Journal of Cheminformatics

METHODOLOGY

Open Access

Learning motif features and topological structure of molecules for metabolic pathway prediction

Jianguo Hu[1†], Yiqing Zhang[1†], Jinxin Xie[1], Zhen Yuan[1], Zhangxiang Yin[1], Shanshan Shi[1], Honglin Li[1,2,3*] and Shiliang Li[1,2*]

Abstract

Metabolites serve as crucial biomarkers for assessing disease progression and understanding underlying pathogenic mechanisms. However, when the metabolic pathway category of metabolites is unknown, researchers face challenges in conducting metabolomic analyses. Due to the complexity of wet laboratory experimentation for pathway identification, there is a growing demand for predictive methods. Various computational approaches, including machine learning and graph neural networks, have been proposed; however, interpretability remains a challenge. We have developed a neural network framework called MotifMol3D, which is designed for predicting molecular metabolic pathway categories. This framework introduces motif information to mine local features of small-sample molecules, combining with graph neural network and 3D information to complete the prediction task. Using a dataset of 5,698 molecules that participate in 11 metabolic pathway categories in the KEGG database, MotifMol3D outperformed state-of-the-art methods in precision, recall, and F1 score. In addition, ablation study and motif analysis have demonstrated the effectiveness and usefulness of the model. Motif analysis, in particular, has shown motif information can actually characterize the main features of specific pathway molecules to a certain extent and enhance the interpretability of the model. An external validation further corroborates this observation. MotifMol3D is an open-source tool that is available at https://github.com/Irena-Zhang/MotifMol3D.git.

Scientific contribution MotifMol3D integrates motif information, graph neural networks, and 3D structural data to enhance feature extraction for small-sample molecules, improving the precision and interpretability of metabolic pathway predictions. The model outperforms state-of-the-art approaches in precision, recall, and F1 score. This work reveals how motif information characterizes pathway-specific molecules, offering novel insights into molecular properties within metabolic pathways.

† Jianguo Hu and Yiqing Zhang have contributed equally to this work.

*Correspondence:
Honglin Li
hlli@ecust.edu.cn
Shiliang Li
slli403@163.com
[1] Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China
[2] Innovation Center for AI and Drug Discovery, School of Pharmacy, East China Normal University, Shanghai 200062, China
[3] Lingang Laboratory, Shanghai 200031, China

Introduction

Metabolites play a crucial role in cellular metabolism, serving as vital biomarkers for assessing disease progression and understanding the underlying pathogenic mechanisms [1]. They regulate enzyme activity to maintain physiological homeostasis, and are essential for human functionality [2]. In order to gain deeper insights into disease mechanisms, metabolomics has emerged as a widely embraced methodology. Through experimental design, sample collection, LC–MS analysis, and subsequent statistical and pathway analyses, metabolomics reveals the biological pathways in which metabolites participate [3]. In the enrichment analysis of differentially expressed metabolites, it is crucial to map them to known biological pathways. If differentially expressed metabolites are not in known biological pathways, they are often ignored. Therefore, there is a growing demand for methods capable of predicting the potential involvement of unknown metabolites in specific biological pathways.

To delve deeper into biological metabolism, several metabolic pathway databases have been developed for the representation, qualitative analysis, and visualization of biological pathways. These databases include the Kyoto Encyclopedia of Genes (KEGG) database [4], MetaCyc database [5], and The Small Molecule Pathway Database (SMPDB) [6]. In the KEGG database [4], the metabolic network has been categorized into eleven groups, such as Carbohydrate Metabolism, Glycan, Xenobiotics, etc., based on molecular interactions, reactions, and relational networks. However, due to the complexity of biological systems, there are still many undiscovered cryptic biological pathways and latent enzymes or molecular compounds within existing pathways. Additionally, it is important to note that a single molecule may participate in multiple biological pathways, while different molecules may also participate in the same pathway [7]. Therefore, in order to predict the metabolic pathways of molecules accurately, it is crucial to understand the role of the chemical structure and physical properties of molecules in different pathways.

Currently, various approaches have been developed to address metabolic pathway-related problems, encompassing topology-based [8, 9], genome information-based [10, 11], machine learning-based [12–16], and deep learning-based methods [1, 17, 18]. These diverse methods are designed to handle tasks such as predicting metabolic pathways, reconstructing metabolic pathways, and predicting missing enzymes in these pathways. For identifying the metabolic pathway categories of a given molecule based on interaction information, Hu et al. proposed a multi-target model utilizing chemical-chemical interactions [15]. Building upon this, Gao et al. further integrated compound-protein interaction and protein–protein interaction to identify pathway categories of proteins [14]. However, due to limited molecular interaction information, there is a need for more general methods. TrackSM is a tool that predicts the pathway categories of unknown small molecules by matching molecular structures [8]. Jia et al. developed a similarity-based Random Forest (RF) model to identify the metabolic pathways to which molecules belong [16].

Recently, graph neural networks have gained attention in the field of pharmacy and show promising applications in molecular structure and drug discovery [19]. Baranwal et al. constructed a multi-class model based on graph convolutional networks to directly predict the metabolic pathway class of molecules [1]. Yang et al. trained a graph attention network to extract local features of molecules for metabolic pathway prediction [18]. Du et al. improved the predictive performance of the model by incorporating additional information on interdependent metabolic pathways [17]. Liu et al. developed a multi-label learning method using attention mechanisms for pathway inference [20]. Bao et al. used transfer learning with a Graph Transformer and CNN for plant metabolic pathway prediction [21]. Another study by Liu et al. introduced a multi-scale neural network with a graph enhancement strategy for better pathway prediction [22]. Although current methods have achieved a relatively good predictive performance, they still lack interpretability in their predictions, which prevents further analysis of the unique characteristics of molecules in different pathways.

In this study, we developed a prediction strategy based on a graph neural network, namely MotifMol3D, for effectively discerning the categories of molecular metabolic pathways. The MotifMol3D framework integrates motif information, graph neural networks, and 3D information to extract local features. It captures molecular characteristics of specific pathways from small molecular samples, enhancing the model's interpretability. The development of this model contributes to researchers' preliminary analysis of molecular mechanisms underlying unknown metabolic pathways.

Methods

A hybrid framework for multi-label classification

We proposed a hybrid framework for metabolic pathway prediction (Fig. 1), combining motif features and a graph attention network (GAT) [23]. The framework includes two feature extraction blocks and a fully connected (FC) layer. In the two feature extraction blocks, one block is responsible for extracting molecular features from 2 and

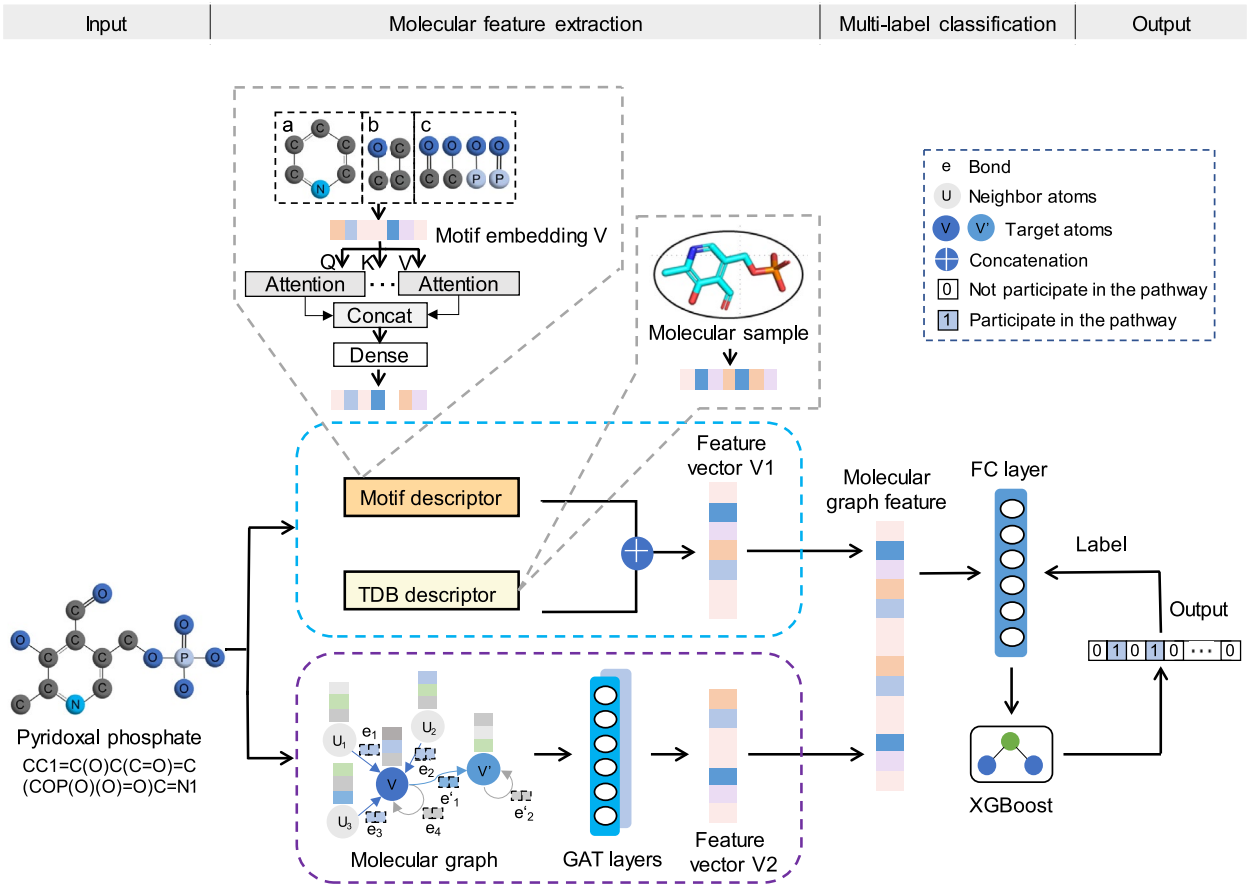

Fig. 1 The structure of MotifMol3D framework for metabolic pathway prediction. The framework comprises four sections: (1) The model takes the SMILES representation of small molecules as input; (2) In the molecular feature extraction module, a heterogeneous motif graph network containing motifs and molecular nodes is established to learn motif features. Additionally, global features of molecules are obtained by introducing a graph attention network; (3) In the multi-label classification module, the features from the output layer are merged and fed into a fully connected (FC) layer, which is then trained with labels from the training set. Finally, XGBoost is employed to predict the molecular pathway categories; (4) The model's ultimate output indicates participation (1) or non-participation (0) in each of the 11 metabolic pathway categories

3D levels, and the other block introduces the graph attention network and combines the bond and node information of the graph to extract the overall features of the molecule. The molecular characteristics derived from the two blocks are concatenated and input into the feedforward network layer, facilitating the extraction of graph features for the categorization of output molecules' pathways. Finally, the graph features are input to XGBoost for the final molecular pathway category prediction.

Feature vector V1

The feature vector V1 consists of the motif descriptor, TDB descriptor, and seven molecular property descriptors. Motifs are functional substructures within chemical compounds that can be identified using SMILES strings. These motifs represent specific arrangements of atoms and bonds that are significant for the compound's properties and behavior [24]. Topological Distance Based 3d Descriptors (TDB descriptors), also known as 3D autoregressive descriptors, provide important 3D structural information by considering the relationship between topology and spatial distance of molecules. These descriptors were generated using seven atomic properties: mass, van der Waals volume, Sanderson electronegativity, polarizability, first ionization potential, I-state, and covalent radius [25]. To generate the TDB descriptor for a compound, its three-dimensional structure was sampled by the Cyndi software [26] and then input into the PaDEL software to calculate the descriptors [27]. Additionally, seven molecular property descriptors related to molar refractivity, rotational bond, aromaticity, and lipophilicity were generated by the RDKit software [28]. The formula for calculating TDB descriptors is:

$$S(d) = \frac{1}{k(d)} \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_i \cdot D_{ij} \cdot X_j \right]_{T_{ij=d}} \quad (1)$$

where n refers to the number of atoms in the molecule, and D_{ij} and T_{ij} are the Euclidean distance and topological distance between atoms i and j, respectively. X_i and X_j represent the corresponding atomic properties, while $k(d)$ indicates the number of atom pairs within the given topological distance d. Seven molecular property descriptors related to refractive index, rotational bonds, aromaticity, and lipophilicity were generated by RDKit. The TDB descriptors and molecular property descriptors were concatenated to generate a 14-dimensional feature vector.

Yu et al. proposed a heterogeneous motif graph network model for molecular graph representation and verified that the relationship at the motif level may contribute to the feature representation of molecular graphs to a certain extent [29]. Inspired by this model, we thought that the motif information of a molecule may characterize the molecule in a small sample dataset to some extent. We divided the compound into bonds and rings, but with a different approach than the previous model. We distinguished the bond directly attached to the aromatic ring from the same bond in the side chain of the aromatic ring. For example, when cleaving pyridoxal phosphate, we obtained three parts: a, b, and c in Fig. 1. Aromatic pyridine is in box a, the carbon–carbon bond and carbon–oxygen bond directly connected to the aromatic ring are in box b, and the carbon–oxygen bond, phosphorus-oxygen bond, and phosphorus-oxygen double bond on the side chain of the aromatic ring are in box c. We employed the term frequency inverse document frequency [30] (*TF-IDF*) value as an index to sort and screen the motifs of compounds. The formula for calculating the *TF-IDF* value corresponding to the motifs of compounds is

$$TF - IDF = C(i)_j \left(log \frac{1+M}{1+N(i)} + 1 \right) \quad (2)$$

where $C(i)_j$ is the number of times that the motif i appears in the molecule j, M is the number of molecules, and $N(i)$ is the number of molecules containing motif i. The TF-IDF values of motifs were calculated based on their frequency within molecules and their occurrence across all molecules. By ranking the TF-IDF values and utilizing grid search, the optimal number of motifs was determined based on multiple evaluation metrics, including accuracy, precision, recall, and F1 score. Ultimately, the top seven motifs were selected to characterize each compound. These motifs were then converted into word embeddings using the Transformer encoder block. The multi-head self-attention mechanism was employed to capture the interactions between the word embeddings. The calculation formula for the multi-head self-attention mechanism is:

$$Head(q,k,v) = W_c \underset{i\in[1,I]}{||} Attention^i \left(W_i^q q, W_i^k k, W_i^v v \right) \quad (3)$$

where $Attention(q,k,v) = softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)v$.

The number of attention heads (I) was set to 4. The word embedding V, represented by $q=k=v$, was processed through a Position-wise Feed-Forward network [31] within each attention head. W represents the learnable weight parameters. Layer normalization was then applied to generate seven feature vectors $V'(v'1, v'2, \ldots, v'7)$. Finally, V' was linearly transformed to produce a 10-dimensional feature vector V. The specific calculation formula for this conversion process is:

$$V_o = W_o \underset{l\in[1,l]}{||} V' \quad (4)$$

The symbol l is the sequence length (the number of motifs), where this study set $l=7$. Motif descriptor, TDB descriptor, and molecular property descriptor are concatenated to generate the 24-dimensional feature vector V1. V1 incorporates information on both the two-dimensional and three-dimensional structure of the compound, enabling the identification of the compound's metabolic pathway.

Feature vector V2

The graph attention network [23] utilized an attention mechanism to learn the local environment of nodes, enabling better feature extraction for compounds. This attention mechanism also provides interpretability when applied to chemical compounds. In a network framework proposed for predicting the synthetic accessibility of organic compounds, a graph attention network was employed [32]. This framework incorporates both point representation and bond information to enhance the feature information of molecules. To generate suitable inputs for the graph-level neural network, the nodes and edges of the molecular graph were processed, resulting in 10 types of atomic features and 4 types of bond features. Unlike the "one-hot" encoding method used in the reference model, the point and edge features of molecular graphs were randomly transformed into word embeddings. In the subsequent two hidden layers, the output feature of

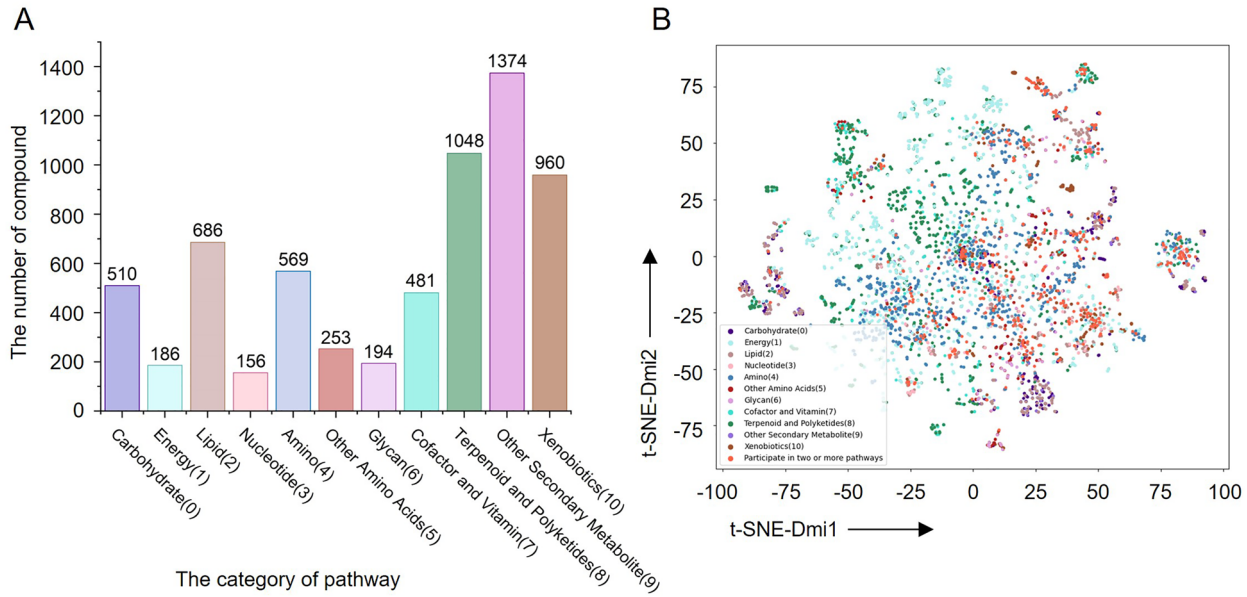

Fig. 2 **A** Distribution of molecules in each metabolic pathway. The data in parentheses correspond to the number of metabolic pathways. **B** Dimensionality reduction visualization of the Morgan fingerprints of small molecules in the metabolic pathway dataset shows overlap in the coverage of chemical space for the pathways

the target point in the first hidden layer was updated by aggregating neighbor information by weighted aggregation. The bond information in the second hidden layer was then combined with the updated node features to enrich the structural information. The calculation formula for this merged information propagation method [32] is

$$h_{(v)}^{(l+1)} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{u\in N_v}\alpha_{vu}^{(k)}W^{(l)}h_{(u)}^{(l)}\right) \tag{5}$$

where σ is the activation function elu [33], v is the target node, u is the first-order neighbor of the target node, K denotes the number of attention heads, W^l represents the learnable weight matrix at level l, and α is the normalized attention coefficient between node v and node u, h^l means the updated node features. The updated nodes then go through two graph attention layers and one graph readout layer to obtain the graph-level feature vector V2, see [32] for more details.

Output layer

The graph feature vector was generated by concatenating feature vector V1 and feature vector V2. This combined vector was then fed into a fully connected layer with a SoftMax activation function to classify the pathway categories of molecules. The output of the fully connected layer was an 11-dimensional feature vector. In the model, a threshold of 0.5 was used to determine whether a compound belongs to a specific pathway.

XGBoost optimization

Referring to XGraphBoost [34], we used the graph feature vector obtained from model training as input for XGBoost. This helped address data label imbalance by using a voting mechanism with multiple weak classifiers. After obtaining graph features using a feed-forward neural network, XGBoost was trained to predict the metabolic pathway category of molecules. The parameters for XGBoost were a maximum tree depth of 60, 30 decision trees, and default values for other parameters.

Results and discussion

Metabolic pathway dataset

In the KEGG database [4], the manually curated metabolic pathway maps related to metabolism are divided into 11 categories: Carbohydrate metabolism (0), Energy metabolism (1), Lipid metabolism (2), Nucleotide metabolism (3), Amino acid metabolism (4), Metabolism of other amino acids (5), Glycan biosynthesis and metabolism (6), Metabolism of cofactors and vitamins (7), Metabolism of terpenoids and polyketides (8), Biosynthesis of other secondary metabolites (9), Xenobiotics biodegradation and metabolism (10). A total of 5764 compounds that participate in metabolic pathways were collected from the KEGG database (March 2022). After

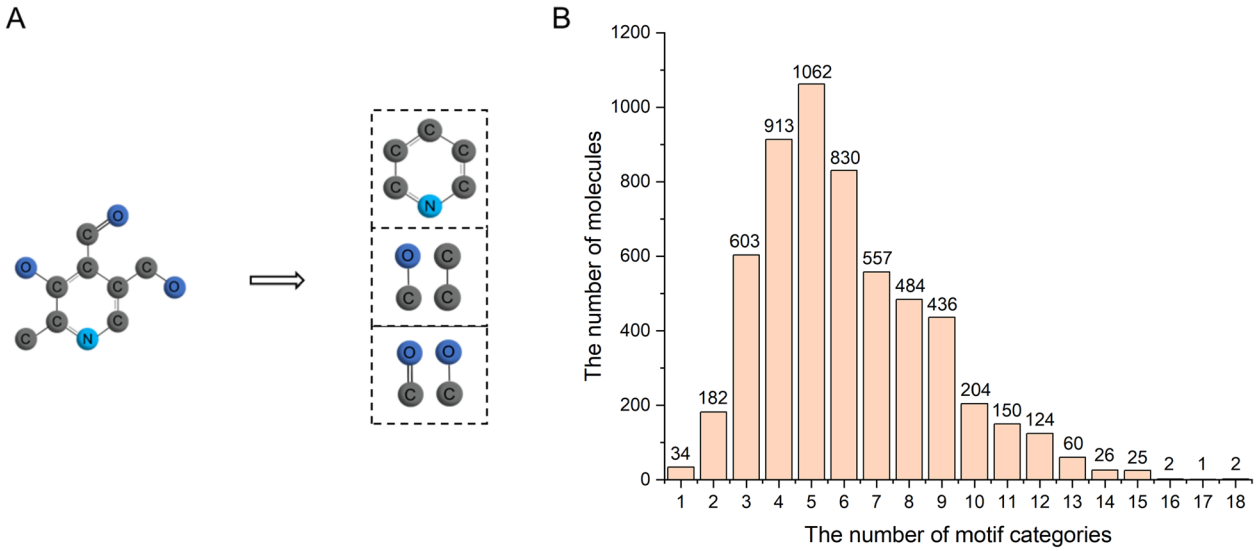

Fig. 3 Construction (**A**) and distribution (**B**) of the motif dictionary

excluding macromolecular substances (proteins, glycans, wax esters), compounds with free radicals, and polymers, finally our dataset contains 5698 molecules. The statistical calculation of small molecule distribution in each pathway (Fig. 2A) revealed that the Biosynthesis of other secondary metabolites (9) pathway contained the largest number of compounds (1374 compounds), while the Nucleotide metabolism (3) pathway harbored the smallest number of compounds (156 compounds). We used the t-SNE method to perform dimensionality reduction visualization of the Morgan fingerprints (radius = 2, nbits = 1024) of small molecules in the dataset (Fig. 2B). The results indicate that the 11 different metabolic pathways do not exhibit a clear clustering trend, suggesting that utilizing the complete structural information from small molecules for pathway clustering may not be optimal. Therefore, in this study, we avoided examining associations between molecules and pathways and instead shifted our focus to elucidating the correlations between specific molecular features and pathways. For model building, the molecules in the dataset were randomly divided into training set (80%), validation set (10%), and test set (10%).

Setting the parameters of motifs

To establish a motif dictionary for molecules, we fragmented them into motifs based on predefined rules. The *TF-IDF* values of specific motifs can be calculated using formula 1, which reflects their uniqueness in the molecule (Fig. 3A). Most molecules (approximately 85.7% or 4885 molecules) contain 3–9 types of motifs. Some molecules have a single motif type (34 molecules), while a few have 16 or more types of motifs. The molecule with the highest number of motif types contains 18 different motifs (Fig. 3B).

After sorting the motifs based on their *TF-IDF* values, we conducted a grid search (motif numbers: 4, 5, 6, 7, 8, 9) to determine the optimal number of motifs for representing each molecule. The model's performance was evaluated using accuracy, precision, recall, and F1 score (Fig. 4). The accuracy remained relatively stable as the number of motifs increased, reaching its highest value at a motif number of 7. The recall and precision exhibited opposite trends with increasing motif numbers. The recall initially decreased and then increased, reaching its lowest value at a motif number of 6. On the other hand, precision initially increased and then decreased, reaching its peak at a motif number of 7. The F1 score followed a similar pattern, peaking at a motif number of 7. Considering precision, recall, and F1 score, the optimal motif number was determined to be 7.

Model performance evaluation

We compared the performance of MotifMol3D with two traditional machine learning models (RF, XGBoost), and four state-of-the-art deeping learning models (GCN-based, GAT-based, MLGL-MP, and Baranwal et al. (2020)). RF and XGBoost were implemented following Du et al. [17]. GCN-based [1] and GAT-based [18] models incorporated local and global features, whereas MLGL-MP [17] utilized GCN and GAT to learn molecular feature representations and pathway interrelationships; meanwhile, Baranwal et al. (2020) [1] further enhances prediction accuracy by combining GCN

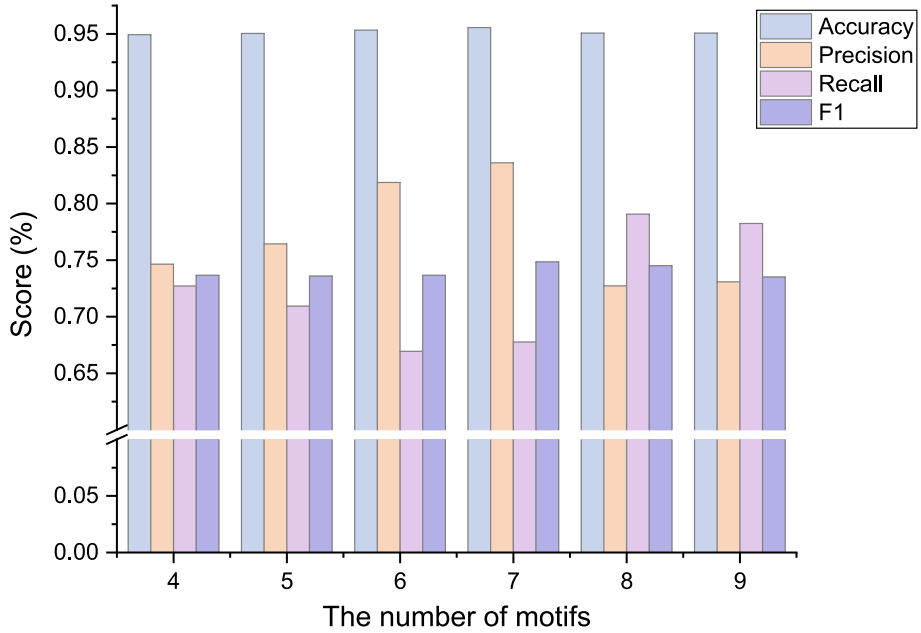

Fig. 4 Model performance under different motif numbers

Table 1 Comparison results of different methods on the metabolic pathway dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1_score (%)
RF	95.70 ± 0.23	73.32 ± 1.26	72.45 ± 1.67	72.88 ± 0.61
XGBoost	95.41 ± 0.29	74.05 ± 1.54	73.3 ± 1.61	73.67 ± 1.52
GCN_based	95.07 ± 0.32	78.22 ± 1.12	70.66 ± 0.72	74.25 ± 0.89
GAT_based	95.25 ± 0.33	79.32 ± 1.09	70.78 ± 0.56	74.81 ± 0.72
MLGL_MP	95.97 ± 0.51	81.37 ± 2.20	77.12 ± 1.82	79.19 ± 1.69
Baranwal et al. (2020)	**96.08 ± 0.21**	82.37 ± 1.08	78.23 ± 1.16	80.01 ± 1.00
MotifMol3D*	95.42 ± 0.42	82.60 ± 1.23	69.67 ± 1.25	75.59 ± 1.02
MotifMol3D	95.72 ± 0.41	**82.86 ± 0.96**	**79.62 ± 0.87**	**81.21 ± 0.53**

MotifMol3D*: MotifMol3D* is the framework of MotifMol3D without XGBoost. Default parameters from the original papers were used for reproduction. The best results are highlighted in bold

with RF. MotifMol3D* is the framework of MotifMol3D without XGBoost. Default parameters from the original papers were used for reproduction.

All methods underwent tenfold cross-validation. MotifMol3D was implemented on Centos Linux using Python 3.7.11 and PyTorch 1.8.1, running on a GPU server with 5 NVIDIA Tesla V100-PCLe-32 GB.

The comparison results of different methods are listed in Table 1. Compared with the classic traditional machine learning methods, MotifMol3D* was superior to RF and XGBoost in precision and F1 indicators. RF and XGBoost were more balanced in precision and recall, while MotifMol3D* had a large difference in precision and recall. Precision rate and recall rate were a pair of contradictory indicators. MotifMol3D* has achieved better performance in precision rate, but poor performance in recall rate. This may be attributed to the ensemble learning nature of XGBoost, where multiple classifiers contribute to a balanced trade-off between accuracy and recall. In the case of class imbalance, models tend to favor the prediction of the majority class, as it dominates the sample distribution, allowing the model to achieve higher accuracy by focusing on the majority class. XGBoost improves the prediction accuracy of the minority class by aggregating multiple weak classifiers and progressively correcting errors made by individual trees. In contrast, deep learning models may converge to local optima during training, leading to a larger discrepancy between accuracy and recall. MotifMol3D, enhanced with XGBoost, effectively combines the strengths of

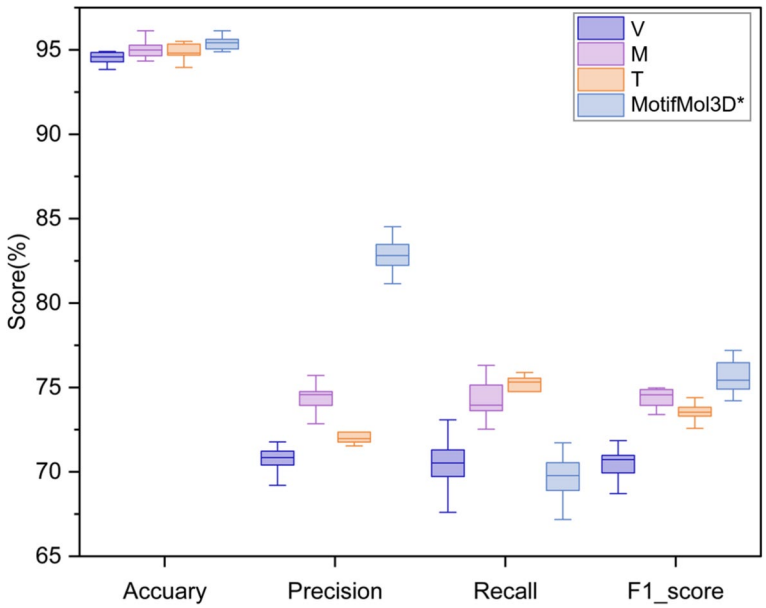

Fig. 5 Ablation study. V is the variant removing the feature vector V1, M is the variant removing the motif descriptors, T is the variant removing the TDB descriptors

both XGBoost and deep learning, resulting in superior performance in precision, recall, and F1 score compared to other deep learning methods. Baranwal et al. (2020), however, performs best in terms of accuracy. Overall, MotifMol3D* was an available graph feature extraction method, and the model performance was better after being combined with XGBoost.

Ablation study

In this section, we evaluated the contribution of each component of MotifMol3D* to the model performance through an ablation study (Fig. 5). Three variants of MotifMol3D* were defined, with the first variant removing the feature vector V1 (denoted as V), the second variant removing the motif descriptors (denoted as M), and the third variant removing the TDB descriptors (denoted as T).

In addition to recall, MotifMol3D* significantly outperformed V in accuracy, precision, and F1 score. Compared to V, MotifMol3D* improved accuracy by 0.96%, precision by 16.94%, and F1 score by 7.2%, while recall decreased by 1.05%. Feature vector V1 included both 2D and 3D information of molecules, indicating that additional dimensional information could improve the predicted metabolic pathways to some extent.

Furthermore, the inclusion of either TDB descriptors (T) or motif information (M) resulted in an improvement in the model's predictive performance. Compared with V, T increased accuracy by 0.42%, precision by 2.03%, recall by 6.43%, and F1 score by 4.20%; M increased accuracy by 0.53%, precision by 5.48%, recall by 5.46%, and F1 score by 5.43%. When comparing T and M, M increased accuracy by 0.11%, precision by 3.39%, decreased recall by 0.93%, and increased F1 score by 1.21%. T and M provided TDB descriptors and molecular motif information, respectively, both of which to some degree improved the model's predictive capability.

Compared with M and T, MotifMol3D* showed improvement in accuracy, precision, and F1 score, except for a decrease in recall. Overall, the simultaneous incorporation of TDB descriptors and motif information improved MotifMol3D*'s predictive performance in metabolic pathways.

Motif analysis

Motifs can be statistically evaluated for the uniqueness of a particular molecule. After selecting the motif number parameter as 7, we investigated the relationship and biological significance between molecular motifs and pathway categories. We calculated the cumulative *TF-IDF* values of motifs for molecules in each pathway category and sorted them in descending order. Table 2 displays the top 7 motifs for each of the 11 biological metabolic pathway categories.

Energy metabolism plays a crucial role in maintaining normal metabolic enzyme activity, which is vital for the growth, development, and reproduction of organisms. It is well-established that dysregulation of energy metabolism is closely associated with various diseases, including

Table 2 The top 7 motifs within each pathway

Pathway	Smiles						
Carbohydrate	'OP'	'CO'	'O=P'	'C1CCO CC1'	'CC'	'C=O'	'CN'
Energy	'OP'	'O=S'	'CS'	'CN'	'OS'	'[N+]=O'	'SS'
Lipid	'C1CC CCC1'	CC'	'OP'	'C1CCC C1'	'C=C'	'[2*]c'	'CN'
Nucleotide	'OP'	'c1cnc nc1'	'c1c[nH] cn1'	'CN'	'O=P'	'C1CC OC1'	'cN'
Amino	'CN'	'OP'	'c1cccc c1'	'cO'	'cC'	'CC'	'C=O'
Other amino	'CN'	'C[Se]'	'OP'	'CP'	'[C-]# [N+]'	'CC'	'O[Se]'
Glycan	'OP'	'CN'	'O=P'	'C1CCO CC1'	'c1cn cnc1'	'CC'	'CO'
Cofactor/ vitamin	'cC'	'OP'	'c1ccnc c1'	'CC'	'CN'	'c1cn cnc1'	'cO'
Terpenoid/PK	'C=C'	'C1CC CCC1'	'CC'	'OP'	'cO'	'c1cc ccc1'	'C1=CC CCC1'
Other secondary metabolite	'c1cc ccc1'	'cO'	'c1ccoc c1'	'CN'	'cC'	'C1CC NCC1'	'c1c [nH]cc1'
Xenobiotics	'c1cc ccc1'	'cCl'	'CCl'	'cO'	'cC'	'cN'	'CN'

obesity [35], type 2 diabetes [36], and cancer [37]. In addition to their roles in DNA and RNA synthesis, energy transfer and storage, signal transduction, and enzyme regulation, nucleotides also serve as essential components of the coenzymes NAD + and ATP, contributing to a wide range of cellular responses [38]. In energy metabolism pathways, the degree of substrate phosphorylation level is closely related to the amount of energy provided. As one of the sub-pathways of the energy metabolism pathway, the oxidative phosphorylation pathway is an efficient source of energy for maintaining growth in many organisms, and its pathway components are sensitive to specific chemical inhibitors [39]. In addition, the unique role of sulfur in organisms is mainly related to redox reactions, and its functions include cell protection and energy metabolism [40]. Therefore, 'OP' and 'OS' have higher weight in energy metabolism. In the nucleic acid metabolic pathway, 'c1cncnc1' and 'c1c[nH]cn1' are two of the top three motifs. These two motifs are the main structural components of pyrimidine and purine, respectively, and the base composed of pyrimidine or purine determines the type and function of nucleotides. Amino acids are the basic structural units of proteins, and their metabolism is closely related to various physiological and pathological conditions. Abnormal amino acid metabolism is associated with various types of cancer, and targeting amino acid metabolism has become a promising strategy for cancer treatment [41]. In the category of amino acid metabolic pathways, the motifs 'CN' and 'C=O' are key groups that characterize the amino and carboxyl groups in amino acids (–NH2 and –COOH, respectively). These two groups regulate amino acid metabolism by adjusting the pH of the cell environment [42]. Glycans are high-molecular-weight compounds composed of multiple monosaccharides, and their metabolism mainly involves modifications of functional groups on the monosaccharides by various enzymes. Common functional groups in glycans biosynthesis and metabolism include hydroxyl (–OH), amino (–NH2), carboxyl (–COOH) and phosphate (–PO4) groups [43]. These groups can participate in various chemical reactions, such as glycosylation, phosphorylation and deamination, forming different polysaccharide structures. The motifs 'OP', 'CN', 'O=P' and 'CO' in the category of polysaccharide biosynthesis and metabolism are related to these functional groups. In general, motif information can characterize the molecular characteristics under a specific pathway to a certain extent.

External validation

KEGG database added a new pathway of pinene, camphor and geraniol degradation pathways under the terpene and polyketone metabolic pathways (map00907) and a new pathway of flavonoid degradation pathway under the other secondary metabolites biosynthesis pathway (map 00946). We collected molecular information from these two pathways and deduplicated it with reference to the training samples. Finally, 14 small molecules

Table 3 Results of the external validation

Label	Smiles	Yes(Y)/No(N)[a]
Pinene, camphor and geraniol degradation pathway		
T0	CC1(C)[C@@H]2CC[C@@]1(C)C(=O)C2	Y
T1	CC1(C)[C@@H]2CC[C@@]1(C)C(=O)C2	Y
T2	CC1(C)[C@H]2CC(=O)O[C@]1(C)CC2=O	Y
T3	CC1(C)[C@H]2CC(=O)[C@]1(C)CC2=O	Y
T4	CC1(C)[C@H]2CC(=O)[C@]1(C)C[C@H]2O	Y
T5	CC1=CC(=O)[C@H](CC(=O)SCCNC(=O)CCNC(=O)[C@H](O)C(C)(C)COP(=O)(O)OP(=O)(O)OC[C@H]2O[C@@H](n3cnc4c(N)ncnc43)[C@H](O)[C@@H]2OP(=O)(O)O)C1(C)C	Y
T6	CC1=CC(=O)[C@H](CC(=O)O)C1(C)C	N
T7	CC12C(=O)CC(CC1=O)C2(C)C	Y
T8	CC1=CC(=O)O[C@H](CC(=O)SCCNC(=O)CCNC(=O)[C@H](O)C(C)(C)COP(=O)(O)OP(=O)(O)OC[C@H]2O[C@@H](n3cnc4c(N)ncnc43)[C@H](O)[C@@H]2OP(=O)(O)O)C1(C)C	Y
T9	CC1(C)C2CC(=O)[C@]1(C)C(O)C2	Y
T10	CC1C(=O)C[C@@H](CC(=O)O)C1(C)C	N
T11	CC1(C)[C@@H]2CC(=O)[C@@]1(C)CC2=O	Y
T12	CC1(C)[C@@H]2CC(=O)O[C@@]1(C)CC2=O	Y
T13	CC1(C)[C@@H]2CC(=O)[C@@]1(C)CC2O	Y
Flavonoid degradation pathway		
O0	O=C(O)Cc1ccc(O)cc1	N
O1	O=C(O)Cc1ccc(O)c(O)c1	N
O2	O=C(O)CCc1ccc(O)cc1	N
O3	Oc1cc(O)cc(O)c1	N
O4	O=C(O)CCc1ccc(O)c(O)c1	N
O5	O=c1c(c2ccc(O)cc2)coc2c([C@@H]3O[C@H](CO)[C@@H](O)[C@H](O)[C@H]3O)c(O)ccc12	Y
O6	CC(=O)/C=C/c1ccc(O)cc1	Y
O7	Oc1ccc([C@H]2COc3cc(O)ccc3C2)cc1	Y
O8	O=C1c2c(O)cc(O)cc2OCC1c1ccc(O)cc1	Y
O9	O=C1c2c(O)cc(O)cc2OC1(O)Cc1ccc(O)c(O)c1	N
O10	O=C1c2c(O)cc(O)cc2OC1(O)Cc1ccc(O)cc1	N
O11	O=C1c2ccc(O)cc2OC[C@H]1c1ccc(O)cc1	Y
O12	O=C1c2ccc(O)cc2OC[C@@H]1c1ccc(O)cc1	Y
O13	Oc1ccc([C@H]2COc3cc(O)ccc3[C@@H]2O)cc1	Y
O14	CC(C(=O)c1ccc(O)cc1O)c1ccc(O)cc1	N
O15	CC(C(=O)c1c(O)cc(O)cc1O)c1ccc(O)cc1	N
O16	O=C(CCc1ccc(O)c(O)c1)c1c(O)cc(O)cc1O	Y
O17	O=C1CC(c2ccc(O)cc2)Oc2c(O)c(O)cc(O)c21	Y
O18	O=C(O)/C(O)=C\C(O)=C1\C(=O)CC(c2ccc(O)cc2)OC1=O	N
O19	O=C1CC(=O)OC(c2ccc(O)cc2)C1	N
O20	O=C(O)CC(=O)CC(O)c1ccc(O)cc1	N
O21	CC(=O)CC(O)c1ccc(O)cc1	N

[a] Yes(Y)/No(N) means that the metabolic pathway of the small molecule is correctly/incorrectly predicted

involved in the terpenoid and ketone metabolic pathway and 22 small molecules involved in the other secondary metabolic biosynthetic pathways were obtained. These small molecules are fed into the model to predict the metabolic pathway outcomes in which they may be involved (Table 3).

In the degradation pathway of pinene, camphor and geraniol, the model accurately predicted the molecular pathways for 12 out of 14 molecules. By looking at the

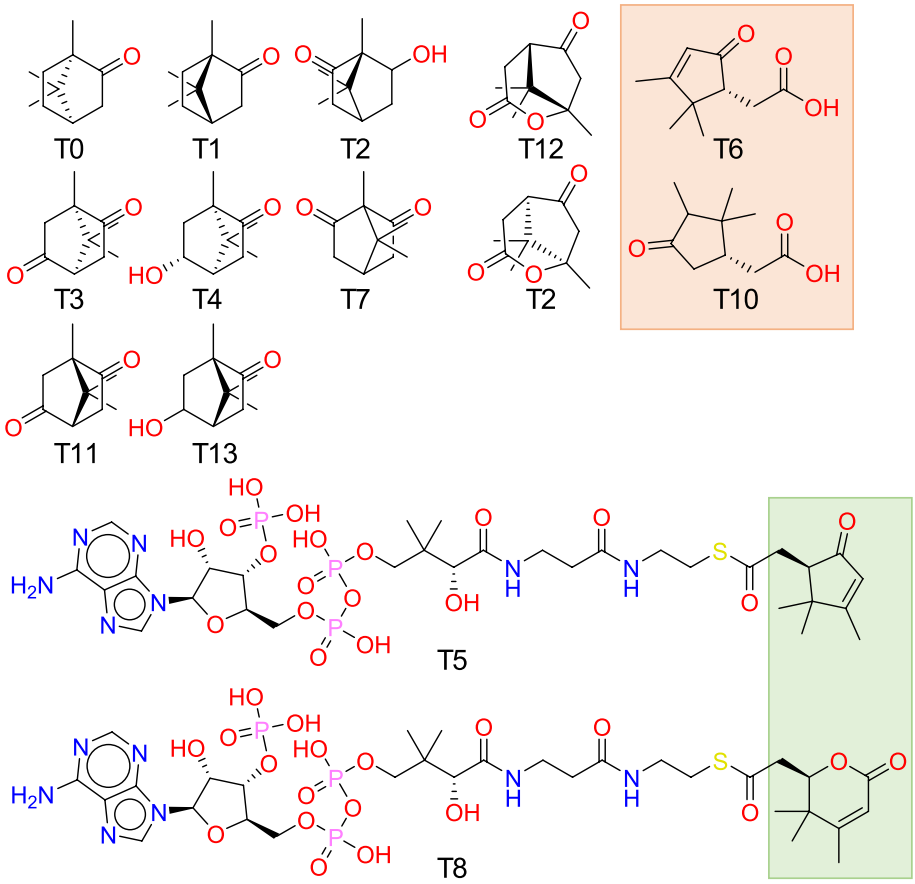

Fig. 6 Compounds in the degradation pathway of pinene, camphor and geraniol. The orange box marks the molecules that failed to predict, and the two groups in the green box are the main differences between the two compounds numbered T5 and T8

structure of these molecules (Fig. 6), we found that eight compounds, numbered T0, T1, T3, T4, T7, T9, T11, and 13, have the same camphor skeleton, distinguished by the presence or absence of carbonyl and hydroxyl groups on the ring, and the relative positions between the carbonyl and hydroxyl groups. The two compounds, numbered T2 and T12, have the same chemical structure, but their chiral conformation is different. The two compounds, numbered T5 and T8, differ only in the chemical groups on the right (see green box in Fig. 6), both are acetyl-CoA on the left. Compounds T6 and T10 (see orange box in Fig. 6) were predicted incorrectly. They were compounds obtained by ring-opening cleavage of camphor compounds, and their structures were relatively simple and lacked the characteristics of original terpenoids, which might be the reason for their failure in prediction by the model.

In the flavonoid degradation pathway, the metabolic pathway categories of 9 molecules were successfully and accurately predicted by the model. The precursor flavonoid molecules of the flavonoid degradation pathway are derived from isoflavone biosynthesis (map00943), flavone and flavonol biosynthesis (map00944), flavonoid biosynthesis (map00941) and puerarine (Fig. 7). It can be seen from the figure that most of the molecules successfully predicted have the basic skeleton of flavonoids (see grey box in Fig. 7), while the molecules that have not been successfully predicted are degraded by the flavonoid molecules after subsequent enzyme reactions, and their molecular structures are significantly different from those of the molecules involved in other secondary metabolites in the biosynthetic pathways where model training is focused. This may be the cause of the failure of model predictions. Similarly, by analyzing the prediction results of each pathway, we found that successfully predicted molecules have higher fragment similarity, while failed predictions have lower similarity to successful molecules or higher similarity to molecules from other pathways (Additional file 2: Table S1). In a given pathway, more frequent similar motif combinations in training data lead to higher prediction success, while novel structures have lower success rates.

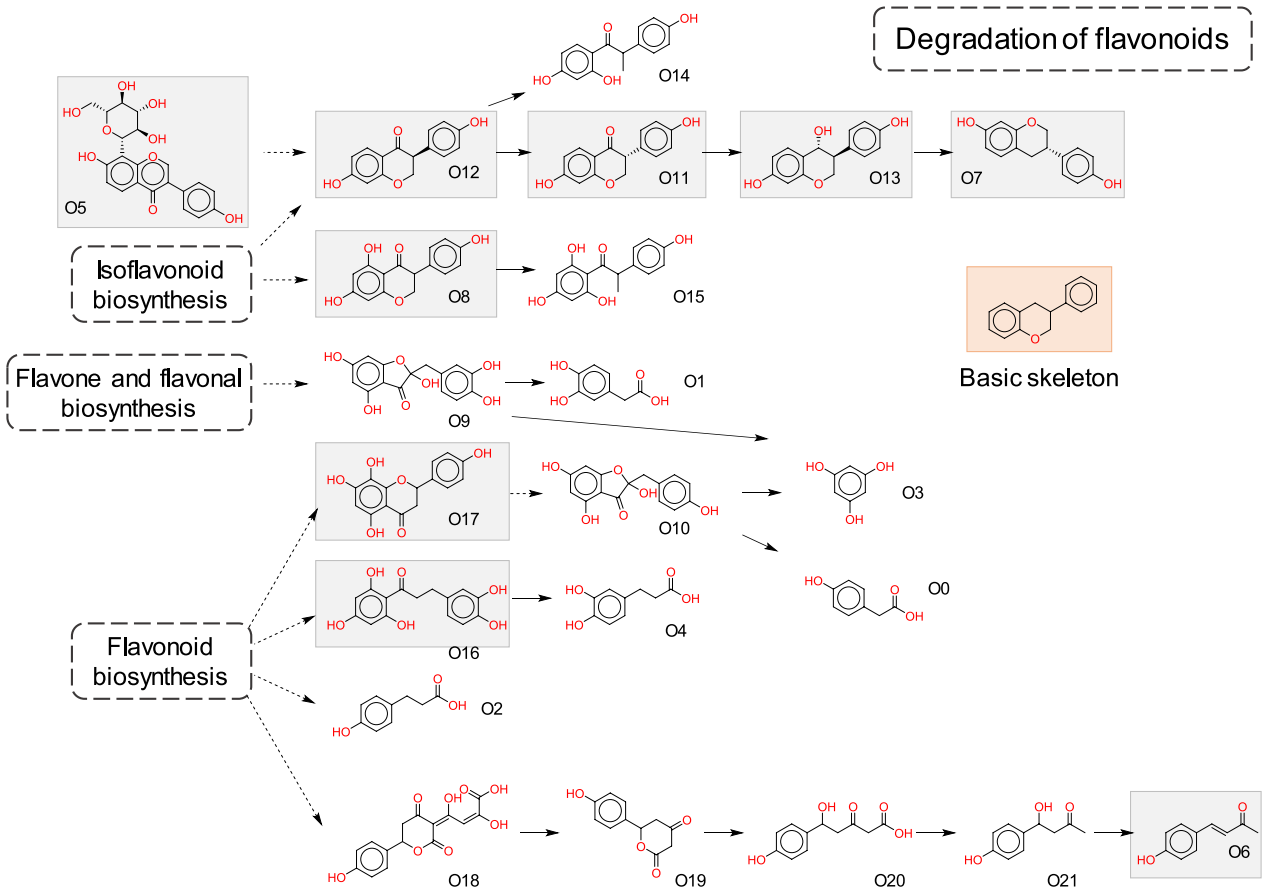

Fig. 7 Compounds in the degradation pathway of flavonoids. The molecules marked in the grey box are those that were successfully predicted, and the molecules not marked are those that were not successfully predicted. The structure in the orange box is the basic skeleton

Overall, this model effectively captures the main structural features associated with specific metabolic pathways in the molecule, so as to identify the classes of metabolic pathways that the molecule may be involved in.

Conclusion

In the article, we propose a hybrid neural network architecture, MotifMol3D, to predict the metabolic pathway categories in which small molecules may participate. Through motif analysis, it was revealed that partial motif information associated with specific pathways in the small-sample dataset can characterize key molecular features within these pathways, thereby enhancing the model's interpretability. Furthermore, external validation also confirmed that MotifMol3D effectively captures the primary structural features in molecules related to specific metabolic pathways. Compared to existing approaches, MotifMol3D exhibited superior performance in terms of precision, recall, and F1-score, establishing itself as the leading model in the field.

In terms of practical applications, the MotifMol3D model can be applied to the pathway analysis of differential metabolites in metabolomics studies (Additional file 1: Figure S1). By effectively predicting the KEGG metabolic pathway categories of compounds, metabolites lacking clear pathway information can be included in further analyses rather than being discarded. Although MotifMol3D cannot directly provide specific reaction or enzyme information, once pathway category information is obtained, we can further explore the potential roles of metabolites using other analytical methods, such as by comparing structurally and functionally similar metabolites in this pathway or combining with other omics analyses to indirectly infer their possible biological functions. Therefore, our model is capable of playing a key role in filling the gaps in metabolic pathway information, aiding further biological exploration.

However, the model does have certain limitations. When the chemical structure of a predicted molecule significantly diverges from the training set, prediction accuracy may decline. This is primarily due to the model relying on structural features in the training data, any gaps in this coverage may impair the model's ability to generalize to unseen compounds. To address this issue, expanding the diversity of the training set by incorporating a wider range of relevant compounds is necessary.

For future improvements, we plan to integrate additional metabolic pathway-related information, such as reaction types, enzyme catalysis characteristics, and upstream/downstream relationships of metabolites, to enhance the model's understanding and generalizability. By exploring these directions, we aim to continually optimize the model's performance and applicability in metabolic pathway prediction.

Abbreviations

KEGG database	Kyoto Encyclopedia of Genes database
SMPDB	The Small Molecule Pathway Database
RF	Random forest
GAT	Graph attention network
GCN	Graph convolutional network
FC layer	Fully connected layer
TDB descriptor	Topological distance based 3d descriptors
TF-IDF	Term frequency inverse document frequency

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00994-6.

Additional file 1: Figure S1.

Additional file 2: Tabel S1.

Author contributions

Jianguo Hu: Conceptualization, Methodology, Software, Writing—Original Draft; Yiqing Zhang: Software, Formal analysis, Writing—Review and Editing; Jinxin Xie and Zhen Yuan: Methodology, Investigation; Zhangxiang Yin and Shanshan Shi: Data Curation, Visualization; Honglin Li and Shiliang Li: Conceptualization, Supervision, Funding acquisition. All authors reviewed the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (grants 82173690 to S.L.L., 82425104 to H.L.); the National Key R&D Program of China (2022YFC3400504); the Fundamental Research Funds for the Central Universities; S.L.L. is also sponsored by the Shanghai Rising-Star Program (23QA1402800).

Availability of data and materials

Project name: MotifMol3D. Project home page: https://github.com/Irena-Zhang/MotifMol3D. The data described in this article are available at https://www.genome.jp/kegg/pathway.html#metabolism [4]. The MotifMol3D model is publicly accessible on the GitHub repository: https://github.com/Irena-Zhang/MotifMol3D.

Declarations

Competing interests

The authors declare no competing interests.

Received: 1 March 2024 Accepted: 21 March 2025
Published online: 21 April 2025

References

1. Baranwal M, Magner A, Elvati P et al (2020) A deep learning architecture for metabolic pathway prediction. Bioinformatics 36(8):2547–2553
2. Noda-Garcia L, Liebermeister W, Tawfik DS (2018) Metabolite–enzyme coevolution: from single enzymes to metabolic pathways and networks. Annu Rev Biochem 87:187–216
3. Ranganathan S, Nakai K, Schonbach C (2018) Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. Elsevier Science Publishers B.V., NLD.
4. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28(1):27–30
5. Karp PD, Riley M, Saier M et al (2000) The ecocyc and metacyc databases. Nucleic Acids Res 28(1):56–59
6. Frolkis A, Knox C, Lim E et al (2010) SMPDB: the small molecule pathway database. Nucleic Acids Res 38(suppl_1):D480–D487
7. Roche-Lima A (2016) Implementation and comparison of kernel-based learning methods to predict metabolic networks. Netw Model Anal Health Inform Bioinform 5(1):26
8. Hamdalla MA, Rajasekaran S, Grant DF et al (2015) Metabolic pathway predictions for metabolomics: a molecular structure matching approach. J Chem Inf Model 55(3):709–718
9. Moriya Y, Shigemizu D, Hattori M et al (2010) PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res 38(suppl_2):W138–W143
10. Faust K, Croes D, Van Helden J (2011) Prediction of metabolic pathways from genome-scale metabolic networks. Biosystems 105(2):109–121
11. Romero P, Wagg J, Green ML et al (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6(1):1–17
12. Cai Y-D, Qian Z, Lu L et al (2008) Prediction of compounds' biological function (metabolic pathways) based on functional group composition. Mol Divers 12:131–137
13. Dale JM, Popescu L, Karp PD (2010) Machine learning methods for metabolic pathway prediction. BMC Bioinform 11(1):1–14
14. Gao Y-F, Chen L, Cai Y-D et al (2012) Correction: Predicting metabolic pathways of small molecules and enzymes based on interaction information of chemicals and proteins. PLOS ONE 7(11). https://doi.org/10.1371/annotation/83922541-168a-4d4f-846a-cb5d127aa7a9
15. Hu L-L, Chen C, Huang T et al (2011) Predicting biological functions of compounds based on chemical-chemical interactions. PLoS ONE 6(12):e29491
16. Jia Y, Zhao R, Chen L (2020) Similarity-based machine learning model for predicting the metabolic pathways of compounds. IEEE Access 8:130687–130696
17. Du B-X, Zhao P-C, Zhu B et al (2022) MLGL-MP: a multi-label graph learning framework enhanced by pathway interdependence for metabolic pathway prediction. Bioinformatics 38(Supplement_1):i325–i332
18. Yang Z, Liu J, Shah HA et al (2022) A novel hybrid framework for metabolic pathways prediction based on the graph attention network. BMC Bioinform 23(5):1–15
19. Gaudelet T, Day B, Jamasb AR et al (2021) Utilizing graph machine learning within drug discovery and development. Brief Bioinform 22(6):bbab159. https://doi.org/10.1093/bib/bbab159
20. Liu XY, Yang HP, Ai CW et al (2023) MVML-MPI: multi-view multi-label learning for metabolic pathway inference. Brief Bioinform 24(6):bbad393. https://doi.org/10.1093/bib/bbad393
21. Bao H, Zhao JH, Zhao XJ et al (2023) Prediction of plant secondary metabolic pathways using deep transfer learning. BMC Bioinform 24(1):348

22. Liu YR, Jiang YQ, Zhang F et al (2024) A novel multi-scale graph neural network for metabolic pathway prediction. IEEE ACM Trans Comput Biol 21(1):178–187
23. Velickovic P, Cucurull G, Casanova A et al (2017) Graph attention networks. Stat 1050(20):10.48550
24. Hirohara M, Saito Y, Koda Y et al (2018) Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinform 19:83–94
25. Klein CT, Kaiser D, Ecker G (2004) Topological distance based 3D descriptors for use in QSAR and diversity analysis. J Chem Inf Comput Sci 44(1):200–209
26. Liu X, Bai F, Ouyang S et al (2009) Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. BMC Bioinform 10:1–14
27. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7):1466–1474
28. Landrum G (2013) RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum 8:5281
29. Yu Z, Gao H. Molecular representation learning via heterogeneous motif graph neural networks. In: International conference on machine learning. PMLR. 2022;25581–94.
30. Ramos J. Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. Citeseer 29–48; 2003.
31. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in neural information processing systems 30; 2017.
32. Yu J, Wang J, Zhao H et al (2022) Organic compound synthetic accessibility prediction based on the graph attention mechanism. J Chem Inf Model 62(12):2973–2986
33. Clevert D, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). In: ICLR 2016; 2016.
34. Deng D, Chen X, Zhang R et al (2021) XGraphBoost: extracting graph neural network-based features for a better prediction of molecular properties. J Chem Inf Model 61(6):2697–2705
35. Kopelman PG (2000) Obesity as a medical problem. Nature 404(6778):635–643
36. DeFronzo RA, Ferrannini E (1991) Insulin resistance. A multifaceted syndrome responsible for NIDDM, obesity, hypertension, dyslipidemia, and atherosclerotic cardiovascular disease. Diabetes Care 14(3):173–194
37. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674
38. Henderson JF, Paterson ARP (1973) Nucleotide metabolism: an introduction. Academic Press, London
39. Foo CS-Y, Pethe K, Lupien A (2020) Oxidative phosphorylation—an update on a new, essential target space for drug discovery in *Mycobacterium tuberculosis*. Appl Sci 10(7):2339
40. Motohashi H, Akaike T (2019) Sulfur-utilizing cytoprotection and energy metabolism. Curr Opin Physiol 9:1–8
41. Lieu EL, Nguyen T, Rhyne S et al (2020) Amino acids in cancer. Exp Mol Med 52(1):15–30
42. Lopez MJ, Mohiuddin SS (2020) Biochemistry, Essential Amino Acids. StatPearls Publishing, Treasure Island
43. Varki A, Cummings RD, Esko JD, et al. Essentials of glycobiology [internet]; 2022.

Publisher's Note