Journal of Hazardous Materials 457 (2023) 131808

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Hi-MGT: A hybrid molecule graph transformer for toxicity identification

Zhichao Tan [a,b], Youcai Zhao [a,b], Tao Zhou [a,b,*], Kunsen Lin [a,b,*]

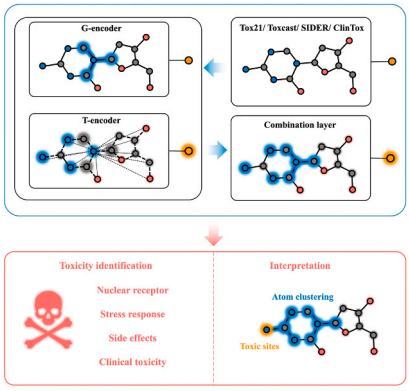
[a] The State Key Laboratory of Pollution Control and Resource Reuse, School of Environmental Science and Engineering, Tongji University, 1239 Siping Road, Shanghai 200092, China

[b] Shanghai Institute of Pollution Control and Ecological Security, 1515 North Zhongshan Rd. (No. 2), Shanghai 200092, China

HIGHLIGHTS

- A novel graph transformer architecture was proposed for toxicity identification.
- Hi-MGT achieves state-of-the-art results on Tox21, Toxcast, SIDER and ClinTox.
- Hi-MGT aggregates local and global information efficiently in molecule graphs.
- A similarity-based toxicity site detection method was proposed.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Dr. S Nan

Keywords:
Toxicity identification
Machine learning
Graph transformer
Toxic site detection

ABSTRACT

Conventional toxicity testing methods that rely on animal experimentation are resource-intensive, time-consuming, and ethically controversial. Therefore, the development of alternative non-animal testing approaches is crucial. This study proposes a novel hybrid graph transformer architecture, termed Hi-MGT, for the toxicity identification. An innovative aggregation strategy, referred to as GNN-GT combination, enables Hi-MGT to simultaneously and comprehensively aggregate local and global structural information of molecules, thus elucidating more informative toxicity information hidden in molecule graphs. The results show that the state-of-the-art model outperforms current baseline CML and DL models on a diverse range of toxicity endpoints and is even comparable to large-scale pretrained GNNs with geometry enhancement. Additionally, the impact of hyperparameters on model performance is investigated, and a systematic ablation study is conducted to demonstrate the effectiveness of the GNN-GT combination. Moreover, this study provides valuable insights into the learning process on molecules and proposes a novel similarity-based method for toxic site detection, which could potentially facilitate toxicity identification and analysis. Overall, the Hi-MGT model represents a significant advancement in the development of alternative non-animal testing approaches for toxicity identification, with promising implications for enhancing human safety in the use of chemical compounds.

* Corresponding authors at: The State Key Laboratory of Pollution Control and Resource Reuse, School of Environmental Science and Engineering, Tongji University, 1239 Siping Road, Shanghai 200092, China.

E-mail addresses: zhou410218380@163.com (T. Zhou), kslin@tongji.edu.cn (K. Lin).

https://doi.org/10.1016/j.jhazmat.2023.131808
Received 2 April 2023; Received in revised form 18 May 2023; Accepted 7 June 2023
Available online 8 June 2023

1. Introduction

Humans are exposed to a vast array of chemical compounds through environmental factors, nutrition, cosmetics, and medication. In order to safeguard against potential harm, these substances must undergo rigorous testing for adverse effects, particularly in relation to their toxicity [18]. The global regulations regarding the management of chemical substances have been strengthened in recent years, with notable emphasis placed on the implementation of the European Union's Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) program [11]. Determining the potential toxicity of chemicals is a critical aspect in deciding whether they are appropriate for extensive use. However, the traditional paradigm for toxicity testing that relies on in vivo animal studies and in vitro techniques is laborious, expensive, and impractical for evaluating an increasing number of compounds [12]. In addition, ethical concerns and the desire to replace animal experimentation have garnered increased attention. The low efficiency and contentious nature of animal testing have led to its gradual phase-out. Consequently, there is an urgent need to develop alternative approaches for toxicity identification that do not involve animal testing.

With the increase of available experimental data, machine learning (ML) has emerged as a promising tool for toxicity identification [4,10], including random forest (RF) [13,17], support vector machine (SVM) [16], k-nearest neighbors (KNN) [5] and many others [8,31]. Nevertheless, conventional machine learning (CML) heavily relies on domain knowledge-based feature engineering for molecular descriptors or fingerprints, which can significantly impact its performance. Nonlinear Support Vector Machines (SVMs), for instance, are capable of handling high-dimensional data but may not be robust in the presence of diverse chemical descriptors [12]. Furthermore, the linear molecule representation employed in these methods is limited in its capacity to describe intricate structural features of molecules. For instance, the MACCS keys, one of the most commonly used structural keys, only encode known functional groups, thus proving inadequate for describing molecules in unfamiliar contexts. Likewise, Morgan's FP, the most widely used Extended Connectivity FPs (ECFPs), can suffer from information loss due to bit collision problems during the hash-map operation [1].

In the past decade, deep learning (DL) has garnered much attention for toxicity identification due to its ability to bypass feature extraction and reach good performance [12]. Particularly, Mayr et al. [18] developed DeepTox using deep neural networks (DNNs) to identify the toxicity of approximately 10k molecules on nuclear receptors and stress responses in the 21st Century (Tox21) Data Challenge, and found that DL excelled in toxicity identification, which outperformed many other CML approaches, like Naive Bayes, SVM, and RF. To further improve the performance of DL in toxicity identification, more attention is paid on how to obtain more expressive representation for molecules due to the limited expression power of traditional descriptors and fingerprints. Among these methods, graph notation has become a state-of-the-art method and DL methods on the graph representations by themselves have become a focus of the current study [1], called Graph Neutral Networks (GNNs). Under the scheme of GNNs, organic molecules represented by their SMILES (Simplified Molecular Input Line Entry Specification) codes are converted into graphs, in which nodes represent atoms, and edges represent bonds [25]. Many expressive GNNs have been developed for the molecule graph, such as Neural FP [7], Attentive FP [28] and HiGNN [33]. Nonetheless, many studies have reported performance degradation as the number of iterations increases and graph models become deeper. This phenomenon is commonly attributed to vanishing gradients and over-smoothing of the representation [26].

To push the boundary of conventional GNNs, researchers [15,20,22,29] have recently explored applying transformer architectures in GNNs, called graph transformers (GTs), in light of the tremendous success of transformers in natural language processing (NLP). A well-designed structure-aware GT is proven to be a potential solution to alleviate performance deterioration in GNNs, as they can easily capture global or long-range information through computing all pairwise node interactions in self-attention block [26]. GTs have already achieved great success, such as topping the leaderboard of the OGB Large-Scale Challenge in the molecular property prediction track [30]. However, their heavy reliance on complex positional encodings (PEs) in node features poses challenges. Learnable PEs significantly increase model parameters, leading to longer training time and requiring more data or complicated pretraining methods, which ultimately limits the wide application of GTs [14,22]. Notably, there has been no previous research on using Transformer-based models for toxicity identification to the best of our knowledge.

This study proposes a novel hybrid graph transformer architecture with limited parameters, called Hi-MGT, for the robust identification of molecule toxicity. Hi-MGT is designed to simultaneously and comprehensively aggregate both local and global structural information of molecules in each layer, thus gathering more meaningful toxicity information hidden in molecule graphs. The performance of Hi-MGT is extensively evaluated on four toxicity benchmark datasets, namely Tox21, Toxcast, SIDER, and ClinTox. The impact of hyperparameters on model performance is discussed, and a systematic ablation study is also conducted to demonstrate the effectiveness of the proposed GNN-GT combination. Furthermore, this study provides insights into the learning process on molecules, shedding light on this black box and providing further evidence of the model's effectiveness. Finally, a novel similarity-based method is proposed for the detection of toxic sites by exploring the relationship between embedded virtual node and toxic sites, which could potentially facilitate toxicity detection and analysis.

2. Methods

2.1. Toxicity datasets and splitting methods

The Tox21, Toxcast, SIDER, and ClinTox datasets are commonly used toxicity benchmark datasets in toxicity identification from MoleculeNet [27]. These four datasets are available in the Support Information and Table S1 provides basic information about them. Besides, in Fig. S1, the composition of each dataset is depicted, highlighting the class imbalance present in each dataset. The maximum molecule length was set to 150 during dataset preparation to prevent the inclusion of overly lengthy molecules. In the data splitting process, all datasets were partitioned into training, validation, and test sets using a standard ratio of 8:1:1. To ensure comprehensive model evaluation, both random splitting and scaffold splitting were applied to all datasets in model comparisons.

2.2. Graph neural networks

2.2.1. Message passing framework

In 2017, Gilmer et al. [9] introduced a general framework for GNNs, called Message Passing Neural Networks (MPNNs), which unifies existing GNNs into a common framework. The message passing framework can be formulated as follows, where $N(v)$ represents all neighboring nodes for node v, M_t is learnable parameter matrix and U_t represents aggregation function.

$$m_v^{t+1} = \sum_{w \in N(v)} M_t\left(h_v^t, h_w^t, e_{vw}\right) \tag{1}$$

$$h_v^{t+1} = U_t\left(h_v^t, m_v^{t+1}\right) \tag{2}$$

$$\hat{y} = R\left(\left\{h_v^T \middle| v \in G\right\}\right) \tag{3}$$

This framework decomposes the learning process into two phases: (i) the message passing phase, as shown in Eqs. (1) and (2). In this phase, each node v obtains its neighborhood information m_v^{t+1} by combining the embeddings of adjacent nodes h_w^t and edge embeddings e_{vw} The updated

node embedding h_v^{t+1} is computed by aggregating the neighborhood information m_v^{t+1} and the former node embeddings h_v^t through U_t. (ii) the readout phase where a feature vector $\hat{y}$ for the entire graph is computed using the readout function R. Commonly used readout functions include summation, mean functions, and the Gated Recurrent Unit (GRU), which is an iterative neural network block.

2.2.2. Graph attention mechanism

In 2017, Veličković et al. [24] proposed graph attention networks (GATs) that extend the attention mechanism to graph-structured data for node classification tasks. Since then, many variant model architectures have been proposed, such as Attentive FP [28], HiGNN [33] and Frag-GAT [32]. The core idea of graph attention mechanism is to obtain a weighted context m_v^{t+1} in the message passing phase, focusing more on task-relevant neighboring atoms and bonds. And (1) can be modified as below

$$m_v^{t+1} = \sum_{w \in N(v)} \alpha_{vw} M_t\left(h_v^t, h_w^t, e_{vw}\right) \tag{4}$$

where α_{vw} is the weights calculated through alignment and weighting, as shown in (5) and (6). The hidden states of target atom and neighboring atom are aligned to obtain a score, followed by SoftMax function to normalize these weights.

$$score_{vw} = Activation(W[h_v, h_w]) \tag{5}$$

$$\alpha_{vu} = SoftMax(score_{vw}) = \frac{exp(e_{vu})}{\sum_{u \in N(v)} exp(e_{vu})} \tag{6}$$

2.3. Position-aware transformer

2.3.1. Transformer encoder

The transformer encoder architecture comprises multiple transformer encoder layers, each consisting of two main components: a self-attention module and a position-wise feed-forward network (FFN). The self-attention mechanism is the backbone of each transformer encoder. In this mechanism, three learnable matrixes W_Q, W_K, W_V are firstly employed to map the original hidden state H to query Q, key K and value V respectively, as shown in Eqs. (8) and (9). Then the updated hidden state is obtained through the scaled dot-product attention in Eq. (7). Specifically, the dot products of the query with all keys QK^T are divided by the square root of the model dimension $\sqrt{d_k}$, and SoftMax function is applied to obtain the weights on the values. This attention mechanism allows each token in the sequence to attend to all other tokens and weigh their contributions according to their relevance to the current token.

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

$$Q = HW_Q, K = HW_K, V = HW_V \tag{8}$$

$$H = \left(h_1^T, h_2^T, \ldots, h_n^T\right)^T \tag{9}$$

Besides, multi-head attention strategy is often adopted to allow the self-attention mechanism to attend jointly to information from different representation subspaces as shown in Eqs. (10) and (11), where W^O is a learnable projection matrix for the output.

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_n)W^O \tag{10}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{11}$$

2.3.2. Positional encodings

The tremendous success of transformers in NLP has spurred a growing interest in leveraging their power to enhance the performance of GNNs. However, applying vanilla transformers to graph learning naively is ineffective, as the transformer architecture treats input as a fully-connected graph and ignores local and structural information. To inject more graph structural information into transformers, a proven effective strategy involves adding positional encodings (PEs) to the self-attention mechanism, which works mainly through two ways: (i) by serving as an additional representation in node features, and (ii) by acting as an attention bias, as illustrated in Eqs. (12) and (13), respectively.

$$H = \left((h_1, PE_1)^T, (h_2, PE_2)^T, \ldots, (h_n, PE_n)^T\right)^T \tag{12}$$

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}} + bias_1 + bias_2 + \ldots\right)V \tag{13}$$

As summarized by Rampášek et al. [21], PEs can be mainly classified into three parts: local PEs, which enable a node to determine its position in a local cluster; relative PEs, which allow two nodes to understand their distances; and global PEs, which provide a node with its global position within the graph.

Inspired by Ying et al. [30] and Maziarka et al. [19], the self-attention mechanism in Hi-MGT was modified to incorporate PEs in the form of $bias_{dist}$, $bias_{adj}$ and $bias_{SP}$, as shown in Eqs. (14) and (15). These biases are 3-D atom pair distance matrix (distance matrix), adjacency matrix, and shortest path distance (SPD) matrix after transformation. Moreover, λ_{att}, λ_{dist}, λ_{adj}, λ_{SP} are scalars weighting corresponding bias, which can be adjusted manually as hyperparameters. Among them, $bias_{adj}$ carries global connectivity information as a global PE, $bias_{SP}$ carries bond-length information as a 2-D relative PE and $bias_{dist}$ carries 3-D information as a 3-D relative PE. Furthermore, when combined with GNNs, more local information is injected into atom embeddings, which works as explicit local PEs after learning of the first core layer. These modifications make the model more aware of graph structure. Therefore, different from the original totally feature-based self-attention, more importance could be given on neighboring nodes.

$$Attention(Q, K, V) = SoftMax(Score(Q, K))V \tag{14}$$

$$Score(Q, K) = \lambda_{att}\frac{QK^T}{\sqrt{d_k}} + \lambda_{dist}bias_{dist} + \lambda_{adj}bias_{adj} + \lambda_{SP}bias_{SP} \tag{15}$$

2.4. Molecule representation

In this work, a total of seven atomic features and three bond features were used to initialize the representations of atoms and bonds, which is detailed in Table S2. All features are one-hot encoded with the atomic features consisting of 38 bits, and the bond features consisting of 6 bits. Notably, a virtual node symbol was set in front of every atomic feature vector, resulting in atomic features of 39 bits in total. The one-hot encoding is generated by listing all the possible categorical variables for the feature and assigning a binary value of 1 or 0 to each variable based on its correspondence with the candidate variables.

2.5. Model design

2.5.1. Model architecture

As an example input using Decitabine, the entire model architecture is summarized in Fig. 1: (i) Firstly, its SMILES notation is converted to a molecule graph through RDkit. Features of every atom and bond are initialized based on Table S2 and stacked to form the atom feature matrix and the bond feature matrix respectively. Then, the adjacency matrix, 3-D relative distance matrix, and SPD matrix are further calculated based on basic graph algorithms and RDkit. Notably, since the molecule length varies in the dataset, zero padding is required to standardize the row size of these matrices in each batch. Particularly, a virtual node not connected with any atoms is embedded in the molecule

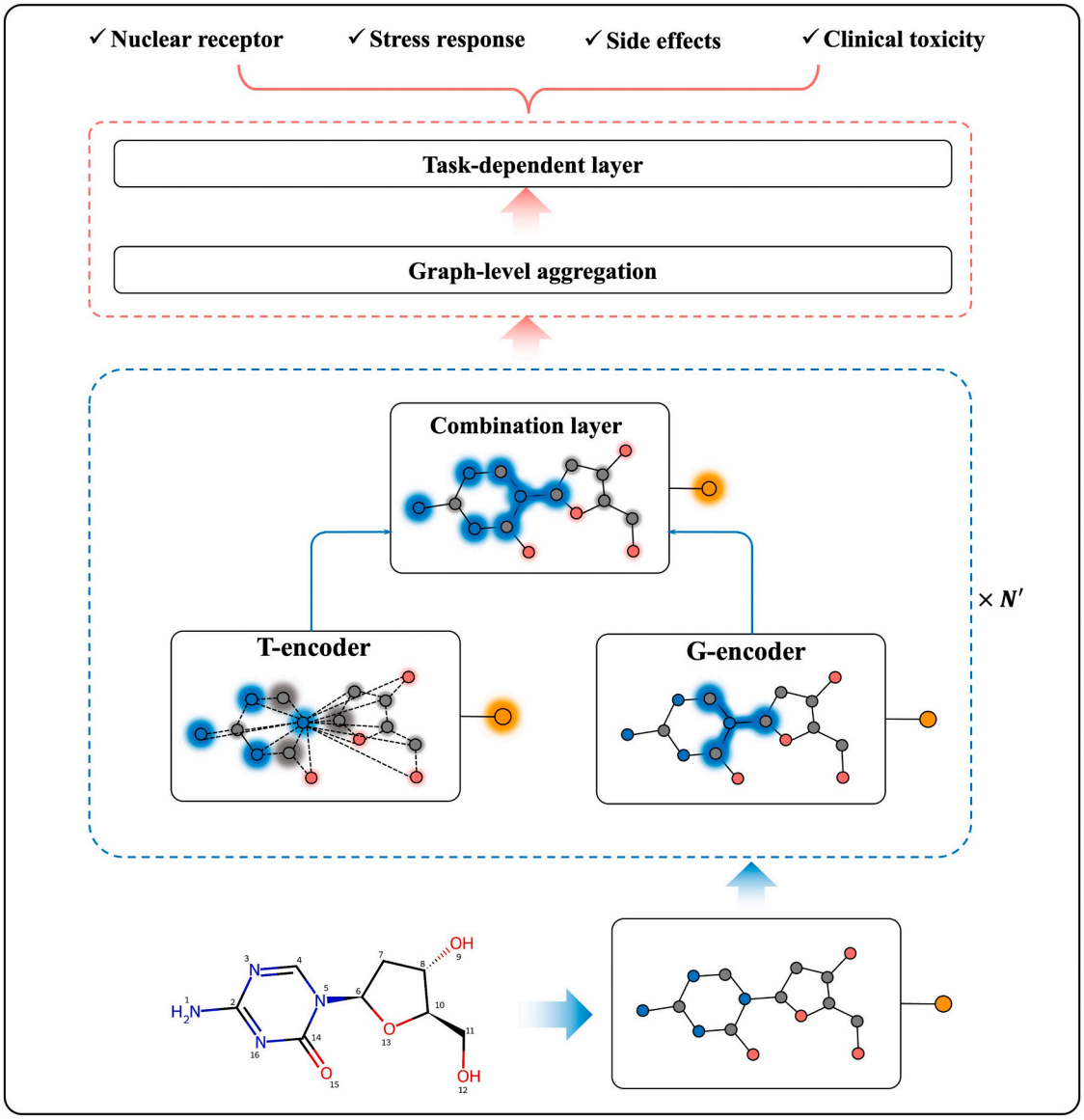


Fig. 1. Model architecture.

graph and set as node 0, which is proven to be an effective strategy for molecule-level aggregation [19]. In Fig. 1, the yellow node outside the frame represents the embedded virtual node. (ii) Atom feature matrix and bond feature matrix are fed into embedding layers to expand their feature dimensions. (iii) All preprocessed molecule graph information is fed into the core learning layer equipped with GNN-GT combination. T-encoder may consist of several transformer encoders. The number of transformer encoders in each core learning layer is set as a hyperparameter N. Then atom embeddings carrying more local information from G-encoder and atom embeddings carrying more global information from T-encoder are combined in combination layer, resulting updated atom embeddings. Similarly, the core learning layer can be duplicated and the total number of core learning layers is also set as a hyperparameter N'. (iv) In final output layer, the molecule representation is obtained through graph-level aggregation or directly from the final representation of the virtual node. Then the molecule representation is fed into task-dependent layer to generate ultimate predictions for different toxicity endpoints.

2.5.2. *GNN-GT combination*

The GNN-GT combination is a crucial component of the entire model architecture, as it allows for the aggregation of two-level graph information. The core learning layer, which incorporates the GNN-GT combination, is depicted in Fig. 2. Intuitively, it consists of three parts: T-encoder, G-encoder and combination layer. At the input, atom embeddings are fed into two encoders simultaneously while bond embeddings are only fed into G-encoder. Layer normalization is then conducted in both encoders to unify feature scale.

In T-encoder, the entire molecule is treated as a fully-connected graph and multi-head self-attention layer is applied to allow for similarity-based global message passing as described in Section 2.3. Particularly, for every atom, all other atoms participate in the message passing based on the similarity calculated from self-attention, which is distance-independent. But some local information is still injected by introduced attention bias. It's followed by FFN layer, a fully-connected neutral network block in T-encoder that outputs the final global-informed atom embeddings. The number of FFN layer in each Transformer encoder is set as a hyperparameter K. As illustrated in subfigure T-encoder, taking the 5th atom in Decitabine (nitrogen) as an example,

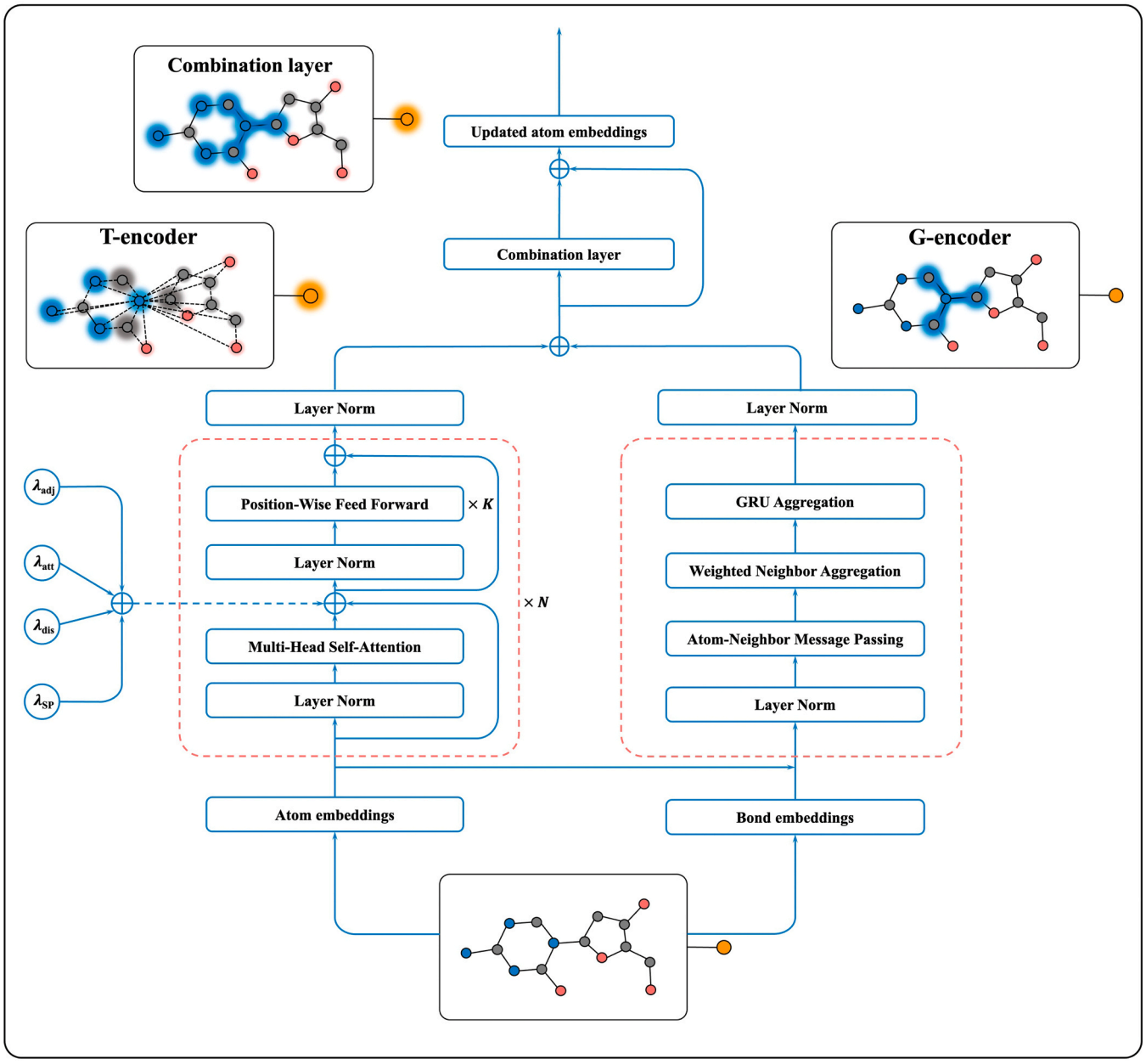


Fig. 2. GNN-GT combination.

it's connected with all other atoms, which represents the connectivity of a fully-connected graph in self-attention mechanism. Each atom is assigned with a certain glowing background to represent a different weight. The direct neighboring atoms is emphasized due to the injected local information while other nitrogens are emphasized because of high similarity relative to the target atom. Especially, the virtual node is also able to exchange message with other atoms in T-encoder.

In G-encoder, local message passing is performed with consideration of local graph structures. For target atom in molecule, neighboring atoms and bonds are filtered out based on adjacency matrix, then message passing occurs as described in Section 2.2.1. This study only employs one-hop graph convolution, so only adjacency atoms participate in the local message passing of G-encoder. Particularly, the isolation of the virtual node makes it not involved in this phase. Finally, GRU block aggregates current embedded atom features with local neighborhood information, forming a new local-informed atom features, which is also followed by layer normalization. As illustrated in subfigure G-encoder, no dash line exists because all connectivity is based on the adjacency matrix of the molecule graph. And only direct neighboring atoms are emphasized under the operation of one-hop graph convolution.

In combination layer, the local-informed and global-informed atom features are combined through residual connection, resulting in more expressive atom embeddings, as illustrated in subfigure Combination layer.

3. Results and discussion

3.1. Performance results on benchmark toxicity datasets

3.1.1. Comparison with other methods on all datasets

The performance of Hi-MGT on toxicity identification was comprehensively compared with other ML and DL methods, including CML models (Logreg, Kernel SVM, XGBoost), GNNs (Weave, GC, Attentive FP) as well as transformer-based models (SMILES-T, ET). Additionally, a large-scale pretrained GNN (GEM) was also included to demonstrate the gap between Hi-MGT with large-scale geometry-based models with

delicate pretraining. The results of these methods were taken from previous study [23,27,6].

Among these datasets, Tox21, Toxcast, SIDER were all trained on a multitask setting, and ClinTox on a single-task setting. Before training, Adam optimizer was adjusted with weights to deal with the class imbalance of toxicity datasets and TPE algorithm was used for hyperparameter optimization. Using the best set of hyperparameters, AUC_ROCs on all testing datasets are summarized in Table 1 where the best results for non-pretrained models are bolded, and these results are also shown in Fig. 3(a) and (b).

As illustrated in Fig. 3(a) and (b), the Hi-MGT model outperforms all other non-pretrained models in all learning tasks except the Tox21 (Random) and SIDER (Random). This result demonstrates that Hi-MGT can accurately identify a broad range of toxicity endpoints by learning molecule graphs. Notably, Hi-MGT outperforms all other non-pretrained models under scaffold splitting, which is a more challenging and realistic splitting method. In particular, on the ClinTox dataset, Hi-MGT achieves a score of 0.982 using random splitting, outperforming the best of the rest, Attentive FP, by 0.04 and even surpassing the average performance of common ML models by 0.2. While ET employs 3D conformers to enhance performance and GEM is also enhanced by geometry information and large-scale pretraining, Hi-MGT still outperforms ET significantly and achieves competitive results with large-scale pretrained GEM.

In summary, the performance in toxicity identification is quite encouraging because Hi-MGT achieves a comparable performance through small-scale training to that of 3D-enhanced models and large-scale pretraining GNNs. Therefore, the molecular toxicity information extracted by Hi-MGT may provide a valuable workaround for addressing many of the problems involved in exploring the large conformational space of molecules [28].

3.1.2. Performance on multitask toxicity prediction

As shown in Fig. 3(c), compared with DeepTox [18], the grand winner in Tox21 Challenge, and baseline model SVM, Hi-MGT achieves competitive results on every toxicity endpoint in Tox21 dataset. Notably, the performance of DeepTox is partly attributed to manually enriched dataset through kernel-based structural and pharmacological analoging (KSPA), but Hi-MGT is only trained on 80% of Tox21 dataset. The performance of Hi-MGT and DeepTox on most endpoints is similar, which demonstrates the great power of Hi-MGT to extract molecule information even though the scale of dataset is limited.

3.1.3. Performance on single-task toxicity prediction

Additionally, to demonstrate the discriminatory ability of Hi-MGT for potential toxic molecules, molecule embeddings before and after transformation on ClinTox is visualized using principle component analysis (PCA) for all 1491 molecules, as depicted in Fig. 4. The initial molecule embeddings are based on Avalon fingerprints, while the transformed embeddings are obtained from the last layer of the optimized model on ClinTox by random splitting. Fig. 4(a) illustrates that there is a significant overlap between toxic and non-toxic molecules in the initial embeddings, making it challenging to distinguish them based on their fingerprint. The distribution curves also exhibit a severe imbalance, further complicating the toxicity identification task. However, after applying the embedding transformation by Hi-MGT, as shown in Fig. 4(b), most molecules are correctly classified, with a clear separation between toxic and non-toxic molecules along the PCA1 dimension. Nonetheless, there are still around 280 molecules falsely categorized as toxic, while only 12 toxic molecules are mistakenly classified as non-toxic, indicating a high recall rate and superior discriminatory ability of Hi-MGT towards potential toxic molecules. Furthermore, an investigation has been conducted on these false negative (FN) molecules to determine why Hi-MGT tends to classify them as non-toxic.

Fig. 4(c) shows two typical molecule structures. The method introduced in section 3.5.3 was used to label the most representative atoms with darker red. For the structure on the left, the carbons connected with a single bond and hydroxyl groups are the most representative atoms for Hi-MGT, which commonly do not possess toxicity. The structure on the right typically contains a nitrogen cation, which is often classified as non-toxic in ClinTox. This could be attributed to the fact that the molecules in the training set, which contain a nitrogen cation, are generally non-toxic, such as [(1R)−2,3-dihydro-1H-inden-1-yl]-prop-2-ynylazanium. Including more typical toxic molecules into datasets is a significant problem that requires further investigation to advance toxicity identification.

Table 1
AUC_ROC on testing datasets for different ML methods.

Methods	Split type	Tox21	Toxcast	SIDER	ClinTox
Logreg	Random	0.794	0.605	0.643	0.722
KernelSVM	Random	0.822	0.669	**0.682**	0.669
XGBoost	Random	0.794	0.640	0.656	0.799
Weave	Random	0.820	0.742	0.581	0.832
GC	Random	0.829	0.716	0.638	0.807
Attentive FP	Scaffold	0.761	0.637	0.606	0.847
	Random	**0.858**	0.805	0.637	0.940
SMILES-T	Scaffold	0.691	0.578	0.504	0.819
ET	Scaffold	0.751	0.623	0.560	0.843
GEM (pretrained)	Scaffold	0.781	0.692	0.672	0.901
Hi-MGT	Scaffold	**0.774**	**0.723**	**0.642**	**0.898**
	Random	0.847	**0.834**	0.647	**0.982**

3.2. Hyperparameter optimization

3.2.1. Optimization results

Hyperparameter optimization was conducted using the TPE algorithm in Optuna with a maximum of 60 search trials for each task. The optimized hyperparameters and their corresponding tuning sections are summarized in Table S3, where decay weight represents the coefficient of the exponential decaying function utilized in distance-based attention calculation. Additionally, graph convolution with or without attention was also considered as a tunable hyperparameter for various aggregation methods of local information in G-encoder.

Fig. S2 depicts the final optimized outcomes of all 12 models and highlights the commonly adopted hyperparameters. These include a model dimension of 64, 3 transformer encoders in the core learning layer, 1 or 2 total layers, sum aggregation, and non-attentive convolution for graph convolution. It is noteworthy that most optimized models utilize the transformer encoder more than once in T-encoder, emphasizing the significance of layer stacking for T-encoder. Surprisingly, three models use the core layer only once without duplication, which deviates from previous studies on GNNs that rely on core layer stacking to achieve more global information [23]. But models with a single core layer still outperform other GNN architectures with stacked layers. This indicates that the core learning layer in Hi-MGT can effectively combine two-level chemical information without relying on sequential aggregation. In contrast to the results in MAT [19], the use of global aggregation through a virtual node is never observed as an optimal in experiments, possibly because of fewer total layers in Hi-MGT, which reduces the opportunity for the virtual node to exchange messages with atoms carrying comprehensive chemical information, particularly the local information aggregated from G-encoder.

3.2.2. The impact of hyperparameters

Fig. 5 is presented to demonstrate the impact of hyperparameters on performance, using the Tox21 dataset. Four critical hyperparameters, namely model dropout, aggregation type, batch size, and lambda for attention, are selected for analysis. It's shown that mean and sum functions are two primary aggregation types that perform well on Tox21, while virtual node aggregation performs relatively poorly. As mentioned above, this result may be attributed to the lack of valid local

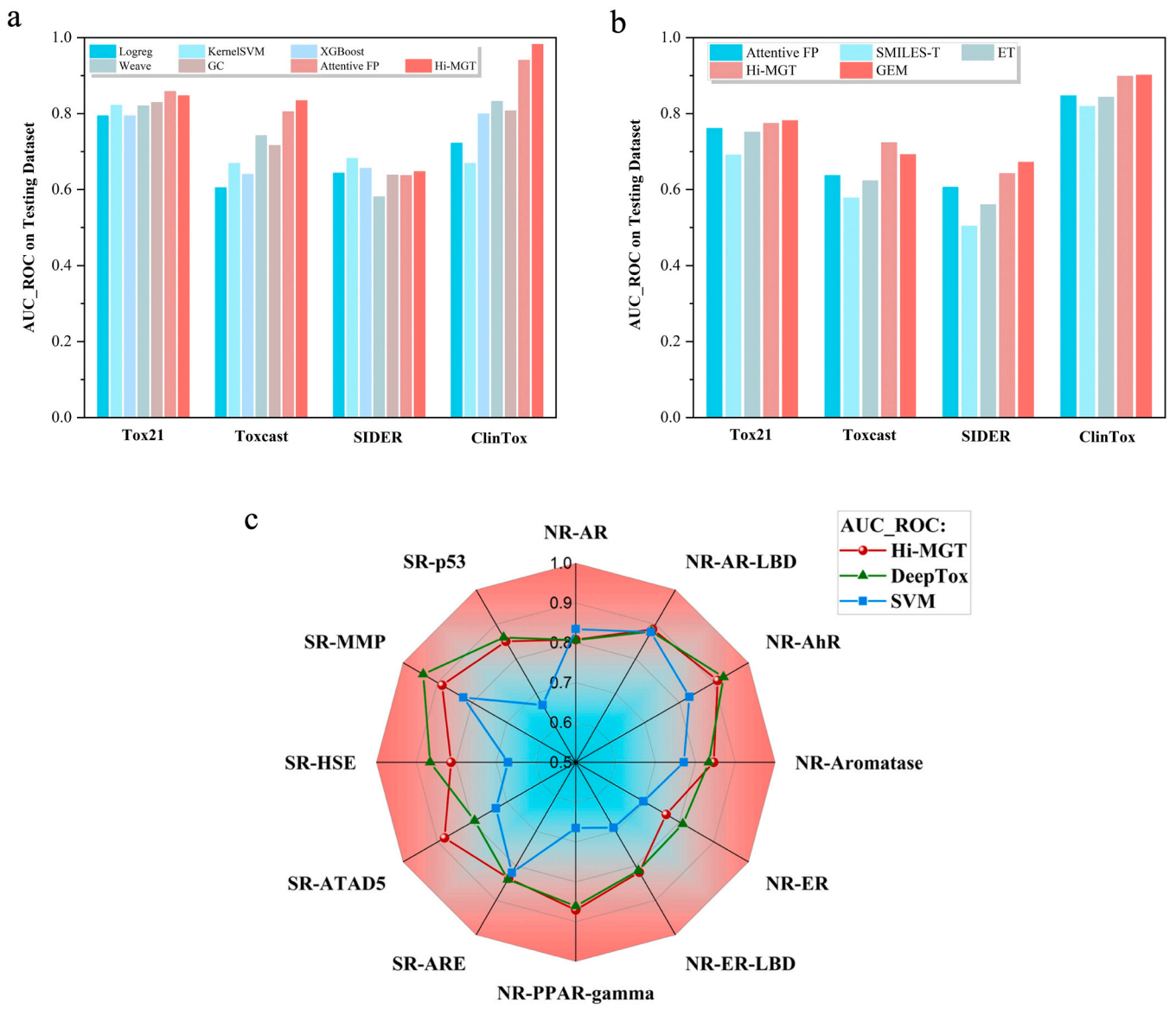


Fig. 3. Performance of Hi-MGT on toxicity benchmark datasets; (a) AUC_ROC on testing datasets in random splitting for different ML methods; (b) AUC_ROC on testing datasets in scaffold splitting for different ML methods; (c) Radar chart of AUC_ROC on different toxicity endpoints of Tox21.

information from the G-encoder, because the virtual node engages in message exchange with other real atoms mainly in T-encoder due to its isolation. In terms of model dropout, the optimal values are mainly distributed from 0.2 to 0.4, which mitigates overfitting on the training set and enhances the generalization ability of Hi-MGT. However, if the dropout rate is too high, model performance deteriorates because of insufficient information. Concerning the batch size, a value below 20 is usually too small to yield satisfactory results. As illustrated in the 3rd row of Fig. 5, a wide range of lambda values yields good performance with contour lines nearly vertical to the x-axis. Thus, the impact of lambda on the model's performance is relatively minor compared to the first three hyperparameters.

3.3. Ablation study

The Hi-MGT model architecture incorporates both global and local information in each core learning layer, which is different from the common approach of GNNs that sequentially fuse information across multiple layers. Results presented in Section 3.1 demonstrate that Hi-MGT outperforms other GNNs with stacked layers, even with only a single layer. This highlights the efficacy of the core layer in Hi-MGT for information fusion. To further investigate the impact of the integration of conventional GNN and position-aware GT with attention bias, ablation study was conducted through isolating these two main components. Additionally, comparison of T-encoder with and without attention bias was also conducted, to further explore the importance of attention bias in achieving position awareness. In comparation, other hyperparameters and training settings were controlled as the same, and scaffold splitting, early stop strategy were used, as described in Section 2.

Table S4 presents the results of 12 ablation models across all datasets. Out of 12 ablated models, suboptimal results are observed in 11 when any one component is removed, suggesting that only T-encoder or G-encoder is usually insufficient for accurate toxicity identification. Furthermore, detailed learning process on Tox21 is also illustrated in Fig. S3, in which Hi-MGT offers sustained improvements on the training and validation sets throughout the training process.

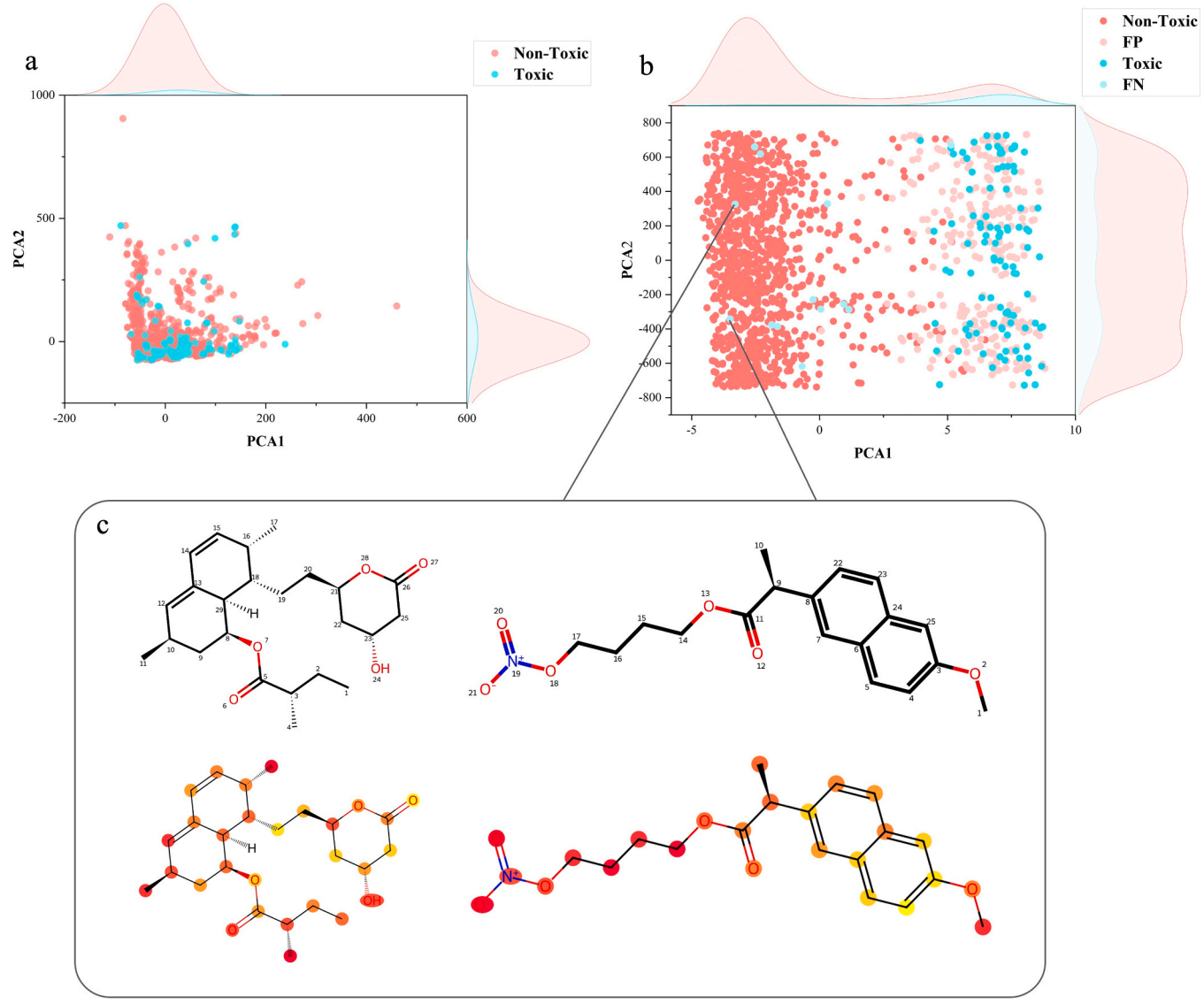


Fig. 4. Molecule embeddings of ClinTox before and after transformation; (a) Scatter plot of dimension-reduced molecule embeddings on ClinTox before transformation; (b) Scatter plot of dimension-reduced molecule embeddings on ClinTox after transformation; (c) Two typical molecule structures falsely classified as "non-toxic" and their colored representative atoms.

3.4. Interpretation analysis

3.4.1. Atom clustering

Heat maps of the atom similarity matrices were constructed for [(1R)− 2,3-dihydro-1H-inden-1-yl]-prop-2-ynylazanium and Decitabine in ClinTox to illustrate the results, as shown in Fig. 6(a) and (b). The heat maps display higher similarity between atom pairs with more intense red color. The heatmap analysis of [(1R)− 2,3-dihydro-1H-inden-1-yl]-prop-2-ynylazanium reveals the presence of three distinct clusters, namely cluster 1 (atoms 1–4), cluster 2 (atoms 5–7), and cluster 3 (atoms 7–13). Among these clusters, cluster 2 and cluster 3 are easy to understand due to their similar features and chemical environments: three similar carbons in cluster 2 and an aromatic system in cluster 3 connected by a conjugated π bond. However, cluster 1 presents a unique case, consisting of three carbons and one positively charged nitrogen atom. Despite its unusual composition, this cluster still adheres to the fundamental chemical principle that electron sharing between atoms leads to similarities in chemical behavior. The positively charged nitrogen atom withdraws electrons from an electron donor formed by a linear triple carbon-carbon bond, which provides a good backbone for electronic transferring, leading to an unusual substructure.

Similarly, the heatmap analysis of Decitabine reveals that despite being composed of different elements, atoms 1–5 and atoms 14–16 exhibit high similarity. The similarity arises from the formation of a large conjugate system with delocalized π electrons, resulting in a planar substructure. The clustering phenomenon observed in the heatmap analysis indicates that Hi-MGT has successfully learned some chemical principles through training and is capable of extracting substructures formed by electron sharing.

3.4.2. Learning process

To better understand the evolving process of cluster 1 in [(1R)− 2,3-dihydro-1H-inden-1-yl]-prop-2-ynylazanium, similarity matrices of each core learning layer were computed for the positive nitrogen atom, as depicted in Fig. 6(e). In the input and initial embedding layers, nitrogen is only similar to atoms 3, 6, and 7, and the similarity mainly comes from the fact that they are all carbon atom connected to two neighboring atoms through single bond.

However, in layer 1, the similarity between positive nitrogen and atom 1–3 is strengthened. The G-encoder in layer 1 aggregates chemical

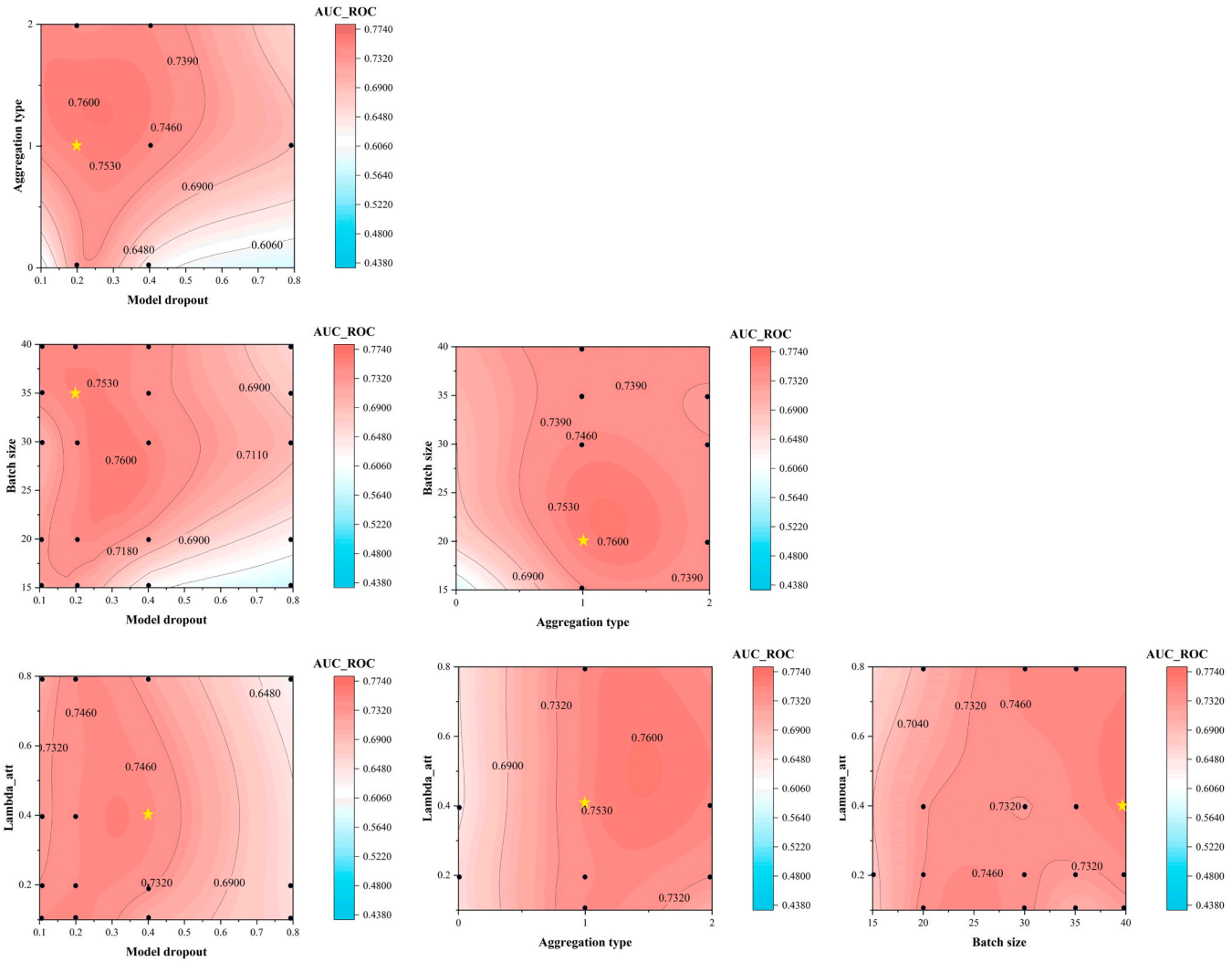


Fig. 5. Hyperparameter optimization on Tox21 (Scaffold splitting). For aggregation type, 0 represents virtual node aggregation, 1 represents sum function and 2 represents mean function.

neighborhood information within 1-hop range to nitrogen, leading to the increased similarity between atom 5 and nitrogen. And the similarity with atom 5 is slightly higher than with atom 3, which may be attributed to a higher formal charge of atom 5. Besides, the similarity between nitrogen and other atoms decays with the increase of distance.

Conversely, the T-encoder in layer 1 relies more on feature-dependent message passing, which can reach relatively far atoms. In this case, atoms 1 and 3 receive more attention from T-encoder, contributing to the formation of cluster 1. And the relation between them is not obvious in G-encoder due to its reliance on one-hop graph convolution in the first layer.

The combination layer further integrates the outputs of G-encoder and T-encoder to form the final output of layer 1. Interestingly, G-encoder gives atom 5 the highest similarity while T-encoder treats it differently. This difference highlights the discrepancy in message passing between G-encoder and T-encoder. In G-encoder, only direct neighboring atoms are included in graph convolution due to one-hop message passing, especially in the first learning layer. So directly connected atoms usually gain more similarity. In contrast, T-encoder can reach far and is more sensitive to global substructures. In layer 2, the connections within cluster 1 are further strengthened. Overall, the evolution of cluster 1 in [(1 R)− 2,3-dihydro-1 H-inden-1-yl]-prop-2-ynylazanium represents a typical learning process involving multiple layers and different types of message passing.

3.4.3. *Toxic sites detection*

In toxicity analysis, identifying potential toxic sites is a critical task because these sites play a significant role in representing the toxicity of the entire molecule and help to elucidate the underlying toxicity mechanism. In this study, a virtual node was embedded as a virtual isolated atom in the molecule graph to represent the molecule's hidden state. It's expected that some hidden relations between the virtual node and toxic sites may exist due to their representational role. To investigate their relationship, atoms are labeled with different colors based on their similarity to the virtual node, as presented in Fig. S4. Notably, in [(1R)− 2,3-dihydro-1 H-inden-1-yl]-prop-2-ynylazanium, atoms 1 and 4 exhibit the highest similarity to the virtual node, which is consistent with chemical intuition regarding toxic sites, as these two atoms possess unique electron atmospheres and occupy an end-point position, which result in a strong attacking ability as electrophile or nucleophile.

Furthermore, more labeled molecules are presented in Fig. 7, including six toxic molecules and six non-toxic molecules. As for toxic molecules, following structures tend to be highlighted: (1) carbonyl group and amine; (2) carbon connected to halogen; (3) carbon connected to sulfur. Surprisingly, these structures are all common reactive sites in organic reactions. For (1), carbonyl group and amine are essential in biochemical reactions to form amino acids. Many biochemical reactions involving amino acid analogs are also critical pathways for toxicity mechanisms. For (2) and (3), Halogen and sulfur

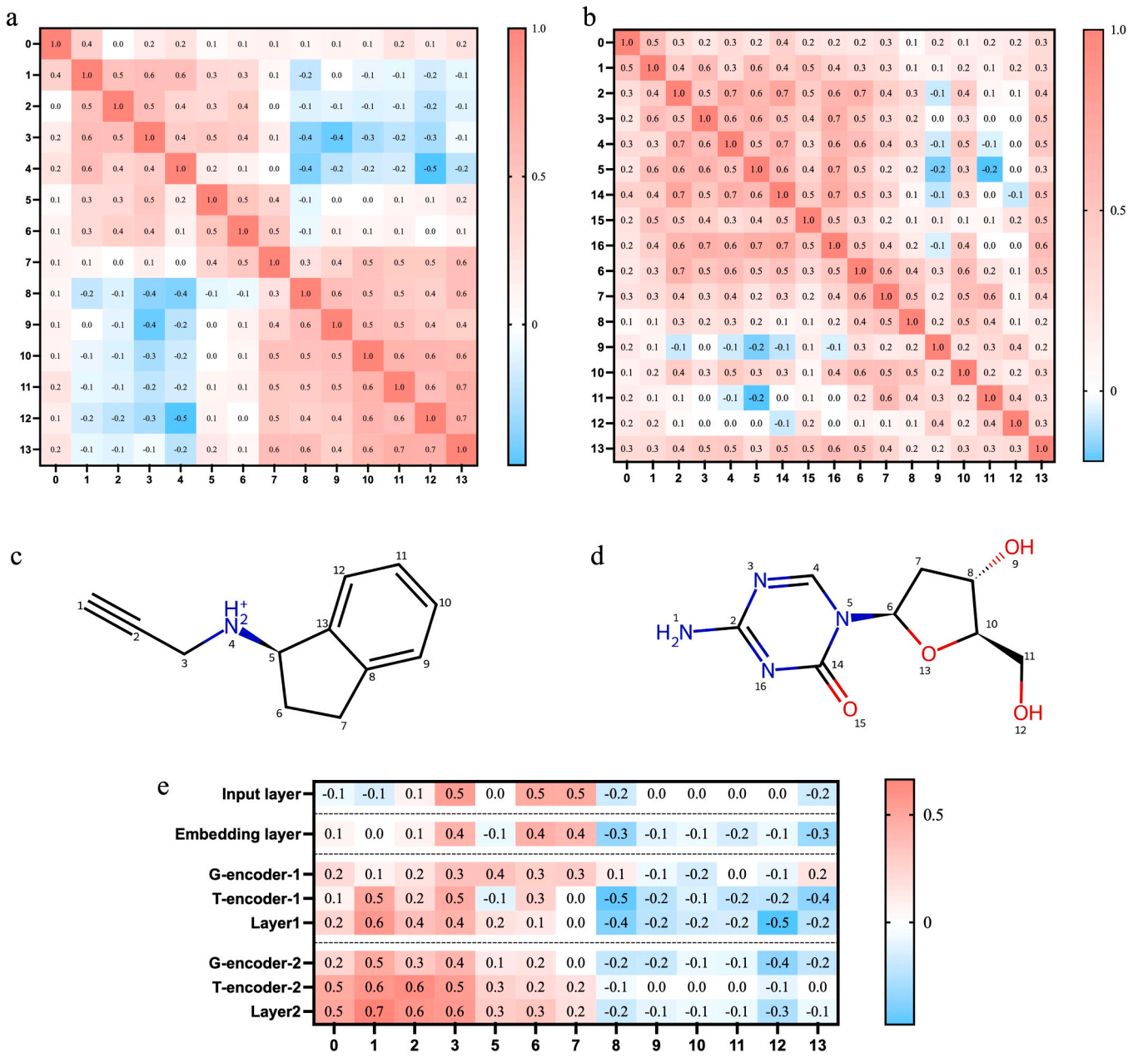


Fig. 6. Heat maps of atom similarity matrix for [(1R)− 2,3-dihydro-1 H-inden-1-yl]-prop-2-ynylazanium and Decitabine; (a) Heat map of atom similarity matrix for [(1R)− 2,3-dihydro-1 H-inden-1-yl]-prop-2-ynylazanium; (b) Heat map of atom similarity matrix for Decitabine; (c) Molecule structure of [(1R)− 2,3-dihydro-1 H-inden-1-yl]-prop-2-ynylazanium; (d) Molecule structure of Decitabine; (e) Similarity between the positively charged nitrogen and other atoms (atom 0 represents the virtual node) during the whole learning process.

possess relatively high electronegativity and large radius, making them good leaving groups in reactions that often result in positive carbon ions with high reactivity. Thus, through similarity analysis with the virtual node, many reactive sites can be detected, and a high similarity between atoms and virtual node indicates higher potential to be toxic. This phenomenon provides us a new method to detect toxic sites, which is distinct from previous study mainly relying on attention weights [25, 33]. For non-toxic molecules, it's obvious that these molecules are labeled with lighter color, representing lack of representative atoms. Due to the limitation of dataset volume, the positive-charged nitrogen tends to be identified as non-toxic sites. How to deal with the scarcity of toxic datasets is a crucial problem for more accurate identification. Besides, the proposed similarity-based toxic sites detection relies on atom-level similarity, which is different from the molecule-level similarity analysis in read-across. But combining the molecule representation learned by Hi-MGT with more structural, physicochemical and derived similarity-based descriptors is a possible way to further enhance the identification [2,3].

Moreover, these molecules also provide some interesting insights. In the sixth molecule, despite having only a slight deviation from a symmetric structure, the labeled colors are noticeably asymmetric, demonstrating Hi-MGT's ability to accurately capture local details. Besides, in toxic molecules, some aromatic carbons or nitrogens are also highlighted with deep red, which is another representation for atom clustering in the conjugated system and indicates that the electrons of toxic sites are not only from themselves but also from the potential conjugated system. In conclusion, the above-mentioned facts demonstrate that virtual nodes can offer valuable insights into potential toxic sites, and Hi-MGT has the capability to learn chemical patterns that align well with chemical intuitions in toxicity identification.

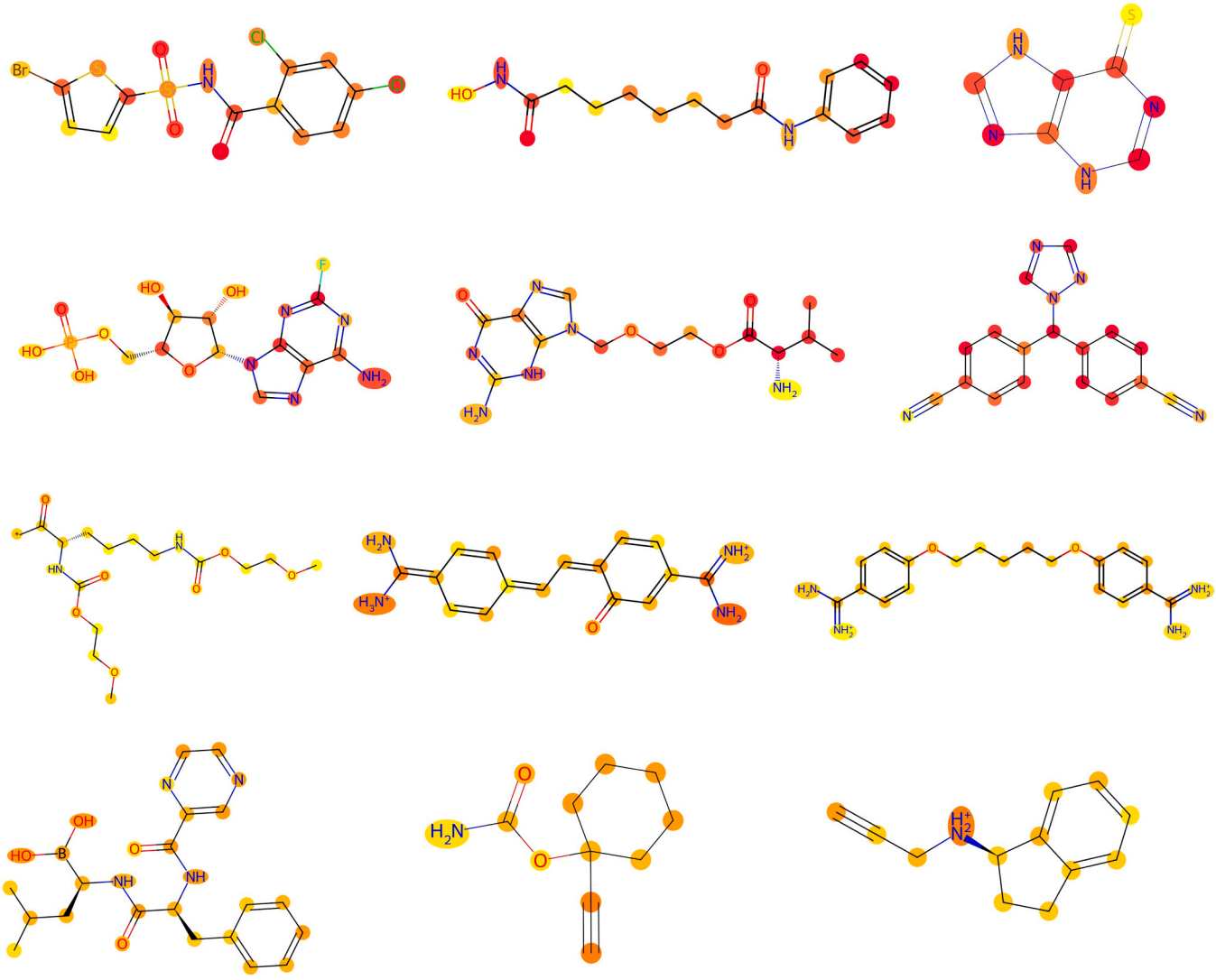


Fig. 7. Typical toxic and non-toxic molecules in ClinTox colored based on similarity between virtual nodes with other atoms. The molecules are divided into two groups: the first six are toxic, while the remaining six are non-toxic. Atoms with higher similarity to the virtual nodes are depicted in redder colors.

4. Conclusions

In this study, a novel toxicity hybrid model architecture, called Hi-MGT was proposed for toxicity identification. GNN-GT combination allows Hi-MGT to simultaneously and comprehensively aggregate local and global structural information of molecules in each layer, resulting in greater power to capture more meaningful toxicity information hidden in molecule graphs. Notably, Hi-MGT outperforms all other nine current CML, GNNs and Transformer-based models obviously in six out of eight tasks and exhibit identification power similar to large-scale pretrained GNNs with 3D geometry enhancement. The AUC_ROC of the prediction on ClinTox reaches 0.982 despite the distinct class imbalance of datasets, with only 12 toxic chemicals not being recalled on ClinTox, demonstrating great potential for applications. The interpretation analysis demonstrates that Hi-MGT accurately extracts many chemical phenomena aligning with chemical intuition, indicating its potential as a powerful tool for toxicity identification and analysis.

Environmental implication

Given the increasing exposure of humans to a diverse range of chemicals, toxic compounds present a significant risk to human safety. Consequently, the identification of the potential toxicity of chemicals is of utmost importance for effective chemical management. Traditional toxicity testing methods, reliant on animal experimentation, are widely regarded as resource-intensive, time-consuming, and ethically controversial. In this context, this study introduces a novel hybrid graph transformer model, termed Hi-MGT, which represents a significant advancement in the development of non-animal testing approaches for toxicity identification. The proposed model displays promising implications for enhancing human safety in the identification of chemical compounds.

CRediT authorship contribution statement

Zhichao Tan: Writing - original draft. **Youcai Zhao:** Conceptualization, Supervision. **Tao Zhou:** Conceptualization, Supervision. **Kunsen Lin:** 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.

Data availability

Data will be made available on request.

Acknowledgment

This work was financially supported by Shanghai Science and Technology Committee 2021 "Science and Technology Innovation Action Plan", Key Technologies and Equipment Development and Demonstration of Intelligent Management System for Hazardous Waste (No. 21DZ1201502) and the China Scholarship Council (No. 202206260111).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2023.131808.

References

[1] An, X., Chen, X., Yi, D., Li, H., Guan, Y., 2022. Representation of molecules for drug response prediction. Brief Bioinform 23 (1). https://doi.org/10.1093/bib/bbab393.
[2] Banerjee, A., Roy, K., 2022. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Mol Divers 26 (5), 2847–2862. https://doi.org/10.1007/s11030-022-10478-6.
[3] Banerjee, A., Roy, K., 2023. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points. Chem Res Toxicol 446–464. https://doi.org/10.1021/acs.chemrestox.2c00374.
[4] Banerjee, P., Siramshetty, V.B., Drwal, M.N., Preissner, R., 2016. Computational methods for prediction of in vitro effects of new chemical structures. J Chemin 8, 51. https://doi.org/10.1186/s13321-016-0162-2.
[5] Cao, D.-S., Huang, J.-H., Yan, J., Zhang, L.-X., Hu, Q.-N., Xu, Q.-S., et al., 2012. Kernel k-nearest neighbor algorithm as a flexible SAR modeling tool. Chemom Intell Lab Syst 114, 19–23. https://doi.org/10.1016/j.chemolab.2012.01.008.
[6] Cremer, J., Sandonas, L.M., Tkatchenko, A., Clevert, D.-Ae, Fabritiis, G.D., 2023. Equivariant graph neural networks for toxicity prediction. ChemRxiv. https://doi.org/10.26434/chemrxiv-2023-9kb55.
[7] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., et al., 2015. Convolutional networks on graphs for learning molecular fingerprints. arXiv 1509, 09292. https://doi.org/10.48550/arXiv.1509.09292.
[8] Fernandez, M., Ban, F., Woo, G., Hsing, M., Yamazaki, T., LeBlanc, E., et al., 2018. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. J Chem Inf Model 58 (8), 1533–1543. https://doi.org/10.1021/acs.jcim.8b00338.
[9] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. arXiv, 1704.01212. https://doi.org/10.48550/arXiv.1704.01212.
[10] Huang, R., Xia, M., Nguyen, D.-T., Zhao, T., Sakamuru, S., Zhao, J., et al., 2016. Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front Environ Sci 3. https://doi.org/10.3389/fenvs.2015.00085.
[11] Jeong, J., Choi, J., 2022. Artificial intelligence-based toxicity prediction of environmental chemicals: future directions for chemical management applications. Environ Sci Technol 56 (12), 7532–7543. https://doi.org/10.1021/acs.est.1c07413.
[12] Jiang, J., Wang, R., Wei, G.W., 2021. GGL-Tox: geometric graph learning for toxicity prediction. J Chem Inf Model 61 (4), 1691–1700. https://doi.org/10.1021/acs.jcim.0c01294.
[13] Koutsoukas St., A., Amand, J., Mishra, M., Huan, J., 2016. Predictive toxicology: modeling chemical induced toxicological response combining circular fingerprints with random forest and support vector machine. Front Environ Sci 4. https://doi.org/10.3389/fenvs.2016.00011.
[14] Kreuzer, D., Beaini, D., Hamilton, W.L., Létourneau, V., Tossou, P., 2021. Rethinking graph transformers with spectral attention. arXiv 2106, 03893.
[15] Li, H., Zhao, D., Zeng, J., 2022. KPGT: Knowledge-guided pre-training of graph transformer for molecular property prediction, In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 857–867.
[16] Li, X., Chen, L., Cheng, F., Wu, Z., Bian, H., Xu, C., et al., 2014. In silico prediction of chemical acute oral toxicity using multi-classification methods. J Chem Inf Model 54 (4), 1061–1069. https://doi.org/10.1021/ci5000467.
[17] Liu, R., Madore, M., Glover, K.P., Feasel, M.G., Wallqvist, A., 2018. Assessing deep and shallow learning methods for quantitative prediction of acute chemical toxicity. Toxicol Sci 164 (2), 512–526. https://doi.org/10.1093/toxsci/kfy111.
[18] Mayr, A., Klambauer, G., Unterthiner, T., Hochreiter, S., 2016. DeepTox: toxicity prediction using deep learning. Front Environ Sci 3. https://doi.org/10.3389/fenvs.2015.00080.
[19] Maziarka, Ł., Danel, T., Mucha, S., Rataj, K., Tabor, J., Jastrzębski, S., 2020. Molecule attention transformer. arXiv. https://doi.org/10.48550/arXiv.2002.08264.
[20] Müller, L., Galkin, M., Morris, C., Rampášek, L., 2023. Attending to graph transformers. arXiv 2302, 04181. https://doi.org/10.48550/arXiv.2302.04181.
[21] Rampášek, L., Galkin, M., Dwivedi, V.P., Luu, A.T., Wolf, G., Beaini, D., 2022. Recipe for a general, powerful, scalable graph transformer. arXiv 2205, 12454. https://doi.org/10.48550/arXiv.2205.12454.
[22] Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., et al., 2020. Self-supervised graph transformer on large-scale molecular data. arXiv 2007, 02835. https://doi.org/10.48550/arXiv.2007.02835.
[23] Sorkun, M.C., Koelman, J.M.V.A., Er, S., 2021. Pushing the limits of solubility prediction via quality-oriented data selection. Iscience 24 (1). https://doi.org/10.1016/j.isci.2020.101961.
[24] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2017. Graph attention networks. arXiv 1710, 10903. https://doi.org/10.48550/arXiv.1710.10903.
[25] Wang, H., Wang, Z., Chen, J., Liu, W., 2022. Graph attention network model with defined applicability domains for screening PBT chemicals. Environ Sci Technol 56 (10), 6774–6785. https://doi.org/10.1021/acs.est.2c00765.
[26] Wu, Z., Jain, P., Wright, M.A., Mirhoseini, A., Gonzalez, J.E., Stoica, I., 2022. Representing long-range context for graph neural networks with global attention. arXiv 2201, 08821. https://doi.org/10.48550/arXiv.2201.08821.
[27] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., et al., 2018. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9 (2), 513–530. https://doi.org/10.1039/c7sc02664a.
[28] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., et al., 2020. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem 63 (16), 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959.
[29] Peisong Niu, Zhou Tian, Wen Qingsong, Sun Liang, Yao Tao Chemistry Guided Molecular Graph Transformer, NeurIPS 2022. AI for Science: Progress and Promises; 2022.
[30] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., Di He, Shen, et al., 2021. Do transformers really perform badly for graph representation. arXiv 2106, 05234. https://doi.org/10.48550/arXiv.2106.05234.
[31] Yu, M.S., Lee, J., Lee, Y., Na, D., 2020. 2-D chemical structure image-based in silico model to predict agonist activity for androgen receptor. BMC Bioinform 21 (Suppl 5), 245. https://doi.org/10.1186/s12859-020-03588-1.
[32] Zhang, Z., Guan, J., Zhou, S., 2021. FraGAT: a fragment-oriented multi-scale graph attention model for molecular property prediction. Bioinformatics 37 (18), 2981–2987. https://doi.org/10.1093/bioinformatics/btab195.
[33] Zhu, W., Zhang, Y., Zhao, D., Xu, J., Wang, L., 2022. HiGNN: a hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.2c01099.