Computer Methods and Programs in Biomedicine 273 (2026) 109088

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

MEMOL: Mixture of experts for multimodal learning through multi-head attention to predict drug toxicity

Jae-Woo Chu [a], Jong-Hoon Park [a], Young-Rae Cho [a,b,*]

[a] Department of Software, Yonsei University Mirae Campus, Yeonsedae-gil 1, Wonju-si, 26493, Gangwon-do, Republic of Korea
[b] Department of Digital Healthcare, Yonsei University Mirae Campus, Yeonsedae-gil 1, Wonju-si, 26493, Gangwon-do, Republic of Korea

ARTICLE INFO

Dataset link: https://github.com/cjw0206/MEMOL

Keywords:
Drug toxicity prediction
Multimodal learning
Mixture of experts
Attention mechanisms
Graph neural networks

ABSTRACT

Background and Objective: Accurate toxicity prediction is essential in drug development, but several challenges remain due to the high complexity of molecular mechanisms. In this regard, toxicity prediction methods using deep learning have emerged as promising solutions, with multimodal approaches by integrating diverse molecular modalities. However, an effective fusion of multiple modalities to achieve precise toxicity predictions still requires further exploration.

Methods: In this study, we propose MEMOL, mixture of experts for multimodal learning through multi-head attention. This novel framework integrates modalities of molecular images, graphs, and fingerprints, and applies a sparse mixture of experts directly into the attention mechanisms. In addition, MEMOL enhances feature extraction and modality fusion through self- and cross-attention mechanisms, leading to improved predictive performance and efficiency.

Results: A comparative experiment on several toxicity benchmark datasets demonstrated that MEMOL outperformed state-of-the-art methods, achieving up to 8.33% higher AUROC and 9.11% higher AUPRC than the second-best model. Furthermore, the multimodal learning analysis showed that the employment of all three modalities consistently resulted in considerable performance improvements, and the ablation study revealed that a sparse mixture of experts enabled effective multimodal integration for drug toxicity prediction.

Conclusion: MEMOL demonstrates superior performance in toxicity prediction, achieving consistent improvements across diverse datasets and offering a robust framework for integrating molecular representations.

1. Introduction

The drug development process is highly structured and complex, involving extensive research, rigorous testing, and collaboration across various fields. The U.S. Food and Drug Administration (FDA) classifies this complex process into five key stages: discovery and development, preclinical research, clinical research, FDA review, and post-market safety monitoring [1].

During the discovery and development phase, researchers identify candidate compounds targeting specific disease mechanisms, aided by high-throughput screening and computer-aided drug design [2]. Promising candidates are then evaluated for metabolism, dosage, and safety before progressing to preclinical studies, where laboratory and animal experiments further assess biological activity and toxicity [3]. Clinical research follows with phased human trials: early stages focus on safety and dosage, while later phases test efficacy and monitor adverse reactions. If successful, a New Drug Application is submitted for FDA review, where trial data and manufacturing quality are examined. Upon approval, the drug enters post-market safety monitoring to track long-term risks [4]. Overall, drug development typically requires 10–15 years and costs $1–3 billion [5].

In particular, toxicity assessment is an essential part of the entire drug development process, playing a crucial role in ensuring safety for patients. However, traditional toxicity evaluation methods face several challenges. These studies are costly, and raise ethical concerns regarding animal testing. To address these limitations, artificial intelligence (AI) has emerged as a powerful tool for analyzing complex datasets and improving predictive accuracy [6,7]. Especially, deep learning approaches have significantly enhanced the performance of toxicity prediction models, with various architectures successfully applied to this domain [8–10].

For example, Transformer-based models are employed in toxicity prediction by processing drug-related sequences [11,12] or image data [13]. By encoding molecular structural formulas or clinical trial data as sequential representations using a Transformer encoder, these models effectively predict various toxicities. Vision Transformer

* Corresponding author at: Department of Software, Yonsei University Mirae Campus, Yeonsedae-gil 1, Wonju-si, 26493, Gangwon-do, Republic of Korea.
E-mail address: youngcho@yonsei.ac.kr (Y.-R. Cho).

https://doi.org/10.1016/j.cmpb.2025.109088
Received 9 April 2025; Received in revised form 9 September 2025; Accepted 24 September 2025
Available online 30 September 2025
0169-2607/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

(ViT) [14], in particular, segments molecular structure images into small patches, converts them into sequential representations, and analyzes inter-patch relationships. This method enables the model to capture spatial and visual patterns, facilitating precise identification of key structural features within molecules. In another example, Graph Neural Networks (GNNs) are applied to molecular graphs, where atoms correspond to nodes and chemical bonds to edges. By leveraging relational information within these graphs, GNNs effectively capture atomic interactions and overall structural properties relevant to toxicity prediction [15,16].

Multimodal learning [17] integrates diverse molecular data sources to provide a comprehensive understanding of drug properties. By fusing complementary information from multiple molecular representations, these approaches capture structural features that may be overlooked when relying on a single modality. For example, molecular images preserve spatial arrangements, molecular graphs encode atomic interactions, and fingerprints (FP) summarize chemical substructures in compact vectors. Leveraging such heterogeneous inputs has shown improved predictive performance in drug toxicity studies [13,18,19]. However, effectively integrating modalities remains challenging due to fundamental differences in data structures and representations.

To address these challenges, we propose a novel framework called MEMOL which is MoE-based multimodal learning with multi-head attention. Unlike previous multimodal models [13,18–20], MEMOL incorporates a sparse Mixture of Experts (MoE) [21–25] directly into the attention mechanisms [26]. In both Multi-head Self-Attention (MSA) and Multi-head Cross-Attention (MCA) [27], the MoE selectively routes modality-specific representations through specialized expert projections before computing attention. This architecture enables the model to dynamically emphasize the most informative representations from each modality and their interactions, going beyond static fusion schemes. By aligning expert specialization with the attention process, MEMOL establishes a new paradigm in toxicity prediction, enhancing both expressiveness and flexibility.

The experimental results demonstrated that MEMOL achieved state-of-the-art performance in drug toxicity prediction, outperforming the second-best model by 8.33% in AUROC and 9.11% in AUPRC. These findings validate the effectiveness of integrating sparse MoE into attention, highlighting both the complementary roles of different modalities and the novelty of expert-guided multimodal attention.

2. Preliminaries

2.1. Multi-head attention

The attention mechanism [27] models pairwise dependencies between elements, enabling the capture of contextual relationships. It is mathematically defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, V are query, key, and value matrices, and $\sqrt{d_k}$ is a scaling factor for numerical stability.

Self-attention uses $Q = XW_Q,\ K = XW_K,\ V = XW_V$, while cross-attention applies $Q = X_1W_Q,\ K = X_2W_K,\ V = X_2W_V$ from distinct sources. Both follow the same formulation but differ in the origin of Q, K, V.

Multi-head attention performs attention H times in parallel with different projections and concatenates the results:

$$\text{Multi_Head}(Q, K, V) = \text{Concat}(A_1, A_2, \dots, A_H), \tag{2}$$

$$A^h = \text{Attention}(Q^h, K^h, V^h), \quad h = 1, 2, \dots, H. \tag{3}$$

Here A^h is the output of the hth head, computed as in Eq. (1) with its own parameters. Concatenation of all heads yields a richer representation for downstream tasks.

2.2. Vision transformer

ViT adapts the Transformer architecture for image data. An input image is divided into non-overlapping $P\times P$ patches, each flattened into a token. Positional embeddings are added to preserve spatial order, and a class token is prepended so the model can output a global representation. This sequence is then processed by a stack of Transformer encoder layers.

Each encoder layer applies multi-head self-attention (MSA) and an MLP with residual connections [28] and layer normalization. The forward propagation is expressed as:

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad l = 1, \dots, L, \tag{4}$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l, \quad l = 1, \dots, L. \tag{5}$$

In Eq. (4), the input z_{l-1} is normalized and passed through MSA, after which a residual connection yields z'_l. Eq. (5) shows a second normalization, followed by the MLP and another residual addition, giving the final z_l. By stacking L layers, the ViT progressively integrates local patch information into a global representation.

2.3. Graph neural network

GNNs are designed to operate on graph-structured data, where nodes represent entities and edges capture their relations. By iteratively propagating information, GNNs learn node representations that reflect both local neighborhoods and global structure. Among the representative models, the Graph Attention Network (GAT) [29] and Graph Isomorphism Network (GIN) [30] are widely studied.

The GAT extends the Graph Convolutional Network [31] by introducing attention to weight neighbors adaptively. The embedding of node i is updated as follows:

$$h'_i = \sigma\left(\sum_{j\in N(i)} \alpha_{ij} W h_j\right), \tag{6}$$

where h_j is the feature of neighbor j, W is a learnable weight matrix, and α_{ij} is the normalized attention score. This allows the model to emphasize more informative neighbors when aggregating features.

The GIN focuses on expressive power for distinguishing graph structures. The embedding of node i is formulated as follows:

$$h'_i = \text{MLP}\left((1+\epsilon)\, h_i + \sum_{j\in N(i)} h_j\right), \tag{7}$$

where ϵ is a learnable parameter controlling the balance between self-information and neighbor aggregation. The summation ensures permutation invariance, and the MLP increases expressive capacity, making GIN theoretically as powerful as the Weisfeiler–Lehman test.

2.4. Molecular fingerprint

FPs convert molecular structures into fixed-length vectors for similarity search, property prediction, and screening. By encoding chemical information in compact form, FPs are widely used in cheminformatics and drug discovery.

Functional Class Fingerprints (FCFP) [32], derived from the Morgan algorithm [33], refine atom labels by iteratively incorporating neighboring atoms and bonds. This captures both structural details and functional roles such as bond types or donor/acceptor groups.

Pharmacophore 2D (Pharm2D) [34] emphasizes pharmacological features that influence molecular activity. It represents patterns such as hydrogen bond donors/acceptors, hydrophobic regions, and charged groups in a two-dimensional framework. By focusing on these functional properties, Pharm2D provides a complementary perspective to structure-based fingerprints.

2.5. Mixture of experts

The MoE is an advanced architectural paradigm that enhances the efficiency of large-scale models through conditional computation, where input data is selectively processed by multiple specialized expert networks. This approach allows the model to scale efficiently while reducing computational resources, making it particularly effective for handling high-dimensional data and intricate relationships. MoE has demonstrated state-of-the-art performance in the fields of large language models, natural language processing, and computer vision. It is especially advantageous for tasks that require processing vast amounts of data while maintaining computational efficiency. MoE can be categorized into dense MoE and sparse MoE based on how the input data are routed to expert networks.

2.5.1. Dense MoE

Dense MoE activates all expert networks simultaneously, aggregating their outputs into a unified representation that encapsulates diverse perspectives and complex features. The output of dense MoE is formulated as follows:

$$F_{\text{MoE}}^{\text{dense}}(x;\Theta,\{W_i\}_{i=1}^{N}) = \sum_{i=1}^{N} G(x;\Theta)_i f_i(x;W_i) \tag{8}$$

$$G(x;\Theta)_i = \text{softmax}(g(x;\Theta))_i \tag{9}$$

where x represents the input data, Θ denotes the parameters of the gating network, and W_i is the parameter of the ith expert network. The term $G(x;\Theta)_i$ corresponds to the gating network's computed weight, which dynamically modulates the contribution of each expert based on the input x. $f_i(x;W_i)$ is the output of the ith expert for input x and $g(x;\Theta)_i$ is the raw gate score of ith expert. These gating weights are normalized using the softmax function, such that $\sum_{i=1}^{N} G(x;\Theta)_i = 1$. During training, the gating function learns to assign higher weights to the most relevant experts, ensuring that the model efficiently leverages their specialized knowledge.

2.5.2. Sparse MoE

Sparse MoE uses a top-K routing strategy, where only a subset of experts is activated for each input, thus lowering computational cost. Its computation is defined as:

$$G(x;\Theta)_i = \text{softmax}(\text{Top_K}(g(x;\Theta),K))_i \tag{10}$$

Here, the Top_K function keeps the K largest gate scores and sets the rest to $-\infty$. After softmax normalization, only the selected K experts receive nonzero weights. The final output is then computed as in Eq. (8), but restricted to the selected experts. This selective routing greatly reduces computation while preserving model capacity, making sparse MoE particularly suitable for complex patterns.

3. Methods

In this section, we introduce a multimodal framework for processing molecular data to predict drug toxicity. Fig. 1-(a) provides an overview of the complete workflow of MEMOL. The input to the model is the Simplified Molecular Input Line Entry System (SMILES) [35] string of a molecule, from which image and graph modalities are derived using RDKit [34], while the FP modality is generated using MolFeat [36]. To extract representative features effectively, modality-specific embedding vectors are obtained using optimized feature extraction techniques. These embeddings are then processed through the MoE self-attention block, which incorporates multi-head self-attention with the MoE, enabling the model to capture intra-modality dependencies. Subsequently, the embeddings are passed through the MoE cross-attention block, where MCA is incorporated with MoE to facilitate inter-modality interactions. MEMOL first learns the relationships between image and graph modalities, and then integrates them with FP representations. After this multimodal fusion process, the model predicts the toxicity profile of the molecular structure, classifying it as either toxic or non-toxic.

3.1. Feature embedding

3.1.1. Image embedding

Using the RDKit library, images of drug molecules with three channels and a size of 224 × 224 were generated from SMILES strings and used as inputs for a pre-trained ViT model. This architecture leverages the feature representation capability of ViT, pre-trained on large-scale datasets, while allowing fine-tuning for specific downstream tasks. Feature extraction is performed using the transformation process previously described in Eqs. (4) and (5).

3.1.2. Graph embedding

To effectively process the graph structure, MEMOL uses GIN and GAT, as described in Section 2.3. Specifically, two GIN layers and two GAT layers are sequentially arranged in GIN-GAT-GIN-GAT order to extract features from a molecular graph. This hybrid approach leverages the strengths of both models: GIN effectively learns discriminative node-level representations, whereas GAT adaptively weights the contributions of neighboring nodes through attention. Finally, to integrate the extracted node-level features into a graph-level representation, global mean pooling is performed over all node embeddings.

3.1.3. Fingerprint embedding

FCFP and Pharm2D, generated by MolFeat, were used in MEMOL for molecular FPs. Because these FPs have high dimensionality, a two-layer autoencoder was designed and pre-trained to effectively compress and encode their essential features. The autoencoder consists of an encoder–decoder structure, where the encoder maps input FPs into a compact latent space, and the decoder reconstructs the compressed features to the original input. The advantage of using an autoencoder is not only to reduce dimensionality but also to retain the most informative features in FPs. After pre-training, only the encoder part of the autoencoder was integrated into the pipeline.

For the pre-training phase, the autoencoder was trained in an unsupervised setting. Specifically, the mean squared error between the input and output is used as the reconstruction loss to ensure that the learned latent space preserves the most salient molecular features. The input to the autoencoder was a 2048-dimensional vector obtained by concatenating the 1024-dimensional FCFP and Pharm2D. The encoder was composed of two fully connected layers, progressively reducing the input to a 256-dimensional latent representation. Conversely, the decoder employs two fully connected layers to restore the compressed representation back to its original 2048-dimensional space.

3.2. MoE attention block

Fig. 1-(b) and (c) illustrate the MoE self-attention block and MoE cross-attention block, respectively. Both modules integrate MoE with multi-head attention, enabling dynamic expert selection while effectively capturing contextual dependencies. The primary distinction between these mechanisms lies in their input configuration:

- **MoE self-attention** processes a single input X, using it to generate the query, key, value, gating networks and residual connection.
- **MoE cross-attention** takes two inputs, X and Y, where the query and residual connection are derived from X, while the key, value and gating networks come from Y.

In particular, in our model, MoE cross-attention is performed twice by swapping X and Y to capture relationships bi-directionally. For example, if we compute MoE cross-attention between image and graph modalities, we first set X and Y as the image and graph, respectively, and then we exchange them, setting X and Y as the graph and image, respectively. Furthermore, since X retains its information through the residual connection, generating the gating network from Y prevents

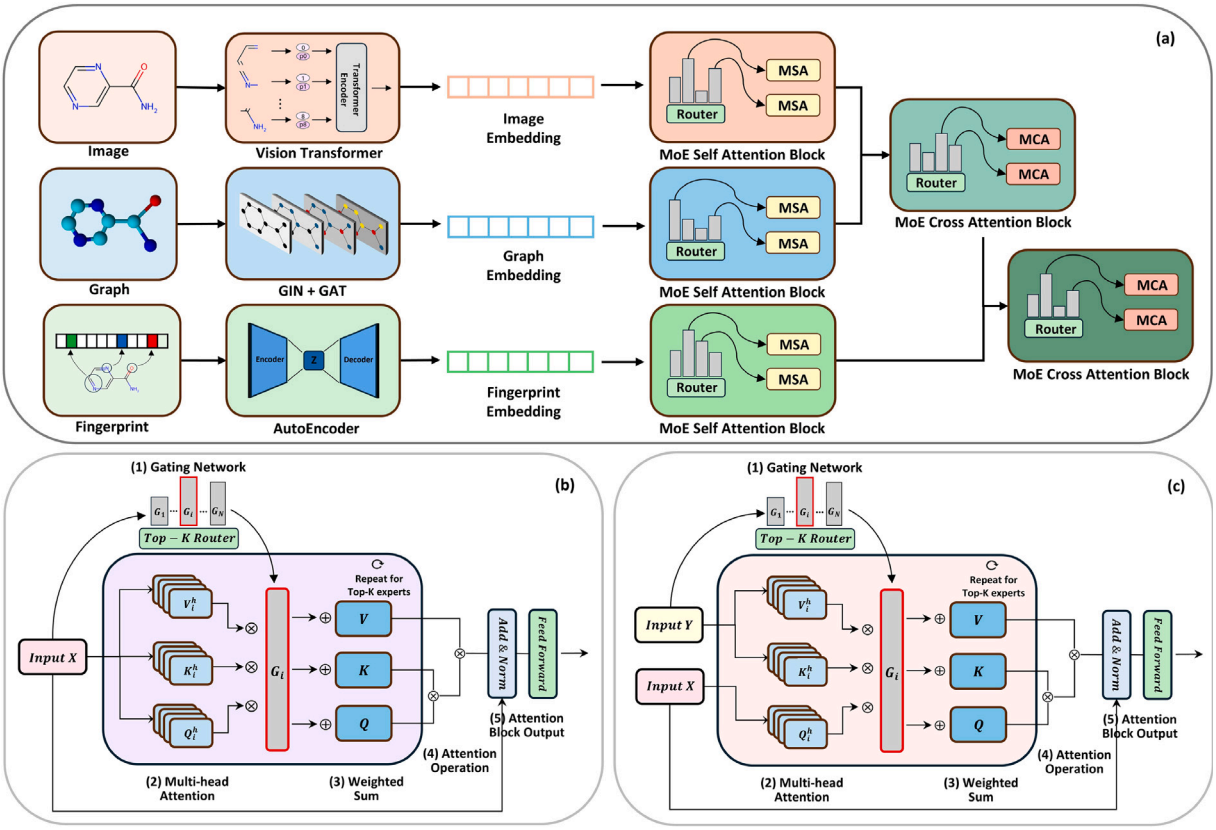


Fig. 1. Illustration of MEMOL architecture. (a) shows the overall framework. MEMOL processes three input modalities – image, graph, and FP – using a ViT, a GIN-GAT hybrid model, and an Autoencoder, respectively. The extracted embeddings pass through MoE self-attention blocks, followed by MoE cross-attention blocks to integrate modalities. (b) and (c) represent an MoE self-attention block and an MoE cross-attention block, respectively. (b) processes a single input X whereas (c) processes two inputs X and Y. In the MoE attention block, following steps are sequentially carried out: (1) A gating network selects top K experts based on computed scores, where the experts are generated from input Y. (2) Each selected expert independently performs multi-head attention, generating query, key, and value representations. (3) Expert outputs are aggregated via weighted summation. (4) Attention operation is performed using the accumulated Q, K, and V representations. (5) A feed-forward network with residual connections and layer normalization enhances stability.

excessive emphasis on the original information of X. Instead, it allows the model to flexibly adjust how X is fused with new information from Y, enabling more effective integration between the two inputs.

Despite this difference in the input structure, both mechanisms share the identical computational process described in Section 2.5.2. In the self-attention case, once the gate scores have been determined, query, key, and value matrices are generated as follows:

$$Q^{i,h} = XW_Q^{i,h}, \quad K^{i,h} = XW_K^{i,h}, \quad V^{i,h} = XW_V^{i,h} \tag{11}$$

where $i \in S$ represents the index of the selected experts, $h \in \{1, \ldots, H\}$ denotes the multi-head index, and $Q^{i,h}$, $K^{i,h}$, $V^{i,h}$ represent the query, key and value matrices, respectively, from the hth head of the ith expert. X refers to the input data and $W_Q^{i,h}$, $W_K^{i,h}$, and $W_V^{i,h}$ are the learnable weight matrices. In the cross-attention case, the only difference is that $K^{i,h} = YW_K^{i,h}$ and $V^{i,h} = YW_V^{i,h}$.

Subsequently, the gate score G_i weights the outputs of each expert across attention heads. The final weighted summation is given by:

$$Q^h = \sum_{i=1}^{S} G_i \cdot Q^{i,h}, \quad K^h = \sum_{i=1}^{S} G_i \cdot K^{i,h}, \quad V^h = \sum_{i=1}^{S} G_i \cdot V^{i,h}, \tag{12}$$

For the selected S experts, all computed values are accumulated, and the attention value X_{attn} is computed using the multi-head attention mechanism, as defined in Eqs. (2), (3). This integration of sparse MoE with attention enables more efficient and specialized feature extraction by dynamically selecting the most relevant experts. The final output then undergoes a residual connection, layer normalization, Feed-Forward Network (FFN), and dropout, as formulated as follows:

$$X_{\text{out}} = \text{Dropout}(\text{FFN}(\text{LayerNorm}(X + X_{\text{attn}}))) \tag{13}$$

The residual connection stabilizes training by preserving the original input while mitigating the vanishing gradient problem. Layer normalization normalizes the feature distributions, improving learning efficiency and accelerating convergence. The FFN, which is composed of two fully connected layers with a ReLU activation function, introduces non-linearity to enhance the expressive power of the model. Finally, dropout prevents overfitting by randomly deactivating neurons during training, thereby improving the generalization performance.

4. Results

4.1. Datasets and experimental settings

This study employs datasets from the Therapeutic Data Commons (TDC) [37] and two additional datasets obtained from previous research. The TDC repository provides 66 AI-ready datasets that are widely recognized for their reliability, enabling direct application to model training without preprocessing. These datasets facilitate various computational tasks, including drug–drug interactions, drug-target interactions, pharmacokinetics, and toxicity prediction, making them fundamental resources in AI-driven drug discovery. For this study, we selected three datasets from TDC:

- hERG evaluates the effects of a drug on the human Ether-à-go-go-Related Gene (hERG) channel, a critical factor in cardiac toxicity assessment.

Table 1
The number of compounds in each dataset before and after preprocessing, along with their distribution across training, validation, and testing sets.

Dataset name	Raw count	Processed count	Train	Valid	Test
hERG	655	643	449	64	130
DILI	475	457	321	47	89
Skin Reaction	404	404	283	40	81
Carcinogenicity	863	768	616	76	76
DILIst	1279	1097	874	112	111

- DILI determines whether a drug induces Drug-Induced Liver Injury (DILI), a major challenge in toxicity prediction.
- Skin Reaction captures adverse dermatological effects, including irritation and allergic responses.

Additionally, two datasets from the literature were incorporated into our experiments, listed below:

- Carcinogenicity [38] contains chemical compounds classified as carcinogenic or non-carcinogenic, sourced from the National Center for Toxicological Research Liver Cancer Database.
- DILIst [39] is an extended version of the FDA DILIrank dataset that incorporates additional data from the literature to improve the prediction of DILI.

During preprocessing, SMILES strings were refined by removing salts, which were identified as separate fragments in the notation using a period ('.') delimiter. Salts are counter-ions commonly added to enhance solubility or stability, but retaining them could introduce redundant molecular representations, leading to potential data noise. Additionally, compounds too large for FP conversion, too small for graph representation, or inorganic were excluded. Following these preprocessing steps, the final dataset was structured as summarized in Table 1. Since the TDC library provides pre-split datasets to facilitate fair comparison with other models, we followed the original partitioning strategy and used a 7:1:2 ratio for training, validation, and testing. In contrast, the Carcinogenicity and DILIst datasets, which were obtained directly from the literature, were divided using an 8:1:1 ratio. For model training, we used the Adam optimizer and conducted extensive hyperparameter tuning to enhance performance. The details of the explored ranges and the final configurations for each dataset are provided in Table S1 of the Supplementary Materials.

To assess the performance of the model, this study employs the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC) as evaluation metrics. AUROC measures classification performance by analyzing the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across various thresholds. TPR represents the proportion of correctly identified positive samples among all actual positives, whereas FPR denotes the proportion of negative samples misclassified as positives. Similarly, AUPRC evaluates model performance by analyzing the balance between precision and recall, making it particularly useful for imbalanced datasets. Both AUROC and AUPRC are widely adopted in classification tasks, providing a comprehensive assessment of model robustness and reliability.

4.2. Comparative study

To validate the effectiveness of MEMOL, we compared its performance with five conventional machine learning models and seven deep learning models. The conventional machine learning models include Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), Support Vector Machine (SVM), and Linear Regression (LR). The deep learning models considered for comparison are as follows:

- ChemBERTa [40] is a transformer-based model designed for molecular property prediction by pre-training on large-scale SMILES string. Leveraging RoBERTa [46] architecture, it learns contextualized chemical representations through masked language modeling, supporting strong transfer to biochemical and pharmacological tasks.
- MolCLR [41] is a self-supervised learning framework that uses GNNs to pre-train on large-scale molecular data via contrastive learning. It generates positive and negative molecular pairs, improving representation learning and downstream molecular property prediction.
- GraphMVP [42] is a multi-view pre-training model that combines 2D molecular graphs and 3D molecular geometric structures. It learns richer molecular representations by integrating topological and spatial information, making it effective even in tasks lacking explicit 3D data.
- MolFormer [43] is a transformer-based model pre-trained on over 1.1 billion unlabeled molecules using SMILES sequences. It applies linear attention and rotary positional embeddings to enhance molecular structure learning, particularly for structure–property relationship prediction.
- GIT-Mol [13] is a multimodal large language model that integrates graph, image, and text modalities for molecular science. Pre-trained on millions of molecules, it achieves notable improvements in molecular property prediction, captioning, and molecule generation tasks.
- AMPred-CNN [44] is a model based on convolutional neural network to process 2D molecular images. It extracts structural patterns using convolutional layers, leveraging image-based feature representations for toxicity prediction.
- CToxPred2 [45] is a cardiotoxicity prediction model that utilizes molecular FPs and descriptors within a semi-supervised learning framework. Trained on 2 million unlabeled molecules, it employs pseudolabeling and submodular optimization to enhance predictive accuracy.

Table 2 compares the performance of various models across five datasets previously described based on AUROC and AUPRC. MEMOL achieved 3.6% higher AUROC and 0.44% higher AUPRC compared to CToxPred2, the second-best performing model on the hERG dataset. For the DILI dataset, MEMOL outperformed MolFormer by 2.64% in AUROC and 1.89% in AUPRC. For the Skin Reaction dataset, MEMOL achieved the highest AUROC, showing a 2.13% improvement over GIT-Mol, and also recorded a 1.87% higher AUPRC compared to GraphMVP. Similarly, for the Carcinogenicity dataset, MEMOL recorded the highest AUROC, outperforming MolFormer by 3.02%. For the DILIst dataset, MEMOL demonstrated an 8.33% improvement in AUROC compared to GIT-Mol and a 9.11% increase in AUPRC over GraphMVP.

Although MEMOL showed lower AUPRC compared with the three other models on the Carcinogenicity, it still indicated a competitive performance. Overall, MEMOL achieved the best performance in all cases except one and outperformed the second-best model by up to 9.11%, highlighting its remarkable predictive capability across diverse tasks. In addition, we conducted further experiments with unified split ratios across all datasets, the details of which are provided in Table S2 of the Supplementary Materials. We also performed statistical significance analysis to verify the robustness of MEMOL's improvements over baseline models. The detailed methodology and results are provided in the Supplementary Materials (see Figure S1-S5).

4.3. Analysis of parameters in MoE

To evaluate the impact of the total number of experts (E) and selected experts by top K (K) in sparse MoE, we conducted experiments by varying E while setting K to 1 and 2, which are commonly used choices for top K selection [47]. Specifically, E was configured as {2, 4,

Table 2
Performance comparison of 13 predictive models across five toxicity benchmark datasets. The models were evaluated using AUROC and AUPRC metrics. The first five models listed are conventional machine learning methods, while the next seven are deep learning-based approaches.

Dataset	hERG		DILI		Skin Reaction		Carcinogenicity		DILIst	
Metric	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
RF	0.8455	0.9349	0.8721	0.8586	0.6706	0.7437	0.7120	0.7762	0.7271	0.6546
LGBM	0.8270	0.9330	0.8812	0.8975	0.6821	0.7303	0.5833	0.6837	0.6787	0.6892
XGB	0.8347	0.9408	0.8878	0.9085	0.7073	0.7729	0.5355	0.6248	0.7300	0.7439
SVM	0.7132	0.8435	0.3868	0.4896	0.7806	0.8016	0.6848	0.7631	0.6683	0.6766
LR	0.4243	0.7289	0.8094	0.7962	0.5957	0.7025	0.3826	0.5392	0.461	0.5085
ChemBERTa [40]	0.8272	0.9202	0.8873	0.9051	0.7749	0.8336	0.6986	0.8307	0.6949	0.6992
MolCLR [41]	0.8428	0.9364	0.9080	0.9152	0.7264	0.7905	0.6304	0.7726	0.6862	0.7006
GraphMVP [42]	0.7449	0.8752	0.8954	0.9237	0.7812	0.8808	0.7087	0.8288	0.7128	0.7488
MolFormer [43]	0.8356	0.9369	0.9211	0.9342	0.8042	0.8596	0.7217	**0.8411**	0.7138	0.7228
GIT-Mol [13]	0.8441	0.9386	0.9060	0.9085	0.8093	0.8774	0.7116	0.8047	0.7303	0.7264
AMPred-CNN [44]	0.7959	0.9152	0.7750	0.7839	0.7079	0.7868	0.6681	0.7550	0.6598	0.6346
CToxPred2 [45]	0.8691	0.9530	0.8721	0.9025	0.7526	0.8061	0.6924	0.7429	0.7149	0.6870
MEMOL	**0.9004**	**0.9572**	**0.9454**	**0.9519**	**0.8265**	**0.8973**	**0.7435**	0.8140	**0.7911**	**0.8170**

Table 3
AUROC results of MEMOL with varying numbers of experts and top K expert selection across five toxicity benchmark datasets. The column K represents the number of experts selected via top K routing, with K set to either 1 or 2. The column E denotes the total number of experts in the model, which varies among {2, 4, 8, 16, 32}. The proposed configuration, where $K = 2$ and $E = 4$, achieved the most balanced performance across all expert configurations.

TopK (K)	Num of experts (E)	hERG	DILI	Skin reaction	Carcinogenicity	DILIst
1	2	0.8863	0.9161	0.7710	0.7145	0.7118
	4	0.8997	0.9434	**0.8361**	**0.7507**	0.7537
	8	0.8752	0.9024	0.7819	0.7087	0.7368
	16	0.8835	0.9009	0.7742	0.7391	0.7385
	32	0.8704	0.8933	0.7851	0.6884	0.7333
2	2	0.8787	0.9272	0.7647	0.7014	0.7557
	4	**0.9004**	**0.9454**	0.8265	0.7435	**0.7911**
	8	0.8838	0.9075	0.7940	0.7319	0.7560
	16	0.8831	0.9277	0.7883	0.6819	0.7235
	32	0.8796	0.9232	0.8112	0.7094	0.7560

8, 16, 32}, and the same five toxicity benchmark datasets from the previous experiments were used for evaluation. The performance of each configuration was measured using the AUROC metric, as summarized in Table 3 and visualized in Fig. 2.

When $K = 1$, the model exhibited the highest AUROC values at $E = 4$ across all datasets. In general, increasing E resulted in a decline in AUROC for most datasets, with the lowest values occurring at $E = 32$ in all cases except for the Skin Reaction and DILIst datasets. This trend was particularly evident in the Carcinogenicity dataset, where AUROC dropped from 0.7507 at $E = 4$ to 0.6884 at $E = 32$.

For $K = 2$, the overall performance was higher than that of $K = 1$ across most experimental settings. Since the experiments were conducted using five toxicity benchmark datasets and five different expert configurations, a total of 25 cases were evaluated. Among these, $K = 2$ outperformed $K = 1$ in 17 out of 25 cases, demonstrating its superiority in performance. This suggests that selecting two experts during inference enhances the robustness of the model.

Especially, when K and E are 2 and 4, respectively, the model achieved the best performance for 3 datasets, and the second-best for 2 datasets. Notably, for the DILIst dataset, our model exhibited a substantial improvement in this setting, achieving an AUROC of 0.7911, which is 4.64% higher than the second-best configuration at $K = 2$, $E = 8$ and $K = 2$, $E = 32$. Similar to observation in $K = 1$, increasing E does not necessarily lead to improved performance in $K = 2$. While $K = 2$ generally led to better results than $K = 1$, performance degradation was still observed as E increased, indicating that a larger number of experts does not always contribute positively and may even introduce noise or redundancy, limiting the robustness of the model.

As illustrated in Fig. 2, AUROC values were lower for all settings where E exceeded four, and the overall trend showed a decline in performance as the number of experts increased. This trend was particularly evident for datasets such as hERG and Carcinogenicity, where excessive expert count led to noticeable performance drops. In summary, the experiments demonstrate that selecting $K = 2$ and $E = 4$ provides the most effective balance between expert selection strategy and predictive performance.

4.4. Impact of multimodal learning

4.4.1. Evaluation of modality combinations

We evaluated the contribution of each modality by conducting experiments with different combinations of modalities. Specifically, we evaluated three scenarios: using only one modality, integrating two modalities, and fusing all three modalities—image, graph, and FP. It should be noted that the model does not apply the cross-attention block when using a single modality. The performance was measured using AUROC and the results are shown in Fig. 3.

When the image modality was used alone, the accuracy of the model exhibited relatively low across most datasets, indicating that molecular images alone do not provide sufficient structural information. However, when image features were removed from the case where all three modalities were used, performance declined in all datasets. This suggests that while images alone are not highly informative, their integration with other modalities improves the accuracy of prediction. Notably, Skin Reaction and DILIst datasets showed a considerable decrease when image features were not included, demonstrating their complementary role.

Similar to images, the graph modality exhibited limited performance when used alone. However, when graphs were added to the integrated image and FP, performance improved across all datasets except hERG. The most notable gains were observed in Skin Reaction,

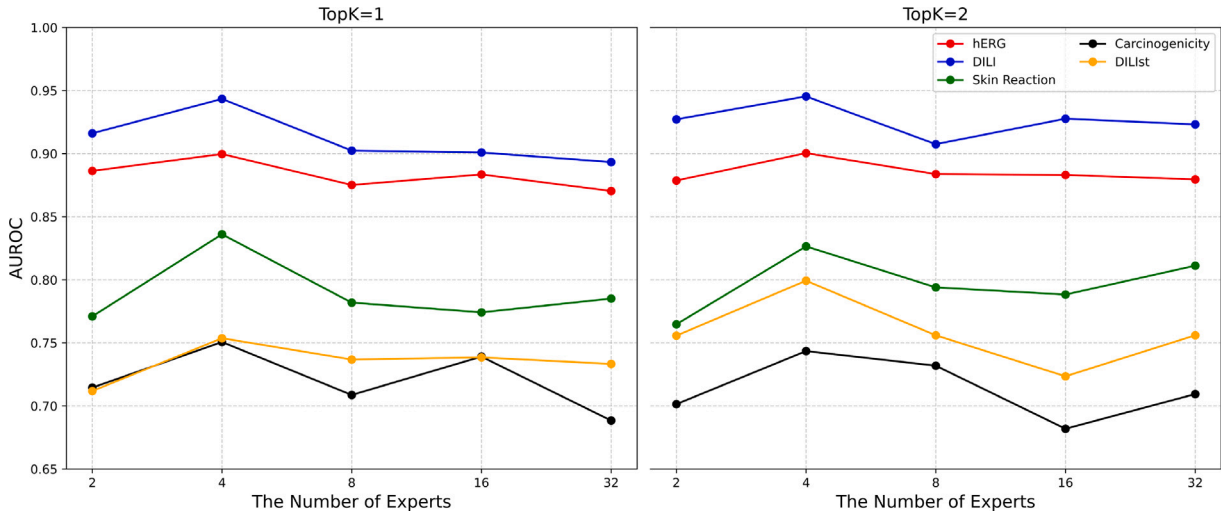


Fig. 2. Visualization of AUROC results of MEMOL with varying numbers of experts across five toxicity benchmark datasets. The left plot corresponds to $K = 1$, where a single expert is selected, while the right plot corresponds to $K = 2$, where two experts are selected.

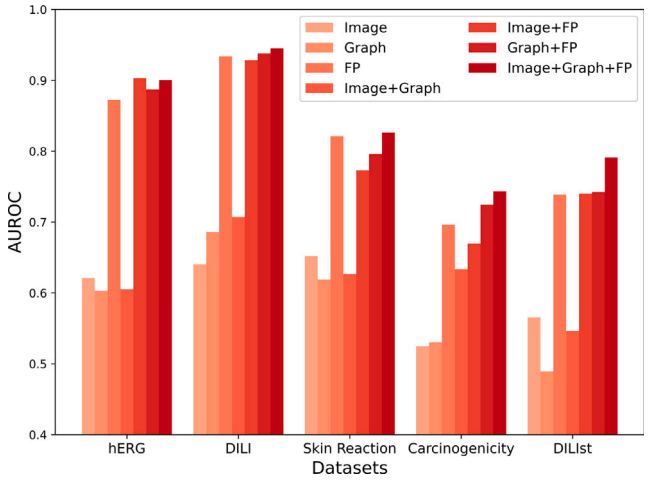


Fig. 3. Comparison of AUROC scores for seven different modality combinations across five benchmark datasets.

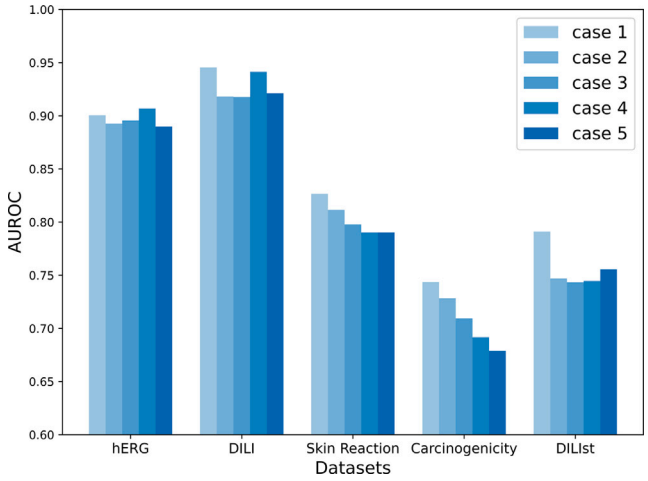


Fig. 4. Comparison of AUROC scores for five modality integration ordering cases across five benchmark datasets.

Carcinogenicity, and DILIst, indicating that incorporating graph-based representations enhances the ability of model to capture structural relationships and molecular interactions relevant to toxicity prediction.

FP performed well even when used alone, indicating their robustness in capturing relevant molecular features. However, integrating FP with images and graphs further improved performance. The best results were consistently observed when FP was integrated with the other two modalities, except hERG, highlighting their strong but complementary role. Although utilizing FP solely provided meaningful toxicity predictions, the use of additional structural information from graphs and images resulted in more comprehensive modeling. Details of the cross-attention weights can be found in Figure S6 of the Supplementary Materials. In conclusion, integrating multiple molecular modalities enables the model to capture diverse structural and chemical properties more effectively.

4.4.2. Assessment of modality integration order

We analyzed the impact of modality integration order by comparing different configurations of multimodal fusion using MoE cross-attention. Hereafter, we define $\otimes$ as MoE cross-attention. In this section, we examined five different cases, each representing a unique approach to fusing image, graph, and FP modalities. The cases are listed below:

- Case 1 ((Image $\otimes$ Graph) $\otimes$ FP): Image and graph modalities were first integrated using MoE cross-attention, followed by FP, which is the approach adopted in the MEMOL.
- Case 2 ((Image$\otimes$FP)$\otimes$Graph): Image and FP were first fused using MoE cross-attention, followed by graph.
- Case 3 ((Graph$\otimes$FP)$\otimes$Image): Graph and FP were first integrated using MoE cross-attention, followed by image.
- Case 4 (Concat(Image $\otimes$ Graph, Image $\otimes$ FP, Graph $\otimes$ FP)): All possible pairs of modalities were processed separately using MoE cross-attention, and the resulting embeddings were concatenated. This approach eliminates the effect of integration order.
- Case 5 (Concat(Image, Graph, FP)): No MoE cross-attention was applied; all three modalities were simply concatenated after the MoE self-attention block.

Fig. 4 presents the experimental results across five datasets by changing the order of modality integration. Case 1 consistently achieved the highest performance, except for hERG, where case 4 performed slightly better. This suggests that integrating image and

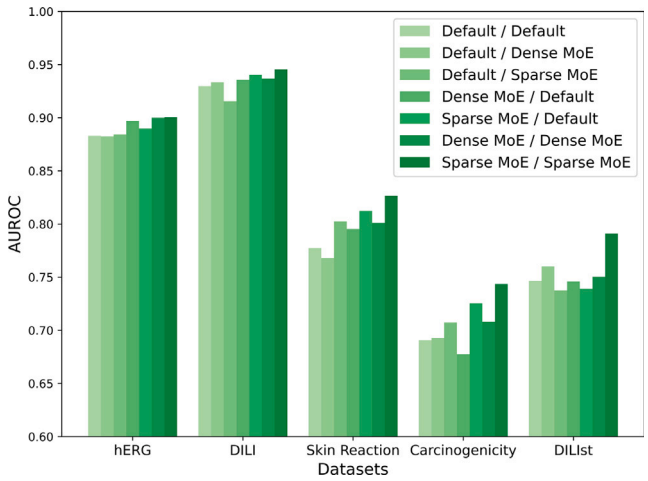


Fig. 5. AUROC results from the ablation study of the proposed model on five benchmark datasets. In the legend, the left part represents the method applied to self-attention, while the right part denotes the approach used for cross-attention.

graph first before incorporating FP, which was identified as the most informative modality in previous experiments, allows the model to extract complementary features more effectively and leads to optimal performance. In contrast, case 2 and case 3, where FP was integrated earlier, showed lower performance across all datasets. This indicates that fusing FP too early can disrupt the structured integration of image and graph features, resulting in suboptimal representations. Case 4, which removed fusion order by processing all modality pairs separately before concatenation, showed competitive results in certain datasets, such as hERG and DILI. However, it generally underperformed compared to case 1, highlighting the advantages of a structured fusion approach. Finally, case 5, which used simple feature concatenation without MoE cross-attention, showed low performance for all datasets. This highlights the importance of explicitly modeling inter-modality interactions. Overall, these experiments demonstrated not only the importance of usage of MoE cross-attention but also the order of modality integration.

4.5. Ablation study

To analyze the effect of MoE on self-attention and cross-attention blocks, we conducted an ablation study comparing different configurations. Specifically, we considered three types of attention: The first type refers to the default, which is a standard multi-head attention without MoE. The second type is multi-head attention using dense MoE where all experts contribute to computation. The last type is multi-head attention using sparse MoE which dynamically selects only a subset of experts. By evaluating various combinations of these configurations, we aim to assess the usage of MoE on attention mechanisms.

Fig. 5 presents the AUROC results using five benchmark datasets for several experimental settings. The results indicate that incorporating MoE with attention mechanisms generally improves performance compared to the default attention mechanism, demonstrating the benefits of expert-based attention mechanisms in multimodal learning. Among these configurations, the setting that applies sparse MoE to both attention blocks achieved the highest performance across all datasets, highlighting the effectiveness of expert selection in balancing computational efficiency and model expressiveness.

Utilizing MoE in both self- and cross-attention blocks proved beneficial. For dense MoE, the configuration using dense MoE for both blocks achieved higher accuracy than those applying it in only one block, except for the DILIst dataset. Similarly, sparse MoE exhibited the same trend as dense MoE. Notably, leveraging sparse MoE was more effective not only in terms of accuracy but also in computational cost (see Figure S7 of the Supplementary Materials for detailed comparisons of computational efficiency across baselines). Furthermore, a sparse MoE ensures that the model selects the most relevant experts, which leads to outperforming other cases.

5. Discussion

In this study, the proposed model, MEMOL, achieved superior performance across multiple benchmark datasets for drug toxicity prediction. Specifically, it outperformed existing models by up to 8.33% in AUROC and 9.11% in AUPRC, demonstrating its strong predictive capability and generalizability. This remarkable performance can be primarily attributed to three key components of the model: (1) multimodal learning, (2) MoE-based attention, and (3) sparse MoE. In this section, we discuss how each of these components contributes to the effectiveness of MEMOL.

One of the core strengths of MEMOL lies in its multimodal architecture that integrates heterogeneous molecular representations, including images, graphs, and FPs. When compared to employing individual modalities alone, the experimental results show that the employment of all three modalities consistently resulted in considerable performance improvements across all benchmark datasets (see Section 4.4.1). This implies that each modality captures complementary features of molecular characteristics, and that a more thorough and accurate understanding of molecular toxicity is made possible by their integration. In particular, FPs reflect chemical and pharmacological attributes, whereas molecular images and graphs mostly encode structural properties. Integrating these diverse viewpoints allows the model to make more accurate predictions.

To improve the effectiveness of multimodal learning, we incorporated the MoE mechanism into multi-head attention blocks. The ablation study demonstrated that the application of MoE generally led to improved performance when compared to the standard multi-head attention configuration (see Section 4.5). This suggests that leveraging expert selection within the attention mechanism enables more adaptive and specialized processing of input features. In particular, the MoE-based attention structure is capable of capturing complex modality-specific characteristics and effectively integrating complementary information while reducing the risk of overlap or conflict across modalities. By dynamically routing information through expert sub-networks, the model achieves enhanced representational capacity.

Furthermore, since MEMOL utilizes sparse MoE instead of dense MoE, the model has advantages not only improved computational efficiency but also high predictive accuracy. In dense MoE, all experts are activated for each input, which can increase computational cost and introduce redundancy in the learned representations. On the other hand, sparse MoE activates only a subset of experts selected based on the input, leading to more efficient computation with minimal impact on performance. Among all tested configurations, applying sparse MoE to both self-attention and cross-attention blocks yielded the highest AUROC scores across the benchmark datasets. These results suggest that selective expert activation enables the model to focus on relevant patterns more effectively, while maintaining a manageable computational load. Overall, sparse MoE appears to be a more suitable choice for multimodal toxicity prediction tasks, where both accuracy and efficiency are crucial.

6. Conclusion

Reliable drug toxicity prediction is a necessary component of drug development, yet it remains a challenging task due to the complexity of molecular mechanisms. In recent years, AI-driven methods have shown great potential in addressing this issue, particularly multimodal approaches that integrate various molecular data representations. Despite

their success, effectively integrating multiple modalities to enhance toxicity prediction still requires further advancement. To overcome these limitations, we propose MEMOL, an MoE-based multimodal framework for drug toxicity prediction through multi-head attention. By incorporating sparse MoE into both self-attention and cross-attention mechanisms, our model effectively refines modality-specific feature selection while facilitating structured cross-modal information exchange, ultimately improving the accuracy of toxicity predictions.

Experiments conducted on five benchmark datasets demonstrated that MEMOL outperforms state-of-the-art methods, surpassing the second-best model by 8.33% in AUROC and 9.11% in AUPRC. In particular, the configuration with $K = 2$ and $E = 4$ yielded the most robust performance across datasets, reinforcing the benefits of careful expert selection. The impact of modality combinations experiments confirmed that utilizing all three modalities is essential for capturing comprehensive molecular characteristics, while also proving the effectiveness of the approach of MEMOL in optimizing the integration order by first fusing image and graph representations before integrating FP representations. The ablation study further validated the effectiveness of sparse MoE in both self-attention and cross-attention mechanisms, highlighting the importance of selecting the most relevant experts for each input.

Future work will explore self-supervised learning on the modalities used in this study to address the lack of labeled data, enabling experiments on larger-scale datasets. Additionally, enhancing model interpretability through explainable AI techniques will help build trust in AI-driven toxicity prediction.

CRediT authorship contribution statement

Jae-Woo Chu: Writing – original draft, Validation, Software, Data curation. **Jong-Hoon Park:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Young-Rae Cho:** Supervision, Investigation, Funding acquisition, Conceptualization.

Ethics statement

This research does not involve human participants and animal studies.

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.

Acknowledgments

This research was supported by National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2025-16067916), Basic Science Research Program through the NRF funded by the Ministry of Education (RS-2024-00460505), and the Regional Innovation System & Education (RISE) program through the Gangwon RISE Center funded by the Ministry of Education and the Gangwon State, Republic of Korea (2025-RISE-10-006).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2025.109088.

Code and data availability

The source code proposed in this research is available at https://github.com/cjw0206/MEMOL.

References

[1] The Drug Development Process, U.S. Food and Drug Administration, 2018, https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process. (Accessed 21 February 2025).

[2] J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there? Nat. Rev. Drug Discov. 5 (12) (2006) 993–996, http://dx.doi.org/10.1038/nrd2199.

[3] J.F. Pritchard, M. Jurima-Romet, M.L. Reimer, E. Mortimer, B. Rolfe, M.N. Cayen, Making better drugs: decision gates in non-clinical drug development, Nat. Rev. Drug Discov. 2 (7) (2003) 542–553, http://dx.doi.org/10.1038/nrd1131.

[4] M. Alomar, A.M. Tawfiq, N. Hassan, S. Palaian, Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future, Ther. Adv. Drug Saf. 11 (2020) 2042098620938595, http://dx.doi.org/10.1177/2042098620938595.

[5] J.-H. Park, Y.-R. Cho, Computational drug repositioning with attention walking, Sci. Rep. 14 (1) (2024) 10072, http://dx.doi.org/10.1038/s41598-024-60756-6.

[6] K. Ogura, T. Sato, H. Yuki, T. Honma, Support vector machine model for hERG inhibitory activities based on the integrated hERG database using descriptor selection by NSGA-II, Sci. Rep. 9 (1) (2019) 12220, http://dx.doi.org/10.1038/s41598-019-47536-3.

[7] Y. Wang, Q. Xiao, P. Chen, B. Wang, In silico prediction of drug-induced liver injury based on ensemble classifier method, Int. J. Mol. Sci. 20 (17) (2019) 4106, http://dx.doi.org/10.3390/ijms20174106.

[8] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, DeepTox: toxicity prediction using deep learning, Front. Env. Sci. 3 (2016) 80, http://dx.doi.org/10.3389/fenvs.2015.00080.

[9] J.Y. Ryu, M.Y. Lee, J.H. Lee, B.H. Lee, K.-S. Oh, DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity, Bioinform. 36 (10) (2020) 3049–3055, http://dx.doi.org/10.1093/bioinformatics/btaa075.

[10] J. Haneczok, M. Delijewski, R. Moldzio, AI molecular property prediction for Parkinson's disease reveals potential repurposing drug candidates based on the increase of the expression of PINK1, Comput. Methods Programs Biomed. 241 (2023) 107731, http://dx.doi.org/10.1016/j.cmpb.2023.107731.

[11] M. Gustavsson, S. Käll, P. Svedberg, J.S. Inda-Diaz, S. Molander, J. Coria, T. Backhaus, E. Kristiansson, Transformers enable accurate prediction of acute and chronic chemical toxicity in aquatic organisms, Sci. Adv. 10 (10) (2024) eadk6669, http://dx.doi.org/10.1126/sciadv.adk6669.

[12] N. Aksamit, A. Tchagang, Y. Li, B. Ombuki-Berman, Hybrid fragment-SMILES tokenization for ADMET prediction in drug discovery, BMC Bioinform. 25 (1) (2024) 255, http://dx.doi.org/10.1186/s12859-024-05861-z.

[13] P. Liu, Y. Ren, J. Tao, Z. Ren, Git-mol: A multi-modal large language model for molecular science with graph, image, and text, Comput. Biol. Med. 171 (2024) 108073, http://dx.doi.org/10.1016/j.compbiomed.2024.108073.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. International Conference on Learning Representation, ICLR, 2021.

[15] J. Chen, Y.-W. Si, C.-W. Un, S.W. Siu, Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network, J. Cheminform. 13 (2021) 1–16, http://dx.doi.org/10.1186/s13321-021-00570-8.

[16] H. Stärk, D. Beaini, G. Corso, P. Tossou, C. Dallago, S. Günnemann, P. Liò, 3D infomax improves gnns for molecular property prediction, in: Proc. 39th International Conference on Machine Learning, ICML, 2022, pp. 20479–20502.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, et al., Multimodal deep learning, in: Proc. 28th International Conference on Machine Learning, ICML, 2011, pp. 689–696.

[18] Z.A. Rollins, A.C. Cheng, E. Metwally, MolPROP: Molecular property prediction with multimodal language and graph fusion, J. Cheminform. 16 (1) (2024) 56, http://dx.doi.org/10.1186/s13321-024-00846-9.

[19] X. Lu, L. Xie, L. Xu, R. Mao, X. Xu, S. Chang, Multimodal fused deep learning for drug property prediction: Integrating chemical language and molecular graph, Comput. Struct. Biotechnol. J. 23 (2024) 1666–1679, http://dx.doi.org/10.1016/j.csbj.2024.04.030.

[20] Y. Qian, X. Li, J. Wu, Q. Zhang, MCL-DTI: using drug multimodal information and bi-directional cross-attention learning method for predicting drug–target interaction, BMC Bioinform. 24 (1) (2023) 323, http://dx.doi.org/10.1186/s12859-023-05447-1.

[21] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, in: Proc. 5th International Conference on Learning Representations, ICLR, 2017.

[22] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, Z. Chen, Gshard: Scaling giant models with conditional computation and automatic sharding, in: Proc. 36th Conference on Neural Information Processing Systems, NeurIPS, 2021.

[23] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, N. Houlsby, Multimodal contrastive learning with limoe: the language-image mixture of experts, in: Proc. 36th Conference on Neural Information Processing Systems, NeurIPS, Vol. 35, 2022, pp. 9564–9576.

[24] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, N. Houlsby, Scaling vision with sparse mixture of experts, in: Proc. 35th Conference on Neural Information Processing Systems, NeurIPS, 2021.

[25] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al., Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025, http://dx.doi.org/10.48550/arXiv.2507.06261.

[26] X. Zhang, Y. Shen, Z. Huang, J. Zhou, W. Rong, Z. Xiong, Mixture of attention heads: Selecting attention heads per token, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2022, pp. 4150–4162.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. 31st Conference on Neural Information Processing Systems, NIPS, 2017.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: Proc. 6th International Conference on Learning Representations, ICLR, 2018.

[30] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? in: Proc. 7th International Conference on Learning Representations, ICLR, 2019.

[31] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proc. 5th International Conference on Learning Representations, ICLR, 2017.

[32] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754, http://dx.doi.org/10.1021/ci100050t.

[33] H.L. Morgan, The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service, J. Chem. Doc. 5 (2) (1965) 107–113, http://dx.doi.org/10.1021/c160017a018.

[34] RDKit: open-source cheminformatics software, 2024, https://www.rdkit.org/. (Accessed 21 February 2025).

[35] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36, http://dx.doi.org/10.1021/ci00057a005.

[36] An open-source hub for all your molecular featurizers, 2025, https://molfeat.datamol.io/, (Accessed 21 February 2025).

[37] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C.W. Coley, C. Xiao, J. Sun, M. Zitnik, Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development, NeurIPS Datasets Benchmarks Track, 2021, http://dx.doi.org/10.48550/arXiv.2102.09548.

[38] T. Li, W. Tong, R. Roberts, Z. Liu, S. Thakkar, DeepCarc: Deep learning-powered carcinogenicity prediction using model-level representation, Front. Artif. Intell. 4 (2021) 757780, http://dx.doi.org/10.3389/frai.2021.757780.

[39] S. Thakkar, T. Li, Z. Liu, L. Wu, R. Roberts, W. Tong, Drug-induced liver injury severity and toxicity (dilist): binary classification of 1279 drugs by human hepatotoxicity, Drug Discov. Today 25 (1) (2020) 201–208, http://dx.doi.org/10.1016/j.drudis.2019.09.022.

[40] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: large-scale self-supervised pretraining for molecular property prediction, 2020, http://dx.doi.org/10.48550/arXiv.2010.09885, arXiv preprint arXiv:2010.09885.

[41] Y. Wang, J. Wang, Z. Cao, A. Barati Farimani, Molecular contrastive learning of representations via graph neural networks, Nat. Mach. Intell. 4 (3) (2022) 279–287, http://dx.doi.org/10.1038/s42256-022-00447-x.

[42] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, J. Tang, Pre-training molecular graph representation with 3D geometry, in: Proc. 10th International Conference on Learning Representations, ICLR, 2022.

[43] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, P. Das, Large-scale chemical language representations capture molecular structure and properties, Nat. Mach. Intell. 4 (12) (2022) 1256–1264, http://dx.doi.org/10.1038/s42256-022-00580-7.

[44] T.T. Van Tran, H. Tayara, K.T. Chong, AMPred-CNN: Ames mutagenicity prediction model based on convolutional neural networks, Comput. Biol. Med. 176 (2024) 108560, http://dx.doi.org/10.1016/j.compbiomed.2024.108560.

[45] I. Arab, K. Laukens, W. Bittremieux, Semisupervised learning to boost hERG, Nav1.5, and Cav1.2 cardiac ion channel toxicity prediction by mining a large unlabeled small molecule data set, J. Chem. Inf. Model. 64 (16) (2024) 6410–6420, http://dx.doi.org/10.1021/acs.jcim.4c01102.

[46] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, http://dx.doi.org/10.48550/arXiv.1907.11692.

[47] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, J. Huang, A survey on mixture of experts in large language models, IEEE Trans. Knowl. Data Eng. (2025) http://dx.doi.org/10.1109/TKDE.2025.3554028.