_{il}} \right) \tag{9}$$

The higher the weight, the more reliable the recovered interaction. By weighted aggregation of the $\boldsymbol{Y}_d$ and $\boldsymbol{Y}_t$, we obtain the final recovered interaction matrix $\boldsymbol{Y}^{dt} \in \mathbb{R}^{n \times m}$:

$$Y^{dt}_{ij} = \frac{W^d_{ij} Y^d_{ij} + W^t_{ij} Y^t_{ij}}{W^d_{ij} + W^t_{ij}},\ i = 1, \cdots, n;\ j = 1, \cdots, m \tag{10}$$

By multiplying with the local imbalance-based weight, interactions recovered by more reliable similarities have more influence on $\boldsymbol{Y}^{dt}$. The values in $\boldsymbol{Y}^{dt}$ are in the range of [0,1]. Furthermore, because "1"s in $\boldsymbol{Y}$ denotes reliable interactions that do not need any update, a correction process is applied to the reconstructed interaction matrix to ensure the consistency of known interactions:

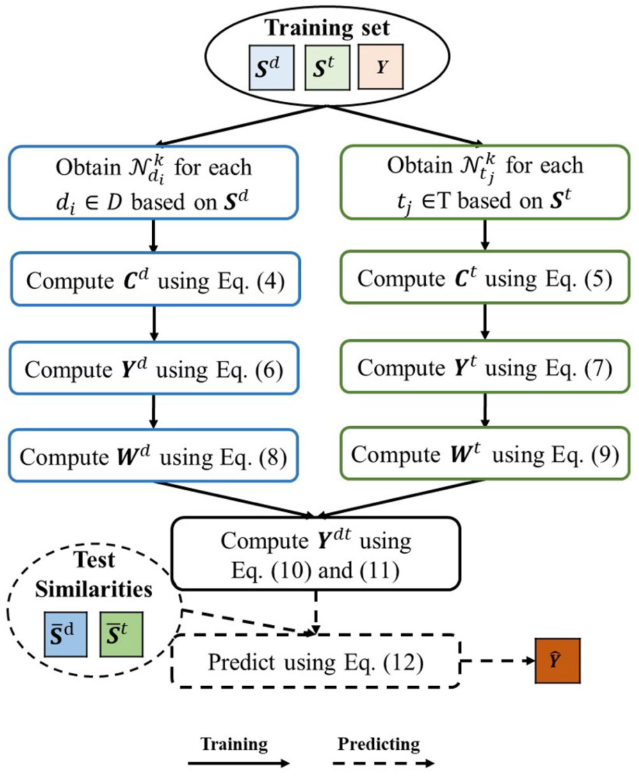
$$\boldsymbol{Y}^{dt} = \max\{\boldsymbol{Y}^{dt}, \boldsymbol{Y}\} \tag{11}$$

where max is the element wise maximum operator.

In the prediction phase, the estimated interaction between drug d_u and target t_v is calculated as:

$$\hat{Y}_{uv} = \begin{cases} \frac{1}{z_{d_u}} \sum\limits_{d_i \in \mathcal{N}^k_{d_u}} \eta^{i'-1} \bar{S}^d_{ui} Y^{dt}_{iv},\ \text{if } d_u \notin D \text{ and } t_v \in T \\ \frac{1}{z_{t_v}} \sum\limits_{t_j \in \mathcal{N}^k_{t_v}} \eta^{j'-1} \bar{S}^t_{vj} Y^{dt}_{uj},\ \text{if } d_u \in D \text{ and } t_v \notin T \\ \frac{1}{z_{uv}} \sum\limits_{(d_i, t_j) \in \mathcal{N}^k_{d_u t_v}} \eta^{i'+j'-2} \left(\bar{S}^d_{ui}\right)^{r_d} \left(\bar{S}^t_{vj}\right)^{r_t} Y^{dt}_{ij} \\ \quad \text{if } d_u \notin D \text{ and } t_v \notin T \end{cases} \tag{12}$$

where $r_d = \min\{1, L^d_{uv}/L^t_{uv}\}$ and $r_t = \min\{1, L^t_{uv}/L^d_{uv}\}$ are the coefficients controlling the weights of drug similarities and target similarities respectively. The smaller r_d is, the larger drug similarities become, as both similarities and r_d are between [0,1]. $L^d_{uv} = \sum_{(d_i, t_j) \in \mathcal{N}^k_{d_u t_v}} C^d_{ij} Y_{ij}$ and $L^t_{uv} = \sum_{(d_i, t_j) \in \mathcal{N}^k_{d_u t_v}} C^t_{ij} Y_{ij}$ are the sum of drug-based and target-based local imbalance of neighbor pairs of (d_u, t_v) respectively. Compared with (2) and (3) in WkNN, there are two improvements made in WkNNIR. The first one is the utilization of recovered interaction matrix with more sufficient interaction information. The second advantage of



Training set
S^d
S^t
Y
Obtain $\mathcal{N}_{d_i}^k$ for each $d_i \in D$ based on S^d
Obtain $\mathcal{N}_{t_j}^k$ for each $t_j \in$T based on S^t
Compute C^d using Eq. (4)
Compute C^t using Eq. (5)
Compute Y^d using Eq. (6)
Compute Y^t using Eq. (7)
Compute W^d using Eq. (8)
Compute W^t using Eq. (9)
Compute Y^{dt} using Eq. (10) and (11)
Test Similarities
$\bar{S}^d$
$\bar{S}^t$
Predict using Eq. (12)
$\hat{Y}$
Training
Predicting

Fig. 2 The workflow of WkNNIR. Solid and dashed arrows denote steps in the training and prediction phases, respectively

WkNNIR is the more reliable similarity assessment via local imbalance by using r_d and r_t in S4.

The workflow of WkNNIR is presented in Fig. 2. In the training phase, it sequentially computes $\mathcal{N}_{d_i}^k$ for each training drug $d_i \in D$, drug-based local imbalance matrix $\boldsymbol{C}^d$, drug-based recovery interaction matrix $\boldsymbol{Y}^d$ and $\boldsymbol{W}^d$ representing weights of $\boldsymbol{Y}^d$, using the training set. Similar variables relating to targets, namely $\mathcal{N}_{t_j}^k$, $\boldsymbol{C}^t$, $\boldsymbol{Y}^t$ and $\boldsymbol{W}^t$, are calculated too. Then, the final interaction matrix $\boldsymbol{Y}^{dt}$ incorporating the recovered interactions in both $\boldsymbol{Y}^d$ and $\boldsymbol{Y}^t$ is obtained according to (10) and (11). In the prediction phase, given test drug and/or target similarities, the estimated interaction matrix $\hat{Y}$ is obtained based on the recovered interaction matrix $\boldsymbol{Y}^{dt}$ according to (12).

5 Ensembles of DTI prediction models

Ensemble methods integrate multiple models that solve the same task, and therefore reduce the generalization error of single models. In this section, we propose three ensemble methods, namely ERS, EGS and ELS, which employ random sampling, global imbalance based sampling, and local imbalance based sampling respectively. The three proposed ensemble methods follow the same framework and can be applied to any DTI prediction method to further improve its accuracy. We firstly introduce the ensemble framework and then describe the sampling strategies used in each method.

5.1 Ensemble framework

The proposed ensemble framework learns multiple models based on diverse sampled training subsets. To adapt it to the DTI prediction task, the proposed framework needs to be modified in the following aspects: adjusting the sampled training subsets generation in the training phase as well as adding dynamic ensemble selection and similarity vector projection processes in the prediction phase.

Algorithm 1 Training ensembles of DTI prediction models.

```
input : training drug set: $D$, training target set: $T$,
        drug similarity matrix: $S^d$, target similarity
        matrix: $S^t$, interaction matrix: $Y$, sampling
        ratio: $R$, ensemble size: $q$
output: ensemble model: $M$, drug subsets: $D'$, target
        subsets: $T'$
1  $M, D', T' \leftarrow \emptyset$ ;
2  Calculate the sampling probability for drugs $p^d$;
3  Calculate the sampling probability for targets $p^t$;
4  for $i \leftarrow 1$ to $q$ do
5      $D^i \leftarrow$ sample $nR$ drugs from $D$ based on $p^d$ ;
6      $T^i \leftarrow$ sample $mR$ targets from $T$ based on $p^t$ ;
7      Construct training subset $S^{di}, S^{ti}, Y^i$ based on $D^i$
       and $T^i$ ;
8      $M_i \leftarrow \text{Train}(S^{di}, S^{ti}, Y^i, D^i, T^i)$ ;
9      $M \leftarrow M \cup M_i$ ;
10     $D' \leftarrow D' \cup D^i$ ;
11     $T' \leftarrow T' \cup T^i$ ;
12 return $M, D', T'$ ;
```

The pseudo code of training an ensemble model is shown in Algorithm 1. In the training phase, the sampling probabilities for drugs and targets, denoted as $\boldsymbol{p}^d \in \mathbb{R}^n$ and $\boldsymbol{p}^t \in \mathbb{R}^m$, are initially computed (Algorithm 1, line 2-3), where $\sum_{i=1}^n p_i^d = 1$ and $\sum_{j=1}^m p_j^t = 1$. The calculation of the sampling probabilities is different in each of the three methods and will be illustrated in Section 5.2. Then, q base models are trained iteratively. For the i-th base model, the nR sized drug subset D^i is sampled from D without replacement according to $\boldsymbol{p}^d$ (Algorithm 1, line 5), i.e. drugs with larger sampling probability have a greater chance to be added to D^i, where R is a user-specified sampling ratio controlling the number of selected drugs. In a similar way, the mR sized target subset T^i is derived from T based on $\boldsymbol{p}^t$ (Algorithm 1, line 6). In the next step, based on the D^i and T^i, we form a training subset which consists of a drug similarity sub-matrix $\boldsymbol{S}^{di}$ retaining similarities between drugs in D^i, a target similarity sub-matrix $\boldsymbol{S}^{ti}$ preserving the similarities between targets in T^i and an interaction sub-matrix $\boldsymbol{Y}^i$ storing interactions involving both D^i and T^i (Algorithm 1, line 7). Finally, the base model M_i is trained on the obtained training subset (Algorithm 1, line 8).

Algorithm 2 Predicting of ensemble DTI prediction.

```
   input : ensemble of DTI prediction model: $M$, drug
           subsets: $D'$, target subsets: $T'$, training drug
           set: $D$, training target set: $T$, test drug-target
           pair: ($d_u$,$t_v$), similarity of $d_u$ to $D$: $s_u^d$,
           similarity of $t_v$ to $T$: $s_v^t$
   output: predicting interaction between $d_u$ and $t_v$: $\hat{Y}_{uv}$
 1 for $i \leftarrow 1$ to $|M|$ do /* Dynamic ensemble
   selection                                          */
 2     if $d_u \notin D$ and $t_v \in T$ then /* S2    */
 3         if $t_v \notin T_i$ then
 4             $M \leftarrow M - M_i$ ;
 5     if $d_u \in D$ and $t_v \notin T$ then /* S3    */
 6         if $d_u \notin D_i$ then
 7             $M \leftarrow M - M_i$ ;
 8 $\hat{Y}_{uv} \leftarrow 0$ ;
 9 for $i \leftarrow 1$ to $|M|$ do /* Predicting      */
10     $s_u^{\prime d} \leftarrow$ Project $s_u^d$ on $D^i$ ;
11     $s_v^{\prime t} \leftarrow$ Project $s_v^t$ on $T^i$ ;
12     $Y_{uv}' \leftarrow$ Predict($M_i, s_u^{\prime d}, s_v^{\prime t}, d_u, t_v$) ;
13     $\hat{Y}_{uv} \leftarrow \hat{Y}_{uv} + Y_{uv}'$ ;
14 $\hat{Y}_{uv} \leftarrow \hat{Y}_{uv}/|M|$ ;
15 return $\hat{Y}_{uv}$ ;
```

The prediction process of the ensemble framework is illustrated in Algorithm 2. The input similarity $\boldsymbol{s}_u^d$ ($\boldsymbol{s}_v^t$) is equivalent to $\boldsymbol{S}_{u\cdot}^d$ ($\boldsymbol{S}_{v\cdot}^t$) if $d_u \in D$ ($t_v \in T$), and $\bar{\boldsymbol{S}}_{u\cdot}^d$ ($\bar{\boldsymbol{S}}_{v\cdot}^t$) otherwise. Dynamic ensemble selection takes place first if the test drug-target pair (d_u, t_v) follows the condition of S2 or S3. Specifically, when (d_u, t_v) follows S2, M_i is discarded if test target $t_v \notin T_i$ (Algorithm 2, line 2-4). Compared with other base models involving t_v in the training process, M_i misses the information of interaction regarding t_v and usually leads to a lower prediction accuracy for (d_u, t_v). Analogously, the models whose corresponding training drug subset does not contain d_u are discarded if (d_u, t_v) follows S3 (Algorithm 2, line 5-7). Dynamic ensemble selection is not applied to S4 ($d_u \notin D$ and $t_v \notin T$), because both d_u and t_v are new emerging drugs and targets for all base models in S4. In the next steps, all retained base models give their prediction, which are eventually averaged to obtain the final predicted score $\hat{Y}_{uv}$ (Algorithm 2, line 9-14). It should be noticed that there are two projection steps to ensure the similarity vectors of the drug and target in the test pair fit the input of each base model (Algorithm 2, line 10-11). As each base model is trained based on a subset of drugs and targets, the similarity vector for d_u and t_v should be projected to the low dimensional space characterized by the drug and target subset used in the corresponding base model. Specifically, the projection of $\boldsymbol{s}_u^d$ on D_i maintains the similarities between d_u and drugs in D_i and deletes other elements in $\boldsymbol{s}_u^d$. For example, given a similarity vector [0.1, 0.2, 0.3, 0.4, 0.5], its projection on drug subset $\{d_1, d_2, d_4\}$ is [0.1, 0.2, 0.4].

5.2 Sampling probability

Sampling probabilities determine the opportunity of each drug and target being used to train base models, which play a key role in the proposed ensemble framework. As we mentioned before, the three proposed ensemble methods employ different sampling probabilities.

ERS adopts the sampling probabilities following the uniform distribution, i.e. $p_i^d = 1/n$ and $p_j^t = 1/m$, where $i = 1, 2, ...n$ and $j = 1, 2, ...m$. In this way, each drug and target has an equal chance to be selected.

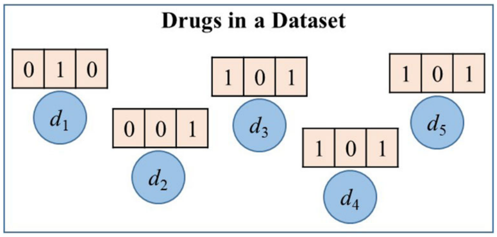
In DTI data, interacting drug-target pairs are heavily outnumbered by non-interacting ones, resulting in an imbalanced distribution within the global interaction matrix. To relieve this global imbalance, EGS forms training subsets by biasing the sampling process to include drugs and targets having more interactions. Moreover, another reason for emphasis on drugs and targets with dense interactions is that they are more informative than others with fewer interactions. In EGS, the sampling probability of each drug (target) is proportional to the number of its interactions:

$$\begin{aligned} p_i^d &= \frac{\sigma + \sum_{j=1}^m Y_{ij}}{n\sigma + \sum_{h=1}^n \sum_{j=1}^m Y_{hj}}, \quad i = 1, 2, ...n \\ p_j^t &= \frac{\sigma + \sum_{i=1}^n Y_{ij}}{m\sigma + \sum_{i=1}^n \sum_{h=1}^m Y_{ih}}, \quad j = 1, 2, ...m \end{aligned} \tag{13}$$

where σ is a smoothing parameter. By using (13), the drugs and targets with more interactions are more likely to be selected in the sampling procedure.

Apart from the global imbalance, the local imbalance could also be used to assess the importance of that drug (target). According to (4), higher C_{ij}^d means that d_i is surrounded by more drugs that have opposite interactivity to t_j. In such cases, correctly predicting Y_{ij} using drug similarities would be difficult. By accumulating the local imbalance (difficulty) of d_i for all interacting targets, we arrive at the local imbalance based importance of d_i:

$$LI_i^d = \sum_{j=1}^m C_{ij}^d [\![Y_{ij} = 1]\!] \tag{14}$$



Drug Sampling Probability

	p_1^d	p_2^d	p_3^d	p_4^d	p_5^d
ERS	0.2	0.2	0.2	0.2	0.2
EGS	0.129	0.129	0.247	0.247	0.247
ELS	0.467	0.244	0.244	0.022	0.022

Fig. 3 The differences of the three sampling probabilities. The top half shows the drugs' location (blue circles) and their interactions (light red rectangles) of a dataset including five drugs and three targets. The table in the bottom half lists the drug sampling probabilities computed by the three ensemble methods on this dataset with $\sigma = 0.1$ and $k = 2$

LI_i^d is a weighted sum version of $\sum_{j=1}^{m} Y_{ij}$, where the interaction which is difficult to be predicted correctly is assigned a higher weight. Compared with EGS only counting the interactions of drugs, (14) emphasizes on drugs having more difficult interactions.

Similarly, the local imbalance based importance of t_j is defined as:

$$LI_j^t = \sum_{i=1}^{n} C_{ij}^t [\![Y_{ij} = 1]\!] \tag{15}$$

Based on the definition of local imbalanced based importance, the key idea in ELS is that it encourages more *difficult* drugs and targets to be learned by more base models, reducing thereby the corresponding error to the greatest extent possible. In ELS, the sampling probability is proportional to the corresponding local imbalance based importance:

$$\begin{aligned} p_i^d &= \frac{\sigma + LI_i^d}{n\sigma + \sum_{h=1}^{n} LI_h^d}, \quad i = 1, 2, \ldots n \\ p_j^t &= \frac{\sigma + LI_j^t}{m\sigma + \sum_{h=1}^{m} LI_h^t}, \quad j = 1, 2, \ldots m \end{aligned} \tag{16}$$

In addition, we exemplify the differences of three sampling probabilities in Fig. 3. In ERS, all drugs have equal chance to be selected. EGS is more likely to choose d_3, d_4 and d_5 that have more interacting targets. ELS will select d_1 with higher probability, because d_1 is near (similar) to drugs (d_2 and d_3) having different interactions and is therefore more difficult to be learned than other drugs.

6 Experiments

In this section, the datasets and the evaluation protocol used in the experiments are presented firstly. Then, the predictive performance and parameter analysis of proposed WkNNIR and three ensembles are reported. Finally, newly discovered interactions found by our methods are presented.

6.1 Dataset

Five benchmark DTI datasets are used in our empirical study. Four of them are gold standard datasets originally provided by [19], of which each corresponds to a target protein family, namely Nuclear Receptors (NR), G-protein coupled receptors (GPCR), Ion Channel (IC) and Enzyme (E). The last dataset (DB), obtained from [58], was derived from DrugBank [59]. Table 1 lists the information about the five datasets. Sparsity is the proportion of interacting drug-target pairs which indicates the global imbalance of the dataset. LI^d and LI^t are the drug-based and target-based local imbalance of the whole dataset respectively:

$$LI^d = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij}^d Y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}} \tag{17}$$

$$LI^t = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij}^t Y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}} \tag{18}$$

where k is the number of neighbors, which is set to 5. Smaller LI^d (LI^t) values indicate more reliable drug (target) similarities and easier S2 (S3) prediction task.

6.2 Evaluation protocol

Three types of cross validation (CV) are conducted to examine the prediction methods in three prediction settings, respectively. In S2, the drug wise CV is applied where one drug fold along with their corresponding rows in $\boldsymbol{Y}$ are separated for testing. In S3, the target wise CV is utilized

Table 1 Statistic of DTI datasets

Dataset	# Drugs	# Targets	# Interactions	Sparsity	LI^d	LI^t
NR	54	26	90	0.064	0.658	0.764
GPCR	223	95	635	0.03	0.707	0.644
IC	210	204	1476	0.035	0.729	0.323
E	445	664	2926	0.01	0.737	0.35
DB	786	809	3681	0.006	0.653	0.528

where one target fold along with their corresponding column in $\boldsymbol{Y}$ are left out for testing. The block wise CV, which splits a drug fold and target fold along with the interactions between them (which is a sub-matrix of $\boldsymbol{Y}$) for testing and uses interactions between remaining drugs and targets for training, is applied to S4. Two repetitions of 10-fold CV are applied to S2 and S3, and two repetitions of 3-fold block wise CV which contains 9 block folds generated by 3 drug folds and 3 target folds are applied to S4.

The Area Under the Precision-Recall curve (AUPR) which heavily punishes highly ranked false positive predictions [60] is used to evaluate the performance of inductive DTI prediction approaches in our experiments. In addition, the Wilcoxon signed rank test at 5% level is utilized to examine the statistically significant difference between our methods and the compared ones.

In the experiments, we firstly compare the proposed WkNNIR with six DTI prediction approaches, namely ALADIN [17], BICTR [18], BLM-MLkNN [26, 30], BLM-NII [10], MLkNNSC [15], NRLMF [9], as well as WkNN which is the baseline of WkNNIR without interaction recovery and local imbalance weighting. All comparing methods except for BICTR either follow the transductive DTI prediction task or work for partial prediction settings. Therefore, we extend those methods to handle the inductive DTI prediction problem and all prediction settings as follows:

- ALADIN [17]: ALADIN works for S2 and S3 within the inductive scheme. As a BLM approach, ALADIN could deal with S4 by using the two-step learning strategy.
- BLM-MLkNN [26, 30]: An individual MLkNN model could only deal with S2 and S3 following the inductive scheme. To deal with S4, the MLkNN is embedded in the BLM and the two-step learning strategy is adopted.
- MLkNNSC [15]: MLkNNSC is initially proposed to predict interactions for test drugs (S2) and test drug-target pairs (S4) and it could work as an inductive approach directly. To handle S3 prediction setting, MLkNNSC is extended via applying clustering for drugs to obtain super-drugs and training MLkNN models for drugs and super-drugs, respectively.
- BLM-NII [10]: BLM-NNI is a transductive method and able to tackle all prediction settings. To enable BLM-NNI to adapt the inductive DTI prediction, we modify its training and prediction process by confining that similarities for test drugs (targets) are available in prediction phase only.
- NRLMF [9]: NRLMF predicts interactions for S2, S3, and S4 in the transductive way. Similar to BLM-NII, NRLMF is altered to an inductive method by excluding similarities for test drugs and targets from the input of the training phase.

Moreover, BICTR trains an ensemble of bi-clustering trees on a reconstructed interaction matrix which is completed by NRLMF, and the input of the tree-ensemble model is the drug and target features. Hence, the similarities are utilized as features in BICTR, i.e. the feature vector of a drug (target) is its similarities to all training drugs (targets).

When it comes to ensemble methods i.e. ERS, EGS and ELS, four comparing DTI methods, namely ALADIN, BLM-NII, MLkNNSC, NRLMF and the two proposed neighborhood based approaches (WkNN and WkNNIR) are utilized as base model. We do not employ BICTR and BLM-MLkNN as base model, because the former one is already an ensemble model and the latter one is the worst method in most cases.

The parameter settings of compared and proposed methods are listed in Table 2. The values for used parameters are selected by performing the inner CV on the training set. Specifically, the 5-fold inner CV is applied to S2 and

Table 2 Parameter settings

Method	Values or ranges of parameters
ALADIN	#models=25, $k\in\{1,2,3,5,7,9\}$, #features$\in\{10,20,50\}$
BICTR	#trees=100, minimal # samples in leaf=1 (2) in S3 and S4 (S2)
BLM-MLkNN	$k\in\{1,2,3,5,7,9\}$
BLM-NII	$\gamma = 1$, $\alpha\in\{0, 0.1,...,1.0\}$, $\lambda\in\{2^{-5}, 2^{-4}, ..., 2^{0}\}$
MLkNNSC	cut-off threshold= 1.1, $k\in\{1,2,3,5,7,9\}$
NRLMF	c=5, k=5, $r\in\{50,100\}$, λ_t, λ_d, α, $\beta\in\{2^{-5}, 2^{-4}, ..., 2^{0}\}$
WkNN	$k\in\{1,2,3,5,7,9\}$, $\eta\in\{0.1,0.2,...,1.0\}$
WkNNIR	$k\in\{1,2,3,5,7,9\}$, $\eta\in\{0.1,0.2,...,1.0\}$
ERS	$q = 30$, $R = 0.95$
EGS	$q = 30$, $R = 0.95$ $\sigma = 1.0(0.1)$ for DB dataset (others)
ELS	$q = 30$, $R = 0.95$, $\sigma = 1.0(0.1)$ for DB dataset (others), $k = 5$

S3, and the 2-fold inner CV is applied to S4 on GPCR, IC, E and DB datasets. For NR dataset containing fewer drugs and targets, splitting NR into fewer folds generates small-sized training sets that may vary the distribution of the whole dataset during the inner CV procedure. This leads to the unreliability of chosen optimal parameter settings. To avoid this issue, we apply the 10-fold inner CV for S2 and S3 and the 3-fold inner CV for S4 on NR dataset.

6.3 Results

In this part, the obtained results comparing the proposed WkNNIR and its baseline WkNN to other competitors are presented and discussed. Next, the results of the three proposed ensemble methods with six different base models are reported.

Table 3 shows the AUPR results for the compared approaches in various prediction prediction settings, where "*" following the numerical results indicates that the corresponding method is statistically different from WkNNIR using Wilcoxon signed rank test at 5% level. WkNNIR achieves the best average rank in all settings. It is significantly superior to other methods in 83/105 cases and does not suffer any significant losses from other competitors. In addition, WkNNIR significantly outperforms WkNN in 13/15 cases, demonstrating the effectiveness of the utilization of the interaction recovery and local imbalance-based weights. Then, we investigate the results in each prediction setting. In S2, WkNNIR is the best method on all datasets. WkNN is the second best method and NRLMF comes next in most cases. In S3, for more difficult datasets (GPCR and DB having higher target-based local imbalance), WkNNIR is the top method, while for easier datasets (IC and E having lower target-based local imbalance) and the small-sized dataset (NR), WkNNIR achieves comparable performance to the corresponding best method without statistically significant difference. In S4, WkNNIR outperforms other competitors on the first three datasets and is slightly inferior to WkNN on E as well as BICTR on DB without significant difference. WkNN outperforms all other six competitors on the first four datasets, which indicates the effectiveness of our proposed neighbor pair based prediction function for S4. Overall, WkNNIR surpasses the compared methods in S2 and S4 and is comparable with state-of-the-art approaches in S3.

As is deducted from the obtained results, the three prediction settings, namely S2, S3, and S4, are not equally challenging. In Table 3, AUPR results of all methods in S4 are immensely lower than that in S2 and S3 on all datasets. This shows that predicting interactions between test drugs and test targets is the most challenging task. Comparing results in S2 and S3, we find that all methods achieve higher AUPR in S3 than S2 on GPCR, IC, E and DB whose LI^t is lower than LI^d. While the performance of all methods, except for BLM-NII, in S3 is inferior to S2 on NR, whose LI^t is higher than LI^d. This indicates that the difficulty of S2 and

Table 3 Results of comparison inductive DTI prediction methods in terms of AUPR

Setting	Dataset	ALADIN	BLMNII	BLM-MLkNN	BICTR	MLkNNSC	NRLMF	WkNN	WkNNIR
S2	NR	0.433*(8)	0.441*(6)	0.436*(7)	0.462*(4)	0.456*(5)	0.513(2)	0.51*(3)	**0.539(1)**
	GPCR	0.306*(8)	0.342*(4.5)	0.326*(7)	0.328*(6)	0.342*(4.5)	0.345*(3)	0.369*(2)	**0.384(1)**
	IC	0.35(4)	0.317*(6)	0.298*(8)	0.359(2)	0.312*(7)	0.343*(5)	0.354*(3)	**0.363(1)**
	E	0.289*(7)	0.26*(8)	0.353*(3)	0.338*(6)	0.34*(5)	0.352*(4)	0.385*(2)	**0.396(1)**
	DB	0.41*(4)	0.202*(8)	0.365*(7)	0.423(2)	0.366*(6)	0.386*(5)	0.413*(3)	**0.425(1)**
	AveRank	6.2	6.5	6.4	4	5.5	3.8	2.6	**1**
S3	NR	0.383*(6)	0.447*(3)	0.37*(8)	0.392*(5)	0.38*(7)	**0.471(1)**	0.443*(4)	0.461(2)
	GPCR	0.517*(5)	0.476*(8)	0.493*(7)	0.539*(3)	0.511*(6)	0.518*(4)	0.541*(2)	**0.577(1)**
	IC	**0.803(1)**	0.787(6)	0.777*(8)	0.799(2)	0.784(7)	0.798(3.5)	0.789*(5)	0.798(3.5)
	E	0.758*(6)	0.77*(4.5)	0.75*(7)	0.77*(4.5)	0.748*(8)	**0.786(1)**	0.776*(3)	0.78(2)
	DB	0.569*(5)	0.433*(8)	0.563*(6)	0.554*(7)	0.579*(4)	0.585(2)	0.581*(3)	**0.595(1)**
	AveRank	4.6	5.9	7.2	4.3	6.4	2.3	3.4	**1.9**
S4	NR	0.095*(7)	0.135(5)	0.07*(8)	0.154(3)	0.117*(6)	0.142(4)	0.159(2)	**0.165(1)**
	GPCR	0.114*(6)	0.121*(5)	0.091*(8)	0.135*(3)	0.1*(7)	0.134*(4)	0.149*(2)	**0.158(1)**
	IC	0.206*(5)	0.176*(6)	0.131*(8)	0.213(4)	0.148*(7)	0.215(3)	0.216*(2)	**0.226(1)**
	E	0.128*(8)	0.147*(5.5)	0.147*(5.5)	0.177*(4)	0.146*(7)	0.198(3)	**0.208(1)**	0.202(2)
	DB	0.248(3)	0.064*(8)	0.183*(7)	**0.257(1)**	0.205*(6)	0.226*(5)	0.247*(4)	0.251(2)
	AveRank	5.8	5.9	7.3	3	6.6	3.8	2.2	**1.4**

The parenthesis is the rank of each method among all competitors

Bold signify the best method(s)

Table 4 Results of comparing inductive DTI prediction methods on the updated gold standard datasets in terms of AUPR

Setting	Dataset	ALADIN	BICTR	NRLMF	WkNN	WkNNIR
S2	NR1	0.527(4)	0.523(5)	0.529(3)	0.547(2)	**0.552(1)**
	GCPR1	0.438(5)	0.447(4)	0.456(3)	0.461(2)	**0.468(1)**
	IC1	0.475(5)	0.509(3)	0.502(4)	0.563(2)	**0.571(1)**
	E1	0.301(5)	0.331(4)	0.362(3)	0.368(2)	**0.382(1)**
	AveRank	4.75	4	3.25	2	**1**
S3	NR1	0.465(5)	0.518(4)	0.52(3)	0.542(2)	**0.565(1)**
	GCPR1	0.874(2)	0.871(3)	**0.888(1)**	0.863(5)	0.866(4)
	IC1	0.75(2)	0.748(3)	0.731(5)	0.747(4)	**0.762(1)**
	E1	0.658(5)	0.687(3)	**0.705(1)**	0.685(4)	0.695(2)
	AveRank	3.5	3.25	2.5	3.75	**2**
S4	NR1	0.207(5)	**0.283(1)**	0.27(3)	0.271(2)	0.259(4)
	GCPR1	0.32(5)	0.339(2)	0.338(3)	0.337(4)	**0.343(1)**
	IC1	0.325(5)	0.356(4)	0.357(3)	**0.395(1)**	0.388(2)
	E1	0.146(5)	0.181(4)	**0.214(1.5)**	**0.214(1.5)**	0.209(3)
	AveRank	5	2.75	2.63	**2.13**	2.5

Bold signify the best method(s)

S3 could be estimated by comparing LI^d and LI^t, e.g. S2 is harder than S3 if LI^d is higher than LI^t, and vice versa. This also verifies the effectiveness of the local imbalance to assess the reliability of drug and target similarities.

As we previously stated, DTI datasets usually contain many missing interactions, e.g. the four gold standard datasets only contain interactions discovered before they were constructed (in 2007). To test the effectiveness of WkNNIR on datasets with fewer missing interactions, we follow the procedure described in [61] to build updated gold standard datasets that include more validated interactions. Specifically, we add newly discovered interactions between drugs and targets in the original datasets recorded in the up-to-date version of KEGG [62], DrugBank [59], ChEMBL [63] and Matador [64] databases. There are 175, 1350, 3201, 4640 interactions in the four updated dataset, denoted as NR1, GPCR1, IC1 and E1 respectively, with 85, 715, 1725 and 1714 new interactions appended.

Table 4 lists the AUPR results of WkNNIR and four competitive comparing methods on the updated gold datasets. BLMNII, BLM-MLkNN and MLkNNSC are not included in the experiments on updated datasets due to their poor performance, as reported in Table 3. In Table 4, we see that WkNNIR is still the best method in S2 and S3. In S4, WkNNIR is slightly inferior to WkNN, because the benefit of the interaction recovery used in WkNNIR is not significant when dealing with datasets having fewer missing interactions. In addition, comparing the results in Tables 3 and 4, we find that the performance on the updated datasets with less missing interactions is usually better than in the original datasets. This verifies that missing interactions indeed hinder DTI prediction methods from achieving better performance.

The average rank of ensemble methods along with their embedded base models in terms of AUPR are summarized in Table 5. The Base column denotes the average ranks of default base models, and the “◦” following the $AveRank$ denotes that the corresponding ensemble method is statistically superior to the base model using Wilcoxon signed rank test at 5% level. The detailed numerical AUPR results are listed in Appendix Tables 13-15. We find that all three ensemble methods achieve better average rank compared to the base models in all prediction settings. ELS is the most effective method and significantly outperforms the base models in all prediction settings. This is because ELS emphasizes on *difficult* drugs and targets by considering local imbalance. EGS aiming to reduce the global imbalance level comes next and its advantage over base models is significant as well. ERS using a totally random sampling strategy is the third one and only achieves significant improvement in S3 and S4.

Furthermore, to check the effectiveness of the proposed ensemble methods on each base model, we calculate the average rank of the three ensemble methods on all datasets for each prediction setting and base model and pick up the best ones to show in Table 6. We divide the employed

Table 5 The average ranks of ensemble methods over six base models and five datasets in three prediction settings

Setting	Base	ERS	EGS	ELS
S2	2.87	2.67	2.25◦	**2.22**◦
S3	3.22	2.37◦	2.22◦	**2.2**◦
S4	3.72	2.43◦	2.24◦	**1.59**◦

Bold signify the best method(s)

Table 6 The best ensemble method for each prediction setting and base model

Base Model	S2	S3	S4
ALADIN	EGS	EGS	EGS
BLMNII	ELS	ELS	ELS
MLkNNSC	ERS	ERS	ERS
NRLMF	EGS	ELS	ELS
WkNN	ELS	ELS	ELS
WkNNIR	ELS	EGS	ELS

base models into two groups based on their performance in Table 3: moderate base models (ALADIN, BLMNII and MLkNNSC) and good base models (NRLMF, WkNN, and WkNNIR). Regarding the moderate base models, ERS, EGS and ELS are most effective on MLkNNSC, ALADIN and BLMNII, respectively. When it comes to good models, ELS usually outperforms the other two ensemble methods. EGS is the top one only for NRLMF in S2 and WkNNIR in S3. This suggests that ELS is more beneficial to base models with better prediction performance.

6.4 Parameter analysis

Here, we analyze the influence of parameter settings on WkNNIR and three ensemble methods.

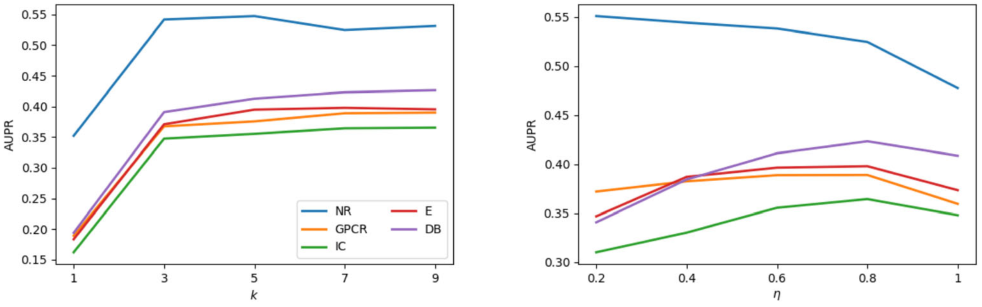
Firstly, we investigate the sensitivity of WkNNIR with respect to k and η in S2 which is shown in Fig. 4. The performance of WkNNIR improves a lot from $k = 1$ to $k = 3$, as more neighbors are exploited. However, for values of k larger than 3, the accuracy of WkNNIR plateaus, indicating that extra neighbors do not provide additional benefits. In terms of η, median values (around 0.6 and 0.8) lead to the best performance on all datasets except for NR. The lower (higher) η diminishes (promotes) the influence of the relatively remote neighbors on the prediction, leading to performance deterioration. In NR, which contains tens of drugs and targets, the performance drops as η increases, indicating that lower η is suitable for the small-sized dataset. Similar accuracy trends with respect to k and η are observed in S3 and S4 as well.

Figure 5 shows the performance of ensemble methods using WkNNIR as the base model under different parameter settings on E dataset in S2. In Fig. 5a concerning the sampling ratio, the performance of all three ensemble methods improve with the increase of R. This is mainly because selecting more drugs and targets contributes to training more accurate base models in them. Figure 5b presents the influence of ensemble size q on each method. ERS achieves better performance when more base models are trained because larger q would increase the chance of important drugs and targets being selected by ERS. Nevertheless, ERS training more than 50 base models cannot even surpass ELS and EGS with smaller ensemble size. EGS and ELS perform more stably under different ensemble sizes. This is because EGS and ELS guarantee that important drugs and targets would be more likely to be chosen and the benefit of training more base models is limited to them. In Fig. 5c, we find that EGS and ELS are insensitive to the smooth parameter σ. At last, as shown in Fig. 5d, ELS reaches a plateau when k which relates to the local imbalance of each drug and target is larger than 3.

6.5 Discovery of new interactions

In this section, we examine the ability of WkNNIR and ensemble methods to discover new reliable DTIs in all prediction settings. For each prediction setting and original gold standard dataset, we conducted cross-validation as illustrated in Section 6.2 and combined the predictions of each fold to obtain the predicted scores of all drug-target pairs in the dataset. Then, we ranked all non-interacting drug-target pairs of the dataset according to their predicted scores and select the top 10 non-interacting pairs as the candidate newly predicted interactions. To verify the reliability of those new interactions, we checked whether they are included in the corresponding updated dataset that incorporate validated interactions from the last version of the KEGG, DrugBank, ChEMBL and Matador databases.

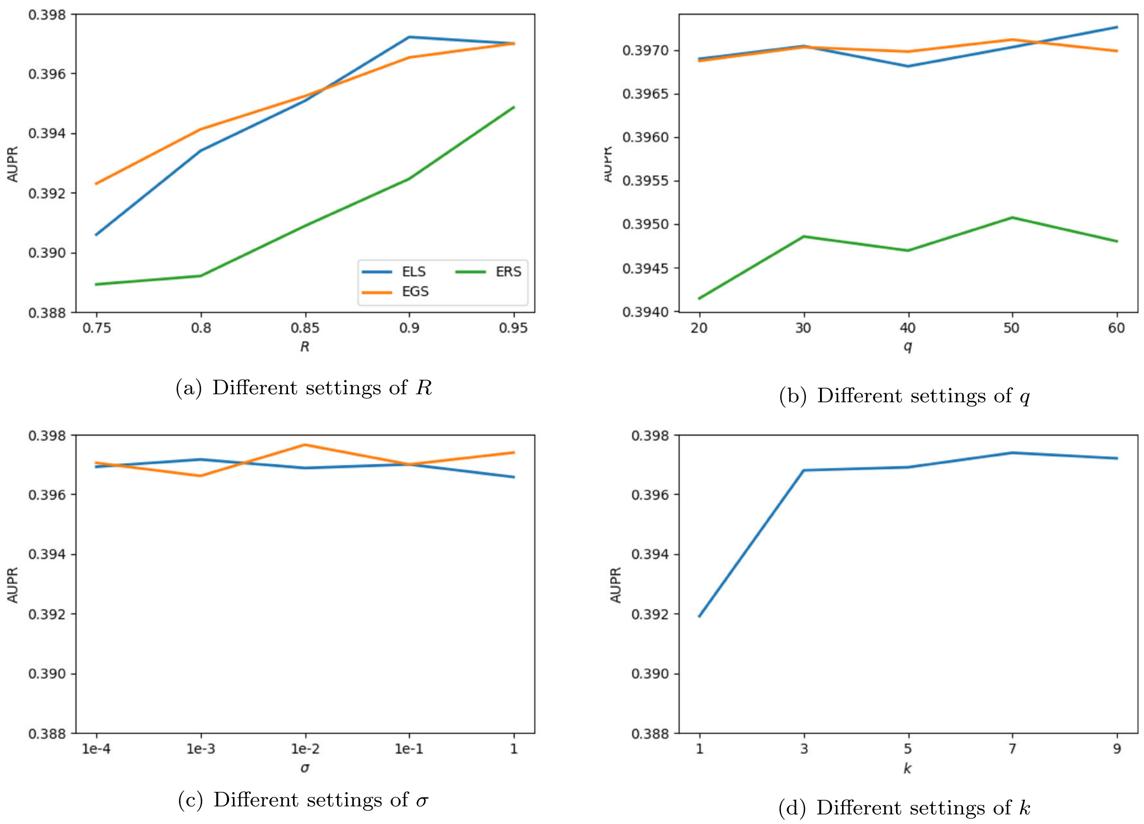
The validated new interactions found by WkNNIR on the four original gold standard datasets are listed in Tables 7, 8, 9 and 10, where K, D, C, M indicate that the corresponding DTI is verified by KEGG, DrugBank, ChEMBL and Matador, respectively. WkNNIR could find at least one new DTI confirmed by external databases in



(a) Different settings of k

(b) Different settings of η

Fig. 4 Performance of WkNNIR with different settings of k and η in S2 where η is set as 0.8 for (a) and k is set as 7 for (b)



(a) Different settings of R

(b) Different settings of q

(c) Different settings of σ

(d) Different settings of k

Fig. 5 Performance of ERS, EGS, ELS using WkNNIR as the base model under different parameter settings on E in S2

each dataset and setting. Furthermore, we notice that there are many new DTIs discovered in only one prediction setting, e.g. for NR dataset D00690-hsa2098 is only found in S2 and D00954-hsa367 is only found in S4. Therefore, it is beneficial to find more new DTIs by examining the predictions from all settings. For example, four new DTIs could be found by looking at only setting S2 in the NR dataset. Nonetheless, nine new DTIs can be discovered if all settings are considered (D00348-hsa5915 is found in both S2 and S3).

Table 7 The validated new interactions predicted by WkNNIR in NR dataset

Setting	Drug	Target	Rank	Database
S2	D00690	hsa2908	3	K, D
	D00348	hsa5915	5	K
	D00348	hsa5916	6	K, D
	D05341	hsa2099	10	C
S3	D00279	hsa5468	8	C
	D00348	hsa5915	9	K
S4	D00954	hsa367	1	D
	D00067	hsa2100	3	K
	D00182	hsa367	9	D
	D00554	hsa2100	10	K

In addition, Table 11 summarizes the total number of validated new DTIs from top X ($X \in \{10, 20, 30\}$) candidate pairs provided by WkNNIR and the other four competitors in all three prediction settings. The new DTIs discovered by more than two prediction settings are counted only once. It should be also mentioned that the maximum

Table 8 The validated new interactions predicted by WkNNIR in GPCR dataset

Setting	Drug	Target	Rank	Database
S2	D04625	hsa154	1	K, D, C
	D02358	hsa154	2	D
	D02147	hsa153	3	D, M
	D00110	hsa1128	4	D
	D02349	hsa154	6	K, D
	D00095	hsa155	7	K
	D00394	hsa3269	8	D
	D00715	hsa1129	10	K, D
S3	D00283	hsa152	1	K, D
	D00283	hsa1814	2	D, M
S4	D01103	hsa1133	4	K
	D06056	hsa3357	6	D
	D06056	hsa3358	7	D
	D06056	hsa3356	8	D

Table 9 The validated new interactions predicted by WkNNIR in IC dataset

Setting	Drug	Target	Rank	Database
S2	D00438	hsa779	1	K, D
	D00319	hsa783	6	D, M
	D00319	hsa786	7	M
	D00553	hsa6329	9	K
	D00553	hsa6334	10	K
S3	D00336	hsa10060	1	D
	D00349	hsa9254	9	D
S4	D00553	hsa6328	4	K
	D00553	hsa6334	5	K
	D01287	hsa11280	6	K
	D04048	hsa11280	7	K
	D01450	hsa11280	8	K
	D00358	hsa6332	9	M

of new DTIs found from top X candidates is $3X$, as we take three predictions settings into consideration. Regarding the top 10 candidate pairs, ALADIN finds the most new DTIs, followed by WkNNIR which discovers four fewer new DTIs among all datasets than ALADIN. However, picking only 10 candidate pairs may be not sufficient for practical applications, especially when a large amount of drug-target pairs need to be tested, e.g. E dataset contains 295,480 drug-target pairs. When one increases the number of candidate pairs, WkNNIR becomes the most effective method detecting at least ten more new DTIs than other competitors. This indicates that WkNNIR outperforms other competitors in terms of the ability to predict new DTIs, particularly when many candidate pairs are considered.

Furthermore, in order to check whether we can find more new interactions by applying ensemble methods to WkNNIR, we collected validated new DTIs discovered by the three proposed ensemble methods using WkNNIR as the base model. Table 12 lists the validated new DTIs only found by ensemble methods. ELS, EGS and ERS find 14, 9 and 12 new DTIs on the four datasets, respectively. This demonstrates that ensemble methods are indeed very promising in discovering new (not yet reported) DTIs.

Table 10 The validated new interactions predicted by WkNNIR in E dataset

Setting	Drug	Target	Rank	Database
S2	D00364	hsa1565	10	K, D
S3	D00947	hsa4129	1	D
	D00005	hsa4128	2	D
	D05458	hsa4128	3	K, D
	D01223	hsa3988	5	M
	D00437	hsa1585	6	M
	D00217	hsa7173	10	M
S4	D00528	hsa5150	3	D
	D00528	hsa50940	4	D

Table 11 The number of validated new DTIs from top 10, 20 and 30 candidate pairs provided by DTI prediction methods

	Dataset	ALADIN	BICTR	NRLMF	WkNN	WkNNIR
Top 10	NR	6	6	6	6	**9**
	GPCR	**18**	15	13	14	14
	IC	**14**	8	13	10	12
	E	10	**11**	10	8	9
	Sum	**48**	40	42	38	44
Top 20	NR	11	17	14	14	**16**
	GPCR	**33**	31	22	22	**33**
	IC	**28**	18	**28**	19	27
	E	18	17	16	16	**25**
	Sum	90	83	80	71	**101**
Top 30	NR	16	20	19	17	**22**
	GPCR	**47**	41	33	37	42
	IC	41	33	**43**	30	37
	E	20	21	27	23	**37**
	Sum	124	115	122	107	**138**

Bold signify the best method(s)

7 Conclusion

In this paper, we propose a new neighborhood-based method (WkNNIR), which is able to deal with all types of interaction prediction settings and successful in handling missing interactions. Furthermore, we propose three ensemble methods, namely ERS, EGS and ELS, that integrate multiple DTI prediction models trained upon various sampled datasets to improve the performance of their embedded base model. Both WkNNIR and ensemble methods were applied to five benchmark DTI prediction datasets. The obtained results affirm the superiority of WkNNIR to other competing methods as well as its baseline WkNN. The performance improvement provided by the ensemble methods to six base models including WkNNIR is also verified. Particularly, ELS using local imbalance based sampling is the most outstanding ensemble approach. Subsequently, we demonstrated that our methods are able to predict reliable new drug–target interactions.

Although only chemical structure similarities and protein sequence similarities are used in this study, there are various kinds of similarities revealing the relationship between drugs (targets) in diverse aspects [13, 65, 66]. It is therefore desirable to investigate the effectiveness of our methods using different types of similarities as well as extending it to multi-modal or multi-view settings, where several types of similarities are integrated and exploited.

Table 12 The validated new interactions only found by ensemble methods with WkNNIR as base model

Dataset	Setting	Drug	Target	Method (Rank)	Database
NR	S4	D00950	hsa367	ERS(6), EGS(7), ELS(9)	D
GPCR	S2	D00790	hsa3269	ERS(10)	D
	S4	D00540	hsa1131	ERS(7), EGS(8), ELS(10)	K, D
		D00394	hsa1133	ERS(8), EGS(10), ELS(9)	D
IC	S2	D06172	hsa6334	ERS(6), ELS(6)	K
		D06172	hsa6329	ERS(7), ELS(5)	K
		D06172	hsa6335	ERS(8), ELS(8)	K
		D06172	hsa6323	ERS(9), ELS(4)	K
		D06172	hsa6328	ERS(10), ELS(7)	K
		D00553	hsa6335	EGS(7), ELS(9)	K
		D00553	hsa6328	EGS(8)	K
		D00553	hsa6323	EGS(9)	K
		D00553	hsa6326	EGS(10), ELS(10)	K
		D06172	hsa6326	ELS(3)	K
	S4	D00438	hsa779	ERS(7), ELS(10)	K, D
E	S2	D00560	hsa1576	ERS(9)	K, D, C
		D00528	hsa5150	ERS(10)	D
		D02441	hsa762	EGS(9), ELS(7)	K, D
		D02441	hsa760	EGS(10), ELS(8)	K, D

Appendix

Table 13 Results of ensemble methods along with their embedded base models in S2 in terms of AUPR

Dataset	Base model	Base	ERS	EGS	ELS
NR	ALADIN	0.433(3)	0.433(3)	**0.446(1)**	0.433(3)
	BLMNII	0.441(2)	0.435(4)	0.438(3)	**0.445(1)**
	MLkNNSC	0.456(4)	**0.495(1)**	0.491(2)	0.481(3)
	NRLMF	**0.513(1)**	0.496(4)	0.506(2)	0.503(3)
	WkNN	0.51(3.5)	0.51(3.5)	0.515(2)	**0.516(1)**
	WkNNIR	0.539(2)	0.526(4)	**0.541(1)**	0.534(3)
GPCR	ALADIN	0.306(4)	**0.317(1)**	0.313(2)	0.311(3)
	BLMNII	0.342(2)	0.341(3)	0.34(4)	**0.344(1)**
	MLkNNSC	0.342(2.5)	**0.352(1)**	0.342(2.5)	0.341(4)
	NRLMF	0.345(4)	**0.358(1)**	0.35(2.5)	0.35(2.5)
	WkNN	0.369(2.5)	**0.37(1)**	0.364(4)	0.369(2.5)
	WkNNIR	**0.384(1)**	0.382(2)	0.377(4)	0.38(3)
IC	ALADIN	**0.35(1)**	0.34(4)	0.346(2.5)	0.346(2.5)
	BLMNII	**0.317(1.5)**	0.316(3.5)	0.316(3.5)	**0.317(1.5)**
	MLkNNSC	0.312(4)	**0.321(1)**	0.317(3)	0.318(2)
	NRLMF	0.343(4)	**0.352(1)**	0.347(3)	0.351(2)
	WkNN	0.354(2)	0.348(4)	**0.356(1)**	0.351(3)
	WkNNIR	**0.363(1)**	0.359(4)	0.361(2)	0.36(3)

Table 13 (continued)

Dataset	Base model	Base	ERS	EGS	ELS
E	ALADIN	0.289(2.5)	0.288(4)	**0.294(1)**	0.289(2.5)
	BLMNII	0.26(4)	0.262(3)	0.264(2)	**0.273(1)**
	MLkNNSC	0.34(4)	**0.353(1)**	0.348(2)	0.344(3)
	NRLMF	0.352(4)	0.365(2.5)	**0.367(1)**	0.365(2.5)
	WkNN	0.385(3.5)	0.385(3.5)	**0.388(1.5)**	**0.388(1.5)**
	WkNNIR	0.396(3)	0.395(4)	**0.397(1.5)**	**0.397(1.5)**
DB	ALADIN	0.41(4)	0.413(2)	0.412(3)	**0.414(1)**
	BLMNII	0.202(4)	**0.211(1)**	0.208(2)	0.203(3)
	MLkNNSC	0.366(4)	**0.378(1)**	0.372(3)	0.374(2)
	NRLMF	0.386(3)	0.375(4)	0.393(2)	**0.394(1)**
	WkNN	0.413(3)	0.394(4)	0.414(1.5)	**0.414(1.5)**
	WkNNIR	**0.425(2)**	0.414(4)	**0.425(2)**	**0.425(2)**

The parenthesis is the rank of each method among all competitors

Bold signify the best method(s)

Table 14 Results of ensemble methods along with their embedded base models in S3 in terms of AUPR

Dataset	Base Model	Base	ERS	EGS	ELS
NR	ALADIN	0.383(4)	**0.411(1)**	0.403(2)	0.393(3)
	BLMNII	0.447(3)	0.445(4)	**0.454(1)**	0.453(2)
	MLkNNSC	0.38(4)	**0.412(1)**	0.405(2)	0.402(3)
	NRLMF	0.471(3.5)	**0.48(1)**	0.471(3.5)	0.476(2)
	WkNN	0.443(4)	0.444(3)	0.45(2)	**0.451(1)**
	WkNNIR	0.461(4)	0.466(3)	0.477(2)	**0.479(1)**
GPCR	ALADIN	**0.517(1)**	0.511(3.5)	0.513(2)	0.511(3.5)
	BLMNII	**0.476(1)**	0.468(4)	0.473(2)	0.472(3)
	MLkNNSC	0.511(4)	**0.525(1)**	0.515(2)	0.514(3)
	NRLMF	0.518(3.5)	0.518(3.5)	**0.527(1.5)**	**0.527(1.5)**
	WkNN	0.541(2)	**0.543(1)**	0.539(4)	0.54(3)
	WkNNIR	0.577(2)	0.571(4)	0.574(3)	**0.579(1)**
IC	ALADIN	**0.803(2)**	0.802(4)	**0.803(2)**	**0.803(2)**
	BLMNII	**0.787(2)**	**0.787(2)**	0.785(4)	**0.787(2)**
	MLkNNSC	0.784(4)	0.791(3)	**0.795(1)**	0.792(2)
	NRLMF	0.798(4)	0.802(2)	0.801(3)	**0.806(1)**
	WkNN	0.789(4)	0.792(3)	0.793(2)	**0.794(1)**
	WkNNIR	0.798(4)	0.799(3)	**0.802(1)**	0.8(2)
E	ALADIN	0.758(4)	**0.761(2)**	**0.761(2)**	**0.761(2)**
	BLMNII	0.77(2)	0.769(3.5)	0.769(3.5)	**0.771(1)**
	MLkNNSC	0.748(4)	**0.757(1)**	0.752(2)	0.751(3)
	NRLMF	0.786(4)	0.791(2.5)	0.791(2.5)	**0.792(1)**
	WkNN	0.776(4)	**0.779(1.5)**	0.778(3)	**0.779(1.5)**
	WkNNIR	0.78(4)	**0.782(1.5)**	0.781(3)	**0.782(1.5)**
DB	ALADIN	0.569(4)	**0.581(1)**	0.58(2)	0.579(3)
	BLMNII	0.433(3)	**0.448(1)**	0.444(2)	0.408(4)
	MLkNNSC	0.579(4)	**0.594(1)**	0.588(2)	0.586(3)
	NRLMF	0.585(3)	0.575(4)	**0.596(1)**	0.594(2)
	WkNN	0.581(2.5)	**0.594(1)**	0.581(2.5)	0.579(4)
	WkNNIR	0.595(2)	0.584(4)	**0.597(1)**	0.594(3)

The parenthesis is the rank of each method among all competitors

Bold signify the best method(s)

Table 15 Results of ensemble methods along with their embedded base models in S4 in terms of AUPR

Dataset	Base Model	Base	ERS	EGS	ELS
NR	ALADIN	0.095(3)	**0.099(1)**	0.097(2)	0.094(4)
	BLMNII	0.135(3.5)	0.135(3.5)	0.142(2)	**0.143(1)**
	MLkNNSC	0.117(4)	0.121(3)	**0.131(1)**	0.123(2)
	NRLMF	0.142(4)	0.149(3)	0.165(2)	**0.166(1)**
	WkNN	0.159(4)	0.161(3)	**0.169(1.5)**	**0.169(1.5)**
	WkNNIR	0.165(3)	0.157(4)	**0.171(1.5)**	**0.171(1.5)**
GPCR	ALADIN	0.114(4)	0.116(3)	**0.121(1)**	0.12(2)
	BLMNII	0.121(3.5)	**0.122(1.5)**	0.121(3.5)	**0.122(1.5)**
	MLkNNSC	0.1(4)	**0.116(1)**	0.108(3)	0.109(2)
	NRLMF	0.134(3)	0.132(4)	0.135(2)	**0.136(1)**
	WkNN	0.149(3.5)	0.149(3.5)	**0.151(1.5)**	**0.151(1.5)**
	WkNNIR	0.158(2.5)	0.154(4)	0.158(2.5)	**0.159(1)**
IC	ALADIN	0.206(4)	0.208(3)	0.21(2)	**0.211(1)**
	BLMNII	0.176(3.5)	0.176(3.5)	0.177(2)	**0.178(1)**
	MLkNNSC	0.148(4)	**0.175(1)**	0.166(2)	0.165(3)
	NRLMF	0.215(4)	**0.223(1)**	0.217(3)	0.219(2)
	WkNN	0.216(3.5)	**0.22(1)**	0.216(3.5)	0.218(2)
	WkNNIR	0.226(4)	**0.229(1.5)**	0.227(3)	**0.229(1.5)**
E	ALADIN	0.128(4)	0.133(3)	**0.136(1)**	0.135(2)
	BLMNII	0.147(4)	**0.148(2)**	**0.148(2)**	**0.148(2)**
	MLkNNSC	0.146(4)	**0.158(1)**	0.15(3)	0.155(2)
	NRLMF	0.198(4)	0.209(2.5)	0.209(2.5)	**0.21(1)**
	WkNN	0.208(4)	**0.211(1.5)**	0.21(3)	**0.211(1.5)**
	WkNNIR	0.202(4)	**0.206(1.5)**	0.204(3)	**0.206(1.5)**
DB	ALADIN	0.248(4)	**0.263(1)**	0.262(2)	0.261(3)
	BLMNII	0.064(4)	**0.075(1)**	0.073(2.5)	0.073(2.5)
	MLkNNSC	0.205(4)	**0.229(1)**	0.224(2)	0.223(3)
	NRLMF	0.226(3)	0.203(4)	**0.23(1.5)**	**0.23(1.5)**
	WkNN	0.247(3)	0.226(4)	0.248(2)	**0.25(1)**
	WkNNIR	0.251(3)	0.249(4)	0.252(2)	**0.253(1)**

The parenthesis is the rank of each method among all competitors

Bold signify the best method(s)

Funding Bin Liu is supported from the China Scholarship Council (CSC) under the Grant CSC No.201708500095. This research is also supported by the Flemish Government (AI Research Program).

Availability of data and code The datasets and code used in this paper are available at https://github.com/intelligence-csd-auth-gr/DTI_WkNNIR.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Ezzat A, Wu M, Li XL, Kwoh CK (2018) Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform 20(4):1337–1357. https://doi.org/10.1093/bib/bby002
2. Chen R, Liu X, Jin S, Lin J, Liu J (2018) Machine learning for drug-target interaction prediction. Molecules 23(9):2208. https://doi.org/10.3390/molecules23092208
3. Dickson M, Gagnon JP (2004) Key factors in the rising cost of new drug discovery and development. Nat Rev Drug Discov 3(5):417–429. https://doi.org/10.1038/nrd1382
4. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve RD productivity: he pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9(3):203–214. https://doi.org/10.1038/nrd3078
5. Jacob L, Vert JP (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24(19):2149–2156. https://doi.org/10.1093/bioinformatics/btn409
6. Opella SJ (2013) Structure determination of membrane proteins by nuclear magnetic resonance spectroscopy. Ann Rev Anal Chem 6(1):305–328. https://doi.org/10.1146/annurev-anchem-062012-092631
7. Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Discov Today 23(8):1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010
8. Ding H, Takigawa I, Mamitsuka H, Zhu S (2013) Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Brief Bioinform 15(5):734–747. https://doi.org/10.1093/bib/bbt056
9. Liu Y, Wu M, Miao C, Zhao P, Li XL (2016) Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput Biol 12(2). https://doi.org/10.1371/journal.pcbi.1004760
10. Mei JP, Kwoh CK, Yang P, Li XL, Zheng J (2013) Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics 29(2):238–245. https://doi.org/10.1093/bioinformatics/bts670
11. van Laarhoven T, Marchiori E (2013) Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PLoS ONE 8(6). https://doi.org/10.1371/journal.pone.0066952
12. Liang Y, Xia LY, Yang ZY, Zhang H (2019) Improved prediction of drug-target interactions using self-paced learning with collaborative matrix factorization. J Chem Inf Model 59(7):3340–3351. https://doi.org/10.1021/acs.jcim.9b00408
13. Thafar MA, Thafar MA, Olayan RS, Olayan RS, Ashoor H, Ashoor H, Albaradei S, Albaradei S, Bajic VB, Gao X, Gojobori T, Gojobori T, Essack M (2020) DTiGEMS+: Drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. J Cheminfo 12(1):1–17. https://doi.org/10.1186/s13321-020-00447-2
14. Mohamed SK, Nováček V, Nounu A, Cowen L (2020) Discovering protein drug targets using knowledge graph embeddings. Bioinformatics 36(2):603–610. https://doi.org/10.1093/bioinformatics/btz600
15. Shi JY, Yiu SM, Li Y, Leung HC, Chin FY (2015) Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering. Methods 83:98–104. https://doi.org/10.1016/j.ymeth.2015.04.036
16. Ezzat A, Zhao P, Wu M, Li XL, Kwoh CK (2017) Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinfo 14(3):646–656. https://doi.org/10.1109/TCBB.2016.2530062
17. Buza K, Peska L (2017) ALADIN: a new approach for drug-target interaction prediction. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, Skopje. https://doi.org/10.1007/978-3-319-71246-8_20
18. Pliakos K, Vens C (2020) Drug-target interaction prediction with tree-ensemble learning and output space reconstruction. BMC Bioinfo 21(1):1V. https://doi.org/10.1186/s12859-020-3379-z
19. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24(13). https://doi.org/10.1093/bioinformatics/btn162
20. Shi JY, Li JX, Chen BL, Zhang Y (2018) Inferring interactions between novel drugs and novel targets via instance-neighborhood-based models. Curr Protein Peptide Sci 19(5):488–497. https://doi.org/10.2174/1389203718666161108093907
21. Liu B, Tsoumakas G (2019) Synthetic oversampling of multi-label data based on local label distribution. In: European conference on machine learning and principles and practice of knowledge discovery in databases (ECML-PKDD 19), Würzburg
22. Liu B, Pliakos K, Tsoumakas G et al (2020) Local imbalance based ensemble for predicting interactions between novel drugs and targets. In: PharML 2020 (Machine Learning for Pharma and Healthcare Applications), Location: online
23. Kurgan L, Wang C (2018) Survey of similarity-based prediction of drug-protein interactions. Curr Med Chem 25. https://doi.org/10.2174/0929867325666181101115314
24. Shi JY, Yiu SM (2015) SRP: a concise non-parametric similarity-rank-based model for predicting drug-target interactions. In: 2015 IEEE international conference on bioinformatics and biomedicine, pp 1636–1641. https://doi.org/10.1109/BIBM.2015.7359921
25. Zhang ML, Zhou ZH (2007) ML-KNN: A lazy learning approach to multi-label learning. Patt Recogn 40(7):2038–2048
26. Zhang W, Liu F, Luo L, Zhang J (2015) Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinfo 16(1):365. https://doi.org/10.1186/s12859-015-0774-y
27. Bleakley K, Yamanishi Y (2009) Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 25(18):2397–2403. https://doi.org/10.1093/bioinformatics/btp433
28. van Laarhoven T, Nabuurs SB, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics 27(21):3036–3043. https://doi.org/10.1093/bioinformatics/btr500
29. Buza K, Nanopoulos A, Nagy G (2015) Nearest neighbor regression in the presence of bad hubs. Knowl-Based Syst 86:250–260. https://doi.org/10.1016/j.knosys.2015.06.010
30. Schrynemackers M, Wehenkel L, Babu MM, Geurts P (2015) Classifying pairs with trees for supervised biological network inference. Molecular BioSystems 11(8):2116–2125. https://doi.org/10.1039/c5mb00174a
31. Stock M, Pahikkala T, Airola A, De Baets B, Waegeman W (2018) A comparative study of pairwise learning methods based

on Kernel ridge regression. Neural Comput 30(8):2245–2283. https://doi.org/10.1162/neco_a_01096

32. Gönen M (2012) Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics 28(18):2304–2310. https://doi.org/10.1093/bioinformatics/bts360
33. Cobanoglu MC, Liu C, Hu F, Oltvai ZN, Bahar I (2013) Predicting drug-target interactions using probabilistic matrix factorization. J Chem Inf Model 53(12):3399–3409. https://doi.org/10.1021/ci400219z
34. Zheng X, Ding H, Mamitsuka H, Zhu S (2013) Collaborative matrix factorization with multiple similarities for predicting drug-Target interactions. In: ACM international conference on knowledge discovery and data mining, pp 1025–1033. https://doi.org/10.1145/2487575.2487670
35. Hao M, Bryant SH, Wang Y (2017) Predicting drug-target interactions by dual-network integrated logistic matrix factorization. Scient Rep 7(1):1–11. https://doi.org/10.1038/srep40376
36. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y (2012) Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 8(5). https://doi.org/10.1371/journal.pcbi.1002503
37. Alaimo S, Pulvirenti A, Giugno R, Ferro A (2013) Drug-target interaction prediction through domain-tuned network-based inference. Bioinformatics 29(16):2004–2008. https://doi.org/10.1093/bioinformatics/btt307
38. Wang W, Yang S, Jing L (2013) Drug target predictions based on heterogeneous graph inference. In: Pacific symposium on biocomputing, pp 53–64
39. Chen X, Liu MX, Yan GY (2012) Drug-target interaction prediction by random walk on the heterogeneous network. Mol BioSyst 8(7):1970–1978. https://doi.org/10.1039/c2mb00002d
40. Fakhraei S, Huang B, Raschid L, Getoor L (2014) Network-based drug-target interaction prediction with probabilistic soft logic. IEEE/ACM Trans Comput Biol Bioinfo 11(5):775–787. https://doi.org/10.1109/TCBB.2014.2325031
41. Ba-Alawi W, Soufan O, Essack M, Kalnis P, Bajic VB (2016) DASPfind: new efficient method to predict drug-target interactions. J Cheminfo 8(1). https://doi.org/10.1186/s13321-016-0128-4
42. Pliakos K, Vens C (2019) Network inference with ensembles of bi-clustering trees. BMC Bioinforma 20(1):1–12. https://doi.org/10.1186/s12859-019-3104-y
43. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
44. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–24. https://doi.org/10.1007/s10994-6226-1
45. Pliakos K, Geurts P, Vens C (2018) Global multi-output decision trees for interaction prediction. Mach Learn 107(8-10):1257–1281. https://doi.org/10.1007/s10994-018-5700-x
46. Zheng Y, Peng H, Zhang X, Gao X, Li J (2018) Predicting drug targets from heterogeneous spaces using anchor graph hashing and ensemble learning. In: Proceedings of the international joint conference on neural networks. https://doi.org/10.1109/IJCNN.2018.8489028
47. Chen T, Guestrin C (2016) XGBOost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. https://doi.org/10.1145/2939672.2939785
48. Pliakos K, Vens C, Tsoumakas G (2019) Predicting drug-target interactions with multi-label classification and label partitioning. IEEE/ACM Trans Comput Biol Bioinfo. https://doi.org/10.1109/tcbb.2019.2951378
49. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8(1):1–13. https://doi.org/10.1038/s41467-017-00680-8
50. Olayan RS, Ashoor H, Bajic VB (2018) DDR: Efficient computational method to predict drug-Target interactions using graph mining and machine learning approaches. Bioinformatics 7(34):1164–1173. https://doi.org/10.1093/bioinformatics/btx731
51. Chu Y, Kaushik AC, Wang X, Wang W, Zhang Y, Shan X, Salahub DR, Xiong Y, Wei DQ (2019) DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. Brief Bioinfo. https://doi.org/10.1093/bib/bbz152
52. Bekker J, Davis J (2020) Learning from positive and unlabeled data: a survey. Mach Learn 109(4):719–760. https://doi.org/10.1007/s10994-020-05877-5
53. Shi JY, Li JX, Lu HM (2016) Predicting existing targets for new drugs base on strategies for missing interactions. BMC Bioinfo 17(Suppl 8):601–609. https://doi.org/10.1186/s12859-016-1118-2
54. Peng L, Zhu W, Liao B, Duan Y, Chen M, Chen Y, Yang J (2017) Screening drug-target interactions with positive-unlabeled learning. Sci Rep 7(1):1–17. https://doi.org/10.1038/s41598-017-08079-7
55. Lin C, Ni S, Liang Y, Zeng X, Liu X (2019) Learning to predict drug target interaction from missing not at random labels. IEEE Trans Nanobiosci 18(3):353–359. https://doi.org/10.1109/TNB.2019.2909293
56. Hattori M, Okuno Y, Goto S, Kanehisa M (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J Amer Chem Soc 125(39):11853–11865. https://doi.org/10.1021/ja036030u
57. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197. https://doi.org/10.1016/0022-2836(81)90087-5
58. Kuang Q, Xu X, Li R, Dong Y, Li Y, Huang Z, Li Y, Li M (2015) An eigenvalue transformation technique for predicting drug-target interaction. Sci Rep 5(1):1–9. https://doi.org/10.1038/srep13867
59. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, MacIejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2018) Drugbank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res 46(D1):D1074–D1082. https://doi.org/10.1093/nar/gkx1037
60. Schrynemackers M, Küffner R, Geurts P (2013) On protocols and measures for the validation of supervised methods for the inference of biological networks. Front Genet 4. https://doi.org/10.3389/fgene.2013.00262
61. Peska L, Buza K, Koller J (2017) Drug-target interaction prediction: A Bayesian ranking approach. Comput Methods Prog Biomed 152:15–21. https://doi.org/10.1016/j.cmpb.2017.09.003
62. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):D353–D361. https://doi.org/10.1093/nar/gkw1092
63. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR (2019) ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D940. https://doi.org/10.1093/nar/gky1075
64. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ,

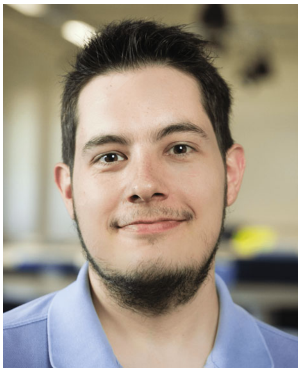
Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R (2008) SuperTarget and Matador: Resources for exploring drug-target relationships. Nucleic Acids Res 36(SUPPL. 1). https://doi.org/10.1093/nar/gkm862

65. Li L, Cai M (2019) Drug target prediction by multi-view low rank embedding. IEEE/ACM Trans Comput Biol Bioinfo 16(5):1712–1721. https://doi.org/10.1109/TCBB.2017.2706267

66. Ding Y, Tang J, Guo F (2020) Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion. Knowledge-Based Sys 106254. https://doi.org/10.1016/j.knosys.2020.106254

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bin Liu received an M.S. degree in computer science from Chongqing University of Posts and Telecommunications, China in 2016. He is currently pursuing a Ph.D. degree in computer science from Aristotle University of Thessaloniki, Greece. His research interests include multi-label learning, class imbalance classification and bioinformatics.



Konstantinos Pliakos is a Postdoc researcher at KU Leuven, Belgium. He received a diploma in Electrical and Computer engineering and a MSc in Computational Intelligence from AUTH in 2011 and 2015, respectively. He received a PhD from KU Leuven in 2019. His research interests include multi-label and multi- target prediction, supervised and semi-supervised learning, dimensionality reduction, recommender systems, biomedical network mining, and drug discovery.

Celine Vens is an associate professor at the Faculty of Medicine of KU Leuven, and itec, an imec research group at KU Leuven, Belgium. She obtained her PhD degree in computer science (machine learning) from the same university in 2007. Her research expertise focuses on multi-label, multi- target, hierarchical prediction, recommender systems, tree ensemble learning, survival analysis and biological network mining.

Grigorios Tsoumakas is an Associate Professor of Machine Learning and Knowledge Discovery at the School of Informatics of the Aristotle University of Thessaloniki (AUTH) in Greece. He received a degree in Computer Science from AUTH in 1999, an MSc in Artificial Intelligence from the University of Edinburgh, United Kingdom, in 2000 and a PhD in Computer Science from AUTH in 2005. His research expertise focuses on supervised learning techniques (ensemble methods, multi-target prediction) and natural language processing (semantic indexing, keyphrase extraction, summarization). He has published more than 100 research papers and according to Google Scholar he has more than 13,000 citations and an h-index of 44. Dr. Tsoumakas is a senior member of the ACM and an action editor of the Data Mining and Knowledge Discovery journal. His honors include receiving the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 10-Year Test of Time Award in 2017. He is an advocate of applied research that matters and has worked as a machine learning and data mining developer, researcher and consultant in several national, international and private sector funded R&D projects. In February 2019 he co-founded Medoid AI, a spin-off company of the Aristotle University of Thessaloniki developing AI products and custom solutions based on machine learning technology.