Journal of Advanced Research 68 (2025) 477–489

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Manuscript

Multi-task aquatic toxicity prediction model based on multi-level features fusion

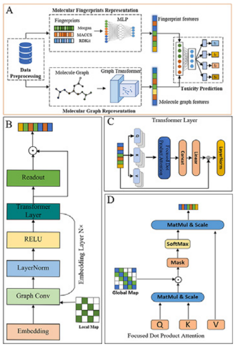
Xin Yang[a,b,1], Jianqiang Sun[c,1], Bingyu Jin[a], Yuer Lu[b], Jinyan Cheng[b], Jiaju Jiang[d], Qi Zhao[a,*], Jianwei Shuai[b,e,*]

[a] School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China
[b] Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325001, China
[c] School of Information Science and Engineering, Linyi University, Linyi 276000, China
[d] College of Life Sciences, Sichuan University, Chengdu 610064, China
[e] Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou 325001, China

HIGHLIGHTS

- Our dataset is sourced from multiple datasets representing different species, ensuring an ample quantity and quality of samples.
- We employ two different representation methods and incorporate a global attention mechanism.
- We extract shared features for multiple tasks through two fully connected layers and create separate output layers for each task.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 18 February 2024
Revised 21 May 2024
Accepted 2 June 2024
Available online 4 June 2024

Keywords:
Acute toxicity
Deep learning
Multi-task model
Molecular fingerprints
Molecular graph features

ABSTRACT

Introduction: With the escalating menace of organic compounds in environmental pollution imperiling the survival of aquatic organisms, the investigation of organic compound toxicity across diverse aquatic species assumes paramount significance for environmental protection. Understanding how different species respond to these compounds helps assess the potential ecological impact of pollution on aquatic ecosystems as a whole. Compared with traditional experimental methods, deep learning methods have higher accuracy in predicting aquatic toxicity, faster data processing speed and better generalization ability.

Objectives: This article presents ATFPGT-multi, an advanced multi-task deep neural network prediction model for organic toxicity.

Methods: The model integrates molecular fingerprints and molecule graphs to characterize molecules, enabling the simultaneous prediction of acute toxicity for the same organic compound across four distinct fish species. Furthermore, to validate the advantages of multi-task learning, we independently construct prediction models, named ATFPGT-single, for each fish species. We employ cross-validation in our experiments to assess the performance and generalization ability of ATFPGT-multi.

Results: The experimental results indicate, first, that ATFPGT-multi outperforms ATFPGT-single on four fish datasets with AUC improvements of 9.8%, 4%, 4.8%, and 8.2%, respectively, demonstrating the superiority of multi-task learning over single-task learning. Furthermore, in comparison with previous

[*] Corresponding authors at: School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (Q. Zhao); Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325001, China (J. Shuai).

E-mail addresses: zhaoqi@lnu.edu.cn (Q. Zhao), shuaijw@wiucas.ac.cn (J. Shuai).

[1] These authors contributed equally to the paper as first authors.

https://doi.org/10.1016/j.jare.2024.06.002

algorithms, ATFPGT-multi outperforms comparative methods, emphasizing that our approach exhibits higher accuracy and reliability in predicting aquatic toxicity. Moreover, ATFPGT-multi utilizes attention scores to identify molecular fragments associated with fish toxicity in organic molecules, as demonstrated by two organic molecule examples in the main text, demonstrating the interpretability of ATFPGT-multi.

Conclusion: In summary, ATFPGT-multi provides important support and reference for the further development of aquatic toxicity assessment. All of codes and datasets are freely available online at https://github.com/zhaoqi106/ATFPGT-multi.

Introduction

Regulatory bodies, such as the United States Environmental Protection Agency and the Organization for Economic Cooperation and Development, stress the significance of performing ecological and environmental risk assessments for both established and emerging chemicals before their production and usage [1,2]. Presently, artificially synthesized chemical substances are pervasive. Some organic compounds pose risks not only to aquatic life, but also have the potential to affect the atmosphere, water, soil, and ecosystems in the natural world. These compounds may persist in the environment for extended periods before undergoing degradation [3,4]. As these organic compounds contaminate the environment, they are concurrently absorbed by aquatic organisms through their skin, digestive system, and respiratory system. Accumulating in the organisms' liver, kidneys, fat, muscles, and brain tissues, these compounds pose significant threats to their survival. For instance, certain organic pollutants may lead to acute or chronic poisoning in aquatic organisms, affecting their growth, reproduction, and behavior [5,6]. Therefore, assessing the potential risks associated with these chemicals is paramount for the continued well-being of humanity.

Assessing aquatic toxicity plays a critical role in evaluating the environmental hazards and risks associated with various chemicals [7,8]. Previously, the acute toxicity of chemicals was determined through tests conducted on various fish species, such as the fathead minnow and zebrafish [9]. Nevertheless, appraising the safety of these chemicals through conventional experimental methods is not only costly but also time-intensive [10]. The sheer volume of these chemicals makes it practically unfeasible to assess each one thoroughly via in vitro or in vivo methods. In recent years, advancements in computational capabilities have led to the widespread application of machine learning techniques, especially deep learning, across various bioinformatics fields. These applications include the analysis of single-cell multi-omics data [11–16], computational toxicology [17–22], miRNA-lncRNA interactions prediction [23–25], metabolite-disease associations prediction [26,27], remote health monitoring [28–30] and circRNA-disease associations prediction [31,32]. These studies provide strong support for the continuous refinement of computational prediction models for the acute toxicity of chemicals. Compared to traditional experimental methods, deep learning methods typically excel in accuracy, especially in handling large-scale, high-dimensional data and extracting complex patterns. These methods often have the capability to automatically learn feature representations of data, and under sufficient training, they can exhibit strong generalization performance on unseen data. Presently, researchers have developed a range of quantitative structure–activity relationship (QSAR) models for the acute toxicity of aquatic organisms based on these technologies. Existing methods for predicting acute toxicity primarily fall into two categories: relying on molecular fingerprints and utilizing graph-based approaches.

Molecular fingerprint methods depict the structure and characteristics of molecules using binary strings, where each string corresponds to a structural element or feature [33,34]. If a molecule possesses that structure or feature, the corresponding binary bit is set to 1. Otherwise, it is set to 0. Using molecular fingerprints as a foundation, researchers have built a series of machine learning models, including regression and classification models, to predict acute toxicity [35,36]. For instance, in 2018, Cao et al. compiled a comprehensive dataset consisting of compounds from 824 crustacean species. They constructed a predictive model for aquatic toxicity by employing six machine learning methods and utilizing seven molecular fingerprints [37]. In 2019, Liu et al. built a diverse dataset encompassing various crustacean species to predict aquatic chemical toxicity and prioritize environmental hazard assessments [38]. In the same year, Ai et al. utilized three conventional machine learning algorithms to develop three ensemble models. Trained on a dataset containing 400 different chemicals, these models effectively pinpointed several molecule structures most pertinent to acute aquatic toxicity [39]. Li et al. employed a comprehensive approach by utilizing median lethal concentrations for 373 organic compounds from the environmental toxicology datasets ECOTOX and EAT5. They constructed five classic machine learning algorithms based on eight types of molecular fingerprints and implemented a multi-classification model that can more accurately classify the acute toxicity of organic compounds to aquatic organisms [40]. However, these methods encounter two significant limitations when dealing with high-dimensional feature data. First, while molecular fingerprints represent chemical structures, they often struggle to capture the complex structural features and interactions within molecules comprehensively. Second, their approach faces challenges in predicting properties of molecules not present in the training data, especially when dealing with rare or novel chemical structures.

Graph-based deep learning techniques depict molecular structures as graphs and utilize graph representation learning algorithms to understand feature representations of molecular structures [41]. In contrast to conventional machine learning approaches relying on molecular fingerprints, graph-based methods excel at capturing the intricate structure of molecules, leading to improved predictive accuracy and generalizability. In recent years, researchers have begun adopting graph neural network-based methods to forecast the toxicity of aquatic organisms. For example, in 2022, Xu et al. gathered 1874 distinct compounds along with their respective labels from ECOTOX and various literature sources. They employed both traditional machine learning techniques and a graph convolutional neural network (GCNN) architecture to develop predictive models [42]. Interestingly, it was observed that GCNN demonstrated superior predictive performance compared to the other tested approaches in their study. However, GCNN still has inherent limitations, such as its limited ability to handle global information or long-range relationships. Furthermore, the current QSAR models limit scalability as they are based on a single or a few fish species' data, making it challenging to design or predict for other fish species.

To tackle the aforementioned challenges, this study introduces a multi-task model named ATFPGT-multi, which integrates molecular fingerprints with a graph neural network featuring a global attention mechanism. ATFPGT-multi addresses the challenge of correlation between multiple tasks by sharing feature extraction layers, simultaneously accommodating the distinctions among different tasks by creating separate output layers for each task. This innovative design empowers our model to yield optimal results in multi-task learning scenarios. Our findings reveal that the proposed method outperforms previous traditional machine learning approaches and GCNN, signifying its potential as a valuable tool for environmental toxicity assessment. This study holds significance in contributing to the classification of acute toxicity in aquatic organisms caused by compounds and the prediction of the hazard levels associated with these compounds.

Materials and methods

Data preparation

We collect data from four species: bluegill sunfish (Lepomis macrochirus, BS), rainbow trout (Oncorhynchus mykiss, RT), fathead minnow (Pimephales promelas, FHM), and sheepshead minnow (Cyprinodon variegatus, SHM) [43–45]. All data originate from ECOTOX database. As these datasets comprise multiple repetitions of experiments conducted on the same species but under varied conditions, we utilize the python package RDKit to standardize and process the chemical structures of all compounds within our dataset. The following steps are undertaken: (1) Data filtering to retain the 96-hour acute toxicity values (96 h-LC50, mg/L). (2) Elimination of salts and inorganic compounds from the dataset. (3) Molecular representation using standardized SMILES notation, followed by merging records for the same molecule after standardization. (4) Delete records of compounds belonging to different categories (including toxic and non-toxic), missing values, and outliers. (5) Classification of the toxicity values of organic compounds in accordance with the EEC 92/32/EEC standard [46], wherein values less than 10 mg/L are categorized as 'toxic' and 'non-toxic' is assigned for values $\geq$ 10 mg/L.

After preprocessing, the final number of compounds for BS, RT, FHM, and SHM are 988, 1246, 938, and 346, respectively. On BS, RT, and SHM, the ratio of toxic to non-toxic instances is approximately 6:4, while on FHM, the ratio of toxic to non-toxic instances is close to 1:1. Fig. 1 illustrates the proportion of each fish species in the total count and the number of toxic versus non-toxic instances for each species.

Model framework

The framework of ATFPGT-multi designed for predicting the toxicity of four fish species, as depicted in Fig. 2A. It primarily consists of four main components: data pre-processing, fingerprints feature representation, molecular graph feature representation based on GNN and transformer, and a fully connected module for toxicity prediction. First, in the fingerprints feature representation module, we extract three different types of fingerprint features for each molecule in datasets, namely Morgan fingerprints [47], MACCS fingerprints [48], and RDKit fingerprints [49]. Since the feature dimensions of different fingerprints are not the same, we concatenate the three types of fingerprint features for each molecule together, forming a higher-dimensional feature representing the molecule's final fingerprint features. Subsequently, a multi-layer perceptron (MLP) network is employed for feature selection from these three different types of fingerprints. Moreover, within the molecular graph feature representation section, a novel graph transformation method is introduced to encode molecular graphs. This approach aims to achieve more robust molecular representations by leveraging the strengths of both GNN and transformer. By combining these techniques, ATFPGT-multi can effectively capture the diverse and complex characteristics of various molecules [50,51]. Finally, the molecular fingerprint features are integrated with the molecular graph features to create the compound's comprehensive features. Following this, a fully connected layer processes the fused features and creating separate outputs for each fish species to achieve multi-task classification.

Featurization stage

The chemical structures in our data are represented in the form of SMILES strings [52]. In order for ATFPGT-multi to accurately identify molecular structures, it is crucial to represent the molecules as accurately as possible. At this phase, we extract two types of features for each compound molecule, namely molecular fingerprint features and molecular graph features. Below, we elaborate on the details and implementation of these two features.

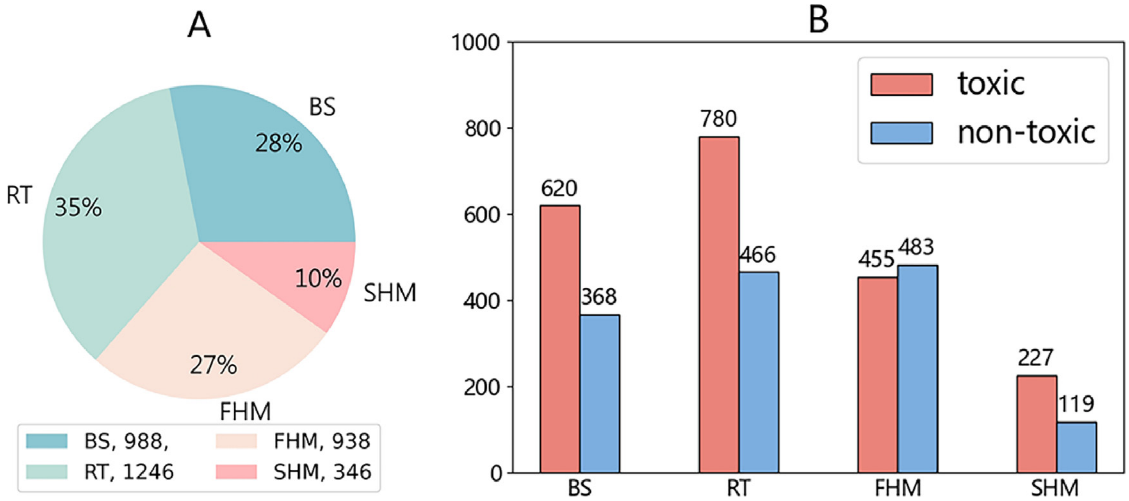


Fig. 1. (A) The amount of data for each fish species and its proportion to the total population. (B) The number of non-toxic and toxic compounds on each fish species.

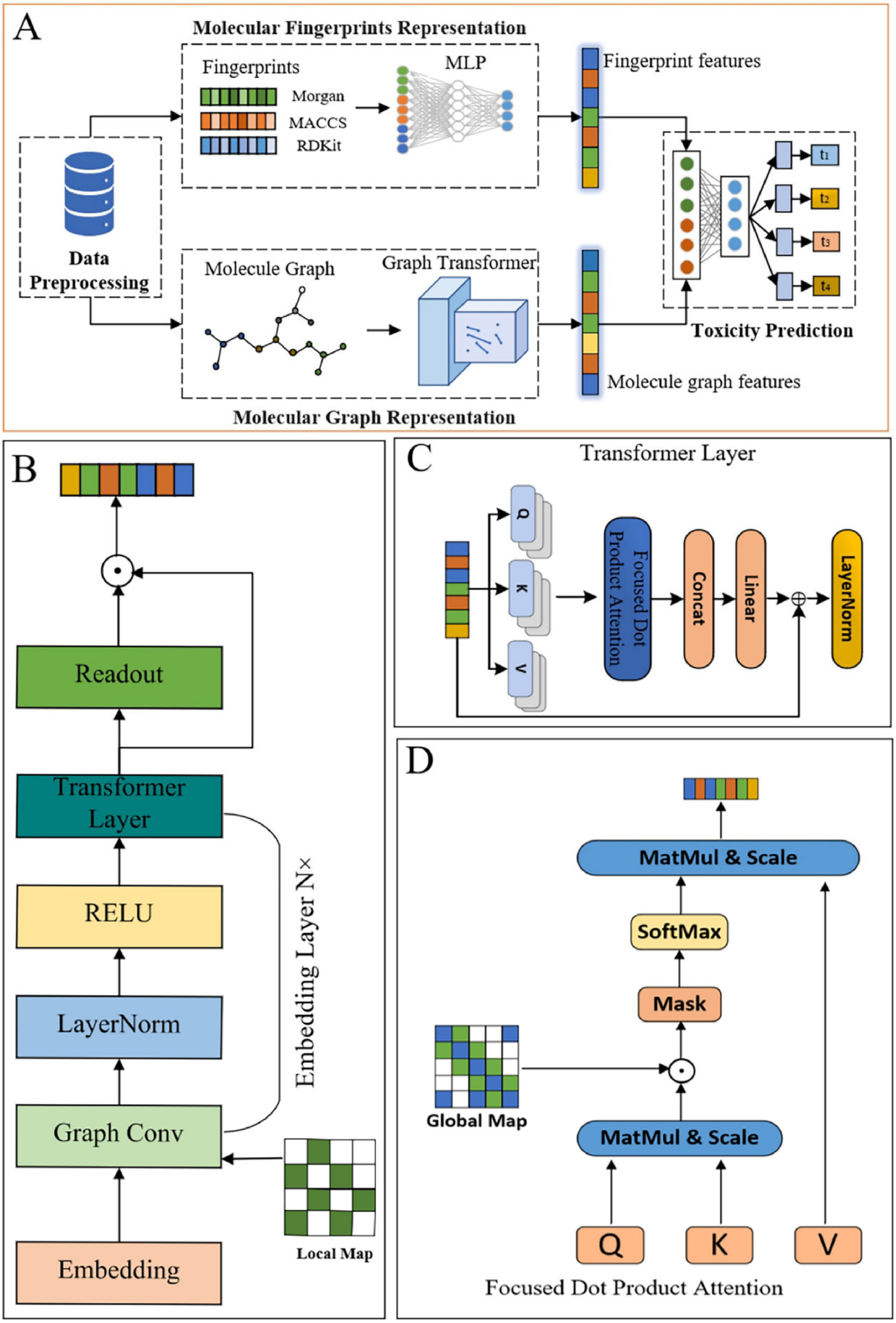


Fig. 2. (A) The overview of ATFPGT-multi. (B) The architecture of the molecular graph representation module, which presents the graph-based representation approach. (C) Shows the transformer layer details based on the multi-head attention mechanism. (D) Describes the process of focused dot product attention layer.

Molecular fingerprint features

Since different molecular fingerprint methods can capture various molecular characteristics, ATFPGT-multi utilizes three distinct types of molecular fingerprints, Morgan fingerprints, MACCS fingerprints, and RDKit fingerprints, to encode compound molecular information. Morgan fingerprints are based on the molecular topology. They encode structural information by recording the neighboring atoms around each atom in the molecule. Morgan fingerprints also take into account the distance information between atoms, making them effective for describing both local and global molecular structures. MACCS is a binary fingerprint based on specific structural fragments or substructures of molecules. They represent molecules as vectors containing 166 bits of binary information, with each bit representing a predefined molecular feature such as rings, bonds, functional groups, etc. The RDKit fingerprints algorithm crafts fingerprints by hashing paths within the molecular graph up to a predetermined length. These fingerprints are designed to capture the structural and chemical characteristics of molecules. Morgan, MACCS, and RDKit fingerprints have found extensive applications and validation in the field of cheminformatics, demonstrating excellent performance in many chemical tasks. These fingerprint methods capture different molecular characteristics, each generating a fixed-length binary vector for every molecule. Each bit in the vector represents a unique molecular substructure. The numerical or binary values represent the presence or absence of specific molecular substructures within the molecule, which are pivotal for understanding molecular activity. The specific information of three types of molecular fingerprints is shown in Table 1.

To represent molecular structure more precisely, feature processing is essential. We perform dimensionality reduction on the generated features to eliminate irrelevant features from the molecular fingerprints. We delineate the specific details and implementation as follows.

In formal terms, let X be the entire fish dataset, where each molecule in the dataset is represented by its SMILES notation. For each molecule x in X , we extract its feature vectors using Morgan fingerprints *M(x)*, MACCS fingerprints *A(x)*, and RDKit fingerprints *R(x)*, concatenating these three vectors together as the molecule's fingerprint feature. The formula is expressed as follows:

$$Z(x) = [M(x), A(x), R(x)] \tag{1}$$

where *Z(x)* represents the final fingerprint feature of the molecule, which is a binary vector of length L, where L is the sum of the dimensions of the three types of fingerprint features. To create a more efficient and concise feature, we utilize a trainable MLP to reduce the dimensionality of the features, retaining only the most relevant features. The output of MLP is then passed to our feature fusion module for toxicity prediction. Formally, we denote $f(Z(x);W)$ as the MLP function with weight parameters W. The definition of MLP is as follows:

$$f(Z(x);W) = ReLU(W_2 * ReLU(W_1 * Z(x) + b_1) + b_2) \tag{2}$$

where W_1 and W_2 are weight matrices, b_1 and b_2 are bias vectors, and * denotes matrix multiplication. The ReLU activation function is defined as $ReLU(z) = \max(0, z)$. This feature processing method facilitates the capture of intricate nonlinear interactions among diverse molecular substructures within molecules, a critical factor for precisely predicting the toxicity of organic compounds to fish in aquatic environments. In summary, the dimensionality reduction of concatenated fingerprint features using MLP is a critical step, as it helps optimize ATFPGT-multi's performance by reducing complexity, improving generalization, and enhancing interpretability.

Table 1
Name, length, and type of molecular fingerprints.

Molecular fingerprints	Bits	Type
Morgan	2048	Circular fingerprints
MACCS	167	Structural features
RDKit	2048	Structural features

Molecular graph feature

Molecular graph features serve as an advanced representation method for describing molecular structures, capturing the relationships between atoms and bonds within a molecule. In this feature representation method, each node in the molecular graph represents an atom, while the edges represent bonds. Additionally, we design two weighted graphs, namely local map (Loc) and global map (M), to encompass more crucial molecular attributes.

Specifically, the molecular graph representation module defines a molecular graph $G = (V, E, Loc, M)$, where atoms are represented as nodes V, bonds are represented as edges E, bond weights are denoted as Loc, and the bond properties and distance information between atoms are represented as M. Loc is employed to characterize the bond information in molecules by allocating weights to each bond type, thereby regulating the propagation of messages between adjacent nodes. In contrast, M not only captures the interactions between atoms but also encodes the properties of the bonds. Furthermore, M incorporates distance information between atoms. Following chemical principles, bonds with a greater number of electrons, such as unsaturated bonds, receive higher weights to amplify the exchange of information between terminal atoms.

As illustrated in Fig. 2B, within the module of molecular graph representation, ATFPGT-multi's capability for representation is enhanced by aggregating node features through graph convolutional layers. The incorporation of LayerNorm and ReLU operations aims to enhance the model's robustness and training stability while introducing non-linearity to better capture intricate relationships within the graph structure. Furthermore, by integrating a transformer layer, the model becomes more adept at capturing essential correlations within the global graph structure, facilitating a comprehensive understanding of information embedded in the graph. Iteratively applying the graph transformer process strengthens the model's ability for multi-level and multi-scale feature abstraction and learning on graph data. The specific implementation details are outlined as follows.

In the graph convolutional layer, Loc is employed to provide edge attributes, facilitating the aggregation of information from neighboring nodes to update node feature representations and derive local atomic embeddings. The process is governed by the local mapping, as depicted in Equation (3).

$$h_i^l = W_1^l h_i^{l-1} + W_2^l \sum_{j \in N(i)} Loc_{j,i} \cdot h_j^{l-1} \tag{3}$$

where, h_i^{l-1} and h_j^{l-1} denote the features of atoms i and j from the l-1th layer, respectively. h_i^l represents the updated embedding of atom i in the lth layer, *N(i)* denotes all neighboring atoms of atom i, while W_1^l, W_2^l are two trainable matrices in the layer l. The bond weight $Loc_{j,i}$ which govern the extent of message propagation is contingent upon the type of bond between atoms i and j. In conclusion, message propagation enables each node to aggregate information from its neighboring nodes and update its feature representation accordingly, thereby enhancing the model's performance.

Additionally, as illustrated in Fig. 2C, following the generation of graph embeddings by the GNN, a transformer layer is introduced. This is a newly designed component that incorporates a multi-head attention mechanism. The objective is to enhance the abstract representation of molecular graph information, making it more robust and better capable of capturing the diversity and complexity among different molecules. Compared to GNN, which integrate local molecular neighborhood information, transformers excel at focusing on the connections between atoms at a global scale, thus capturing superior global structural information of molecules. In this work, we enhance the transformer's performance using a global map to better understand the molecule's overall environment through more concentrated attention. We introduce a focused dot product attention layer to enhance the performance of the traditional transformer, as illustrated in Fig. 2D. The formula is defined as follows:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}} \cdot M)V \tag{4}$$

In the above equation, $Q,K,V \in R^{N_{atoms} \times d}$ represent the query vector, key vector, and value vector in the transformer. Moreover, this equation utilizes M to weigh affinity matrices. This method enables better focus on atom pairs with stronger interactions while reducing attention on irrelevant atom pairs. It enhances the performance of the transformer in understanding complex molecular structures. Finally, by merging the graph embeddings generated by the GNN with the information processed through the transformer layer, the ultimate representation is formed. Once the embedding module completes atom feature learning, the readout function calculates attention scores for each atom and combines atom embeddings into fully connected prediction layers to generate molecular embeddings [53].

In summary, our molecule graph representation module can provide more precise structural information about molecules, thereby producing accurate molecular feature representations that provide more information for prediction task. The combination of GNN and transformer allows for the integration of local and global information, leading to more comprehensive and effective molecular embeddings. This synergistic approach enables ATFPGT-multi to capture intricate structural patterns, long-range dependencies, and holistic molecular properties, ultimately improving predictive performance. Furthermore, to better control the balance between nodes and their neighboring nodes, we use the sigmoid function to weigh the relative importance of node features and neighboring node features. This approach improves the performance of ATFPGT-multi.

Fully connect neural network

In this section, we will concatenate the molecular fingerprint features and molecular graph features together, and linearly transform the features through fully connected layers to obtain a more concise and precise feature representation. Then, an independent fully connected output is created for each fish dataset to achieve multi-task classification. By sharing certain layers and parameters of the model, ATFPGT-multi is able to learn shared representations. This shared representation can capture common features across multiple tasks, thereby enhancing the generalization capability of ATFPGT-multi. This approach enhances the model's generalization performance for each task. ATFPGT-multi can adapt better to the features and characteristics of each task, thus improving overall performance. Each neuron in a layer is connected to every neuron in the preceding layer, allowing the network to capture complex relationships in the input data. Each connection has a weight and bias, enabling neurons to learn complex feature representations and effectively model input data through linear transformations and nonlinear activations. The formula for the fully connected neural network is shown as follows:

$$y = f(Wc + b) \tag{5}$$

where y represents the output of the fully connected neural network layers, f is the ReLU activation function, W is the weight matrix, c represents the fused molecular features, and b is the bias term. In the last layer, we apply a sigmoid activation function for each task, optimized using the binary cross-entropy loss function (BCELoss). The equation for BCELoss is can be calculated by Equation (6).

$$BCELoss = -\frac{1}{Z}\sum_{i=1}^{Z}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{6}$$

where Z represents the total number of compound samples, y_i is the category of the i-th sample, and p_i is the prediction value of the $i-$th compound, typically a probability value with the range of 0 to 1. Moreover, we utilize Adam for optimizing our model. The Adam optimization algorithm has advantages such as adaptive learning rate, low memory requirements, stability, and simple parameter adjustment.

Performance evaluation

To comprehensively assess the performance of ATFPGT-multi, we employ a five-fold cross-validation procedure. In this process, the dataset is divided into five equally sized subsets, with four subsets used as training sets and one subset used as a validation set during each iteration of training. This process is repeated five times, ensuring that each subset is used as a validation set once. By using multiple subsets of the data for training and validation, cross-validation helps assess how well the model generalizes to unseen data. If the model performs consistently well across all folds, it suggests good generalization ability. Additionally, we repeat the entire experimental procedure 100 times to ensure randomness in the experiment and obtain reliable experimental results. The following metrics are used to evaluate the predictive performance of ATFPGT-multi, including accuracy (ACC), recall (RE), and precision (PR). The calculation formulas for these metrics are as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

where TP represents the correctly predicted toxic compounds, TN represents the correctly predicted non-toxic compounds, FP represents the non-toxic compounds incorrectly predicted as toxic, and FN represents the toxic compounds incorrectly predicted as non-toxic. The area under the receiver operating characteristic curve (AUC) is also calculated to reflect the classifier's discriminative ability, and it is mainly used for comparing and selecting modeling methods and hyperparameters.

Experimental detail

Each organic compound is represented as a string using SMILES notation. First, the SMILES is converted into three different molecular fingerprints, which are then concatenated together. Subsequently, a MLP is employed to reduce the dimensionality, yielding a 256-dimensional embedding. Simultaneously, the SMILES is transformed into a molecular graph using RDKit. Then implementing GNN to aggregate the local information of the molecule, followed by using transformer to aggregate global information to obtain atom embeddings, and then calculating attention scores for each atom through a readout layer, and combining atom embeddings to generate molecular graph embeddings. The concatenated molecule fingerprint embedding and molecular graph embedding serve as the input for a shared fully connected layer. ATFPGT-multi is achieved by creating separate classification heads for each task. ATFPGT-multi is pre-trained for 30 epochs using the Adam optimizer with a learning rate of 0.0005. In the specific implementation, we initially use random search for preliminary exploration to narrow down the range of candidate hyperparameters. Then, we perform grid search within this narrowed range for more detailed tuning, further refining the hyperparameters to achieve optimal performance. A cosine learning rate scheduler is employed to adjust the learning rate during training. The entire framework is implemented using PyTorch. Additional fine-tuning details are provided in Table 2.

Results

Multi-task learning boosted performance

The framework of ATFPGT-multi is a novel deep learning architecture designed to combine data from multiple categories, implementing a multi-task learning strategy. Additionally, to validate whether multi-task learning methods improve toxicity predictive performance, we establish separate prediction models (ATFPGT-single) for each fish species. ATFPGT-single is a variant of ATFPGT-multi, where the former has only one classification head in the output layer, enabling predictions for only one classification task, whereas the latter comprises four parallel classification heads, allowing simultaneous predictions for four classification tasks. As shown in Table 3, the AUC values of ATFPGT-multi on four fish datasets are 0.932, 0.928, 0.881, and 0.906, respectively, which are 9.8 %, 4 %, 4.8 %, and 8.2 % higher than that of ATFPGT-single. The ACC values of ATFPGT-multi are 0.869, 0.854, 0.779, and 0.887, respectively, representing increases of 10.4 %, 3.9 %, 2.6 %, and 17.3 % compared to that of ATFPGT-single. The PRE values of ATFPGT-multi are 0.865, 0.866, 0.902, and 0.7, respectively, exhibiting improvements of 13.5 %, 11.2 %, 19.6 %, and 4 % over that of ATFPGT-single. The RE values of ATFPGT-multi are 0.801, 0.77, 0.729, and 0.75, showing improvements of 22.7 %, 6 %, and 10.6 % over that of ATFPGT-single on BS, RT, and SHM datasets, respectively. However, it exhibits a 6.7 % decrease compared to that of ATFPGT-single on FHM dataset. This could be attributed to a higher number of non-toxic samples in the entire sample space than toxic samples. ATFPGT-multi might be more inclined to predict samples as non-toxic, emphasizing accurate predictions of non-toxic instances. Based on the above result analysis, it can be concluded that the joint learning on relevant tasks in ATFPGT-multi allows for a better understanding of the inter-relationships among tasks, thereby enhancing overall performance.

For a more intuitive depiction of ATFPGT-multi's performance, in Fig. 3, we compare the performance of ATFPGT-multi and ATFPGT-single on each fish dataset using bar charts. On RT, BS, and SHM datasets, ATFPGT-multi outperforms ATFPGT-single in all evaluation metrics. On FHM dataset, except RE, the other three evaluation metrics of ATFPGT-multi are also higher compared to ATFPGT-single. Based on the above analysis, we conclude that the multi-task learning allows models to share parameters across multiple tasks, which can enhance ATFPGT-multi's generalization ability. By sharing underlying representations, ATFPGT-multi can transfer features and knowledge learned from one task to others, thereby improving overall performance. Cross-species learning with shared representations can enhance model performance by sharing knowledge and representations, facilitating knowledge transfer and transfer learning. Through a series of experiments,

Table 2
The hyper-parameters for ATFPGT-multi model.

Parameter	**Description**	**Range**
batch_size	Input batch size	{32, 64}
lr	Learning rate	{0.0005}
train_epoch	Training epoch	{30, 50}
dropout	Dropout ratio	{0.1, 0.2}
hidden_size	Size of hidden layers in MLP	{512}
attn_layers	Number of embedding layers	{4}
atten_head	Number of attention heads for transformer	{4}
output_dim	Hidden size of embedding layer	{256}
D	Hidden size of readout layer	{2}

Table 3
Comparative analysis of ATFPGT-multi and other methods under 5-fold CV on four fish species datasets.

Fish species	Models	AUC	ACC	PR	RE
BS	ATFPGT-multi	**0.932**	**0.869**	**0.865**	**0.801**
	ATFPGT-single	0.849	0.787	0.762	0.653
	GCN-multi	0.796	0.674	0.599	0.880
	GCN-single	0.875	0.826	0.727	0.880
	ANN-KRFP	0.775	0.794	0.709	0.840
	SVM-KRFP	0.780	0.800	0.709	0.850
	RF-KRFP	0.761	0.787	0.673	0.850
RT	ATFPGT-multi	**0.928**	**0.854**	**0.866**	**0.770**
	ATFPGT-single	0.892	0.822	0.779	0.725
	GCN-multi	0.804	0.719	0.675	0.799
	GCN-single	0.835	0.787	0.643	0.872
	SVM-KRFP	0.794	0.827	0.667	0.922
	RF-MACCS	0.822	0.849	0.714	0.929
	RF-KRFP	0.800	0.831	0.679	0.922
FHM	ATFPGT-multi	**0.881**	**0.779**	**0.902**	**0.729**
	ATFPGT-single	0.841	0.759	0.754	0.781
	GCN-multi	0.744	0.505	0.870	0.454
	GCN-single	0.847	0.770	0.812	0.734
	SVM-PubChem	0.787	0.787	0.788	0.787
	SVM-KRFP	0.832	0.833	0.812	0.851
	RF-PubChem	0.762	0.764	0.738	0.787
SHM	ATFPGT-multi	**0.906**	**0.887**	**0.700**	**0.750**
	ATFPGT-single	0.837	0.756	0.671	0.678
	GCN-multi	0.693	0.604	0.733	0.566
	GCN-single	0.859	0.881	0.625	0.947
	ANN-CDK	0.743	0.797	0.625	0.860
	RF-PubChem	0.786	0.831	0.688	0.884
	SVM-CDKExt	0.786	0.831	0.688	0.884

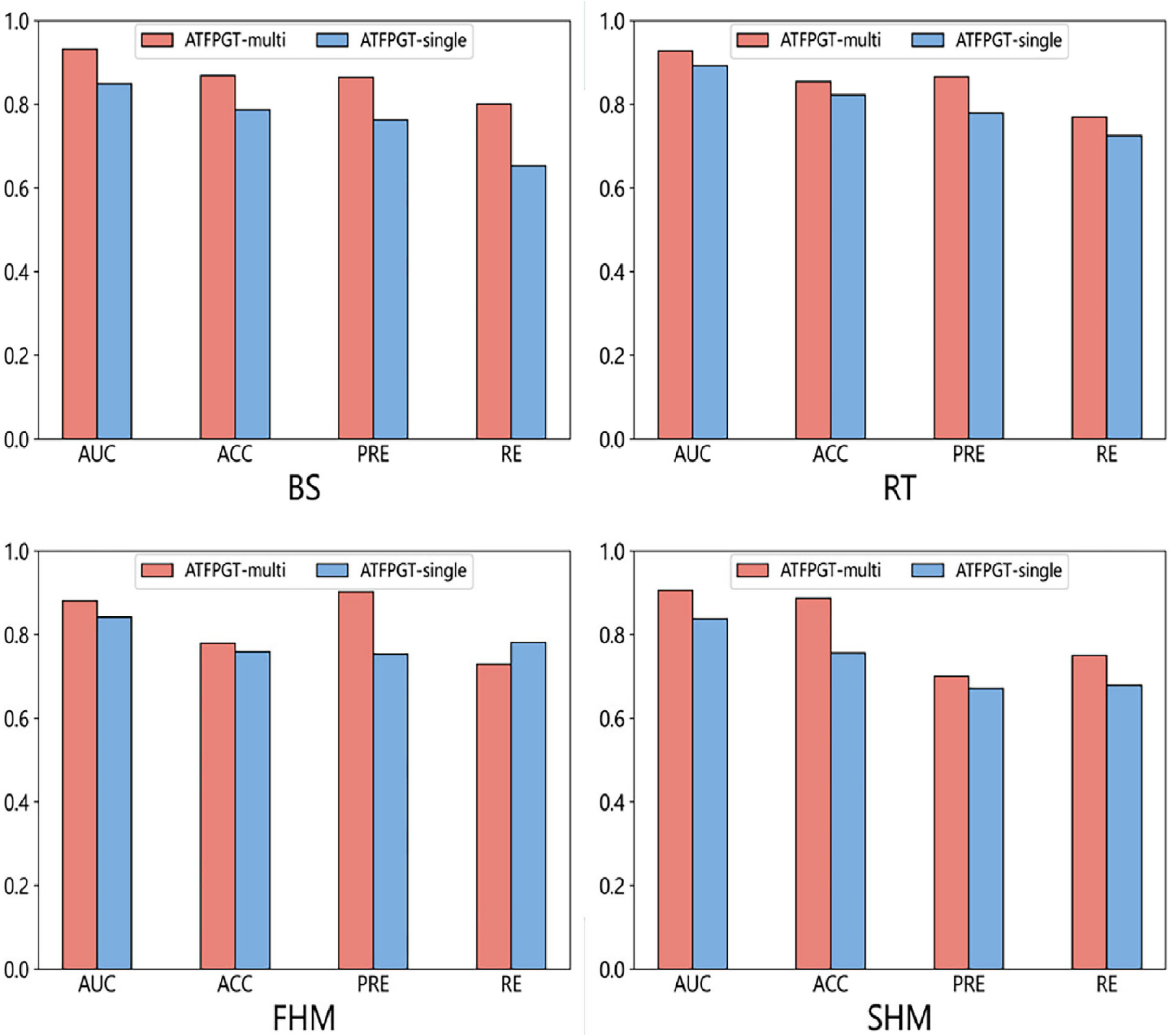


Fig. 3. Comparative analysis of ATFPGT-multi and ATFPGT-single for four fish species datasets.

the effectiveness of our approach in improving model performance, particularly in enhancing the model's toxicity prediction capability, has been demonstrated.

Comparison with other methods

To evaluate the performance of ATFPGT-multi further, we draw inspiration from previous study where researchers not only constructed 36 classification models for four datasets using four molecular fingerprints and nine machine learning algorithms but also developed single-task GCN (GCN-single) and multi-task GCN (GCN-multi) models [32], as shown in Table 3. From their results, we select the top-performing three machine learning models for each fish dataset. Specifically, for BS dataset, we gather ANN-KRFP, SVM-KRFP and RF-KRFP. On RT dataset, we choose SVM-KRFP, RF-KRFP and RF-MACCS. For FHM dataset, we select SVM-PubChem, RF-PubChem and SVM-KRFP. For SHM dataset, the chosen algorithms include ANN-CDK, RF-PubChem and SVM-CDKExt. Additionally, we chose GCN-multi and GCN-single models for each dataset.

We compare ATFPGT-multi with these classical methods. On BS dataset, the AUC of ATFPGT-multi is 20.3 %, 19.5 %, 30.2 %, 6.5 %, and 17.1 % higher than that of ANN-KRFP, SVM-KRFP, RF-KRFP, GCN-single, and GCN-multi, respectively. On RT dataset, the AUC of ATFPGT-multi surpasses that of SVM-KRFP, RF-KRFP, RF-MACCS, GCN-single, and GCN-multi by 16.9 %, 16 %, 12.9 %, 11.1 %, and 15.4 %. For FHM dataset, ATFPGT-multi's AUC outperforms that of SVM-PubChem, RF-PubChem, SVM-KRFP, GCN-single, and GCN-multi by 11.9 %, 15.6 %, 5.9 %, 4 %, and 18.4 %, respectively. On SHM dataset, the AUC of ATFPGT-multi is higher than that of ANN-CDK, RF-PubChem, SVM-CDKExt, GCN-single, and GCN-multi by 21.9 %, 15.3 %, 15.3 %, 5.5 %, and 30.7 %, respectively. The above analysis results indicate that ATFPGT-multi enhances the model's generalization ability, improves data efficiency, and strengthens the predictive capability for toxicity through multiple feature representations.

Additionally, from Fig. 4, we can observe that ATFPGT-multi achieves the highest values for key evaluation metrics, including AUC, and ACC, compared to other algorithms on BS, RT, FHM and SHM datasets. In conclusion, through the above analysis, it is evident that ATFPGT-multi exhibits superior overall performance compared to other models. This further elucidates its advantages in feature extraction and multi-task learning.

Comparing with existing prediction methods, as shown in Table 4, ATFPGT-multi performs better on the key evaluation metric AUC than the comparative algorithms. This indicates that

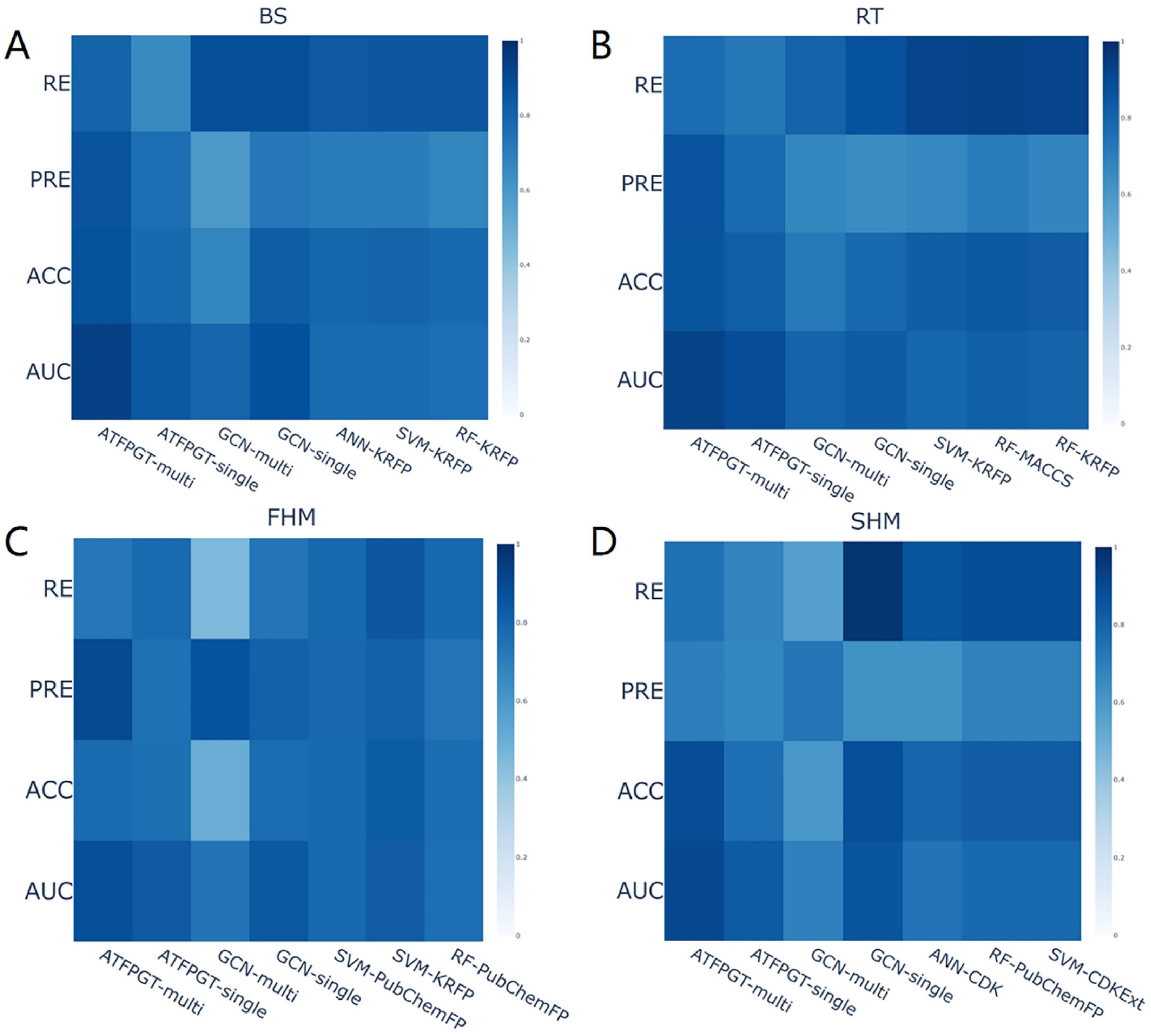


Fig. 4. The performance of ATFPGT-multi and the comparative algorithms for each fish species datasets.

Table 4
The AUC performance of different methods under 5-fold CV on four fish species datasets.

Models	BS	RT	FHM	SHM
ATFPGT-multi	**0.932**	**0.928**	**0.881**	**0.906**
ATFPGT-single	0.849	0.892	0.857	0.837
GCN-multi	0.796	0.804	0.744	0.693
GCN-single	0.875	0.835	0.847	0.859
SVM	0.78	0.794	0.832	0.786
RF	0.761	0.822	0.762	0.786
ANN	0.775	0.792	0.768	0.743

ATFPGT-multi can integrate information from different tasks, thereby improving the efficiency of data utilization, especially when there is correlation between tasks. Multi-task learning can enhance ATFPGT-multi's generalization ability, enabling it to perform better on new data. To some extent, it addresses the limitations of existing models that can only predict a single task and the risk of overfitting. In addition, multi-task learning facilitates knowledge transfer between tasks, so that knowledge learned from one task can contribute to the learning of another task.

Ablation studies on ATFPGT-multi

The comprehensive evaluation results of ATFPGT-multi indicate that the multi-task approach achieves the best predictive performance. To investigate the importance of each component in the ATFPGT-multi structure, a series of ablation experiments are conducted.

- ATFPGT-FP: it only uses three molecular fingerprint features of Morgan, MACCS and RDKit as input for the toxicity prediction.
- ATFPGT-GT: it only applies molecular graph features as input for the toxicity prediction.

We show the performance of these models in Table 5. The results show that ATFPGT-FP achieves AUC values of 0.807, 0.8, 0.725, and 0.758 on four fish datasets, respectively, which are lower than ATFPGT-multi by 15.5 %, 16 %, 21.5 %, and 19.5 %. ATFPGT-GT obtains AUC values of 0.885, 0.884, 0.846, and 0.873 on the four fish datasets, respectively, which are lower than ATFPGT-multi by 5.3 %, 5 %, 4.1 %, and 3.8 %. Specifically, solely using molecular graph representation features would result in a

Table 5
Comparison analysis between ATFPGT-multi and its ablation experiments on four fish datasets.

Fish species	Method	AUC	ACC	PRE	RE
BS	ATFPGT-multi	**0.932**	**0.869**	**0.865**	**0.801**
	ATFPGT-FP	0.807	0.762	0.690	0.710
	ATFPGT-GT	0.885	0.825	0.743	0.847
RT	ATFPGT-multi	**0.928**	**0.854**	**0.866**	**0.770**
	ATFPGT-FP	0.800	0.739	0.709	0.702
	ATFPGT-GT	0.884	0.821	0.782	0.834
FHM	ATFPGT-multi	**0.881**	**0.779**	**0.902**	**0.729**
	ATFPGT-FP	0.725	0.676	0.807	0.750
	ATFPGT-GT	0.846	0.772	0.767	0.781
SHM	ATFPGT-multi	**0.906**	**0.887**	**0.700**	**0.750**
	ATFPGT-FP	0.758	0.711	0.611	0.666
	ATFPGT-GT	0.873	0.809	0.731	0.868

decrease in model performance. Molecular fingerprints typically represent specific molecular fragments, accurately capturing subtle differences between different molecules. Similarly, relying solely on the molecular fingerprint encoding module would also lead to significant performance degradation, emphasizing the importance of the molecular graph representation module in capturing global molecular information and the correlation between atoms.

Additionally, experimental evidence demonstrates that our proposed feature fusion method outperforms in toxicity prediction. Through exploring the impact of different components on the model, it is found that ATFPGT-multi, which simultaneously adopts two molecular representation methods, exhibits the best performance. This indicates that these two modules complement each other in representing molecular features, demonstrating their potential in the field of multi-species acute toxicity prediction.

Exploring different types of molecular fingerprints

To explore the impact of various combinations of molecular fingerprints on the model, we investigate seven distinct combinations of molecular fingerprints to evaluate ATFPGT-multi, and the comparative results are presented in Fig. 5. To facilitate the identification of the model with the best fingerprint combination, we calculate the average evaluation metrics for each model on four datasets. Specifically, when using Morgan, MACCS, and RDKit alone, the models are named ATFPGT-Morgan, ATFPGT-MACCS, and ATFPGT-RDKit, respectively. Their AUC values are 0.827, 0.798, and 0.816, respectively. These values are lower than that of ATFPGT-multi by 11.9 %, 17.2 %, and 13.4 %. Among them, ATFPGT-Morgan exhibits the best performance, while ATFPGT-MACCS performs the worst. Next, we explore pairwise combinations of the three molecular fingerprints: Morgan + MACCS (ATFPGT-MorMAC), Morgan + RDKit (ATFPGT-MorR), and MACCS + RDKit (ATFPGT-MACR). The AUC values for these combinations are 0.868, 0.871, and 0.815, respectively. These values are lower than that of ATFPGT-multi by 6.6 %, 6.2 %, and 13.5 %, respectively. Among these, ATFPGT-MA yields the best results, while ATFPGT-MR performs the lowest. When utilizing all three types of molecular fingerprint patterns simultaneously, the model's performance surpasses that of all the aforementioned combinations of molecular fingerprints. These results suggest that these fingerprints are complementary in capturing different molecular features. In conclusion, our study indicates that a multi-fingerprint

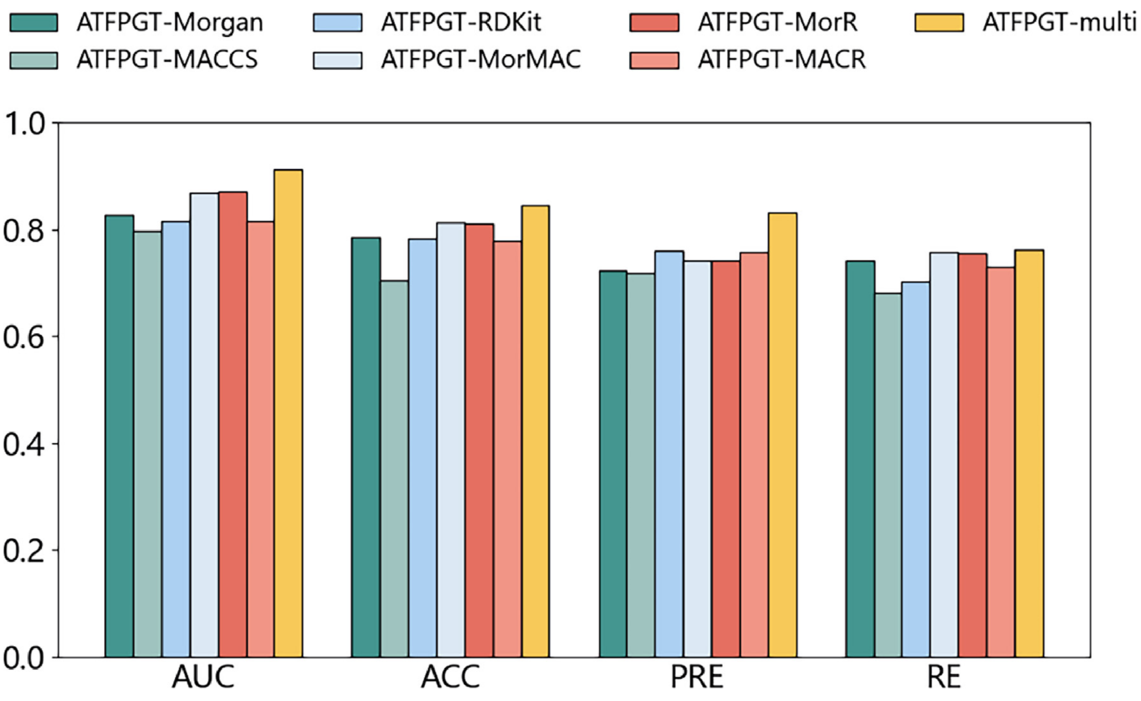


Fig. 5. Comparison analysis between ATFPGT-multi and its six variant models on four fish datasets.

approach is beneficial for toxicity prediction, with Morgan + MACCS + RDKit combination yielding the best results.

Model interpretability

Traditional models often lack interpretability, making it difficult to intuitively understand the relationship between molecular structure and toxicity. To address this challenge, we conduct a comprehensive interpretability case study of ATFPGT-multi. Through visualizing the atomic-level attention weights, we further investigate the contribution of individual atoms to molecular toxicity. Take the example of four compounds, CC(C)CC(C)c1sccc1NC(=O)c1cn(C)nc1C(F)(F)F, CC(C)[C@@H](Nc1ccc(C(F)(F)F)cc1Cl)C(=O)OC(C#N)c1cccc(Oc2ccccc2)c1, Cn1cc(C(=0)Nc2ccccc2-c2cc(F)c(F)c(F)c2)c(C(F)F)n1 and COC(=O)c1nc(−c2ccc(Cl)c(0C)c2F)cc(N)c1Cl. As shown in Fig. 6A, we observe that three fluorine atoms in the sample molecules exhibit high attention weights. In organic molecules, compounds containing fluorine atoms often possess elevated electronegativity. This may enhance interactions with other atoms, leading to alterations in biological activity. Additionally, some atoms, such as cyano groups and aldehyde groups, known to influence molecular toxicity, also receive significant attention [54]. Similarly, in Fig. 6B, we observe that six Cl atoms in the sample molecules receive high attention. This significantly influences the molecule's polarity, thereby affecting membrane permeability and negatively impacting cellular stability [55]. The same experimental conclusion can also be observed in Fig. 6C and D.

By delving deeper into these findings, several noteworthy observations can be made. First, ATFPGT-multi demonstrates remarkable capabilities in recognizing the significance of different functional moiety in influencing toxicity, which may vary depending on the dataset. This adaptability implies that the model has broad prospects in various molecular property predictions, especially in toxicity prediction tasks. Second, the global attention mechanism employed by ATFPGT-multi helps elucidate the relationship between molecular structure and properties. Through in-depth analysis, we enhance our understanding of the potential mechanisms of ATFPGT-multi and further demonstrate its promise in the field of acute toxicity prediction in biology. For example, attention scores can reveal the extent to which different parts of a molecule contribute to toxicity, helping chemical designers focus and adjust atomic groups that may contribute to toxicity. By modifying the molecular structure in a targeted way, safer and more environmentally friendly compounds can be developed. Moreover, identifying relevant fragments of molecules can speed up the chemical evaluation process, allowing researchers to more quickly assess the potential toxicity and environmental impact of compounds. This helps improve the efficiency and accuracy of chemical assessments and promotes a faster and more comprehensive understanding of chemical safety.

Discussion and conclusion

In this study, we collect data from four different fish species, each corresponding to a distinct task. We introduce ATFPGT-multi, a model capable of jointly predicting toxicity across all four tasks. This model integrates multi-level fingerprints and a graph transformer architecture, showcasing excellent performance in predicting acute toxicity. Our comprehensive evaluation indicates that ATFPGT-multi outperforms all previous methods. Additionally, to validate the feasibility of multi-task learning, we train individual ATFPGT-single models for each task and evaluate them using various classification metrics, further confirming the reliability and stability of multi-task learning.

The outstanding predictive ability of ATFPGT-multi is primarily attributed to our dataset, which originates from multiple datasets representing various species, ensuring an ample quantity and quality of samples. Next, we employ two different representation

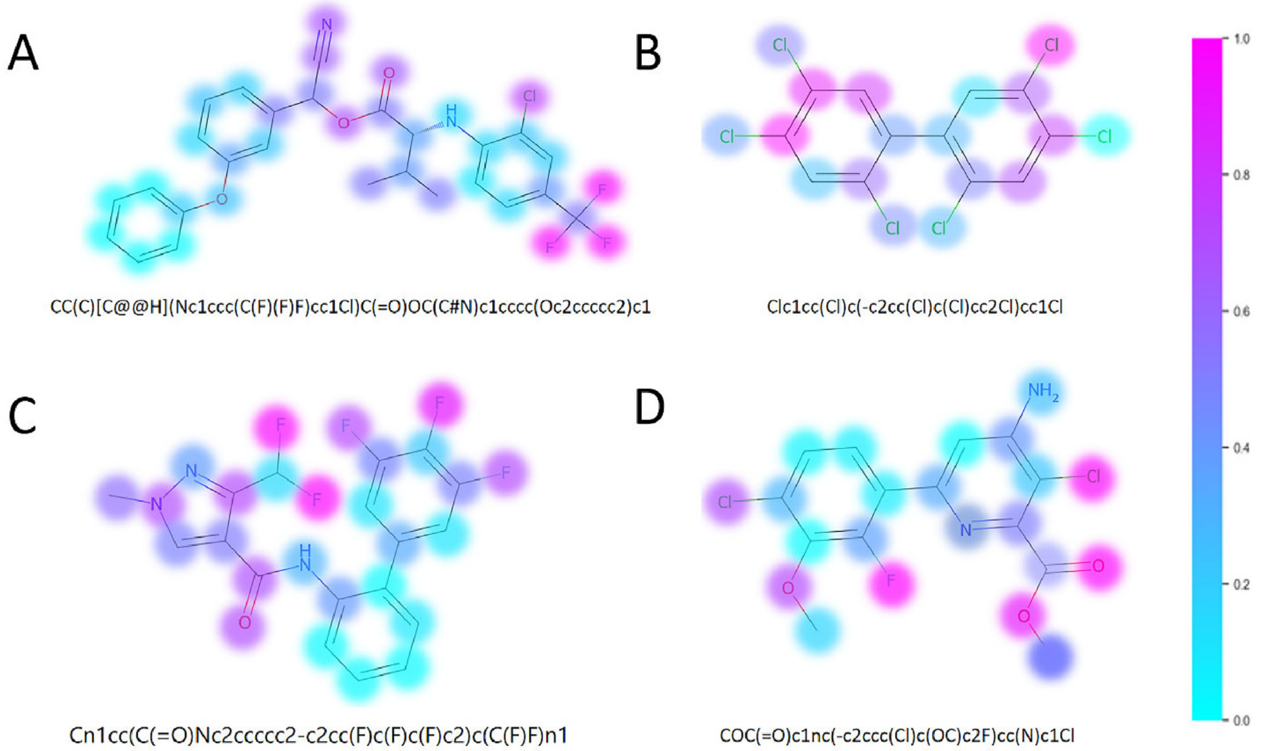

A
CC(C)[C@@H](Nc1ccc(C(F)(F)F)cc1Cl)C(=O)OC(C#N)c1cccc(Oc2ccccc2)c1
B
Clc1cc(Cl)c(-c2cc(Cl)c(Cl)cc2Cl)cc1Cl
C
Cn1cc(C(=O)Nc2ccccc2-c2cc(F)c(F)c(F)c2)c(C(F)F)n1
D
COC(=O)c1nc(-c2ccc(Cl)c(OC)c2F)cc(N)c1Cl
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 6. Visualization of atomic attention weights. (A) The exemplar molecule selected from BS dataset is presented. (B) The exemplar molecule chosen from RT dataset is displayed. (C) The exemplar molecule selected from FHM dataset is presented. (D) The exemplar molecule selected from SHM dataset is presented.

methods and incorporate a global attention mechanism, allowing for the precise capture of toxic motifs and structures within molecules that influence biological activity. Finally, we extract shared features for multiple tasks through two fully connected layers and create separate output layers for each task. This design ensures that our model, in multi-task learning, not only considers the correlations between different tasks but also adequately accounts for their differences. These factors collectively contribute to ATFPGT-multi achieving optimal results in toxicity prediction.

However, ATFPGT-multi still has some limitations and challenges. The performance may be affected during training due to the imbalance in the number of samples in each task. The current dataset only covers compound data from four fish species, while the aquatic biodiversity is extensive, lacking sufficient coverage of compounds from different species. Additionally, given the complexity of the model, it requires higher computational resources and longer training times. Fine-tuning and optimizing the model architecture, parameters, and training process will enhance ATFPGT-multi's predictive performance on a wider range of aquatic toxicity endpoints. Furthermore, further exploration of multitask learning applications is possible. For example, predicting the toxicity of aquatic organisms under different environmental conditions and exploring ATFPGT-multi's transfer learning capabilities will provide us with deeper insights. In addition, our model contributes to the safer development of chemicals and the protection of aquatic ecosystems. First, the accuracy and reliability of ATFPGT-multi provide crucial tools for chemical design and evaluation. By predicting the toxicity of molecules, researchers and decision-makers can identify potential environmental risks earlier and take appropriate measures to mitigate these risks. Second, by gaining a deeper understanding of the relationship between molecular structure and toxicity, ATFPGT-multi can assist in developing safer alternatives or improving existing products, thereby reducing adverse environmental impacts. Overall, in comparison to previous methods, ATFPGT-multi demonstrates the best performance in predicting aquatic toxicity. This not only emphasizes the crucial role of various molecular representation methods but also highlights the advantages of multi-task learning. This outcome holds significant implications for assessing environmental hazards in aquatic ecosystems.

Funding

This work is supported by the Ministry of Science and Technology of the People's Republic of China (STI2030-Major Projects2021ZD0201900), National Natural Science Foundation of China (Grant Nos. 12,090,052 and 11805091), Natural Science Foundation of Liaoning Province (Grant No. 2023-MS-288).

CRediT authorship contribution statement

Xin Yang: Data curation, Investigation, Methodology, Software, Writing – original draft. **Jianqiang Sun:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration. **Bingyu Jin:** Data curation, Investigation, Software. **Yuer Lu:** Data curation, Investigation, Software. **Jinyan Cheng:** Formal analysis, Investigation, Methodology. **Jiaju Jiang:** Software, Visualization, Writing – review & editing. **Qi Zhao:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Jianwei Shuai:** Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by Biomedical Big Data Intelligent Computing Center of Oujiang Lab.

References

[1] Casalegno M, Benfenati E, Sello G. An automated group contribution method in predicting aquatic toxicity: the diatomic fragment approach. Chem Res Toxicol 2005;18:740–6.

[2] Cunha DL, Mendes MP, Marques M. Environmental risk assessment of psychoactive drugs in the aquatic environment. Environ Sci Pollut Res 2018;26:78–90.

[3] Melvin SD, Lanctôt CM, Doriean NJC, Bennett WW, Carroll AR. NMR-based lipidomics of fish from a metal(loid) contaminated wetland show differences consistent with effects on cellular membranes and energy storage. Sci Total Environ 2019;654:284–91.

[4] Zenker A, Cicero MR, Prestinaci F, Bottoni P, Carere M. Bioaccumulation and biomagnification potential of pharmaceuticals with a focus to the aquatic environment. J Environ Manage 2014;133:378–87.

[5] Grabicova K, Grabic R, Fedorova G, Fick J, Cerveny D, Kolarova J, et al. Bioaccumulation of psychoactive pharmaceuticals in fish in an effluent dominated stream. Water Res 2017;124:654–62.

[6] Kullmann L, Habedank F, Kullmann B, Tollkühn E, Frankowski J, Dorow M, et al. Evaluation of the bioaccumulation potential of alizarin red S in fish muscle tissue using the European eel as a model. Anal Bioanal Chem 2020;412:1181–92.

[7] Belanger SE, Lillicrap AD, Moe SJ, Wolf R, Connors K, Embry MR. Weight of evidence tools in the prediction of acute fish toxicity. Integr Environ Assess Manag 2023;19:1220–34.

[8] Schmidt S, Schindler M, Faber D, Hager J. Fish early life stage toxicity prediction from acute daphnid toxicity and quantum chemistry. SAR QSAR Environ Res 2021;32:151–74.

[9] Ankley GT, Villeneuve DL. The fathead minnow in aquatic toxicology: past, present and future. Aquatic Toxicol (Amsterdam, Netherlands) 2006;78:91–102.

[10] Schüürmann G, Ebert R, Kühne R. Quantitative read-across for predicting the acute fish toxicity of organic compounds. Environ Sci Tech 2011;45:4616–22.

[11] Hu H, Feng Z, Lin H, Cheng J, Lyu J, Zhang Y, et al. Gene function and cell surface protein association analysis based on single-cell multiomics data. Comput Biol Med 2023;157:106733.

[12] Hu H, Feng Z, Lin H, Zhao J, Zhang Y, Xu F, et al. Modeling and analyzing single-cell multimodal data with deep parametric inference. Brief Bioinform 2023;24:bbad005.

[13] Meng R, Yin S, Sun J, Hu H, Zhao Q. scAAGA: Single cell data analysis framework using asymmetric autoencoder with gene attention. Comput Biol Med 2023;165:107414.

[14] He Z, Gao K, Dong L, Liu L, Qu X, Zou Z, et al. Drug screening and biomarker gene investigation in cancer therapy through the human transcriptional regulatory network. Comput Struct Biotechnol J 2023;21:1557–72.

[15] Zhang Y, Song C, Zhang Y, Wang Y, Feng C, Chen J, et al. TcoFBase: a comprehensive database for decoding the regulatory transcription co-factors in human and mouse. Nucl Acids Res 2022;50(D1):D391–401.

[16] Guo J, Fang S, Wu Y, Zhang J, Chen Y, Liu J, et al. CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. Nucleic Acids Res 2019;47(W1):W516–22.

[17] Wang T, Sun J, Zhao Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism. Comput Biol Med 2023;153:106464.

[18] Chen Z, Zhang L, Sun J, Meng R, Yin S, Zhao Q. DCAMCP: A deep learning model based on capsule network and attention mechanism for molecular carcinogenicity prediction. J Cell Mol Med 2023;27:3117–26.

[19] Chen Z, Jiang Y, Zhang X, Zheng R, Qiu R, Sun Y, et al. ResNet18DNN: prediction approach of drug-induced liver injury by deep neural network with ResNet18. Brief Bioinform 2021;23(1):bbab503.

[20] Chen Z, Cao Y, He S, Qiao Y. Development of models for classification of action between heat-clearing herbs and blood-activating stasis-resolving herbs based on theory of traditional Chinese medicine. Chin Med 2018;13:1–11.

[21] Chen Z, Zhao M, You L, Zheng R, Jiang Y, Zhang X, et al. Developing an artificial intelligence method for screening hepatotoxic compounds in traditional Chinese medicine and Western medicine combination. Chin Med 2022;17(1):58.

[22] Chen Z, Jiang Y, Zhang X, Zheng R, Qiu R, Sun Y, et al. The prediction approach of drug-induced liver injury: response to the issues of reproducible science of artificial intelligence in real-world applications. Brief Bioinform 2022;23(4):bbac196.

[23] Liu H, Ren G, Chen H, Liu Q, Yang Y, Zhao Q. Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized. Knowl-Based Syst 2020;191:105261.
[24] Zhang L, Yang P, Feng H, Zhao Q, Liu H. Using Network Distance Analysis to Predict lncRNA-miRNA Interactions. Interdisciplinary sciences, computational life sciences 2021;13:535–45.
[25] Wang W, Zhang L, Sun J, Zhao Q, Shuai J. Predicting the potential human lncRNA-miRNA interactions based on graph convolution network with conditional random field. Brief Bioinform 2022;23:bbac463.
[26] Gao H, Sun J, Wang Y, Lu Y, Liu L, Zhao Q, et al. Predicting metabolite-disease associations based on auto-encoder and non-negative matrix factorization. Brief Bioinform 2023;24:bbad259.
[27] Sun F, Sun J, Zhao Q. A deep learning method for predicting metabolite-disease associations via graph neural network. Brief Bioinform 2022;23: bbac266.
[28] Zhu F, Niu Q, Li X, Zhao Q, Su H, Shuai J. FM-FCN: a neural network with filtering modules for accurate vital signs extraction. Research 2024. Article 0361.
[29] Zhu F, Shuai Z, Lu Y, Su H, Yu R, Li X, et al. oBABC: A one-dimensional binary artificial bee colony algorithm for binary optimization. Swarm Evol Comput 2024;87:101567.
[30] Zhu F, Ding J, Li X, Lu Y, Liu X, Jiang F, et al. Systems, MEAs-Filter: a novel filter framework utilizing evolutionary algorithms for cardiovascular diseases diagnosis. Health Information Science and Systems 2024;12:8.
[31] Wang CC, Han CD, Zhao Q, Chen X. Circular RNAs and complex diseases: from experimental results to computational models. Brief Bioinform 2021;22: bbab286.
[32] Zhao Q, Yang Y, Ren G, Ge E, Fan C. Integrating Bipartite Network Projection and KATZ Measure to Identify Novel CircRNA-Disease Associations. IEEE Trans NanoBiosci 2019;18:578–84.
[33] Nguyen-Vo TH, Nguyen L, Do N, Le PH, Nguyen TN, Nguyen BP, et al. Predicting Drug-Induced Liver Injury Using Convolutional Neural Network and Molecular Fingerprint-Embedded Features. ACS Omega 2020;5:25432–9.
[34] Teng S, Yin C, Wang Y, Chen X, Yan Z, Cui L, et al. MolFPG: Multi-level fingerprint-based Graph Transformer for accurate and robust drug toxicity prediction. Comput Biol Med 2023;164:106904.
[35] Gajewicz-Skretna A, Kar S, Piotrowska M, Leszczynski J. The kernel-weighted local polynomial regression (KwLPR) approach: an efficient, novel tool for development of QSAR/QSAAR toxicity extrapolation models. J Cheminf 2021;13:9.
[36] Rajabi M, Shafiei F. QSAR models for predicting aquatic toxicity of esters using genetic algorithm-multiple linear regression methods. Comb Chem High Throughput Screen 2019;22:317–25.
[37] Cao Q, Liu L, Yang H, Cai Y, Li W, Liu G, et al. In silico estimation of chemical aquatic toxicity on crustaceans using chemical category methods. Environ Sci Processes Impacts 2018;20:1234–43.
[38] Liu L, Yang H, Cai Y, Cao Q, Sun L, Wang Z, et al. In silico prediction of chemical aquatic toxicity for marine crustaceans via machine learning. Toxicol Res 2019;8:341–52.
[39] Ai H, Wu X, Zhang L, Qi M, Zhao Y, Zhao Q, et al. QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. Ecotoxicol Environ Saf 2019;179:71–8.
[40] Li X, Liu G, Wang Z, Zhang L, Liu H, Ai H. Ensemble multiclassification model for aquatic toxicity of organic compounds. Aquatic Toxicol (Amsterdam, Netherlands) 2023;255:106379.
[41] Wei L, Ye X, Xue Y, Sakurai T, Wei L. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. Brief Bioinform 2021;22:bbab041.
[42] Xu M, Yang H, Liu G, Tang Y, Li W. In silico prediction of chemical aquatic toxicity by multiple machine learning and deep learning approaches. J Appl Toxicol 2022;42:1766–76.
[43] Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, et al. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. J Chem Inf Model 2012;52:3099–105.
[44] Li F, Fan D, Wang H, Yang H, Li W, Tang Y, et al. In silico prediction of pesticide aquatic toxicity with chemical category approaches. Toxicol Res 2017;6:831–42.
[45] Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, et al. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties, Bioinformatics (Oxford, England), 2019;35: 1067-69.
[46] Joaquin BF, Heidi O, Birgit S. European List of Notified Chemical Substances-In support of Directive 92/32/EEC, the 7th amendment to Directive 67/548/EEC.
[47] Zhong S, Guan X. Count-based Morgan fingerprint: a more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties. Environ Sci Tech 2023;57:18193–202.
[48] Yin Z, Ai H, Zhang L, Ren G, Wang Y, Zhao Q, et al. Predicting the cytotoxicity of chemicals using ensemble learning methods and molecular fingerprints. J Appl Toxicol 2019;39:1366–77.
[49] Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. J Cheminf 2020;12:51.
[50] Tan Z, Zhao Y, Zhou T, Lin K. Hi-MGT: A hybrid molecule graph transformer for toxicity identification. J Hazard Mater 2023;457:131808.
[51] Yun S, Jeong M, Yoo S, Lee S, Yi SS, Kim R, et al. Graph transformer networks: learning meta-path graphs to improve GNNs. Neural networks: Off J Int Neural Network Soc 2022;153:104–19.
[52] Toropov AA, Di Nicola MR, Toropova AP, Roncaglioni A, Dorne J, Benfenati E. Quasi-SMILES: Self-consistent models for toxicity of organic chemicals to tadpoles. Chemosphere 2023;312:137224.
[53] Gao J, Shen Z, Xie Y, Lu J, Lu Y, Chen S, et al. TransFoxMol: predicting molecular property with focused attention. Brief Bioinform 2023;24:bbad306.
[54] Johnston NR, Strobel SA. Principles of fluoride toxicity and the cellular response: a review. Arch Toxicol 2020;94:1051–69.
[55] Fang Y, Zhang Q, Yang H, Zhuang X, Deng S, Zhang W, et al. Molecular contrastive learning with chemical element knowledge graph. Proc AAAI Conf Artif Intel 2022;36: 3968–76.