frontiers in Pharmacology

TECHNOLOGY AND CODE
published: 05 September 2019
doi: 10.3389/fphar.2019.00971

ATC-NLSP: Prediction of the Classes of Anatomical Therapeutic Chemicals Using a Network-Based Label Space Partition Method

Xiangeng Wang†, Yanjing Wang†, Zhenyu Xu, Yi Xiong and Dong-Qing Wei**

State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Jianfeng Pei, Peking University, China

Reviewed by:
Cao Dongsheng, Central South University, China
Quan Zou, University of Electronic Science and Technology of China, China

***Correspondence:**
Yi Xiong
xiongyi@sjtu.edu.cn
Dong-Qing Wei
dqwei@sjtu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: *10 June 2019*
Accepted: *29 July 2019*
Published: *05 September 2019*

Citation:
Wang X, Wang Y, Xu Z, Xiong Y and Wei D-Q (2019) ATC-NLSP: Prediction of the Classes of Anatomical Therapeutic Chemicals Using a Network-Based Label Space Partition Method. Front. Pharmacol. 10:971. doi: 10.3389/fphar.2019.00971

Anatomical Therapeutic Chemical (ATC) classification system proposed by the World Health Organization is a widely accepted drug classification scheme in both academic and industrial realm. It is a multilabeling system which categorizes drugs into multiple classes according to their therapeutic, pharmacological, and chemical attributes. In this study, we adopted a data-driven network-based label space partition (NLSP) method for prediction of ATC classes of a given compound within the multilabel learning framework. The proposed method ATC-NLSP is trained on the similarity-based features such as chemical–chemical interaction and structural and fingerprint similarities of a compound to other compounds belonging to the different ATC categories. The NLSP method trains predictors for each label cluster (possibly intersecting) detected by community detection algorithms and takes the ensemble labels for a compound as final prediction. Experimental evaluation based on the jackknife test on the benchmark dataset demonstrated that our method has boosted the absolute true rate, which is the most stringent evaluation metrics in this study, from 0.6330 to 0.7497, in comparison to the state-of-the-art approaches. Moreover, the community structures of the label relation graph were detected through the label propagation method. The advantage of multilabel learning over the single-label models was shown by label-wise analysis. Our study indicated that the proposed method ATC-NLSP, which adopts ideas from network research community and captures the correlation of labels in a data driven manner, is the top-performing model in the ATC prediction task. We believed that the power of NLSP remains to be unleashed for the multilabel learning tasks in drug discovery. The source codes are freely available at https://github.com/dqwei-lab/ATC.

Keywords: drug classification, multilabel classification, label correlation, label space partition, label propagation

INTRODUCTION

The Anatomical Therapeutic Chemical (ATC) Classification System (MacDonald and Potvin, 2004), maintained by the World Health Organization Collaborating Centre for Drug Statistics Methodology, is the most widely accepted and canonical scheme for drug categorization. This system assigns different group labels for drugs based on the organ or systems where they take effect and/

or their therapeutic, pharmacological, and chemical attributes. The ATC system is a strict hierarchy, including five levels of classification, and for the first level, there are 14 main groups: 1) alimentary tract and metabolism (coded by **A**); 2) blood and blood-forming organs (coded by **B**); 3) cardiovascular system (coded by **C**); 4) dermatologicals (coded by **D**); 5) genitourinary system and sex hormones (coded by **G**); 6) systemic hormonal preparations, excluding sex hormones and insulins (coded by **H**); 7) anti-infectives for systemic use (coded by **J**); 8) antineoplastic and immunomodulating agents (coded by **L**); 9) musculoskeletal system (coded by **M**); 10) nervous system (coded by **N**); 11) antiparasitic products, insecticides, and repellents (coded by **P**); 12) respiratory system (coded by **R**); 13) sensory organs (coded by **S**); and 14) various (coded by **V**). Given a new compound, prediction of its ATC classes can provide us with deeper insights into its therapeutic indications and side effects, thus accelerating both basic research and drug development (Hutchinson et al., 2004; Dunkel et al., 2008).

Traditionally, identification of ATC classes for a new drug using experimental methods is both time- and resource-consuming. Therefore, *in silico* prediction of ATC classes of a compound by machine learning techniques is a hot field in drug discovery and development. Previous studies (Dunkel et al., 2008; Wu et al., 2013) formulate the prediction of ATC classes as a single-label learning task, which is suggested to be inappropriate due to the multilabel nature of this biological system (Chou, 2013). Within the multilabel learning framework, Cheng et al. (2017b) proposed a multilabel predictor iATC-mISF, which utilized multilabel Gaussian kernel regression and three types of features (chemical–chemical interaction, structural similarity, and fingerprint similarity). The iATC-mISF has been upgraded as iATC-mHyb (Cheng et al., 2017a) by further incorporating drug ontological information. Besides one-dimensional representation of features, inspired by the histograms of oriented gradients (HoG) method proposed by the computer vison community (Dalal and Triggs, 2005), Nanni and Brahnam (2017) reshaped the features into two-dimensional matrix and performed slightly better than iATC-mISF. Continuing in this direction, the same group (Lumini and Nanni, 2018) applied pretrained convolutional neural networks models on the two-dimensional feature matrix as a featurizer and achieved best performance among the previously published methods on this task.

Typically, multilabel (ML) classification algorithms are classified into three major groups: algorithm adaptation, problem transformation, and ensembles of multilabel classifier (EMLC) (Wan et al., 2017). Algorithm adaptation methods incorporate specific tricks that modify traditional single-label learning algorithms into multilabel ones. The representative algorithm of this group is ML-*k*NN (Zhang and Zhou, 2005). For the problem transformation method, it converts multilabel learning problem into one or more single-label problems. The common strategies for such a transformation include binary relevance, classifier chains, label ranking, and label powerset (LP) (Read et al., 2011). LP trains models on each possible subset of label sets (Gibaja and Ventura, 2014). For a dataset with high cardinality in the large label set, LP is prone to be overfitting because of the exponentially increased number of subsets. To tackle the overfitting nature of label powerset, (Tsoumakas et al., 2011) proposed the RA*k*EL*d* method, which divides the label set into *k* disjoint subsets and use label powerset in these subsets. One major drawback of RA*k*EL*d* is that the *k* is arbitrarily chosen without incorporating the label correlations, which can be possibly learnt from the training data. The **n**etwork-based **l**abel **s**pace **p**artition (NLSP) (Szymański et al., 2016) is an EMLC built upon ML. This NLSP method divides the label set into *k* small-sized label sets (possibly intersecting) by a community detection method, which can incorporate the label correlation structures in the training set, such that it finally learns *k* representative ML classifiers. As a result, NLSP tackles much less subsets compared to LP on the original label set and selects *k* in a data-driven manner. For more detailed explanation of multilabel learning, refer to (Zhang and Zhou, 2014; Moyano et al., 2018).

In this study, we adopted an NLSP method to explore the correlation among labels. Our NLSP method was evaluated on a benchmark dataset (Chen et al., 2012) by the jackknife test. The proposed method demonstrates its superiority over other state-of-the-art approaches by our experimental results. The main strength of our method hinges on two aspects. On the one hand, the NLSP clusters the label space into subspaces and utilizes the correlation among labels. On the other hand, the ensemble learning nature of NLSP on the overlapping subspace could further improve model performance. Interesting patterns on the label relation graph were also detected by NLSP. In addition, the label-wise analysis of the best NLSP model was performed to provide experimental biologists with more insights.

MATERIALS AND METHODS

Benchmark Dataset and Sample Formulation

We utilized the same dataset as the previous study (Cheng et al., 2017b) to facilitate model comparison. This dataset consists of 3,883 drugs, and each drug is labeled with at least one or more of 14 main ATC classes. It is a tidy dataset where no missing value and contradictory record. The UpSet visualization technique (Lex et al., 2014) was used for quantitative analysis of interactions of label sets.

Then, we adopted the same method provided by (Cheng et al., 2017b) to represent the drug samples. The dataset can be formulated in set notation as the union of elements in each class: $\mathbb{S} = \mathbb{S}_1 \bigcup \mathbb{S}_2 \ldots \bigcup \mathbb{S}_{14}$ (1), and a sample D can be represented by concatenating the following three types of features.

1. A 14-dimentional vector, $D^{\mathbf{Int}} = [\Phi_1 \Phi_2 \Phi_3 \ldots \Phi_{14}]^T$ (2), which represents its maximum interaction score Φ_i (Kotera et al., 2012) with the drugs in each of the 14 $\mathbb{S}_i$.
2. A 14-dimentional vector, $D^{\mathbf{StrSim}} = [\Psi_1 \Psi_2 \Psi_3 \ldots \Psi_{14}]^T$ (3) which represents its maximum structural similarity score Ψ_i (Kotera et al., 2012) with the drugs in each of the 14 $\mathbb{S}_i$.
3. A 14-dimentional vector, $D^{\mathbf{FigSim}} = [\mathrm{T}_1 \mathrm{T}_2 \mathrm{T}_3 \ldots \mathrm{T}_{14}]^T$ (4), which represents its molecular fingerprint similarity score T_i (Xiao et al., 2013) with the drugs in each of the 14 $\mathbb{S}_i$.

Therefore, a given drug D is formulated by:

$$D = D^{\mathbf{Int}} \oplus D^{\mathbf{StrSim}} \oplus D^{\mathbf{FigSim}} = [@_1 @_2 @_3 \ldots @_{42}]^T \quad (5)$$

Where $\oplus$ represents the symbol for orthogonal sum and where

$$@_u = \begin{cases} \Phi_u (1 \le u \le 14) \\ \Psi_u (15 \le u \le 28) \\ \mathrm{T}_u (29 \le u \le 42) \end{cases} \quad (6)$$

For more details, refer to Cheng et al. (2017b).

Measuring Label Correlation

In order to evaluate the correlation between two labels, we calculated the bias corrected Cramér's V statistic for all the label pairs (Bergsma, 2013). Cramér's V (sometimes referred to as Cramér's phi and denoted as φc) statistic is a measure of association between two nominal variables, ranging from 0 to 1 (inclusive). The bias corrected Cramér's V statistic is given by (here n denotes sample size and χ^2 stands for the chi-square statistic without a continuity correction for a contingency table with r rows and c columns)

$$\tilde{V} = \sqrt{\frac{\tilde{\varphi}^2}{\tilde{m}}} \quad (7)$$

where

$$\tilde{\varphi}^2 = \max(0, \varphi^2 - \frac{(r-1)(c-1)}{n-1})) \quad (8),$$

$$\varphi^2 = \frac{\chi^2}{n} \quad (9)$$

and

$$\tilde{m} = \min(\tilde{r} - 1, \tilde{c} - 1) \quad (10),$$

$$\tilde{r} = r - \left(\frac{(r-1)^2}{n-1} \right) \quad (11),$$

$$\tilde{c} = c - \left(\frac{(\mathrm{c}-1)^2}{n-1} \right) \quad (12).$$

Network-Based Label Space Partition

The NLSP is a newly proposed multilabel learning method and has achieved top performance in some predictive tasks (Szymański et al., 2016). In this study, we adopted the data-driven NLSP method for prediction of ATC classes of a compound. NLSP divides the predictive modeling task into the training and classification phase.

In the training phase, four steps are preformed:

1. Establishing a label co-occurrence graph on the training set. The label co-occurrence graph G has the label set L as the vertex set and the edge between two vertices (labels) exists if at least one sample S in training set D_{train} is assigned by these two labels l_i and l_j together (here l_i, l_j denote labels of the set L_s, which stands for the assigned label set of a sample S; || || stands for the cardinality of a given set):

$$E = \left\{ \{l_i, l_j\} : (\exists (S, L_s) \in \mathrm{D}_{train}) \left(l_i \in L_s \wedge l_j \in L_s \right) \right\} \quad (13)$$

We can also easily assign weights to G by defining a counting function $w: L \to \mathbb{N}$:

$$w(l_i, l_j) = \text{number of sample } S \text{ that have both labels assigned} = \left\| \left\{ S : (S, L_s) \in \mathrm{D}_{train} \wedge l_i \in L_s \wedge l_j \in L_s \right\} \right\| \quad (14)$$

2. Detecting community on the label co-occurrence graph. There are various community detection algorithms. In this study, we utilized the following two methods to identify communities because both of the two methods have linear time complexity:

   a) **Largest modularity using incremental greedy search (Louvain method)** (Blondel et al., 2008): This method is based on greedy aggregation of communities, beginning with communities with single convex and merging the communities iteratively. In each step, two communities are merged when the merging makes the highest contribution to modularity. The algorithm halts when there is no merge that could increase current modularity. This method is frequently referred as "Louvain method" in the network research community. The detailed explanation of this method is described in **Supplementary Method S1**.

   b) **Multiple async label propagation (LPA)** (Raghavan et al., 2007): This method assigns unique tags to every vertex in a graph and then iteratively updates the tags of every vertex. This update reassigns the tag of the majority of neighbors to the central vertex. The updating order of vertices shuffled at each iteration. The algorithm is stopped when all vertices have tags identical to the dominant tag in proximity. The detailed description of LPA is appended in **Supplementary Method S2**.

3. For each community C_i, corresponding training set D_i is created by taking the original dataset with label columns presented in L_i.
4. For each community, a base predictor b_i is learnt on the training set D_i. In this study, we compared the performance of five types of base predictors:

   (a) **Extremely randomized trees (ERT)** (Geurts et al., 2006; Li et al., 2019) is an ensemble method that adds more randomness compared to random forests by the random top–down splitting of trees instead of computing the locally optimal cut-point for each feature under consideration. This increase in randomness allows to reduce the variance of the model a bit, at the expense of a slightly greater increase in bias.

(b) **Random forests (RF)** (Breiman, 2001) is an ensemble method that combines the probabilistic predictions of a number of decision tree-based classifiers to improve the generalization ability over a single estimator.
(c) **Support vector machine (SVM)** (Cortes and Vapnik, 1995) is a widely used classification algorithm which tries to find the maximum margin hyperplane to divide samples into different classes. Incorporated by kernel trick, this method could handle both linear and no-linear decision boundary.
(d) **Extreme gradient boosting (XGB)** (Chen and Guestrin, 2016) is a newly proposed boosting method, which has achieved state-of-the-art performance on many tasks with tabular training data (Chen et al., 2018). Traditional gradient boosting machine is a meta algorithm to build an ensemble strong learner from weak learners such as decision trees, while XGB is an efficient and distributed implementation of gradient boosting machine.
(e) **Multilayer perceptron (MLP)** (Ruck et al., 1990) is a supervised learning algorithm which could learn nonlinear models. It has one or more nonlinear hidden layers between the input and output. For each hidden layer, different numbers of hidden neurons can be assigned. Each hidden neuron yields a weighted linear summation of the values from the previous layer, and the nonlinear activation function is followed. The weights are learnt through backpropagation algorithm or variations upon it.

In the classification phase, we just perform predication on all communities detected in the training phase and fetch the union of assigned labels:

$$b(S) = \bigcup_{j=1}^{k} b_i(S) \tag{15}$$

Parameter Tuning

There are two layers of hyperparameters tunable for NLSP:

1. The base learner: we chose five types of base learners.

   (a) Extremely randomized trees: we tuned the hyperparameter of number of trees at [500, 1000], other hyperparameters are at the default values.
   (b) Random forests: we tuned hyperparameter of number of trees at [500, 1,000], other hyperparameters are at the default values.
   (c) Support vector machine: we tuned the hyperparameter of C (penalty) at [0.01, 0.1, 1, 10, 100], we chose the radial basis function with gamma value of $\frac{1}{N_{features}} = \frac{1}{42}$, other hyperparameters are at the default values.
   (d) Extreme gradient boosting: we tuned the hyperparameter of number of trees at [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], other hyperparameters are at the default values.
   (e) Multilayer perceptron: We tuned the hyperparameter of hidden layer sizes at [50, 100, 200, 500, 1,000], other hyperparameters are at the default values.

2. The cluster: for each type of base learner, we try to compare two community detection methods.

   (a) Largest modularity using incremental greedy search (Blondel et al., 2008).
   (b) Multiple async label propagation (Raghavan et al., 2007).

Performance Measures of Multilabel Learning

Evaluation of a multilabel learning model is not a trivial task (Zhang et al., 2015; Yuan et al., 2016; Zhang et al., 2017; You et al., 2018; Xiong et al., 2019; You et al., 2019). Inspired by the definition of Chou *et al.* (Chou, 2013) and practice of Madjarov et al. (2012), we utilized the following five metrics to evaluate the multilabel learning models throughout this work.

$$\begin{cases} Aiming = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k^*\right\|}\right) \\ Coverage = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k\right\|}\right) \\ Accuracy = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}{\left\|\mathbb{L}_k \cap \mathbb{L}_k^*\right\|}\right) \\ Absolute\ True = \frac{1}{N}\sum_{k=1}^{N}\left(\mathbb{L}_k, \mathbb{L}_k^*\right) \\ Hamming\ loss = \frac{1}{N}\sum_{k=1}^{N}\left\|\mathbb{L}_k \ominus \mathbb{L}_k^*\right\| \end{cases} \tag{16}$$

where N is the total number of samples, M is the total number of labels, $\bigcup$ represents union in set theory and $\bigcap$ represents intersection in set theory, $\mathbb{L}_k$ denotes the true label set of k-th sample, $\mathbb{L}_k^*$ means the predicted label vector of k-th sample, $\ominus$ stands for the symmetric difference between two sets, and

$$\Delta(\mathbb{L}_k, \mathbb{L}_k^*) = \begin{cases} 1, if\ all\ the\ labels\ in\ \mathbb{L}_k\ equal\ \mathbb{L}_k^* \\ 0, otherwise \end{cases} \tag{17}$$

In order to avoid the zero-divisor problem generated by all negative predictions, we add a pseudo-number 1 to 0 divisors in the calculation of the aiming metric. These above metrics have been used in a series of studies (Cheng et al., 2017a; Cheng et al., 2017b; Nanni and Brahnam, 2017).

Performance Measures of Single-Label Learning

Apart from the metrics in the multilabel framework, we also utilized the following metrics to assess the single-label classification models.

$$\begin{cases} Accuracy = \dfrac{TP+TN}{TP+TN+FN+FP} \\ Specificity = \dfrac{TN}{TN+FP} \\ Recall = \dfrac{TP}{TP+FN} \\ F1 = \dfrac{2TP}{2TP+FP+FN} \end{cases} \quad (18)$$

where *TP*, *TN*, *FN*, and *TN* are true positives, true negatives, false positives, and false negatives for the prediction of each label, respectively. These metrics have widely been used in a large number of bioinformatics applications recently (Feng et al., 2017; Niu and Zhang, 2017; Sun et al., 2017; Wang et al., 2017; Xu et al., 2017; He et al., 2018; Li et al., 2018; Pan et al., 2018; Qiao et al., 2018; Xiong et al., 2018; Xu et al., 2018; Zhang et al., 2018; Bian et al., 2019; Wei et al., 2019a; Wei et al., 2019b; Zou et al., 2019). In addition, we also calculated the area under the receive operating characteristic curve (AUC) by the trapezoidal rule.

Model Validation Method

There are mainly three methods to evaluate the generalization ability of a classification model, such as the independent testing method, *k*-fold cross validation, and the jackknife method. In order to fairly compare our proposed model with previous works on the same benchmark dataset, we utilized the jackknife method for the model validation in the multilabel learning framework. Jackknife is a resampling method for parameter estimation. The jackknife estimation of a parameter is constructed by calculating the parameter for each subsample omitting the *i*-th observation and then takes the mean value of these parameters as final estimation.

In the model validation of single-label analysis, we utilized 10 times repeated 10-fold cross validation (10 × 10-fold CV) method. In *k*-fold cross validation (CV), the sample set is randomly partitioned into *k* subsets with equal size. Of the *k* subsets, one subset is selected as the validation data for testing the model, and the remaining $k-1$ subsets are used for training. The cross-validation process is then repeated *k* times (the folds), with each of the *k* subsets used exactly once as the validation data. The 10-fold cross-validation is proven to be a better alternative of jackknife method in terms of bias, variance, and computation complexity (Kohavi, 1995). We also repeated 10-fold CV 10 times in shuffled benchmark dataset to further reduce the estimation variance.

RESULTS AND DISCUSSION

Label Correlation Analysis

One major advantage of multilabel learning framework is the explicit exploitation of label correlations (Zhang and Zhou, 2014). We calculated bias corrected Cramér's V statistics for all the label pairs and depicted them in a heatmap manner (**Figure 1A**), and the UpSet visualization of label intersections is depicted in **Figure 1B**. The results indicated that 46 drugs are both labeled as ATC category 4 (dermatologicals) and ATC category 12 (respiratory system), 43 drugs are both labeled as ATC category 13 (sensory organs) and ATC category 7 (anti-infectives for systemic use), which can be explained by the fact that many widely applied corticosteroids, such as dexamethasone, betamethasone, and fluocortolone, can be used both in dermatology and respirology medicine. We also found that several label sets are correlated, especially for ATC category 4 (dermatologicals) and ATC category 13 (sensory organs), of which the Cramér's V statistic

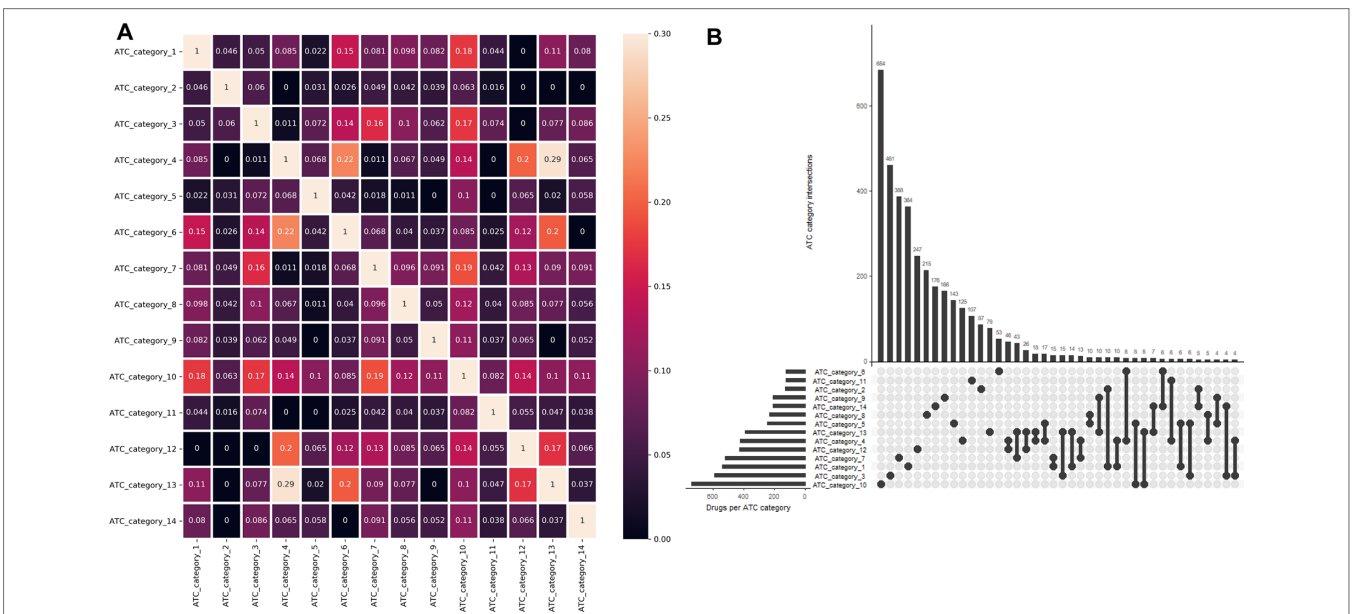

FIGURE 1 | Label correlation landscape. **(A)** The pair wise visualization of Cramér's V statistics for all the labels in a heatmap manner. **(B)** The UpSet visualization of label intersections. The horizontal bar shows the number of drugs per ATC category, and the vertical bar shows the number of drugs per ATC category intersection.

is 0.29. Details about the pairwise intersection numbers of drugs and the pairwise Cramér's V statistics between all the labels are shown in **Table S1** and **Table S2**.

Multilabel Performance Comparison

Table 1 shows the prediction performances based on the jackknife test among different methods on the benchmark dataset. We found the absolute true value of almost all our NLSP-based methods performed better than that of other methods, which is the most stringent metric for multilabel learning. Among all the NLSP-based models, the NLSP-XGB-LPA performs the best, consistently better than all the other methods trained on benchmark dataset, in terms of aiming, coverage, accuracy, and absolute true. As for the value of absolute true, our NLSP-XGB-LPA has boosted ~11.67% compared to the best deep learning model trained on the same benchmark dataset (Lumini and Nanni, 2018). As for the clusterer, we found that the LPA method performs consistently better than the Louvain method in all the NLSP-based models (**Figure S1**), so we append the suffix of "-LPA" to all the NLSP-based models. We then trained the final NLSP-XGB-LPA model on the full benchmark dataset using previous optimized hyperparameters. This model can be accessed through https://github.com/dqwei-lab/ATC.

Label Community Analysis

One major innovation of NLSP method is the construction of label relation graph, which is built on the concept of label co-occurrence (Szymanski and Kajdanowicz, 2019). The communities detected in the label relation graph will not only help to improve the classification performance but also provide us with deeper insights of the intrinsic label structure. We extracted the community membership information from the final model of NLSP-XGB-LPA (shown in **Figure 2**). We found that there are two communities detected, in which ATC category 8 (anti-infectives for systemic use) lies in a unique community. In terms of medicinal chemistry and clinical pharmacotherapeutics, anti-infectives for systemic use are structure variant and usage limited compared to other 16 types of drugs. For example, daptomycin (DB00080) is one of the anti-infectives for systemic use, which is composed of an unusual molecular structure of lipopeptide with limited indications for skin and skin structure infections caused by Gram-positive infections, *S. aureus* bacteremia, and right-sided *S. aureus* endocarditis (Henken et al., 2010). The community membership learnt from benchmark dataset is surprising but intuitive. This result suggests the potential pattern extraction power of network-based machine learning models in terms of pharmacology.

Single-Label Analysis

Apart from multilabel learning metrics, it is often useful to evaluate multilabel learning models in a label-wise manner (Michielan et al., 2009; Mayr et al., 2016). We utilized the parameters of the best-performing model of NLSP-XGB-LPA and conducted 10 times repeated 10-fold cross-validation (10×10-fold CV) because the jackknife test is rather time consuming. The details are listed in **Table 2**. We found that our NLSP-XGB-LPA performs well in all the single-label subtasks of ATC prediction, especially for the label of "anti-infectives for systemic use," reaching an AUC at 0.9946. Compared to a dedicated single-label classification system for cardiovascular system (Gurulingappa et al., 2009), our best-performing multilabel model boosted the value of accuracy from 0.8947 into 0.9490.

TABLE 1 | Comparison with other state-of-the-art multilabel predictors.

Method	DL[a]	Aiming	Coverage	Accuracy	Absolute true	Hamming loss
EnsANet_LR ⊕ DO[c] (τ = 0.25) (Lumini and Nanni, 2018)	Yes	0.7957	0.8335	0.7778	0.7090	Not available
EnsANet_LR ⊕DO[c] (τ = 0.5) (Lumini and Nanni, 2018)	Yes	0.9011	0.7162	0.7232	0.6871	
EnsLIFT (Nanni and Brahnam, 2017)	No	0.7818	0.7577	0.7121	0.6330	
iATC-mHyb[c] (Cheng et al., 2017a)	No	0.7191	0.7146	0.7132	0.6675	
Chen et al. (Chen et al., 2012)	No	0.5076	0.7579	0.4938	0.1383	
iATC-mISF (Cheng et al., 2017b)	No	0.6783	0.6710	0.6641	0.6098	
NLSP-ERT-LPA	No	0.7948	0.7691	0.7578	0.7213	0.03817
NLSP-RF-LPA	No	0.8072	0.7889	0.7778	0.7489	**0.03427**
NLSP-SVM-LPA	No	0.7844	0.7529	0.7370	0.6925	0.04322
NLSP-XGB-LPA	No	**0.8135**[b]	**0.7950**	**0.7828**	**0.7497**	0.03429
NLSP-MLP-LPA	No	0.7958	0.7858	0.7591	0.7090	0.04032

[a] *DL denotes whether this model is a deep learning-based method.*

[b] *The bold value stands for the best value of specific metrics.*

[c] *These models are trained on a modified benchmark dataset, whose metrics are not comparable to our model.*

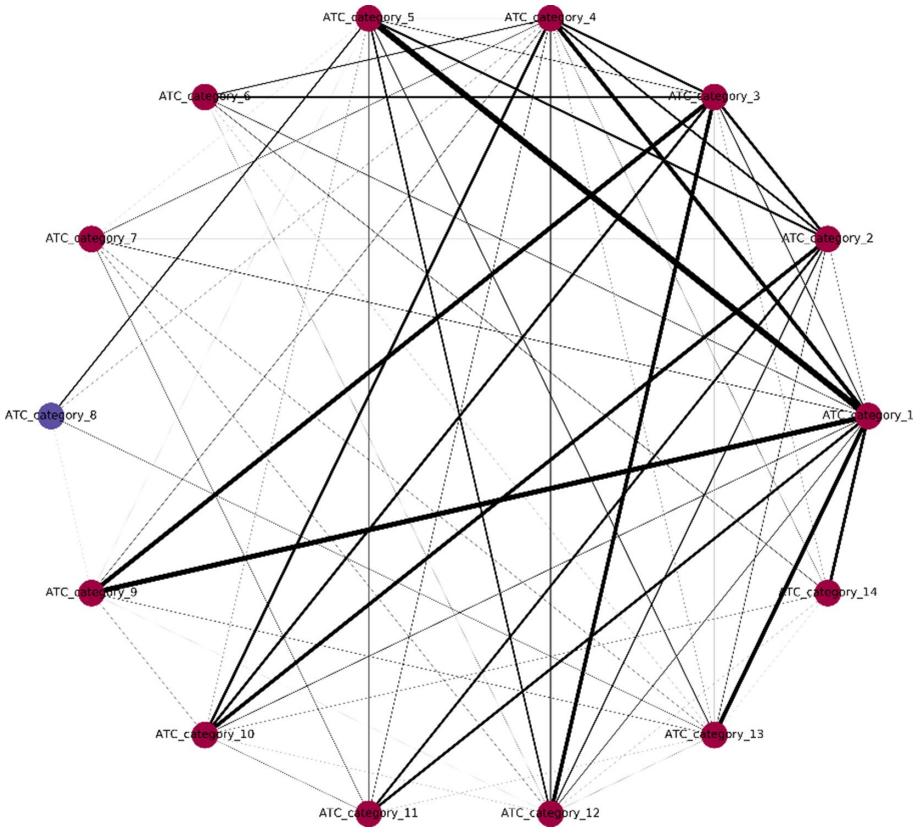

FIGURE 2 | Label relation graph. Different colors stand for different communities. The line width represents the weight between two labels. Communities are detected by multiple async label propagation method, while the weight represents the frequency of label co-occurrence.

TABLE 2 | Label-wise analysis of best-performing multilabel learning model.

Predictive label	Accuracy	Specificity	Recall	F1 score	AUC	Evaluation method
Alimentary tract and metabolism	0.9269	0.7312	0.7549	0.7406	0.9550	10 × 10-fold CV
Blood and blood forming organs	0.9793	0.7754	0.5644	0.6430	0.9493	10 × 10-fold CV
Cardiovascular system	0.9490	0.8371	0.8274	0.8306	0.9752	10 × 10-fold CV
Dermatologicals	0.9403	0.7966	0.6038	0.6845	0.9472	10 × 10-fold CV
Genitourinary system and sex hormones	0.9691	0.8148	0.6682	0.7294	0.9539	10 × 10-fold CV
Systemic hormonal preparations, excluding sex hormones and insulins	**0.9867**[a]	0.8227	0.7605	0.7816	0.9940	10 × 10-fold CV
Anti-infectives for systemic use	0.9793	**0.9276**	**0.9170**	**0.9215**	**0.9946**	10 × 10-fold CV
Antineoplastic and immunomodulating agents	0.9792	0.8683	0.7724	0.8126	0.9804	10 × 10-fold CV
Musculoskeletal system	0.9820	0.8707	0.7836	0.8209	0.9842	10 × 10-fold CV
Nervous system	0.9511	0.8581	0.8913	0.8733	0.9825	10 × 10-fold CV
Antiparasitic products, insecticides and repellents	0.9863	0.8312	0.7358	0.7714	0.9803	10 × 10-fold CV
Respiratory system	0.9573	0.8432	0.7516	0.7923	0.9720	10 × 10-fold CV
Sensory organs	0.9492	0.8206	0.6367	0.7140	0.9487	10 × 10-fold CV
Various	0.9717	0.7681	0.6997	0.7241	0.9703	10 × 10-fold CV
Cardiovascular system (Gurulingappa et al., 2009)	0.8947		Not available			100 × bootstrapping
Cardiovascular system (Gurulingappa et al., 2009)	0.7712					Test set
SuperPred (Dunkel et al., 2008)	0.676[b]					Jackknife

[a] *The bold value stands for the best value of specific metrics.*

[b] *The mean accuracy of flattened 850 ATC classes.*

CONCLUSION

Based upon the NLSP method, we have achieved the state-of-the-art performance on the benchmark dataset using the similarity-based features such as chemical–chemical interaction and structural and fingerprint similarities of a compound to other compounds belonging to the different ATC categories. Label community and single-label analysis were also performed on the benchmark dataset. There are three major conclusions can be reached. First, compared to dedicated single-label models (Dunkel et al., 2008; Gurulingappa et al., 2009), multilabel learning framework could improve the performance on single-label metrics by incorporating label correlation information. Second, compared to feature engineering tricks (Nanni and Brahnam, 2017; Lumini and Nanni, 2018), the introduction of new method such as NLSP could generate more performance improvement. Third, at least in the ATC prediction task, the NLSP method, which adopts ideas from network research community and captures the correlation of labels in a data-driven manner, can perform better than the models based on deep learning techniques, especially in the absolute true rate metric. The idea behind NLSP method is fascinating, and the power of NLSP remains to be unleashed for the multilabel learning tasks in drug discovery.

Although the NLSP method was the first time to be applied to the multilabel classification task in pharmacology and achieved good performance in the preliminary results, there are shortcomings in several aspects in this study. First, the similarity-based features are not recalculated for the specific communities detected by the NLSP methods. Second, the rigidity of the model validation can be improved by the independent external dataset. Last but not the least, the number of communities detected by NLSP on this drug classification problem is too low, which may be not an ideal dataset for proving the predictive power of the NLSP-based method. These problems can be addressed in the further studies.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or **Supplementary Files**.

AUTHOR CONTRIBUTIONS

YX, D-QW and XW contributed conception and design of the study; XW and YW organized the database; XW, YW and ZX performed the statistical analysis; XW wrote the first draft of the manuscript; XW and YW wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This work was supported by the funding from National Key Research Program (contract no. 2016YFA0501703), National Natural Science Foundation of China (grant no. 31601074, 61872094, 61832019), and Shanghai Jiao Tong University School of Medicine (contract no. YG2017ZD14).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.00971/full#supplementary-material

REFERENCES

Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* 42 (3), 323–328. doi: 10.1016/j.jkss.2012.10.002

Bian, Y., Jing, Y., Wang, L., Ma, S., Jun, J. J., and Xie, X. Q. (2019). Prediction of orthosteric and allosteric regulations on cannabinoid receptors using supervised machine learning classifiers. *Mol. Pharm.* 16 (6), 2605–2615. doi: 10.1021/acs.molpharmaceut.9b00182

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.-Theory E.* 2008 (10), P10008. doi: 10.1088/1742-5468/2008/10/P10008

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi: 10.1023/A:1010933404324

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA: ACM). doi: 10.1145/2939672.2939785

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23 (6), 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Chen, L., Zeng, W. M., Cai, Y. D., Feng, K. Y., and Chou, K. C. (2012). Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical–chemical interactions and similarities. *PLoS One* 7 (4), e35254. doi: 10.1371/journal.pone.0035254

Cheng, X., Zhao, S. G., Xiao, X., and Chou, K. C. (2017a). iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget* 8 (35), 58494–58503. doi: 10.18632/oncotarget.17028

Cheng, X., Zhao, S. G., Xiao, X., and Chou, K. C. (2017b). iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics* 33 (3), 341–346. doi: 10.1093/bioinformatics/btw644

Chou, K. C. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* 9 (6), 1092–1100. doi: 10.1039/c3mb25555g

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi: 10.1007/BF00994018

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings*, 886–893. doi: 10.1109/CVPR.2005.177

Dunkel, M., Gunther, S., Ahmed, J., Wittig, B., and Preissner, R. (2008). SuperPred: drug classification and target prediction. *Nucleic Acids Res.* 36, W55–W59. doi: 10.1093/nar/gkn307

Feng, P., Zhang, J., Tang, H., Chen, W., and Lin, H. (2017). Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions. *Interdiscip. Sci.* 9 (4), 540–544. doi: 10.1007/s12539-016-0193-4

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63 (1), 3–42. doi: 10.1007/s10994-006-6226-1

Gibaja, E., and Ventura, S. (2014). Multi-label learning: a review of the state of the art and ongoing research. *WIREs Data Mining Knowl. Discov.* 4 (6), 411–444. doi: 10.1002/widm.1139

Gurulingappa, H., Kolarik, C., Hofmann-Apitius, M., and Fluck, J. (2009). Concept-based semi-automatic classification of drugs. *J. Chem. Inf. Model.* 49 (8), 1986–1992. doi: 10.1021/ci9000844

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* 19 (1), 306. doi: 10.1186/s12859-018-2321-0

Henken, S., Bohling, J., Martens-Lobenhoffer, J., Paton, J. C., Ogunniyi, A. D., Briles, D. E., et al. (2010). Efficacy profiles of daptomycin for treatment of invasive and noninvasive pulmonary infections with *Streptococcus pneumoniae*. *Antimicrob. Agents Chemother.* 54 (2), 707–717. doi: 10.1128/AAC.00943-09

Hutchinson, J. M., Patrick, D. M., Marra, F., Ng, H., Bowie, W. R., Heule, L., et al. (2004). Measurement of antibiotic consumption: a practical guide to the use of the anatomical thgerapeutic chemical classification and definied daily dose system methodology in Canada. *Can. J. Infect. Dis.* 15 (1), 29–35. doi: 10.1155/2004/389092

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," (Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.).

Kotera, M., Hirakawa, M., Tokimatsu, T., Goto, S., and Kanehisa, M. (2012). The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.* 802, 19–39. doi: 10.1007/978-1-61779-400-1_2

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* 20 (12), 1983–1992. doi: 10.1109/TVCG.2014.2346248

Li, X., Xu, Y., Lai, L., and Pei, J. (2018). Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. *Mol. Pharm.* 15 (10), 4336–4345. doi: 10.1021/acs.molpharmaceut.8b00110

Li, Y., Niu, M., and Zou, Q. (2019). ELM-MHC: an improved MHC identification method with extreme learning machine algorithm. *J. Proteome Res.* 18 (3), 1392–1401. doi: 10.1021/acs.jproteome.9b00012

Lumini, A., and Nanni, L. (2018). Convolutional neural networks for ATC classification. *Curr. Pharm. Des.* 24 (34), 4007–4012. doi: 10.2174/1381612824666181112113438

MacDonald, K., and Potvin, K. (2004). Interprovincial variation in access to publicly funded pharmaceuticals: a review based on the WHO Anatomical Therapeutic Chemical Classification System. *Can. Pharm. J.* 137 (7), 29–34. doi: 10.1177/171516350413700703

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognit.* 45 (9), 3084–3104. doi: 10.1016/j.patcog.2012.03.004

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3, 80. doi: 10.3389/fenvs.2015.00080

Michielan, L., Terfloth, L., Gasteiger, J., and Moro, S. (2009). Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J. Chem. Inf. Model.* 49 (11), 2588–2605. doi: 10.1021/ci900299a

Moyano, J. M., Gibaja, E. L., Cios, K. J., and Ventura, S. (2018). Review of ensembles of multi-label classifiers: models, experimental study and prospects. *Inf. Fusion* 44, 33–45. doi: 10.1016/j.inffus.2017.12.001

Nanni, L., and Brahnam, S. (2017). Multi-label classifier based on histogram of gradients for predicting the anatomical therapeutic chemical class/classes of a given compound. *Bioinformatics* 33 (18), 2837–2841. doi: 10.1093/bioinformatics/btx278

Niu, Y., and Zhang, W. (2017). Quantitative prediction of drug side effects based on drug-related features. *Interdiscip. Sci.* 9 (3), 434–444. doi: 10.1007/s12539-017-0236-5

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2018). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 34 (9), 1473–1480. doi: 10.1093/bioinformatics/btx822

Qiao, Y., Xiong, Y., Gao, H., Zhu, X., and Chen, P. (2018). Protein–protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinformatics* 19 (1), 14. doi: 10.1186/s12859-018-2009-5

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* 76 (3), 036106. doi: 10.1103/PhysRevE.76.036106

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Mach. Learn.* 85 (3), 333–359. doi: 10.1007/s10994-011-5256-5

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., and Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Trans. Neural Netw.* 1 (4), 296–298. doi: 10.1109/72.80266

Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* 18 (1), 277. doi: 10.1186/s12859-017-1700-2

Szymanski, P., and Kajdanowicz, T. (2019). Scikit-multilearn: a scikit-based Python environment for performing multi-label classification. *J. Mach. Learn. Res.* 20 (1), 209–230.

Szymański, P., Kajdanowicz, T., and Kersting, K. (2016). How is a data-driven approach better than random choice in label space division for multi-label classification? *Entropy* 18 (8), 282. doi: 10.3390/e18080282

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2011). Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* 23 (7), 1079–1089. doi: 10.1109/TKDE.2010.164

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17 (17–18), 1700262. doi: 10.1002/pmic.201700262

Wang, N. N., Huang, C., Dong, J., Yao, Z. J., Zhu, M. F., Deng, Z. K., et al. (2017). Predicting human intestinal absorption with modified random forest approach: a comprehensive evaluation of molecular representation, unbalanced data, and applicability domain issues. *RSC Adv.* 7 (31), 19007–19018. doi: 10.1039/C6RA28442F

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2019a). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35 (8), 1326–1333. doi: 10.1093/bioinformatics/bty824

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019b). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*. doi: 10.1093/bioinformatics/btz408

Wu, L., Ai, N., Liu, Y., Wang, Y., and Fan, X. (2013). Relating anatomical therapeutic indications by the ensemble similarity of drug sets. *J. Chem. Inf. Model.* 53 (8), 2154–2160. doi: 10.1021/ci400155x

Xiao, X., Min, J. L., Wang, P., and Chou, K. C. (2013). iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theor. Biol.* 337, 71–79. doi: 10.1016/j.jtbi.2013.08.013

Xiong, Y., Qiao, Y., Kihara, D., Zhang, H. Y., Zhu, X., and Wei, D. Q. (2019). Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates. *Curr. Drug Metab.* 20 (3), 229–235. doi: 10.2174/1389200219666181019094526

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. Q. (2018). PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9, 2571. doi: 10.3389/fmicb.2018.02571

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Xu, Y., Chen, P., Lin, X., Yao, H., and Lin, K. (2018). Discovery of CDK4 inhibitors by convolutional neural networks. *Future Med. Chem.* 11, 165–177 doi: 10.4155/fmc-2018-0478

You, R., Yao, S., Xiong, Y., Huang, X., Sun, F., Mamitsuka, H., et al. (2019). NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Res.* 47 (W1), W379–W387. doi: 10.1093/nar/gkz388

You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., and Zhu, S. (2018). GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* 34 (14), 2465–2473. doi: 10.1093/bioinformatics/bty130

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. (2016). DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 32 (12), i18–i27. doi: 10.1093/bioinformatics/btw244

Zhang, M. L., and Zhou, Z. H. (2005). A k-nearest neighbor based algorithm for multi-label classification. *2005 IEEE International Conference on Granular Computing, Vols 1 and 2*, 718–721. doi: 10.1109/GRC.2005.1547385

Zhang, M. L., and Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* 26 (8), 1819–1837. doi: 10.1109/TKDE.2013.39

Zhang, W., Liu, F., Luo, L., and Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinformatics* 16, 365. doi: 10.1186/s12859-015-0774-y

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA–protein interactions. *PLoS Comput. Biol.* 14 (12), e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, W., Zhu, X., Fu, Y., Tsuji, J., and Weng, Z. (2017). Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. *BMC Bioinformatics* 18 (Suppl 13), 464. doi: 10.1186/s12859-017-1875-6

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. *RNA* 25 (2), 205–218. doi: 10.1261/rna.069112.118

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.