Multi-label Classification with XGBoost for Metabolic Pathway Prediction

Hyunwhan Joe[1] and Hong-Gee Kim[1,2*]

[1]Biomedical Knowledge Engineering Lab., Seoul National University, Seoul, Republic of Korea.
[2*]School of Dentistry and Dental Research Institute, Seoul National University, Seoul, Republic of Korea.

*Corresponding author(s). E-mail(s): hgkim@snu.ac.kr;
Contributing authors: hyunwhanjoe@snu.ac.kr;

Abstract

Background: Metabolic pathway prediction is one possible approach to address the problem in system biology of reconstructing an organism's metabolic network from its genome sequence. Recently there have been developments in machine learning-based pathway prediction methods that conclude that machine learning-based approaches are similar in performance to the most used rule-based method, PathoLogic. One issue is that previous studies evaluated PathoLogic without taxonomic pruning which significantly improves its performance.
Results: In this study, we update the evaluation results from previous studies to demonstrate that PathoLogic with taxonomic pruning outperforms previous machine learning-based approaches and that further improvements in performance need to be made for them to be competitive. Furthermore, we introduce mlXGPR, a XGBoost-based metabolic pathway prediction method based on the multi-label classification pathway prediction framework introduced from mlLGPR and evaluate it on single-organism and multi-organism benchmarks. Our results indicate that mlXGPR outperform other previous pathway prediction methods including PathoLogic with taxonomic pruning in terms of hamming loss, precision and F1 score on single organism benchmarks.
Conclusions: The results from our study indicate that the performance of machine learning-based pathway prediction methods can be substantially improved and can even outperform PathoLogic with taxonomic pruning.

Keywords: Metabolic Pathway Prediction, BioCyc, XGBoost

1 Introduction

A fundamental prerequisite in comprehending an organism's metabolism is the realization of an encompassing model of the metabolic interactions that occur in the organism [1]. An example of such a model is a Pathway/Genome Database (PGDB) that describes an organism's genes, proteins and metabolic and regulatory networks [2]. Initially, PGDBs were constructed through literature-based manual curation but this approach was not scalable [3]. This led to hybrid approaches where PGDBs are initially generated then refined through manual curation afterwards [4].

The PGDB creation workflow used by Pathway Tools [2], a software environment that is used to create and manage PGDBs, consists of two main steps with additional post-processing steps afterwards which can be seen in **Figure** 1. The first step is the PGDB generation step where the schema, replicons, genes and proteins of a PGDB are generated from an organism's annotated genome. The next step is the pathway prediction step which is divided into two sub-steps. The first sub-step performs reactome inference where the set of enzyme-catalyzed metabolic reactions occurring in an organism are predicted. The second sub-step is pathway inference where, based on the predicted reactome, the pathways occurring in the organism are predicted. Only metabolic pathways are predicted instead of other types of biochemical pathways such as signaling pathways. Metabolic pathway prediction in the literature commonly refers to either predicting the metabolic pathways that a molecule is associated with [5][6][7] or the metabolic pathways occurring in an organism based on its annotated genome [1][8][9]. This work will focus on the latter and assumes that the reactome is already inferred and provided. Lastly, pathway prediction can also be differentiated into predicting pathways from a reference database and predicting unobserved novel pathways (pathway discovery) [1] and this work focuses on the former.

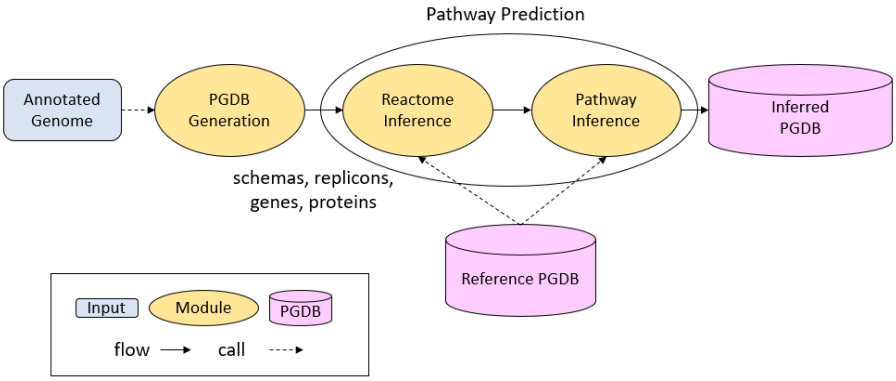

Fig. 1 Workflow of PGDB Creation

PathoLogic is a pathway prediction algorithm developed by SRI International that is used by Pathway Tools. PathoLogic predicts metabolic pathways in MetaCyc [10], a curated multi-organism metabolic pathway database, from an organism's annotated genome. It assigns scores to each metabolic pathway in MetaCyc, where a higher

score reflects a higher likelihood that the pathway is present in the target organism. Afterwards, the decision to include or reject the pathway is completed through a sequence of defined rules [8]. While PathoLogic has gone through several iterations and updates to improve its accuracy, it has several limitations. One limitation is that since the rules defined are hard-coded, it makes the algorithm relatively inflexible to maintain and extend. Another limitation is that the pathway scoring system is ad-hoc and does not reflect actual mathematical probabilities.

As a response to these limitations, Dale et al. [1] introduced the first study that evaluated multiple machine learning-based metabolic pathway prediction methods. Their results demonstrated that machine learning methods were able to perform as well as PathoLogic with the best performing ML-based approach achieving a small improvement over PathoLogic. Despite the promising results from the study, PathoLogic is still used as the main engine for Pathway Tool's prediction algorithm. Recently, there has been several studies which updated the pioneer study with new datasets, features and methodologies [11][12][13].

One of the studies mlLGPR [11], made a novel contribution of modeling the prediction task as multi-label classification compared to other studies which modeled it as binary classification. Multi-label classification is where more than one class label can be predicted which differs from traditional classification where only one label is predicted [14]. Modeling the prediction task as multi-label classification allowed the training dataset used in mlLGPR to be more compact allowing for more organisms to be used for training. For example, Aljarbou et al. [12] has 4,979 instances covering 20 organisms and DeepRF [13] had 172,380 instances covering 60 organisms. mlLGPR's multi-label modeling allows for its dataset to be smaller with 15,000 instances but is able to cover 15,000 organisms. mlLGPR uses a binary relevance approach [15] where the multi-label learning process is divided into independent binary classifiers for each pathway label allowing for the possibility of parallel training. What also differentiated mlLGPR with other pathway prediction studies such as [1], [12], [13] is that for their evaluation methodology they used a completely separate evaluation dataset which they did not use for training and hyperparameter tuning. Another novel contribution from the mlLGPR study was that it was the first machine learning-based pathway prediction method to be evaluated also on multi-organism genomes such as symbionts and microbiomes. The evaluation results from mlLGPR were also similar to other studies on single-organism genomes showing similar performance to PathoLogic.

A limitation of previous machine learning-based metabolic pathway prediction methods was that the feature engineering task involving designing and testing features was a time consuming task. As a response to this limitation, representational learning approaches [16] such as pathway2vec [9] and triUMPF [17] were introduced to generate features to be used for prediction. While the research direction and results from the two studies are promising, they shared similar problems with mlLGPR in their evaluation methodology for single organism genomes. The common issue is that PathoLogic is evaluated without using taxonomic pruning. MetaCyc pathways can be assigned a taxonomic range for which they can occur and PathoLogic utilizes these ranges when deciding on whether to include or reject a pathway. Taxonomic pruning was introduced to improve the performance of PathoLogic by removing false positives

[8]. While the mlLGPR study acknowledges that PathoLogic was evaluated without using taxonomic pruning for the single-organism benchmark it does not give the reason for not applying it when it improves performance. This is an issue because evaluating PathoLogic without using taxonomic pruning for single organism genomes can lead to potentially lower results which can be misleading as a benchmark.

In this study, we provide two contributions to the problem of metabolic pathway prediction from annotated genomes. The first contribution is that we evaluate PathoLogic with taxonomic pruning on the single organism prediction benchmark to provide a more accurate pathway prediction benchmark. Our results show that PathoLogic with taxonomic pruning showed a significant increase on the four evaluation metrics for the majority of the single organism datasets. In addition, we observed that the evaluation datasets introduced in the mlLGPR study shares characteristics of tabular datasets with its mixed feature data types. Recent studies have shown that tree ensemble models such as XGBoost tend to outperform deep learning prediction models when applied to tabular datasets [18][19]. With these observations, for our second contribution we introduce a XGBoost-based pathway prediction method termed mlXGPR based on the multi-label classification prediction framework introduced by mlLGPR and evaluated it on single organism and multi-organism benchmark datasets. mlXGPR on the single organism benchmark outperformed the other prediction methods including PathoLogic with taxonomic pruning for three of the evaluation metrics hamming loss, precision and F1 score.

2 Methods

The workflow for mlXGPR is similar to the multi-label classification for metabolic pathway prediction workflow introduced in the mlLGPR study. The first step is the feature engineering step which takes the training and evaluation datasets and transforms them into feature vectors. The mlLGPR study introduced five different feature groups which are enzymatic reaction abundance (AB), reaction evidence (RE), pathway evidence (PE), pathway commons (PC) and possible pathways (PP) where AB is the main feature group that can be combined with other feature groups. After the training dataset is transformed into feature vectors, we use k-fold cross validation and grid search to tune the hyperparameters of our prediction model. Once the hypermeters are chosen for the final prediction model, the whole training dataset is then used for training the model. The trained model is then evaluated on the benchmark datasets and then can be to deployed to predict new datasets. One difference between mlLGPR and mlXGPR is that mlXGPR uses XGBoost as the prediction model instead of logistic regression as used in mlLGPR. Another difference is that mlLGPR does not use cross-validation for hyperparameter tuning but used one split to tune its hyperparameters. The workflow for mlXGPR can be seen in **Figure** 2.

2.1 Definitions and problem formulation

In this study, we will use the conventions introduced in the mlLGPR study [11]. All vectors are column vectors which are denoted by boldface lowercase letters (e.g. $\mathbf{x}$). A subscript character to a vector, $\mathbf{x}_i$, denotes the i-th element of $\mathbf{x}$ while a superscript,

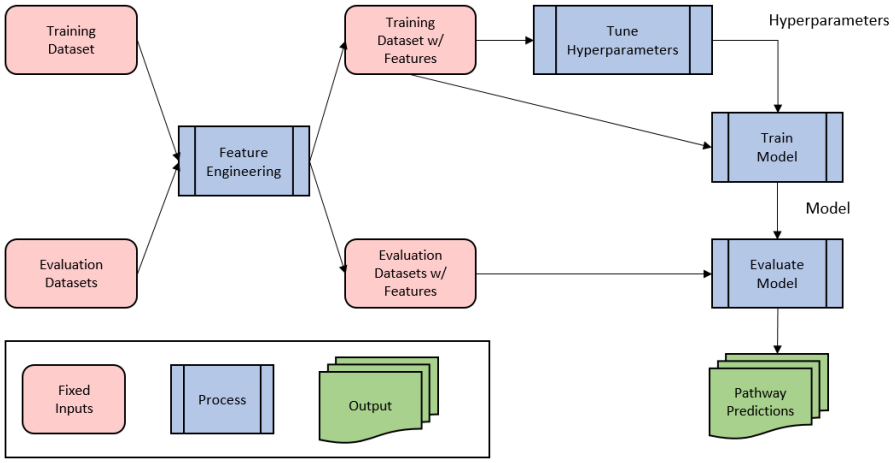

Fig. 2 mlXGPR Workflow

$\mathbf{x}^{(i)}$, denotes an index to a sample. In addition, calligraphic letters (e.g. $\mathcal{S}$) are used to represent sets and $|.|$ will be used to denote set cardinality.

A multi-label pathway dataset consisting of n samples can be defined as $\mathcal{S} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) : 1 < i \leqslant n\}$. $\mathbf{x}_i$ is a vector that corresponds to the abundance of each enzymatic reaction e, which is an element of the set $\mathcal{E} = \{e_1, e_2, ..., e_r\}$, having r possible reactions. The abundance of an enzymatic reaction $e_l^{(i)}$, for a sample i can be defined as $a_l^{(i)} \in \mathbb{R}_{\geq 0}$. The class labels $\mathbf{y}^{(i)} = [y_1^{(i)}, ..., y_t^{(i)}] \in \{0, 1\}^t$ is a vector of size t. Its elements corresponds to pathway labels derived from a reference pathway database $\mathcal{Y}$. A sample for the multi-label pathway dataset used can be seen in **Table** 1

$\mathcal{X} = \mathbb{R}^r$ is defined as the r-dimensional input space. Each sample $\mathbf{x}^{(i)} \in \mathcal{X}$ is transformed into an m-dimensional vector by a transformation function $\Phi : \mathcal{X} \to \mathbb{R}^m$. The transformation function is obtained from the Feature engineering process (see Section Features engineering). In summary, the metabolic pathway prediction task can be defined as given a multi-label dataset $\mathcal{S}$, learn a hypothesis function $f : \Phi(\mathbf{x}^{(i)}) \to 2^{|\mathcal{Y}|}$, such that it can classify metabolic pathway labels accurately for an unseen sample $\mathbf{x}^*$.

Table 1 Sample of Multi-label Pathway Dataset.
The number of pathways is independent from the number of enzymatic reactions.

Input Enzymatic Reaction Abundances				
EC-1	EC-1.1	...	EC-6.6.1.1	EC-6.6.1.2
2	9	...	1	0
Output Presence of Pathways				
VALSYN	ARG-PRO	...	PWY-7081	PW-721
1	0	...	1	0

2.2 Feature engineering

Five types of feature vectors were designed and introduced in the mlLGPR study [11]. Each feature vector is created through 5 transformation sub-processes i)- enzymatic reactions abundance (ϕ^{ab}), ii)- reactions evidence (ϕ^{re}), iii)- pathways evidence (ϕ^{pe}), iv)- pathway common (ϕ^{pc}) and v)- possible pathways (ϕ^{pp}). The enzymatic reaction abundance transformation maps to a r-dimensional vector that denotes the total occurrence of each enzymatic reaction in an organism. Each enzymatic reaction is identified by its Enzyme Commission (EC) number [20]. The reaction evidence transformation maps to a vector that represents the properties of the enzymatic reactions for each sample. The pathway evidence transformation maps to a vector whose features expands on core PathoLogic rules to also include enzyme presence, pathway gaps, network connectivity and etc. The possible pathway transformation maps to a vector which holds for each pathway two representations. The first is a boolean representation, whether each pathway is present or not, from enzymatic reaction information, and is decided by a user-defined threshold. The second is a numeric representation which represents the probabilities for each pathway whether they are present or not based off enzymatic reaction information. Each transformation maps a sample to a different vector which are concatenated into a m-dimensional feature vector $\Phi(\mathbf{x}^{(i)}) = [\phi^{ab}(\mathbf{x}^{(i)}), \phi^{re}(\mathbf{x}^{(i)}), \phi^{pe}(\mathbf{x}^{(i)}), \phi^{pc}(\mathbf{x}^{(i)}), \phi^{pp}(\mathbf{x}^{(i)})]$. The number of features for each feature group can be seen in **Table** 2.

Table 2 Number of features for each feature group

Feature Group	Number of Features
Enzymatic Reaction Abundance	3650
Reaction Evidence	68
Pathway Evidence	32
Pathway Commons	3650
Possible Pathways	5052

2.3 Prediction Model and Multi-label Learning Process

XGBoost is a machine learning algorithm that utilizes gradient boosted decision trees [21] where each tree is trained to predict the pseudo-residuals of the previous tree based on a pre-defined objective function [22]. One of the key factors in XGBoost's success and popularity is innovations in scalability such as optimizations in handling sparse data, weighted quantile sketch calculations and parallel/distributed computing [23]. Recently, XGBoost version 1.6 started to provide native support for multi-label classification which allows for the efficient training of classifiers on many class labels. Before this addition, studies used outside libraries such as scikit-multilearn [24],[25],[26] or sklearn.MultiOutputClassifier for multi-label classification [27]. The multi-label learning process using the binary relevance approach breaks down into t independent binary classification tasks, where each task corresponds to one class label.

2.4 Experimental Setup

In this section, we describe the experimental setup to evaluate mlXGPR's pathway prediction performance across multiple datasets including single and multi-organisms. The single organism golden dataset consists of six Tier 1 PGDBs from BioCyc which are EcoCyc(v21) [28], HumanCyc(v19.5) [4], AraCyc(v18.5) [29], YeastCyc(v19.5), LeishCyc(v19.5) [30] and TrypanoCyc(v18.5) [31] which were used in previous benchmarks [11],[9],[17]. BioCyc is a PGDB Web portal that contains thousands of PGDBs and divides PGDBs into tiers based on the manual curation involved [32]. Tier 1 is the highest quality PGDB in BioCyc and the requirement is at least one person year worth of literture-based curation. LeishCyc and TrypanoCyc are currently Tier 2 but the versions used during the mlLGPR study were Tier 1 at the time. Basic statistical information for each PGDB can be seen in **Table** 3. For the multi-organism benchmark dataset we used the Critical Assessment of Metagenome Interpretation (CAMI) initiative low complexity dataset [33] used in the triUMPF study [17].

Table 3 Dataset Statistics

Dataset	Instances	Enzymatic Reactions	Pathways
EcoCyc	1	719	307
HumanCyc	1	693	279
AraCyc	1	1034	510
YeastCyc	1	544	229
LeishCyc	1	292	87
TrypanoCyc	1	512	175
CAMI	40	1083	674
Synset 2	15000	3650	2526

mlXGPR's performance was compared to three representative pathway prediction methods. The results from PathoLogic version 21 without taxonomic pruning and set to the default settings in the mlLGPR study will be included in our evaluation. We also evaluated PathoLogic version 22 with taxonomic pruning and default settings to showcase the improvements in performance with taxonomic pruning. Version 22 was used instead of version 21 because it is not available to be downloaded anymore. One difference between the two versions is that PathoLogic version 22 predicts pathways from MetaCyc v22 which removed 7 pathways from MetaCyc v21. MinPath is another well known pathway prediction method that uses integer programming to predict the minimum set of pathways [34]. We did not include MinPath in our evaluation because it had too many false positives leading to low precision as can be seen in the mlLGPR study. For the representative machine learning-based pathway prediction methods we included both results from the mlLGPR and triUMPF [17] studies. The models from Aljarbou et al. [12] and DeepRF [13] were not used in the evaluation because both models are binary classifiers instead of multi-label and are trained using different datasets making it difficult to accurately compare. In addition, from the best of our knowledge the datasets and source code used in both studies are not open source which make comparing their performances even more difficult. For the performance metrics,

we used the Hamming loss [35], precision, recall and F1 score to match the metrics used in the previous studies.

For training, we used the corrupted synthetic dataset Synset 2 that was constructed and used for training in the mlLGPR study. Synset 2 was constructed from MetaCyc version 21 and contains 2526 metabolic pathways and 3650 enzymatic reactions including incomplete ones such as EC 1.2.3-. The dataset was generated by randomly selecting pathways for each synthetic sample based on the Poisson distribution with mean value equal to 500. The corruption process is done by randomly retaining/inserting/removing enzymatic reactions from each selected pathway based on earlier defined constraints. The dataset was corrupted to reflect errors that could occur from upstream data analysis on experimental data. Synset 2 consists of 15,000 synthetic samples as can be seen in **Table** 3. An abalation test on the five feature groups (AB, RE, PE, PP and PC) was done with Synset2 and a combination of +AB+RE+PE feature groups yielded the highest prediction performance with +AB+RE performing the second highest. mlXGPR uses only the +AB+RE feature groups because the PE feature group uses different features for each pathway label and XGBoost does not natively support this type of multi-label classification. XGBoost only supports multi-label classification where the features are the same throughout each label.

For evaluation, mlXGPR uses 6-fold cross validation grid search to determine the optimal hyperparameters for the max depth and number of estimators. We used the Scikit-Learn API for XGBoost and the options for the max depth was {2,4,6,8} and {16,17,18,19} for the number of estimators. The options for the number of estimators was chosen by pre-testing with early stopping. The final model was trained using all of Synset 2 with max depth set to 4 and the number of estimators set to 19 based on the highest average F1 score from grid search. In addition we also used 'hist' for the tree method because it was fastest among the other options and all the options had similar results. The 'hist' option is an approximate tree method similar to the method used in LightGBM [36] which is another well known gradient boosting decision tree method. All tests were conducted on an Ubuntu 20.04.5 server with dual Intel Xeon CPU E5-2640 v4. Lastly, Python 3.9 and XGBoost 1.7 were used to obtain the experimental results.

3 Results

Table 4 shows the pathway prediction performance results for mlXGPR and the four other methods. In terms of the Hamming loss, precision and F1 score, mlXGPR outperformed the other methods. mlXGPR particularly performed well in terms of precision. PathoLogic with taxonomic pruning had the highest recall on EcoCyc, AraCyc, HumanCyc and YeastCyc while mlLGPR had the highest recall on LeishCyc and TrypanoCyc. PathoLogic with taxonomic pruning outperformed PathoLogic without taxonomic pruning in all metrics on all the PGDBs except with TrypanoCyc where pruning only improved the precision. Since PathoLogic version 22 is a different version to the one used in the mlLGPR study, we also tested PathoLogic version 22 without taxonomic pruning and saw similar improvements in performance with taxonomic pruning.

Table 4 Performance of each prediction algorithm on six single organism T1 PGDBs. ↓ indicates that a lower score is better while for ↑ a higher score is better. The best performing method is bold for each metric.

Metrics & Methods	*EcoCyc*	*HumanCyc*	*AraCyc*	*YeastCyc*	*LeishCyc*	*TrypanoCyc*
Hamming loss (↓)						
PathoLogic	0.0610	0.0633	0.1188	0.0424	0.0368	0.0424
PathoLogic+Pruning	0.0372	0.0424	0.0649	0.0257	0.0234	0.0530
mlLGPR	0.0804	0.0633	0.1069	0.0550	0.0380	0.0590
triUMPF	0.0435	0.0954	0.1560	0.0649	0.0443	0.0776
mlXGPR	**0.0170**	**0.0218**	**0.0451**	**0.0174**	**0.0071**	**0.0127**
Precision (↑)						
PathoLogic	0.7230	0.6695	0.7011	0.7194	0.4803	0.5480
PathoLogic+Pruning	0.8105	0.7688	0.8502	0.8106	0.6667	0.6589
mlLGPR	0.6187	0.6686	0.7372	0.6480	0.4731	0.5455
triUMPF	0.8662	0.6080	0.7377	0.7273	0.4161	0.4561
mlXGPR	**0.9889**	**0.9746**	**0.9692**	**0.9744**	**0.9600**	**0.9497**
Recall (↑)						
PathoLogic	0.8078	0.8423	0.7176	0.8734	0.8391	0.7829
PathoLogic+Pruning	**0.9055**	**0.8817**	**0.8235**	**0.9345**	0.6437	0.4857
mlLGPR	0.8827	0.8459	0.7314	0.8603	**0.9080**	**0.8914**
triUMPF	0.7590	0.3835	0.3529	0.3319	0.7126	0.6229
mlXGPR	0.8697	0.8244	0.8020	0.8297	0.8276	0.8629
F1 Score (↑)						
PathoLogic	0.7631	0.7460	0.7093	0.7890	0.6109	0.6447
PathoLogic+Pruning	0.8554	0.8214	0.8367	0.8682	0.6550	0.5592
mlLGPR	0.7275	0.7468	0.7343	0.7392	0.6220	0.6768
mlXGPR	**0.9255**	**0.8932**	**0.8777**	**0.8962**	**0.8889**	**0.9042**

We also evaluated mlXGPR's performance on complex multi-organism genomes such as the CAMI low complexity dataset. MetaPathways v2.5 [37] was used to create the benchmark CAMI environment PGDB (ePGDB) which are PGDBs for microbial communities [38]. MetaPathways utilizes a modified version of PathoLogic for pathway prediction. mlXGPR was compared with two other pathway prediction methods mlLGPR and triUMPF and the results can be seen in **Table** 5. PathoLogic was not included in the comparison since MetaPathways uses it to create the ePGDB. The results for mlLGPR and triUMPF were taken from the triUMPF study. triUMPF achieved the lowest Hamming loss 0.0436 and the highest sample average F1 score 0.5864. mlLGPR had the highest sample average recall 0.7827 but lowest sample average precision 0.357 in comparison. mlXGPR was the opposite with the highest sample average precision 0.8074 but the lowest sample average recall 0.2485 which also contributed to it having the lowest sample average F1 score 0.3789 among the three methods. The results for mlXGPR are consistent with the single-organism benchmark in its high precision but lower recall. One limitation of the CAMI ePGDB as a benchmark is that it is automatically generated using MetaPathways but the predictions have not been curated so it can be said that the results demonstrate more how similar the other prediction methods are with MetaPathways and PathoLogic than their actual prediction performance. This also is a possible explanation for triUMPF's higher performance since it was trained on mostly Tier 3 BioCyc PGDBs which are generated

from PathoLogic and have no curation. Currently there is still a lack of highly curated ePGDBs that can be used for multi-organism pathway prediction benchmarks.

Table 5 Performance of mlLGPR, triUMPF and mlXGPR on the multi-organism community dataset CAMI. ↓ indicates that a lower score is better while for ↑ a higher score is better. The best performing method is bold for each metric. The sample average is calculated for the average precision, recall and F1 score.

Metrics	*mlLGPR*	*triUMPF*	*mlXGPR*
Hamming loss (↓)	0.0975	**0.0436**	0.0496
Average Precision (↑)	0.3570	0.7027	**0.8074**
Average Recall (↑)	**0.7827**	0.5101	0.2485
Average F1 score (↑)	0.4866	**0.5864**	0.3789

4 Conclusions

In this study, we introduce a XGBoost-based metabolic pathway prediction method called mlXGPR that uses a binary relevance approach where the multi-label classifier is decomposed into independent binary classifiers for each pathway label. mlXGPR is based on mlLGPR which introduced an approach that modeled the metabolic pathway inference problem as a multi-label classification problem. mlXGPR was motivated by previous pathway prediction studies in that they were not compared properly with PathoLogic using taxonomic pruning. This is why we attempted to apply XGBoost, a SOTA supervised learning method for tabular data to the problem of multi-label pathway prediction. We trained a XGBoost prediction model with tuned hyper-parameters and compared its performance with three representative metabolic pathway prediction methods on single organism and multi-organism genome benchmark datasets. The results was that mlXGPR outperformed the other methods on three of the four evaluation metrics which are Hamming loss, precision and F1 score for single-organism datasets.

While we were able to improve the performance of machine learning-based pathway prediction to outperform PathoLogic using taxonomic pruning, mlXGPR still shares the common issue with mlLGPR in that its performance is reliant on feature information that is manually curated. This is why representational learning-based pathway prediction approaches are promising but currently their performance still need improvement. Since the binary relevance approach that mlXGPR uses supposes that the class labels are independent, one future direction is evaluating alternative multi-classification approaches which utilizes the correlations between class labels such as in classifier-chains [15]. Another potential direction for future studies in machine learning-based pathway prediction is if the datasets and source code from other studies such as [12] and DeepRF [13] become open, their methodologies can be evaluated on the mlLGPR benchmark datasets. The reverse can also be done with evaluating the methodologies used in mlLGPR, triUMPF and mlXGPR on the different datasets

used in these studies. This would allow for a more comprehensive evaluation of the performance of machine learning-based pathway prediction models.

Acknowledgments. Hyunwhan Joe would like to thank members of the Biomedical Knowledge Engineering Lab. for the discussions and helpful comments that they gave.

Declarations

- Funding: Not applicable.
- Conflict of interest/Competing interests: The authors declare that they have no competing interests.
- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and materials: The data used in the study is available at https://github.com/hyunwhanjoe/mlXGPR/
- Code availability: The source code used in the study is available at https://github.com/hyunwhanjoe/mlXGPR/
- Authors' contributions: HGK and HJ provided the initial ideas. HJ preprocessed the data and designed the methods. HJ wrote the code and conducted the experiments. HGK and HJ wrote the manuscript. All authors read and approved the final manuscript.

References

[1] Dale, J.M., Popescu, L., Karp, P.D.: Machine learning methods for metabolic pathway prediction. BMC Bioinformatics **11**(1), 15 (2010) https://doi.org/10.1186/1471-2105-11-15

[2] Karp, P.D., Paley, S.M., Midford, P.E., Krummenacker, M., Billington, R., Kothari, A., Ong, W.K., Subhraveti, P., Keseler, I.M., Caspi, R.: Pathway Tools version 24.0: Integrated Software for Pathway/Genome Informatics and Systems Biology (2015) https://doi.org/10.48550/ARXIV.1510.03964

[3] Karp, P.D.: The EcoCyc Database. Nucleic Acids Research **30**(1), 56–58 (2002) https://doi.org/10.1093/nar/30.1.56

[4] Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., Karp, P.D.: Computational prediction of human metabolic pathways from the complete human genome. Genome Biology **6**(1), 2 (2004) https://doi.org/10.1186/gb-2004-6-1-r2

[5] Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., Kanehisa, M.: PathPred: An enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Research **38**(Web Server), 138–143 (2010)

[6] Baranwal, M., Magner, A., Elvati, P., Saldinger, J., Violi, A., Hero, A.O.: A deep learning architecture for metabolic pathway prediction. Bioinformatics **36**(8), 2547–2553 (2020) https://doi.org/10.1093/bioinformatics/btz954

[7] Jia, Y., Zhao, R., Chen, L.: Similarity-Based Machine Learning Model for Predicting the Metabolic Pathways of Compounds. IEEE Access **8**, 130687–130696 (2020)

[8] Karp, P.D., Latendresse, M., Caspi, R.: The Pathway Tools Pathway Prediction Algorithm. Standards in Genomic Sciences **5**(3), 424–429 (2011) https://doi.org/10.4056/sigs.1794338

[9] M A Basher, A.R., Hallam, S.J.: Leveraging heterogeneous network embedding for metabolic pathway prediction. Bioinformatics **37**(6), 822–829 (2021) https://doi.org/10.1093/bioinformatics/btaa906

[10] Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D.: The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Research **48**(D1), 445–453 (2020) https://doi.org/10.1093/nar/gkz862

[11] M. A. Basher, A.R., McLaughlin, R.J., Hallam, S.J.: Metabolic pathway inference using multi-label classification with rich pathway features. PLOS Computational Biology **16**(10), 1008174 (2020) https://doi.org/10.1371/journal.pcbi.1008174

[12] Aljarbou, Y.S., Haron, F.: Determining the Presence of Metabolic Pathways using Machine Learning Approach. International Journal of Advanced Computer Science and Applications **11**(8) (2020)

[13] Shah, H.A., Liu, J., Yang, Z., Zhang, X., Feng, J.: DeepRF: A deep learning method for predicting metabolic pathways in organisms based on annotated genomes. Computers in Biology and Medicine **147**, 105756 (2022)

[14] Tsoumakas, G., Katakis, I.: Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining **3**(3), 1–13 (2007)

[15] Zhang, M.-L., Li, Y.-K., Liu, X.-Y., Geng, X.: Binary relevance for multi-label learning: An overview. Frontiers of Computer Science **12**(2), 191–202 (2018)

[16] Bengio, Y., Courville, A., Vincent, P.: Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013)

[17] Basher, A.R.M.A., Mclaughlin, R.J., Hallam, S.J.: Metabolic Pathway Prediction Using Non-Negative Matrix Factorization with Improved Precision. Journal of Computational Biology **28**(11), 1075–1103 (2021) https://doi.org/10.1089/cmb.2021.0258

[18] Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information Fusion **81**, 84–90 (2022) https://doi.org/10.1016/j.inffus.2021.11.011

[19] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems **35**, 507–520 (2022)

[20] Bairoch, A.: The ENZYME database in 2000. Nucleic Acids Research **28**(1), 304–305 (2000)

[21] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5) (2001)

[22] Sagi, O., Rokach, L.: Approximating XGBoost with an interpretable decision tree. Information Sciences **572**, 522–542 (2021) https://doi.org/10.1016/j.ins.2021.05.055

[23] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[24] Szymanski, P., Kajdanowicz, T.: Scikit-multilearn: A scikit-based python environment for performing multi-label classification. J. Mach. Learn. Res. **20**(1), 209–230 (2019)

[25] Chen, S., Xiao, L.: Predicting and characterising persuasion strategies in misinformation content over social media based on the multi-label classification approach. Journal of Information Science, 016555152311699 (2023) https://doi.org/10.1177/01655515231169949

[26] Zhang, J., Wang, Z., Wang, H.-Y., Chung, C.-R., Horng, J.-T., Lu, J.-J., Lee, T.-Y.: Rapid Antibiotic Resistance Serial Prediction in Staphylococcus aureus Based on Large-Scale MALDI-TOF Data by Applying XGBoost in Multi-Label Learning. Frontiers in Microbiology **13**, 853775 (2022) https://doi.org/10.3389/fmicb.2022.853775

[27] Piter, C.A., Hadi, S., Yulita, I.N.: Multi-label classification for scientific conference activities information text using extreme gradient boost (xgboost) method. In: 2021 International Conference on Artificial Intelligence and Big Data Analytics, pp. 1–5 (2021). IEEE

[28] Keseler, I.M., Gama-Castro, S., Mackie, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Kothari, A., Krummenacker, M., Midford, P.E., Muñiz-Rascado, L., Ong, W.K., Paley, S., Santos-Zavaleta, A., Subhraveti, P., Tierrafría, V.H., Wolfe, A.J., Collado-Vides, J., Paulsen, I.T., Karp, P.D.: The EcoCyc Database in 2021. Frontiers in Microbiology **12**, 711077 (2021)

[29] Mueller, L.A., Zhang, P., Rhee, S.Y.: AraCyc: A Biochemical Pathway Database for Arabidopsis. Plant Physiology **132**(2), 453–460 (2003)

[30] Doyle, M.A., MacRae, J.I., De Souza, D.P., Saunders, E.C., McConville, M.J., Likić, V.A.: LeishCyc: A biochemical pathways database for Leishmania major. BMC Systems Biology **3**(1), 57 (2009)

[31] Shameer, S., Logan-Klumpler, F.J., Vinson, F., Cottret, L., Merlet, B., Achcar, F., Boshart, M., Berriman, M., Breitling, R., Bringaud, F., Bütikofer, P., Cattanach, A.M., Bannerman-Chukualim, B., Creek, D.J., Crouch, K., De Koning, H.P., Denise, H., Ebikeme, C., Fairlamb, A.H., Ferguson, M.A.J., Ginger, M.L., Hertz-Fowler, C., Kerkhoven, E.J., Mäser, P., Michels, P.A.M., Nayak, A., Nes, D.W., Nolan, D.P., Olsen, C., Silva-Franco, F., Smith, T.K., Taylor, M.C., Tielens, A.G.M., Urbaniak, M.D., van Hellemond, J.J., Vincent, I.M., Wilkinson, S.R., Wyllie, S., Opperdoes, F.R., Barrett, M.P., Jourdan, F.: TrypanoCyc: A community-led biochemical pathways database for Trypanosoma brucei. Nucleic Acids Research **43**(D1), 637–644 (2015)

[32] Karp, P.D., Billington, R., Caspi, R., Fulcher, C.A., Latendresse, M., Kothari, A., Keseler, I.M., Krummenacker, M., Midford, P.E., Ong, Q., Ong, W.K., Paley, S.M., Subhraveti, P.: The BioCyc collection of microbial genomes and metabolic pathways. Briefings in Bioinformatics **20**(4), 1085–1093 (2019)

[33] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M., Hansen, L.H., Sørensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Don Kang, D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G.G.Z., Cuevas, D.A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H.-P., Göker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A.E., Rattei, T., McHardy, A.C.: Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nature Methods **14**(11), 1063–1071 (2017) https://doi.org/10.1038/nmeth.4458

[34] Ye, Y., Doak, T.G.: A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. PLoS Computational Biology **5**(8), 1000465 (2009)

[35] Wu, X.-Z., Zhou, Z.-H.: A unified view of multi-label performance measures. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3780–3788 (2017)

[36] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.:

Lightgbm: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3149–3157 (2017)

[37] Konwar, K.M., Hanson, N.W., Bhatia, M.P., Kim, D., Wu, S.-J., Hahn, A.S., Morgan-Lang, C., Cheung, H.K., Hallam, S.J.: MetaPathways v2.5: Quantitative functional, taxonomic and usability improvements. Bioinformatics **31**(20), 3345–3347 (2015) https://doi.org/10.1093/bioinformatics/btv361

[38] Konwar, K.M., Hanson, N.W., Pagé, A.P., Hallam, S.J.: MetaPathways: A modular pipeline for constructing pathway/genome databases from environmental sequence information. BMC Bioinformatics **14**(1), 202 (2013) https://doi.org/10.1186/1471-2105-14-202