Predicting Combinative Drug Pairs via Multiple Classifier System with Positive Samples Only

Jian-Yu Shi , Jia-Xin Li , Kui-Tao Mao , Jiang-Bo Cao , Peng Lei , Hui-Meng Lu , Siu-Ming Yiu

PII: S0169-2607(18)30604-7
DOI: https://doi.org/10.1016/j.cmpb.2018.11.002
Reference: COMM 4812

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 1 May 2018
Revised date: 24 October 2018
Accepted date: 12 November 2018

Highlights

- Five heterogeneous features are extracted to characterize drugs and drug pairs.
- A two-layer MCS is designed to ensemble features for predicting drug combination.
- MCS consists of one-class SVMs and is trained by only approved drug combinations.
- Combining modes and targeting pathways of drugs in combination are investigated.

Predicting Combinative Drug Pairs via Multiple Classifier System with Positive Samples Only

Jian-Yu Shi [*1], Jia-Xin Li [1], Kui-Tao Mao[2], Jiang-Bo Cao [1], Peng Lei[3] , Hui-Meng Lu[1] and Siu-Ming Yiu[4]

1. jianyushi@nwpu.edu.cn, 279367149@qq.com, 791596353@qq.com, School of Life Science, Northwestern Polytechnical University, China

2. 284652735@qq.com, School of Computer Science, Northwestern Polytechnical University, China

3. leipengml@163.com, Department of Chinese Medicine, Shaanxi Provincial People's Hospital, China

4. smyiu@cs.hku.hk, Department of Computer Science, the University of Hong Kong, Hong Kong, China

* Corresponding authors: jianyushi@nwpu.edu.cn

Abstract

Background and Objective: Due to the synergistic effects of drugs, drug combination is one of the effective approaches for treating complex diseases. However, the identification of drug combinations by dose-response methods is still costly. It is promising to develop supervised learning-based approaches to predict potential drug combinations on a large scale. Nevertheless, these approaches have the inadequate utilization of heterogeneous features, which causes the loss of information useful to classification. Moreover, they have an intrinsic bias, because they assume unknown drug pairs as non-combinations, of which some could be real drug combinations in practice.

Methods: To address above issues, this work first designs a two-layer multiple classifier system (TLMCS) to effectively integrate heterogeneous features involving anatomical therapeutic chemical codes of drugs, drug-drug interactions, drug-target interactions, gene ontology of drug targets, and side effects. To avoid the bias caused by labelling unknown samples as negative, it then utilizes the one-class support vector machines, (which requires no negative instance and only labels approved drug combinations as positive instances), as the member classifiers in TLMCS. Last, both a 10-fold cross validation (10-CV) and a novel prediction are performed to validate the performance of TLMCS.

Results: The comparison with three state-of-the-art approaches under 10-CV exhibits the superiority of TLMCS, which achieves the area under the receiver operating characteristic curve = 0.824 and the area under the precision-recall curve = 0.372.

Moreover, the experiment under the novel prediction demonstrates its ability, where 9 out of the top-20 predicted combinative drug pairs are validated by checking the published literature. Furthermore, for each of the newly-validated drug combinations, this work analyses the combining mode of the member drugs and investigates their relationship in terms of drug targeting pathways.

Conclusions: The proposed TLMCS provides an effective framework to integrate those heterogeneous features and is trained by only positive samples such that the bias of taking unknown drug pairs as negative samples can be avoided. Furthermore, its results in the novel prediction reveal five types of drug combinations and three types of drug relationships in terms of pathways.

Keywords: drug combination, multiple classifier system, heterogeneous features, one-class classification

1 Introduction

The cause of complex diseases (e.g. cancer, HIV, cardiovascular disease) is always sophisticated and the corresponding treatment based on individual drugs bears very limited efficacy[1]. As a promising approach, drug combination has been widely exploited in treating complex diseases, due to its synergistic effect of drugs. The synergy of drugs means that the combined efficacy is greater than the sum of efficacies caused by individual drugs in the combination[2].For example, two individual drugs, *chlorpromazine* and *pentamidine*, show no antitumor activity, but their combination stops tumor growth more effectively than *paclitaxel*, which is a regular drug designed for controlling tumor growth[3].

The identification or quantification of synergistic drug combination majorly depends on dose-response methods[2]. However, existing dose-response methods can only work in expensive and limited conditions, and their results vary greatly even with the same set of drugs. Thus, it is impossible to screen potential combinations among a large scale of drugs for diverse diseases, because the number of possible combinations is extremely large. For example, considering the pairwise combinations among *m* drugs, the combinatorial number is equal to $m(m-1)/2$. As the number of approved drug combinations is increasing[4], computational methods (especially supervised learning-based approaches) have been proven as an inspiring approach to be a complementation to dose-response methods[2]. They are able to not only predict potential drug combination on a large scale but also provide a manner to reveal the underlying mechanism of drug combination[5-9].

Current computational approaches can be roughly categorized into two groups[10], disease-driven and drug-driven. Disease-driven approaches can infer potential combinations among multiple drugs by exploiting the disease-associated genes and targets in pathways, or protein interaction network[5, 6]. They are usually customized for specific diseases since they heavily depend on how well the diseases are understood. However, complex diseases are always poorly-understood. Besides, these approaches are not flexible enough to integrate other information (e.g. pharmacology or clinic phenotype) to enhance themselves.

Focusing on drugs rather than diseases, drug-driven approaches can infer potential combinations among a large scale of drugs. They characterize each drug pair as a feature vector, which captures various attributes of the drug pair[7-9], such as side effects and drug targets. They are helpful to understand complex diseases. Feature extraction and classifier model are two crucial factors for drug-driven approaches. However, existing drug-driven approaches are lacking of effective integration of heterogeneous features. For example, when k kinds of heterogeneous features are available, for each drug pair, some of the current approaches collapse from two original n-dimensional feature vectors of member drugs into a 1-dimensional similarity value (e.g. Jaccard Index) and assembles k feature-derived similarity values into a new k-dimensional feature vector[11]. Obviously, the collapse would cause the loss of information useful to classification. Other approaches simply concatenate the features into one high-dimensional vector, in which the number of entries is equal to the sum of those in the features[12]. However, such a naïve concatenation always leads to a time-consuming training and increases the risk of potential overfitting. More remarkably, all the existing drug-driven approaches assume the unknown drug pairs as negative samples, which possibly contain real drug combinations in fact. For among these approaches, this intrinsic bias is unavoidable.

To address these two issues, this work proposes a novel drug-driven approach of predicting potential pairwise combination among drugs. For the first issue, inspired by the successful cases of multiple classifiers in biological or medical applications[13], it designs a novel multiple classifier system (MCS) to obtain a better integration of heterogeneous features. For the second issue, it leverages a one-class classifier to obtain an unbiased training based on only positive samples (approved combinational drug pairs).

The remaining sections are organized as follows. Section Method presents the problem formulation of drug combination prediction, the extraction of five heterogeneous features, and a novel two-layer multiple classifier system (TLMCS), which ensembles those features by a set of support vector machines (SVM). Section Results provides the collection of both drug combinations and their heterogeneous features, the comparison with other state-of-the-art approaches under cross-validation by the implementation of TLMCS using regular binary SVMs, as well as the novel prediction of drug combination by the implementation of TLMCS using one-class SVMs. Finally, we elucidate the advantages of our TLMCS, its potential improvements and remaining issues in Section Discussion.

2 Materials and Methods

2.1 Overview

As recommended in a series of recent publications[14, 15], there are general guidelines for developing a biomedical or medical predictor according to machine learning[16]. These rules lead us to answer five ‘how to’ questions: (i) how to build a benchmark dataset containing enough samples; (ii) how to formulate the samples by an effective mathematical encoding, which reflects the intrinsic correlation between the samples and their targets (labels); (iii) how to develop or utilize a powerful algorithm based on the encoded samples to build the predictor; (iv) how to apply a cross-validation to evaluate the predictor’s performance in a fair manner; (v) how to deploy a publicly accessible and user-friendly web-server for the predictor.

In the following sections, we shall describe how to deal with the first four questions one-by-one when designing our TLMCS. In our future work, we will make efforts to develop an implementation of web-server for TLMCS.

2.2 Dataset

We collected the entries of drug combination from Drug Combination Database (DCDB)**[4]**. The original dataset involves 759 drugs, contains 945 approved combinative drug pairs, of which the combined therapeutic effects were validated to be better than any of their member drugs in clinics.

To extract drug features, we exploit three kinds of information sources of drugs[10], including drug properties, drug targets and clinical observations. Drug properties involve Anatomical Therapeutic Chemical (ATC) classification codes and drug-drug interactions (DDI). Drug targets contain drug-target interactions (DTI) and Go Ontology terms of drug targets (GO)[17]. Clinical observations are just side effects (SE). We collect ATC codes and DDI from DrugBank**[18]**, DTI and GO terms from DCDB[4] and UniProt[19], and assembled SE from both SIDER**[20]** and OFFSIDES**[21]**.

Out of 759 drugs, 596 have one or more ATC codes. Because a part of drugs has no ATC code available, we also applied the ATC predictor, SPACE**[22]**, to obtain predicted ATC codes for 48 drugs, which have no ATC code but have the structural notations of the Simplified Molecular-Input Line-Entry System (SMILES). In total, except for the drugs having no structural notation of SMILES, the ATC codes of 644 drugs were extracted.

Furthermore, to validate the TLMCS with heterogeneous features, we only picked the drugs having all of GO-, SE-, DDI-, DTI-, and ATC-based features and the drug pairs in which they participate. Ultimately, our datasets contain 378 drugs and 71,253 drug pairs, of which 298 are approved drug pairs (positive instances) and 70,955 are unknown pairs. After removing replicate and missing entries of features, we finally extracted the DDI-based feature involving 1,225 additional drugs, the DTI-based feature involving 472 targets, the GO-based feature involving 3,142 GO terms, the first SE-based feature involving 3,161 SIDER entries, and the second SE-based feature involving 8,799 OFFSIDE entries. Since there are common entries between SIDER and OFFSIDE, their concatenation (the final SE-based feature) contains 10,408 entries of side effects.

Unlike other works[8, 11], we didn’t calculate the pairwise similarities of drugs based on their chemical structures. The major reason is that chemical-structure features cannot discriminate synergistic drug pairs from other pairs. See also Section

3.2. In addition, a part of drugs attending drug combinations are not small molecular drugs, but biotech drugs, including peptide drugs and protein drugs. The chemical structure-based similarity is not appropriate when the dataset is a mixture of the usual small-molecular drugs, the biotech macro-molecular drugs and those small-molecular drugs having no benzene ring. For example, the combination (DCDB ID: DC005627, for treating *Suspected Heparin-Induced Thrombocytopenia*) consists a peptide drug *Desirudin* (DrugBank ID: DB11095, chemical formula: $C_{287}H_{440}N_{80}O_{110}S_6$) and a small molecule *Argatroban* (DrugBank ID: DB00278, chemical formula: $C_{23}H_{36}N_6O_5S$). Also, another combination (DCDB ID: DC001547, for treating *Follicular Lymphoma*) is totally comprised of two protein drugs (monoclonal antibodies), *Ibritumomab tiuxetan* (DrugBank ID: DB00078, chemical formula: $C_{6382}H_{9830}N_{1672}O_{1979}S_{54}$) and *Rituximab* (DrugBank ID: DB00073, chemical formula:$C_{6416}H_{9874}N_{1688}O_{1987}S_{44}$). In addition, some small-molecular drugs have no regular benzene ring in their chemical structure. For example, *Ferrous sulfate* (DrugBank ID: DB13257, Chemical formula: FeO_4S) can be jointly taken with a small-molecular drug *Folic acid* (Chemical formula: $C_{19}H_{19}N_7O_6$) to treat *Megaloblastic Anemia* (DCDB ID: DC007260), it, however, has no benzene ring and even no carbon atom. Obviously, their greatly varied sizes, as well as significantly different structural units, would results in illogical pairwise drug similarities based on chemical structures.

2.3 Problem formulation

Based on the assumption that synergistic/combinative drug pairs are similar to each other and different from antagonistic or ineffective drug pairs, we model the prediction of drug combination as a classification problem, by treating all pairs between drugs as the instances, and labelling combinative drug pairs as positive instances and other drug pairs as unknown instances (or negative instances in existing approaches). The crucial factors in classification are feature extraction and classifier.

Given a set of *m* drugs, denoted as $D=\{d_i\}, i=1,2,...,m$, of which each drug d_i is represented as an *n*-dimensional feature vector $\mathbf{f}_i=[f_{i,1}, f_{i,2},..., f_{i,n}]^T \in \mathbf{R}^{n\times 1}$. Let $P_{i,j}$ be the drug pair (d_i,d_j). Considering the symmetry that $P_{i,j}=P_{j,i}$, we define the feature vector $\mathbf{F}_{i,j}$ of $P_{i,j}$ as

$$\mathbf{F}_{i,j} = \mathbf{f}_i + \mathbf{f}_j \ . \tag{1}$$

The definition of the feature vector of a drug pair is also illustrated in Figure 1.

The feature vectors of labeled instances and their labels are utilized to train a model, named classifier. When given the feature vector $\mathbf{F}_{x,y}$ of an unlabelled pair $P_{x,y}$, the trained classifier discriminates how likely $P_{x,y}$ is a positive instance. In other words, the score $s_{x,y}$ generated by the classifier indicates the confidence of $P_{x,y}$ being a combinative drug pair.

2.4 Heterogeneous Features of Drugs

The corresponding ATC-, DDI-, DTI-, GO-, and SE-based features drug features are described as follows.

(1) According to the first level of ATC codes of drugs, the pairwise ATC-based drug similarities can be defined as,

$$S_{i,j}^{ATC} = \frac{\left| A_i \cap A_j \right|}{\left| A_i \cup A_j \right|} \tag{2}$$

where A_i is the set of the first-level ATC codes of d_i and $|\cdot|$ is the size of a set. All pairwise similarities are organized into a semantic similarity matrix $\mathbf{S}^{ATC}$. In the previous work[10], this similarity is directly taken as the features of drug pairs. However, this extraction assumes that drugs are totally independent to each other and ignores the underlying structure in the similarity matrix. That is to say, it collapses the information containing in the feature.

In this work, we think that there is a certain structural relationship among drugs, which can be captured by a graph. After regarding the ATC-based drug similarity matrix as the adjacent matrix of ATC-based drug-drug association graph, of which nodes are drugs and edges between nodes are their ATC associations, we can model the structural relationship among drugs. Therefore, to characterize drugs in such a graph, we apply Singular Value Decomposition on the ATC-based similarity matrix to extract drug features as follows:

$$\mathbf{f}^{ATC} = \mathbf{U}\sqrt{\boldsymbol{\Sigma}} \tag{3}$$

where $\mathbf{U}$ is the left singular-vector matrix and $\boldsymbol{\Sigma}$ is the diagonal singular-value matrix of $\mathbf{S}^{ATC}$. The *i*-th row of $\mathbf{f}^{ATC}$ denotes the ATC-based feature vector of d_i , which reflects the diverse topological features of drug nodes in the ATC-based graph.

(2) For DDI, DTI, GO terms, or SE, drug features can be calculated in a common way that counts the occurrence of a list of unique entries. Thus, a drug can be generally represented as a binary occurrence vector, of which each element denotes an entry in the list occurs or not. The entry names vary with different features. For example, they are drug names in DDI, Uniprot identities of targets in DTI, term identities of drug targets in GO and clinical names in SE. These feature are briefly described as follows.

The interactions between the drugs in D and *s* different drugs in another drug set $D_s = \{d_p\}$ are used to define the DDI-based feature, $\mathbf{f}_i^{DDI} = \left[f_{i,1}, f_{i,2}, ..., f_{i,s} \right]$, where $f_{i,p} = 1$ if drug $d_i \in D$ interacts with drug $d_p \in D_s$ or $f_{i,p} = 0$ if not.

Likewise, denote *v* unique targets interacting with *m* drugs as $T = \{t_1, t_2, ..., t_v\}$. The target profile of d_i is directly defined as its feature vector, $\mathbf{f}_i^{DTI} = \left[f_{i,1}, f_{i,2}, ..., f_{i,v} \right]$, where $f_{i,p} = 1$ if drug d_i interacts with target t_p or $f_{i,p} = 0$ if not.

For proteins, GO defines a set of terms describing gene functions and the relationships between these terms**[17]**. Each target of a drug may have one or more GO terms. Denote *w* unique terms involved in *v* targets as $GO = \{g_1, g_2, ..., g_w\}$. The GO terms of the targets of d_i are directly used to define its GO-based feature vector as, $\mathbf{f}_i^{GO} = \left[f_{i,1}, f_{i,2}, ..., f_{i,w} \right]$, where $f_{i,p} = 1$ if the targets of d_i has term g_p, or $f_{i,p} = 0$ if not.

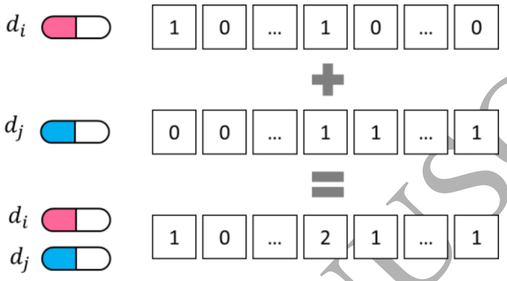
The SE-based feature extracts the entries of side effects from both SIDER**[20]** and OFFSIDES**[21]**. Two sets of side effects are concatenated into an enhanced one, which contains a set of unique entries of side effect. The SE-based feature vector of d_i is defined as $\mathbf{f}_i^{SE} = \left[f_{i,1}, f_{i,2}, ..., f_{i,u} \right]$, where $f_{i,p} = 1$ if the *p*-th entry of side effects is observed while taking the therapy of drug d_i , $f_{i,p} = 0$ otherwise.

An example is shown to illustrate the GO-based feature extraction in Figure 1. Suppose that a dataset contains two drugs, *Ethambutol* (labeled by DCC0056 in DCDB[4]) and *Pyrazinamide* (DCC0062). *Ethambutol* has 11 GO terms of its targets and *Pyrazinamide* has 7 as follows.

- *Ethambutol* has: GO:0005576; GO:0005618; GO:0005829; GO:0005886; GO:0005887; GO:0009247; GO:0016021; GO:0040007; GO:0046677; GO:0052636; GO:0071766.

- *Pyrazinamide* has: GO:0004318; GO:0005618; GO:0005829; GO:0005835; GO:0005886; GO:0006633; GO:0040007.

Since sharing 4 GO terms, they can be coded into two 14-d vectors, of which the dimension is the number of unique GO terms. After sorting them, we generate a list of unique GO terms as: (1) GO:0004318; (2) GO:0005576; (3) GO:0005618; (4) GO:0005829; (5) GO:0005835; (6) GO:0005886; (7) GO:0005887; (8) GO:0006633; (9) GO:0009247; (10) GO:0016021; (11) GO:0040007; (12) GO:0046677; (13) GO:0052636; (14) GO:0071766. The following figure shows their feature vectors and the feature vector of their pair (see also Formula 1).



GO Feature	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14
Ethambutol	0	1	1	1	0	1	1	0	1	1	1	1	1	1
Pyrazinamide	1	0	1	1	1	1	0	1	0	0	1	0	0	0
Drug Pair	1	1	**2**	**2**	1	**2**	1	1	1	1	**2**	1	1	1

Fig. 1 - Illustration of features of a drug pair.

2.5 Multiple Classifier System

The proposed drug features not only derive from heterogeneous information sources, but also they have distinct forms in terms of calculation. In details, the entries of $\mathbf{f}^{DDI}$, $\mathbf{f}^{DTI}$, $\mathbf{f}^{SE}$ and $\mathbf{f}^{GO}$ are of binary, sparse, and high-dimensional features, and $\mathbf{f}^{ATC}$ is a real-valued feature. To exploit all these features, the integrated feature vector can be obtained by directly concatenating them sequentially. However, due to its higher dimension (i.e. is significantly greater than the number of samples), the concatenated feature definitely causes the time-consuming training of classifier and increases the risk of overfitting, which leads to a poor generalization on newly coming instances.

In order to address such issues, we design a two-layer multiple classifier system (TLMCS), of which the first layer handles heterogeneous features parallelly by a set of member classifiers and the second one having only one classifier integrates the outputs of the former layer to generate the final result (Fig. 2). Compared with

Boosting or Bagging, which provide a serial ensemble architecture of multiple classifiers, TLMCS defines a parallel architecture of member classifiers, which can be implemented by parallel processors to achieve real-time performance. The theoretical analysis of MCS can be found in [23, 24].

Generally, for the given testing drug pair $P_{x,y}$, its N feature vectors $\{\mathbf{F}_{x,y}^{(k)}\}$ are firstly calculated and input into a set of member classifiers $\{C_k^1\}, k=1,...,N,$ in the first layer of our TLMCS simultaneously. Next, corresponding N feature vectors, the scores $\{s_{x,y}^{(k)}\}$ of $P_{x,y}$ being a positive instance are output by $\{C_k^1\}$. After that, these output scores are sequentially integrated into a new vector $\mathbf{G}_{x,y}$ which is regarded as the input feature vector of $P_{x,y}$ for the only classifier C^2 in the second layer. Lastly, the probabilistic output of C^2 is taken as the final score $s_{x,y}^{final}$ of $P_{x,y}$, which indicates how likely d_x and d_y can be a combinative drug pair.

Diverse classifiers (e.g. logistical regression and support vector machines) have been popularly applied in many areas of bioinformatics, such as combinative drug prediction[8], rare disease variants analysis[25], disease-gene identification[26], gene classification[27] and protein domain classification[28]. We choose Support Vector Machines (SVM) as the member classifiers in our TLMCS, because SVM is able to cope with not only the regular binary classification but also the one-class classification, which trains the classifier with positive instances only [29].

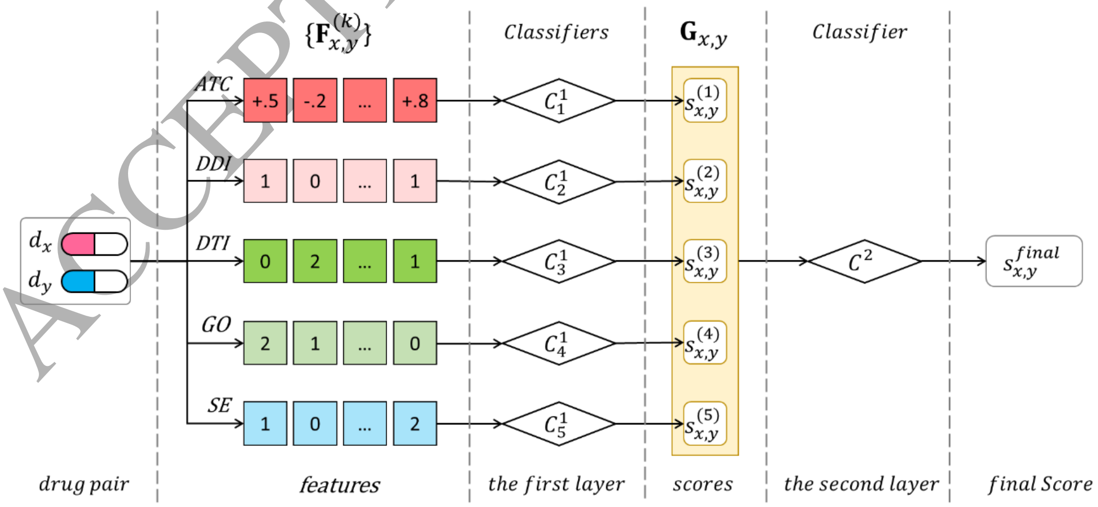


Fig. 2 - The architecture of multiple classifier system.

In the context of predicting drug combination, approved combinative drug pairs are labelled as positive instances. Remarkably, unknown drug pairs are regarded as negative instances when compared with other approaches under cross validation, while they are just unknown instances when applied to the prediction in a real application. Therefore, we provide one implementation of TLMCS by binary SVMs for binary classification, which was adopted by former approaches, as well as another implementation by one-class SVMs for novel prediction in the real application, where we only know approved combinative drugs[4].

2.6 Support Vector Machines as Member Classifiers

Given a set of training instances $\left\{(\mathbf{x}_i, y_i): i = 1...,M\right\}$ where $\mathbf{x}_i$ is the *i*-th instance (feature vector) and $y_i \in \{-1,1\}$ is its class label, predicting potential combinative drugs is modelled as a classification problem. The label denotes a positive instance if $y_i = 1$, otherwise a negative instance in the scenario of binary classification or an unknown instance in the scenario of one-class classification. The binary SVM and the one-class SVM were designed for the two scenarios respectively [29, 30] and their brief introduction as follows.

A binary SVM constructs discriminant function by solving the following primal optimization problem[30]

$$\begin{array}{ll} \min\limits_{w,b,\xi} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{M}\xi_i \\ s.t. & y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1-\xi_i \\ & \xi_i \geq 0, i = 1,...,M \end{array} \tag{4}$$

where $\phi(\mathbf{x}_i)$ maps $\mathbf{x}_i$ into a higher-dimensional space and C >0 is the regularization parameter. Due to the possible high dimensionality of the vector variable **w**, we usually turn the above problem to its dual problem to solve,

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha} \quad , \\ s.t. & \mathbf{y}^T\boldsymbol{\alpha} \\ & 0 \leq \alpha_i \leq C, i = 1,...M \end{array} \tag{5}$$

where **Q** is an M by M positive semidefinite matrix, $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel function. According to the primal-dual relationship, the optimal **w** satisfies $w = \sum_{i=1}^{M} y_i\alpha_i\phi(\mathbf{x}_i)$. Thus, for a given testing instance x and its feature vector $\mathbf{f}_x$, its confidence score of being a positive instance is defined as

$$s(\mathbf{x})=\sum_{i=1}^{M} y_i\alpha_i \,\mathrm{K}(\mathbf{x}_i,\mathbf{x})+b\,. \tag{6}$$

In a one-class SVM, the support vector model is trained on data that has only one class, which is called “desired” class or positive class. It infers the properties of desired instances from these properties and can predict where unknown instances are like or unlike the desire examples. One-class SVM is especially appropriate for the predicting task, which gives a few positive instances and many unlabelled instances (e.g. prediction of drug combination).

Technically, the one-class SVM estimates the support of a high-dimensional distribution of desired instances[29]. Compared to the regular binary SVM which learns the optimal separating hyperplane, it learns a sphere with the minimal volume containing all the desired instances. It solves the following primal optimization problem

$$\begin{array}{ll} \min\limits_{w,\xi,\rho} & \frac{1}{2}\mathbf{w}^T\mathbf{w}-\rho+\frac{1}{\upsilon M}\sum_{i=1}^{M}\xi_i\,. \\ s.t. & \mathbf{w}^T\phi(\mathbf{x}_i)\geq\rho-\xi_i \\ & \xi_i\geq 0,\upsilon\geq 0, i=1,\ldots,M \end{array} \tag{7}$$

Similarly, its dual problem is

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \\ s.t. & \mathbf{1}^T\boldsymbol{\alpha}=1, \\ & 0\leq\alpha_i\leq\frac{1}{\upsilon M}, i=1,\ldots M \end{array}\,. \tag{8}$$

Thus, for the given testing instance *x* and its feature vector $\mathbf{f}_x$, its confidence score of being a positive/desired instance is defined as

$$s(\mathbf{x})=\sum_{i=1}^{M}\alpha_i \,\mathrm{K}(\mathbf{x}_i,\mathbf{x})-\rho\,. \tag{9}$$

2.7 Assessment

The subsampling-based test, e.g. k-fold cross-validation (k-CV) and its extreme case jackknife test, is often used for evaluating the performance of a statistical predictive method. For a given benchmark dataset, the predictor under k-CV generates slightly different predictions with random samplings each time, whereas it surely yields a unique prediction under the jackknife test [16]. Due to the uniqueness of the jackknife test, it has been increasingly recognized and widely adopted by investigators when testing the power of various predictive methods[31]. However, the computational cost of the jackknife test is always greatly high. Because the results achieved by k-CV and the jackknife test respectively are not significantly different in the case of enough samples, we adopt k-CV to evaluate the performance of our TLMCN in the experiment.

Take 10-CV as an example (Fig.3). First, the whole dataset is randomly split into 10 subsets of approximately equal size. Secondly, the whole 10-CV is performed in 10 rounds. In each round, one subset is taken as the testing set while nine other subsets are merged together as the training set. Remarkably, there is no overlapping between the training set and the testing set in any round of 10-CV. In other words, the testing set acts as an independent dataset in each round of CV. The classifier's performance in each round is measured. All the rounds have different testing sets so as to take each subset as the testing set in turn. Finally, the cross-validation averages the measures of prediction in all the rounds to derive a more accurate estimate of the classifier's predictive performance. Usually, the whole procedure of 10-CV is repeated n times (i.e. n=20) under different random seeds to obtain a robust estimation of the classifier's performance.

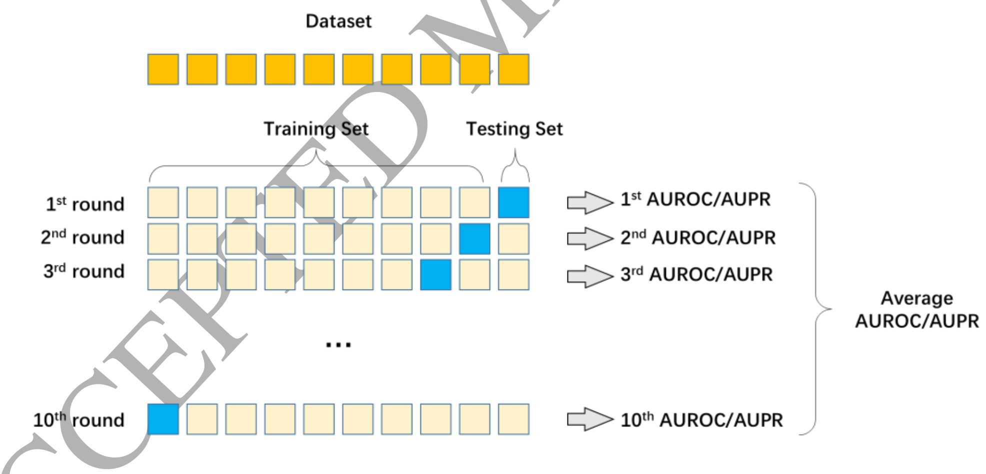


Fig. 3 - Illustration of 10-fold cross-validation.

Generally, there are two groups of metrics to measure the performance of predictive approaches. Working for the single-label classification, the first group usually contains Accuracy, Sensitivity, Specificity, and Matthews correlation coefficient. All of them are derived from the numbers of true positives, true negatives, false positives and false negatives. Developed for the multi-label classification, the second group usually contains Accuracy, Aiming, Absolute-True, Absolute-False, and Coverage. All the metrics in this group are determined by the relationship between the set of the real labels and the set of the predicted labels[32]. Our problem is just a case of the single-label classification, in which a sample is a drug pair and has only one label.

We pay more attention to the illustrate the predictive ability of our TLMCS as its discrimination threshold is varied, whereas the first group of metrics cannot meet the demand. Therefore, we leverage both the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve to make such an illustration and adopt the areas under these two curves (denoted as AUROC and AUPR respectively) to measure the performance of our TLMCS.

3 Results

3.1 Comparison with state-of-the-art approaches

The effectiveness of our TLMCS was first validated by comparing it with two state-of-the-art approaches**[8, 11]**. The former one utilizes only the SE-based features**[8]**. In contrast, the latter**[11]** simply collapses the original *n*-dimension feature vectors of drug pairs into 1-dimension similarity values and assembles all the feature-derived similarity values into new *k*-dimension feature vectors, where *k* is the number of feature types (k=5 here). In addition, we compared TLMCS with the naïve approach, which concatenates five kinds of feature vectors into a long feature vector of 15,625 dimensions.

As the previous works recommend[8],[10],[33], ten-fold cross-validation (10-CV) on the drug pairs were performed to assess the predicting approach. The performance of 10-CV was measured by both the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR)[34, 35]. To obtain a robust result, the whole procedure was repeated 20 times under different random seeds. The final performance was reported by the averages of both AUROC and AUPR.

We used LibSVM[36] as the implementation of SVM and selected the Radial Basis Function (RBF) $K(\mathbf{f}_j,\mathbf{f}_i)=\exp(-\gamma\|\mathbf{f}_j-\mathbf{f}_i\|^2)$ as its kernel function. LibSVM provides both regular and one-class implementations of SVM. The regular binary SVM was adopted when running all the approaches in this section, while the one-class SVM was used in the next section.

To achieve the best values of both the penalty *C* of regularization and the shape factor γ of RBF in SVM, we performed a grid search, which is expended by $C=\{0.1,1,10,100,1000\}$ and $\gamma=\{0.0001,0.001,\ 0.01,0.1,1\}$, with an inner 10-CV[37]. The best pairs of values for the state-of-the-art approaches[8, 11] and the naïve concatenation approach, are (C=10, γ=0.001) , (C=1, γ=0.001) , and (C=10, γ=0.0001) . In TLMCS, for ATC, DDI, DTI, GO and SE features, the best values of *C* in the first layer are 1, 100, 1, 100 and 10 respectively, the best value of *C* is 1 in the second layer, and γ=0.001 for the SVMs in both layers. The comparison of prediction performance is illustrated in Fig. 3.

According to the comparison, we draw several significant observations as follows. First, more features are needed and their integration can improve the prediction. For example, compared with the approach presented by Huang et al. which only applied the SE-based feature[8], all the other approaches integrating heterogeneous features achieve better prediction.

Secondly, different integrating manners of diverse features generate varied prediction performances. In details, the naïve concatenation is better than Cheng's approach[11]. The underlying reason is that the simple collapse of the heterogeneous feature vectors in Cheng's approach would cause the loss of information useful to classification, while the naïve concatenation uses the original features. Thus, compared with Cheng's approach, the naïve concatenation achieves the better prediction. However, the feature concatenation causes the very high-dimensional feature vectors, so as to induce the time-consuming training and the possible over-fitting issue.

Finally, our TLMCS wins the best and is significantly superior to other approaches with AUROC=0.824 and AUPR=0.372. It allocates each type of features to each classifier in its first layer without collapsing feature vectors, then aggregates the classified results generated by those classifiers as the final small k-dimensional feature vectors. Obviously, TLMCS utilizes the feature vectors as fully as possible in its first layer and has no need to train the classifier model with the highly-dimensional concatenation of heterogeneous features.

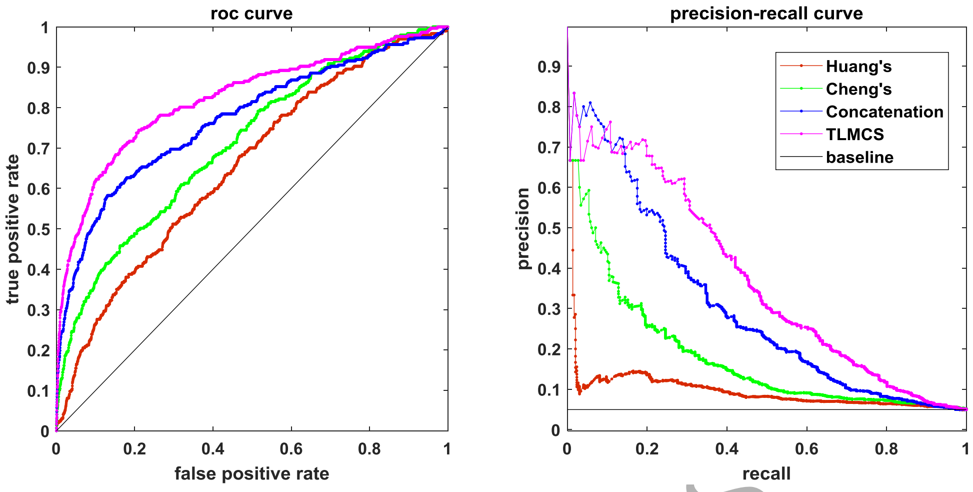


Fig. 4 - Comparison of state-of-the-art approaches under fair sampling. The values of AUROC (in the left panel) achieved by Huang et. al [8], Cheng and Zhao[11], the Concatenation and our TLMCS are 0.656, 0.713, 0.771, and **0.824**, while the values of AUPR (in the right panel) achieved by them are 0.108, 0.238, 0.313, and **0.372** respectively.

3.2 Novel Prediction with Positive Samples Only

The former approaches treat the unknown drug pairs as negative samples, of which some samples however could be drug combinations in a real application. More remarkably, the dataset we adopted contains no negative sample. Therefore, modelling the prediction of drug combinations as a binary classification has an intrinsic bias in clinics. In this case, modelling the problem as a one-class classification, which uses only positive samples to train the classifier, is more appropriate to infer synergistic drugs. Technically, we trained one-class SVM with positive samples only to perform novel prediction to discover potential combinative drug pairs among unknown drug pairs.

The novel experiment contained the following steps. First, we adopted one-class SVM as the classifiers in our TLMCS and trained them only with approved drug combinations and without negative samples. Then, by running the trained TLMCS on all the unknown drug pairs, we ranked them according to their decision scores generated by the TLMCS. Last, we output the top-100 unknown drug pairs (see also Supplementary 'SI_NovelPrediction.xlsx') and validated the first 20 pairs by checking published literature and searching the internet manually.

Table 1. Novel prediction and Validation

Rank	Drug 1	Drug 2	PubMed	Chemical-Structure Similarity	Remarks
1	**Hydroxyurea** (DB01005)	**Posaconazole** (DB01263)	PMID: 24403304	0.0318	Assistance
2	Cetirizine (DB00341)	Epinephrine (DB00668)	PMID: 27999602 PMID: 15617665 PMID: 12113226 PMID: 24335343	0.3133	Assistance & Clinical Suggestion[1]
3	Hydroxyurea (DB01005)	Hydralazine (DB01275)	PMID: 9259423	0.0562	Analogue
5	Bupivacaine (DB00297)	Pregabalin (DB00230)	PMID: 25940854	0.1949	**Synergistic Partner (in two drugs)**
7	Cetirizine (DB00341)	Tobramycin (DB00684)	PMID: 17102679	0.2264	**Synergistic Partner (in six drugs)**
14	**Topiramate** (DB00273)	**Dextromethorphan** (DB00514)	PMID: 26441400	0.2338	**Synergistic Partner (in two drugs)**
18	**Olmesartan** (DB00275)	**Celecoxib** (DB00482)	PMID: 16939632	0.2614	Assistance
19	Levetiracetam (DB01202)	Ceftriaxone (DB01212)	PMID: 24071615	0.1388	**Synergistic Partner (in two drugs)**
20	Dobutamine (DB00841)	Chlorothiazide (DB00880)	PMID: 23115539	0.1704	Compatible

Remarkably, 9 out of top-20 predicted drug combinations are validated (Table 1). According to how the drugs in a drug pair of interest work together, these drug pairs are grouped into five cases as follows:

[1] https://medlineplus.gov/druginfo/meds/a698026.html

(1) 'Synergistic Partner', which denotes the explicit evidence of enhancing therapeutic effect achieved by a drug pair or multiple drugs containing the drug pair;

(2) 'Assistance', which indicates the drug pair attend at the different phases in a long-term treatment or work as auxiliaries in a complex treatment;

(3) 'Analogue', which denotes the drugs in the pair of interest have similar therapeutic effects and can be the alternative to each other;

(4) 'Compatible', which means the drugs in the pair of interest can be compatibly combined with another drug respectively.

(5) 'Clinical suggestion', which denotes a clinical suggestion of being a drug combination, given by experienced doctors. For example, the pair of '*Cetirizine*' and '*Epinephrine*'.

We examined whether or not drugs' chemical structures are important to drug combination. First, by applying the fingerprint in Chemistry Development Kit (CDK) to represent the chemical structure of each small-molecular drug into a 1024-dimensional feature vector, we calculated chemical-structure similarities between drugs by Jaccard Index. Then, we checked into the training set. Two groups of similarities corresponding to combinative drug pairs (positive class) and other drug pairs (negative class) were checked respectively. As shown in Fig. 5-a, the distributions of these two groups of similarities have no significant difference because their distribution shapes are similar and their average similarities are close (0.2598 and 0.2417 respectively). In addition, we calculated the average occurrence of each entry of CDK fingerprint for two classes respectively and found a greatly high correlation (r=0.9928) between two sets of average occurrences (Fig. 5-b). This finding shows that two classes of drug pairs in the training set have no significance over all the fingerprint entries on average. Last, we checked into the nine newly-detected drug pairs in the novel prediction (Table 1). All the chemical-structure similarities of these drug pairs are very small (<0.32). Especially, four pairs of synergistic drug combinations have the similarity values varying from 0.1388 to 0.2264. These results demonstrate that drug chemical structures are trivial when discriminating synergistic drug combinations from drug pairs.

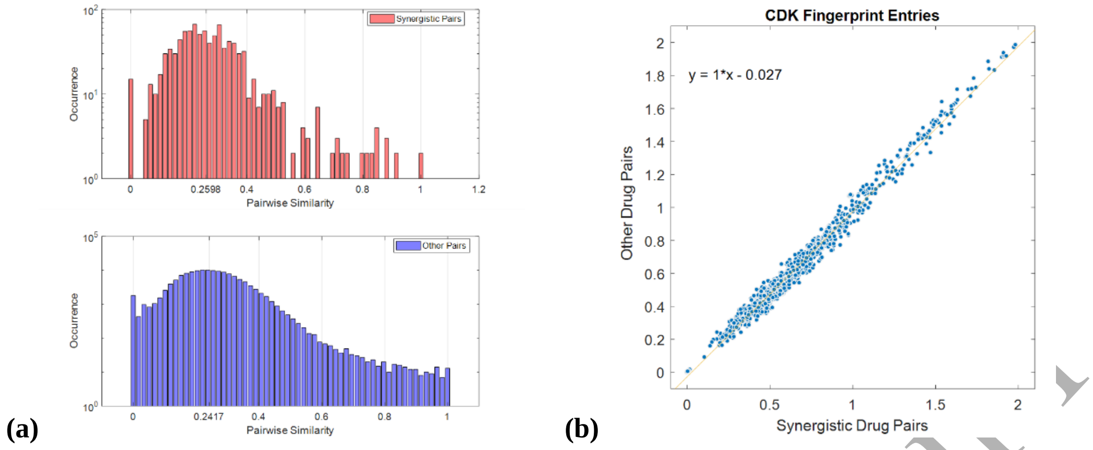


Fig. 5 – Analysis on drugs' chemical structures by CDK fingerprint. (a) Two groups of similarities corresponding to synergistic drug pairs (positive class) and other drug pairs (negative class); (b) the distribution of the average occurrence of each entry of CDK fingerprint w.r.t two classes.

Furthermore, to dig out why the nine manually validated drug pairs can be considered as drug combinations, we searched the genetic pathways for their drug components in KEGG. Five out of them have only one drug component having pathways, one pair out of them have no pathway, and three pairs have both drug components having pathways in KEGG. The full information about pathways of the predicted drug pairs can be found in Supplement "SI_Pathway.docx". Those drug pairs, in which both of drug components have the pathways found in KEGG, were analysed further (Table 2). We defined the relationship of those pathways by the class names of pathway provided by KEGG BRITE hierarchy. In detail, if some of the pathways targeted by two drugs are same (e.g. *Topiramate* and *Dextromethorphan*), their relationship is identified as 'Same'. If their pathways are different but share the first two class levels in terms of BRITE hierarchy, the drug pairs are considered as 'Related' (e.g. *Olmesartan* and *Celecoxib*). If their pathways are different but only share the first class level, they are regarded as 'Cross-talking' (e.g. *Hydroxyurea* and *Posaconazole*). This investigation reveals that if a drug pair is possibly a drug combination, then its drug components could have same, related or cross-talking pathways.

In summary, the validation results in both Table 1 and Table 2 demonstrates the ability of our approach to screen potential drug combinations.

Table 2. Relevant Pathways in Predicted Drug Combinations

Rank	Drug Pair	KEGG Pathway(the first two levels)	Relationship
1	Hydroxyurea (KEGG: D00341)	**hsa00230**: **Metabolism**[2]-> Nucleotide metabolism **hsa00240**: **Metabolism**-> Nucleotide metabolism **hsa00480**: **Metabolism**-> Metabolism of other amino acids hsa04115: Cellular Processes-> Cell growth and death	Cross-talking
	Posaconazole (KEGG: D02555)	**ko00100**: **Metabolism**-> Lipid metabolism	
14	Topiramate (KEGG: D00537)	**hsa04020: Environmental Information Processing-> Signal transduction** **hsa04080: Environmental Information Processing-> Signaling molecules and interaction** **hsa04720: Organismal Systems-> Circulatory system** **hsa04724: Organismal Systems-> Nervous system** hsa04727: Organismal Systems-> Nervous system hsa04730: Organismal Systems-> Nervous system	Same
	Dextromethorphan (KEGG: D03742)	**hsa04020: Environmental Information Processing-> Signal transduction** **hsa04080: Environmental Information Processing-> Signalling molecules and interaction** **hsa04720: Organismal Systems-> Circulatory system** **hsa04724: Organismal Systems-> Nervous system**	
18	Olmesartan (KEGG: D01204)	**hsa04020**: **Environmental Information Processing-> Signal transduction** **hsa04080**: **Environmental Information Processing-> Signalling molecules and interaction** hsa04270: Organismal Systems-> Circulatory system hsa04614: Organismal Systems-> Endocrine system	Related
	Celecoxib (KEGG: D00567)	hsa00590: Metabolism-> Lipid metabolism **hsa04370**: **Environmental Information Processing-> Signal transduction**	

4 Discussion

4.1 Our contributions

Predicting drug combination remains a challenge. In this paper, based on supervised learning, we have proposed a novel drug-driven approach to predict potential pairwise combinations of drugs on a large scale. This work focused on three aspects.

First, to distinguish synergistic drug pairs better, we have extracted five kinds of heterogeneous features, including ATC-, DDI-, DTI-, GO-, and SE-based features. Then, to achieve a better utilization of a set of heterogeneous features, we have designed a novel two-layer multiple classifier system

[2] Bold texts in table denote the matched pathways

(TLMCS). Owing to a faster training as well as a well-fitted learning model, the elaborate TLMCS is able to provide an effective framework to integrate those heterogeneous features. After that, the superiority of our TLMCS implemented by regular binary SVMs has been demonstrated by comparing with two state-of-the-art approaches as well as the naïve feature concatenation.

More importantly, the predicting power of our TLMCS implemented by one-class SVMs, which are trained by positive samples only and overcome the intrinsic bias of the former approaches, has been further demonstrated in a real application by the novel prediction validated by checking literature manually. Finally, for each of the validated novel drug combinations, we've investigated the mode of the member drugs working together and investigated their relationship in terms of drug targeting pathways. Both these combining modes and the pathway relationships between combinative drugs are helpful to uncover the underlying mechanism of how synergistic drug combinations generate.

4.2 Future technical improvement

So far, we have only implemented the first four steps of building a biomedical predictor. For current predictive approaches, people tend to develop their versions of web servers, which have a user-friendly interface and can be accessed in public[38]. The tendency contributes many computational tools to academic or industrial communities[39] [38, 40-44]. Nowadays, in the form of web-server, many predictive approaches have significantly enhanced their impacts on diverse areas, especially on medical science [45] and medicinal chemistry [46]. As a result, we will make efforts to provide a web-server for the proposed TLMCS in our future work.

On the other side, we will focus on improving the features of drug pairs and the predictive model. In the feature improvement, being aware of the fact that there are missing entries of drug targets, drug-drug interactions, GO terms and side effects in the original database[4], we plan to leverage both predictive approaches (e.g. for DTI [47-50], DDI [51-53], GO annotation [54], side effects[55-57]) and web-servers (e.g. for DDI [38], ATC [39] and DTI [40-44]) to dig out more missing entries. Moreover, we shall integrate both pathway dynamics and drug response to model the mechanism of action for the identified synergistic drug pairs.

In the model improvement, we plan to design a better TLMCS, which optimally accommodates more heterogeneous features, by analysing the dominant entries and the diversity of heterogeneous features. Furthermore, we will integrate other proteins, especially upstream and downstream proteins of the drug targets in pathways and other drug-binding proteins, and will perform a systematic analysis based on a specific protein-protein interaction network to provide a better elucidation of why and how two or more drugs can generate synergistic effects.

4.3 Issues about dataset

In this section, we highlight several issues about building a benchmark dataset of drug combinations.

Note that the benchmark dataset used contains no negative sample because medical researchers or clinical doctors always record or publish successful cases (positive samples). When comparing TLMCS with those state-of-the-art approaches, which are all supervised approaches, we made a compromise by regarding the unlabelled samples as negative samples even though a few of them are possibly positive samples. The training with such a set of labelling samples would result in biased classifiers/predictors, but the bias is common to all the approaches. In the field of machine learning, one usually supposes that the samples are independent and identically distributed (i.i.d.). Upon such an assumption, the results generated by an independent validation set and the cross-validation have no significant difference. Thus, the results generated by cross-validation can reflect the superiority of our approach.

Absolutely, it is important to build a new validation set without overlapping with the training set because the samples cannot be perfectly i.i.d. We're struggling for it as well as finding explicitly labeled negative samples, and will make efforts to build more datasets in our future work. However, the task is very difficult. For a drug pair, we cannot say easily whether it has a synergistic effect, because the determination of synergistic drugs needs enough pieces of disease-specific and dose-specific clinical evidence. FDA orange book shows that the combination of Aspirin and Caffeine showing its synergistic effect on different diseases should require different drug doses respectively (e.g. *severe pain* treated by Aspirin 356.4 mg and Caffeine 30 mg, *tension headache* treated by Aspirin 325 mg and Caffeine 40 mg, and *pain of acte* *musculoskeletal disorder* treated by Aspirin 385 mg and Caffeine 30 mg). In the case of no observed synergistic effect between drugs for a disease, we probably find their synergistic effect by adjusting their dosages or try them for other diseases. Nevertheless, because such a trial takes much time and money in reality, clinical doctors cannot search it for all the dosages across different diseases, so as to the number of drug combinations is quirt less than the number we expect.

Finally, we consider the learning methodology when the benchmark is updated with a set of labelled negative samples or an additional validation dataset. Currently, only positive samples are explicitly labeled, binary classification is inappropriate to cope with the prediction of drug combinations, but one-class classification, which only uses positive samples to train the predictor, is quite qualified for this task. If the situation mends, (e.g. we obtain a set of explicitly labeled negative samples, even a small number), either semi-supervised or transductive learning becomes a better solution for predicting drug combination. Semi-supervised learning is appropriate to the case that the testing samples come from the same distribution as that of the training samples, while transductive learning can work for the case that the distributions of the training samples and the testing samples are different or not related. In fact, transductive learning can allow that the testing samples come from an arbitrary distribution.

In summary, although the collection of drug combinations is tough, except for positive samples and unlabelled samples, a new benchmark dataset containing a substantial number of negative samples is definitely needed. With the help of appropriate learning methods (e.g. transductive learning), it is expected that we can achieve a better prediction of drug combinations.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61872297), the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University (No. ZZ2018170, ZZ2018235, China National Training Programs of Innovation and Entrepreneurship for Undergraduates (No. 201710699330), and the Program of Peak Experience of NWPU (2016).

Competing Interests

The authors have no competing interests to declare.

References

[1] J. Jia, F. Zhu, X.H. Ma, Z.W.W. Cao, Y.X.X. Li, Y.Z. Chen, Mechanisms of drug combinations: interaction and network perspectives, Nature Reviews Drug Discovery, 8 (2009) 111-128.

[2] D. Chen, X. Liu, Y. Yang, H. Yang, P. Lu, Systematic synergy modeling: understanding drug synergy from a systems biology perspective, Bmc Syst Biol, 9 (2015) 56.

[3] A.A. Borisy, P.J. Elliott, N.W. Hurst, M.S. Lee, J. Lehar, E.R. Price, G. Serbedzija, G.R. Zimmermann, M.A. Foley, B.R. Stockwell, C.T. Keith, Systematic discovery of multicomponent therapeutics, Proceedings of the National Academy of Sciences of the United States of America, 100 (2003) 7977-7982.

[4] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, X. Chen, DCDB 2.0: a major update of the drug combination database, Database : the journal of biological databases and curation, 2014 (2014) bau124.

[5] P. Li, J.X. Chen, J.A. Wang, W. Zhou, X. Wang, B.H. Li, W.Y. Tao, W. Wang, Y.H. Wang, L. Yang, Systems pharmacology strategies for drug discovery and combination with applications to cardiovascular diseases, J Ethnopharmacol, 151 (2014) 93-107.

[6] L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, S.T. Wong, DrugComboRanker: drug combination discovery based on target network analysis, Bioinformatics, 30 (2014) i228-236.

[7] X.M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. van Noort, P. Bork, Prediction of drug combinations by integrating molecular and pharmacological data, PLoS computational biology, 7 (2011) e1002323.

[8] H. Huang, P. Zhang, X.A. Qu, P. Sanseau, L. Yang, Systematic prediction of drug combinations based on clinical side-effects, Sci Rep, 4 (2014) 7160.

[9] K.F. Pang, Y.W. Wan, W.T. Choi, L.A. Donehower, J.C. Sun, D. Pant, Z.D. Liu, Combinatorial therapy discovery using mixed integer linear programming, Bioinformatics, 30 (2014) 1456-1463.

[10] J.Y. Shi, J.X. Li, K. Gao, P. Lei, S.M. Yiu, Predicting combinative drug pairs towards realistic screening via integrating heterogeneous features, Bmc Bioinformatics, 18 (2017) 409.

[11] F. Cheng, Z. Zhao, Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties, Journal of the American Medical Informatics Association : JAMIA, 21 (2014) e278-286.

[12] L. Chen, B.Q. Li, M.Y. Zheng, J. Zhang, K.Y. Feng, Y.D. Cai, Prediction of Effective Drug Combinations by Chemical Interaction, Protein Interaction and Target Enrichment of KEGG Pathways, BioMed Research International, (2013).

[13] X. Cheng, S.G. Zhao, X. Xiao, K.C. Chou, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget, 8 (2017) 58494-58503.

[14] B. Liu, L. Fang, R. Long, X. Lan, K.C. Chou, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics, 32 (2016) 362-369.

[15] B. Liu, F. Yang, D.S. Huang, K.C. Chou, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, Bioinformatics, 34 (2018) 33-40.

[16] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, Journal of Theoretical Biology, 273 (2011) 236-247.

[17] M.A. Harris, J.I. Deegan, J. Lomax, M. Ashburner, S. Tweedie, S. Carbon, S. Lewis, C. Mungall, J. Day-Richter, K. Eilbeck, J.A. Blake, C. Bult, A.D. Diehl, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, G. Binkley, J.M. Cherry, K.R. Christie, M.C. Costanzo, Q. Dong, S.R. Engel, D.G. Fisk, J.E. Hirschman, B.C. Hitz, E.L. Hong, C.J. Krieger, S.R. Miyasato, R.S. Nash, J. Park, M.S. Skrzypek, S. Weng, E.D. Wong, K.K. Zhu, D. Botstein, K. Dolinski, M.S. Livstone, R. Oughtred, T. Berardini, D.H. Li, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, R. Huntley, N. Mulder, V.K. Khodiyar, R.C. Lovering, S. Povey, R. Chisholm, P. Fey, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, K. Van Auken, M.G. Giglio, L. Hannick, J. Wortman, M. Aslett, M. Berriman, V. Wood, H. Jacob, S. Laulederkind, V. Petri, M. Shimoyama, J. Smith, S. Twigger, P. Jaiswal, T. Seigfried, D. Howe, M. Westerfield, C. Collmer, T. Torto-Alalibo, E. Feltrin, G. Valle, S. Bromberg, S. Burgess, F. McCarthy, G.O. Consortium, The Gene Ontology project in 2008, Nucleic Acids Res, 36 (2008) D440-D444.

[18] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, D.S. Wishart, DrugBank 3.0: a comprehensive resource for 'omics' research on drugs, Nucleic Acids Res, 39 (2011) D1035-1041.

[19] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-A-Jee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L.G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W.Z. Li, W.D. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G.Y. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C.N. Arighi, L. Arminski, C.M. Chen, Y.X. Chen, J.S. Garavelli, H.Z. Huang, K. Laiho, P. McGarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q.H. Wang, Y.Q. Wang, L.S. Yeh, J. Zhang, U. Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res, 45 (2017) D158-D169.

[20] M. Kuhn, I. Letunic, L.J. Jensen, P. Bork, The SIDER database of drugs and side effects, Nucleic Acids Res, 44 (2016) D1075-D1079.

[21] N.P. Tatonetti, P.P. Ye, R. Daneshjou, R.B. Altman, Data-driven prediction of drug effects and interactions, Science translational medicine, 4 (2012) 125ra131.

[22] Z.Y. Liu, F.F. Guo, J.Y. Gu, Y. Wang, Y. Li, D. Wang, L. Lu, D. Li, F.C. He, Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources, Bioinformatics, 31 (2015) 1788-1795.

[23] T.K. Ho, J.J. Hull, S.N. Srihari, Decision Combination in Multiple Classifier Systems, Ieee T Pattern Anal, 16 (1994) 66-75.

[24] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, Ieee T Pattern Anal, 20 (1998) 226-239.

[25] S. Wang, Y.C. Zhang, W.R. Dai, K. Lauter, M. Kim, Y.Z. Tang, H.K. Xiong, X.Q. Jiang, HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS, Bioinformatics, 32 (2016) 211-218.

[26] B.L. Chen, M. Li, J.X. Wang, X.Q. Shang, F.X. Wu, A fast and high performance multiple data integration algorithm for identifying human disease genes, BMC medical genomics, 8 (2015).

[27] J.M. Ma, M.N. Nguyen, J.C. Rajapakse, Gene Classification Using Codon Usage and Support Vector Machines, Ieee Acm T Comput Bi, 6 (2009) 134-143.

[28] J.-Y. Shi, S.-M. Yiu, Y.-N. Zhang, F.Y.-L. Chin, Effective Moment Feature Vectors for Protein Domain Structures, Plos One, 8 (2013).

[29] B. Scholkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput, 13 (2001) 1443-1471.

[30] C. Cortes, V. Vapnik, Support-Vector Networks, Mach Learn, 20 (1995) 273-297.

[31] Z. Hajisharifi, M. Piryaiee, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, Journal of Theoretical Biology, 341 (2014) 34-40.

[32] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Molecular bioSystems, 9 (2013) 1092-1100.

[33] Y. Park, E.M. Marcotte, Flaws in evaluation schemes for pair-input computational predictions, Nat Methods, 9 (2012) 1134-1136.

[34] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: the 23rd international conference on Machine learning, 2006, pp. 233-240.

[35] Y. Jiao, P. Du, Performance measures in evaluating machine learning based bioinformatics predictors for classifications, Quantitative Biology, 4 (2016) 320-330.

[36] C.-C. Chang, C.-J. Lin, LIBSVM : a library for support vector machines. , ACM Transactions on Intelligent Systems and Technology, 2 (2011) 27:21-27:27.

[37] D.S. Cao, L.X. Zhang, G.S. Tan, Z. Xiang, W.B. Zeng, Q.S. Xu, A.F. Chen, Computational Prediction of Drug-Target Interactions Using Chemical, Biological, and Network Features, Mol Inform, 33 (2014) 669-681.

[38] H. Luo, P. Zhang, H. Huang, J. Huang, E. Kao, L. Shi, L. He, L. Yang, DDI-CPI, a server that predicts drug-drug interactions through implementing the chemical-protein interactome, Nucleic Acids Res, 42 (2014) 46-52.

[39] X. Cheng, S.G. Zhao, X. Xiao, K.C. Chou, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics, 33 (2017) 341-346.

[40] J.L. Min, X. Xiao, K.C. Chou, iEzy-drug: a web server for identifying the interaction between enzymes and drugs in cellular networking, Biomed Res Int, 2013 (2013) 701317.

[41] X. Xiao, J.L. Min, P. Wang, K.C. Chou, iGPCR-drug: a web server for predicting interaction between GPCRs and drugs in cellular networking, Plos One, 8 (2013) e72234.

[42] Y.N. Fan, X. Xiao, J.L. Min, K.C. Chou, iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking, Int J Mol Sci, 15 (2014) 4915-4937.

[43] X. Xiao, J.L. Min, P. Wang, K.C. Chou, iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints, J Theor Biol, 337 (2013) 71-79.

[44] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, K.C. Chou, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach, Journal of biomolecular structure & dynamics, 33 (2015) 2221-2233.

[45] K.C. Chou, Impacts of Bioinformatics to Medicinal Chemistry, Med Chem, 11 (2015) 218-234.

[46] K.C. Chou, An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science, Current topics in medicinal chemistry, 17 (2017) 2337-2358.

[47] J.Y. Shi, S.M. Yiu, Y.M. Li, H.C.M. Leung, F.Y.L. Chin, Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering, Methods, 83 (2015) 98-104.

[48] J.-Y. Shi, Z. Liu, H. Yu, Y.-J. Li, Predicting Drug-Target Interactions via Within-Score and Between-Score, BioMed Research International, 2015, Article ID 350983, 9 pages (2015).

[49] J.-Y. Shi, J.-X. Li, H.-M. Lu, Predicting existing targets for new drugs base on strategies for missing interactions, Bmc Bioinformatics, 17 (2016) 282.

[50] J.-Y. Shi, J.-X. Li, B.-L. Chen, Y. Zhang, Inferring Interactions between Novel Drugs and Novel Targets via Instance-Neighborhood-Based Models, Current Protein & Peptide Science, 19 (2018) 488-497.

[51] P. Zhang, F. Wang, J. Hu, R. Sorrentino, Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects, Sci Rep, 5 (2015) 12339.

[52] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, X. Li, Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data, Bmc Bioinformatics, 18 (2017) 18.

[53] H. Yu, K.-T. Mao, J.-Y. Shi, H. Huang, Z. Chen, K. Dong, S.-M. Yiu, Predicting and Understanding Comprehensive Drug-Drug Interactions via Semi-Nonnegative Matrix Factorization, Bmc Syst Biol, 12 (2018).

[54] C. Lu, J. Wang, Z.L. Zhang, P.Y. Yang, G.X. Yu, NoisyGOA: Noisy GO annotations prediction using taxonomic and semantic similarity, Comput Biol Chem, 65 (2016) 203-211.

[55] W. Zhang, X. Yue, F. Liu, Y.L. Chen, S.K. Tu, X.N. Zhang, A unified frame of predicting side effects of drugs by using linear neighborhood similarity, Bmc Syst Biol, 11 (2017).

[56] W.P. Lee, J.Y. Huang, H.H. Chang, K.T. Lee, C.T. Lai, Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources, Ieee Access, 5 (2017) 20449-20462.

[57] W. Zhang, X.R. Liu, Y.L. Chen, W.J. Wu, W. Wang, X.H. Li, Feature-derived graph regularized matrix factorization for predicting drug side effects, Neurocomputing, 287 (2018) 154-162.