Journal of Cheminformatics

RESEARCH Open Access

Enhancing multi-task in vivo toxicity prediction via integrated knowledge transfer of chemical knowledge and in vitro toxicity information

Minsu Park[1], Yewon Shin[2], Hyunho Kim[1,3*] and Hojung Nam[1,2,4*]

Abstract

The evaluation of potential drug toxicity is a crucial step in early drug development. in vivo toxicity assessment represents a key challenge that must be addressed before advancing to clinical trials. However, traditional in vivo experiments primarily rely on animal models, raising concerns regarding cost, time efficiency, and ethical considerations. To address these challenges, various computational approaches have been developed to support in vivo toxicity evaluations, though these methods often demonstrate limited generalizability due to data scarcity. In this study, we propose MT-Tox, a knowledge transfer-based multi-task learning model specifically designed for in vivo toxicity prediction that overcomes data scarcity. Our model implements a sequential knowledge transfer strategy across three stages: general chemical knowledge pretraining, in vitro toxicological auxiliary training, and in vivo toxicity fine-tuning. This hierarchical approach significantly improves model performance by systematically leveraging information from both chemical structure and toxicity data sources. MT-Tox outperforms baseline models across three in vivo toxicity endpoints: carcinogenicity, drug-induced liver injury (DILI), and genotoxicity. Through ablation studies and attention analyses, we demonstrate that each knowledge transfer technique makes meaningful contributions to the prediction process. Finally, we demonstrate the real-world application of our model as a prediction tool for early-stage drug discovery through comprehensive DrugBank database screening.

Scientific contribution: We propose a knowledge transfer framework that integrates chemical and in vitro toxicological information to enhance in vivo toxicity prediction in low-data regimes. Our model provides dual-level interpretability across chemical and biological domains through attention mechanism. Moreover, we demonstrate our model's applicability by screening the DrugBank database, simulating practical toxicity screening scenarios in drug development.

Keywords Graph neural networks, Transfer learning, Multi-task learning, In vivo toxicity

*Correspondence:
Hyunho Kim
hyunho.kim@kitox.re.kr
Hojung Nam
hjnam@gist.ac.kr
Full list of author information is available at the end of the article

Introduction

The failure of drug development is often caused by toxicity, making early-stage screening for potential toxicity a critical step [1, 2]. Conventionally, drug toxicity is assessed through in vitro studies, which are subsequently followed by in vivo experiments on animal models before clinical trials. However, these methods have been noted for their limitations in terms of cost, time, and ethical concerns, particularly regarding in vivo assessments [3]. As the number of drug candidates continues to grow, there is an increasing need for a more efficient approach to reduce failures in in vivo toxicity assessments.

Therefore, various computational approaches have been studied using machine learning (ML) and deep learning (DL) methods to predict in vivo toxicity [4–9]. Although the growing interest in computational approaches has led to the accumulation of toxicity data, well-organized in vivo toxicity datasets are still limited for effectively training ML and DL models. Thus, to ensure robust performance, it is important to design ML and DL models considering the data scarcity.

In such a low-data regime, leveraging auxiliary information through transfer learning (TL) or multi-task learning (MTL) can be useful to improve model performance [10, 11]. In particular, TL has been widely explored as a strategy to enhance the performance of downstream tasks by utilizing molecular structure-based pre-training [12–15]. However, several studies have suggested that such structure-based pre-trained models may suffer from an objective gap with downstream tasks [16, 17]. In contrast, MTL aims to improve the performance of several tasks by training them jointly in a unified framework [18]. Leveraging this property, recent toxicity prediction studies have attempted to incorporate toxicity data from different levels, such as in vitro and in vivo toxicity [19, 20] or toxicity across multiple conditions [21–23] to enhance the performance on tasks with limited data. However, since MTL relies on the simultaneous training of all tasks to enable knowledge sharing, it lacks mechanisms for explicitly modeling task-specific knowledge transfer, especially when training is imbalanced across tasks. When dealing with MTL of heterogeneous datasets that have different biological levels, such as in vitro and in vivo toxicity, this imbalance may amplify the instability of training, result in performance degradation on tasks.

To overcome these limitations, we proposed MT-Tox, a model that combines TL and MTL to enable in vivo toxicity-specific knowledge transfer. To effectively guide in vivo toxicity prediction, MT-Tox utilizes both chemical and in vitro toxicity knowledge data with specifically designed strategies that align with the characteristics of each information (Fig. 1). The training procedure of MT-tox consists of three stages: (1) general chemical knowledge pre-training, (2) in vitro toxicological auxiliary training, and (3) in vivo toxicity fine-tuning (Fig. 1A). In the general chemical knowledge pre-training stage, the model learns functional group representations from ChEMBL [24], one of the large-sized compound databases to provide a pre-trained graph encoder. In the in vitro toxicological auxiliary training stage, the model undergoes multi-task learning on 12 toxicity assays from the Tox21 dataset [25], allowing the model to acquire contextual information related to in vitro toxicity. Finally, during the in vivo toxicity fine-tuning stage, the model incorporates pre-trained in vitro toxicity context using a cross-attention mechanism to refine predictions for each in vivo toxicity endpoint in a multi-task learning (Fig. 1B). This design is inspired by the concept of in vitro to in vivo extrapolation (IVIVE), which models the biological relationships between in vitro and in vivo toxicity to estimate the systematic process of in vivo toxicity [26]. As suggested by IVIVE approaches, our model selectively transfers useful in vitro toxicity information to specific in vivo toxicity via the cross-attention mechanism, which is expected to improve the predictive performance of in vivo toxicity.

Materials and methods

Datasets

The datasets used in this study consist of two pre-training datasets: ChEMBL and Tox21, and a fine-tuning set with three types of in vivo toxicity endpoints. First, the ChEMBL dataset, a large-scale collection of bioactive compound information, was utilized as a pre-training set to train general molecular structural knowledge. We used a version of the ChEMBL dataset processed by GuacaMol [27], which is widely used as a pre-training benchmark. The Simplified Molecular Input Line Entry System (SMILES) strings in the dataset were standardized using the 'StandardizeSmiles' function from RDKit [28]. Also, we also excluded inorganic compounds which do not have carbon atoms and those with molecular weights greater than 1,000 Da to focus on small-molecule drug prediction. After the pre-processing steps, the final processed dataset contained a total of 1,578,896 compounds. Second, the Tox21 dataset contains diverse in vitro bioassay data related to toxicity. We selected this dataset to provide supplementary information about the in vitro toxicity context of input query molecules for improving in vivo toxicity prediction performance. To this end, we collected bioactivity data from 12 assays corresponding to the Tox21 (2014) challenge from the National Center for Advancing Translational Sciences (NCATS) [25, 29]. The bioactivity for each compound was processed based on the 'ASSAY_OUTCOME' field in the aggregated

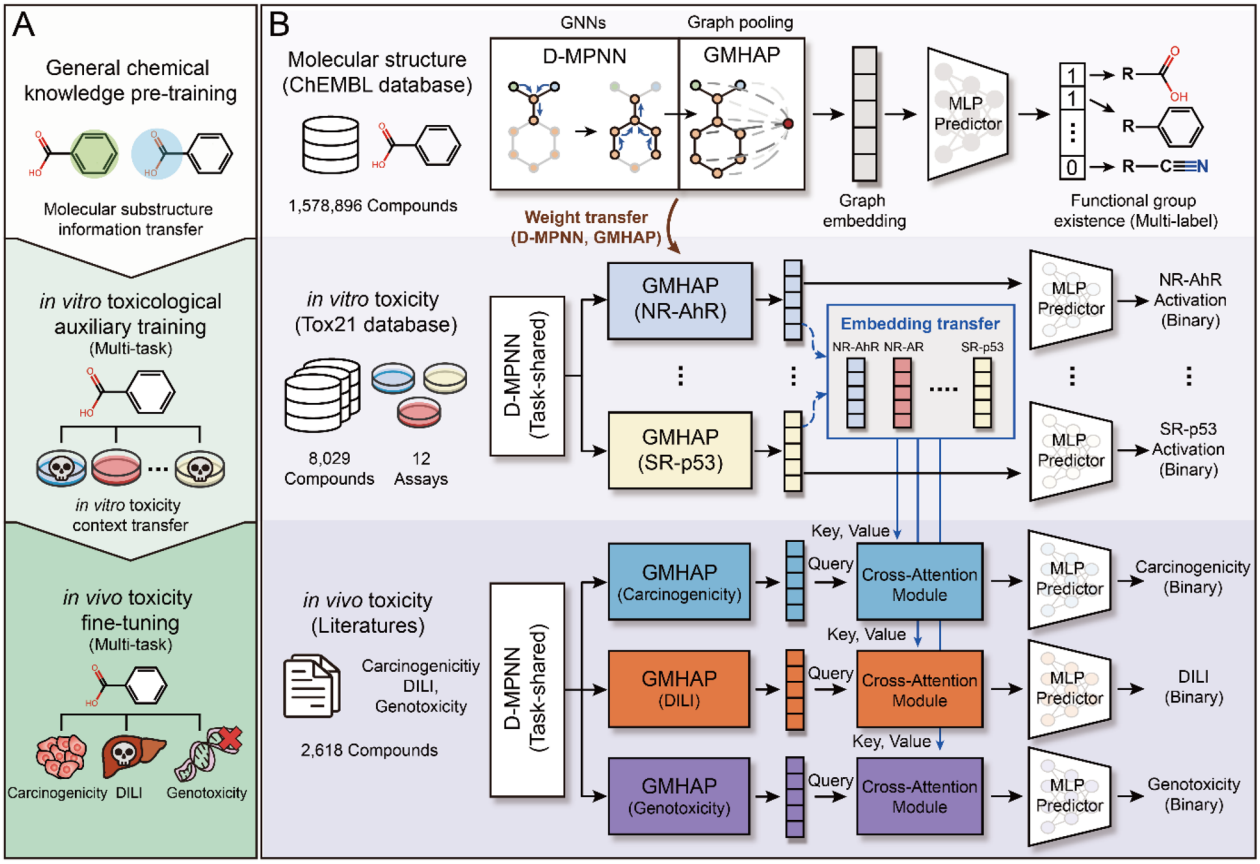

Fig. 1 The research overview of the MT-Tox framework. (**A**) The general workflow of the MT-Tox framework. MT-Tox consists of 3 types of learning stages: General chemical knowledge pre-training, in vitro toxicological auxiliary training, and in vivo toxicity fine-tuning. (**B**) A detailed explanation for each stage is provided

readout profile of each assay in the dataset. Compounds having an 'Active' label were assigned as positive, other compounds with 'Inactive' were labeled negative. Cases with 'Inconclusive' were removed. Subsequently, the same standardization and duplicate removal procedures were applied to the SMILES strings in this dataset. The final processed dataset consists of 8,029 compounds with 12 assay endpoints. The detailed statistics for each assay are provided in Table 1.

Finally, to construct the fine-tuning dataset, we collected three in vivo toxicity endpoints from diverse sources [20, 30–35]. It consists of structurally diverse molecules with carcinogenicity, drug-induced liver injury (DILI), and genotoxicity outcomes. We applied additional rigorous pre-processing to SMILES and activity to ensure consistency across the diverse in vivo data sources. The additional pre-processing procedure was developed based on the approaches described in previous studies [36, 37]. Specifically, the SMILES strings were standardized using RDKit through normalization, reionization, principal fragment extraction, uncharging and charge

Table 1 Data statistics for the Tox21 dataset

Assay ID	Number of positives	Number of negatives
NR-AhR	820	6103
NR-AR	280	6853
NR-AR-LBD	331	7310
SR-ARE	1036	4898
NR-aromatase	770	5782
NR-ATAD5	428	7138
NR-ER	453	6933
NR-ER-LBD	1413	5630
SR-HSE	442	6134
SR-MMP	1232	5170
SR-p53	489	6645
NR-PPAR	341	6569

normalization, and removal of stereochemical configurations. For principal fragment extraction, we adopted the parent compound extraction strategy from PubChem [38], defining the principal fragment as the substructure that accounts for more than 70% of the total atoms and contains at least one carbon atom. Afterward, duplicate compounds were removed using the first 14 characters of the InChIKey. Additional duplicate molecules were removed through tautomer standardization. For the duplicated compounds, only those demonstrating consistent labels in more than two-thirds of instances were retained, while inconsistent entries were excluded. After SMILES standardization, we filtered out molecules having less than four atoms to remove simple molecules such as CO_2 or water. The final constructed dataset contains 2,618 compounds. Additionally, to further evaluate the generalization performance of the model, we built an external test set from both independent sources and distinct sources within the fine-tuning set [31, 33, 39–41]. Compounds with SMILES strings overlapping with the fine-tuning set were removed from the external test sets to ensure a fair comparison. Table 2 summarizes the statistics of the final fine-tuning and external test sets.

Model architecture of MT-Tox

Our model consists of three sequential stages: general chemical knowledge pre-training, in vitro toxicological auxiliary training, and in vivo toxicity fine-tuning. Across all training steps, we used the identical graph neural network (GNN)-based method with slight variations according to the objectives of each stage.

Backbone model architecture

For the backbone architecture, we adopted the Bayes-hERG architecture [42] to represent the graph structure of the compounds. This model consists of three main components: directed message-passing neural network (D-MPNN), global multi-head attentive pooling (GMHAP), and a multi-layered perceptron (MLP) predictor. D-MPNN [43] is a type of GNN that updates node representations by passing messages along directed edges. Based on these updated node features, the GMHAP module aggregates node information into a single graph embedding using a self-attention mechanism. Then, the final graph embedding from GMHAP is passed through an MLP predictor, which is composed of MLP with rectified linear unit (ReLU) activation functions, to calculate the prediction score. Further details on the backbone architecture are described in Sect. 1 of Additional file 1.

Table 2 Data statistics for the in vivo toxicity fine-tuning dataset

Data type	Toxicity type	Number of positives	Number of negatives
Fine-tuning set	Carcinogenicity	511	517
	DILI	781	597
	Genotoxicity	286	675
External test set	Carcinogenicity	114	113
	DILI	15	17
	Genotoxicity	11	18

General chemical knowledge pre-training model architecture

The general chemical knowledge pre-training step aims to learn the general structural features of compounds, enabling the model to achieve stable learning in subsequent stages. For this step, we employed the graph-level self-supervised task proposed in the study of Rong et al. [14]. The model architecture follows the same design as the backbone, consisting of D-MPNN, GMHAP, and an MLP predictor. Thus, the general chemical knowledge pre-training model gets the molecular graphs as input features and calculates the graph embeddings through D-MPNN and GMHAP. This graph embedding passed through an MLP predictor, which is composed of three-layered MLP with ReLU activation functions and a dropout rate of 0.1. After calculating the logit values, the predictor calculates the 85 functional group probabilities with a sigmoid activation function.

In vitro toxicological auxiliary training model architecture

The in vitro toxicological auxiliary training step aims to learn bioactivity information from Tox21 assays that can potentially enhance in vivo toxicity prediction. As demonstrated in prior studies [44, 45], sharing information across the Tox21 endpoints through multi-task learning could lead to more generalizable models compared to training separate single-task models. Thus, we adapted the backbone architecture to accommodate multi-task learning. The D-MPNN module is shared across tasks, while the GMHAP and MLP predictor modules are designed to be task-specific. We made the GMHAP layer task-specific since different in vitro assays may be sensitive to distinct molecular substructures. This multi-task design allows the model to capture the task-specific compound variations. Finally, the task-specific graph embeddings from the GMHAP layer pass through each task-specific MLP predictor to predict the in vitro toxicity. The predictor consists of two-layered MLP with ReLU activation functions and a dropout rate of 0.15. Since this model is used to transfer the graph embeddings to the subsequent in vivo toxicity prediction model, we employed a two-layered MLP predictor to maximize the information of the graph embeddings. Using the

calculated logit values, the predictor predicts the toxicity probabilities of 12 in vitro toxicity endpoints with the softmax activation function.

In vivo toxicity fine-tuning model architecture

In the fine-tuning stage, the model leverages information learned from the previous two stages to perform multi-task in vivo toxicity prediction. As shown in Table 2, the number of samples for each fine-tuning task is highly limited. Therefore, to improve performance through task-level knowledge sharing, we also adopted multi-task learning in the fine-tuning step. The model architecture at this stage extends the backbone architecture by incorporating a cross-attention module. This additional cross-attention layer selectively integrates information from the Tox21 auxiliary trained model into the graph embeddings of the fine-tuning model. The detailed formulation is as follows:

$$h_{vivo}^{t} = LayerNorm\left(g_{vivo}^{t}\right) \in R^{1\times d},$$

$$H_{vitro} = LayerNorm(G_{vitro}) \in R^{12\times d},$$

$$Q^{t} = h_{vivo}^{t} W_{Q}^{t}, K^{t} = H_{vitro} W_{K}^{t}, V^{t} = G_{vitro} W_{V}^{t},$$

$$Att^{t} = softmax\left(\frac{Q^{t}(K^{t})^{T}}{\sqrt{d}}\right),$$

$$h_{aggregate}^{t} = Att^{t} \cdot V^{t} \in R^{1\times d},$$

$$h_{final}^{t} = ReLU\left(FFN\left(g_{vivo}^{t} + h_{aggregate}^{t}\right)\right) \in R^{1\times d}$$

Let g_{vivo}^{t} and G_{vitro} denote the graph embeddings for an in vivo task t in the fine-tuning model and the set of graph embeddings obtained from the in vitro toxicological auxiliary model, respectively. The graph embedding is defined as the output generated after the input graph passes through the D-MPNN and GMHAP modules. Then, the cross-attention mechanism computes attention by treating h_{vivo}^{t} as query, and the H_{vitro} and G_{vitro} as key and value, respectively. To reduce the potential attention bias caused by numerical scale differences between embeddings, we applied layer normalization to each embedding before entering the attention layer. The $h_{aggregate}^{t}$ is then calculated through the cross-attention mechanism, this vector is finally added to the g_{vivo}^{t} to produce the final representation. Thus, the resulting embedding h_{final}^{t} selectively integrates the in vivo toxicity information and relevant Tox21 assay embedding information, allowing the model to capture the specific context for each in vivo toxicity task. Finally, h_{final}^{t} is passed through the MLP predictor which consists of three-layered MLP with a dropout rate of 0.1 and ReLU activation. After calculating logit values, the final toxicity probabilities of three in vivo toxicity endpoints are calculated with the softmax function.

Training scheme

In the general chemical knowledge pre-training step, the ChEMBL dataset is split into training and validation sets at a 9:1 ratio with a random split. For model training, the model is trained using MultiLabelSoftMarginLoss in PyTorch [46] as the loss function. The D-MPNN and GMHAP modules trained in this step are then utilized as pre-trained modules in later learning stages. After the pre-training step, we performed in vitro toxicological auxiliary training. The dataset was split into training and validation sets at a 9:1 ratio with scaffold split. For training, we employed a weighted binary cross-entropy loss, where the weights were determined based on the ratio of positive and negative samples for each task. Finally, in the in vivo toxicity fine-tuning step, we split datasets using two schemes; fivefold nested cross-validation and a scaffold split with an 8:1:1 ratio for training, validation and test set for each task. For the loss function, we employed a weighted binary cross-entropy loss based on the ratio of positive and negative samples in each task. Furthermore, we applied an exponential moving average [47, 48] of parameters to stabilize the training process. For the model parameters, the D-MPNN and GMHAP modules were initialized with the pre-trained weights from the general chemical knowledge pre-training step. Also, when incorporating the Tox21 auxiliary embeddings via cross-attention, the parameters of the Tox21 auxiliary model were frozen, so that only the parameters of the fine-tuning model are updated during training. For the details about the model implementation, the graph input feature and hyperparameters are described in **Table S1-3** in the Additional file 1.

Evaluation metrics

Since the general chemical knowledge pre-training task is multi-label classification, we used macro-f1, micro-f1, and coverage error as the evaluation metrics. The macro-f1 metric is calculated by averaging the f1 score of each class and the micro-f1 metric is calculated by the entire prediction results. The coverage error is a metric that indicates how many ranked scores are needed to cover all true labels in multi-label classification. Therefore, as the coverage error is closer to the average number of labels, it implies that the model predicts accurately across all classes.

We used the conventional binary classification metrics for in vitro toxicological auxiliary training and in vivo toxicity fine-tuning tasks for the performance evaluation: balanced accuracy (BAC), area under the receiver operating characteristic curves (AUC), and area under the precision-recall curve (AUPR). Additionally, we used the sum of ranks for each evaluation metric to assess the overall performance across all tasks. The detailed formulas of evaluation metrics are explained in Sect. 3 of Additional file 1.

Results

Performance evaluation of general chemical knowledge and toxicological auxiliary training

The task used for general chemical knowledge pre-training is a multi-label classification task designed to predict whether the input chemical contains 85 functional groups. On average, there are 195,975 positive instances per functional group across the ChEMBL dataset, and each compound contains an average of 10.527 functional groups. The performance of the model on the validation set achieved a macro-f1 score of 0.963, a micro-f1 score of 0.999, and a coverage error of 10.534. This indicates that the pre-trained model has effectively learned molecular structure information. The in vitro toxicological auxiliary training model is a multi-task prediction model for 12 Tox21 endpoints, and its performance was compared with individual random forest (RF) models trained on extended connectivity fingerprints (ECFP) and ECFP with molecular descriptors calculated from RDKit for each assay. The average AUPR of the auxiliary training model used in this study is 0.338, while the RF model with ECFP achieved 0.306 and the RF model with ECFP combined with 195 RDKit-derived molecular descriptors achieved 0.320 (Figure S1 in the Additional file 1). This demonstrates that the auxiliary training model has successfully learned the information contained in the Tox21 dataset. Detailed performance comparisons for Tox21 data are provided in Sect. 4 of Additional file 1.

Performance comparison with diverse knowledge transfer models

We conducted a prediction performance comparison with baseline models based on various knowledge transfer methods, including TL and MTL. For TL baselines, the models selected for comparison include two representative large-scale pre-trained models: ChemBERTa-2 [49] and GraphMVP [13]. These models utilize diverse molecular representations and incorporate pre-training strategies based on structural information. We also compared our model to two representative MTL models, specifically MTDNN-FP [19] and MGA [20]. MTDNN-FP is an ECFP-based deep learning model designed to improve clinical toxicity prediction by MTL of in vitro, in vivo, and clinical toxicity data. MGA is a multi-task graph attention-based model that simultaneously predicts 33 various types of toxicity endpoints. Both models aim to enhance toxicity prediction performance across several biological levels in a unified framework through MTL. The key distinction of our model is its differential treatment of in vitro and in vivo toxicity data, while baseline models use both data types simultaneously during training. For fair comparison, we retrained all baseline models with the same datasets used in our study. Specifically, in the MTL baseline models, we selected two training configurations using only multi-task learning of in vivo data (in vivo), and using both in vivo and Tox21 data (in vitro + in vivo).

First, we compared the model performance across the entire dataset using fivefold nested cross-validation (Table S4 in Additional file 2). Among the baseline models, the three multi-task graph attention-based models including MGA (in vivo), MGA (in vitro + in vivo), MT-Tox consistently achieved high predictive performance across all tasks. This demonstrates that our model architecture which is based on graph-based multi-task learning is effective for the in vivo toxicity prediction. Second, we compared model performance using a scaffold split to examine its generalization ability across molecular structure. As a result, MT-Tox achieved the best performance in five out of nine metrics and ranked highest in the sum of ranks (Table 3 and Table S5 in Additional file 2). Among the baseline models, the TL based methods (GraphMVP, ChemBERTa-2) showed good performance on genotoxicity. This highlights the importance of large-scale chemical knowledge-based pre-training in the low- data regime. However, while the ChemBERTa-2 model achieved the second-best performance on the genotoxicity prediction task, it showed weak performance on other tasks. This observation is consistent with previous reports that structure-based pre-trained models often fail to generalize across diverse downstream tasks [16, 17]. For the multi-task baseline models, it is noteworthy that models trained with in vivo data only generally achieved better performance than those trained with both in vivo and Tox21 data. This suggests that naively applying multi-task learning to datasets having different biological levels may lead to negative transfer in specific tasks. In contrast, our model achieved robust performance across all tasks, suggesting that the sequential knowledge transfer strategy proposed in this study can offer more effective learning than simple multi-task learning of heterogeneous toxicity datasets.

We also compared prediction performance on the external test set in the same comparison setting (Table 4 and **Table S6** in Additional file 2). MT-Tox achieved

Table 3 Performance comparison of MT-Tox with baseline models on the scaffold split test set

Model	Carcinogenicity			DILI			Genotoxicity		
	BAC	AUPR	AUC	BAC	AUPR	AUC	BAC	AUPR	AUC
TL1[a]	**0.608 ± 0.017**[ns]	**0.677 ± 0.027**[ns]	**0.674 ± 0.022**[**]	0.543 ± 0.029	0.643 ± 0.031	0.589 ± 0.034	0.594 ± 0.017	0.514 ± 0.030	0.645 ± 0.028
TL2[b]	0.557 ± 0.021	0.625 ± 0.018	0.548 ± 0.020	0.508 ± 0.038	0.589 ± 0.042	0.514 ± 0.049	0.605 ± 0.021	0.544 ± 0.022	0.687 ± 0.014
MTL1[c]	0.561 ± 0.053	0.630 ± 0.040	0.614 ± 0.038	0.501 ± 0.023	0.649 ± 0.033	**0.613 ± 0.021**	0.545 ± 0.066	0.447 ± 0.065	0.670 ± 0.059
MTL2[d]	0.552 ± 0.051	0.642 ± 0.038	0.627 ± 0.037	0.540 ± 0.049	0.658 ± 0.036	0.610 ± 0.032	0.514 ± 0.049	0.404 ± 0.074	0.627 ± 0.076
MTL3[e]	0.566 ± 0.024	0.633 ± 0.037	0.627 ± 0.030	**0.577 ± 0.032**[ns]	0.648 ± 0.024	0.609 ± 0.026	0.631 ± 0.038	0.486 ± 0.033	0.686 ± 0.026
MTL4[f]	0.589 ± 0.029	0.632 ± 0.026	0.637 ± 0.026	0.576 ± 0.030	0.632 ± 0.027	0.600 ± 0.026	0.590 ± 0.041	0.469 ± 0.037	0.655 ± 0.028
MT-Tox	0.603 ± 0.052	0.658 ± 0.039	0.630 ± 0.035	0.571 ± 0.018	**0.677 ± 0.027**[ns]	**0.613 ± 0.017**[ns]	**0.666 ± 0.024**[*]	**0.577 ± 0.037**[**]	**0.707 ± 0.017**[**]

[a] : TL baseline model 1 (GraphMVP)

[b] : TL baseline model 2 (ChemBERTa-2)

[c] : MTL baseline model 1 (MTDNN-FP (in vivo))

[d] : MTL baseline model 2 (MTDNN-FP (in vitro + in vivo))

[e] : MTL baseline model 3 (MGA (in vivo))

[f] : MTL baseline model 4 (MGA (in vitro + in vivo))

[ns] : $p \geq 0.05$ (one-sided Mann–Whitney U test between the 1st and 2nd best models)

[*] : $0.01 \leq p < 0.05$ (one-sided Mann–Whitney U test between the 1st and 2nd best models)

**: $0.001 \leq p < 0.01$ (one-sided Mann–Whitney U test between the 1st and 2nd best models)

Table 4 Performance comparison of MT-Tox with baseline models on the external test set

Model	Carcinogenicity			DILI			Genotoxicity		
	BAC	AUPR	AUC	BAC	AUPR	AUC	BAC	AUPR	AUC
TL1[a]	0.638 ± 0.018	0.739 ± 0.011	0.712 ± 0.025	0.522 ± 0.042	0.505 ± 0.028	0.541 ± 0.033	0.593 ± 0.040	0.553 ± 0.032	0.655 ± 0.049
TL2[b]	0.611 ± 0.011	0.734 ± 0.007	0.681 ± 0.013	0.532 ± 0.058	0.556 ± 0.085	0.537 ± 0.100	**0.703 ± 0.019**[**]	**0.589 ± 0.018**[*]	**0.746 ± 0.017**[***]
MTL1[c]	0.526 ± 0.012	0.582 ± 0.018	0.573 ± 0.018	0.494 ± 0.011	0.587 ± 0.023	0.524 ± 0.029	0.499 ± 0.012	0.337 ± 0.026	0.502 ± 0.027
MTL2[d]	0.511 ± 0.014	0.579 ± 0.026	0.554 ± 0.027	0.497 ± 0.023	0.587 ± 0.015	0.529 ± 0.020	0.492 ± 0.008	0.331 ± 0.019	0.496 ± 0.028
MTL3[e]	0.660 ± 0.015	0.729 ± 0.016	0.739 ± 0.013	0.596 ± 0.038	0.651 ± 0.048	0.676 ± 0.055	0.617 ± 0.031	0.547 ± 0.025	0.637 ± 0.032
MTL4[f]	0.647 ± 0.018	0.727 ± 0.014	0.733 ± 0.016	0.628 ± 0.037	**0.675 ± 0.035**[ns]	**0.709 ± 0.032**[ns]	0.640 ± 0.038	0.511 ± 0.040	0.635 ± 0.038
MT-Tox	**0.662 ± 0.010**[ns]	**0.778 ± 0.017**[**]	**0.772 ± 0.014**[**]	**0.648 ± 0.040**[ns]	0.662 ± 0.016	0.699 ± 0.024	0.654 ± 0.040	0.554 ± 0.024	0.669 ± 0.019

[a] : TL baseline model 1 (GraphMVP)

[b] : TL baseline model 2 (ChemBERTA-2)

[c] : MTL baseline model 1 (MTDNN-FP (in vivo))

[d] : MTL baseline model 2 (MTDNN-FP (in vitro + in vivo))

[e] : MTL baseline model 3 (MGA (in vivo))

[f] : MTL baseline model 4 (MGA (in vitro + in vivo))

[ns] : $0.05 \leq p$ (one-sided Mann–Whitney U test between the 1st and 2nd best models)

[*] : $0.01 \leq p < 0.05$ (one-sided Mann–Whitney U test between the 1st and 2nd best models)

**: $0.001 \leq p < 0.01$ (one-sided Mann–Whitney U test between the 1st and 2nd best models)

***: $p < 0.001$ (one-sided Mann–Whitney U test between the 1st and 2nd best models)

more balanced performance than baseline models as it achieved the lowest sum of ranks in all metrics. Although the ChemBERTa-2 model achieved the best performance on the genotoxicity task, it showed relatively lower performance on other tasks, resulting in the third-best ranking in the sum of ranks. To further investigate this discrepancy, we compared the confusion matrices of MT-Tox and ChemBERTa-2 (**Table S7** in Additional file 1). For the genotoxicity, the performance of MT-Tox was comparable when considering the data quantity and potential bias of the external dataset. Furthermore, considering the small size of the external test set, we conducted additional comparison with random and data-driven baselines to demonstrate that the performance of our model was not obtained by chance (See **Sect. 6** in Additional file 1). As a result, MT-Tox model achieved superior performance across all evaluation metrics compared to the baseline models (**Table S8** in the Additional file 2). Specifically, the MT-Tox model achieved both high average performance and low variance, it demonstrates

stable and robust performance even under limited data conditions.

To ensure the generalizability of our model, we further conducted a performance comparison with previously published large-scale ADMET multi-task models (**Sect. 7** in the Additional file 1). Despite using a relatively smaller number of endpoints in multi-task learning, MT-Tox achieved comparable performance on the external test set (**Table S9** in the Additional file 2). Furthermore, because DILI has been extensively studied in previous research, we compared our model against recently published study [37], DILIPredictor, which incorporates both DILI-related in vitro assay data and pharmacokinetic information from an IVIVE perspective (See **Table S10-11** and **Sect. 8** in Additional file 1). Despite our model is using only chemical structure and general in vitro toxicity assay data, it achieved substantially better performance. This emphasizes the efficiency of our knowledge transfer approach. In conclusion, our MT-Tox model is robust and generalizable across in vivo toxicity endpoints compared to other knowledge transfer-based methods.

Quantitative assessment of individual knowledge transfer components in MT-Tox

Our MT-Tox incorporates two types of knowledge transfer to enhance in vivo toxicity prediction. To quantify their impacts, we analyzed the individual contributions of each knowledge transfer component (Fig. 2). First, we conducted an ablation study to examine how each component affects overall model performance (Fig. 2A). For this analysis, we evaluated four model variants: (1) removal of general chemical knowledge pre-training (W/o ChEMBL P.T), (2) removal of in vitro toxicological auxiliary training (W/o Tox21 A.T), (3) removal of both components with multi-task learning (W/o Both (MTL)), and (4) removal of both components with single-task learning for each endpoint (W/o Both (STL)). The results show that the full MT-Tox model consistently performs better than other variants across tasks, indicating that both types of knowledge transfer independently contribute to enhancing model performance. Specifically, the W/o ChEMBL P.T model exhibited a significant drop in performance for genotoxicity and increased variance across tasks. This suggests that general chemical knowledge pre-training provides a better initialization for the molecular graph encoder. It enhances the stability of model learning even in tasks with limited data. However, in the case of DILI, we observed performance increases when removing chemical knowledge pre-training (W/o ChEMBL P.T, W/o Both, W/o Both (STL)). This counterintuitive finding can be attributed to negative transfer, where the general chemical knowledge learned from ChEMBL introduces structural biases that are detrimental to DILI prediction. Unlike other toxicity endpoints that rely on direct structure–activity relationships, DILI primarily results from complex metabolic and immune-mediated processes that are poorly correlated with general chemical structural features [50, 51]. The pre-trained chemical representations, optimized for bioactivity prediction tasks, appear to encode irrelevant or misleading patterns for DILI, thereby interfering with the model's ability to learn DILI-specific features from the limited training data. This suggests that for certain toxicity endpoints with weak structure–activity relationships, domain-specific training without broad chemical pre-training may be more beneficial. For the case of W/o Tox21 A.T model which is a multi-task learning model using in vivo toxicity with chemical knowledge transfer, it showed a performance decrease across all tasks compared to MT-Tox, meaning that the auxiliary information from Tox21 contributes positively to the prediction performance for each task. Finally, the W/o Both (MTL) model, which indicates naive multi-task learning without both knowledge transfer components, performed consistently better than the W/o Both (STL). This finding provides support for the validity of our model using multi-task learning in the context of in vivo toxicity prediction.

We conducted an additional analysis to further investigate how knowledge transfer influences the model prediction. Since the in vitro toxicity knowledge transfer occurs at the graph embedding level, we analyzed how the embedding space varies across tasks depending on the presence of auxiliary information (Fig. 2B). As a result, the W/o Tox21 A.T model exhibits entangled distribution of embeddings across the three tasks, while the MT-Tox model demonstrates clearly separated embedding spaces for each task. This indicates that the information from Tox21 is effectively incorporated into the graph representation, specifically for the downstream tasks. Therefore, the injection of Tox21 information plays a crucial role in distinct toxicity representations, thereby reducing task confusion and improving overall performance in multi-task learning.

Dual-level interpretability through attention mechanisms revealing toxicological relevance

Our MT-Tox incorporates two types of attention modules, namely GMHAP and cross-attention modules. In this section, we demonstrated how each attention module can be applied to support the interpretation of the model prediction (Fig. 3).

First, the GMHAP module provides interpretability into the model prediction by revealing which atoms receive higher attention during the graph embedding process. Given that the MT-Tox model incorporates

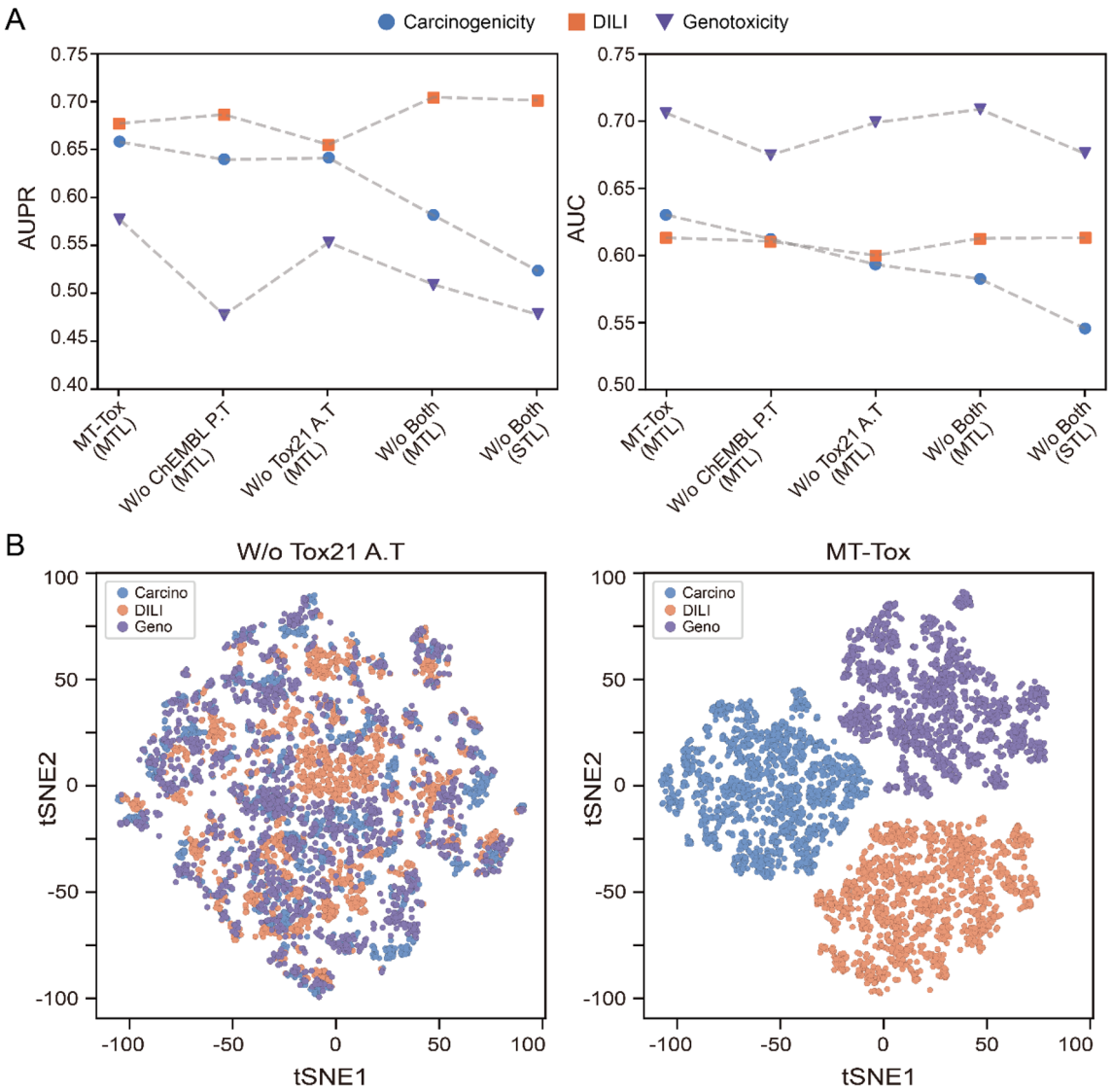

Fig. 2 Analysis of contributions of knowledge components. (A) Ablation studies of MT-Tox by removing general chemical knowledge and in vitro toxicological information. (B) 2D t-SNE plots of graph embeddings on MT-Tox and W/o Tox21 A.T models. The t-SNE components were calculated by using the whole compounds in the fine-tuning set

a pre-training task focused on identifying functional groups in the molecule, we hypothesized that the model would show higher attention to the substructural information during inference. To validate this hypothesis, we analyzed the attention value distributions for true positive compounds in the carcinogenicity prediction task (Fig. 3A). The results revealed that the MT-Tox model achieved higher attention variance in most attention heads compared to the W/o ChEMBL P.T model. This implies that the MT-Tox model is more focused on specific atoms within the molecule. This behavior can be interpreted as a result of the general chemical knowledge pre-training task, which involves predicting functional groups in the molecules. Furthermore, to examine whether the focused atom attention regions correspond to functional groups, we visualized the attention values on actual molecular structures in the carcinogenicity task (Fig. 3B). As a result, the W/o ChEMBL P.T model exhibits uniformly distributed attention across the molecule, whereas the MT-Tox model shows a focused attention pattern on specific functional groups. For example, head 7 has strong attention to nitrogen atoms such as nitroso and azobenzene substructures and head 8 focuses on benzene rings. These functional groups have been previously identified as structural alerts for carcinogenicity [52]. Consequently, these results imply that general

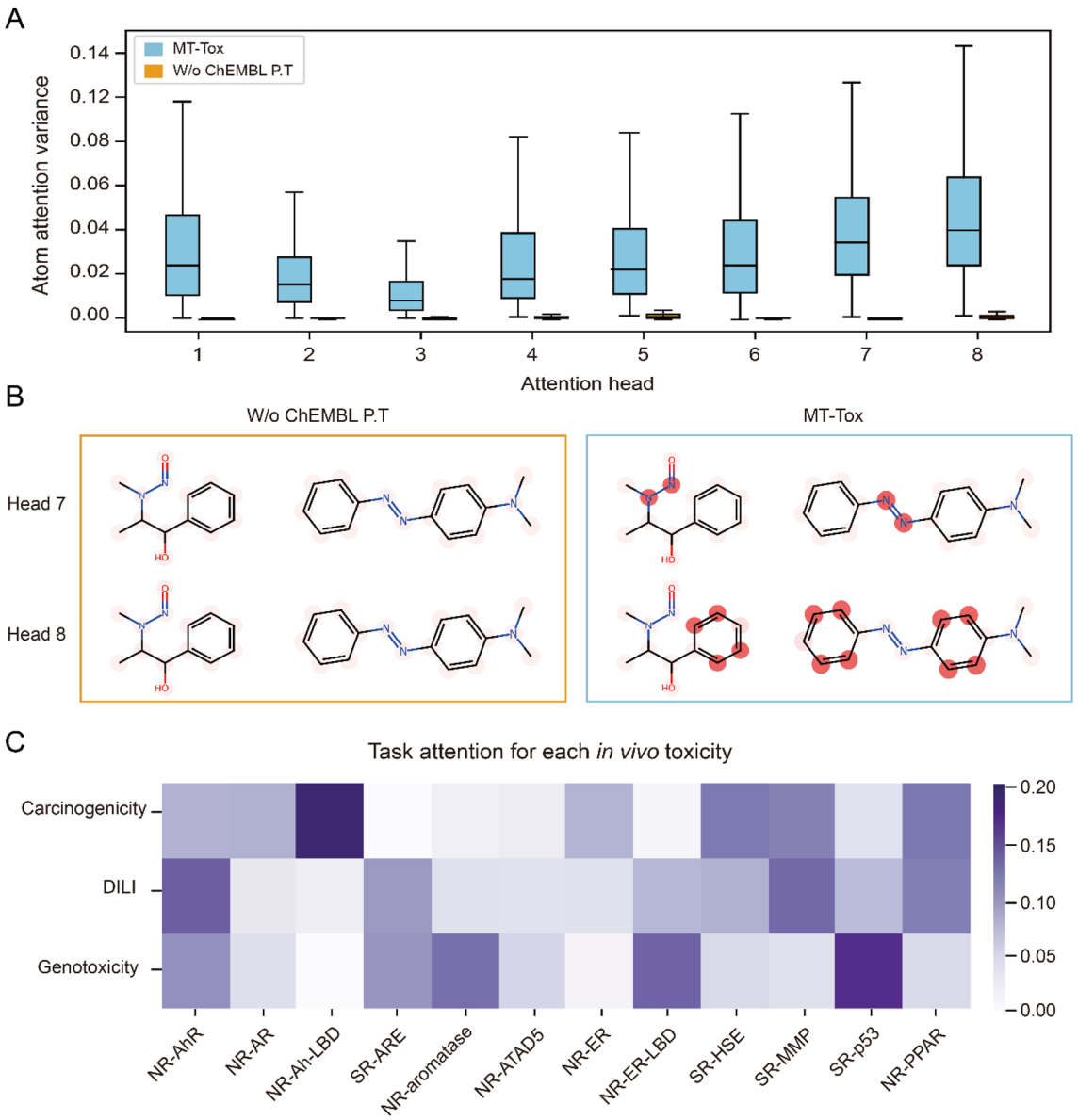

Fig. 3 Visualization of model interpretability analysis results. (A) Box plot illustrates the distribution of atom-level attention variances across attention heads in the GMHAP module for MT-Tox (skyblue) and W/o ChEMBL P.T. (orange) (B) Attention visualization examples of true positive molecules for carcinogenicity task in attention heads 7 and 8. The intensity of the red color indicates the magnitude of the attention value. The darker red color represents the higher attention value of the corresponding atom (C) The heatmap shows the average attention scores of Tox21 endpoints from the cross-attention modules with the true positive compounds for each in vivo toxicity

chemical knowledge pre-training enables the GMHAP module to respond sensitively to functional group level features across attention heads. This enhances the model's interpretability about the molecular structure in the context of in vivo toxicity prediction.

The second aspect of interpretability comes from the cross-attention module, enabling identification of which Tox21 features are primarily utilized in each in vivo toxicity prediction. We visualized a heatmap with the average attention values computed from true positive samples for each task (Fig. 3C). The results show that each in vivo task selectively attends to different Tox21 assays, indicating task-specific utilization of in vitro toxicity information. For instance, the model focused on NR-AR-LBD for carcinogenicity, NR-AhR and SR-MMP for DILI, and SR-p53 for genotoxicity. These targets are known to be associated with corresponding toxicity [53–56]. This observation suggests that each in vivo toxicity selectively

attends to specific Tox21 assays, potentially reflecting their underlying biological relevance. In conclusion, the cross-attention mechanism not only enhances model performance, but also provides interpretability by revealing which auxiliary information is utilized in the prediction process.

Validation as a drug development tool through toxicity screening

To prove the applicability of the proposed model, we performed in vivo toxicity predictions on compounds collected from the DrugBank [57] database (Fig. 4). To conduct this experiment, we employed the MT-Tox model to predict the in vivo toxicities across five drug categories in DrugBank: nutraceutical, approved, investigational, withdrawn, and experimental (Fig. 4A).

First, we compared the predicted scores among four groups (nutraceutical, approved, investigational, and withdrawn) to analyze the inter-group differences (Fig. 4B). The predicted scores were consistently lowest for nutraceuticals across all types of toxicity, and generally increased in the order of approved, investigational, and withdrawn. This tendency aligns with the potential toxicity associated with the different stages of drug development. For example, as nutraceuticals typically consist of dietary supplements or food additives that can be consumed daily, they are expected to exhibit minimal toxicity. Approved and investigational drugs are also expected to have a low risk of in vivo toxicity since they have already passed preclinical stages. In contrast, withdrawn drugs are discontinued due to the safety issues or lack of commercial viability, and therefore, they may have higher predicted toxicity compared to others. Furthermore, the DILI prediction scores for investigational compounds were significantly higher than those for nutraceuticals and approved drugs (one-sided Mann–Whitney U test, p-value: 0.001). Since DILI labels are typically assigned during the clinical phase, investigational compounds that are still in clinical trials may show higher DILI scores than approved drugs. Taken together, these results demonstrate that MT-Tox provides discriminative toxicity predictions across different DrugBank categories. This supports the reliability of MT-Tox as a predictive model for toxicity evaluation in drug development. Second, we performed a toxicity screening on compounds in the Experimental group from DrugBank (Fig. 4C). We selected this group for in vivo toxicity screening since it includes compounds being tested for target efficacy or at the preclinical stage. The prediction results showed that several compounds were highly predicted to be toxic in terms of DILI and genotoxicity. In addition, we examined structures of compounds predicted as positive. For this analysis, we manually selected compounds with prediction scores exceeding 0.8 for each toxicity endpoint (Fig. 4D-F). For the carcinogenicity case, the compounds contained azobenzene or nitroso groups, which are well-known structural alerts for carcinogenic potential [43] (Fig. 4D). In particular, Ranimustine (DB13832) has been studied as a DNA alkylating agent and may be potentially related to carcinogenicity. In the case of DILI, compounds containing nitrobenzene moieties were commonly predicted to have high scores. This functional group has been reported to be associated with hepatotoxicity [58] (Fig. 4E). Notably, Aranidipine (DB09229) is annotated in DrugBank with evidence of hepatocellular hypertrophy in repeated dose studies in rats and mice. Lastly, the compounds in genotoxicity were found to be predominantly nucleobase derivatives (Fig. 4F). These structures are commonly found in anticancer agents that target DNA, and may carry a risk of genomic damage. For instance, 8-azaguanine (DB01667) has antineoplastic effects but has also been reported to be associated with purine nucleotide biosynthesis [59].

Additionally, to further evaluate the toxicity screening capability of MT-Tox, we conducted a toxicity screening experiment on recently reported drugs having toxicity (see **Sect. 9** in Additional File 1). For this experiment, 13 drugs that reported as causing DILI between 2022 and 2025 were collected and screened using our model. As shown in **Table S12**, the results show that MT-Tox correctly predicted nine out of the 13 compounds as toxic. Notably, three true positive compounds (zelnecirnon, ocaliva, orelabrutinib) were predicted as non-toxic by majority of baseline models trained on the same dataset (**Table S13**). This result highlights the ability of our model to correctly recognize DILI-associated toxic structures. Summing up these results, our model captures structurally meaningful features associated with each toxicity type, supporting its applicability as a screening tool for structural filtering in the early-stage drug development process.

Discussion

In vivo toxicity could be influenced by multiple factors such as molecular structure, host-specific genetic and biological conditions, protein-binding interactions, and systemic drug exposure levels. Therefore, accurate prediction of in vivo toxicity requires the integration of diverse mechanistic information, such as in vitro assay results and pharmacokinetic data. However, identifying combination of in vitro assays that reflect in vivo toxicity mechanism is challenging, as the selection process of relevant assays is often biased and uncertain. This complexity is exemplified by our observation of low toxicity label correlation between overlapping compounds in the Tox21 and in vivo toxicity datasets (See **Table S14** in

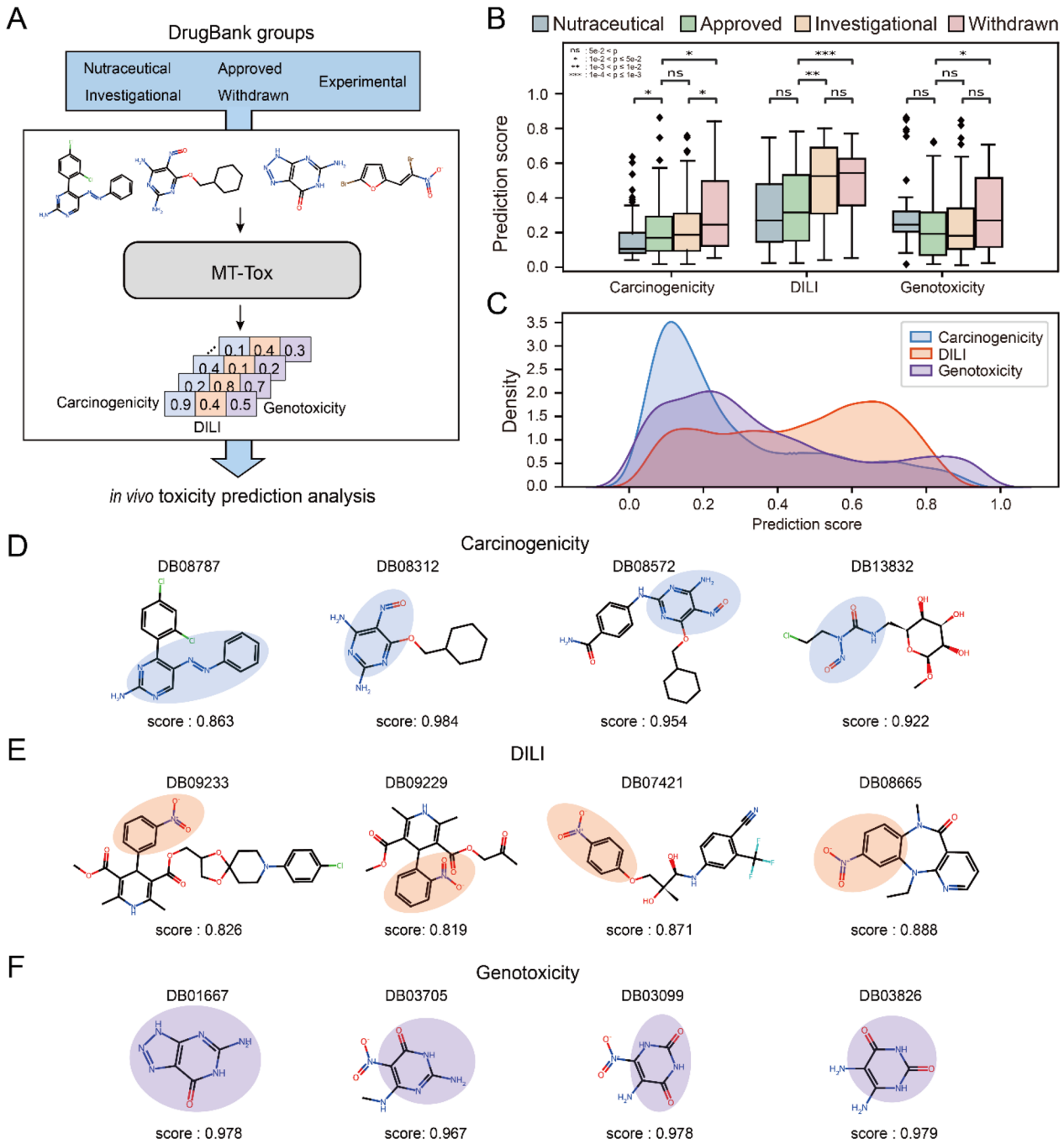

Fig. 4 Analysis results of real-world application to the DrugBank screening. (A) Workflow of the model prediction with the DrugBank dataset. (B) Box plot illustrates the distribution of the predicted in vivo toxicity scores for the four DrugBank groups. Compounds were randomly sampled for each group (n = 54) to match the number of compounds in the withdrawn category, which had the fewest number of compounds. The asterisks indicate the statistical significance results from the one-sided Mann–Whitney U test. (C) Density of the predicted in vivo toxicity scores from the experimental group in the DrugBank. The heatmap shows the average attention scores of Tox21 endpoints from the cross-attention modules with the true positive compounds for each in vivo toxicity. (D-F) Visualization of the high-scored compounds based on the prediction scores for carcinogenicity, DILI, and genotoxicity. Each molecule is labeled with its DrugBank ID and the predicted scores for the corresponding task. The highlighted regions indicate the common functional groups in the high-scored compounds

Additional file 2). In this context, the knowledge transfer framework of our model provides a clear advantage. By applying cross-attention for in vitro toxicological knowledge transfer at the graph embedding level, it allows the model to identify useful signals even in noisy auxiliary data. Ablation studies and task attention analyses further confirmed that Tox21-derived information contributes meaningfully to the model performance of each in vivo toxicity prediction.

Furthermore, the cross-attention mechanism in MT-Tox provides flexibility for incorporating additional features. In MT-Tox architecture, the auxiliary information is injected by the key–value in the cross-attention module. Therefore, integrating an auxiliary information module pre-trained with the same backbone into the cross-attention mechanism allows model extension without any additional explicit changes to the fine-tuning model. As a proof of concept, we conducted a case study applying MT-Tox to DILI prediction by incorporating pharmacokinetic properties (see **Sect. 10** in Additional File 1). Specifically, we used Cmax total [37] and half-life [60] data as auxiliary inputs. The model performances for pharmacokinetics and in vivo fine-tuning set are described in the **Table S15-16** in the Additional file 1. As a result, we observed consistent performance gains across all evaluation metrics, that demonstrates not only the ease of incorporating auxiliary information, but also suggest that integrating in vivo–relevant features such as pharmacokinetics may further enhance the model's predictive capability.

While we have discussed about the various advantages of our model, there is still room for improvement. First, despite incorporating various knowledge transfer strategies to address the low-data regime, the model may still be vulnerable to a restricted applicability domain. This limitation could be mitigated by integrating out-of-distribution handling techniques like Bayesian neural networks [61] or transductive algorithms [62]. Second, while we confirmed the effectiveness of our knowledge transfer scheme throughout the study, its contribution to DILI prediction was comparatively suboptimal. We found that this phenomenon is related to the complexity of DILI, which manifests as a clinical side effect with multiple subtypes such as hepatitis, cirrhosis [63] and diverse mechanisms [64]. Employing additional in vitro toxicity proxies and fine-grained in vivo phenotype modeling could help build models with higher resolution and improved accuracy. Finally, the in vitro auxiliary information used in our study was limited to the 12 bioassays from the Tox21 challenge, which may be insufficient to reflect the diversity of in vitro bioactivity information. We expect that incorporating data containing diverse compound-gene/protein level information such as the Comparative Toxicogenomics Database (CTD) [65] could enhance the resolution of in vitro auxiliary information and lead to a more advanced model for in vivo toxicity prediction. These improvements will further enhance the robustness and generalizability of the model in more complex scenarios.

Conclusion

In this study, we proposed a knowledge transfer framework for improving in vivo toxicity prediction named MT-Tox. MT-Tox utilizes two sequential knowledge transfer strategies: general chemical knowledge pre-training and in vitro toxicological information training. The performance comparison showed that MT-Tox achieved improved and more balanced performance across all tasks than other baseline models based on MTL and TL. Furthermore, in the knowledge transfer components and attention analyses, we confirmed that each component contributes to performance improvement through different mechanisms. The general chemical knowledge transfer provides a stable starting point for model training by pre-training a GNN that understands the substructural information of molecules. Meanwhile, the in vitro toxicological information transfer enhanced the graph representation specific to in vivo toxicity by transferring the in vitro toxicity context through a cross-attention mechanism. Finally, our analysis of the DrugBank compounds prediction demonstrated the potential applicability of MT-Tox as a screening tool in real-world drug discovery settings. Through this analysis, we prioritized candidate compounds that may warrant further toxicological assessment and suggested potential structural features associated with specific toxicity risks.

In conclusion, our work demonstrated meaningful progress in in vivo toxicity prediction by leveraging diverse in vitro data, while also offering interpretability for the prediction process. We expect that further utilization of MT-Tox may support toxicity screening and molecular optimization in the early drug discovery process.

Abbreviations

ML	Machine learning
DL	Deep learning
TL	Transfer learning
MTL	Multi-task learning
STL	Single-task learning
IVIVE	in vitro to in vivo extrapolation
DILI	Drug induced liver injury
SMILES	Simplified Molecular Input Line Entry System
GNN	Graph neural networks
D-MPNN	Directed message-passing neural network
GMHAP	Global multi-head attentive pooling
MLP	Multi-layered perceptron
ReLU	Rectified linear unit
RF	Random forest
ECFP	Extended connectivity fingerprints
BAC	Balanced accuracy
AUPR	Area under the precision-recall curve

AUC Area under the receiver operating characteristics curves

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01110-4.

Additional file 1. Additional description of the Materials and Methods, Results and Discussion with figures and tables.

Additional file 2. Excel sheet for full prediction performance metrics for model comparison. It contains detailed performance results (BAC, AUPR, AUC, Sum of rank) for all models evaluated on internal and external test sets in the main text and Additional file 1.

Acknowledgements
We appreciate the high-performance GPU computing support of HPC-AI Open Infrastructure via GIST SCENT.

Author contributions
MP and HK, HN conceptualized the study. MS implemented the model and conducted experiments. MS and YS conducted the external performance comparison and prepared the initial draft. HK and HN revised the manuscript.

Funding
This work was supported by a grant of the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (grant number: RS-2024-00460010), and supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2025-02304253).

Availability of data and materials
The data and source code for MT-Tox are available at our GitHub repository (https://github.com/GIST-CSBL/MT-Tox) under the PolyForm Noncommercial License 1.0.0. Parts of the source code in our code were adapted from BayeshERG repository (https://github.com/GIST-CSBL/BayeshERG). All trained model weights and generated data using our source code are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). The raw datasets for ChEMBL, Tox21 and DrugBank used in this study are publicly available. GuacaMol benchmark dataset [https://github.com/BenevolentAI/guacamol] ChEMBL 24 (used in GuacaMol) [https://www.ebi.ac.uk/chembl/] Tox21 public data [https://tripod.nih.gov/pubdata/] DrugBank (release version 5.1.13) [https://go.drugbank.com/].

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Buk-Gu, Gwangju 61005, Republic of Korea. [2]AI Graduate School, Gwangju Institute of Science and Technology (GIST), Buk-Gu, Gwangju 61005, Republic of Korea. [3]Present Address: Division of Advanced Predictive Research, Korea Institute of Toxicology, 141 Gajeong-Ro, Yuseong-Gu, Daejeon 34114, Republic of Korea. [4]Center for AI-Applied High Efficiency Drug Discovery (AHEDD), Gwangju Institute of Science and Technology (GIST), Buk-Gu, Gwangju 61005, Republic of Korea.

Received: 21 May 2025 Accepted: 13 October 2025
Published online: 12 November 2025

References

1. Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS (2016) Failure of investigational drugs in late-stage clinical development and publication of trial results. JAMA Intern Med 176:1826–1833
2. Kramer JA, Sagartz JE, Morris DL (2007) The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. Nat Rev Drug Discov 6:636–649
3. Akhtar A (2015) The flaws and human harms of animal experimentation. Camb Q Healthc Ethics 24:407–419
4. Jiang C, Yang H, Di P, Li W, Tang Y, Liu G (2019) In silico prediction of chemical reproductive toxicity using machine learning. J Appl Toxicol 39:844–854
5. Liu R, Madore M, Glover KP, Feasel MG, Wallqvist A (2018) Assessing deep and shallow learning methods for quantitative prediction of acute chemical toxicity. Toxicol Sci 164:512–526
6. Seal S, Mahale M, García-Ortegón M, Joshi CK, Hosseini-Gerami L, Beatson A, Greenig M, Shekhar M, Patra A, Weis C et al (2025) Machine learning for toxicity prediction using chemical structures: pillars for success in the real world. Chem Res Toxicol 38:759–807
7. Tonoyan L, Siraki AG (2024) Machine learning in toxicological sciences: opportunities for assessing drug toxicity. Front Drug Disc 2024:4
8. Wang MWH, Goodman JM, Allen TEH (2021) Machine learning in predictive toxicology: recent applications and future directions for classification models. Chem Res Toxicol 34:217–239
9. Xu Y, Pei J, Lai L (2017) Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. J Chem Inf Model 57:2672–2685
10. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, Lai L, Pei J (2020) Transfer learning for drug discovery. J Med Chem 63:8683–8694
11. Simões RS, Maltarollo VG, Oliveira PR, Honorio KM (2018) Transfer and multi-task learning in QSAR modeling: advances and challenges. Front Pharmacol 2018:9
12. Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, Wang F, Wu H, Wang H (2022) Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell 4:127–134
13. Liu S, Wang H, Liu W, Lasenby J, Guo H, Tang J: Pre-training Molecular Graph Representation with 3D Geometry. pp. arXiv:2110.07728; 2021:arXiv:2110.07728.
14. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J (2020) Self-supervised graph transformer on large-scale molecular data. Adv Neural Inf Process Syst 33:12559–12571
15. Wang Y, Wang J, Cao Z, Barati Farimani A (2022) Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 4:279–287
16. Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, Fan X, Chen H (2023) Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nat Mach Intell 5:542–553
17. Lu Y, Jiang X, Fang Y, Shi C (2021) Learning to pre-train graph neural networks. Proc AAAI Conf Artif Intell 35:4276–4284
18. Zhang Y, Yang Q (2022) A Survey on Multi-Task Learning. IEEE Trans Knowl Data Eng 34:5586–5609
19. Sharma B, Chenthamarakshan V, Dhurandhar A, Pereira S, Hendler JA, Dordick JS, Das P (2023) Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. Sci Rep 13:4908
20. Wu Z, Jiang D, Wang J, Hsieh C-Y, Cao D, Hou T (2021) Mining toxicity information from large amounts of toxicity data. J Med Chem 64:6924–6936
21. Jain S, Siramshetty VB, Alves VM, Muratov EN, Kleinstreuer N, Tropsha A, Nicklaus MC, Simeonov A, Zakharov AV (2021) Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods. J Chem Inf Model 61:653–663
22. Lui R, Guan D, Matthews S (2023) Mechanistic task groupings enhance multitask deep learning of strain-specific Ames mutagenicity. Chem Res Toxicol 36:1248–1254
23. Martínez MJ, Sabando MV, Soto AJ, Roca C, Requena-Triguero C, Campillo NE, Páez JA, Ponzoni I (2022) Multitask deep neural networks for Ames mutagenicity prediction. J Chem Inf Model 62:6342–6351
24. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera Juan F et al (2023) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52:D1180–D1192
25. Huang R: A Quantitative High-Throughput Screening Data Analysis Pipeline for Activity Profiling. In *High-Throughput Screening Assays in*

Toxicology. Edited by Zhu H, Xia M. New York, NY: Springer New York; 2016: 111–122
26. Han P, Li X, Yang J, Zhang Y, Chen J (2024) Advancing toxicity predictions: a review on in vitro to in vivo extrapolation in next-generation risk assessment. Environ Health 2:499–513
27. Brown N, Fiscato M, Segler MHS, Vaucher AC (2019) GuacaMol: Benchmarking Models for de Novo Molecular Design. J Chem Inf Model 59:1096–1108
28. RDKit: Open-source cheminformatics [https://www.rdkit.org]
29. Huang R, Xia M, Nguyen D-T, Zhao T, Sakamuru S, Zhao J, Shahane SA, Rossoshek A, Simeonov A: Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Frontiers in Environmental Science* 2016, Volume 3 - 2015.
30. Garcia M, Morger A, Norinder U, Buesen R, Landsiedel R, Volkamer A, Kirchmair J, Mathea M (2021) ChemBioSim: enhancing conformal prediction of in vivo toxicity by use of predicted bioactivities. J Chem Inform Model 61:3255–3272
31. Li T, Tong W, Roberts R, Liu Z, Thakkar S (2021) DeepDILI: deep learning-powered drug-induced liver injury prediction using model-level representation. Chem Res Toxicol 34:550–565
32. Ma H, An W, Wang Y, Sun H, Huang R, Huang J (2021) Deep graph learning with property augmentation for predicting drug-induced liver injury. Chem Res Toxicol 34:495–506
33. Yoo JW, Kruhlak NL, Landry C, Cross KP, Sedykh A, Stavitskaya L (2020) Development of improved QSAR models for predicting the outcome of the in vivo micronucleus genetic toxicity assay. Regul Toxicol Pharmacol 113:104620
34. Zhang C, Cheng F, Li W, Liu G, Lee PW, Tang Y (2016) In silico prediction of drug induced liver toxicity using substructure pattern recognition method. Mol Inform 35:136–144
35. Zhang L, Ai H, Chen W, Yin Z, Hu H, Zhu J, Zhao J, Zhao Q, Liu H (2017) CarcinoPred-EL: novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. Sci Rep 7:2118
36. Dolciami D, Villasclaras-Fernandez E, Kannas C, Meniconi M, Al-Lazikani B, Antolin AA (2022) CanSAR chemistry registration and standardization pipeline. J Cheminform 14:28
37. Seal S, Williams D, Hosseini-Gerami L, Mahale M, Carpenter AE, Spjuth O, Bender A (2024) Improved detection of drug-induced liver injury by integrating predicted in vivo and in vitro data. Chem Res Toxicol 37:1290–1305
38. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker Benjamin A, Thiessen Paul A, Yu B et al (2024) PubChem 2025 update. Nucleic Acids Res 53:D1516–D1525
39. Fan D, Yang H, Li F, Sun L, Di P, Li W, Tang Y, Liu G (2018) In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. Toxicol Res 7:211–220
40. Fradkin P, Young A, Atanackovic L, Frey B, Lee LJ, Wang B (2022) A graph neural network approach for molecule carcinogenicity prediction. Bioinformatics 38:i84–i91
41. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L (2015) Deep learning for drug-induced liver injury. J Chem Inf Model 55:2085–2093
42. Kim H, Park M, Lee I, Nam H (2022) BayeshERG: a robust, reliable and interpretable deep learning model for predicting hERG channel blockers. Brief Bioinform. https://doi.org/10.1093/bib/bbac211
43. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M et al (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59:3370–3388
44. Mayr A, Klambauer G, Unterthiner T, Hochreiter S (2015) DeepTox: toxicity prediction using deep learning. Front Environ Sci 2016:3
45. Yuan Li Y, Chen L, Pu C, Zang C, Yan Y, Chen Y, Zhang Y, Liu H (2023) Co-model for chemical toxicity prediction based on multi-task deep learning. Mol Inform 42:2200257
46. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al: PyTorch: An Imperative Style, High-Performance Deep Learning Library. pp. arXiv:1912.01703; 2019:arXiv:1912.01703.
47. Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. SIAM J Control Optim 30:838–855
48. Tarvainen A, Valpola H (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv Neural Inform Process Syst 30:1
49. Ahmad W, Simon E, Chithrananda S, Grand G, Ramsundar B: ChemBERTa-2: Towards chemical foundation models. *arXiv preprint arXiv:220901712* 2022.
50. Atienzar FA, Blomme EA, Chen M, Hewitt P, Kenna JG, Labbe G, Moulin F, Pognan F, Roth AB, Suter-Dick L et al (2016) Key challenges and opportunities associated with the use of in vitro models to detect human DILI: integrated risk assessment and mitigation plans. BioMed Res Int 2016:9737920
51. Shin HK, Kang M-G, Park D, Park T, Yoon S (2020) Development of prediction models for drug-induced cholestasis, cirrhosis, hepatitis, and steatosis based on drug and drug metabolite structures. Front Pharmacol 2020:11
52. Benigni R, Bossa C, Tcheremenskaia O (2013) Nongenotoxic carcinogenicity of chemicals: mechanisms of action and early recognition through a new set of structural alerts. Chem Rev 113:2940–2957
53. Chiu WA, Guyton KZ, Martin MT, Reif DM, Rusyn I (2018) Use of high-throughput in vitro toxicity screening data in cancer hazard evaluations by IARC monograph working groups. ALTEX - Alternatives to animal experimentation 35:51–64
54. Hsieh J-H, Smith-Roe SL, Huang R, Sedykh A, Shockley KR, Auerbach SS, Merrick BA, Xia M, Tice RR, Witt KL (2019) Identifying compounds with genotoxicity potential using Tox21 high-throughput screening assays. Chem Res Toxicol 32:1384–1401
55. Khadka KK, Chen M, Liu Z, Tong W, Wang D (2020) Integrating adverse outcome pathways (AOPs) and high throughput *in vitro* assays for better risk evaluations, a study with drug-induced liver injury (DILI). ALTEX 37:187–196
56. Peraza MA, Burdick AD, Marin HE, Gonzalez FJ, Peters JM (2005) The toxicology of ligands for peroxisome proliferator-activated receptors (PPAR). Toxicol Sci 90:269–295
57. Knox C, Wilson M, Klinger-Christen M, Franklin M, Oler E, Wilson A, Pon A, Cox J, Chin NE, Strawbridge-Seth A et al (2023) DrugBank 60: the DrugBank Knowledgebase for 2024. Nucleic Acids Res 52:1265–1275
58. Li H, Zhang Z, Yang X, Mao X, Wang Y, Wang J, Peng Y, Zheng J (2019) Electron deficiency of nitro group determines hepatic cytotoxicity of Nitrofurantoin. Chem Res Toxicol 32:681–690
59. Kim N, Choi J-W, Song AY, Choi WS, Park H-R, Park S, Kim I, Kim HS (2019) Direct potentiation of NK cell cytotoxicity by 8-azaguanine with potential antineoplastic activity. Int Immunopharmacol 67:152–159
60. Lombardo F, Berellini G, Obach RS (2018) Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 1352 drug compounds. Drug Metab Dispos 46:1466–1477
61. Gal Y, Ghahramani Z: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning* (Maria Florina B, Kilian QW eds.), vol. 48. pp. 1050--1059. Proceedings of Machine Learning Research: PMLR; 2016:1050--1059.
62. Netanyahu A, Gupta A, Simchowitz M, Zhang K, Agrawal P: Learning to Extrapolate: A Transductive Approach. In *The Eleventh International Conference on Learning Representations*; 2023.
63. Shin HK, Chun H-S, Lee S, Park S-M, Park D, Kang M-G, Hwang S, Oh J-H, Han H-Y, Kim W-K, Yoon S (2022) ToxSTAR: drug-induced liver injury prediction tool for the web environment. Bioinformatics 38:4426–4427
64. Skat-Rørdam J, Lykkesfeldt J, Gluud LL, Tveden-Nyborg P (2025) Mechanisms of drug induced liver injury. Cell Mol Life Sci 82:213
65. Davis AP, Wiegers TC, Sciaky D, Barkalow F, Strong M, Wyatt B, Wiegers J, McMorran R, Abrar S, Mattingly CJ (2024) Comparative toxicogenomics database's 20th anniversary: update 2025. Nucleic Acids Res 53(D1):D1328–D1334

Publisher's Note