Computers in Biology and Medicine 164 (2023) 106904

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

MolFPG: Multi-level fingerprint-based Graph Transformer for accurate and robust drug toxicity prediction

Saisai Teng [a,b], Chenglin Yin [a,b], Yu Wang [a,b], Xiandong Chen [c], Zhongmin Yan [a,b,**], Lizhen Cui [a,b,***], Leyi Wei [a,b,*]

[a] School of Software, Shandong University, Jinan, China
[b] Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, China
[c] Shaoxing Healthcare Security Bureau, China

ARTICLE INFO

Keywords:
Molecular fingerprint
Graph transformer
Drug toxicity prediction

ABSTRACT

Drug toxicity prediction is essential to drug development, which can help screen compounds with potential toxicity and reduce the cost and risk of animal experiments and clinical trials. However, traditional handcrafted feature-based and molecular-graph-based approaches are insufficient for molecular representation learning. To address the problem, we developed an innovative molecular fingerprint Graph Transformer framework (MolFPG) with a global-aware module for interpretable toxicity prediction. Our approach encodes compounds using multiple molecular fingerprinting techniques and integrates Graph Transformer-based molecular representation for feature learning and toxic prediction. Experimental results show that our proposed approach has high accuracy and reliability in predicting drug toxicity. In addition, we explored the relationship between drug features and toxicity through an interpretive analysis approach, which improved the interpretability of the approach. Our results highlight the potential of Graph Transformers and multi-level fingerprints for accelerating the drug discovery process by reliably, effectively alarming drug safety. We believe that our study will provide vital support and reference for further development in the field of drug development and toxicity assessment.

1. Introduction

Drug development is a time-consuming and expensive process. Identifying compounds with potential toxicity in the early stages through toxicity prediction can avoid significant investment in such compounds and improve the efficiency of drug development [1–3]. Furthermore, accurate prediction and assessment of potential drug toxicity help to minimize negative consequences on human health during clinical trials and post-market release [4–6]. However, these traditional experimental procedures involve many trials and tests and have a high consumption rate, a lengthy cycle, and a low economic return [7,8]. Thus, the artificial intelligence-based drug toxicity prediction approach has become particularly popular [9–12], as it may assist researchers in evaluating the toxicity of pharmaceuticals more rapidly and precisely, thereby allowing for greater control over the safety and efficacy of drugs [13–15]. In addition, AI-based drug toxicity prediction [16,17] can assist drug experts in better identifying the pharmacodynamic and pharmacokinetic features of medications, hence expediting the development of novel drugs [18–21].

From the perspective of molecular representations, existing artificial intelligence-based drug toxicity prediction approaches can be generally categorized into two classes: (1) Handcrafted features-based approaches; and (2) Graph-based drug toxicity prediction approaches. Handcrafted features-based approaches include primarily chemical descriptor-based and molecular fingerprint-based features approaches. The former requires expressing the structure and chemical properties of drug molecules with numerical descriptors, followed by the application of machine learning algorithms to predict the toxicity of drugs. Chemical descriptors may include molecular weight, lipid solubility, hydrophilicity, charge, etc [22]. The latter is to represent the structure and features of the drug molecule with a binary string, where each string refers to a structure or feature, and if the drug molecule has the structure or feature, the corresponding binary bit is 1; otherwise, it is 0 [23]. Molecular representation based on handcrafted features relies on

* Corresponding author. School of Software, Shandong University, Jinan, China.
** Corresponding author. School of Software, Shandong University, Jinan, China.
*** Corresponding author. School of Software, Shandong University, Jinan, China.
E-mail addresses: yzm@sdu.edu.cn (Z. Yan), clz@sdu.edu.cn (L. Cui), weileyi@sdu.edu.cn (L. Wei).

https://doi.org/10.1016/j.compbiomed.2023.106904
Received 23 February 2023; Received in revised form 20 March 2023; Accepted 10 April 2023
Available online 14 May 2023

traditional approaches such as Random Forest (RF) [24,25], eXtreme Gradient Boosting (XGBoost) [26,27], Support Vector Machine (SVM) [28–31], and deep neural networks (DNNs) [32–35] and so on [36–39]. However, the above approaches have two limitations. First, molecular fingerprints and chemical descriptors are representations of chemical structures. Although containing certain chemical information, they often cannot capture more complex structural features and interactions in molecules. Second, it is challenging to analyze the interpretability of the prediction results using this approach, making it difficult to recognize this approach in the actual application of drug development. Finally, the approach's inability to predict properties of molecules that have not been previously encountered in the training data may be particularly relevant when predicting rare or novel chemical structures.

Graph-based approaches represent the molecular structure as a graph and learn the feature representation of the drug's molecular structure using graph representation learning algorithms [40–42]. Compared to traditional approaches [43] based on molecular fingerprints and chemical descriptors, graph representation-based approaches can better capture the structure and relationships between drugs, enhancing the accuracy and generalizability of predictions. In recent years, drug toxicity prediction approaches based on graph representations have achieved good performance in some related fields [32,44,45]. For instance, researchers have employed graph neural networks (GNNs)-based [46] approaches to predict several types of medication toxicity and achieved good prediction results. K. M. Quinn et al. [44] proposed using molecular graph representation and graph convolutional neural network (GCN) to predict the products and by-products of chemical reactions with high prediction accuracy. C. Chen et al. [47] introduced a drug molecular toxicity prediction approach based on graph representation and Bayesian machine learning, integrating molecular graph representation with chemical descriptors to predict the likelihood that drug molecules will cause liver harm. Although existing graph-based approaches have achieved significant progress in accelerating data-driven toxicity prediction, they still suffer from the following intrinsic problems: (1) Graph-based approaches require large-scale datasets, and it is challenging to acquire robust molecular representations from insufficient datasets using the existing GNNs. (2) Traditional graph network-based toxicity prediction approaches only consider local structures, which can capture the structural information of molecules to a certain extent but lack a global perspective and ignores the role of long-range atoms. (3) The information captured by graph-based and handcrafted feature-based molecular representations is distinct or even complementary, and existing approaches have not explored the optimal combination.

In this study, we proposed an innovative molecular fingerprint Graph Transformer framework named MolFPG with global awareness for interpretable toxicity prediction. To learn more robust and information-rich molecular representations, we integrate fingerprint-based and graph-based representations and propose a novel architecture of Graph Transformer. The robustness of our approach can be attributed to the following aspects: (1) we utilized multiple molecular fingerprinting strategies to encode compounds, allowing our approach to consider the attributes of drug molecules more comprehensively; (2) we utilized the Graph Transformer architecture to represent molecular graph data, which has a powerfully expressive and learning capability to learn different types of features adaptively; (3) we added a global-aware attention mechanism to make the framework more robust. Our results demonstrate that the proposed approach outperforms several state-of-the-art models and has the potential to become a valuable tool for drug development and toxicity assessment.

2. Materials and methods

2.1. Dataset

In this study, we utilized two distinct toxicity datasets, one for classification and the other for regression, to more comprehensively assess the performance of our proposed approach in terms of its generalizability across different types of toxicity prediction tasks. For the classification task, we selected the Ames Mutagenicity dataset [48], which is used to predict the mutagenic potential of chemical compounds. For the regression task, we selected the Acute Toxicity LD50 dataset [49], which is used to predict the acute toxicity level of compounds to experimental animals.

Table 1 shows an overview of the sample characteristics for two datasets, including task types, number of molecules, partitioning methods, and evaluation metrics. Furthermore, to investigate the distribution of the number of atoms within molecules, density distribution graphs were plotted, as shown in Fig. 1. Notably, the number of atoms in both datasets followed a normal distribution, and statistical information such as mean, variance, and peak value could be derived from the distribution graphs. These findings offer valuable insights into the structure and chemical properties of the molecules and have significant implications for further research in the field.

2.2. Problem definition

We consider a problem of drug toxicity prediction, given a training set $\mathscr{D} = (\mathbf{x}_i, y_i) * i = 1^n$ of n drug molecules, where $\mathbf{x}_i$ represents the concatenated fingerprint and graph representation of the i-th drug molecule, and y_i represents its true toxicity value. We consider two different tasks in this problem: classification and regression. For the classification task, our goal is to classify each drug molecule as having either high or low toxicity, i.e., map $f(\mathbf{x})$ to a binary classification result. For the regression task, our goal is to predict the specific toxicity value of each drug molecule, i.e., map $f(\mathbf{x})$ to a real number. We use cross-entropy as the loss function for the classification task and mean squared error as the loss function for the regression task. Therefore, our objective is to minimize the loss function and learn a more accurate toxicity prediction model.

2.3. The architecture of MolFPG

Fig. 2 illustrates the overall architecture of our MolFPG. Our proposed MolFPG framework can be divided into four main parts: data preprocessing, molecular fingerprint representation, molecular graph representation, and toxicity prediction. The workflow of MolFPG is described as follows:

During the data preprocessing stage, molecules are transformed into three different types of fingerprints, namely Morgan, Maccs and RDKit fingerprints [50], which are concatenated together as the input of the molecular fingerprint representation module. Additionally, atoms are regarded as nodes and bonds as edges to construct the molecular topology graph, which serves as the input of the molecular graph representation module.

In the molecular fingerprint representation module, we utilize a Multi-layer Perceptron (MLP) network [51] to extract and represent features from three different types of fingerprints. The MLP network generates more abstract and informative representations of molecular fingerprints, which encode multi-level features of molecules and support downstream toxicity prediction.

In the molecular graph representation module, we introduce a newly-designed global-aware Graph Transformer module to encode molecular graphs for more robust molecular representations. This module integrates the advantages of GNNs [52] and Transformers [53],

Table 1
Sample characteristics for AMES and LD50 datasets.

Dataset	Data type	Task type	Number of drugs	Split	Metric
AMES	SMILES	1	7278	Scaffold	AUROC
LD50	SMILES	1	7385	Scaffold	RMSE

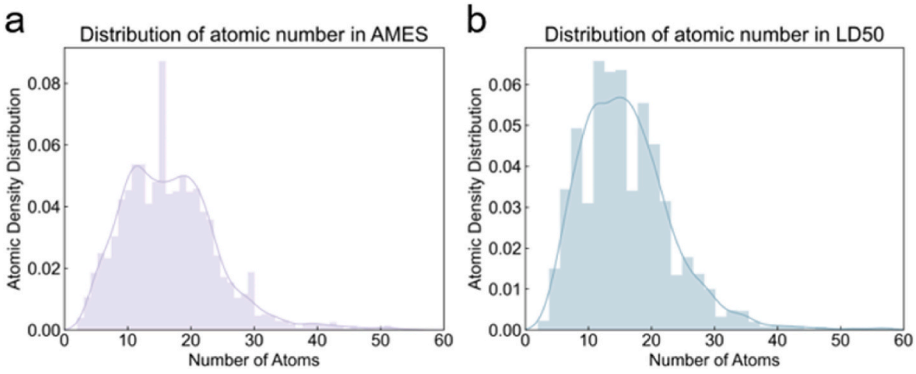

Fig. 1. Density distribution of the number of atoms for AMES and LD50 datasets.

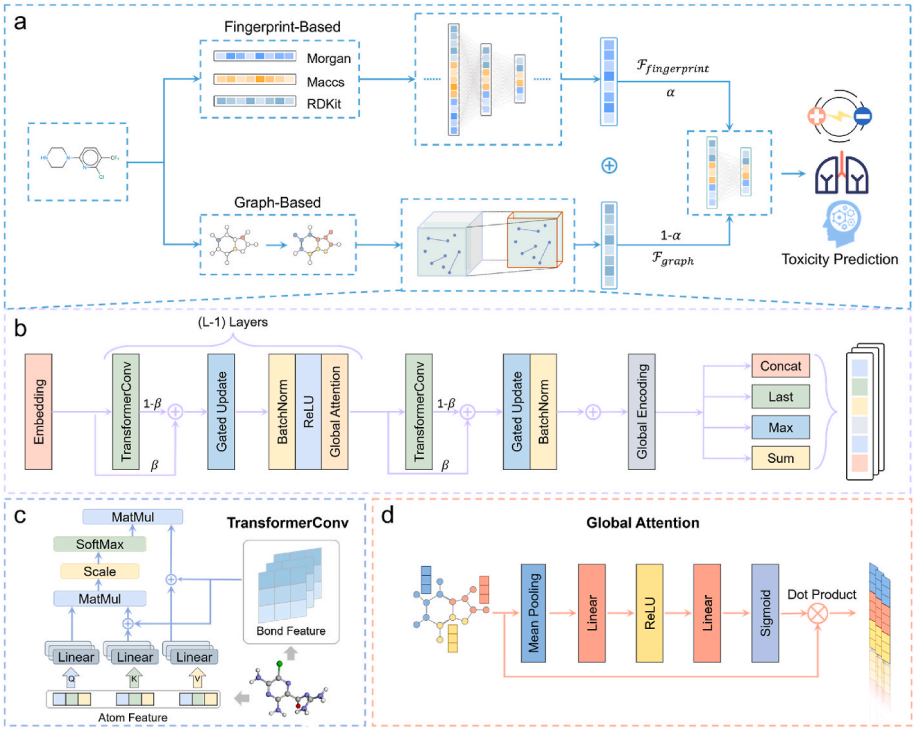

Fig. 2. The overview of the MolFPG framework. (a) Illustrates the overall workflow of the MolFPG architecture. (b) The architecture of the global attention Graph Transformer module, which presents the graph-based representation approach. (c) shows the convolutional details based on the scaled dot-product attention. (d) describes the process of our global attention module.

considering the global information of the molecular structure and having a strong adaptive ability to better capture the differences and complexities among different molecules. The combination of these two modules provides a powerful tool for accurately and efficiently predicting the toxicity of chemicals.

In the toxicity prediction stage, we employ an adaptive attention mechanism to balance the importance of molecular fingerprint representation and molecular graph representation. The adaptive attention mechanism is a dynamic weight allocation process that adaptively adjusts the weight of each representation based on the feature information of each molecule, enabling the model to better capture the molecular feature information.

2.4. Molecular fingerprint encoding module

In our work, we employed three different fingerprint encoding techniques, namely Morgan fingerprints, Maccs fingerprints, and RDKit fingerprints, to represent the molecular structure information. Each encoding technique generates a fixed-length binary vector representation for each molecule, with each bit in the vector corresponding to a unique molecular substructure [54]. The presence or absence of a substructure in a compound determines the value of the corresponding bit in its fingerprint vector. The Morgan fingerprint captures local substructures around each atom in the molecule by generating circular substructures up to a specified radius [55]. In contrast, the Maccs fingerprint uses a predefined set of keys to encode molecular substructures [56]. The RDKit fingerprint algorithm generates a fingerprint by hashing the paths in the molecular graph up to a specified length [57]. The resulting bit vectors provide a compact representation of molecular structures that be used as input for molecular fingerprint encoding modules in toxicity prediction tasks.

Formally, let X be a dataset of drugs with molecular structures represented as SMILES strings. For each molecular x in X, we apply each fingerprint encoding method to obtain three binary vector representations: Morgan fingerprint $M(x)$, Maccs fingerprint $A(x)$, and RDKit fingerprint $R(x)$. The concatenation of these three fingerprint vectors

forms the input representation for our toxicity prediction model:

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

where $Z(x)$ is a fixed-length binary vector of length L, and each component of $Z(x)$ represents the presence or absence of a unique molecular substructure in the compound x.

To further process the input fingerprint vector $Z(x)$, we apply a multi-layer perceptron (MLP) with ReLU activation functions to learn a non-linear mapping from the input space to a high-dimensional feature space. The output of the MLP is then passed to our feature fusion module for toxicity prediction. Formally, let $f(Z(x);\ W)$ be the MLP function with weight parameters W, and $h(x)$ be the output of the MLP for a given compound x. The MLP function is defined as:

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

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 $ReLU(z) = max(0, z)$. The output $h(x)$ of the *MLP* is given by $h(x) = f(Z(x);\ W)$, and represents the learned high-dimensional feature representation of the drug x. The feature representation, which captures the intricate and non-linear interactions among diverse molecular substructures within the drugs, is indispensable in accurately predicting the toxicity of the compound.

2.5. Global attention Graph Transformer encoding module

In the Graph Transformer architecture proposed in our study, we first transform each node feature vector $\mathbf{x}_i$ with two weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, and the attention coefficients $\alpha_{i,j}$ are computed via multi-head dot product attention using $\mathbf{W}_3$ and $\mathbf{W}_4$. To facilitate the aggregation of information from neighboring nodes and update the node's feature representation accordingly, our model adopts a message-passing mechanism. Specifically, the transformed feature vector $\mathbf{x}_i^{'}$ for node i is given by:

$$\mathbf{x}_i^{'} = \beta_i \mathbf{W}_1 \mathbf{x}_i + (1 - \beta_i)\left(\sum_{j \in \mathcal{N}(i)} \alpha_{i,j}\left(\mathbf{W}_2 \overrightarrow{x}_j + \mathbf{W}_6 \mathbf{e}_{ij}\right)\right) \tag{3}$$

where the attention coefficients $\alpha_{i,j}$ are computed as follows:

$$\alpha_{i,j} = \mathrm{softmax}\left(\frac{(\mathbf{W}_3 \mathbf{x}_i)^{\top}\left(\mathbf{W}_4 \mathbf{x}_j + \mathbf{W}_6 \mathbf{e}_{ij}\right)}{\sqrt{d}}\right) \tag{4}$$

and β_i is computed using sigmoid function applied to the concatenation of three vectors:

$$\beta_i = \mathrm{sigmoid}\left(\mathbf{w}_5^{\top}[\mathbf{W}_1 \mathbf{x}_i, \mathbf{m}_i, \mathbf{W}_1 \mathbf{x}_i - \mathbf{m}_i]\right) \tag{5}$$

where $\mathbf{m}_i$ is a weighted sum of the Transformer feature vectors of i's neighbors, with the weight coefficients given by the attention coefficients $\alpha_{i,j}$ and the weight matrix $\mathbf{W}_6$. The computation of $\mathbf{m}_i$ is as follows:

$$\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}\left(\mathbf{W}_2 \mathbf{x}_j + \mathbf{W}_6 \mathbf{e}_{ij}\right) \tag{6}$$

where $\mathbf{e}_{ij}$ is the edge feature vector between node i and node j. Overall, the message-passing mechanism enables each node to aggregate information from its neighbors and update its feature representation accordingly, which can improve the performance of our toxicity prediction task.

In the representation of molecular structures, local features are not the only crucial factors for property prediction, but also the interactions between different atoms in a molecule. To better utilize global information, we introduce a novel global attention mechanism that captures global features by performing mean-pooling on the features of all atoms, and computes attention weights for each atom using Eq. (7):

$$\mathrm{Attn}_{global} = \mathrm{sigmoid}\left(\mathrm{w}_8^{\top}\mathrm{ReLU}\left(\mathrm{w}_7^{\top}\mathrm{rep}_{mean} + \mathrm{b}_1\right) + b_2\right) \tag{7}$$

where rep_{mean} indicates the representation of the molecule by mean pooling. This global attention mechanism not only enhances the capability of modeling atom interactions but also comprehensively captures complex properties in molecular structures, further improving the accuracy and robustness of our molecular property prediction model.

Overall, our Graph Transformer framework enables each node to efficiently aggregate and integrate information from its neighbors, resulting in a refined feature representation that is more informative for our toxicity prediction task. Specifically, Eq. (3) computes the updated feature representation of a node as a weighted combination of its features and the aggregated features from its neighbors. The weights used for aggregation are computed using Eq. (4), which allows for effective modeling of complex relationships between nodes. Moreover, the balance between the node's features and the aggregated features from its neighbors is controlled by the sigmoid function, which is defined in Eq. (5). This effectively enables the node to weigh the relative importance of its features and those of its neighbors in the prediction task. This approach effectively captures local interactions and dependencies between nodes in the molecular structure, leading to improved predictive performance.

2.6. Feature fusion module and toxicity prediction

Our toxicity prediction model leverages a feature fusion module to effectively integrate molecular fingerprint encoding and graph-based feature representation. To generate the final toxicity prediction score for the molecule, we employ an attention mechanism to appropriately weigh the contributions of both types of representations [58]. Specifically, we first compute the attention weight as follows:

$$\mathrm{Attn}_i = \mathrm{sigmoid}\left(\mathbf{w}_1^{\top}[\mathrm{G}_i, \mathrm{F}_i, \mathrm{G}_i - \mathrm{F}_i]\right) \tag{8}$$

Here, G_i denotes the graph-based feature representation of node i, and F_i represents the feature obtained from the molecular fingerprint encoding module. We then calculate the final toxicity prediction score for the entire molecule as a weighted sum of the feature representations of each node, where the weight for each node is given by its attention weight:

$$\mathbf{Tox}_i = mlp(\mathrm{Attn}_i \mathrm{F}_i + (1 - \mathrm{Attn}_i) G_i) \tag{9}$$

Overall, our MolFPG framework integrates the technologies of molecular fingerprint and molecular graph representation, enabling more comprehensive and accurate feature learning and prediction through the synergistic action of the fingerprint encoding module and the molecular graph encoding module. Molecular fingerprint technology captures various chemical features of molecules, such as 3D spatial information, specific rings, and functional groups, while molecular graph representation describes the topological structure information of molecules, such as chemical bond types, molecular conformation, groups, and functional groups. By combining the fingerprint-based representation from the molecular fingerprint encoding module with the graph-based feature representation learned by the Graph Transformer architecture, our feature fusion module enables a more comprehensive analysis of molecular features associated with toxicity. This synergistic fusion of complementary representations leads to improved predictive performance and a deeper understanding of toxicity's complex mechanisms.

2.7. Experiment setup

In the experimental settings of MolFPG, we utilized the PyTorch framework and carefully designed and selected parameters to ensure the reliability and accuracy of the experimental results. In terms of optimizer function selection, we considered the differences among experimental tasks and attempted various optimizer functions, including

Adam, Adagrad, and RMSprop. We also conducted a grid search on the learning rate and regularization coefficient, with a learning rate range of {1e-5, 5e-5, ….., 1e-3, 5e-3} and a regularization coefficient range of {1e-6, 1e-5, 1e-4}, adjusting and attempting them according to the differences among experimental tasks. For the selection of batch size, we considered the size and complexity of the dataset and conducted multiple experiments within the range of {32, 64, 128} to determine the optimal batch size. We set the embedding dimension size of all baseline models to 300 to ensure the fairness of the experimental results.

In terms of loss function, we set different loss functions for different experimental tasks, using binary cross-entropy (BCE) loss function for classification tasks and mean squared error (MSE) loss function for regression tasks. To prevent overfitting, we employed early stopping techniques, which helped to stop training after reaching a certain performance level. In our experiments, we employed an NVIDIA GeForce RTX 3090 graphics card, which features 24 GB of memory and 10,496 CUDA cores. Furthermore, we recommend a minimum of 64 GB system memory to accommodate the model's requirements. For researchers with limited resources, it is suggested to explore options such as reducing model size, lowering training complexity, or utilizing pre-trained models to optimize the use of the MolFPG model.

3. Results

3.1. Performance results of MolFPG on benchmark datasets

We evaluated the performance of MolFPG using two toxicity datasets for drug discovery, which included classification and regression tasks. To better evaluate the model's generalization ability on out-of-distribution samples, we followed previous works [59] and employed a scaffold splitting strategy [60] to partition the training, validation, and test datasets in an 8:1:1 ratio. The scaffold-based splitting is a more challenging approach that partitions molecules based on their substructures [61]. By doing so, we aim to provide a more rigorous evaluation of the model's performance in handling diverse molecular structures.

As illustrated in Table 2, in the AMED classification dataset, MolFPG achieves the best AUC of 0.868 and 0.851, surpassing the previous best-performing method by 2.6% and 1.7%. The accomplishment highlights the significant advantage of MolFPG in handling classification tasks, which can be attributed to its effective feature extraction and representation learning techniques, as well as its unique algorithmic design. Additionally, in the LD50 regression dataset, MolFPG again outperforms the aforementioned models, as evidenced by a significantly lower RMSE of 0.935 and 0.697. The exceptional performance of MolFPG in regression tasks likely stems from its in-depth understanding of and improved feature capture abilities for complex chemical structure data. Furthermore, MolFPG's superior performance can be attributed to its unique ability to effectively capture multi-level information. This is accomplished through the fusion of multiple molecular fingerprint information and the integration of global-wise attention in the Graph Transformer module. In comparison to other models, our approach enables MolFPG to more accurately and comprehensively capture the nuances of molecular properties, resulting in superior predictive performance. Overall, our results demonstrate the superiority of MolFPG over existing models in predicting molecular toxicity for classification and regression tasks. These findings have significant implications for the pharmaceutical industry, as they highlight the potential of MolFPG for accelerating drug discovery and reducing the need for animals.

Table 2
Overall performance of MolFPG and state-of-the-art methods.

Methods	AMES (ROC–AUC)		LD50 (RMSE)	
	Valid	Test	Valid	Test
RF [24]	0.812	0.803	1.264	0.823
XGBoost [26]	0.765	0.763	1.282	0.902
Attentive FP [62]	0.846	0.825	1.12	0.879
GCN [63]	0.832	0.832	1.046	0.792
GraphSAGE [64]	0.823	0.835	1.019	0.827
GAT [65]	0.845	0.833	1.067	0.778
GIN [46]	0.846	0.837	1.028	0.794
MolFPG	**0.868**	**0.851**	**0.935**	**0.679**

3.2. Ablation studies of MolFPG

3.2.1. Analysis of ablation studies on MolFPG architecture components

To investigate the contribution of various components in the MolFPG architecture on its performance, a series of ablation studies were conducted. Specifically, we employed variants of MolFPG as follows: (1) removing the molecular fingerprint encoding module (w/o FP), (2) removing the global attention Graph Transformer encoding module (w/o GT), and (3) Using all of the components as a baseline for comparison of ablation experiments (MolFPG).

As shown in Fig. 3a, the results of our ablation studies show that each component of MolFPG contributes significantly to its overall performance in predicting drug toxicity. Specifically, removing the molecular fingerprint encoding module (w/o FP) led to a substantial drop in performance, indicating that the information encoded in fingerprints plays a crucial role in capturing the nuances of molecular properties. Similarly, removing the global attention Graph Transformer encoding module (w/o GT) also resulted in a significant decrease in performance, highlighting the importance of this module in capturing global dependencies and effectively fusing information from different regions of the molecular graph. The ROC curves of the AMES classification dataset and the scatter plot of the regression dataset in Fig. 4 also support the above observation.

Moreover, our ablation experiments demonstrate the effectiveness of the proposed multi-level fusion approach in MolFPG. By comparing the performance of MolFPG with the variants that lack one of the two encoding modules, we observe that MolFPG, which leverages both molecular fingerprint encoding and global attention Graph Transformer encoding, achieves the best performance. It indicates that the combination of these two modules leads to a more comprehensive and accurate representation of molecular properties, enabling MolFPG to better capture the underlying patterns in drug toxicity. Overall, our ablation studies demonstrate the effectiveness and importance of each component in the MolFPG architecture, demonstrating its potential for predicting molecular properties in drug discovery.

3.2.2. Exploring the effects of different types of molecular fingerprints

In order to explore the impact of different types of molecular fingerprints on MolFPG's performance in predicting drug toxicity, we evaluated the performance of MolFPG using seven different combinations of molecular fingerprints. As illustrated in Fig. 3b, the Maccs fingerprint exhibits slightly inferior performance compared to the other two types of fingerprints when used alone, and its combination with the other two fingerprints also yields inferior results. This may be attributed to the relatively small number of bit positions in the Maccs fingerprint, resulting in the inadequate encoding of molecular substructure information. However, when Morgan, Maccs, and RDKit fingerprints are used in combination, the resulting model shows improved performance, indicating the complementary nature of these fingerprints in capturing diverse molecular features. Specifically, the Morgan fingerprint, which encodes circular substructures, can capture the 3D spatial information of molecules and their local chemical environments, while the RDKit fingerprint, which encodes path-based substructures [50], is effective in capturing the molecular topology and bond connectivity. In contrast to the Morgan and RDKit fingerprints, the Maccs fingerprint is based on a predefined set of substructural keys representing specific molecular features [56]. Overall, our results indicate that a multi-fingerprint approach is advantageous for predicting drug toxicity, with the Morgan + Maccs + RDKit combination being the most effective.

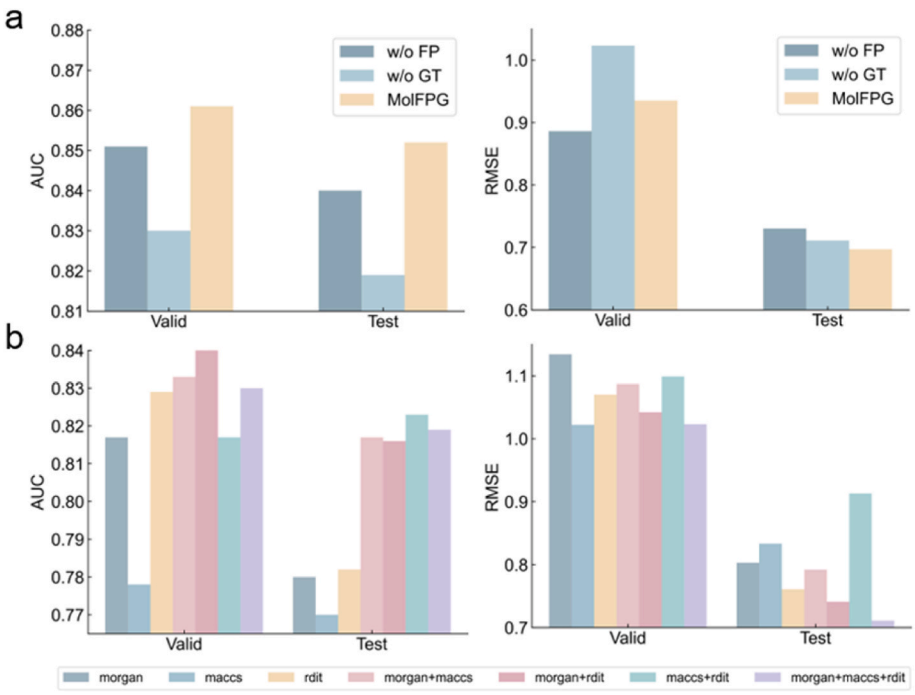

Fig. 3. Ablation results on the variants of MolFPG. (a) The AUC results on AMES datasets. The larger the value, the better the effect (left). The RMSE results on LD50 datasets. The smaller the value, the better the effect (right). (b) The influence of various molecular fingerprint combinations on performance. On the left is the AMES dataset, and on the right is the LD50 dataset.

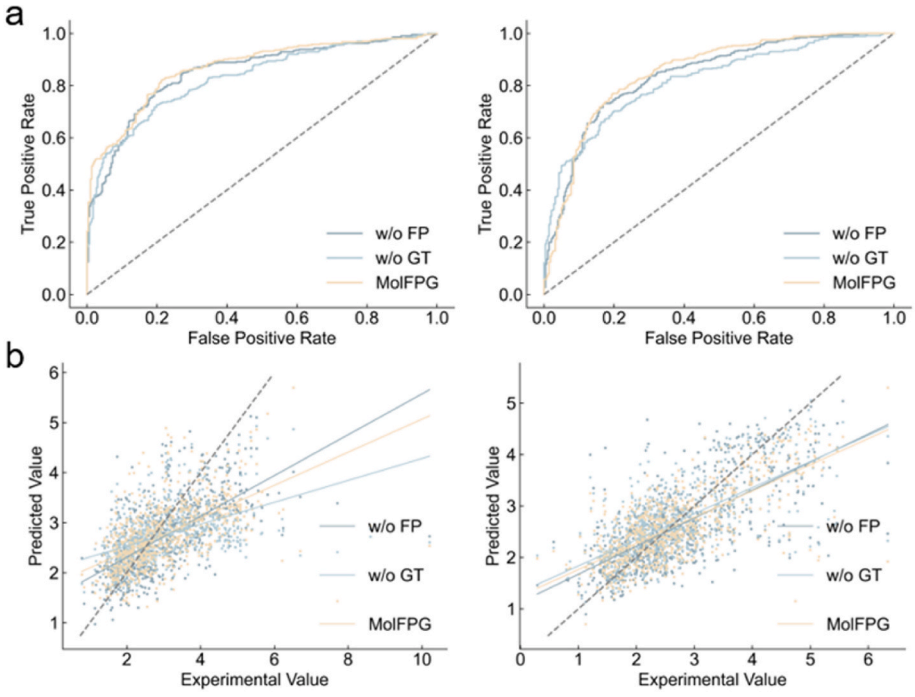

Fig. 4. Ablation results on the variants of MolFPG. (a) The ROC curves for MolFPG and its variants on both the AMES validation dataset (left) and the AMES test dataset (right). (b) The scatter plots of LD50 regression with the left representing the validation dataset and the right representing the test dataset. The folded line represents the trend line of the scattered points.

3.3. Evaluating the robustness of MolFPG

After exploring the efficacy of each component in the MolFPG framework, the robustness of the architecture was evaluated, which is a crucial factor for its broad applicability. To assess the robustness of the MolFPG architecture, the PASP method proposed by Li et al. [66] was employed to perturb the LD50 dataset. The perturbation technique, based on molecular fingerprint similarity calculations, modifies molecular substructures in a natural manner, without breaking the chemical rules within the molecule. Based on this, we can measure the robustness of the model by inputting both the original molecules and perturbed molecules into the prediction model and analyzing the differences in the

output. Specifically, we constructed three levels of perturbed molecules by setting different thresholds, with molecular structure similarity thresholds of 0.8–1, 0.5–0.8, and 0.3–0.5, and differences in molecular properties less than 0.2. Finally, from the LD50 dataset, consisting of 7385 molecules, 1230 molecules meeting the criteria were selected. The effect score, which quantifies the robustness of the model, was obtained by calculating the RMSE between the model predictions of the original and perturbed molecules, as well as the difference in the properties of the original and perturbed molecules.

As illustrated in Table 3, compared to traditional fingerprint-based models and graph-based models, the effect score of the MolFPG framework was lower. This may be attributed to the integration of three types of fingerprint features and structural information based on molecular topology graphs, resulting in an architecture that is highly tolerant to perturbations in molecular structure. In conclusion, the robustness evaluation of the MolFPG architecture demonstrates its high tolerance to changes in molecular structure, highlighting its potential for real-world applications. The integration of multiple molecular fingerprints and graph-based structural information provides a more comprehensive understanding of the underlying mechanisms of drug toxicity.

3.4. Evaluating the interpretability of MolFPG

Traditional molecular toxicity prediction models often lack transparency, making it challenging to understand the decision-making process and the relationship between molecular structure and properties [22,67,68]. To address this challenge, we conducted interpretability case studies on both the AMES and LD50 datasets using the MolFPG model. Specifically, we randomly selected three drug molecules with prediction accuracy up to 99% in the AMES dataset and calculated atomic attention weights based on their hidden layer states. The attention weights can be regarded as a measure of the relevance of the atomic properties to the molecule. Furthermore, due to the use of multi-layer graph convolutions in our framework, the attention weights of each atom also contain information about its surrounding neighbors. Similarly, in the LD50 regression dataset, we selected three molecules with complete agreement between the predicted and experimental values. Please note that as the molecular fingerprint-based models are unable to obtain atomic-level representations, our experiment was conducted on the molecular graph-based module.

As shown in Fig. 5, the results demonstrate that our MolFPG model can assign high attention weights to nitrogen heterocyclic groups, unstable radicals, and oxidative groups that impact molecular toxicity in the AMES results. Similarly, in the LD50 dataset results, it can be clearly observed that our model accurately recognizes toxic groups containing sulfur. By delving deeper into these results, we can derive several noteworthy observations. Firstly, the MolFPG model exhibits a remarkable capability in discerning the significance of distinct functional groups impacting toxicity, which may differ across datasets. This adaptability implies that the model holds promise in a wide array of toxicity prediction tasks. Secondly, the attention mechanism employed by the MolFPG model facilitates the elucidation of relationships between molecular substructures and their corresponding properties, shedding light on the fundamental molecular mechanisms governing toxicity. Lastly, the interpretability proffered by the model's attention weights serves as an aid for researchers and medicinal chemists in the identification of toxophores or pharmacophores, ultimately guiding the development of safer and more efficacious drugs. Through the in-depth analysis, we augment our comprehension of the MolFPG model's underlying mechanisms and further demonstrate its potential in providing invaluable insights into the interplay between molecular structure and properties within the realm of drug toxicity prediction.

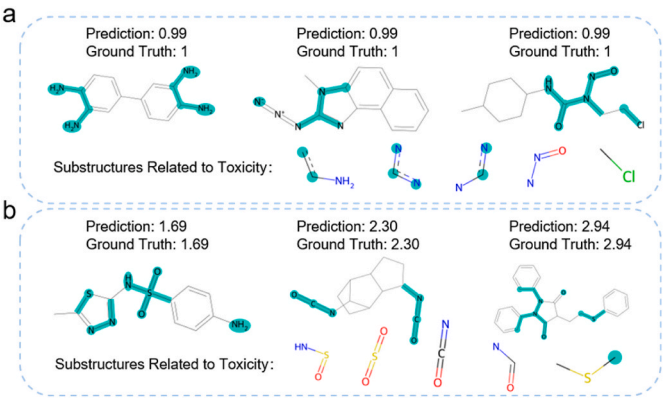

Fig. 5. Interpretability analysis. The attention weight visualization of molecules in the (a) AMES classification prediction task and (b) LD50 regression prediction task. The coloring of atoms is determined by their respective degree of importance.

Table 3
Robustness effect scores in the perturbed LD50 dataset.

Methods	Effect score		
	Level 1	Level 2	Level 3
RF	0.352	0.465	0.542
XGBoost	0.368	0.432	0.573
Attentive FP	0.366	0.494	0.589
GCN	0.297	0.441	0.595
GraphSAGE	0.323	0.351	0.582
GAT	0.261	0.379	**0.536**
GIN	0.298	0.365	0.598
MolFPG	**0.226**	**0.329**	0.627

4. Conclusion

The MolFPG model, which integrates a multi-level fingerprint-based Graph Transformer architecture, has demonstrated remarkable accuracy and robustness in predicting drug toxicity. Our comprehensive evaluations have shown that MolFPG outperforms state-of-the-art methods. Extensive ablation studies have validated the importance of different model components, such as the multi-level fingerprint encoding module and global attention Graph Transformer encoding module, in achieving the superior performance of MolFPG. Moreover, the analysis of the impact of different types of molecular fingerprints on the model performance has revealed that a multi-fingerprint approach, particularly the combination of Morgan, Maccs, and RDKit fingerprints, is advantageous in capturing various molecular features and achieving the best performance. Additionally, the robustness experiments further verify the reliability and stability of the MolFPG architecture, even in perturbations and unseen samples, maintaining high accuracy and robustness in drug toxicity prediction. The interpretability case studies further demonstrate the ability of the MolFPG model to assign different weights to atoms based on target properties, providing essential implications for medicinal chemists to explore the relationship between substructures and molecular properties.

Overall, the MolFPG model represents a significant step forward in leveraging fingerprint-based and graph-based representations for improving drug toxicity prediction, with broad implications for drug discovery and development. However, some limitations and generalizability concerns need to be considered. The model is dependent on high-quality datasets, and its performance might be affected by biases, noise, or lack of comprehensiveness in the data. Additionally, the applicability of MolFPG to other drug-related properties, such as bioactivity and metabolic stability, remains to be assessed. Despite these limitations, the MolFPG model remains a promising tool for drug toxicity prediction and can be further refined to address these concerns in the future.

Funding

The work was supported by the Natural Science Foundation of China (Nos. 62071278 and 62072329) and Natural Science Foundation of Shandong Province (ZR2020ZD35).

Author contributions

S.T. designed the project, performed computational studies, analyzed data, wrote the draft, and revised the manuscript. **C.Y.** assisted in drawing the framework graph. **Y.W.** reviewed and revised the manuscript. **L.W., L.C., X.C., Z.Y.** reviewed the manuscript.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "MolFPG: Multi-level Fingerprint-based Graph Transformer for Accurate and Robust Drug Toxicity Prediction".

Acknowledgments

The authors acknowledge the anonymous reviewers for reviewing the manuscript.

References

[1] R.A. Thompson, E.M. Isin, Y. Li, R. Weaver, L. Weidolf, I. Wilson, A. Claesson, K. Page, H. Dolgos, J.G. Kenna, Risk assessment and mitigation strategies for reactive metabolites in drug discovery and development, Chem. Biol. Interact. 192 (2011) 65–71.
[2] K. Sachdev, M.K. Gupta, A comprehensive review of computational techniques for the prediction of drug side effects, Drug Dev. Res. 81 (2020) 650–670.
[3] F. Yamashita, M. Hashida, In silico approaches for predicting ADME properties of drugs, Drug Metabol. Pharmacokinet. 19 (2004) 327–338.
[4] V. Kumar, N. Sharma, S. Maitra, In vitro and in vivo toxicity assessment of nanoparticles, Int. Nano Lett. 7 (2017) 243–256.
[5] G. Daston, D.J. Knight, M. Schwarz, T. Gocht, R.S. Thomas, C. Mahony, M. Whelan, SEURAT: safety evaluation ultimately replacing animal testing—recommendations for future research in the field of predictive toxicology, Arch. Toxicol. 89 (2015) 15–23.
[6] R. Su, H. Yang, L. Wei, S. Chen, Q. Zou, A multi-label learning model for predicting drug-induced pathology in multi-organ based on toxicogenomics data, PLoS Comput. Biol. 18 (2022), e1010402.
[7] A. Worth, J. Barroso, S. Bremer, J. Burton, S. Casati, S. Coecke, R. Corvi, B. Desprez, C. Dumont, V. Gouliarmou, Alternative methods for regulatory toxicology–a state-of-the-art review, JRC Sci Policy Rep EUR 26797 (2014) 1–470.
[8] D. Krewski, D. Acosta Jr., M. Andersen, H. Anderson, J.C. Bailar III, K. Boekelheide, R. Brent, G. Charnley, V.G. Cheung, S. Green Jr., Toxicity testing in the 21st century: a vision and a strategy, J. Toxicol. Environ. Health 13 (2010) 51–138.
[9] X. Pan, X. Lin, D. Cao, X. Zeng, P.S. Yu, L. He, R. Nussinov, F. Cheng, Deep Learning for Drug Repurposing: Methods, Databases, and Applications, Wiley Interdisciplinary Reviews: Computational Molecular Science, 2022, e1597.
[10] X. Zeng, X. Tu, Y. Liu, X. Fu, Y. Su, Toward better drug discovery with knowledge graph, Curr. Opin. Struct. Biol. 72 (2022) 114–126.
[11] X. Zeng, H. Xiang, L. Yu, J. Wang, K. Li, R. Nussinov, F. Cheng, Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework, Nat. Mach. Intell. 4 (2022) 1004–1016.
[12] C. Cao, J.H. Wang, D. Kwok, F.F. Cui, Z.L. Zhang, D. Zhao, M.J. Li, Q. Zou, webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study, Nucleic Acids Res. 50 (2022) D1123–D1130.
[13] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, R.K. Tekade, Artificial intelligence in drug discovery and development, Drug Discov. Today 26 (2021) 80.
[14] A. Omer, P. Singh, N. Yadav, R. Singh, An overview of data mining algorithms in drug induced toxicity prediction, Mini-Rev. Med. Chem. 14 (2014) 345–354.
[15] L. Zhang, Y. Yuan, J. Yu, H. Liu, SEMCM: a self-expressive matrix completion model for anti-cancer drug sensitivity prediction, Curr. Bioinf. 17 (2022) 411–425.
[16] X. Zeng, F. Wang, Y. Luo, S. Kang, J. Tang, F.C. Lightstone, E.F. Fang, W. Cornell, R. Nussinov, F. Cheng, Deep generative molecular design reshapes drug discovery, Cell Rep. Med. 4 (2022), 100794.
[17] G. Xiong, Z. Wu, J. Yi, L. Fu, Z. Yang, C. Hsieh, M. Yin, X. Zeng, C. Wu, A. Lu, ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties, Nucleic Acids Res. 49 (2021) W5–W14.
[18] K. Soni, Y. Hasija, Artificial intelligence assisted drug research and development, in: 2022 IEEE Delhi Section Conference (DELCON), IEEE, 2022, pp. 1–10.
[19] B. Song, F. Li, Y. Liu, X. Zeng, Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison, Briefings Bioinf. 22 (2021) bbab282.
[20] X. Zeng, S. Zhu, W. Lu, Z. Liu, J. Huang, Y. Zhou, J. Fang, Y. Huang, H. Guo, L. Li, Target identification among known drugs by deep learning from heterogeneous networks, Chem. Sci. 11 (2020) 1775–1797.
[21] X. Lin, Z. Quan, Z. Wang, H. Huang, X. Zeng, A novel molecular representation with BiGRU neural networks for learning atom, Briefings Bioinf. 21 (2020) 2099–2111.
[22] L. Zhang, H. Zhang, H. Ai, H. Hu, S. Li, J. Zhao, H. Liu, Applications of machine learning methods in drug toxicity prediction, Curr. Top. Med. Chem. 18 (2018) 987–997.
[23] E. Gawehn, J.A. Hiss, G. Schneider, Deep learning in drug discovery, Mol. Info. 35 (2016) 3–14.
[24] B. Gellin, J.F. Modlin, R.F. Breiman, Vaccines as tools for advancing more than public health: perspectives of a former director of the National Vaccine Program office, Clin. Infect. Dis. 32 (2001) 283–288.
[25] Y. Zhang, Y. Wang, Z. Gu, X. Pan, J. Li, H. Ding, Y. Zhang, K. Deng, Bitter-RF: a random forest machine model for recognizing bitter peptides, Front. Med. 10 (2023).
[26] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, Xgboost: extreme gradient boosting, R package version 0.4-2 1 (2015) 1–4.
[27] F. Wang, Y. Ding, X. Lei, B. Liao, F. Wu, Machine learning and deep learning strategies in drug repositioning, Curr. Bioinf. 17 (2022) 217–237.
[28] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[29] H. Zhang, Q. Zou, Y. Ju, C. Song, D. Chen, Distance-based support vector machine to predict DNA N6-methyladenine modification, Curr. Bioinf. 17 (2022) 473–482.
[30] F.Y. Dao, H. Lv, M.J. Fullwood, H. Lin, Accurate identification of DNA replication origin by fusing epigenomics and chromatin interaction information, Research vol. 2022 (2022). ID 9780293.
[31] Z. Tao, Y. Li, Z. Teng, Y. Zhao, A method for identifying vesicle transport proteins based on LibSVM and MRMD, Comput. Math. Methods Med. 2020 (2020), 8926750.
[32] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. AspuruGuzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, Adv. Neural Inf. Process. Syst. 28 (2015).
[33] Y. Yang, L. Chen, Identification of drug-disease associations by using multiple drug and disease networks, Curr. Bioinf. 17 (2022) 48–59.
[34] Y. Li, G. Qiao, K. Wang, G. Wang, Drug-target interaction predication via multi-channel graph neural networks, Briefings Bioinf. 23 (2022).
[35] H. Li, Y. Gong, Y. Liu, H. Lin, G. Wang, Detection of transcription factors binding to methylated DNA by deep recurrent neural network, Briefings Bioinf. 23 (2022).
[36] F. Wang, Y. Ding, X. Lei, B. Liao, F.-X. Wu, Machine learning and deep learning strategies in drug repositioning, Curr. Bioinf. 17 (2022) 217–237, https://doi.org/10.2174/1574893616666211119093100.
[37] Y. Li, G. Qiao, X. Gao, G. Wang, Supervised graph co-contrastive learning for drug-target interaction prediction, Bioinformatics 38 (2022) 2847–2854.
[38] Q. Liu, J. Wan, G. Wang, A survey on computational methods in discovering protein inhibitors of SARS-CoV-2, Briefings Bioinf. 23 (2022).
[39] L. Wei, X. Ye, T. Sakurai, Z. Mu, L. Wei, ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning, Bioinformatics 38 (2022) 1514–1524.
[40] Y. Chen, T. Ma, X. Yang, J. Wang, B. Song, X. Zeng, MUFFIN: multi-scale feature fusion for drug–drug interaction prediction, Bioinformatics 37 (2021) 2651–2658.
[41] C. Li, J. Wang, Z. Niu, J. Yao, X. Zeng, A spatial-temporal gated attention module for molecular property prediction based on molecular geometry, Briefings Bioinf. 22 (2021), bbab078.
[42] L. Wei, X. Ye, Y. Xue, T. Sakurai, L. Wei, ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism, Briefings Bioinf. 22 (2021), bbab041.
[43] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discov. Today 20 (2015) 318–331.
[44] C.W. Coley, W. Jin, L. Rogers, T.F. Jamison, T.S. Jaakkola, W.H. Green, R. Barzilay, K.F. Jensen, A graph-convolutional neural network model for the prediction of chemical reactivity, Chem. Sci. 10 (2019) 370–377.
[45] S. Lee, M. Lee, K. Gyak, S.D. Kim, M. Kim, K. Min, Novel solubility prediction models: molecular fingerprints and physicochemical features vs graph convolutional neural networks, ACS Omega 7 (2022) 12268–12277.
[46] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How Powerful Are Graph Neural Networks?, 2018 arXiv preprint arXiv:1810.00826.
[47] D.P. Williams, S.E. Lazic, A.J. Foster, E. Semenova, P. Morgan, Predicting drug-induced liver injury with Bayesian machine learning, Chem. Res. Toxicol. 33 (2019) 239–248.
[48] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P.W. Lee, Y. Tang, In silico prediction of chemical Ames mutagenicity, J. Chem. Inf. Model. 52 (2012) 2840–2847.
[49] H. Zhu, T.M. Martin, L. Ye, A. Sedykh, D.M. Young, A. Tropsha, Quantitative structure– activity relationship modeling of rat acute toxicity by oral exposure, Chem. Res. Toxicol. 22 (2009) 1913–1921.
[50] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754.
[51] D.E. Rumelhart, G.E. Hinton, J.L. McClelland, A General Framework for Parallel Distributed Processing, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, 1986, p. 26.

[52] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Network. 20 (2008) 61–80.
[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 30.
[54] Y. Yang, D. Gao, X. Xie, J. Qin, J. Li, H. Lin, D. Yan, K. Deng, DeepIDC, A prediction framework of injectable drug combination based on heterogeneous information and deep learning, Clin. Pharmacokinet. 61 (2022) 1749–1759.
[55] H.L. Morgan, The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service, J. Chem. Doc. 5 (1965) 107–113.
[56] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL keys for use in drug discovery, J. Chem. Inf. Comput. Sci. 42 (2002) 1273–1280.
[57] A.P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L.J. Bellis, M. De Veij, A.R. Leach, An open source chemical structure curation pipeline using RDKit, J. Cheminf. 12 (2020) 1–16.
[58] D. Wang, Z. Zhang, Y. Jiang, Z. Mao, D. Wang, H. Lin, D. Xu, DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism, Nucleic Acids Res. 49 (2021) e46.
[59] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, J. Leskovec, Strategies for Pre-training Graph Neural Networks, 2019 arXiv preprint arXiv:1905.12265.
[60] B. Ramsundar, P. Eastman, P. Walters, V. Pande, Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More, O'Reilly Media, 2019.
[61] X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu, H. Wang, Geometry-enhanced molecular representation learning for property prediction, Nat. Mach. Intell. 4 (2022) 127–134.
[62] Y. Lei, J. Hu, Z. Zhao, S. Ye, Drug-Target Interaction Prediction Based on Attentive FP and Word2vec, Intelligent Computing Theories and Application: 18th International Conference, ICIC 2022, Xi'an, China, August 7–11, 2022, Proceedings, Part II, Springer, 2022, pp. 507–516.
[63] T.N. Kipf, M. Welling, Semi-supervised Classification with Graph Convolutional Networks, 2016 arXiv preprint arXiv:1609.02907.
[64] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Adv. Neural Inf. Process. Syst. 30 (2017).
[65] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph Attention Networks, 2017 arXiv preprint arXiv:1710.10903.
[66] Y. Li, C. Hsieh, R. Lu, X. Gong, X. Wang, P. Li, S. Liu, Y. Tian, D. Jiang, J. Yan, An adaptive graph learning method for automated molecular interactions and properties predictions, Nat. Mach. Intell. 4 (2022) 645–651.
[67] S. Dara, S. Dhamercherla, S.S. Jadav, C.M. Babu, M.J. Ahsan, Machine learning in drug discovery: a review, Artif. Intell. Rev. 55 (2022) 1947–1999.
[68] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. 32 (2020) 18069–18083.