World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3077

PREDICTING DRUG TOXICITY ON TOX21 WITH A MULTI-TASK GNN-TRANSFORMER MODEL

YuHui Li
School of pharmacy, Shenyang Pharmaceutical University, Benxi 117004, Liaoning, China.

Abstract: As research into medicinal chemical synthesis deepens, an increasing number of novel drug molecules with promising clinical therapeutic potential are emerging. During drug development, many compounds are discarded due to excessive toxic side effects, while traditional toxicity testing faces challenges of high cost and lengthy timelines. To enable rapid toxicity assessment by researchers, this paper proposes a drug toxicity prediction model (TGT) based on GNN-Transformer. The model was constructed and validated using the Tox21 dataset, with caffeine selected for practical generalisation testing. The Tox21 dataset encompasses toxicity test results for diverse compounds, providing high-quality data. The model architecture leverages graph neural networks (GNN) to process molecular graph-structured data, effectively capturing complex topological relationships and chemical information within molecules. This transforms molecular graph structures into meaningful node feature representations. The Transformer component, with its exceptional sequence modelling capabilities, further learns from GNN-extracted features, capturing long-range dependencies between molecular structures. Through training and optimisation, the model demonstrated commendable performance in toxicity prediction tasks, achieving an average AUC of 0.7488 on the validation set. Its high accuracy in predicting drug toxicity was further validated through practical application on caffeine molecules, establishing it as an efficient and precise predictive tool for early-stage drug safety assessment.The GNN-Transformer drug toxicity prediction model proposed in this study enhances prediction reliability by integrating multi-task learning with interpretability techniques. It serves as an auxiliary pre-screening tool for drug development, thereby helping to shorten the research and development cycle and reduce costs.

Keywords: GNN-Transformer model; Drug toxicity prediction; Tox21 dataset; Molecular structure data

1 INTRODUCTION

With the continuous advancement of pharmaceutical technology, an increasing number of novel drug molecules with clinical potential are being successfully synthesised. During drug development, accurate assessment of human toxicity is paramount. However, traditional methods face significant challenges: despite many candidate drugs demonstrating favourable safety profiles in animal models, over 30% of drugs ultimately fail due to human toxicity emerging during clinical trials. Statistics indicate that completing a full toxicity testing regimen for a single compound takes approximately 3.5 years on average, with global annual expenditure on toxicological experiments reaching around US$3 billion [1]. Conventional in vivo and in vitro toxicity assessment methods are not only time-consuming and labour-intensive but also entail ethical controversies.

To address this predicament, the US Food and Drug Administration (FDA) and the National Institute of Environmental Health Sciences (NIEHS), among other agencies, jointly launched the Tox21 initiative in 2008. This programme aims to develop faster, more efficient novel methods for toxicity prediction [2]. The initiative's core achievement is the Tox21 database, which collates test data for approximately 10,000 environmental chemicals and pharmaceuticals across 12 toxicity-related pathways (such as nuclear receptor signalling and stress response pathways) [3], providing invaluable resources for computational toxicology research.

In recent years, leveraging artificial intelligence to predict drug toxicity has emerged as a research hotspot. Numerous studies have attempted to apply machine learning (ML) and deep learning (DL) algorithms to analyse large-scale toxicity datasets for early identification of potential risks [4]. For instance, in a Tox21-based predictive model, Idakwo et al. employed methods such as random forests (RF) and support vector machines (SVM), yet achieved accuracy (ACC) ranging only between 0.58 and 0.79 [5]. Analysis suggested that data imbalance may have constrained performance. In 2016, Mayr et al.'s DeepTox model utilised deep neural networks (DNN), achieving AUC values exceeding 0.90 in two out of twelve prediction tasks and demonstrating overall superior performance [6]. Li Jianqing et al.'s model, constructed using XGBoost, also attained an average AUC of 0.84 [1].

Despite these successful attempts, existing methods exhibit notable limitations: (1) most models lack validated robustness under missing atomic feature scenarios; (2) common input formats (e.g., SMILES strings) inadequately capture complex interactions between atoms and functional groups, potentially yielding numerous false positives/negatives; (3) models generally suffer from weak interpretability and generalisation capabilities.

Concurrently, graph neural networks (GNNs) have emerged as powerful graph representation learning tools, finding extensive application in fields such as drug discovery [7]. However, their inherent limitations in message passing mechanisms—including restricted layer depth and susceptibility to over-smoothing—leave room for improvement in representational capacity. Conversely, Transformer models have achieved breakthroughs across multiple domains through their robust global information capture capabilities. However, when processing molecular structures, they exhibit insufficient local feature extraction, particularly when performance is susceptible to limited or noisy data. Integrating both approaches can better harness molecular structural features, thereby enhancing prediction reliability and robustness while reducing false positive/false negative outcomes [8].

This paper therefore aims to investigate the influence of stereochemical structures in key pharmaceutical compounds upon compound toxicity, whilst endeavouring to establish a high-performance GNN-Transformer coupled model to provide guidance for contemporary drug toxicity assessment.

2 CONSTRUCTION OF THE TOXIC GNN-TRANSFORMER (TGT) MODEL

The workflow for our model construction is shown in Figure 1.

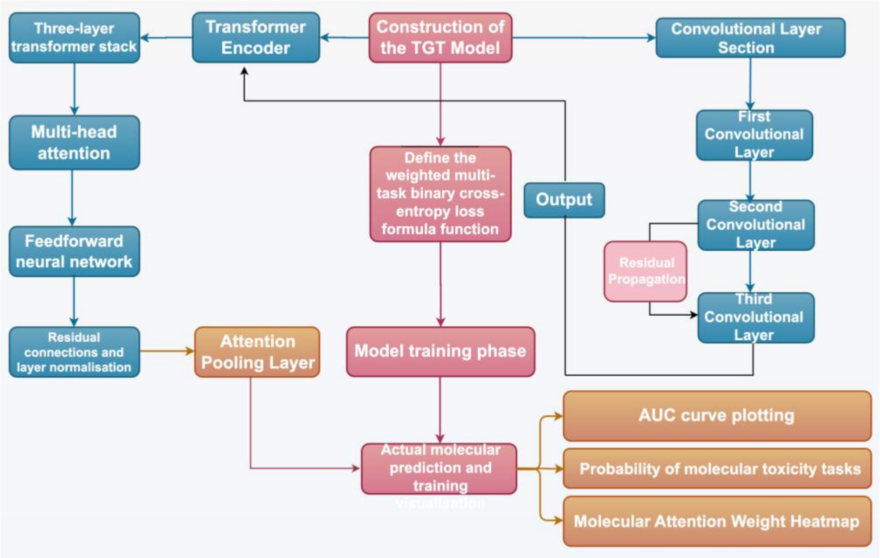
Three-layer transformer stack
Transformer Encoder
Construction of the TGT Model
Convolutional Layer Section
Multi-head attention
Define the weighted multi-task binary cross-entropy loss formula function
Output
First Convolutional Layer
Second Convolutional Layer
Residual Propagation
Third Convolutional Layer
Feedforward neural network
Residual connections and layer normalisation
Attention Pooling Layer
Model training phase
AUC curve plotting
Actual molecular prediction and training
Visualisation
Probability of molecular toxicity tasks
Molecular Attention Weight Heatmap

Figure 1 Technology Roadmap

2.1 Graph Neural Networks

To accelerate model training while ensuring training efficacy, we constructed a three-layer graph convolutional architecture for molecular local feature extraction. To mitigate the issue of neuron death, LeakyReLU was employed as the activation function with a negative slope of 0.2. Concurrently, we introduced residual connections from the second layer with a random dropout rate of 0.2 to enhance model robustness and prevent overfitting. Consequently, for the lth graph convolutional layer:

$$H^{(l)} = \sigma\left(\text{GraphConv}\left(H^{(l-1)}, Ei\right)\right) + H^{(l-1)} \tag{1}$$

where $\sigma = \text{LeakyReLU}(0.2)$.

2.2 Transformer Encoder

Following local feature extraction by the GNN, the graph data is first converted into a dense sequence format processable by the Transformer, thereby achieving semantic alignment between graph and sequence.

The Transformer encoder comprises three stacked layers of identical Transformers, each consisting of multi-head attention, a feedforward neural network (FFN), residual connections, and layer normalisation. Recognising the inherent topological structure of molecular frameworks, we employ the natural topology of molecular graphs to replace conventional positional encoding. Each encoder layer incorporates eight attention heads. The multi-head attention formula is as follows:

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{head}_1, \dots, \text{head}_{\,h})W^O \tag{2}$$

$$\text{head}_i = \text{softmax}\left(\frac{QW_i^Q\left(KW_i^K\right)^T}{\sqrt{d_k}}\right)VW_i^V \tag{3}$$

where $W_i^Q, W_i^K \in \mathbb{R}^{128\times16}, W_i^V \in \mathbb{R}^{128\times16}, W^O \in \mathbb{R}^{128\times128}$, and the key dimension $d_k = 16$.

In FFN, the number of hidden layers is quadrupled, expanding the dimensionality to 512 dimensions to enhance the model's non-linear expressive capability. Residual connections and layer normalisation are then applied to capture global interactions and stabilise the model training process.

2.3 Attention Pooling Layer

The attention pooling layer aggregates node-level features from the Transformer output into graph-level representations, incorporating learnable weights to dynamically fuse local and global features across twelve tasks. Its specific construction steps are as follows:

(1).The output of the Transformer encoder serves as the input for this section, namely:

$$H_{\text{trans}} \in \mathbb{R}^{B\times N_{max}\times d} \tag{4}$$

where B means the batch size, N_{max} represents the maximum number of atoms within a batch, and d is the feature dimension, with a value of 128.

(2).Through a two-layer fully connected network, node features are mapped to scalar attention scores a_i, calculated as follows:

$$a_i = W_2 \cdot \tanh\,(W_1 h_i + b_1) + b_2 \tag{5}$$

where W_1, b_1 are the parameters for the first layer, W_2, b_2 are the parameters for the second layer, and h_i is the feature vector for the ith node.

(3).Perform softmax normalisation on all nodes for each sample, ensuring that only unmasked, valid nodes participate in the computation.

(4).Perform weighted aggregation on the normalised nodes to form the final graph-level representation, with an output dimension of $\mathbb{R}^{B\times d}$.

Then construct twelve task classifiers, each of which outputs the predicted probability for its respective task.

2.5 Defining the Loss Function

To achieve a final output of 12 tasks and balance positive and negative samples within the dataset, we constructed a weighted multi-task binary cross-entropy loss formula. For the kth task, its loss function $\mathcal{L}^{(k)}$ is defined as:

$$\mathcal{L}^{(k)} = \frac{1}{N_{\text{valid}}^{(k)}} \sum_{i=1}^{B} \mathcal{M}_i^{(k)} \left[y_i^{(k)} \log \sigma\left(\hat{y}_i^{(k)}\right) + \frac{N_{neg}^{(k)}}{N_{\text{pos}}^{(k)} + \epsilon} \left(1 - y_i^{(k)}\right) \log\left(1 - \sigma\left(\hat{y}_i^{(k)}\right)\right) \right] \tag{6}$$

where $N_{valid}^{(k)}$ means the number of valid samples, $\sigma(x)$ represents the Sigmoid function, $N_{\text{pos}}^{(k)}$ indicates the number of positive samples for the kth task in the training set, $N_{neg}^{(k)}$ denotes the number of negative samples for the kth task in the training set, $y_i^{(k)}$ signifies the true label for the ith sample of the kth task, and $\hat{y}_i^{(k)}$is the predicted probability value for the ith sample of the kth task.

According to the dataset, the number of valid samples for all 12 tasks is greater than zero. Summing the loss values for each task yields the total loss value $\mathcal{L}_{total}$.

2.6 Model Training

The model is trained using the training set, whilst simultaneously calculating the AUC value for each toxicity task based on the validation and test sets. Learning rate adjustments and automatic iterative optimisation of the model are performed according to these AUC values. Concurrently, to prevent overfitting, an early stopping mechanism is introduced: training is terminated if the validation AUC fails to improve for ten consecutive iterations. The final output comprises the task with the highest average AUC across all twelve tasks, alongside the individual AUC values for each task at that point.

3 MODEL RESULTS AND ANALYSIS

3.1 Data Preprocessing

The compound toxicity dataset utilised in this experiment originates from the official Tox21 project website. This dataset maintains high data quality standards, has been in prolonged use, and possesses considerable reliability and modelling value. It encompasses bioactivity data for over 12,000 chemical substances, covering 12 distinct cell nucleus receptor-related toxicity prediction endpoints such as apoptosis, DNA damage, and endocrine disruption. These are represented as binary values (0 or 1), where 1 indicates activity and 0 denotes no activity.

We performed missing value removal and deduplication on the original Tox21 dataset, ultimately obtaining a dataset containing 7,823 compounds. The cleaned dataset was then partitioned into training, validation, and test sets in an 8:1:1 ratio.

To incorporate chemical molecular descriptors from the dataset into graph neural networks (GNNs), we standardise atomic features on the cleaned data and generate edge indices to represent the graph-based relationships within molecular structures, thereby conforming to the input format required by neural networks. To enhance the robustness of the constructed model, during the training phase, 10% of atomic features are randomly masked with a 30% probability. Gaussian noise is injected for data augmentation, and NaN values in the labels are replaced with -1 to denote invalid samples. Concurrently, the output dimension is adjusted to [1,12].

3.2 AUC Value Predicted by the Model

Area Under Curve (AUC) is defined as the area enclosed by the ROC curve and the coordinate axes. It serves as a core metric for evaluating the performance of binary classification models, indicating the model's ability to distinguish between positive and negative samples. Its value ranges between [0, 1]. By calculating the AUC value on the test set, it functions as the final ranking evaluation metric. AUC serves to evaluate whether a model possesses discriminative capability. General guidelines for assessing model performance using AUC are as follows: if 0.8 > AUC ≥ 0.7, performance is considered good; if 0.7 > AUC ≥ 0.6, performance is deemed average; and if 0.6 > AUC ≥ 0.5, performance is regarded as poor.

To provide clearer observation of the model training process, we visualise the changes in AUC and loss values during training, as depicted in Figure 2.

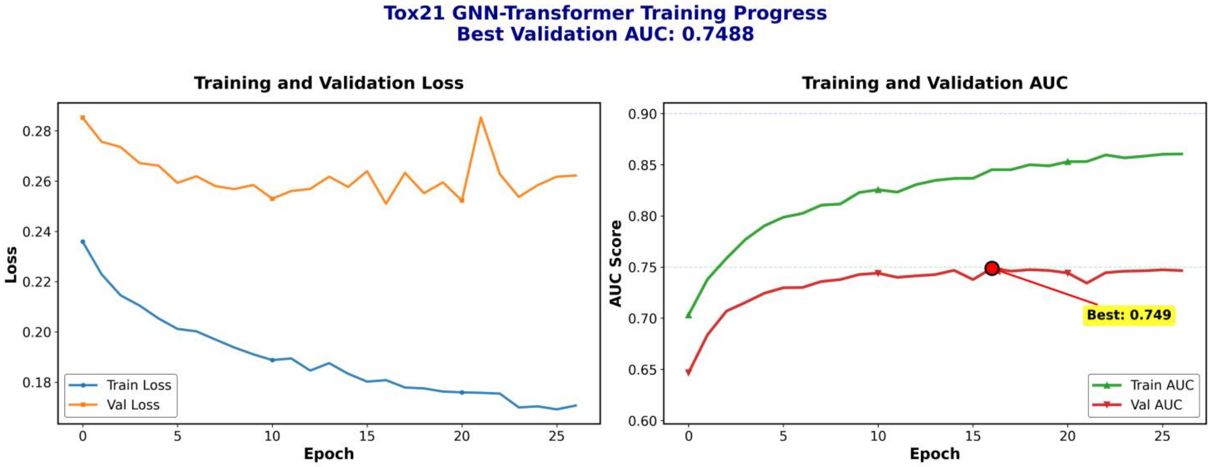

Figure 2 AUC Value Variation and Loss Value Variation Diagram

In Figure 2, the blue line represents the loss on the training set, the yellow line represents the loss on the validation set, the green line shows the change in the model's AUC value on the training set, and the red line shows the change in the model's AUC value on the validation set.

The model was ultimately deployed to make predictions on the test set. The AUC values for each task on the test set are presented in Tables 1 and 2.

Table 1 AUC for the First Six Tasks

Model	AR	AR-LBD	AhR	Aromatase	ER	ER-LBD
Our study	0.78	0.76	0.78	0.71	0.64	0.68
AMAZIZ	0.77	0.85	0.91	0.82	0.81	0.81
dmlab	0.83	0.82	0.78	0.84	0.77	0.77
T	0.68	0.85	0.91	0.83	0.78	0.81
microsomes	-	-	0.90	-	0.79	0.83
filipsPL	0.74	0.74	0.89	0.78	0.77	-

Table 2 AUC for the last six tasks

Model	PPAR-γ	ARE	ATAD5	HSE	MMP	p53
Our study	0.75	0.61	0.72	0.73	0.74	0.65
AMAZIZ	0.83	0.81	0.83	0.84	0.95	0.84
dmlab	0.83	0.77	0.8	0.86	0.95	0.88
T	0.82	0.8	0.81	0.81	0.94	0.85
microsomes	0.72	0.8	0.81	-	-	0.83
filipsPL	-	0.76	-	0.77	0.93	0.82

3.3 Analysis of Model Prediction Results

As shown in Figure 2, during the initial training phase, both the training set loss and validation set loss decreased, while their respective AUC values increased. This indicates that the model was learning effectively and performing well. As training progressed, the rate of decrease in the validation set loss slowed, and its AUC value stabilised, suggesting the model was approaching overfitting. At this point, the early stopping mechanism was activated, halting further training. The final model achieved an optimal AUC value of 0.7488 upon completion of training, indicating favourable predictive performance.

Tables 1 and 2 demonstrate our model's robust predictive capability, with the highlighted values representing the highest AUC scores. In the AR task, our model ranked second.

3.4 Application of the Model

To further validate the model's practicality and validity, we applied it to the caffeine molecule.
First, we input CN1C=NC2=C1C(=O)N(C(=O)N2C)C as the initial data into the model, then ran the model following the aforementioned construction procedure.
Subsequently, the attention weight heatmap for the caffeine molecule was generated, as shown in Figure 3.

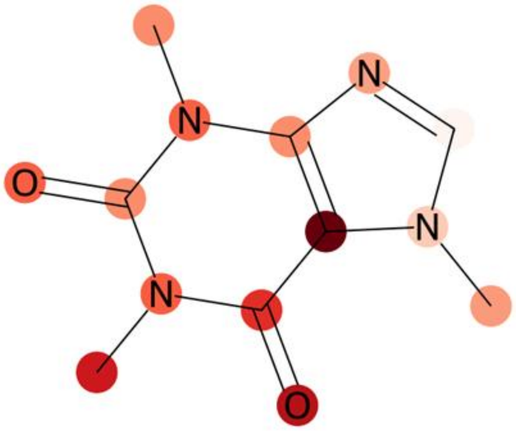

Figure 3 Attention Weight Heatmap of Caffeine Molecules

In Figure 3, darker colours indicate higher attention weights and greater contribution to the prediction of toxic activity.
To further enhance the model's interpretability and reliability, we took the NR-AR toxicity task as an example and plotted the absolute gradient values of each feature dimension for every atom in the caffeine molecule relative to the prediction outcome. This illustrates the contribution strength of each feature dimension for each atom to the current prediction task, as shown in Figure 4.

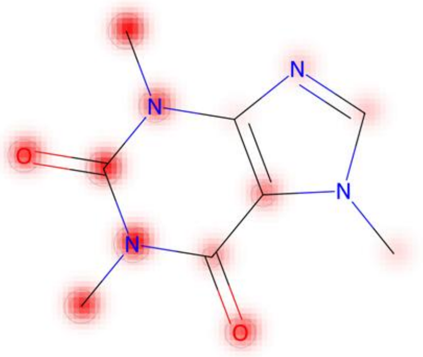

Figure 4 Plot of the Absolute Gradient Values for Each Feature Dimension in the Caffeine Molecule On The Prediction Results

In Figure 4, each small square represents a feature dimension. The darker the colour, the greater the contribution of that feature dimension for the atom in question to the prediction results for the NR-AR toxicity task.
The predicted results for the caffeine molecule across 12 toxicity tasks using the constructed TGT model are presented in Table 3. A probability closer to 1 indicates a higher likelihood of activity for that task metric, and consequently a higher overall toxicity probability.

Table 3 Predicted Results for the Caffeine Molecule Toxicity Task

Toxicity prediction endpoint	Probability
NR-AR	0.42413792
NR-AR-LBD	0.23928782
NR-AhR	0.30647254
NR-Aromatase	0.14343938
NR-ER	0.3543046
NR-ER-LBD	0.20357095

Toxicity prediction endpoint	Probability
NR-PPAR-γ	0.08609559
SR-ARE	0.2497363
SR-ATAD5	0.20533149
SR-HSE	0.13517842
SR-MMP	0.08458616
SR-p53	0.24755009

The results indicate that the probability for each task is less than 0.5, suggesting a low likelihood of caffeine exhibiting cytotoxicity, which is consistent with common sense. As demonstrated by the aforementioned predictions for caffeine, our model performs well in predicting toxicity across 12 tasks, delivering robust predictive capabilities and enabling relatively accurate toxicity predictions for novel drugs.

4 CONCLUSIONS AND OUTLOOK

This study constructs a novel drug toxicity prediction model based on the Tox21 dataset, employing a GNN-Transformer coupled architecture. It integrates dynamic data augmentation with early stopping mechanisms, enhances interpretability through attention pooling and gradient visualisation, and captures key chemical features relevant to molecular toxicity within a multi-task learning framework. This model addresses shortcomings in traditional approaches concerning global dependency modelling and local structural perception integration, enhancing feature learning comprehensiveness and prediction reliability to better align with practical toxicological scenarios. Key advantages include: firstly, improved model generalisation through dynamic data augmentation; secondly, enhanced feature representation by integrating GNN's local structural awareness with Transformer's global relationship modelling; Third, it incorporates an early stopping mechanism to suppress overfitting and accelerate convergence. This model serves as a complementary pre-experimental tool to traditional drug toxicity testing, providing insights for drugs with unknown toxicity profiles. It contributes to shortening drug development cycles and reducing experimental costs. Nevertheless, the model retains certain limitations. It may underfit in tasks with extremely sparse labels (e.g., the ARE task). Furthermore, while atomic random perturbations enhance robustness, they may moderately reduce prediction accuracy, potentially contributing to the model's suboptimal AUC performance. Future work could design more effective sample weighting or learning strategies for sparse tasks, seeking a better balance between stability enhancement and accuracy preservation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Li Jianqing, Wang Tianqin, Teng Yuefa, et al. Machine learning prediction model for emerging pollutants-induced activities of 12 nuclear receptors. Chromatography, 2025, 43(08): 959-970.

[2] Mori P M, Bortner D C, Touny E L, et al. Unusual and novel mitochondrial-driven mechanism of toxicity from Tox21 dataset. Free Radical Biology and Medicine, 2025, 240(S1): S29.

[3] Luo X, Zhang L, Sakamuru S, et al. Systematic evaluation of Tox21 compounds that target β-adrenergic receptors and their role in cardiotoxicity. Toxicology and applied pharmacology, 2025, 505117567.

[4] Thomas RS, Paules RS, Simeonov A, et al. The US Federal Tox21 Program: A strategic and operational plan for continued leadership. ALTEX, 2018, 35(2): 163-168.

[5] Gabriel I, Sundar T, Joseph L, et al. Structure－activity relationship-based chemical classification of highly imbalanced Tox21 datasets. Journal of Cheminformatics, 2020, 12(1): 66.

[6] Mayr A, EMayr A, EKlambauer G, et al. DeepTox: Toxicity Prediction using Deep Learning. Frontiers in Environmental Science, 2016, 3.

[7] Sun Y, Zhu D, Wang Y, et al. GTC: GNN-Transformer co-contrastive learning for self-supervised heterogeneous graph representation. Neural Networks, 2025, 181, 106645.

[8] Wei X H, Xu Z W, Wang X W, et al. Harnessing Geospatial and Temporal Information:GNN-Transformer Application to MJO Prediction. Journal of Jilin University (Natural Science Edition), 2025, 63(01): 67-75.