frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 November 2019
doi: 10.3389/fbioe.2019.00306

STS-NLSP: A Network-Based Label Space Partition Method for Predicting the Specificity of Membrane Transporter Substrates Using a Hybrid Feature of Structural and Semantic Similarity

Xiangeng Wang[1,2†], *Xiaolei Zhu*[3†], *Mingzhi Ye*[1], *Yanjing Wang*[1], *Cheng-Dong Li*[1], *Yi Xiong*[1*] *and Dong-Qing Wei*[1,2*]

[1] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Joint Laboratory of International Cooperation in Metabolic and Developmental Sciences, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China, *[2] Peng Cheng Laboratory, Shenzhen, China,* *[3] School of Sciences, Anhui Agricultural University, Hefei, China*

OPEN ACCESS

Edited by:
Wen Zhang,
Huazhong Agricultural University, China

Reviewed by:
Xiaoyang Jing,
Toyota Technological Institute at Chicago, United States
Balachandran Manavalan,
Ajou University, South Korea
Zhenhua Li,
National University of Singapore, Singapore

***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 Bioinformatics and Computational Biology,
a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *05 August 2019*
Accepted: *17 October 2019*
Published: *06 November 2019*

Citation:
Wang X, Zhu X, Ye M, Wang Y, Li C-D, Xiong Y and Wei D-Q (2019) STS-NLSP: A Network-Based Label Space Partition Method for Predicting the Specificity of Membrane Transporter Substrates Using a Hybrid Feature of Structural and Semantic Similarity.
Front. Bioeng. Biotechnol. 7:306.
doi: 10.3389/fbioe.2019.00306

Membrane transport proteins play crucial roles in the pharmacokinetics of substrate drugs, the drug resistance in cancer and are vital to the process of drug discovery, development and anti-cancer therapeutics. However, experimental methods to profile a substrate drug against a panel of transporters to determine its specificity are labor intensive and time consuming. In this article, we aim to develop an *in silico* multi-label classification approach to predict whether a substrate can specifically recognize one of the 13 categories of drug transporters ranging from ATP-binding cassette to solute carrier families using both structural fingerprints and chemical ontologies information of substrates. The data-driven network-based label space partition (NLSP) method was utilized to construct the model based on a hybrid of similarity-based feature by the integration of 2D fingerprint and semantic similarity. This method builds predictors for each label cluster (possibly intersecting) detected by community detection algorithms and takes union of label sets for a compound as final prediction. NLSP lies into the ensembles of multi-label classifier category in multi-label learning field. We utilized Cramér's V statistics to quantify the label correlations and depicted them via a heatmap. The jackknife tests and iterative stratification based cross-validation method were adopted on a benchmark dataset to evaluate the prediction performance of the proposed models both in multi-label and label-wise manner. Compared with other powerful multi-label methods, ML-*k*NN, MTSVM, and RA*k*EL*d*, our multi-label classification model of NLPS-RF (random forest-based NLSP) has proven to be a feasible and effective model, and performed satisfactorily in the predictive task of transporter-substrate specificity. The idea behind NLSP method is intriguing and the power of NLSP remains to be explored for the multi-label learning problems in bioinformatics. The benchmark dataset, intermediate results and python code which can fully reproduce our experiments and results are available at https://github.com/dqwei-lab/STS.

Keywords: membrane transporter, substrate specificity, structural fingerprint, chemical ontology, multi-label classification

INTRODUCTION

Membrane transport proteins, also known as transporters or carriers, are a diverse and large group of proteins that transport various hydrophilic molecules, encompassing ions and small molecules across lipid bilayers within a cell or between cells, thus playing crucial roles in various biological functions, such as binding with small molecules in extracellular space, which is the key component to determine the bioavailability and biological activity of chemicals, i.e., their adverse and therapeutic effects (International Transporter et al., 2010). In recent years, a number of efflux and influx transporters from ATP binding cassette (ABC) (Chen et al., 2016) and solute carrier (SLC) (Nyquist et al., 2017) families have attracted significant interest, since they are of vital importance in determining the ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of a wide range of drugs and xenobiotics. More importantly, membrane proteins are the major media of multi-drug resistance in cancer (Szakács et al., 2006; Fletcher et al., 2010). For example, multi-drug resistance protein 1 (MDR1; aka P-glycoprotein and ABCB1) is overexpressed in many malignant neoplasms and its expression can also be induced by chemotherapy. The overexpression of MDR1 has proven to be correlated with drug resistance in breast, prostate and lung cancer (Holohan et al., 2013). To make things worse, widely-applied targeted drugs such as nilotinib, imatinib, sunitinib, and erlotinib are also identified as regulators and substrates for specific transporters. Thus, understanding the specificity of transporter substrates (identification of potential transporters for existing and novel drug molecules at the early phase of drug discovery process) is not only momentous to the discovery and development of safe and efficacious drugs but also helpful to identify potential drug resistance in anti-cancer therapeutics. However, experimental methods to profile compounds against a panel of transporters are time- and resource-consuming. It should be of high value to develop *in silico* classification models to predict the specificity of membrane transporter substrates.

Generally, two major categories of computational approaches are utilized to predict potential transporters involved in membrane transport of chemicals (Shaikh et al., 2017). The first type of approaches are receptor-based methods, which evaluate the interaction details between transporters and drug molecules via available three-dimensional structures of macromolecules. However, these approaches are hindered by the scarcity of the high-resolution structures of membrane transporters, which are generally difficult to be resolved by experimental technologies. The second category of approaches are ligand-based methods (Chakraborty et al., 2017), via the structural likeness of ligands to known substrates. The most commonly applied ligand-based approach is the quantitative structure-activity relationship (SAR or QSAR) model, which aims to build a mapping from molecular descriptors of ligands to biological functions (e.g., whether the compound is a specific transporter substrate). Many SAR or QSAR models have been built to classify substrates and non-substrates for a specific type of transporters, such as P-glycoprotein (P-gp/MDR1/ABCB1) (Huang et al., 2007; Wang et al., 2011; Poongavanam et al., 2012; Li et al., 2014), BCRP/ABCG2 (Zhong et al., 2011; Hazai et al., 2013; Gantner et al., 2017), MRP1/ABCC1 (Lingineni et al., 2017) by a variety of machine learning models, including linear models, neural networks, support vector machines (SVM), and etc. Li et al. (2014) developed the naïve Bayesian classifier to predict potential P-gp substrates using simple molecular properties, topological descriptors, and structural fingerprints on a compiled dataset of 423 P-gp substrates and 399 non-substrates. Hazai et al. (2013) developed an SVM classification model for prediction of BCRP substrates on a dataset composed of 164 BCRP substrates and 99 non-substrates.

However, these traditional QSAR models only consider a single type of carrier at a time. With the ever-accumulating high-quality data of various drug transporters, it is superior to assign a compound into the maximum possible number of transporters. The failure of clinical trials on MDR1 inhibitors such as tariquidar (Pusztai et al., 2005) and zosuquidar (Cripe et al., 2010) also suggests that, in order to block the potential drug efflux of cancer cell entirely, we need to consider the specificity of as much transporters as possible in the design phase of new drugs. Thanks to the efforts conducted by Mak et al. (2015), the interaction data on various types of transporters and their substrates and modulators were curated on Metrabase database exploited for QSAR modeling. In addition to the data from Metrabase, Shaikh et al. (2017) further retrieved data of ABCG2, MDR1 and MRP1 from the literature, to construct a benchmark dataset of substrates and non-substrates of the 13 transporters from ABC and SLC families. In their recent study (Shaikh et al., 2017), they employed proteochemometric (PCM) modeling technique to enable simultaneous consideration of multiple transporters. They built PCM- and QSAR-based predictive models for the transporter-substrate specificity of pharmaceutically important membrane transporters. In those models, the physicochemical, topological descriptors of ligand molecules, MACCS and variants of Morgan fingerprints were used as input features.

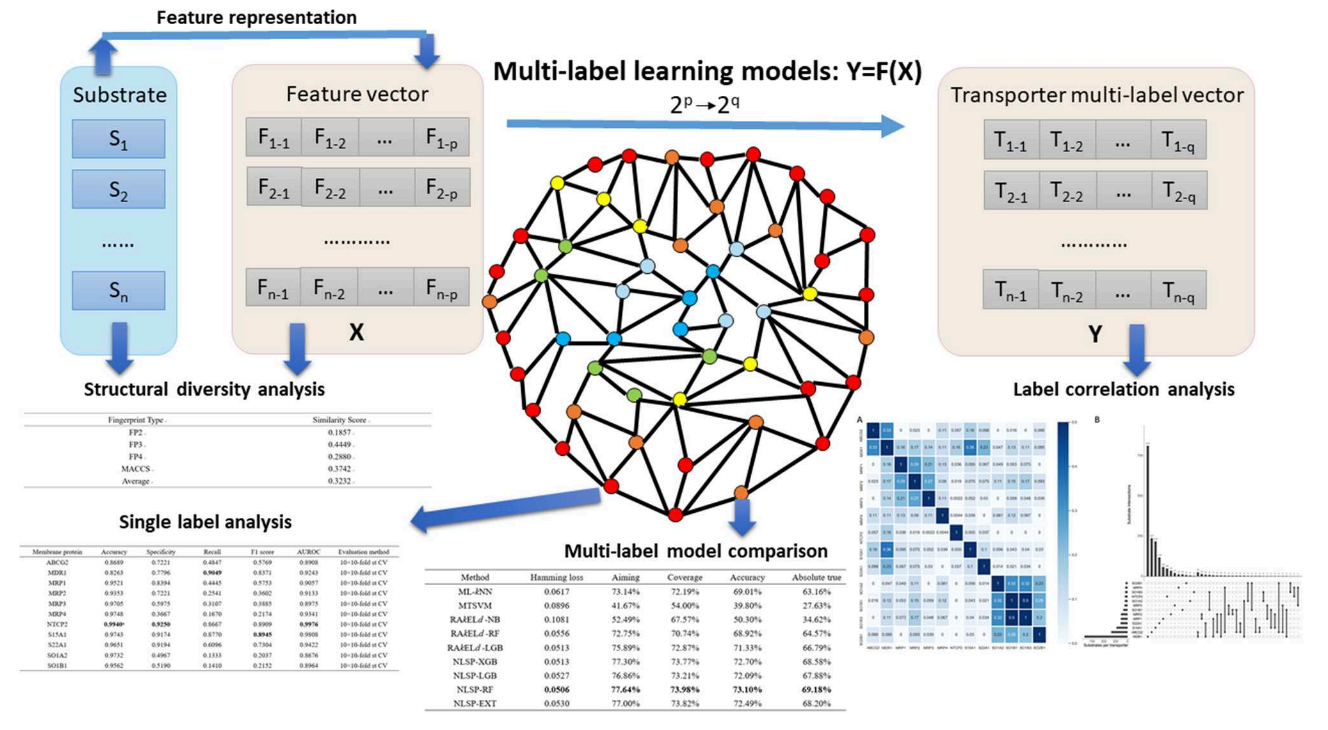
Inspired by the successful application of multi-label classification systems in the classification of drugs (Chen et al., 2014), we formulated the problem of transporter-substrate specificity as a multi-label classification task since some compounds can be substrates of more than one transporters. Typically, multi-label classification (MLC) models are divided into three major groups: algorithm adaptation, problem transformation, and ensembles of multi-label classifier (EMLC). Algorithm adaptation methods incorporate specific tricks that convert traditional single-label learning classifiers into multi-label ones. The representative model of this group is ML-*k*NN (Zhang and Zhou, 2005). For the problem transformation method, it converts multi-label learning tasks into one or several single-label problems. For example, label powerset (LP) is a method of problem transformation, which trains models on each possible subset of label sets (Gibaja and Ventura, 2014). For a dataset with high cardinality in label set, LP is prone to overfitting because of the exponentially increased number of subsets. To tackle the overfitting nature of label powerset, Tsoumakas et al. (2011) try to segment the label space into subspaces and apply label powerset in these subspaces. They proposed the RA*k*EL*d* method, which cuts the label set into *k* disjoint 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 training data. The Network-based Label Space Partition (NLSP) (Szymanski et al., 2016) is an EMLC built upon LP, and it divides the label sets into *n* small-sized label sets (possibly intersecting) by community detection method which can incorporate the label correlation structures in training set, such that learning *k* representative LP classifiers. As a result, NLSP tackles much less subsets compared to LP and selects *k* in a data-driven manner. For a more detailed explanation of multi-label learning, refer to Zhang and Zhou (2014), Moyano et al. (2018).

In the present study, we developed an *in-silico* method for predicting the Specificity of membrane Transporter Substrates based on the Network-based Label Space Partition algorithm, termed STS-NLSP, which has both unleashed the correlation among labels and integrated two types of similarity-based features. Specifically, a given compound substrate was classified as one or more of the following classes of transporters (Shaikh et al., 2017): (i) ABCG2; (ii) MDR1; (iii) MRP1; (iv) MRP2; (v) MRP3; (vi) MRP4; (vii) NTCP2; (viii) S15A1; (ix) S22A1; (x) SO1A2; (xi) SO1B1; (xii) SO1B3; (xiii) SO2B1. In order to represent the information of substrates, we not only used the structural fingerprints, but also employed their biological information (i.e., chemical ontology), extracted from the ChEBI database (Degtyarenko et al., 2008). Then, we compared our NLSP-based methods to three different types of multi-label classification methods constructed on identical features. Our results demonstrated that the NLSP-RF model yielded out consistently better performance than another two types of methods using the jackknife test on the benchmark dataset, and we chose it as our final STS-NLSP. Label-wise analysis, validated via iterative stratification, of the final models was also performed for the convenience of experimental biologists. The major steps in the article are summarized in **Figure 1**.

RESULTS AND DISCUSSION

Structural Diversity Analysis

In the total of 1, 846 structural different substrates on the benchmark dataset, we calculated the similarity scores of four types of fingerprints, FP2, FP3, FP4, MACCS, and their average similarity score (SS) for each pair (1, 702,935 different pairs in total) of substrates. The higher the score was between two substrates, the more similar they were each other. Listed in **Table 1** were the average values of all pairs for the four type of similarity scores, and the average of these four types. The results demonstrated that the dataset of substrates was structurally different and diverse in terms of 2D fingerprints. We could thus put more confident on the representativeness of this dataset. The average similarity score of FP2 was lowest among the four types of fingerprints. Since the four types of fingerprints presented distinct attributes of the molecules, we used the average similarity



Fingerprint Type	Similarity Score
FP2	0.1857
FP3	0.4449
FP4	0.2880
MACCS	0.3742
Average	0.3232

Membrane protein	Accuracy	Specificity	Recall	F1 score	AUROC	Evaluation method
ABCG2	0.8689	0.7221	0.4847	0.5769	0.8908	10×10-fold st CV
MDR1	0.8263	0.7796	**0.9049**	0.8371	0.9243	10×10-fold st CV
MRP1	0.9521	0.8394	0.4445	0.5753	0.9057	10×10-fold st CV
MRP2	0.9353	0.7221	0.2541	0.3602	0.9133	10×10-fold st CV
MRP3	0.9705	0.5975	0.3107	0.3885	0.8975	10×10-fold st CV
MRP4	0.9748	0.3667	0.1670	0.2174	0.9341	10×10-fold st CV
NTCP2	**0.9940**	**0.9250**	0.8667	0.8909	**0.9976**	10×10-fold st CV
S15A1	0.9743	0.9174	0.8770	**0.8945**	0.9808	10×10-fold st CV
S22A1	0.9651	0.9194	0.6096	0.7304	0.9422	10×10-fold st CV
SO1A2	0.9732	0.4967	0.1333	0.2037	0.8676	10×10-fold st CV
SO1B1	0.9562	0.5190	0.1410	0.2152	0.8964	10×10-fold st CV

Method	Hamming loss	Aiming	Coverage	Accuracy	Absolute true
ML-*k*NN	0.0617	73.14%	72.19%	69.01%	63.16%
MTSVM	0.0896	41.67%	54.00%	39.80%	27.63%
RA*k*EL*d* -NB	0.1081	52.49%	67.57%	50.30%	34.62%
RA*k*EL*d* -RF	0.0556	72.75%	70.74%	68.92%	64.57%
RA*k*EL*d* -LGB	0.0513	75.89%	72.87%	71.33%	66.79%
NLSP-XGB	0.0513	77.30%	73.77%	72.70%	68.58%
NLSP-LGB	0.0527	76.86%	73.21%	72.09%	67.88%
NLSP-RF	**0.0506**	**77.64%**	**73.98%**	**73.10%**	**69.18%**
NLSP-EXT	0.0530	77.00%	73.82%	72.49%	68.20%

FIGURE 1 | Major steps in the article. Substrates, which were confirmed structural diverse, were featurized into numerical vectors, combined with corresponding transporter multi-label vectors, and then fed into different multi-label learning models. Label correlation analysis provided us insights on the interaction among transporters. To facilitate researchers working on specific membrane transporter, NLSP-RF, with consistently better multi-label performance metrics, was selected after multi-label model comparison for the transporter-wise (single label) analysis. For more detailed description, refer to the subsequent parts in this article.

Table 1 | The average SS of all pairs of substrates on the benchmark dataset for the four types of fingerprints.

Fingerprint type	Similarity score
FP2	0.1857
FP3	0.4449
FP4	0.2880
MACCS	0.3742
Average	0.3232

score to represent their 2D fingerprint similarity for each pair of substrates.

Label Correlation Analysis

One primary merit of multi-label learning vis-à-vis single-label learning framework is the explicit utilization of label correlations (Zhang and Zhou, 2014). Bias corrected Cramér's V statistics were calculated for all the possible label pairs and depicted in **Figure 2A**. The UpSet visualization (Lex et al., 2014) of label-set intersections is shown in **Figure 2B**. We found 25 substrates are both transported by MDR1 and ABCG2, which is intuitive because MDR1 and ABCG2 are both in the superfamily of ATP-binding cassette transporters. One major common substrate of MDR1 and ABCG2 is gefitinib (Maemondo et al., 2010), which is the first-line targeted chemotherapy agent for non–small-cell lung cancer. Elevated MDR1 and ABCG2 expression has been demonstrated to confer acquired resistance in in EGFR-expressing cancer cells (Chen et al., 2011). The medical implications of co-transport of MDR1 and ABCG2 in cancer has been already noticed by clinicians and basic researchers. We also found several label sets are correlated, especially for SOB1B1 and SOB1B3, of which the Cramér's V statistic is 0.5. Details about the pair-wise intersection numbers of substrates and the pair-wise Cramér's V statistics between all the transporters are shown in **Tables S1, S2**.

Multi-Label Model Comparison

We compared the prediction performance of NLSP-based models to another three classification methods (i.e., ML-*k*NN, MTSVM and RA*k*EL*d*-based models) on the identification of specificity of transporter substrates. The classification performances of all the models on the benchmark dataset using jackknife test were shown in **Table 2**. We found NLSP-RF (random forest-based NLSP) is consistently better than the other models in all the five predefined multi-label measures. On the other hand, we found all the NLSP-based methods perform consistently better than other models, and the MTSVM is the most unsatisfactory model. For the RA*k*EL*d*-based methods, we found the choice of base-learners will have huge impact on the model performance. Therefore, we selected the NLSP-RF as the classification engine to construct the final prediction model. To get deeper insights of this predictive task, we compared the mean feature importance (Gini index) of structural similarity- and semantic similarity-based features on the final prediction model. We found the structural similarity-based features are significantly ($p < 10^{-7}$) more important than semantic similarity-based features (**Figure 3**), suggesting the selectivity of chemicals among different transporters majorly hinges on the 2D structure of chemicals.

Single-Label Analysis

As for experimental biologist working on specific membrane protein, it is useful to evaluate multi-label learning models for each label respectively (Michielan et al., 2009; Mayr et al., 2016). We utilized the hyperparameters of the best-performing multi-label model of NLSP-RF and performed 10 times repeated 10-fold stratified cross validation (10 × 10-fold st CV) (Sechidis et al., 2011), because the jackknife test is rather time-consuming and tends to overestimate different performance measures (Kohavi, 1995), The details are listed in **Table 3**. We found NLSP-RF perform well in all the single-label subtasks from the viewpoint of accuracy and AUROC, but perform worse in the prediction subtask of MRP2, MRP3, MRP4, SO1A2, SOB1B1 in view of F1 score, which is intuitive because our benchmark dataset is highly-imbalanced for these five proteins. We also compared our model with the previous results from Shaikh et al. (2017). Although we did not manually collect equal-sized negative data for each transporter, our model performs similarly well except for the subtasks suffering from imbalance learning problem.

Comparison With Previous Studies

In this article, the benchmark dataset proposed by Shaikh et al. (2017) was compiled and implemented to test our multi-label classification method. The differences between our method and Shaikh's method (Shaikh et al., 2017) were summarized in **Table 4**. To our best knowledge, it is the first study incorporating the prediction of the specificity of membrane transporter substrates into multi-label learning framework, whereas previously published methods were constructed as single-label systems. Compared with the single-label systems, it is much trickier to develop predictive models within multi-label learning framework. In the single-label systems, a balanced dataset of substrates (positive samples) and non-substrates (negative samples) were usually constructed for each single transporter, which can result in overestimated prediction performance than the actual cases where the number of substrates is significantly lower than that of non-substrates. It has been noticed that an increasing number of compounds are simultaneously assigned as substrates of multiple (two or more different) transporters. Using the multi-label system, our model extends the discriminative classes from 1 to 13 at a time.

Although it is much more complicated and challenging to deal with, our proposed model based on the multi-label system has two main advantages. Firstly, it can simultaneously predict multiple transporters of a given compound as the substrate. Secondly, it does not need prepare the datasets of non-substrates for each single transporter, as the single-label system does, because one positive instance of one transporter could possibly be a negative sample for another. Especially, the single-label systems will take a lot of labor work to manually collect the same number of non-substrates with the increasing available substrates. The multi-label systems can avoid the labor work to build the datasets of non-substrates due to its innate negative

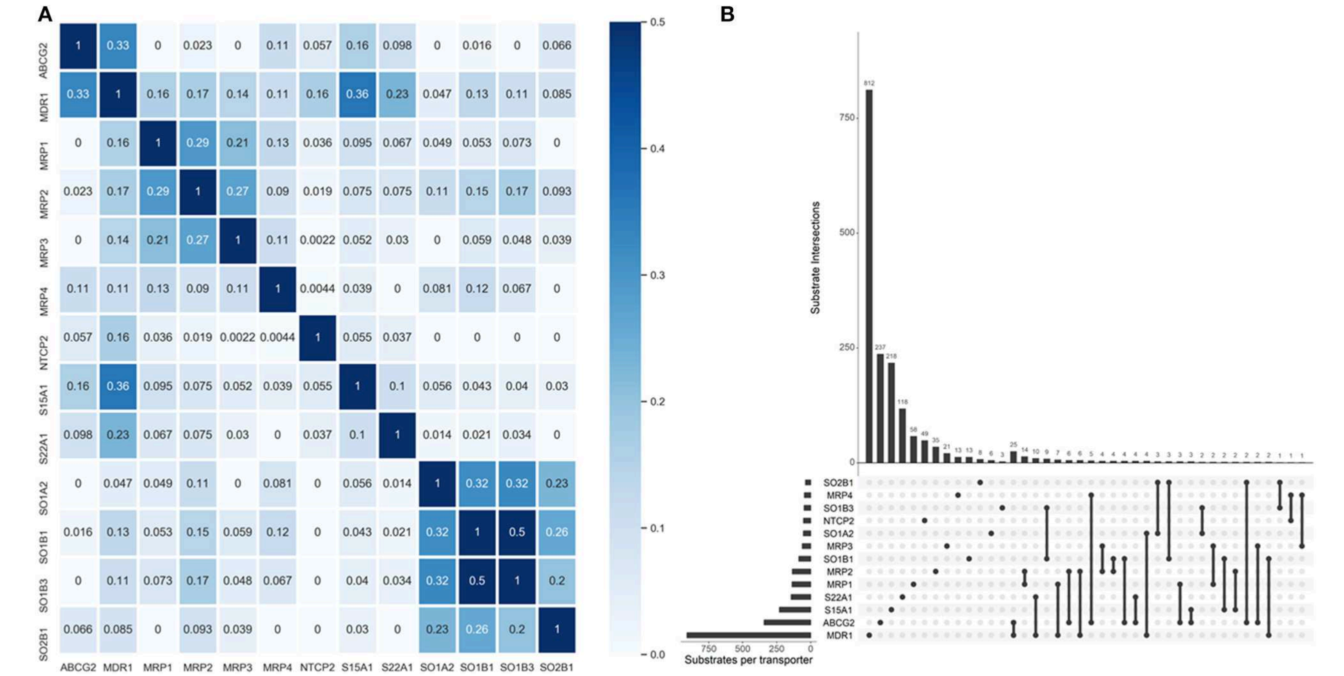


FIGURE 2 | Label correlation landscape. **(A)** The pair-wise heatmap visualization of Cramér's V statistics. **(B)** The UpSet visualization of label intersections. The horizontal bars show the number of substrates per transporter and the vertical bars show the number of substrates per transporter category intersection. The filled dots denote the transporter whose exclusive substrates are counted in the corresponding vertical bars. The vertical lines stand for the intersection of substrates of specific transporters. More dots they encompass, more intersections are considered for the tallying of the corresponding vertical bars.

TABLE 2 | Performance comparison of various multi-label classification methods.

Method	Hamming loss	Aiming	Coverage	Accuracy	Absolute true
ML-*k*NN	0.0617	73.14%	72.19%	69.01%	63.16%
MTSVM	0.0896	41.67%	54.00%	39.80%	27.63%
RA*k*EL*d* -NB	0.1081	52.49%	67.57%	50.30%	34.62%
RA*k*EL*d* -RF	0.0556	72.75%	70.74%	68.92%	64.57%
RA*k*EL*d* -LGB	0.0513	75.89%	72.87%	71.33%	66.79%
NLSP-XGB	0.0513	77.30%	73.77%	72.70%	68.58%
NLSP-LGB	0.0527	76.86%	73.21%	72.09%	67.88%
NLSP-RF	**0.0506**	**77.64%**	**73.98%**	**73.10%**	**69.18%**
NLSP-EXT	0.0530	77.00%	73.82%	72.49%	68.20%

The bold value stands for the best value of specific metrics in these models.

nature among labelset. We believe that the multi-label system proposed in our study will further benefit the research about the specificity of membrane transporter substrates, especially for the drug resistance screening in cancer research.

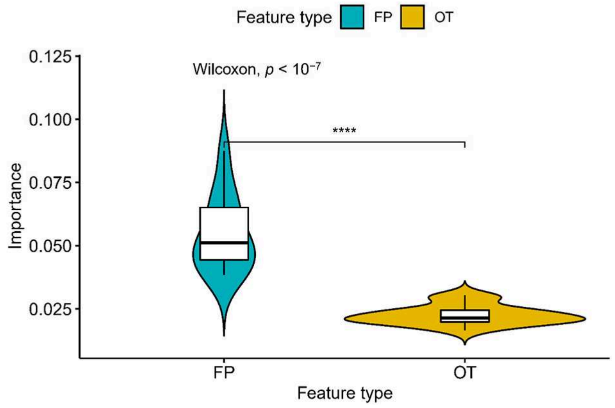


FIGURE 3 | Comparison of feature importance between structural similarity- and semantic similarity-based features. "FP," fingerprint, stands for structural similarity-based features. "OT," ontology, stands for semantic similarity-based features. ****$p < 0.0001$.

MATERIALS AND METHODS

Benchmark Dataset

We utilized the same benchmark dataset proposed by Shaikh et al. (2017) to evaluate the performance of the proposed models, which contains 2,293 small molecules classified into 13 main classes of transporter substrates. The chemical structures of those small molecules were identified by Simplified Molecular Input Line Entry Specification (SMILES). The detailed composition of the benchmark dataset was listed in **Table 5**. Thus, the benchmark dataset $\mathbb{S}$ can be formulated as

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \ldots \cup \mathbb{S}_i \cup \ldots \cup \mathbb{S}_{12} \cup \mathbb{S}_{13} \quad (1)$$

TABLE 3 | Label-wise analysis of best-performing multi-label learning model.

Membrane protein	Accuracy	Specificity	Sensitivity	CCR	F1 score	AUROC	Evaluation method
ABCG2	0.8689	0.7221	0.4847	0.6034	0.5769	0.8908	10 × 10-fold st CV
MDR1	0.8263	0.7796	**0.9049**	0.8422	0.8371	0.9243	10 × 10-fold st CV
MRP1	0.9521	0.8394	0.4445	0.6419	0.5753	0.9057	10 × 10-fold st CV
MRP2	0.9353	0.7221	0.2541	0.4881	0.3602	0.9133	10 × 10-fold st CV
MRP3	0.9705	0.5975	0.3107	0.4541	0.3885	0.8975	10 × 10-fold st CV
MRP4	0.9748	0.3667	0.1670	0.2668	0.2174	0.9341	10 × 10-fold st CV
NTCP2	**0.9940**[a]	**0.9250**	0.8667	0.8958	0.8909	**0.9976**	10 × 10-fold st CV
S15A1	0.9743	0.9174	0.8770	**0.8972**	**0.8945**	0.9808	10 × 10-fold st CV
S22A1	0.9651	0.9194	0.6096	0.7645	0.7304	0.9422	10 × 10-fold st CV
SO1A2	0.9732	0.4967	0.1333	0.3150	0.2037	0.8676	10 × 10-fold st CV
SO1B1	0.9562	0.5190	0.1410	0.330	0.2152	0.8964	10 × 10-fold st CV
ABCG2	0.76	0.756	0.764	0.76	0.77	Not available	5-fold cv[b]
MDR1	0.776	0.798	0.751	0.775	0.761		5-fold cv[b]
MRP1	0.826	0.844	0.812	0.828	0.841		5-fold cv[b]
MRP2	0.814	0.886	0.746	0.816	0.804		5-fold cv[b]
MRP3	0.869	0.855	0.885	0.87	0.868		5-fold cv[b]
MRP4	0.905	0.857	0.949	0.903	0.914		5-fold cv[b]
NTCP2	0.93	0.93	0.93	0.93	0.93		5-fold cv[b]
S15A1	0.847	0.819	0.869	0.844	0.864		5-fold cv[b]
S22A1	0.844	0.875	0.813	0.844	0.84		5-fold cv[b]
SO1A2	0.711	0.979	0.419	0.699	0.581		5-fold cv[b]
SO1B1	0.776	0.726	0.829	0.777	0.784		5-fold cv[b]

[a] *The bold value stands for the best value of specific metrics in the model of NLSP-RF.*
[b] *5-fold cv results are from Shaikh et al. (2017).*

TABLE 4 | Methodological differences between Shaikh's method, and our present method (STS-NLSP).

Difference	Shaikh's method (Shaikh et al., 2017)	STS-NLSP
Learning framework	Single-label learning	Multi-label learning
Machine learning method	SVM, random forest, etc.	NLSP
Dataset distribution	A balanced number of substrates and non-substrates for each single transporter, respectively	Substrates categorized into 13 transporters with an imbalanced distribution (910 substrates for a majority of transporter MDR1, and 39 substrates for a minority of transporter SO2B1)
Features	Molecular descriptors, molecular fingerprints and Sequence-based descriptors for transporter proteins	Average similarity score fingerprints, and semantic similarity
Evaluation metrics	Recall, Specificity, Precision, Accuracy, F1 score, MCC	Aiming, Coverage, Accuracy, Absolute True, Absolute False
Validation method	Five-fold cross validation and independent test using an unseen external set	Jackknife test

where the subset $\mathbb{S}_i$ includes the samples from the i-th transporter ($i = 1, 2, \ldots, 13$), and ∪ stands for the symbol for "union" in the set theory.

Measuring Label Correlation

In order to evaluate the association between two labels, we calculated the bias corrected Cramér's V statistic for all the label pairs (Bergsma, 2013). Cramér's V (also referred to as Cramér's phi, denoted as ϕ_c) statistic is a measure of association between two categorical variables, ranging from 0 to 1 (inclusive). But it is shown that sample Cramér's V tends to overestimate the correlation compared to its population counterpart (Bergsma, 2013). 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).

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

where

$$\tilde{\varphi}^2 = \max\left(0, \varphi^2 - \frac{(r-1)\,(c-1)}{n-1}\right), \qquad (3)$$

TABLE 5 | Anatomy of the benchmark dataset $\mathbb{S}$ according to the 13 classes of transporter substrates (see Equation 1). See **Supporting Information** for further explanation.

Subset	Name	Description	Substrates
$\mathbb{S}_1$	ABCG2	ATP-binding cassette subfamily G member 2 (BCRP)	344
$\mathbb{S}_2$	MDR1	Multidrug resistance protein 1 (P-glycoprotein 1)	910
$\mathbb{S}_3$	MRP1	Multidrug resistance-associated protein 1	138
$\mathbb{S}_4$	MRP2	Multidrug resistance-associated protein 2	136
$\mathbb{S}_5$	MRP3	Multidrug resistance-associated protein 3	63
$\mathbb{S}_6$	MRP4	Multidrug resistance-associated protein 4	47
$\mathbb{S}_7$	NTCP2	Sodium/taurocholate cotransporter	53
$\mathbb{S}_8$	S15A1	Solute carrier family 15 member 1 (peptide transporter 1)	230
$\mathbb{S}_9$	S22A1	Solute carrier family 22 member 1 (organic cation transporter 1)	144
$\mathbb{S}_{10}$	SO1A2	Solute carrier organic anion transporter family member 1A2	54
$\mathbb{S}_{11}$	SO1B1	Solute carrier organic anion transporter family member 1B1	87
$\mathbb{S}_{12}$	SO1B3	Solute carrier organic anion transporter family member 1B3	48
$\mathbb{S}_{13}$	SO2B1	Solute carrier organic anion transporter family member 2B1	39
Number of total virtual substrates			2,293[a]
Number of total structural different substrates			1,846[b]

[a] *The number of virtual substrates is calculated as follows: for a structurally same substrate, its contribution to the total number of virtual substrates is 2 if it occurs in two different classes of transporter substrates; that is 3 if it occurs in three different classes of transporter substrates; and so forth.*

[b] *Of the 1,846 structural different substrates, 1,591 belong to one class, 145 to two classes, 62 to three classes, 28 to four classes, 12 to five classes, and 4 to six classes, 3 to seven classes, and 1 to nine classes. Refer to **Supporting Information** for elaborated information of substrates listed in each of 13 classes.*

$$\varphi^2 = \frac{\chi^2}{n} \tag{4}$$

and

$$\tilde{m} = \min(\tilde{r} - 1, \tilde{c} - 1)\ , \tag{5}$$

$$\tilde{r} = r - \left(\frac{(r-1)^2}{n-1}\right)\ , \tag{6}$$

$$\tilde{c} = c - \left(\frac{(c-1)^2}{n-1}\right)\ . \tag{7}$$

Feature Representation

We are to describe the effective formulization of samples in the training and testing datasets in this section. Now, let us address this from both structural and biological (i.e., chemical ontology) angles.

Features to Reflect Structural Similarity

The simple 2D fingerprint was chosen to represent the structural characteristics of small molecules, since it not only has high efficiency on the measurement of inter-molecular structural similarity, but also it has achieved effectiveness in similarity search, virtual screening and QSAR studies, despite its neglect of information about the target-ligand interactions, in comparison to 3D shape and docking methods (Duan et al., 2010; Xiao et al., 2013). In this study, four different types of fingerprints were generated by Open Babel (O'Boyle et al., 2011), which are MACCS, FP2, FP3, and FP4, on the basis of SMILES for each substrate. These fingerprints were binary strings, which encode the presence or absence of sub-structural fragments. Given two substrates, their fingerprint similarity was defined by Tanimoto coefficient (Keum et al., 2016),

$$TC = \frac{c}{a + b - c} \tag{8}$$

where a and b are the number of bits set in substrate bit-strings, c strands for the number of bits shared by two substrates. The structural similarity score between any pair of two substrates was calculated by the average Tanimoto coefficients of the four types of fingerprints between them. A specific sample is formulated as a 13-D vector via its maximum structural similarity score with those in each of the 13 subsets,

$$D^{StrSim} = [\alpha_1\ \alpha_2\ \alpha_3\ \ldots\ \alpha_{13}]^T \tag{9}$$

where α_1 denotes its maximum structural similarity score with the substrates in the subset $\mathbb{S}_1$, α_2 for that in the subset $\mathbb{S}_2$, and so on.

Features to Reflect Semantic Similarity

In the present study, we utilized the ontology information of compounds, named as ChEBI ontology (Degtyarenko et al., 2008), which was similar to gene ontology, to incorporate the semantic information. ChEBI provides an ontology database of chemical entities with curated biological annotations. The ChEBI ontology information was retrieved from ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/ ("chebi.obo," July 2017). Theoretically, ontologies are limited vocabularies can be conceived as graph structures consisting of "terms" forming the node set and "relations" of two terms forming the edge set. It consists of three separate subontologies, of which the roots will be "chemical entity," "role," and "subatomic particle," respectively (Hastings et al., 2013). As has been stated in a series of studies (Pesquita et al., 2009; Ferreira and Couto, 2010; Couto and Silva, 2011; Couto and Pinto, 2013), there are various ways to measure semantic similarity relying on information content (IC) between two entities based on an ontology. Given any compound which corresponds to a term c on the ChEBI ontology, let $p(c)$ be the usage frequency of the term c in some corpus. The information content of a term can be given by

$$IC(c) = -logp\,(c) \tag{10}$$

given two compounds c_1 and c_2, the following formula was used to measure the semantic similarity between them:

$$sim_{Lin}\,(c_1,\ c_2) = \frac{2 \times IC(c_{MICA})}{IC\,(c_1) + IC(c_2)} \tag{11}$$

where $MICA$ is their most informative common ancestor of both c_1 and c_2. A specific sample is formulated as a 13-D vector via its maximum semantic similarity score with those in each of the 13 subsets.

$$D^{SemSim} = [\beta_1\ \beta_2\ \beta_3\ \ldots\ \beta_{13}]^T \tag{12}$$

where β_1 means its maximum semantic similarity score with the substrates in the subset $\mathbb{S}_1$, β_2 for that in the subset $\mathbb{S}_2$, and so on.

Multi-Label Classification Methods

Network Based Label Space Partition

The NLSP is a newly proposed multi-label learning method and has achieved top performance in many predictive tasks (Szymanski et al., 2016). This method has also recently reached the top performance in the drug classification and enzyme-substrate selectivity prediction tasks by our group (Shan et al., 2019; Wang et al., 2019). Inspired by these current advances, we adopted the data-driven NLSP method for the prediction of specificity of membrane transporter substrates. NLSP divides the predictive modeling into training and classification phase

The training phase is divided into four parts. We firstly establish a label co-occurrence graph on the training set, which can be weighted or not. Then we detect the community on the label co-occurrence graph. There are various community detection algorithms. In this study, we utilized the largest modularity using incremental greedy search (Blondel et al., 2008) method and multiple async label propagation (Raghavan et al., 2007) to fulfill this task. Thirdly, for each community, a corresponding training set is generated by selecting the original dataset with label columns presented in the community. Finally, for each community, a base predictor is learnt on the training set. In this study, we compared the performance of five types of base predictors:

1. **Extremely randomized trees (ERT)** (Geurts et al., 2006) is a tree-based 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 reduces the variance of the model a bit, at the expense of a slightly greater increase in bias.
2. **Random forests (RF)** (Breiman, 2001; Manavalan et al., 2014, 2018b; Lv et al., 2019; Ru et al., 2019) is a tree-based ensemble method that combines the probabilistic predictions of a number of decision tree-based classifiers to improve the generalization ability over a single estimator.
3. **Support vector machine (SVM)** (Chang and Lin, 2011; Xiong et al., 2011, 2012; Sun et al., 2014; Manavalan and Lee, 2017; Manavalan et al., 2018d; Zhang et al., 2018; Meng et al., 2019) 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.
4. **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 sequentially from weak learners such as decision trees s, while XGB is an efficient and distributed implementation of gradient boosting machine.
5. **LightGBM (LGB)** (Ke et al., 2017; Xu et al., 2017; Liao et al., 2018) is another cutting-edge implementation of gradient boosting decision trees. Two innovative techniques, gradient-based one-side sampling and exclusive feature bundling are incorporated in the model training process, which has proven to achieve almost similar accuracy as XGB with up to over 20 times speed-up.

In the classification phase, we just perform predication on all the communities identified in the training phase and fetch the union of assigned labels. For more technical details refer to Szymanski et al. (2016).

Benchmark Methods

Inspired by the recent study (Cheng et al., 2017), we compared NLSP-base methods with another three cutting-edge multi-label classification methods, ML-*k*NN (Zhang and Zhou, 2007), MLTSVM (Chen et al., 2016) and RA*k*EL*d*-based methods (Tsoumakas et al., 2011). ML-*k*NN is a lazy learning model based on traditional *k*NN (Fukunaga and Hostetler, 1973). For a new data instance, it firstly finds the top-*k* closest samples in the training set. Secondly, it calculates the number of each label in the *k* samples. Thirdly, based on the aforementioned label number, it estimates the label probability by naïve Bayes method. Finally, the label probability is generated by maximum *a posteriori* estimation. MLTSVM is a variation of twin support vector machine designed for multi-label scenario proposed by Chen et al. (2016). As for twin support vector machine (Khemchandani and Chandra, 2007), it relaxes the parallel constrain of separating hyperplane in SVM thus boosting the training speed (Joachims, 1998). RA*k*EL*d* (RAndom *k* labELsets) is proposed by Tsoumakas et al. (2011) to overcome the overfitting problem of LP method.

RA*k*EL*d* divides the label space into *k* disjoint subsets and trains an ensemble of LP classifiers on each subset. Experiments shows that RA*k*EL*d* improves the performance over LP by a considerable margin and is among the best-performing methods especially for application domains with large number of labels (Tsoumakas et al., 2011).

Model Evaluation Method

The widely applied model validation methods are *k*-fold cross-validation, leave-one-out cross-validation (or called as jackknife test), and independent tests (or called as holdout method) (Chou and Zhang, 1995; Kohavi, 1995; Niu and Zhang, 2017; Han et al., 2018; Zhang et al., 2018; Aparo et al., 2019). Jackknife test uses a single instance from the sample set as the validation data, and the remaining samples as the training data. This process is iterated until each sample in the sample set is used as the validation case.

As for *k*-fold cross-validation (CV), the sample set is segmented randomly into *k* exclusive subsets with equal size. One subset of the *k* subsets is selected as the validation data, and the remnant *k*-1 subsets are as training data. This process is then repeated *k* time, until each of the *k* subsets used as the validation data for one time. A single estimation metric is finally generated by averaging the results from *k* folds. Typically for the classification task, the CV is often performed in stratified manner, which partitions a dataset so that the proportion of samples of each class in each fold equals to that in the whole dataset. Stratified CV is proven to improve CV in terms of bias and variance (Kohavi, 1995). But the Stratified CV for multi-label learning task is male-defined. Experiments on multi-label learning task either utilize presplit training/test set accompanying a benchmark dataset or the unstratified version of cross-validation and holdout method (Madjarov et al., 2012; Zhang and Zhou, 2014). This situation will possible lead to a scenario where the test set is absent of even single positive example of rare labels, causing the zero-divisor problem of various multi-label evaluation metrics. Commonly, researchers avert this problem via the removal of all the rare labels (Heider et al., 2013; Riemenschneider et al., 2016; Xing et al., 2019), which is suboptimal because the rare events are often of greater importance compared to common ones (Taleb, 2007). Two possible interpretations of multi-label stratification exist. One treats the distinct labelsets as unique classes, while another considers each label independently of the rest. The number of distinct labelsets often grow exponentially with the number of labels, which means the first interpretation is not applicable of the task at hand. The next interpretation was thus utilized in this article. Inspired by the study of Sechidis et al. (2011), we utilized 10 times repeated 10-fold iterative stratification cross-validation to validate our best performing multi-label method in a label-wise manner. The basic idea of this method is to iteratively sample each label, respectively in a greed manner. In the whole process, the rare labels are treated in priority to avoid zero-divisor problem and grasp instances with greater importance. The pseudocode of iterative stratification is given by **Algorithm 1**.

Algorithm 1: Iterative Stratification ($\mathbf{D}, n, r_1, \ldots, r_k$)

```
Input: A dataset, $\mathbf{D}$, consists of a set of labels $\mathbf{L} = \{l_1, .., l_q\}$,
designated number of folds $k$, required proportion of samples in each
fold, $r_1, \ldots, r_k$ (e.g. in 5-fold CV, $k = 5, r_j = 0.2, j = 1 \ldots 5$)
Output: Exclusive subsets $S_1, \ldots, S_k$ of $\mathbf{D}$
// Generate the required number of samples at each fold
for $j \leftarrow 1$ to $k$ do
    $c_j \leftarrow |\mathbf{D}| r_j$
// Generate the required number of samples of each label at each fold
for $i \leftarrow 1$ to $|L|$ do
    // Calculate the samples of each label in the initial set
    $D^i \leftarrow \{(D, L) \in \mathbf{D} : l_i \in L\}$
    for $j \leftarrow 1$ to $k$ do
      $c_j \leftarrow |D^i| r_j$
while $|D| > 0$ do
    // Identify the label with the fewest (but at least one) remaining
     samples,
    // Break ties randomly
    $D^i \leftarrow \{(D, L) \in \mathbf{D} : l_i \in L\}$
    $l \leftarrow argmin_i \left(|D^i|\right) \bigcap \{i : D^i \neq \varnothing\}$
    foreach $(D, L) \in D^l$ do
      // Identify the fold(s) with the largest number of required samples
      for this label
      // Break ties by considering the largest number of required
      samples, break further ties randomly
      $M \leftarrow \mathrm{argmax}_{j=1 \ldots k}(c_j^i)$
      if $|M| = 1$ then
       $m \in M$
      else
       $M' \leftarrow argmax_{j \in M}(c_j)$
       if $|M'| = 1$ then
         $m \in M'$
       else
         $m \leftarrow randomElementOf(M')$
      $S_m \leftarrow S_m \bigcup \{(D, L)\}$
      $D \leftarrow D\{(D, L)\}$
      // Update desired number of examples
      foreach $l_i \in \mathbf{L}$ do
       $c_m^i \leftarrow c_m^i - 1$
       $c_m \leftarrow c_m - 1$
return $S_1, \ldots, S_k$
```

Performance Metrics for Multi-Label Learning

Multi-label classification algorithms have widely been used in various bioinformatic applications (Zou et al., 2013; Yuan et al., 2016; Wan et al., 2017; You et al., 2018, 2019). Inspired by a set of five metrics established by Chou (2013) and the recommendation of Madjarov et al. (2012), we used the following five metrics to evaluate our multi-label learning model:

$$
\begin{cases}
Aiming = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\|\mathbb{L}_k \bigcap \mathbb{L}_k^*\|}{\|\mathbb{L}_k^*\|}\right) \\
Coverage = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\|\mathbb{L}_k \bigcap \mathbb{L}_k^*\|}{\|\mathbb{L}_k\|}\right) \\
Accuracy = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\|\mathbb{L}_k \bigcap \mathbb{L}_k^*\|}{\|\mathbb{L}_k \bigcup \mathbb{L}_k^*\|}\right) \\
Absolute\ True = \frac{1}{N}\sum_{k=1}^{N}\Delta\left(\mathbb{L}_k, \mathbb{L}_k^*\right) \\
Hamming\ loss = \frac{1}{N}\sum_{k=1}^{N}\| \mathbb{L}_k \ominus \mathbb{L}_k^* \|
\end{cases}
\tag{13}
$$

where N denotes the total number of samples, M stands for 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, $\ominus$ is the symmetric difference between two sets, and

$$\Delta\left(\mathbb{L}_k,\ \mathbb{L}_k*\right) = \begin{cases} 1, \textit{if all the labels in } \mathbb{L}_k \textit{ equal } \mathbb{L}_k* \\ 0, \textit{ otherwise} \end{cases} \tag{14}$$

These above metrics have been widely used in bioinformatic applications (Cheng et al., 2017).

Performance Metrics for Single-Label Learning

Apart from the metrics in the multi-label framework, we also utilized the following metrics to asses our methods in a label-wise manner (He et al., 2018; Manavalan et al., 2018a,c, 2019; Qiao et al., 2018; Xiong et al., 2018, 2019; Xu et al., 2018; Zhang et al., 2018a,b,c, 2019a,b,c; Bian et al., 2019; Su et al., 2019; Wei et al., 2019; Zeng et al., 2019; Zhu et al., 2019; Zou et al., 2019).

$$\begin{cases} Accuracy = \dfrac{TP + TN}{TP + TN + FN + FP} \\ Specificity = \dfrac{TN}{TN + FP} \\ Sensitivity = \dfrac{TP}{TP + FN} \\ F1 = \dfrac{2}{\dfrac{1}{\dfrac{TP}{TP + FP}} + \dfrac{1}{Sensitivity}} \\ CCR = \dfrac{Sensitivity + Specificity}{2} \end{cases} \tag{15}$$

where TP, TN, FN, TN are true positives, true negatives, false positives and false negatives for the prediction of each label respectively. In addition, the area under the receive operating characteristic curve (AUROC) were also calculated by trapezoidal rule.

CONCLUSION

Accurate prediction of the specificity of substrates for a panel of membrane transporters is of pivotal importance both in the ADMET profiling of drugs and the therapeutics of various cancers. The active drug efflux mediated via transporters lies in the junction of pharmacokinetics and pharmacodynamics. Novel chemicals are impossible to take any effect on cancers if they can be transported out of malignant cells even with satisfactory pharmacokinetic properties and potent *in vivo* anti-cancer activity. In addition, cancer stem cells are characterized by the expression of various transporters, which provides a vicious mechanism enabling cancer recurrence even many years after initial therapy. Identifying compounds without affinity to membrane transporters are prerequisite to the eradication of latent cancer stem cells. The aim of this study is to develop multi-label classification models to predict the classes of transporters given a substrate compound. This method utilized a hybrid of similarity-based features based on structural fingerprints and chemical ontologies. It was shown that the integration of 2D fingerprint and semantic similarity was a feasible and effective way to identify the specificity of a transporter substrate molecule. Various multi-label classification models such as ML-*k*NN, MTSVM, RA*k*EL*d* and NLSP were tested and compared on the benchmark dataset. NLSP-RF was finally selected for constructing the prediction model. To our best knowledge, this article is the first study to apply the multi-label system into the task of predicting of the specificity of membrane transporter substrates.

However, due to the imbalanced nature of classes on the benchmark dataset, our multi-label prediction system preforms unsatisfactory on the proteins of MRP2, MRP3, MRP4, SO1A2, and SOB1B1 in view of F1 score. In the next step, we will make efforts to address the imbalanced datasets via high throughput screens to boost the prediction performance on the specificity of membrane transporter substrates and deploy the optimized final model on a dedicate webserver for clinical and pharmacological usage. Our ultimate objective is to develop pan-transporter inhibitors for anti-cancer therapeutics.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://pubs.acs.org/doi/full/10.1021/acs.jcim.6b00508.

AUTHOR CONTRIBUTIONS

YX, D-QW, and XW contributed conception and design of the study. XW and YW organized the database. XW, YW, XZ, and MY performed the statistical analysis. XW wrote the first draft of the manuscript. XW, YW, and C-DL 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/fbioe.2019.00306/full#supplementary-material

REFERENCES

Aparo, A., Bonnici, V., Micale, G., Ferro, A., Shasha, D., Pulvirenti, A., et al. (2019). fast subgraph matching strategies based on pattern-only heuristics. *Interdiscip. Sci.* 11, 21–32. doi: 10.1007/s12539-019-00323-0

Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* 42, 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, 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. Theor. Exp.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/p10008

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

Chakraborty, C., George Priya Doss, C., Zhu, H., and Agoramoorthy, G. (2017). Rising strengths Hong Kong SAR in bioinformatics. *Interdiscip. Sci.* 9, 224–236. doi: 10.1007/s12539-016-0147-x

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Chen, G., Liu, J., Chen, W., Xu, Q., Xiao, M., Hu, L., et al. (2016). A 20(S)-protopanoxadiol derivative overcomes multi-drug resistance by antagonizing ATP-binding cassette subfamily B member 1 transporter function. *Oncotarget* 7, 9388–9403. doi: 10.18632/oncotarget.7011

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

Chen, L., Lu, J., Zhang, N., Huang, T., and Cai, Y. D. (2014). A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes. *Mol. Biosyst.* 10, 868–877. doi: 10.1039/c3mb70490d

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, CA: ACM).

Chen, W.-J., Shao, Y.-H., Li, C.-N., and Deng, N.-Y. (2016). MLTSVM: a novel twin support vector machine to multi-label learning. *Pattern Recognit.* 52, 61–74. doi: 10.1016/j.patcog.2015.10.008

Chen, Y.-J., Huang, W.-C., Wei, Y.-L., Hsu, S.-C., Yuan, P., Lin, H. Y., et al. (2011). Elevated BCRP/ABCG2 expression confers acquired resistance to gefitinib in wild-type EGFR-expressing cells. *PLoS ONE* 6:e21428. doi: 10.1371/journal.pone.0021428

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

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

Chou, K. C., and Zhang, C. T. (1995). Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* 30, 275–349. doi: 10.3109/10409239509083488

Couto, F. M., and Pinto, H. S. (2013). The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *J. Bioinf. Comput. Biol.* 11:17. doi: 10.1142/S0219720013710017

Couto, F. M., and Silva, M. J. (2011). Disjunctive shared information between ontology concepts: application to Gene Ontology. *J. Biomed. Semantics* 2:5. doi: 10.1186/2041-1480-2-5

Cripe, L. D., Uno, H., Paietta, E. M., Litzow, M. R., Ketterling, R. P., Bennett, J. M., et al. (2010). Zosuquidar, a novel modulator of P-glycoprotein, does not improve the outcome of older patients with newly diagnosed acute myeloid leukemia: a randomized, placebo-controlled trial of the Eastern Cooperative Oncology Group 3999. *Blood* 116, 4077–4085. doi: 10.1182/blood-2010-04-277269

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 36, D344–D350. doi: 10.1093/nar/gkm791

Duan, J., Dixon, S. L., Lowrie, J. F., and Sherman, W. (2010). Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* 29, 157–170. doi: 10.1016/j.jmgm.2010.05.008

Ferreira, J. D., and Couto, F. M. (2010). Semantic similarity for automatic classification of chemical compounds. *PLoS Comput. Biol.* 6:1000937. doi: 10.1371/journal.pcbi.1000937

Fletcher, J. I., Haber, M., Henderson, M. J., and Norris, M. D. (2010). ABC transporters in cancer: more than just drug efflux pumps. *Nat. Rev. Cancer* 10, 147–156. doi: 10.1038/nrc2789

Fukunaga, K., and Hostetler, L. (1973). Optimization of k nearest neighbor density estimates. *IEEE Trans. Inf. Theor.* 19, 320–326. doi: 10.1109/TIT.1973.1055003

Gantner, M. E., Alberca, L. N., Mercader, A. G., Bruno-Blanch, L. E., and Talevi, A. (2017). Integrated application of enhanced replacement method and ensemble learning for the prediction of BCRP/ABCG2 substrates. *Curr. Bioinf.* 12, 239–248. doi: 10.2174/1574893611666151109193016

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 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, 411–444. doi: 10.1002/widm.1139

Han, S., Cai, H., Che, D., Zhang, Y., Huang, Y., and Xie, M. (2018). Metrical consistency NMF for predicting gene-phenotype associations. *Interdiscip. Sci.* 10, 189–194. doi: 10.1007/s12539-017-0224-9

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., et al. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 41, D456–D463. doi: 10.1093/nar/gks1146

Hazai, E., Hazai, I., Ragueneau-Majlessi, I., Chung, S. P., Bikadi, Z., and Mao, Q. (2013). Predicting substrates of the human breast cancer resistance protein using a support vector machine method. *BMC Bioinf.* 14:130. doi: 10.1186/1471-2105-14-130

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

Heider, D., Senge, R., Cheng, W., and Hüllermeier, E. (2013). Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics* 29, 1946–1952. doi: 10.1093/bioinformatics/btt331

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* 13, 714–726. doi: 10.1038/nrc3599

Huang, J., Ma, G., Muhammad, I., and Cheng, Y. (2007). Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. *J. Chem. Inf. Model.* 47, 1638–1647. doi: 10.1021/ci700083n

International Transporter, C., Giacomini, K. M., Huang, S. M., Tweedie, D. J., Benet, L. Z., Brouwer, K. L., et al. (2010). Membrane transporters in drug development. *Nat. Rev. Drug Discov.* 9, 215–236. doi: 10.1038/nrd3028

Joachims, T. (1998). "Text categorization with support vector machines: learning with many relevant features," in *European Conference on Machine Learning* (Dortmund), 137–142.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach), 3146–3154.

Keum, J., Yoo, S., Lee, D., and Nam, H. (2016). Prediction of compound-target interactions of natural products using large-scale drug and protein information. *BMC Bioinf.* 17 (Suppl. 6):219. doi: 10.1186/s12859-016-1081-y

Khemchandani, R., and Chandra, S. (2007). Twin support vector machines for pattern classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 905–910. doi: 10.1109/Tpami.2007.1068

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Montreal), 1137–1145.

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

Li, D., Chen, L., Li, Y., Tian, S., Sun, H., and Hou, T. (2014). ADMET evaluation in drug discovery. 13. Development of *in silico* prediction models for P-glycoprotein substrates. *Mol. Pharm.* 11, 716–726. doi: 10.1021/mp400450m

Liao, Z. J., Wan, S. X., He, Y., and Zou, Q. (2018). Classification of small GTPases with hybrid protein features and advanced machine learning techniques. *Current Bioinformatics* 13, 492–500. doi: 10.2174/1574893612666171121162552

Lingineni, K., Belekar, V., Tangadpalliwar, S. R., and Garg, P. (2017). The role of multidrug resistance protein (MRP-1) as an active efflux transporter on blood-brain barrier (BBB) permeability. *Mol. Divers.* 21, 355–365. doi: 10.1007/s11030-016-9715-6

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

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

Maemondo, M., Inoue, A., Kobayashi, K., Sugawara, S., Oizumi, S., Isobe, H., et al. (2010). Gefitinib or chemotherapy for non–small-cell lung cancer with mutated EGFR. *N. Engl. J. Med.* 362, 2380–2388. doi: 10.1056/NEJMoa0909530

Mak, L., Marcus, D., Howlett, A., Yarova, G., Duchateau, G., Klaffke, W., et al. (2015). Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. *J. Cheminform.* 7:31. doi: 10.1186/s13321-015-0083-5

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., Govindaraj, R. G., Shin, T. H., Kim, M. O., and Lee, G. (2018a). iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* 9:1695. doi: 10.3389/fimmu.2018.01695

Manavalan, B., and Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* 33, 2496–2503. doi: 10.1093/bioinformatics/btx222

Manavalan, B., Lee, J., and Lee, J. (2014). Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS ONE* 9:e106542. doi: 10.1371/journal.pone.0106542

Manavalan, B., Shin, T. H., Kim, M. O., and Lee, G. (2018b). AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.* 9:276. doi: 10.3389/fphar.2018.00276

Manavalan, B., Shin, T. H., Kim, M. O., and Lee, G. (2018c). PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.* 9:1783. doi: 10.3389/fimmu.2018.01783

Manavalan, B., Shin, T. H., and Lee, G. (2018d). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

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

Meng, C., Wei, L., and Zou, Q. (2019). SecProMTB: support vector machine-based classifier for secretory proteins using imbalanced data sets applied to *Mycobacterium tuberculosis*. *Proteomics* 19:e1900007. doi: 10.1002/pmic.201900007

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, 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

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

Nyquist, M. D., Prasad, B., and Mostaghel, E. A. (2017). Harnessing solute carrier transporters for precision oncology. *Molecules* 22:E539. doi: 10.3390/molecules22040539

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open babel: an open chemical toolbox. *J. Cheminform.* 3:33. doi: 10.1186/1758-2946-3-33

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* 5:e1000443. doi: 10.1371/journal.pcbi.1000443

Poongavanam, V., Haider, N., and Ecker, G. F. (2012). Fingerprint-based *in silico* models for the prediction of P-glycoprotein substrates and inhibitors. *Bioorg. Med. Chem.* 20, 5388–5395. doi: 10.1016/j.bmc.2012.03.045

Pusztai, L., Wagner, P., Ibrahim, N., Rivera, E., Theriault, R., Booser, D., et al. (2005). Phase II study of tariquidar, a selective P-glycoprotein inhibitor, in patients with chemotherapy-resistant, advanced breast carcinoma. *Cancer* 104, 682–691. doi: 10.1002/cncr.21227

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 Bioinf.* 19:14. 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* 76:036106. doi: 10.1103/PhysRevE.76.036106

Riemenschneider, M., Senge, R., Neumann, U., Hüllermeier, E., and Heider, D. (2016). Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. *BioData Min.* 9:10. doi: 10.1186/s13040-016-0089-1

Ru, X., Li, L., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome Res.* 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

Sechidis, K., Tsoumakas, G., and Vlahavas, I. (2011). "On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases*, eds D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis (Berlin Heidelberg: Springer), 145–158. doi: 10.1007/978-3-642-23808-6_10

Shaikh, N., Sharma, M., and Garg, P. (2017). Selective fusion of heterogeneous classifiers for predicting substrates of membrane transporters. *J. Chem. Inf. Model.* 57, 594–607. doi: 10.1021/acs.jcim.6b00508

Shan, X., Wang, X., Li, C.-D., Chu, Y., Zhang, Y., Xiong, Y. I., et al. (2019). Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method. *J. Chem. Inf. Model*. 2019:9b00749. doi: 10.1021/acs.jcim.9b00749

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods* 166, 91–102. doi: 10.1016/j.ymeth.2019.02.009

Sun, Y., Xiong, Y., Xu, Q., and Wei, D. (2014). A hadoop-based method to predict potential effective drug combination. *Biomed Res. Int.* 2014:196858. doi: 10.1155/2014/196858

Szakács, G., Paterson, J. K., Ludwig, J. A., Booth-Genthe, C., and Gottesman, M. M. (2006). Targeting multidrug resistance in cancer. *Nat. Rev. Drug Discov.* 5, 219–234. doi: 10.1038/nrd1984

Szymanski, 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* 18:282. doi: 10.3390/e18080282

Taleb, N. N. (2007). Black swans and the domains of statistics. *Am. Stat.* 61, 198–200. doi: 10.1198/000313007x219996

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2011). Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* 23, 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* 17:262. doi: 10.1002/pmic.201700262

Wang, X., Wang, Y., Xu, Z., Xiong, Y., and Wei, D. (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

Wang, Z., Chen, Y., Liang, H., Bender, A., Glen, R. C., and Yan, A. (2011). P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J. Chem. Inf. Model.* 51, 1447–1456. doi: 10.1021/ci2001583

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

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

Xing, L., Lesperance, M., and Zhang, X. (2019). Simultaneous prediction of multiple outcomes using revised stacking algorithms. *Bioinformatics*. 2019:btz531. doi: 10.1093/bioinformatics/btz531

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10 (Suppl. 1):S20. doi: 10.1186/1477-5956-10-S1-S20

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, 229–235. doi: 10.2174/1389200219666181019094526

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. (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

Xiong, Y., Xia, J., Zhang, W., and Liu, J. (2011). Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. *PLoS ONE* 6:e28440. doi: 10.1371/journal.pone.0028440

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.* 2018:478. 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, 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, 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, i18–i27. doi: 10.1093/bioinformatics/btw244

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Briefings Bioinf.* 2019:bbz080. doi: 10.1093/bib/bbz080

Zhang, M.-L., and Zhou, Z.-H. (2005). "A k-nearest neighbor based algorithm for multi-label classification," in *IEEE International Conference on Granular Computing* (Beijing), 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, 1819–1837. doi: 10.1109/tkde.2013.39

Zhang, M. L., and Zhou, Z. H. (2007). ML-KNN: a lazy learning approach to multi-label leaming. *Pattern Recognit.* 40, 2038–2048. doi: 10.1016/j.patcog.2006.12.019

Zhang, N., Sa, Y., Guo, Y., Lin, W., Wang, P., and Feng, Y. M. (2018). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinf.* 13, 50–56. doi: 10.2174/1574893611666160608102537

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug-drug interactions. *Inf. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2019:2931546. doi: 10.1109/TCBB.2019.2931546

Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., and Li, X. (2018a). Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing* 287, 154–162. doi: 10.1016/j.neucom.2018.01.085

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018b). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhang, W., Yu, C., Wang, X., and Liu, F. (2019c). Predicting CircRNA-disease associations through linear neighborhood label propagation method. *IEEE Access* 7, 83474–83483. doi: 10.1109/ACCESS.2019.2920942

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

Zhang, Y., Liu, J., Liu, X., Hong, Y., Fan, X., Huang, Y., et al. (2018). GC[Formula: see text]NMF: a novel matrix factorization framework for gene-phenotype association prediction. *Interdiscip. Sci.* 10, 572–582. doi: 10.1007/s12539-018-0296-1

Zhong, L., Ma, C. Y., Zhang, H., Yang, L. J., Wan, H. L., Xie, Q. Q., et al. (2011). A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG-SVM method. *Comput. Biol. Med.* 41, 1006–1013. doi: 10.1016/j.compbiomed.2011.08.009

Zhu, X., He, J., Zhao, S., Tao, W., Xiong, Y., and Bi, S. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *Saccharomyces cerevisiae*. *Briefings Funct. Genomics*. 2019:elz018. doi: 10.1093/bfgp/elz018

Zou, Q., Chen, W. C., Huang, Y., Liu, X. R., and Jiang, Y. (2013). Identifying multi-functional enzyme by hierarchical multi-label classifier. *J. Comput. Theor. Nanos.* 10, 1038–1043. doi: 10.1166/jctn.2013.2804

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Conflict of Interest: 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.