JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 4577−4586

pubs.acs.org/jcim

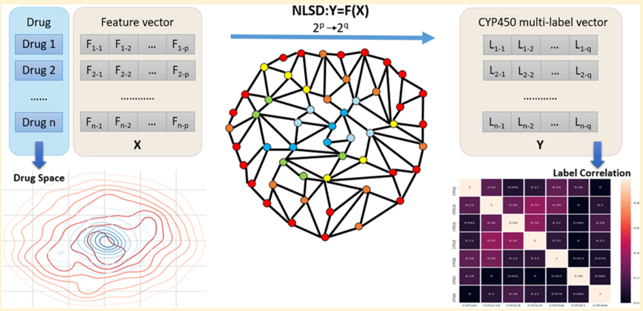
Prediction of CYP450 Enzyme−Substrate Selectivity Based on the Network-Based Label Space Division Method

Xiaoqi Shan,[†] Xiangeng Wang,[†] Cheng-dong Li,[†] Yanyi Chu,[†] Yufang Zhang,[†] Yi Xiong,[*,†] and Dong-Qing Wei[*,†,‡]

[†]State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, and Joint Laboratory of International Cooperation in Metabolic and Developmental Sciences, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China

[‡]Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nanshan District, Shenzhen, Guangdong 518055, China

Supporting Information



ABSTRACT: A drug may be metabolized by multiple cytochrome P450 (CYP450) isoforms. Predicting the metabolic fate of drugs is very important to prevent drug−drug interactions in the development of novel pharmaceuticals. Prediction of CYP450 enzyme−substrate selectivity is formulized as a multilabel learning task in this study. First, we compared the performance of feature combinations based on four different categories of features, which are physiochemical property descriptors, mol2vec descriptors, extended connectivity fingerprints, and molecular access system key fingerprints on modeling. After identifying the best combination of features, we applied seven different multilabel models, which are multilabel *k*-nearest neighbor (ML-*k*NN), multilabel twin support vector machine, and five network-based label space division (NLSD)-based methods (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM). All of the six models (ML-*k*NN, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM) in this paper exhibit better performances than the previous work. Besides, NLSD-XGB achieves the best performance with the average top-1 prediction success of 91.1%, the average top-2 prediction success of 96.2%, and the average top-3 prediction success of 98.2%. When compared with the previous work, NLSD-XGB shows a significant improvement over 11% on top-1 in the 10 times repeated 5-fold cross-validation test and over 14% on top-1 in the 10 times repeated hold-out method. To the best of our knowledge, the network-based label space division model is first introduced in drug metabolism and performs well in this task.

INTRODUCTION

Cytochrome P450 (CYP450) enzymes were first discovered in rat liver microsomes in 1958, named after the common properties of the absorption peak at 450 nm produced by the mixture of these enzymes and carbon monoxide.[1] CYP450 enzymes were widely found in bacteria, fungi, plants, and animals. In the human body, the CYP450 enzyme is responsible for the redox process of endogenous substrates and exogenous compounds such as fatty acids, steroids, toxins, and 90% of commonly used drugs, which play an important role in drug efficacy and toxicity.[2] So far, 57 CYP450 genes have been identified in the human genome.[3] All CYP450 genes can be mainly divided into CYP1, CYP2, and CYP3 subfamilies according to the similarity of amino acid sequences.[4] In recent years, accumulating studies[5−7] have confirmed that CYP450 gene polymorphism is one of the main reasons leading to drug metabolism heterogeneity between different individuals in clinical practice bringing great trouble to physicians.

Seven CYP450 isoforms (1A2, 2C8, 2C9, 2C19, 2D6, 2E1, and 3A4) are in charge of the majority of all possible metabolism routes.[8] The flexibility of CYP450 conformations makes it difficult to predict substrates of CYP450 using structure-based methods, such as molecular docking, molecular dynamics simulating, and pharmacophore mapping.[9] Alternative approaches, such as ligand-based methods, are utilized to predict enzyme−substrate selectivity based on structural similarity between ligands and known substrates. The most commonly used ligand-based approach is the quantitative structure−activity relationship (SAR or QSAR) model, which provides a quantitative estimate of the reactivity of each potential metabolic site. Moreover, a large number of machine learning methods, which can be considered as upgraded QSAR, have been widely used to predict CYP450 substrates.[8,10]

For the study of CYP450 and substrate selection specificity, one drug molecule may be metabolized by multiple CYP450 isoforms, which can be formulated as a multilabel classification problem. Multilabel classification issues in biology remain to

Received: September 6, 2019
Published: October 11, 2019

DOI: 10.1021/acs.jcim.9b00749

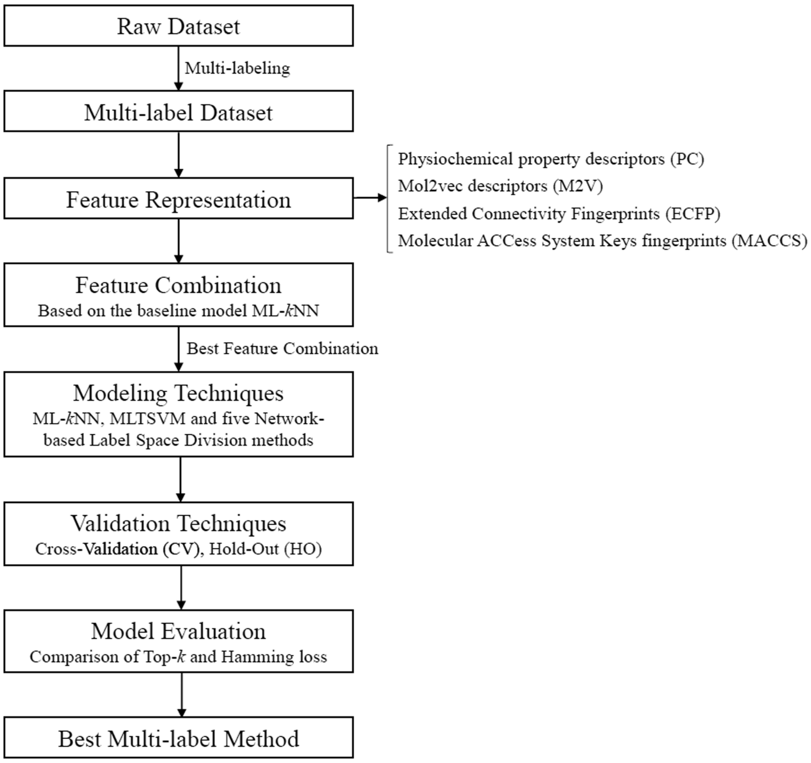

Raw Dataset
Multi-labeling
Multi-label Dataset
Feature Representation
Physiochemical property descriptors (PC)
Mol2vec descriptors (M2V)
Extended Connectivity Fingerprints (ECFP)
Molecular ACCess System Keys fingerprints (MACCS)
Feature Combination
Based on the baseline model ML-kNN
Best Feature Combination
Modeling Techniques
ML-kNN, MLTSVM and five Network-based Label Space Division methods
Validation Techniques
Cross-Validation (CV), Hold-Out (HO)
Model Evaluation
Comparison of Top-k and Hamming loss
Best Multi-label Method

Figure 1. Flowchart of the multilabel modeling process for the prediction of CYP450 isoform specificity.

be challenging in the field of machine learning, in terms of feature selection and model building.[11,12] Compared with the single-label classification strategy, the multilabel classification strategy can decipher the relationship among predictive labels to improve model performance. Michielan et al.[13] used ct-SVM, multilabel *k*-nearest neighbor (ML-*k*NN), and CPG-NN methods to classify 580 CYP450 substrates on two datasets with five and seven classes, respectively. The correct predictions of ct-SVM/5 class and CPG-NN/5 class models on the test set were 77.5−96.6 and 75.6−97.1%, respectively. These results indicated that the multilabel method can achieve a consistent reflection of the real metabolic information. Wei et al.[4] used ML-*k*NN, BP-MLL, and RankSVM methods to classify 77 CYP450 substrates into five classes. Their results showed that the correct prediction of the five enzyme−substrates reached accuracy greater than 80%. Hunt et al.[14] developed a seven-class random forest (7-class RF) model to predict the major metabolizing isoforms for a compound, and the model had a 76% success rate with top-1 and an 88% success rate with top-2. They provided a relatively complete dataset, which contains 484 compounds and 1299 pairs of compounds/CYP450 isoforms. However, they did not take the full advantage of multilabel classification techniques on this dataset. Therefore, our study uses the relatively canonical data from Tyzack et al.[10] to establish multilabel classification models with seven classes, expecting to get a better result.

Algorithm adaptation, problem transformation, and ensembles of multilabel classifier (EMLC) are three major types of multilabel classification models. Algorithm adaptation methods utilized various tricks to modify single-label learning algorithms into multilabel ones. The canonical method of this group is ML-*k*NN.[15] Problem transformation methods convert the multilabel learning problem into several single-label tasks. Label powerset (LP) is a traditional method of problem transformation that builds models on each possible subset of label sets.[16] For a dataset with many labels in the label set, LP tends to be overfitting because the number of subsets will exponentially grow when the cardinality of the set increases linearly. To ameliorate the overfit attribute of the label-powerset method, Tsoumakas et al.[17] divided the label space into subspaces and used the label-powerset method in these subspaces. The RA*k*EL*d* method is designed by this principle, which segments the label set into *k* nonoverlapping subsets. One main weakness 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 network-based label space division (NLSD)[18] is an EMLC built upon LP, and it divides the label sets into *n* small-sized label sets (possibly intersecting) by the community detection method, which can incorporate the correlation among labels in the training set. It finally learns *k* representative LP classifiers. As a result, NLSD tackles much less subsets compared to LP and selects *k* in a data-driven manner. The NLSD-based algorithm has been successfully used in our other study for the prediction of the anatomical therapeutic chemical classes of a given compound.[19] A more detailed explanation of multilabel learning can be found in refs 20, 21.

In this present study, we applied the network-based label space division method to exploit the label correlation structures in the dataset. Our experimental results indicated that the NSLD-based models have reached the top performance in the benchmark dataset in comparison with the state-of-the-art methods. The main advantage of our method relies on two aspects. On the one hand, the NLSDs divide the label space into subspaces by a network-based algorithm, which not only avoids overfitting by the traditional label-powerset method but also provides a way to utilize the correlation among labels. On the other hand, the ensemble learning nature

DOI: 10.1021/acs.jcim.9b00749

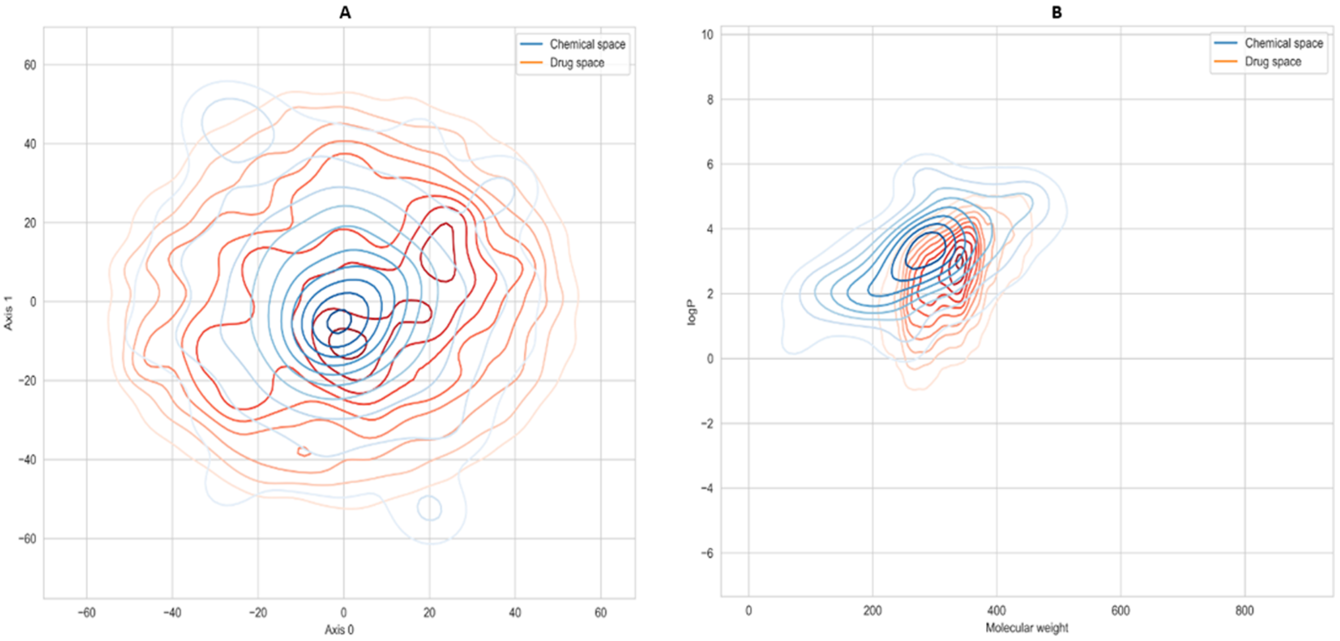


Figure 2. Contour plot of drug space over chemical space. The red lines represent the drug space of our dataset, and the blue lines represent the chemical space. (A) Drug space over chemical space in terms of dimensionally reduced structural attributes (ECFP 2048). (B) Drug space over chemical space in terms of physicochemical attributes (molecular weight and log *P*).

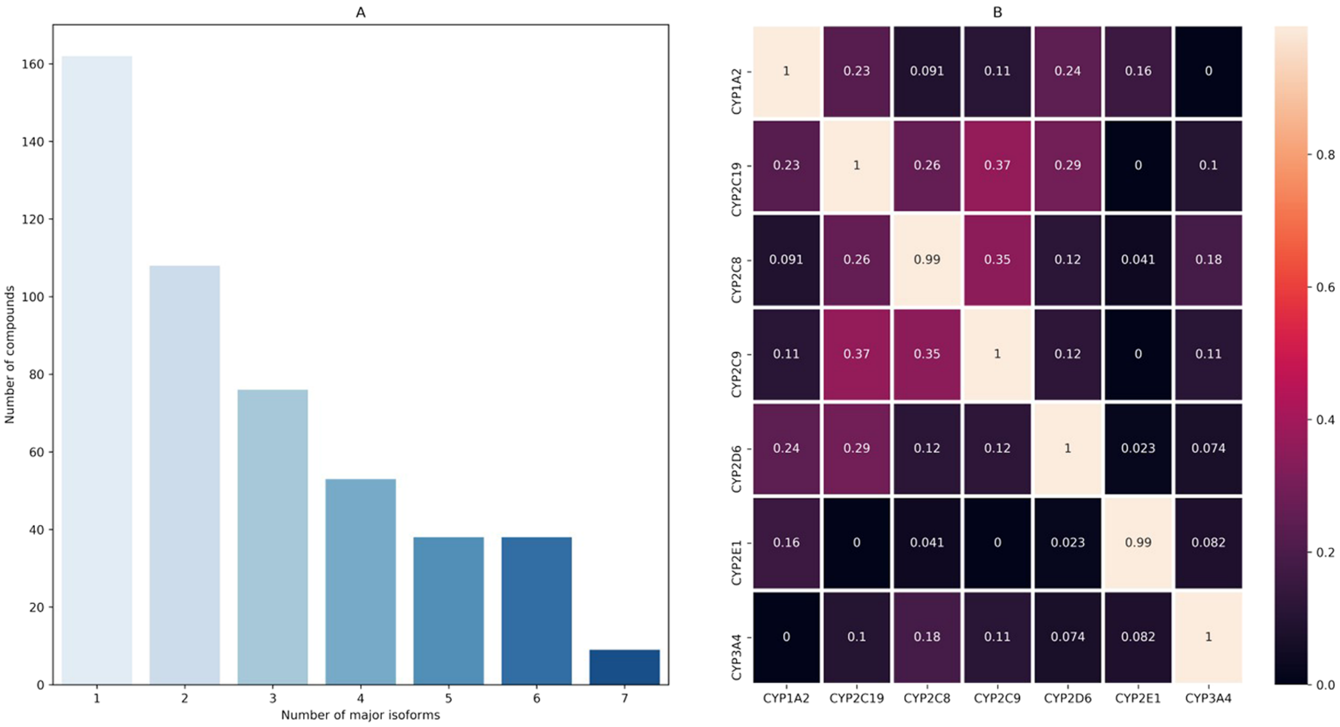


Figure 3. Distribution of the numbers of isoforms identified as major metabolizing enzymes for each compound, and the pairwise correlation of different label pairs. (A) Bar plot representing the number of compounds, which can be metabolized by different types of CPY450 isoforms. (B) Heatmap showing the bias-corrected Cramér's *V* statistics of different label pairs. The bias-corrected Cramér's *V* statistic lies in the interval of [0, 1], the higher the value, the stronger the correlation between the two labels.

of NLSDs on the overlapping subspace could further improve model performance.

■ METHODS

In our study, the raw dataset[10] was first integrated to build a multilabel dataset, and then four categories of features for 15 different feature combinations were evaluated for modeling to obtain the optimal one. We adopted 10 times repeated 5-fold cross-validation (CV) as the same model evaluation method consistent with the previous work.[14] In addition, we considered it necessary to separate an independent test set from the whole dataset, so we adopted the hold-out (HO) method (training set/test set = 85:15%). Next, ML-*k*NN, multilabel twin support vector machine (MLTSVM), and five

NLSDs (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM) were chosen to build multilabel classification models. Finally, top-k and hamming loss of each model were compared to select the best model. The workflow of our study is illustrated in Figure 1.

Dataset. In this study, the raw dataset we used contains 484 compounds and 1299 compound/isoform pairs, in which 484 compounds can be metabolized by at least one CYP450 enzyme from seven CYP450 enzymes, including 1A2, 2C8 2C9, 2C19, 2D6 2E1, and 3A4. We developed multilabel classifiers on this dataset.

The dataset used herein was generated during the data collection from the published models of P450 regioselectivity by Tyzack et al.[10,14] Five thousand small molecular compounds randomly selected from the ZINC database[22] were considered as chemical space, and the dataset herein was regarded as drug space. Figure 2 shows the contour plot of drug space over compound space after dimensional reduction of extended connectivity fingerprints (ECFP) as well as over molecular weight and log P. It shows that the dataset used in this work is evenly distributed in the small molecule compound space both in terms of structural and physicochemical attributes. The distribution of compounds is generated by the t-distribution random adjacency embedding (t-distributed stochastic neighbor embedding) algorithm,[23] which reduced the 2048-dimensional ECFP fingerprints into a two-dimensional (2D) subspace for density estimation and visualization. Distribution of the numbers of isoforms identified as major metabolizing enzymes for each compound is shown in Figure 3A. One basic assumption of multilabel learning is that we can exploit the label correlation to improve model performance. We calculated the bias-corrected Cramér's V statistics[24] on the label sets, and the details of the pairwise correlation of labels are depicted in Figure 3B. We can notice the correlations among different labels, which suggests that the predictive model will benefit from multilabel learning models.

In the process of labeling, if a compound can be metabolized by an isoform, we label it as 1; otherwise, we label it as 0. If a compound can be metabolized by CYP1A2, 2C9, 2C19, and 3A4 but cannot be metabolized by CYP2C8, 2D6, and 2E1, then the true classification of the compound will be represented as 1011001 (the encoding order is CYP1A2, 2C8, 2C9, 2C19, 2D6, 2E1, and 3A4).[4] Since this multilabel dataset contains the metabolic information of 484 compounds/7 enzymes, the label of multilabel classification model on this dataset is a 484 × 7-dimensional matrix.

Feature Representation. Four types of features, including physiochemical property descriptors (PC), mol2vec descriptors (M2V), extended connectivity fingerprints (ECFP), and molecular access system (MACCS) key fingerprints of all compounds, were calculated and considered as features for modeling. Two hundred features of physiochemical properties were generated by RDKit in Python (http://www.rdkit.org/). Three hundred and thirteen features of physiochemical properties were generated by molecular operating environment software. Besides, duplicated features were removed.

Mol2vec is a natural language processing-inspired technique, which considers compound substructures derived from Morgan's algorithm as "words" and compound structures as "sentences". The Word2vec algorithm is applied to the compound corpus to obtain the high dimensional embedding of substructures, in which the substructure vectors related to chemistry occupy the same part of the vector space.[25] Mol2vec is an unsupervised method, which is first trained on the unlabeled dataset to obtain the feature vectors of substructures. Then, the feature vectors are summed up to obtain the composite vectors.[26] The composite vectors were considered as mol2vec descriptors, and 300 mol2vec descriptors were used for modeling in our study.

The MACCS key is a fast method for substructure screening in the molecular database. The MACCS key fingerprint is often used to calculate chemical similarity.[27] The public version of the MACCS key contains 166 bits, where each bit corresponds to the existence of a specific molecular signature, such as the carbonyl group (No. 154 bond).

ECFP 4 and ECFP 6 are among the best-performing fingerprints both in the virtual screen of separating actives from decoys and in ranking diverse structures by similarity.[28] It is worth using more than 1024-bit fingerprints due to the improved performance and reduced hash collision. Therefore, 2048-bit ECFP 4 fingerprints were selected for modeling in our study.

In our study, ECFP had 2048 bits while MACCS key fingerprints had 166 bits. Descriptors with zero variance were removed.[9] In total, 478 physiochemical property descriptors, 300 mol2vec descriptors, 2048 ECFP fingerprints, and 166 MACCS key fingerprints were used as input features for model training. The values of each descriptor were normalized to range between 0 and 1 by subtracting the minimum value of the descriptor and dividing by the range.[9] Besides, we have combined four categories of features with 15 kinds of feature combinations in total and built multilabel models based on different feature combinations to select the best feature combination for constructing the final model.

Modeling Techniques. In our work, several multilabel classification algorithms have been employed to construct our classification models. First, we used the same method as WhichP450[14] to perform 5-fold cross-validation on the whole dataset (training set and test set undivided); Second, a completely blind external validation set (test set) is built to give a final evaluation of the model. Therefore, we divided the training set and test set on the complete dataset by 85:15% by considering the limited number of samples on the dataset. To reduce the bias of the splits on the dataset, we have randomly shuffled the dataset, and the modeling processes were replicated 10 times in both methods. Multilabel k-nearest neighbor (ML-kNN),[29] multilabel twin support vector machine (MLTSVM),[30] and five network-based label space division methods (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM) were used in this study.[18]

Multilabel k-Nearest Neighbor (ML-kNN). ML-kNN was used as the baseline method in this study.[29] It is a lazy learning method on the basis of traditional kNN.[31] For a given new sample, it first finds the top-k closest samples in the training set. Second, it calculates the number of each label in the k samples. Third, on the basis of the aforementioned label number, it estimates the label probability by the naïve Bayes method. Finally, the label probability is generated by maximum a posteriori estimation. This method is currently widely used in the multilabel prediction task and can achieve satisfactory performance,[4,32,33] so we use it as the baseline method.

Multilabel Twin Support Vector Machine (MLTSVM). MLTSVM is a variation of twin support vector machine designed for a multilabel scenario proposed recently.[30] For the twin support vector machine,[34] it relaxes the parallel constraint

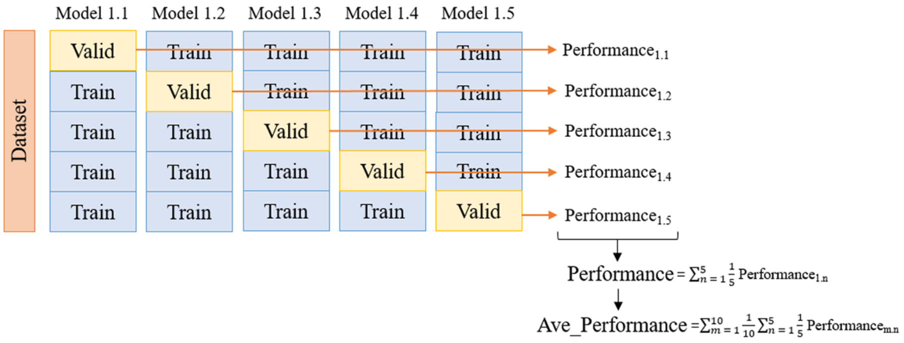

Model 1.1
Model 1.2
Model 1.3
Model 1.4
Model 1.5
Dataset
Valid
Train
Performance$_{1.1}$
Performance$_{1.2}$
Performance$_{1.3}$
Performance$_{1.4}$
Performance$_{1.5}$
Performance = $\sum_{n=1}^{5} \frac{1}{5}$ Performance$_{1.n}$
Ave_Performance = $\sum_{m=1}^{10} \frac{1}{10} \sum_{n=1}^{5} \frac{1}{5}$ Performance$_{m.n}$

Figure 4. Process of cross-validation (CV) methodology.

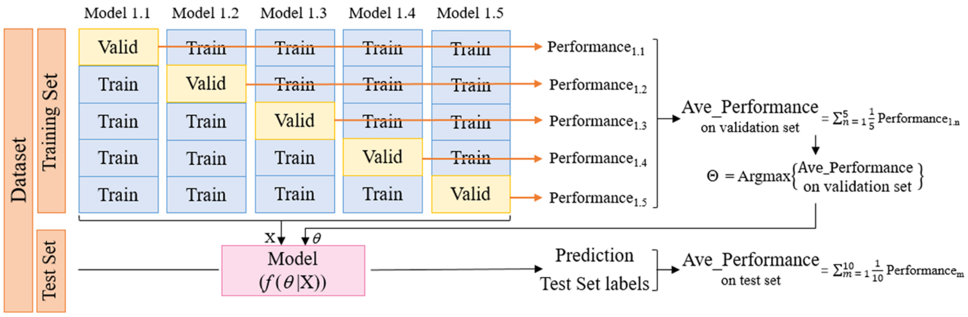


Figure 5. Process of hold-out (HO) methodology.

of separating hyperplane in SVM thus boosting the training speed.[35]

Network-Based Label Space Division (NLSD-X). The Network-based label space division (NLSD-X) was used for our multilabel prediction,[18] where X stands for a classification algorithm as the base classifier. Unlike the well-established label space partitioning method for supervised multilabel learning-Ra*k*EL,[17] which divides the label space into predefined subsets, NLSD-X takes the advantage of mature community detection approaches from the social network research field, and partitions the label space in a data-driven manner. This method can perform well in various multilabel classification benchmark datasets.[18] This method divided predictive modeling into training and classification phases.

In the training phase, four steps are preformed:

1. Establishing a label co-occurrence graph on the training set.
2. Detecting community on the label co-occurrence graph.
3. For each community L_i, and corresponding training set D_i is created by taking the original dataset with label columns present in L_i.
4. For each community, a base predictor h_i is learnt on the training set D_i.

In the classification phase, we just perform prediction on all communities detected in the training phase and fetch the union of assigned labels $h(\overline{x}) = \cup_{i=0}^{k} h_i(\overline{x})$.

We used five algorithms that include multilayer perceptron (MLP),[36] extreme gradient boosting,[37] extra tree,[38] random forest,[39] and support vector machines[35] to obtain the models named NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM.

Validation Techniques. Two methodologies were used in our study. The first was the same as that of the WhichP450[14] method, which performed 5-fold cross-validation on the whole dataset (as shown in Figure 4). In the WhichP450 method,[14] the whole dataset was divided into five parts, four parts of which were considered as a training set to be used in model training, and the remaining part was considered as a validation set used for model validation. The result of the model evaluation was the average of the 5-fold cross-validation results with 10 times repeated by the model verified on the validation set. We explained the importance of dividing a separate test set above. The second method was hold-out methodology, which was to conduct 5-fold cross-validation on the training set and conduct the model evaluation on the completely blind test set (as shown in Figure 5). In this methodology, a validation set was selected to optimize the parameters in each of the cross-validation within the same model.[13] The best parameters were adjusted using 5-fold cross-validation. Finally, a multilabel classification model with optimal parameters was trained on the whole training set. The result of the model evaluation was the average of the 5-fold cross-validation results with 10 times repeated by the model verified on the independent test set.

Model Evaluation Metrics. We adopted the hamming loss and top-*k* for multilabel classification model evaluation. The hamming loss represents the proportion of estimated false prediction labels. If one of the five labels is predicted incorrectly, the hamming loss is 0.2. Besides, the top-*k* performance evaluation metrics used in WhichP450[14] were also utilized. The top-*k* accuracy of each model was assessed, whereby a successful prediction was deemed to be one where at least one isoform predicted in the top-*k* ranked isoforms matched any of the observed isoforms.[14] In other word, top-*k* means that at least one class was predicted correctly in the top-*k* classes. We needed to compare the top-*k* values of this study with those of the previous study, so here we added the meaning of top-*k* in detail. As shown in Table S1, in this case, top-1 means that the class of Class3A4 was predicted correctly. Top-2 means that at least one class was predicted correctly in Class3A4 and Class2D6. Top-3 means that at least one class was predicted correctly in Class3A4, Class2D6, and Class2C19. Top-1, top-2, and top-3 were used to evaluate models in the previous study of WhichP450.[14]

RESULTS AND DISCUSSION

We built multilabel models with 15 different feature combinations to select the best feature combination based on the baseline model ML-*k*NN. After identifying the best combination of features, we applied seven different multilabel models using these features, which are ML-*k*NN, MLTSVM, and five network-based label space division methods (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM). In model comparison, top-*k* and hamming loss on these seven models were compared in both CV and HO methods, which helped us to choose the best multilabel model.

Predictive Power of Different Kinds of Feature Combinations. We used 478 physiochemical descriptors, 300 mol2vec descriptors, 2048 ECFP fingerprints, and 166 MACCS key fingerprints for model training. Besides, we have combined the four categories of features and built multilabel baseline models based on ML-*k*NN for different feature combinations to select the best feature combination. Table 1 shows hamming loss and top-*k* of models based on ML-*k*NN with 15 different feature combinations in the CV method, which were also used to compare with those of the previous work. Each column represents the average of top-*k* and hamming loss on the model obtained by 10 times repeated 5-fold cross-validation.

Table 1. Hamming Loss and Top-*k* of Multilabel Baseline Models Based on ML-*k*NN with 15 Different Feature Combinations

		top-*k*		
feature combinations (ML-*k*NN)	hamming loss	top-1%	top-2%	top-3%
PC + M2V + ECFP + MACCS	0.2586	79.0	89.9	95.1
PC + M2V + ECFP	0.2592	80.2	90.9	95.9
PC + M2V + MACCS	0.2567	77.9	88.3	93.4
PC + ECFP + MACCS	0.2590	79.2	90.0	95.1
M2V + ECFP + MACCS	0.2670	77.3	89.9	95.2
PC + M2V	0.2619	75.8	87.5	92.8
PC + ECFP	0.2634	78.9	90.4	95.3
PC + MACCS	0.2568	76.2	87.7	93.0
M2V + ECFP	0.2804	74.0	87.6	92.9
M2V + MACCS	0.2673	76.2	88.1	93.2
ECFP + MACCS	0.2671	76.5	89.3	94.4
PC	0.2666	75.0	86.9	92.3
M2V	0.2620	77.5	88.4	93.7
ECFP	0.2961	66.3	82.4	90.2
MACCS	0.2678	76.3	88.5	93.3
WhichP450[14]		76.3	88.4	93.3

There is a total of four categories of features used for modeling. Hamming loss has little difference, and we choose the best combination of features by top-*k* values, which are the same evaluation indexes as the previous work.[14] Table 1 shows that the ML-*k*NN model based on mol2vec descriptors performs best among the models built with only one category of features, where top-*k* values are already higher than those of the previous work. Mol2vec provides the state-of-the-art performance for classification and regression of various datasets, especially on less training dataset probably, for which our multilabel dataset with 484 compounds may be suitable.[26] When physiochemical descriptors and MACCS key fingerprints are respectively used for the features of the ML-*k*NN model, the top-*k* values of which are almost the same as the previous work. The description above illustrates that the features we select are effective. Besides, it is necessary to combine features to further boost the prediction performance. First, the model performance of ECFP is not good when it is used for modeling alone, but the model performance is significantly improved after combining ECFP features with other categories of features. Second, it can be further proved by Table 1 that using a combination of all categories of features for modeling does not give the best performance. In contrast, among these 15 different feature combinations, the combination of physiochemical property descriptors, mol2vec descriptors, and ECFP fingerprints used for modeling achieves the highest top-*k* values, of which the average top-1 prediction success is 80.2%, the average top-2 prediction success is 90.9%, and the average top-3 prediction success is 95.9%. Therefore, we chose physiochemical descriptors, mol2vec descriptors, and ECFP fingerprints as the final feature combination to build seven multilabel models.

The reason behind the best-performing feature combination can be interpreted as follows. ECFPs are canonical features representing 2D structures and are of pivotal importance in QASR modeling. However, the isoforms of CYP450 share similar substrate structures, thus the introduction of physiochemical descriptors further improves the modeling performance. In addition, mol2vec descriptors can be interpreted as semantic features for chemicals, so the incorporation of these descriptors provides orthogonal clues of isoform selectivity.

Hyperparameter Tuning. We tuned the following three types of hyperparameters for NLSD.

1. The community detection method: we try to compare the two cluster methods of the largest modularity and label propagation.[40]
2. The base learner: five types of base learners were chosen
    (i) Multilayer perceptron (MLP): the hyperparameter of hidden layer sizes was tuned at [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 500, 1000], and the other hyperparameters were set at the default values.
    (ii) Extreme gradient boosting (XGB): the hyperparameter of a number of trees was tuned at [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], and the other hyperparameters were set at the default values.
    (iii) Extremely randomized trees (EXT): the hyperparameter of a number of trees was tuned at [10, 20, 50, 100, 200, 300, 500], and the other hyperparameters were set at the default values.
    (iv) Random forests (RF): the hyperparameter of a number of trees was tuned at [10, 20, 50, 100, 200, 300, 500], and the other hyperparameters were set at the default values.
    (v) Support vector machine (SVM): the hyperparameter of C (penalty) was tuned at [0.01, 0.1, 0.5, 0.8, 1, 10, 20, 100], the radial basis function was chosen, and the other hyperparameters were set at the default values.
3. The problem-transforming method: we try to compare the two problem-transforming methods of label powerset[20] and classifier chain.[41]

Table 2. Hamming Loss and Top-*k* of Multilabel Models Based on ML-*k*NN with 10 Rounds of 5-Fold Cross-Validation in the CV Method

		top-*k*		
ML-*k*NN (CV)	hamming loss	top-1%	top-2%	top-3%
valid_1	0.2580	79.7	91.5	96.7
valid_2	0.2716	76.7	89.5	95.5
valid_3	0.2650	78.3	89.3	93.8
valid_4	0.2659	77.9	90.1	94.6
valid_5	0.2556	78.9	87.8	94.8
valid_6	0.2583	80.6	93.0	96.5
valid_7	0.2657	80.4	89.0	95.8
valid_8	0.2606	76.9	90.1	95.5
valid_9	0.2606	78.1	90.3	94.6
valid_10	0.2677	78.1	88.8	95.0
average	0.2629 ± 0.0062	78.5 ± 1.3	89.9 ± 1.4	95.3 ± 0.8
WhichP450[14]		76.3 ± 5.4	88.4 ± 4.9	93.3 ± 3.4

Table 3. Average of hamming Loss and Top-*k* of Multilabel Models Based on ML-*k*NN, MLTSVM, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM with 10 Rounds of 5-Fold Cross-Validation in the CV Method Compared with 7-Class RF

model (CV)	ave_ham loss	ave_top-1%	ave_top-2%	ave_top-3%
ML-*k*NN	0.2629 ± 0.0048	78.5 ± 1.3	89.9 ± 1.4	95.3 ± 0.8
MLTSVM	0.3016 ± 0.0068			
NLSD-MLP	0.2490 ± 0.0032	83.8 ± 1.3	93.8 ± 1.1	97.4 ± 0.3
NLSD-SVM	0.2419 ± 0.0034	84.7 ± 0.8	95.4 ± 0.5	98.1 ± 0.5
NLSD-RF	0.2256 ± 0.0042	86.6 ± 0.6	94.6 ± 0.3	97.5 ± 0.3
NLSD-EXT	0.2195 ± 0.0025	87.2 ± 0.6	95.4 ± 0.5	97.7 ± 0.6
NLSD-XGB	0.2313 ± 0.0041	87.6 ± 0.4	95.1 ± 0.6	97.6 ± 0.4
WhichP450[14]		76.3 ± 5.4	88.4 ± 4.9	93.3 ± 3.4

Model Evaluation and Comparison with Other Methods. After identifying the best combination of features, we applied seven different multilabel models using these features, which are ML-*k*NN, MLTSVM, and five network-based label space division methods (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM). Top-*k* and Hamming loss on these seven multilabel models were compared in both CV and HO methods.

In CV methodology, the top-*k* and hamming loss results for ML-*k*NN were built and validated using the same method, 5-fold cross-validation, as WhichP450[14] are detailed in Table 2. Table 2 shows the model evaluation with 10 rounds of 5-fold cross-validation on the modeling process intuitively and in detail. Each column represents the average of top-*k* and hamming loss obtained by 5-fold cross-validation on the randomly shuffled dataset. The average of these 10 results was considered as our final model evaluation.

The seven-class random forest used in WhichP450[14] was conducted with the 5-fold cross-validation only once. From this, we could see that the performance of ML-*k*NN in our study was more stable than 7-class RF. As shown in Table 2, in the ML-*k*NN model, the average top-1 prediction success is 78.5%, the average top-2 prediction success is 89.9%, and the average top-3 prediction success is 95.3%, which shows that ML-*k*NN performs better than the model of the previous study. Besides, with the best feature combination, we established seven multilabel models in the CV method and compared top-*k* and hamming between the models, as shown in Table 3.

Table 3 shows the comparisons of the average of hamming loss and top-*k* on ML-*k*NN, MLTSVM, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM with 10 rounds of 5-fold cross-validation in the CV method and the previous work. The MLTSVM method runs very slowly on this small dataset, so it is not recommended for use in real applications. In addition, MLTSVM does not perform very well on this problem, so we do not think it is necessary to compare the top-*k* on it with other models. Except for MLTSVM, the top-*k* values of our remaining six models are all higher than those of the previous work and all six models are more stable. NLSD-XGB achieves the best performance with the average top-1 prediction success, 87.6%, the average top-2 prediction success, 95.1%, and the average top-3 prediction success, 97.6%. The results of 10 repetitions of hamming loss and top-*k* on ML-*k*NN, MLTSVM, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM with 10 rounds of 5-fold cross-validation in the CV method and the previous work are detailed in Tables S2−S8 of the Supporting Information.

Figure 6 shows the top-*k* values line graph of six models—ML-*k*NN, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM, which are compared with those of the model in the previous work, that is, the seven-class random forest model in the CV method. We can see that the top-*k* values of our six models are all higher than those of 7-class RF, among which five network-based label space division methods (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM) perform better in our dataset and NLSD-XGB has the best performance in the CV method.

In HO methodology, the whole dataset was divided into the training set and the test set. Similarly, with the best feature combination, we established seven multilabel models in the HO method and compared top-*k* and hamming loss between

DOI: 10.1021/acs.jcim.9b00749

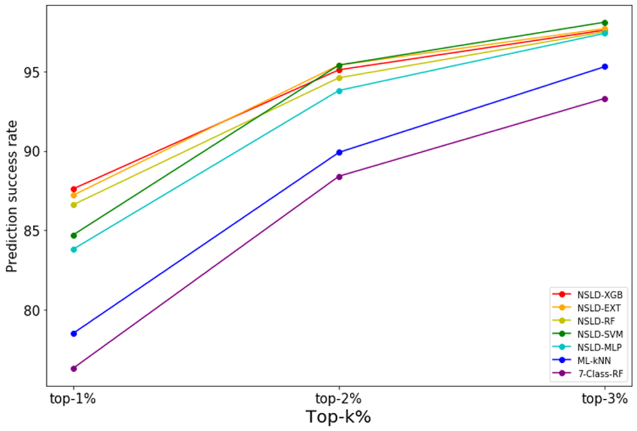


Figure 6. Top-*k* value line graph of ML-*k*NN, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, NLSD-SVM, and 7-class RF in the CV method.

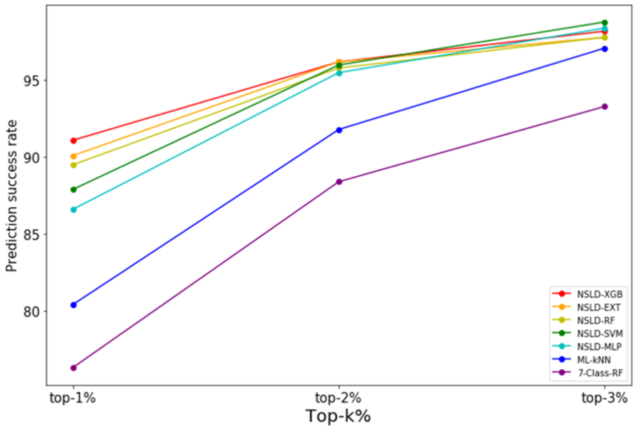


Figure 7. Top-*k* value line graph of ML-*k*NN, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM in the HO method and 7-class RF-CV.

the models, as shown in Table 4. The average of top-*k* and hamming loss was obtained by 5-fold cross-validation on the given randomly shuffled training set. Then, the average of these 10 results validated on the separate test set was considered as our final model evaluation.

Table 4 shows the comparisons of the average of hamming loss and top-*k* on ML-*k*NN, MLTSVM, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM with 10 rounds of 5-fold cross-validation in the HO method and 7-class RF-CV (previous work).[14] Except for MLTSVM, the top-*k* values of our remaining six models are all higher than those of the previous work,[14] and all six models are more stable. From Table 4, we can see that there is no overfitting tendency in all models, and the classification model of NLSD-XGB achieves the best performance with the average top-1 prediction success, 91.1%, the average top-2 prediction success, 96.2%, and the average top-3 prediction success, 98.2%. The results of 10 repetitions of hamming loss and top-*k* on ML-*k*NN, MLTSVM, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM with 10 rounds of 5-fold cross-validation in the HO method and the previous work[14] are detailed in Tables S9−S15 of the Supporting Information. This conclusion is consistent with the result in the CV method that NLSD-XGB can also be regarded as the best method on this dataset with 10 rounds of 5-fold cross-validation in the HO method.

Figure 7 shows the top-*k* value line graph of six models—ML-*k*NN, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM in the HO method, which is compared with those of the previous work's model[14]—seven-class random forest model.

In the 10 repetitions on this dataset, top-1 was used as the most stringent evaluation index. It can be concluded that NLSD-XGB performs best in the CV and HO validation techniques. In addition, the performance of NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM are similar, so we cannot give the exact NLSD-X method, which performs best in tackling multilabel classification problems. However, it may be proved that the NLSD method is superior to other common multilabel classification algorithms in our study. It is worth mentioning that the performance of NLSD based on deep learning algorithm (NLSD-MLP) on this dataset is not better than those of other NLSD methods based on other machine learning algorithms (XGB, EXT, RF, and SVM) in the multilabel learning task, the conclusion of which is consistent with the research of Raies et al.[42] For the isoform-wise metrics, the accuracy, specificity, recall, F_1 score, and area under the curve are listed in Table S17.

Finally, the network-based label space division model has been proved to perform well in our study. There may be several reasons as follows: the NLSD-X method used the network inspired and data-driven algorithm to divide the label space into subspaces and used the well-established base classifier for each partition. This method improved the performance of the label-powerset method and binary relevance by converting the label sets into subspaces with significantly reduced set cardinality, thus preventing overfitting. In addition, the NLSD-X method assumed that the label

Table 4. Average of Hamming Loss and Top-*k* of Multilabel Models Based on ML-*k*NN, MLTSVM, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM with 10 Rounds of 5-Fold Cross-Validation in the HO Method Compared with 7-Class RF-CV

model (HO)	ave_ham loss on training set	ave_ham loss on test set	ave_top-1%	ave_top-2%	ave_top-3%
ML-*k*NN	0.2647 ± 0.0061	0.2648 ± 0.0229	80.4 ± 5.5	91.8 ± 2.9	97.1 ± 1.7
MLTSVM	0.2950 ± 0.0146	0.2990 ± 0.0287			
NLSD-MLP	0.2457 ± 0.0084	0.2489 ± 0.0253	86.6 ± 3.9	95.5 ± 2.0	98.4 ± 1.6
NLSD-SVM	0.2430 ± 0.0087	0.2431 ± 0.0312	87.9 ± 3.0	96.0 ± 1.3	98.8 ± 1.0
NLSD-RF	0.2302 ± 0.0066	0.2196 ± 0.0221	89.5 ± 2.6	95.8 ± 1.4	97.8 ± 0.9
NLSD-EXT	0.2230 ± 0.0060	0.2143 ± 0.0262	90.1 ± 2.8	96.2 ± 1.0	97.8 ± 1.1
NLSD-XGB	0.2387 ± 0.0057	0.2221 ± 0.0143	91.1 ± 2.1	96.2 ± 2.3	98.2 ± 1.1
WhichP450[14]			76.3 ± 5.4	88.4 ± 4.9	93.3 ± 3.4

relationship existing in the training data is representative, and our dataset satisfies the hypothesis.

CONCLUSIONS

In this study, the raw dataset we used contains 484 compounds and 1299 compound/isoform pairs. After data labeling, we obtained the multilabel dataset with a 484 × 7-dimensional label. Then, the physiochemical descriptors, mol2vec descriptors, ECFP, and MACCS key fingerprints of all compounds were calculated and considered as features for modeling. Besides, we have combined the four categories of features into 15 different feature combinations to select the best feature combination. Considering ML-*k*NN as the baseline model, the model with mol2vec descriptors performs best among the models built with only one category of features, of which the average top-1 prediction success is 77.5%, the average top-2 prediction success is 88.4%, and the average top-3 prediction success is 93.7%. Only based on mol2vec, the top-*k* values of the benchmark model are better than the previous work.[14] Among 15 different feature combinations, the combination of physiochemical property descriptors, mol2vec descriptors, and ECFP fingerprints used for modeling achieves the best top-*k* values, of which the average top-1 prediction success is 80.2%, the average top-2 prediction success is 90.9%, and the average top-3 prediction success is 95.9%. Therefore, we chose physiochemical descriptors, mol2vec descriptors, and ECFP fingerprints as the final feature combination to build seven multilabel models.

Next, we applied seven different multilabel models with 10 rounds of 5-fold cross-validation using these features, which are ML-*k*NN, MLTSVM, and five network-based label space division methods (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM). In model comparison, top-*k* and hamming loss on these seven multilabel models were compared in both CV and HO methods.

In CV methodology, NLSD-XGB achieves the best performance with the average top-1 prediction success of 87.6%, the average top-2 prediction success of 95.1%, and the average top-3 prediction success of 97.6%.

In HO methodology, NLSD-XGB achieves the best performance with the average top-1 prediction success, 91.1%, the average top-2 prediction success, 96.2%, and the average top-3 prediction success, 98.2%. The six models (ML-*k*NN, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM) in this work all produce better performances than the previous work.[14] Besides, when compared with the previous work,[14] NLSD-XGB shows a significant improvement over 11% on top-1 in the CV method and over 14% on top-1 in the HO method. Thus, we considered NLSD-XGB as our best method on this dataset finally.

To the best of our knowledge, MLTSVM[30] and network-based label space division models (NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM)[18] were first applied in the field of drug metabolism. The top-1 values of five network-based label space division models were all significantly improved compared with those of the previous work such as WhichP450.[14]

Finally, the network-based label space division model has been proved to perform well in our study. Although we have achieved the state-of-the-art performance in the task, several drawbacks still exist. First, the dataset we used is relatively small, which limits the generalization ability of our models. Second, we only considered five commonly used base classifiers as base models for label space partition, without making a complete survey of possible classifiers.

Apart from all of the aforementioned drawbacks, the merits of the network-based label space division model are still ill-explored in multilabel prediction tasks in biomedical research. We suggest that robust testing and comparison should be performed on this method on various tasks specifically for biomedical research.

ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00749.

> Explanation of top-*k* (Table S1); results of 10 repetitions of hamming loss and top-*k* on models with 10 rounds of 5-fold cross-validation in cross-validation and hold-out methods (Tables S2–S15); the average of hamming loss and top-*k* of multilabel models based on ML-*k*NN, NLSD-MLP, NLSD-XGB, NLSD-EXT, NLSD-RF, and NLSD-SVM with 1000 rounds of 5-fold cross-validation in the CV method (Table S16); the isoform-wise metrics for NLSD-XGB with 10 rounds of 5-fold cross-validation (Table S17) (PDF)
>
> Details of the datasets used in this study (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: xiongyi@sjtu.edu.cn. Phone/Fax: +86 21-34204573 (Y.X.).
*E-mail: dqwei@sjtu.edu.cn (D.-Q.W.).

ORCID

Xiangeng Wang: 0000-0002-2510-7985
Yi Xiong: 0000-0003-2910-6725
Dong-Qing Wei: 0000-0003-4200-7502

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

REFERENCES

(1) Klingenberg, M. Pigments of Rat Liver Microsomes. *Arch. Biochem. Biophys.* **1958**, *75*, 376–386.

(2) Feiters, M. C.; Rowan, A. E.; Nolte, R. J. M. From Simple to Supramolecular Cytochrome P450 Mimics. *Chem. Soc. Rev.* **2000**, *29*, 375–384.

(3) Nelson, D. R.; Kamataki, T.; Waxman, D. J.; Guengerich, F. P.; Estabrook, R. W.; Feyereisen, R.; Gonzalez, F. J.; Coon, M. J.; Gunsalus, I. C.; Gotoh, O.; et al. The P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers, Early Trivial Names of Enzymes, and Nomenclature. *DNA Cell Biol.* **1993**, *12*, 1–51.

(4) Zhang, T.; Dai, H.; Liu, L. A.; Lewis, D. F. V.; Wei, D. Classification Models for Predicting Cytochrome P450 Enzyme-Substrate Selectivity. *Mol. Inf.* **2012**, *31*, 53–62.

(5) Preissner, S. C.; Hoffmann, M. F.; Preissner, R.; Dunkel, M.; Gewiess, A.; Preissner, S. Polymorphic Cytochrome P450 Enzymes

(Cyps) and Their Role in Personalized Therapy. *PLoS One* **2013**, No. e82562.

(6) Hong, J. Y.; Yang, C. S. Genetic Polymorphism of Cytochrome P450 as a Biomarker of Susceptibility to Environmental Toxicity. *Environ. Health Perspect.* **1997**, *105*, 759−762.

(7) Seredina, T. A.; Goreva, O. B.; Talaban, V. O.; Grishanova, A. Y.; Lyakhovich, V. V. Association of Cytochrome P450 Genetic Polymorphisms with Neoadjuvant Chemotherapy Efficacy in Breast Cancer Patients. *BMC Med. Genet.* **2012**, *13*, 45.

(8) Xiong, Y.; Qiao, Y.; Kihara, D.; Zhang, H.-Y.; Zhu, X.; Wei, D.-Q. Survey of Machine Learning Techniques for Prediction of the Isoform Specificity of Cytochrome P450 Substrates. *Curr. Drug Metab.* **2019**, *20*, 229−235.

(9) Li, X.; Xu, Y.; Lai, L.; Pei, J. Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharm.* **2018**, *15*, 4336−4345.

(10) Tyzack, J. D.; Hunt, P. A.; Segall, M. D. Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations. *J. Chem. Inf. Model.* **2016**, *56*, 2180−2193.

(11) Zou, Q.; Chen, W.; Huang, Y.; Liu, X.; Jiang, Y. Identifying Multi-Functional Enzyme by Hierarchical Multi-Label Classifier. *J. Comput. Theor. Nanosci.* **2013**, *10*, 1038−1043.

(12) Zhang, W.; Zhu, X.; Fu, Y.; Tsuji, J.; Weng, Z. Predicting Human Splicing Branchpoints by Combining Sequence-Derived Features and Multi-Label Learning Methods. *BMC Bioinf.* **2017**, *18*, No. 464.

(13) Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Comparison of Multilabel and Single-Label Classification Applied to the Prediction of the Isoform Specificity of Cytochrome P450 Substrates. *J. Chem. Inf. Model.* **2009**, *49*, 2588−2605.

(14) Hunt, P. A.; Segall, M. D.; Tyzack, J. D. Whichp450: A Multi-Class Categorical Model to Predict the Major Metabolising Cyp450 Isoform for a Compound. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 537−546.

(15) Zhang, M.-L.; Zhou, Z.-H. In *A K-Nearest Neighbor Based Algorithm for Multi-Label Classification*, 2005 IEEE International Conference on Granular Computing, IEEE, 2005; pp 718−721.

(16) Gibaja, E.; Ventura, S. Multi-Label Learning: A Review of the State of the Art and Ongoing Research. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery* **2014**, *4*, 411−444.

(17) Tsoumakas, G.; Katakis, I.; Vlahavas, I. Random K-Labelsets for Multilabel Classification. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1079−1089.

(18) Szymański, P.; Kajdanowicz, T.; Kersting, K. How Is a Data-Driven Approach Better Than Random Choice in Label Space Division for Multi-Label Classification? *Entropy* **2016**, *18*, No. 282.

(19) Wang, X.; Wang, Y.; Xu, Z.; Xiong, Y.; Wei, D.-Q. ATC-NLSP: Prediction of the Classes of Anatomical Therapeutic Chemicals Using a Network-Based Label Space Partition Method. *Front. Pharmacol.* **2019**, *10*, No. 971.

(20) Zhang, M.-L.; Zhou, Z.-H. A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819−1837.

(21) Moyano, J. M.; Gibaja, E. L.; Cios, K. J.; Ventura, S. Review of Ensembles of Multi-Label Classifiers: Models, Experimental Study and Prospects. *Inf. Fusion* **2018**, *44*, 33−45.

(22) Irwin, J. J.; Shoichet, B. K. Zinc − a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(23) van der Maaten, L.; Hinton, G. Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(24) Bergsma, W. A Bias-Correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* **2013**, *42*, 323−328.

(25) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 2013, arXiv:1301.3781. arXiv.org e-Print archive. https://arxiv.org/abs/1301.3781.

(26) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.

(27) Cao, Q.; Liu, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Estimation of Chemical Aquatic Toxicity on Crustaceans Using Chemical Category Methods. *Environ. Sci.: Processes Impacts* **2018**, *20*, 1234−1243.

(28) O'Boyle, N. M.; Sayle, R. A. Comparing Structural Fingerprints Using a Literature-Based Similarity Benchmark. *J. Cheminf.* **2016**, *8*, No. 36.

(29) Zhang, M.-L.; Zhou, Z.-H. Ml-Knn: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recognit.* **2007**, *40*, 2038−2048.

(30) Chen, W.-J.; Shao, Y.-H.; Li, C.-N.; Deng, N.-Y. Mltsvm: A Novel Twin Support Vector Machine to Multi-Label Learning. *Pattern Recognit.* **2016**, *52*, 61−74.

(31) Fukunaga, K.; Hostetler, L. Optimization of K Nearest Neighbor Density Estimates. *IEEE Trans. Inf. Theory* **1973**, *19*, 320−326.

(32) Zhang, W.; Liu, F.; Luo, L.; Zhang, J. Predicting Drug Side Effects by Multi-Label Learning and Ensemble Learning. *BMC Bioinf.* **2015**, *16*, No. 365.

(33) Liu, G.-P.; Li, G.-Z.; Wang, Y.-L.; Wang, Y.-Q. Modelling of Inquiry Diagnosis for Coronary Heart Disease in Traditional Chinese Medicine by Using Multi-Label Learning. *BMC Complementary Altern. Med.* **2010**, *10*, No. 37.

(34) Khemchandani, R.; Chandra, S. In *Twin Support Vector Machines for Pattern Classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2007; pp 905−910.

(35) Joachims, T. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Proceedings of ECML-98, 10th European Conference on Machine Learning, Springer, 1998.

(36) Zhang, Z.; Lyons, M.; Schuster, M.; Akamatsu, S. In *Comparison between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron*, Proceedings International Conference Automatic Face and Gesture Recognition, IEEE, 1998.

(37) Chen, T.; Guestrin, C. In *Xgboost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016; pp 785−794.

(38) Bremer, K. Branch Support and Tree Stability. *Cladistics* **1994**, *10*, 295−304.

(39) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *Newsletter of the R Project*, 2002, Vol. *2*, pp 18−22.

(40) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.: Theory Exp.* **2008**, No. P10008.

(41) Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier Chains for Multi-Label Classification. *Mach. Learn.* **2011**, *85*, 333.

(42) Raies, A. B.; Bajic, V. B. In Silico Toxicology: Comprehensive Benchmarking of Multi-Label Classification Methods Applied to Chemical Toxicity Data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, No. e1352.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on October 22, 2019, with an error in the affiliations. The corrected version was reposted October 29, 2019.

DOI: 10.1021/acs.jcim.9b00749