Supervised learning techniques to predict compounds in pathway modules based on molecular properties

Hayat Ali Shah, Juan Liu, Zihui Yang, Xiaolei Zhang, and Jing Feng

Abstract—Machine learning algorithms provide significant indications in metabolomics to predict chemical compounds in metabolic pathways and their modules. The modules in the metabolic pathway are subnetworks of functionally related genes based on rules such as protein-protein interactions, co-regulated expression, coordinated physiological activity, and successive reaction steps. All modules in the metabolic pathway are not functional due to missing reaction steps. Fully functional module are helpful to improve the diseases process, drug discover, and prediction of missing reaction. The structural mapping of chemical compounds with the pathway module is necessary to predict unknown reaction step. The main purpose of this paper to predict the chemical compounds in pathway modules and their classes. Here, we proposed machine learning algorithms extra tree classifier (ETC) to learn the molecular and atomic properties of chemical compounds to predict pathway modules. Our method predicts and maps chemical molecules to the metabolic pathway module and their classes as binary and multi-label classification problems. The overall prediction rate of the classifier 98.59%, indicating extra tree classifier features are more interpretable and have a high predictive performance on a variety of tasks.

Index Terms—Chemical compound, machine learning, metabolic pathway, pathway module, binary and multi-labeled classification

1 INTRODUCTION

The metabolic pathway is the series of biochemical reactions between the start substrates and the final product, where the product of one enzyme is the substrates of subsequent enzymes. The metabolic pathway is decomposed into a subnetwork (module) of functionally related genes based on rules such as phenotypic features, protein to protein interaction, and co-regulated expression. These modules are categorized into three types: pathway module, signature module, and reaction module existing in public pathway databases. Several public databases are available such as KEGG [1], [2], MetaCyc [3], [4], Brenda [5], Rhea [6], comprising major components of the pathway module, including chemical compounds, reactions, and enzyme-substrates binding. However, some hidden reactions, enzymes, chemical compounds have not been detected in the pathways modules, which is a great barrier to understanding pathway modules' function. The mapping of small molecules' with their corresponding pathway modules is essential to predict missing elements of the pathway modules.

A variety of methods have been established in vivo, to analyze the role of compounds in metabolic pathways based on physical, biological or biochemical experiments. However, these methods lead to the problems of high cost and low efficiency, and high-throughput equipments are required for the analysis of compounds in metabolic pathways. To overcome, these limitations machine learning methods have incredible ability to adapt structure

and chemical activity of compounds for the prediction of metabolic pathways. Until now, many researchers approached to machine learning methods for the prediction of chemical compounds with the metabolic pathway classes. Amongst them, Cai et al. [7] proposed a single-label nearest neighbor algorithm(NNA) to classify compounds into pathway types existing in KEGG. Similar approaches adopted by Hu et al. [8] replaced the K-nearest neighbor (KNN) algorithm with AdaBoost to predict the metabolic pathway compounds. Later, random forest [9] as a multi-label classifier was adopted by Macchiarulo et al. [10], Barnawal et al [11], and Jia et al. [12] for the prediction of metabolic pathway classes. Barnawal et al. [11] developed the random forest with graph convolution neural network (GCN) to predict metabolic pathway classes, and Jia et al. [12] targeted the actual metabolic pathway to which chemical compound belongs. However, these all methods only consider the association of compounds with the metabolic pathway based on biological function. In addition, a metabolic pathway is a large network consisting of several modules. All pathway modules in the metabolic pathway are not fully functional owing to the lack of experimentally identified compounds. These unknown compound may cause misleading interpretation and non-functional of modules. Activate the non-functional modules, it is necessary to predict which type of compounds link with the pathway module. Besides, the analysis of compounds with pathway functional unit (module) may improve the understanding of diseases, diagnosing and drug discovery.

In this article we develop machine learning method for prediction of query compound, whether belong or not to pathway modules. We define feature vectors representing the chemical transformation patterns of compound using MACCS (Molecular Access System) chemical fingerprints.

- Email: hayatali@whu.edu.cn (Hayat Ali Shah)
- Email: liujuan@whu.edu.cn (Juan Liu) (Corresponding author)
- Email: zhy@whu.edu.cn (Zihui Yang)
- Email: 2020202110014@whu.edu.cn (Xiaolei Zhang)
- Email: gfeng@whu.edu.cn (Jing Feng)

We apply an ensemble Extra trees classifier (ETC) [13] has been widely applied in bioinformatics for a variety of computational task, including diagnosing diseases [14], tumor segmentation [15], cyber-attack detection [16], time-series classification [17]. The purpose of ensemble machine learning is to combine different base estimators with improving robustness and generalizability over a single estimator.

The focus of this paper comprising two main contributions. First, we apply machine learning algorithms ETC to predict the metabolic pathways and pathway modules to which query compounds belong. In term of machine learning these predictions turns to binary classification, where our model classifies the data into two possible outcomes (a) compounds belong to the pathway module, (b) query compound does not belong to the pathway modules. Furthermore, to enhance the process of diagnosing and drug discovery, it is essential to analyze the structure, chemical behavior and biological function of compounds. Based on structure and common biological function, these compounds are classified into 10 different categories, such as carbohydrates, energy and lipid metabolism. Besides, some compounds are participating in multi classes for a different function such as Glyceraldehyde-3 Phosphate belong to Carbohydrate metabolism, Energy metabolism, Biosynthesis of terpenoids and polyketides, and Biosynthesis of other secondary metabolites. For multi classes classification we develop ETC to predicts the classes of pathway modules. The possible probabilistic outcomes of classifier will show the real class of compounds. The performance of our proposed classifier is also competitive with previous work and other ensemble machine learning classifiers.

2 MATERIALS

The chemical compounds datasets of metabolic pathways and modules were retrieved from KEGG databases (<http://www.genome.jp/kegg/module.html> and <https://www.kegg.jp/kegg/pathway.html>) accessed in October 2020. For the binary classification, we retrieved 6664 different compounds; 4612 were associated with binary classification, whether link or not with the pathway modules. The remaining compounds were discarded from this experiment based on similar structure. For binary classification, our model access each compound $2(n)$ times, where n shows the number of compounds in datasets. The access of all compounds in binary classification are 9224. Furthermore, 1985 different compounds were retrieved for multi-class classification from the KEGG module database (<https://www.genome.jp/kegg/module.html>) accessed in October 2020. These compounds were associated with 10 different pathway module classes, including carbohydrate metabolism (CM), energy metabolism (EM), lipid metabolism (LM), nucleotide metabolism (NM), amino acid metabolism (AM), glycan metabolism (GM), metabolism of cofactors and vitamins (MCV), biosynthesis of terpenoids and polyketides (BTP), biosynthesis of other secondary metabolites (BOSM), and xenobiotic biodegradation (XB). The overall label of compounds for these classes are 2908, including single label and multi-label classes. The statistics of label compound of each class shown in figure 3A. Each compound's possible

outcome could be $10(n)$, and for all compounds, the classifier can be classified 19850 possible outcomes.

Besides, these compounds were downloaded in SMILES (simplified molecular-input line-entry system) format. During retrieving, many compounds were not available in SMILES format. The SDF (structure-data file) format of these compounds translated into SMILES structure using online Chemoinformatics tools. Some compounds are involved in more than one pathway module, recommending these compounds as a multi-label classification problem.

3 METHODS

We developed a machine learning ensemble extra tree classifier for predicting the pathway module and classes of modules using molecular and atomic properties as features. Classifier takes compounds in SMILES format along with 166 MACCS keys and 7 additional molecular descriptors, including molecular weight, rotatable bonds, ring counts, lipophilicity, aromaticity, and polarizability. These hyperparameters were applied by Baranwal et al. [11] for the prediction of metabolic pathway classes. To yield the optimal performance, hyperparameters of the ETC were determined and optimized through grid searching [18] shown in section 4.5. Our algorithm averages tree predictions by splitting the input space randomly. The chemical compounds datasets were splatted into training and testing datasets for learning and prediction purposes. The ETC learn the atom and molecular properties, structure of compounds, and predict the appropriate classes for data. The process of preprocessing, splitting of data, and prediction shown in Figure 1. This method was applied for both binary and multi-class classification problems. First, our model classified data into two classes $C \in (\text{pathway module})$ or $C \notin (\text{pathway module})$, where output either belongs or not with the pathway module in the metabolic pathway.

The second, For multi-class classification, our classifier classify inputs chemical compounds across single or multi classes of pathway modules. In this case, the output of the ensemble classifier is a probability distribution over ten different classes. More precisely, the pathway module classes can be expressed by numbers $C \in \{0, 1, 2, 3, \dots, 9\}$, where the numbers from 0 to 9 shows the number of module classes in KEGG. This shows that each module classes belong to the query compound. Multi-class classification having a set of classes C , where $(C = c_1, c_2, c_3, \dots, c_{10})$, the task of a classifier is to assign c from the set of classes C to query compound on all dataset samples, $x = (d_1, d_2, d_3, \dots, d_m)$ consisting of n attribute values $d_i \in D_m$. For classification, the model needs to be trained on training data. The D is divided into DT and $\widehat{DT}$ training, and testing data, respectively. The model M train on DT and predict $\widehat{DT}$, where can map unseen query compound to corresponding classes $c \in C$. The model M predicts module classes based on the likelihood of the query compound belonging to the class.

As a machine learning model, we developed ETC that aggregates the outputs of multiple correlated decision trees collected in a forest to output its classification. Each decision tree is constructed from the training sample and provided with k features. Each decision tree selects the best feature for splitting the data, based on mathematical criteria. The

iteration of ETC starts to select m features randomly as a candidate set of splitting features. Within each of these features, F_i , with $i \in (1, \dots, m)$ draws a single random cutpoint equally from the interval $(\min(F_i), \max(F_i))$, evaluates the performance of this feature with this cutpoint regarding entropy. Finally, features paired with their randomly selected cutpoint and select the best cutpoint. The given formula can measure the entropy of the data:

$$E(S) = -\sum_i^h p_i \log_2 p_i \quad i \in (1, 2, \dots, 10) \quad (1)$$

Where S is the number of unique class labels, and p_i is simply the frequentest probability of a class i . In an account of the pathway module, $n \in (1, \dots, 10)$ the input data is labeled across ten different classes. The data disorder can be reduced in our target classes by the information gain.

$$IG(Y, X) = E(Y) - E\left(\frac{Y}{X}\right) \quad (2)$$

Where, Y is the target pathway module class and X is the input of the classifier. To calculate the reduction of uncertainty, subtract the entropy of Y given X from Y entropy; the more information is ($C = c_1, c_2, c_3, \dots, c_{10}$) gained about Y from X , the greater uncertainty reduction.

Fig. 1. A comprehensive process for constructing the single label and multi-label classification model and its evaluation. The two types of compounds were extracted from the KEGG pathway module and KEGG pathway databases. The SDF files of compounds were converted into SMILES molecular structure format by using online Cheminformatics tools. Then, each compound is labeled by its corresponding classes according to chemical-chemical properties. After preprocessing, ETC selects the random simple from data and construct DT, get a prediction for each simple. Finally, our model perform voting to predict the result, and select the appropriate class for the query compound.

4 EXPERIMENTS

We performed binary and multi-class classification based on two different types of datasets for predicting pathway modules in the metabolic pathway and module classes by developing ensemble classifiers. We also compared our ensemble model with RF, and collection of classifiers, including KNN, DT, RF. All ensemble classifiers trained on

similar molecular descriptors of compounds data sets with their own hyperparameters. We tuned the hyperparameters for all classifiers by performing a grid search over the set of possible hyperparameters settings to achieve high performance.

4.1 Extra Trees Classifier

The Extra Trees algorithms work by making an enormous number of unpruned decision trees from the training datasets. The final prediction is established by using majority voting on account of classification. The ETC working is different from other tree-based ensemble methods. It splits the node randomly by picking cut points, which will decrease variance better than other randomization strategies. Based on random splitting, the execution time of the ETC is faster. Owing to the computational efficiency, the Extra trees algorithm has massive applications for classification and regression [16], [19], [20].

4.2 Group of Classifiers

We integrated a group of classifiers, including KNN, DT, RF, for both module and module classes to achieve better performance than ETC. The hyperparameters of each classifier were selected from their possible hyperparameters. For good performance, the hyperparameters of each classifier tuned by grid search methods. However, these integrated classifiers' performance worse than ETC and single RF in both experiments.

4.2.1 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a supervised learning algorithm used for classification and regression problems in bioinformatics [21], [22], [23]. It is assuming every data point near to each other is falling in the same class. KNN algorithms classify new data points based on a similarity measure by a majority vote to its neighbors. We integrated KNN with the number of neighbors 1, leaf size 6, and the value of power parameter (P) is 1; this is equivalent to using Manhattan-distance (1) [24].

4.2.2 Decision Tree (DT)

A decision tree belongs to the supervised learning algorithms for regression and classification problems. The DT [25], [26] tries to solve the problems, by using tree representation, where each of the trees relates to attributes, and each leaf node corresponds to the class label. The leaf nodes are the final nodes or decision nodes of the model. The DT creates a training model to predict class by learning decision rules inferred from prior training data. In this experiment, DT implemented the hyperparameters criterion "Gini," The maximum depth of the tree is 10.

4.3 Random Forest (RF)

Random forest is an ensemble of DTs, trained with the bagging method used for classification and regression that operate by constructing a group of DTs to predict the mode of classes [27]. Generally, the bagging method is the combination of a learning model to increase the overall

performance. We apply the RF algorithm on both experiments to compare the abilities with our model. The hyperparameters of the RF is similar to the ETC, shown in the supplementary material. However, RF subsamples the input data with replacement, whereas ETC uses the whole original sample. The performance of RF with binary and multi-label classification is not good as an ETC.

4.4 Implementation

All these classifiers models are implemented in Python 3.8 with Keras library on an Intel(R) Core TM i7-4600U CPU @ 2.70 and 2.10 GHz Processor with a 64-bit operating system, x64 based processor. The structure of compounds were encoded to corresponding physio chemical structure by MACCS keys and RDKit. For ensemble classifiers, we used the readily available implementation in the Scikit-learn module. The hyperparameters of classifiers were tuned and optimized by Grid search optimization method.

4.5 Hyperparameters Optimization

Like other machine learning algorithms, the Extra trees classifier's performance can depend greatly on the selection of different hyperparameters, such as the number of estimators, criterion, maximum features, depth, etc. The accuracy, precision, and recall of the Extra trees classifier with default hyperparameters were 97.94, 84.67, and 85.43, respectively. To maximize our classifier's performance, we perform hyperparameters optimization via the Grid Search [18]. We tuned the most important hyperparameters for our data and selected the number of estimators (200), maximum features (1.0), maximum depth (60) of the decision tree, and the criterion (entropy). As a consequence of our experiments, the classifier shows high performance based on the hyperparameters' best selection. After optimizing hyperparameters, the accuracy, precision, and recall are 98.59, 90.70, and 91.71, respectively. The illustration of grid search with different hyperparameters is shown in Figures 4A, 4B, and 4C.

Fig 4A

Fig 4B

Fig 4C

Figure 4. The figures show the statistics of classifier performance with different hyperparameters.

4.6 Prediction of pathway module (binary- classification)

Prediction of pathway module is a binary classification problem in machine learning. Binary classification datasets labeled with "Zero" and "Ones", where 1 and 0 represents the chemical compounds belong or not with pathway modules respectively. For binary classification, we retrieved 4614 biochemical compounds in SMILES structure, 2117 compounds do not belong to any pathway module in metabolic pathways noted it negatives, and the remaining 2497 belong to pathway modules marked it positives in our data. These datasets are randomly selected for the training, testing, and validation of the classifier. The prediction of the classifier belongs to ($n = 1, 2$) possible outcome classes. The datasets were split into training and testing data for training and validations. The model predicts the query compound belongs to the pathway module or not, based on the input data. After preprocessing the data, extra tree classifier was implemented to predict the pathway module. After prediction the pathway modules we also implemented RF and different integrated classifiers to compare the metric performance with other ensemble classifiers. The metric score of the ETC and other ensemble classifiers shown in table 1.

4.6.1 Comparison with Other Ensemble Classifiers

We also adopted other widely used machine learning classifiers in bioinformatics RF, collections of other classifiers, including KNN, DT, and RF. We evaluated our model for the prediction of pathway modules in the metabolic pathway. Besides, we compared our model with existing methods for the prediction of the binary classification problems. Jia et al. [12] used the same type of data sets for the prediction of actual metabolic pathways, to calculate the specificity (SP), sensitivity (SN), accuracy (ACC), precision, F1-measure [28], [29], and Matthews correlation coefficient (MCC) [30] according to the given formulas:

The accuracy of binary classification is defined as follows:

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Here, the accuracy is the fraction correctly predicted of all query compounds associated with the pathway module in the metabolic pathway. The model is also needed observed precision and recall for performance measurement.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall(SN) = \frac{TP}{FP + FN} \quad (5)$$

$$F1 - measure = 2 \times \frac{Precision}{Recall} \quad (6)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TN + FP) \times (TP + FN) \times (TP + FP)}} \quad (7)$$

Here, true positive (TP), the chemical compound belongs to pathway module and model declared it is belonged with the module, true negative (TN), the compound is not present in pathway module and model declared it is not-link. False-positive (FP), the compound is not-belonging with pathway module and model shows, it is belong with pathway module, false negative (FN) module declared it is not-link, but it is the part of pathway module.

Fig 2 (A). ETC PRAUC

Fig 2 (B). RF, PRAUC

Fig 2(C). Classifiers PRAUC

Fig 2 (D). ETC, AUROC

TABLE 1
Comparison of classifiers for pathway module prediction in metabolic pathways

Methods	Reference	Accuracy	Precision	Recall	MCC	F1-Score
Random forest (local ten-fold cross validation)	Jia et al [12]	93.42	94.37	92.49	86.86	93.35
Random forest (Global ten-fold cross-validation)	Jia et al [12]	93.29	93.99	92.49	86.59	93.24
Ensemble classifiers(KNN, DT, RF)	compared in this experiments	92.49	92.22	92.29	84.98	92.24
Extra trees classifier	Our	94.11	94.05	94.05	88.23	94.05
Random forest	compared in this experiments	93.23	93.23	93.23	86.47	93.23

Fig 2(E). RF, AUROC

Fig 2(F). Classifiers, AUROC

Figure 2. Shows the pathway module prediction performance of classifiers with PR and ROC curve.

Prior studies frequently focused on metabolomics on ensemble RF instead of other classifiers to predict binary, multi-label, and multi-class classification based on its effective performance. Therefore, we implemented RF alone and ensemble with other classifiers on our data, compared with ETC. The bold value of the arguments shows the high

performance of the classifier in table 1. The ETC has high-performance metrics as compared to prior work and as well as other classifier used in these experiments. We did a precision-recall curve (PR) and receiver operating characteristic curve (ROC) analysis on three different ensemble models mentioned in figure 2. The obtained curve, AUROCs, and AUPRs show that the ETC performance is higher than the other two ensemble RF and integrated classifiers. The curves also show that the classifier’s performance is worse than the other two methods. As a consequence of these analyses, the ETC is more related to the atom and molecular properties features in pathways modules.

4.7 Prediction of Pathway Module (Multi-Class Classification)

We performed second experiments for the predictions of compounds in multiple pathway module classes. In terms of machine learning classification, module class classification turned into multi-class classification, where inputs were categorized into multi classes. In our experiment, data sets belong to ($N = 1, 2, 3, \dots, 10$) ten different classes. The query compound either belongs to a single class or multiple classes based on input labeled. The data divided into positive and negative samples. The positive samples are all the points in class i , and let the negative sample be all the points not in class. For the prediction of pathway module classes, we used 1985 labeled compounds $L \in (0, 1, 2, \dots, 9)$. The classifier predicts the probabilistic outcome in a single class or multi-classes based on the input labeled. The compound in datasets and each class’ performance statistics are shown in Figures 3A and 3B respectively. Our model shows high performance (precision, recall, f1-score) for each class. The accuracy for multi-class classification problems as follows:

$$Accuracy = \frac{\sum_{i=1}^N \sum_{c=1}^{10} (TruePrediction)_{i,c}}{N \times 10} \times 100\% \quad (8)$$

Here N , shows the total number of compounds in datasets, c represents the ten classes of pathway modules. The accurate class prediction is 1 if the model correctly predicts the label for the i^{th} compound of the pathway module class c . The performance metrics of our algorithm compared with other machine learning algorithms, for multi-class classification illustrated in table 2.

Fig 3A. Number of compounds in each class of pathway modules

Fig 3B. Class-wise performance statistics of ETC .

4.7.1 Comparison with Other Multi-Class Ensemble Classifiers

Extra Trees Classifier shown state-of-the-art metrics performance for the prediction of pathway module classes. We evaluated our model with other ensemble and group of machine learning classifiers. our model is also compared with existing methods, which used similar datasets to predict metabolic pathway classes [8], [11]. The performance metrics of these models shown in table 2.

We compared ETC with previous researcher works and other classifiers in the current experiment on multi-class classification data. We evaluated our model to calculate accuracy, sensitivity, and precision by the formulas shown in 4.7. The performance of the ETC is higher than other methods in all metrics performance terms. Let us assume that our model prediction is given in table 3. The true positive (TP), false positive (FP), true negative (TN), and false-negative (FN) can be calculated based on actual classes and predicted classes of pathway modules.

The table shows the actual classes, assuming predicted classes, and confusion metrics of the chemical compounds. CM, EM, LM, NM, AM, GM, MCV, BTP, BOSM, and XB

TABLE 2
Comparison of the multi-class classifier for the prediction of pathway module classes

Methods	Reference	Accuracy	Precision	Recall
Random forest	Barnwal et al. [11]	97.58	83.69	83.63
AdaBoost	Hu et al [8]	94.64	77.97	67.83
Ensemble classifiers(KNN, LR, RF)	compared in this experiments	98.14	87.44	88.94
Extra trees classifier	Our Proposed Model	98.59	90.70	91.71
Random forest	compared in this experiments	97.14	87.44	88.94

TABLE 3
classifiers prediction of pathway module classes

Chemical Compounds	Actual classes	Predicted classes	True Positive	False-positive	True negative	False-negative
Glycerone	CM & EM	CM & LM	CM	LM	NM, AM, GM, MCV, BTP, BOSM, XB	EM
Phosphate	EM, GM, BTP, BOSM	EM, BOSM, MCV	EM, BOSM	MCV	NM, AM, GM, BOSM	CM, BTP
Glyceraldehyde 3-phosphate	CM	CM	CM	-	EM, LM, NM, AM, GM, MCV, BTP, BOSM, XB	-
Alpha-D-glucose	CM, BOSM	EM, LM, NM, AM, GM, MCV, BTP, XM	-	EM, LM, NM, AM, GM, MCV, BTP, XM	-	CM, BOSM

are the classes of pathway modules in KEGG, described in section 2. Table 3 presents four query compounds and their corresponding metrics values. This method is repeated for all the compounds in the datasets, and the cumulative statistics for TPs, FPs, TNs, and FNs are used to evaluate the performance.

5 DISCUSSION

In this paper, we predict the functional unit of gene sets (module) in the metabolic pathway, where compounds link with pathway modules. The compounds data retrieved from two different databases of KEGG, pathway module, and metabolic pathway databases. We distributed this problem into two classification problems, the pathway module and its classes. In machine learning, both problems, prediction of module and classes of module recommended as a single and multi-class classification problem, respectively. For both types of classification, we adopted ensemble ETC based on the compounds' SMILES molecular structure. The classifier makes the number of DTs from the training datasets and predicts using adaptive voting. We then choose several

based classifiers, including KNN, random forest, decision tree, and design an ensemble adaptive voting algorithm to improve the prediction accuracy. However, the performance of random forest and ensemble classifiers is not good as ETC. Further, we also compared ETC performance with prior published work. Resultant, our model showed better performance than others in all terms of accuracy, precision, and recall, shown in Tables 1 and 2.

Overall, the ETC classifier's implementation is easy to process the metabolomics for drug designing, synthesizing new reactions predictors, and predicting enzymes. Interaction of molecule with the functional parts of the metabolic pathways. The ETC classifier can be made to train and predict based on atom-bond specification, biological function to predict reactant pair from the available chemical compound dataset, to determine unknown reaction for the optimization, reconstruction, of metabolic pathways.

6 CONCLUSION

This article proposed a machine learning ensemble classifier ETC to predict pathway modules in the metabolic pathway based on the chemical-chemical interaction, molecular structure, and physical descriptors of chemical compounds. The experimental results proved, our ETC reached state-of-the-art performance on both module and module-classes classification problems. This article only predicts the pathway module and classes of the module where chemical compounds map. This work can be extended to predict the metabolic pathways and functional unit of genes set in pathway based on chemical reactions. Further, hybrid datasets of chemical compounds and reactions to predict the metabolic pathway and modules is another future work in metabolomics.

Funding

This work was funded by the National Key R&D Program of China (No.2019YFA0904303), the Major Projects of Technological Innovation in Hubei Province (2019AEA170), and the Frontier Projects of Wuhan for Application Foundation (2019010701011381).

REFERENCES

- [1] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 27, no. 1, pp. 29–34, 1999.
- [2] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, and M. Tanabe, "New approach for understanding genome variations in KEGG," *Nucleic acids research*, vol. 47, no. D1, pp. 590–595, 2019.
- [3] G. Pessi, J. Y. Choi, J. M. Reynolds, D. R. Voelker, and C. B. Mamoun, "Exploring metabolomic approaches to analyse phospholipid biosynthetic pathways in Plasmodium," *Proceedings of the National Academy of Sciences, USA*, vol. 101, pp. 6206–6211, 2004.
- [4] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, and L. A. Mueller, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic acids research*, vol. 44, no. D1, pp. 471–480, 2016.
- [5] S. Placzek, I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, and D. Schomburg, "BRENDA in 2017: new perspectives and new tools in BRENDA," *Nucleic acids research*, p. gkw952, 2016.
- [6] T. Lombardot, A. Morgat, K. B. Axelsen, L. Aimò, N. Hyka-Nouspikel, A. Niknejad, A. Ignatchenko, I. Xenarios, E. Coudert, and N. Redaschi, "Updates in Rhea: SPARQLing biochemical reaction data," *Nucleic acids research*, vol. 47, no. D1, pp. 596–600, 2019.
- [7] Y.-D. Cai, Z. Qian, L. Lu, K.-Y. Feng, X. Meng, B. Niu, G.-D. Zhao, and W.-C. Lu, "Prediction of compounds' biological function (metabolic pathways) based on functional group composition," *Molecular diversity*, vol. 12, no. 2, pp. 131–137, 2008.
- [8] L.-L. Hu, C. Chen, T. Huang, Y.-D. Cai, and K.-C. Chou, "Predicting biological functions of compounds based on chemical-chemical interactions," *PloS one*, vol. 6, no. 12, p. e29491, 2011.
- [9] A. Ng and K. Soo, "Random forests," pp. 117–127, 2018.
- [10] A. Macchiarulo, J. M. Thornton, and I. Nobeli, "Mapping human metabolic pathways in the small molecule chemical space," *Journal of chemical information and modeling*, vol. 49, no. 10, pp. 2272–2289, 2009.
- [11] M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, and A. O. Hero, "A deep learning architecture for metabolic pathway prediction," *Bioinformatics*, vol. 36, no. 8, pp. 2547–2553, 2020.
- [12] Y. Jia, R. Zhao, and L. Chen, "Similarity-based machine learning model for predicting the metabolic pathways of compounds," *IEEE Access*, vol. 8, pp. 130 687–130 696, 2020.
- [13] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [14] B. Baranidharan, A. Pal, and P. Muruganandam, "Cardio-Vascular Disease Prediction based on Ensemble technique enhanced using Extra Tree Classifier for Feature Selection."
- [15] M. Goetz, C. Weber, J. Bloecher, B. Stieltjes, H.-P. Meinzer, and K. Maier-Hein, "Extremely randomized trees based brain tumor segmentation," *Proceeding of BRATS challenge-MICCAI*, pp. 006–011, 2014.
- [16] M. R. C. Acosta, S. Ahmed, C. E. Garcia, and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks," *IEEE access*, vol. 8, pp. 19 921–19 933, 2020.
- [17] L. Wehenkel, D. Ernst, and P. Geurts, "Ensembles of extremely randomized trees and some generic applications," *Proceedings of robust methods for power system state estimation and load forecasting*, 2006.
- [18] A. Orlenko, D. Kofink, L.-P. Lyytikäinen, K. Nikus, P. Mishra, P. Kuukasjärvi, P. J. Karhunen, M. Kähönen, J. O. Laurikka, and T. Lehtimäki, "Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning," *Bioinformatics*, vol. 36, no. 6, pp. 1772–1778, 2020.
- [19] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement," *Information*, vol. 11, no. 6, p. 332, 2020.
- [20] R. Shafique, A. Mehmood, and G. S. Choi, "Cardiovascular Disease Prediction System Using Extra Trees Classifier," 2019.
- [21] M. Shen, Y. Xiao, A. Golbraikh, V. K. Gombur, and A. Tropsha, "An in silico screen for human S9 metabolic turnover using k-nearest neighbor QSPR method," *J. Med. Chem.*, vol. 46, pp. 3013–3020, 2003.
- [22] Z. Yao and W. L. Ruzzo, "A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data," in *BMC bioinformatics*, vol. 7, 2006, pp. 1–11.
- [23] C. R. de Magalhães, R. Carrilho, D. Schrama, M. Cerqueira, A. M. R. da Costa, and P. M. Rodrigues, "Mid-infrared spectroscopic screening of metabolic alterations in stress-exposed gilthead seabream (*Sparus aurata*)," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.
- [24] K. M. Ponnimoli and D. S. Selvamuthukumaran, "Analysis of Face Recognition using Manhattan Distance Algorithm with Image Segmentation," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 7, pp. 18–27, 2014.
- [25] J. M. Dale, L. Popescu, and P. D. Karp, "Machine learning methods for metabolic pathway prediction," *BMC bioinformatics*, vol. 11, no. 1, p. 15, 2010.
- [26] M. Cuperlovic-Culf, "Machine learning methods for analysis of metabolic data and metabolic pathway modeling," *Metabolites*, vol. 8, no. 1, p. 4, 2018.
- [27] D. Murrugarra, A. Veliz-Cuba, B. Aguilar, S. Arat, and R. Laubacher, "Modeling stochasticity and variability in gene regulatory networks," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2012, no. 1, p. 5, 2012.

- [28] A. Shaikh, N. Mahoto, F. Khuhawar, and M. Memon, "Performance evaluation of classification methods for heart disease dataset," *Sindh University Research Journal-SURJ (Science Series)*, vol. 47, no. 3, 2015.
- [29] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.
- [30] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

Hayat Ali Shah received the M.S. degree in Computer Science from Virtual University of Pakistan, Lahore, in 2018. He is currently pursuing Ph.D. degree in Software Engineering with Wuhan University, China. His research interest include bioinformatics, Multiple sequence alignment tools, Simulated alignment, and machine learning methods in prediction and reconstruction of metabolic pathway.

Juan Liu Juan Liu received her B.S. and Ph.D. degrees in computer science from Wuhan University, Wuhan in 1991 and 1996 respectively. She is now a professor in School of Computer Science at Wuhan University, Wuhan. Her research interests include machine learning, data mining, bioinformatics, pattern recognition, and artificial intelligence methods for medicine.

Zihui Yang is a Ph.D. candidate in the School of Computer Science, Wuhan University, Wuhan. His current research interests include synthetic biology, machine learning, metabolic pathway reconstruction and metabolic flux analysis.

Xiaolei Zhang received the Bachelor's degree in Computer Science from Northeast Normal University, in 2020. She is currently pursuing M.S. degree in Computer Science with Wuhan University, China. Her research interest includes bioinformatics, metabolic pathway optimization, bioretrosynthesis and deep learning.

Jing Feng received his Ph.D. degree in software engineering from Wuhan University, Wuhan, in 2007. His current research interests include bioinformatics, pattern recognition, data mining, machine learning, and artificial intelligence methods in medicine.