nature communications

Article

https://doi.org/10.1038/s41467-025-60989-7

ToxACoL: an endpoint-aware and task-focused compound representation learning paradigm for acute toxicity assessment

Received: 24 November 2024

Accepted: 30 May 2025

Published online: 01 July 2025

Check for updates

Jiang Lu[1,2,3,7], Lianlian Wu[1,2,7], Ruijiang Li[2], Mengxuan Wan[4], Jun Yang[5], Peng Zan[4], Hui Bai[6] ✉, Song He[2] ✉ & Xiaochen Bo[1,2,3] ✉

Multi-species acute toxicity assessment forms the basis for chemical classification, labelling and risk management. Existing deep learning methods struggle with diverse experimental conditions, imbalanced data, and scarce target data, hindering their ability to reveal endpoint associations and accurately predict data-scarce endpoints. Here we propose a machine learning paradigm, Adjoint Correlation Learning, for multi-condition acute toxicity assessment (ToxACoL) to address these challenges. ToxACoL models endpoint associations via graph topology and achieves knowledge transfer via graph convolution. The adjoint correlation mechanism encodes compounds and endpoints synchronously, yielding endpoint-aware and task-focused representations. Comprehensive analyses demonstrate that ToxACoL yields 43%-87% improvements for data-scarce human endpoints, while reducing training data by 70% to 80%. Visualization of the learned top-level representation interprets structural alert mechanisms. Filled-in toxicity values highlight potential for extrapolating animal results to humans. Finally, we deploy ToxACoL as a free web platform for rapid prediction of multi-condition acute toxicities.

Today, explosive growth, massive production, widespread application, and long-term emissions of various chemicals have induced a huge threat to human health and the environment[1–3]. Multi-species acute systemic toxicity assessment forms the basis for chemical classification, labeling and risk management[4,5]. Acute toxicity assessment is typically conducted as the initial phase of the entire safety assessment[6–9] and is usually mandatory to the regulatory procedures of new chemicals, directly determining whether the chemicals can enter subsequent industrial use or clinical trials[10–12]. Acute toxicity evaluates the unwanted effects that occur either immediately or at a short time interval after a single or multiple administration of a substance[13]. In some cases, acute systemic toxicity data may be used to establish doses for longer-term studies, identify target organs for toxicity, and assess the hazard of accidental ingestions of chemical contaminants[14]. Due to ethical and legal restrictions, large-scale experimental testing on humans or certain unconventional species, such as wildlife, is not feasible. Conventional animal testing remains the primary method for acute toxicity assessment[15,16]. These tests involve diverse conditions, such as various species, administration routes, and assessment indicators[17]. However, the experimental results

[1]Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, People's Republic of China. [2]Department of Advanced & Interdisciplinary Biotechnology, Academy of Military Medical Sciences, Beijing, People's Republic of China. [3]Institute of Advanced Technology and Equipment, Xi'an Jiaotong University, Xi'an, People's Republic of China. [4]Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronics Engineering and Automation, Shanghai University, Shanghai, People's Republic of China. [5]Department of Cell Biology, School of Life Sciences, Central South University, Changsha, People's Republic of China. [6]Clinical Translational Research Center, Beijing Tsinghua Changgung Hospital, Tsinghua University, Beijing, People's Republic of China. [7]These authors contributed equally: Jiang Lu, Lianlian Wu. ✉e-mail: huibai13@hotmail.com; hes1224@163.com; boxc@bmi.ac.cn

for the same compound vary significantly under different species and different testing conditions. When extrapolating test doses and toxicity from experimental species to humans and other unconventional species, significant gap in physiological structures and metabolic mechanisms often lead to inaccurate toxicity assessments[18]. Especially in drug discovery, toxicity to the human body often appears in clinical trials or after market launch although no toxicity was found in preclinical animal testing, leading to the failure of drug discovery. A classic example is troglitazone. No toxicity was found in vitro and in animal testing. Shortly after it was approved for the treatment of type 2 diabetes, it was withdrawn from the market due to its hepatotoxicity. This emphasizes the importance of accurately predicting or extrapolating the toxicity of compounds to humans.

Modern toxicology emphasizes the 3Rs (replacement, reduction, and refinement) principle in animal testing, seeking alternative methods for toxicity assessment[19,20]. Computational methods have emerged as powerful tools for replacing animal testing. Among them, machine learning (ML) and deep learning (DL) have demonstrated significant advantages over traditional quantitative structure–activity relationship (QSAR) statistical methods in terms of prediction scope and accuracy, especially in the context of rapidly growing toxicity data[21–24]. For the convenience of ML modeling, several ML-ready toxicity databases have been established, providing extensive acute toxicity data resources. For example, in our previous work, TOXRIC database[25], we collected 59 multi-species, multi-endpoint acute toxicity values, covering over 80,000 compounds. Luechtefeld et al.[4,26] constructed a database of 10,000 chemicals and 800,000 toxicological studies, focusing on acute oral toxicity among other endpoints.

There were several reports indicating that ML trained on masses of toxicity data is so good at predicting some toxicities and sometimes outperforms expensive animal studies[16,27–29]. Studies have developed ML algorithms for acute toxicity assessment, encompassing single-task learning (STL), multi-task learning (MTL) methods, and consensus modeling. In acute toxicity prediction under STL paradigms, existing studies have reported that random forest (RF) exhibits optimal performance among traditional ML algorithms including support vector machines (SVM), artificial neural networks (ANN), message-passing neural networks (MPNN), gradient boosting (GB), Xgboost (XGB), and generalized linear models (GLM)[30–32]. Luechtefeld et al.[28] developed read-across structure–activity relationship (RASAR) models that employ binary fingerprints and Jaccard similarity to construct chemical adjacency matrices, with data Fusion RASAR improves predictive accuracy by integrating multi-property features to train RF. While graph-based algorithms employing molecular graph inputs, such as graph convolution network (GCN) and attentive fingerprint (Attentive FP)[33], have shown better performance over RF and other traditional ML algorithms[34,35]. However, they lack the ability to integrate multi-condition acute toxicity data for training, resulting in poor performance on small-sized endpoints with insufficient data. MTL models recognize the correlations between multiple endpoints and train on multi-condition toxicity data. They extract molecule representations via a shared encoder and use a multi-channel regression module to simultaneously predict multi-condition acute toxicity. Previous studies have proven that multi-task deep neural network (MT-DNN) and multi-task graph convolution network (MT-GCN) can effectively improve average performance across all endpoints compared to all STL implementations for acute toxicity prediction[17,32,36,37]. Their superiority over STL indirectly indicates that the shared learning strategy among diverse endpoints can mutually boost each other[38]. Additionally, consensus models integrate multiple model (STL or MTL) outputs, demonstrating better performance compared to individual models. Jain et al.[17] developed a consensus framework named DLCA that integrates multiple MT-DNNs with different feature inputs, showing improved performance compared to individual models. Mansouri et al.[5] constructed consensus models for five endpoints of acute toxicity, and proposed a weight-of-evidence (WoE) approach with weighted integration to generate consensus predictions of the five endpoints. Effective performance improvement was achieved at single endpoints.

However, due to the inherent complexities of acute toxicity endpoints, such as diverse experimental conditions, scarce target endpoint data, and imbalanced endpoint data, existing ML methods struggle to reveal relationships among multi-condition endpoints, fail to accurately predict data-scarce endpoints (especially in humans), and have not explored the extrapolation patterns between species. Firstly, acute toxicity in vivo experiments involve various species (e.g., mouse, rabbit, and dog, etc.), administration routes (e.g., intravenous, skin, oral, etc.), and measurement indicators like median lethal dose (LD_{50}), lethal dose low (LDLo) and toxic dose low (TDLo). These diverse experimental conditions across studies make it difficult to model the relationships among multi-condition endpoints using a unified mathematical logic or general methodology. Additionally, ethical and legal restrictions make it difficult to obtain training data for target endpoints, such as humans and certain unconventional species. Large-scale reference data for expensive experimental species are also scarce, resulting in extreme data imbalance across endpoints. While existing MTL models effectively improve the average performance across all endpoints, they struggle to make accurate predictions for data-scarce target endpoints. This may be due to the varying toxicity intensity of compounds across species, which creates a significant gap between endpoints[39]. This further emphasizes the importance of identifying and modeling the relationships between multi-condition endpoints. Moreover, such modeling supports the exploration of extrapolation patterns between species. Understanding the differences in toxicity responses across species and identifying which species exhibit toxicity patterns most similar to humans are key questions that warrant further exploration.

To address these challenges, We introduce a machine learning paradigm, Adjoint Correlation Learning, for multi-species acute toxicity assessment of compounds, named as ToxACoL. We first collect multi-species, multi-condition acute toxicity data from some public chemical databases, such as TOXRIC[25] and PubChem[40], which involve various test species, administration routes, and measurement indicators. Based on these data, ToxACoL is designed to apply graph topology to model relationships between multi-condition endpoints, incorporating an adjoint correlation mechanism to process and integrate multi-endpoint information with compound representations in parallel. In ToxACoL, graph convolution is employed to propagate information on multi-endpoints and their relationships, while a feed-forward network with residual connections is used to propagate compound embeddings. The two branches interact through a correlation operation to learn endpoint-aware compound representations. Comprehensive analyses on a 59-endpoints acute toxicity dataset demonstrate that ToxACoL effectively captures relationships among multi-condition endpoints and achieves balanced performance across them. By learning the relationships across endpoints, ToxACoL significantly improves prediction accuracy for data-scarce endpoints, including human-oral-TDLo, women-oral-TDLo, and man-oral-TDLo, with performance increases of 56%, 87%, and 43%, respectively, compared to state-of-the-art methods. Meanwhile, ToxACoL reduces the required training data for sparse endpoints by ~70–80%, aligning with the global 3Rs principle of reducing animal testing. Then, tested on two additional benchmark datasets (115-endpoint dataset and study of Mansouri et al.[5]), ToxACoL demonstrates robust performance compared to representative acute toxicity prediction models. AD analysis depicts the chemical space where ToxACoL achieves reliable predictions. Furthermore, analysis of ToxACoL's top-level representations assists in identifying structural alerts, uncovering potential mechanisms of acute toxicity, and providing insights for extrapolating animal test results to humans. Finally, an online platform for the prediction of acute toxicity was developed that integrates the ToxACoL model, serving as a freely accessible and useful resource for regulatory applications.

Results

ToxACoL models relationships among multi-condition endpoints by introducing the adjoint correlation mechanism

ToxACoL introduces the adjoint correlation mechanism to parallelly learn multi-condition labels and multi-type sample information (Fig. 1a), achieving good performance in multi-condition acute toxicity assessment. This paradigm insists on bidirectional learning from compounds (samples) and multi-condition endpoints (labels) simultaneously. By capturing endpoint-to-endpoint dependencies, the relationships between multi-condition endpoints are captured and a multi-condition endpoint graph is constructed. The adjoint correlation mechanism facilitates interaction between the two learning branches, i.e., compounds and endpoints, at each layer, enabling the extraction of endpoint-aware and task-focused compound representations.

The construction of the endpoint graph is based on existing studies about large-scale acute toxicity datasets, that involve multi-condition acute toxicity endpoints, and each endpoint provides information of test species, administration route, and measurement indicator. Taking our previous study, the TOXRIC database[25] for example, which includes 59 various toxicity endpoints with 80,081 unique compounds represented using SMILES strings, and 122,594 usable toxicity measurements described by continuous values with a unified toxicity chemical unit: −log(mol/kg). The 59 acute toxicity endpoints involve a total of 15 test species, 8 administration routes, and 3 measurement indicators, ensuring comprehensive information coverage. However, the size of samples across 59 endpoints is highly imbalanced and sparse (Fig. 1b). Some endpoints involve tens of thousands of available measurement samples, while some endpoints only contain about 100 samples, with a high rate of missing data (Supplementary Table 1). To address this issue, we constructed an endpoint graph to capture inter-endpoint relationships and leverage multi-condition endpoint information to enhance performance on data-scarce endpoints. For each pair of endpoints, we first count their shared training compounds, and then calculate the Pearson correlation coefficients (PCC)[41] of toxicity measurements based on their shared compounds if the quantity of these shared compounds is acceptable (Fig. 1c). Two endpoints are considered dependent, and thus an edge is formed between them, only if the shared compound count exceeds a certain threshold and their toxicity measurements are highly correlated (see the “Methods” section). Based on this deduction, we can construct an acute toxicity endpoint graph, where nodes represent toxic endpoints and edges represent the dependency between two endpoints. Considering that each endpoint contains three attributes: test species, administration route, and measurement indicator, we separately encodes each attribute into a one-hot subvector and then concatenates the three subvectors to initialize graph node features, serving as the initial endpoint embeddings (Fig. 1d).

Next, the adjoint correlation learning is applied between the compounds and the acute toxicity endpoint graph. Avalon fingerprints is adopted as the initial representation for compounds, as previous studies[17,42] have shown that it is the cutting-edge feature for predicting acute toxicity. The adjoint correlation layer is designed to process the compounds and the endpoint graph in parallel, where compound embeddings are processed by feed-forward layers with residual connections and endpoint embeddings are processed by graph convolution. Critically, the new endpoint embeddings are correlated with the compound embeddings, and then accumulated onto the previous compound embeddings in the form of residuals. The compound embeddings after the residual operation, as well as the new endpoint graph after graph convolution, are used together as inputs for the next adjoint correlation layer. In this way, multiple adjoint correlation layers are cascaded sequentially. The endpoint embeddings outputted by the topmost graph convolution are treated as toxicity regressor weights for the corresponding endpoints. These weights are then combined with a top regression layer to assess the toxicity intensity for multiple endpoints.

ToxACoL successfully balances performance across multi-condition endpoints

We first evaluate the predictive performance of ToxACoL and existing state-of-the-art methods under benchmark setting of the 59-endpoint acute toxicity dataset from TOXRIC, which include single-task deep neural networks (ST-DNN)[17], single-task random forest (ST-RF)[17], graph attention network (GAT)[43], graph convolution network (GCN)[44], attentive fingerprint (Attentive FP)[33,43], multi-task deep neural networks (MT-DNN)[17], multi-task graph convolution network (MT-GCN)[17], and deep learning consensus architecture (DLCA)[17,42]. These baseline models have been extensively studied in[17,42–44], and the MTL and consensus models among them still maintain the leading performance for acute toxicity assessment to this day. We maintained the same experimental settings as these baselines to ensure fairness, such as selecting the same Avalon fingerprints as our molecular features and adopting the same 5-fold dataset split for cross-validation. The performance is evaluated using the metrics of determination coefficient (R^2) and root-mean-squared error (RMSE), which respectively reflect the fitness or deviation between the predicted toxicity intensity by computational models and the ground-truth intensity.

We first compared the average performance of these models on all the 59 endpoints via 5-fold cross-validation (Fig. 2a, Supplementary Tables 3 and 4). The results indicated that ToxACoL yielded an improvement in overall average performance to other baseline models. Concretely, ToxACoL achieved an averaged R^2 of 0.5843 and a smaller averaged RMSE of 0.6396, surpassing the previously best-performing algorithm (DLCA). Considering the acute toxicity data is very imbalanced, we investigated the performance of different models in dealing with the dilemma of data imbalance between endpoints. A ridge diagram (Fig. 2b) fitting the performance of each model on all 59 endpoints was drawn based on kernel density estimation (KDE)[45], to intuitively present the overall distribution of toxicity estimation performance of each model on all endpoints. On the one hand, the overall performance of ToxACoL is superior to other baselines. On the other hand, the performance distribution of ToxACoL is more concentrated with a significantly smaller standard deviation than other models, indicating that ToxACoL can more robustly and evenly handle the data-imbalanced multi-condition acute toxicity evaluation tasks.

To further compare these models on single endpoints, we statistically analyzed their performance rankings on each endpoint and visualized the proportion of these rankings via stacked histograms (Fig. 2c). Our ToxACoL ranked in the top two on the vast majority of endpoints. Based on these endpoint-wise rankings, we further made a Friedman and Nemenyi test[46] to intuitively display the averaged performance ranking gap of different models on 59 endpoints via the critical difference (CD) diagram (Fig. 2d). It can be seen that ToxACoL ranked ahead of other baseline models concerning both R^2 and RMSE, demonstrating its holistic superiority to other baseline models. Finally, the endpoint-wise performance of all models were summarized in increasing order of their toxicity measurement samples (Fig. 2e), and it can be observed that ToxACoL surpassed other baseline models on most toxic endpoints.

To verify the indispensability of the adjoint correlation mechanism in ToxACoL, we provided an ablation study to compare the standard ToxACoL and its variant. In this variant model, the feed-forward layer and graph convolution layer in ToxACoL no longer interact at any layer, but independently handle compound embeddings and endpoint embeddings, respectively. We tested six sets of network structures with different depths and recorded the average R^2 through 5-fold

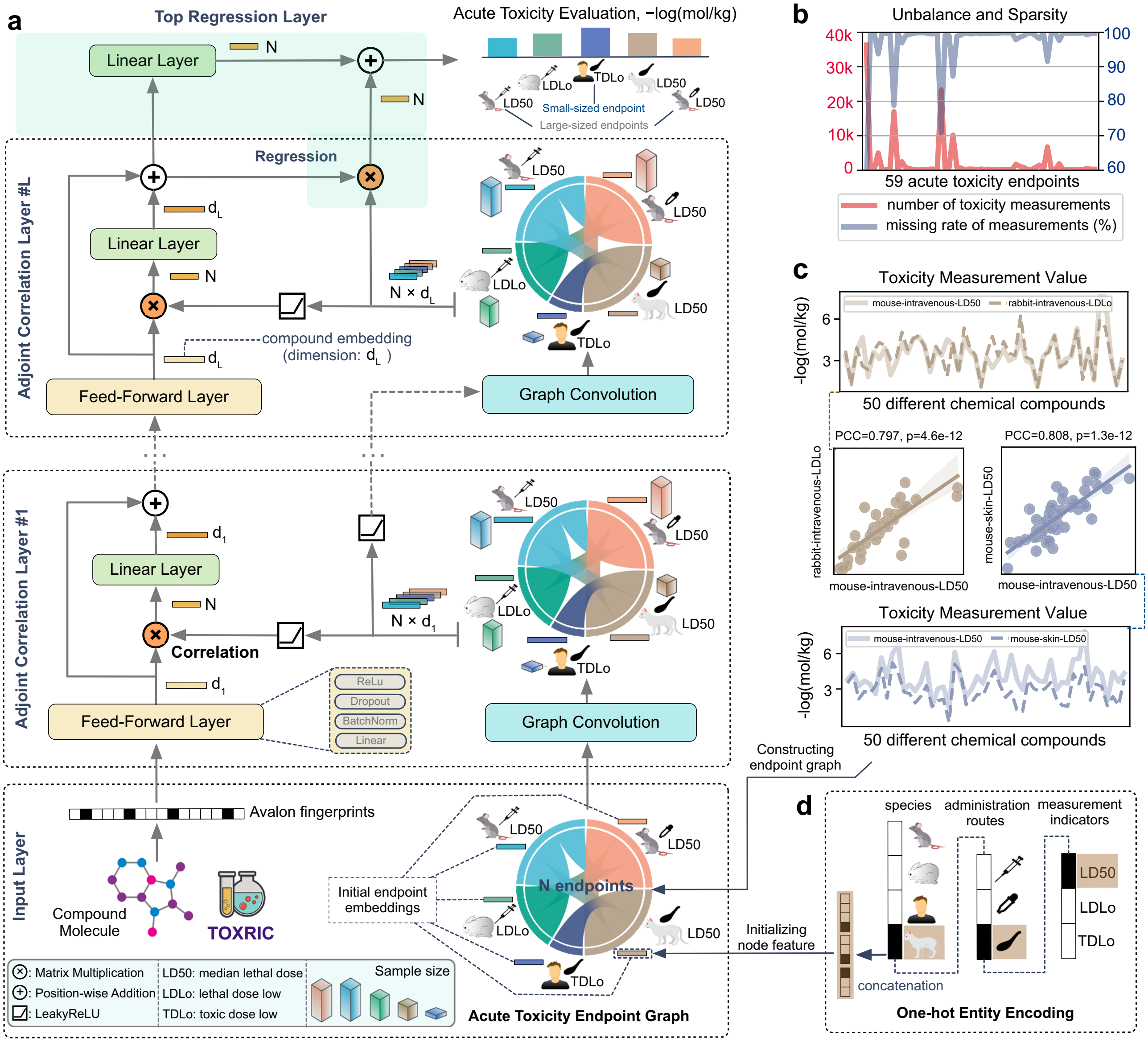


Fig. 1 | High-level overview of ToxACoL. a ToxACoL workfolow. The training toxicity measurements were leveraged to explore pairwise dependencies between endpoints and the acute toxicity endpoint graph was constructed based on these dependencies. Each adjoint correlation layer comprising a residual network layer and a graph convolution layer was designed to process compound embeddings and endpoint embeddings parallelly, and the two branches internally interact via a correlation operation. After a cascade of multiple adjoint correlation layers, the embedding of each endpoint outputted by the topmost graph convolution layer will serve as the toxicity regressor for the corresponding endpoint, and then perform the toxicity regression with the top-level compound embedding, finally outputting toxicity intensity value concerning the corresponding endpoint. **b** Illustration of data imbalance and data sparsity of the large-scale multi-condition acute toxicity dataset. **c** Two examples for calculating pairwise dependencies between endpoints, which were based on the training compounds shared by the two endpoints. The dependency was evaluated via a two-sided Pearson correlation coefficient (PCC) analysis. There exists a significant correlation between mouse-intravenous-LD50 and rabbit-intravenous-LDLo, as well as for mouse-intravenous-LD50 and mouse-skin-LD50. The center line in the correlation plots represents the regressed line and the error band denotes the confidence interval of 0.95 for linear regression. **d** The one-hot entity encoding strategy encompassing three endpoint attributes was developed for initializing endpoint embeddings in graph. Credits: the icons of bottles, chemicals, and animals including mouse, rabbit, cat, and man, along with illustrations of administration tools including spoon, syringe, and dropper, are sourced from https://creazilla.com/. Source data are provided as a Source Data file.

cross-validation. We found that the standard ToxACoL consistently outperformed its variant regardless of network depth (Supplementary Fig. 1). Moreover, as the network becomes deeper, the performance of the variant model has sharply declined, which is mainly caused by the inherent over-smoothing problem existing in conventional GCN[47–49]. On the contrary, the standard ToxACoL appears less sensitive to increasing network layers, and thus the performance gap between the two models widens. It is because the adjoint correlation mechanism can guide ToxACoL to learn better endpoint-aware representations at feed-forward layers, and the extra gradients from adjoint correlation can regularize the learning of endpoint embeddings in GCN. In brief, introducing the adjoint correlation mechanism can effectively overcome the intractable over-smoothing of graph convolution at deeper layers.

ToxACoL achieves a substantial improvement in performance on data-scarce endpoints especially human-related endpoints

In real world, the primary focus of acute toxicity assessments for compounds is on humans and unconventional species (such as specific wildlife). However, toxicity data for these species are

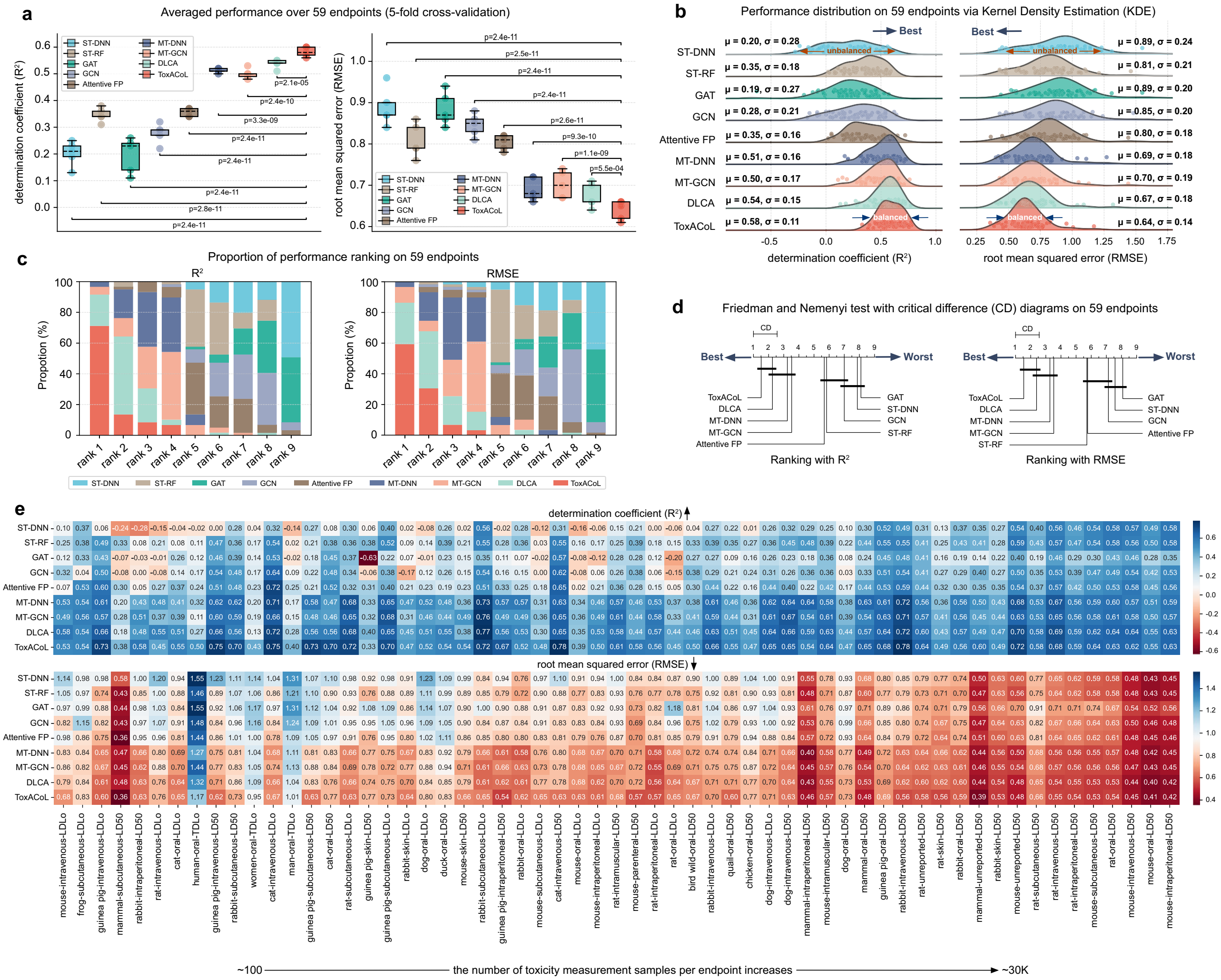


Fig. 2 | Performance comparison for multi-condition acute toxicity estimation on the 59-endpoint dataset. a Average R^2 and RMSE on all toxic endpoints via 5-fold cross-validation. The two-sided Wilcoxon signed-rank test was selected to compute the significant difference between ToxACoL and other baselines across all endpoints. It can be seen that the *p*-values are small, indicating that the improvements by our ToxACoL are statistically significant. The five dots on each box plot represent the results of five cross-validation experiments; the center line in the box represents the median among the five results, excluding outliers; the lower and upper bounds of the box represent the first (Q1) and third (Q3) quartiles, respectively; the lower and upper bounds of the whiskers represent the minima and maxima, excluding outliers, respectively. **b** Overall performance distribution of different models on 59 endpoints, fitted using Kernel density estimation (KDE). The 59 dots in the ridge plot represent the endpoint-wise performance of the corresponding method on the 59 endpoints. The more concentrated their distribution and the smaller their standard deviation, the more balanced the model's performance on all endpoints. **c** The proportion of different models in performance rankings on all 59 toxic endpoints. **d** The Friedman and Nemenyi test with the critical difference (CD) for all models. The CD diagrams illustrate the average performance ranking of each model on 59 endpoints, calculated based on R^2 and RMSE. The length of the horizontal thick line segments is shorter than the CD value, indicating that the differences between the two models covered by these thick line segments are not significant. **e** The heatmap of endpoint-wise performance achieved by all models. All endpoints were arranged from left to right in ascending order of their sample sizes of toxicity measurements. Source data are provided as a Source Data file.

difficult to obtain, resulting in severe data scarcity for these target endpoints. This challenge is frequently overlooked in existing studies, leading to poor predictive performance for human-related endpoints. Next, we specifically assess ToxACoL's competitiveness in addressing data-scarce endpoints, with a particular focus on human-related endpoints. To explicitly distinguish between small/large-sized endpoints, we quantitatively classified all endpoints, where endpoints with less than 200 toxicity measurements were considered small-sized endpoints, and those with more than 1000 measurements were treated as large-sized endpoints. In total, there are 21 small-sized endpoints and 11 large-sized endpoints in the 59-endpoint dataset from TOXRIC. The three human-related endpoints, human-oral-TDLo, women-oral-TDLo, and man-oral-TDLo, are typical small-sized endpoints, with only 140, 156, and 163 available toxicity measurement samples, respectively. In addition, the assessment indicator inside the three endpoints is TDLo, which only appears in them, while the other 56 animal endpoints do not contain it. Therefore, the three human endpoints have significant semantic and biological gaps with other endpoints, making it difficult to effectively transfer knowledge learned from animal endpoints to the three human endpoints, which is why other baseline models perform poorly on the human endpoints.

We first compared the average R^2 of these models after 5-fold cross-validation on the human-related endpoints (Fig. 3a, c, and e). ToxACoL brought significant performance improvements in the three human endpoints. The R^2 achieved by ToxACoL on the three endpoints are 0.50, 0.43, and 0.40, respectively,

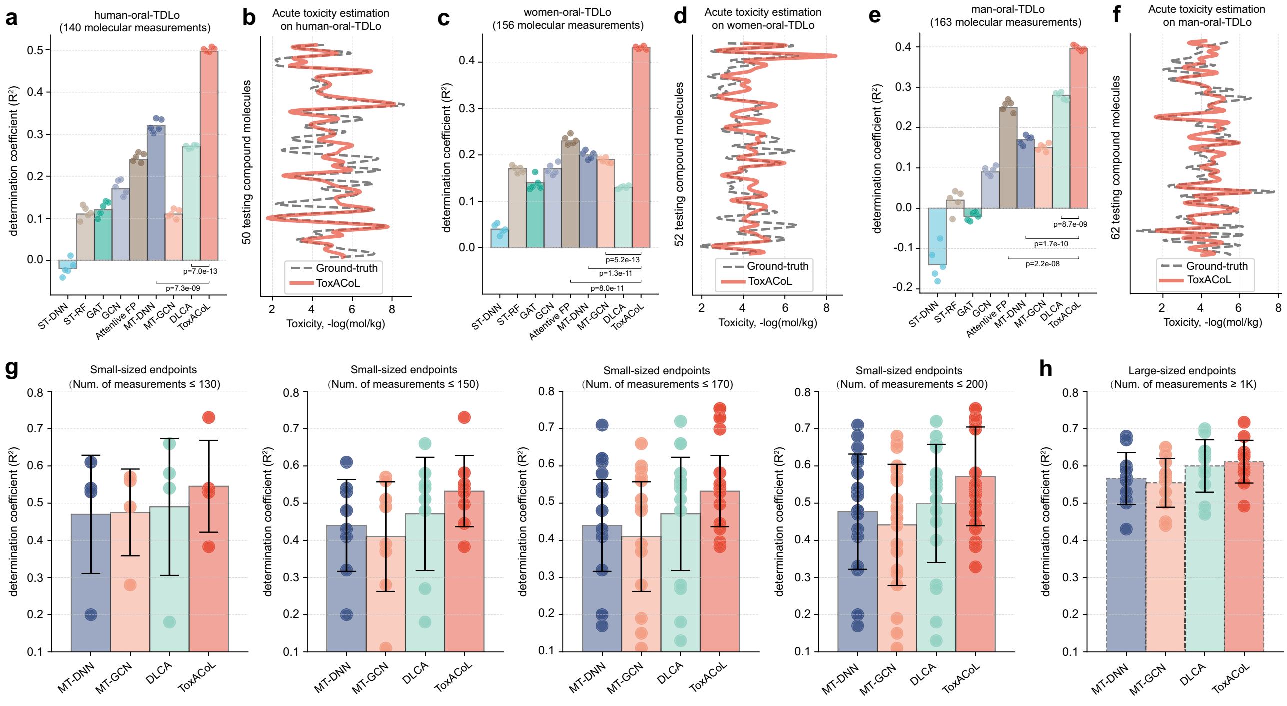


Fig. 3 | The performance comparison between ToxACoL and baseline models on small-sized acute toxic endpoints. a, **c**, and **e** Average R^2 of different models over 5-fold cross-validation on human-oral-TDLo, women-oral-TDLo, and man-oral-TDLo. Their significant differences were analyzed on the basis of a two-sided Student *t*-test. **b**, **d**, and **f** Acute toxicity estimation curves of ToxACoL for testing compounds at three human-related endpoints. Here, ToxACoL was trained using four folds of the whole toxicity dataset, and the testing compounds are all from the remaining one test fold. **g** Comparison between ToxACoL and advanced baseline methods on more small-sized endpoints. Taking the first subgraph for example, it considered the 4 endpoints ($n = 4$) with sample size of measurements <130, and so on for the following three subgraphs. The dots on the bar represent R^2 values at single endpoints, and the bar with the error bar denotes the mean R^2 value with standard deviation over the n small-sized endpoints (from left to right, $n = 4$, 8, 14, 21, respectively). **h** Comparison between ToxACoL and advanced baseline methods on 11 large-sized endpoints ($n = 11$). The bar with the error bar represents the mean R^2 value with standard deviation over the 11 large-sized endpoints. Source data are provided as a Source Data file.

outperforming the previous state-of-the-art results by a large margin (56% for human-oral-TDLo, 87% for women-oral-TDLo, and 43% for man-oral-TDLo). We selected one of the 5-fold cross-validation experiments to directly present the acute toxicity estimation results of ToxACoL on the test fold (Fig. 3b, d, and f). The toxicity intensity values predicted by ToxACoL can maintain a good consistency with the ground-truth toxicity intensity of the test compounds, and for some certain compounds, ToxACoL predicted them accurately, despite the scarce training measurement samples and the isolated endpoint attributes for the three human endpoints.

The performance of ToxACoL on more small-sized endpoints was further examined (Fig. 3g, Supplementary Fig. 2a). We set different thresholds for the sample size of toxicity measurements to observe the average R^2 of different models on all the small-sized endpoints whose sample size of measurements is less than the threshold. The improvement from ToxACoL on these small-sized endpoints is universal and significant, leading by nearly 15% on the average R^2 over 21 small-sized endpoints. Moreover, the R^2 achieved by ToxACoL is more balanced over all small-sized endpoints (with a more concentrated dot distribution and smaller standard deviation on histograms). We also compared the average R^2 on the 11 large-sized endpoints and found that ToxACoL had little performance gaps to the other advanced baseline models (Fig. 3h, Supplementary Table 5, Supplementary Fig. 2b), as the large sample size of toxicity measurements was sufficient to saturate all models. This further indicates that the advantages of our adjoint correlation learning mechanism over other baseline models mainly focus on small-sized acute toxic endpoints.

ToxACoL exhibits strong potential in reducing required training data for data-scarce endpoints

Collecting acute toxicology measurements via in vivo animal testing is time-consuming, labor-intensive, and expensive[12], and its sample size will not be too large due to the 3Rs principle, especially for data-scarce experimental subjects such as humans. Therefore, reducing the number of training measurements required for AI models on these small-sized endpoints is of great significance. This section mainly focuses on the acute toxicity estimation performance of ToxACoL on fewer training measurements for data-scarce endpoints.

In this test, the quantity of training measurements for the 21 small-sized endpoints was randomly reduced at one certain reduction ratio, while their test sets remained unchanged. We evaluated six different reduction schemes, which reduced the sample size of training measurement of small-sized endpoints to 80%, 50%, 40%, 30%, 20%, and 10% of their original size, respectively. Correspondingly, we retrained and evaluated our ToxACoL six times based on these reduced training datasets. Meanwhile, we also evaluated the performance of other state-of-the-art methods when the sample size of the small-sized endpoints was reduced to 30%.

An exciting experimental conclusion is that compared to other methods, ToxACoL only needs to use 20–30% of their training measurements on small-sized endpoints to match the previously optimal performance for the state-of-the-art methods (Fig. 4, Supplementary Tables 6 and 7, Supplementary Fig. 3), showing its strong potential in reducing animal toxicity testing. Taking the human-oral-TDLo for example (Fig. 4a), the original sample size of training measurements averaged over 5-fold cross-validation was 95, and ToxACoL achieved an R^2 of 0.38 with only 30% of the training measurements (≈29

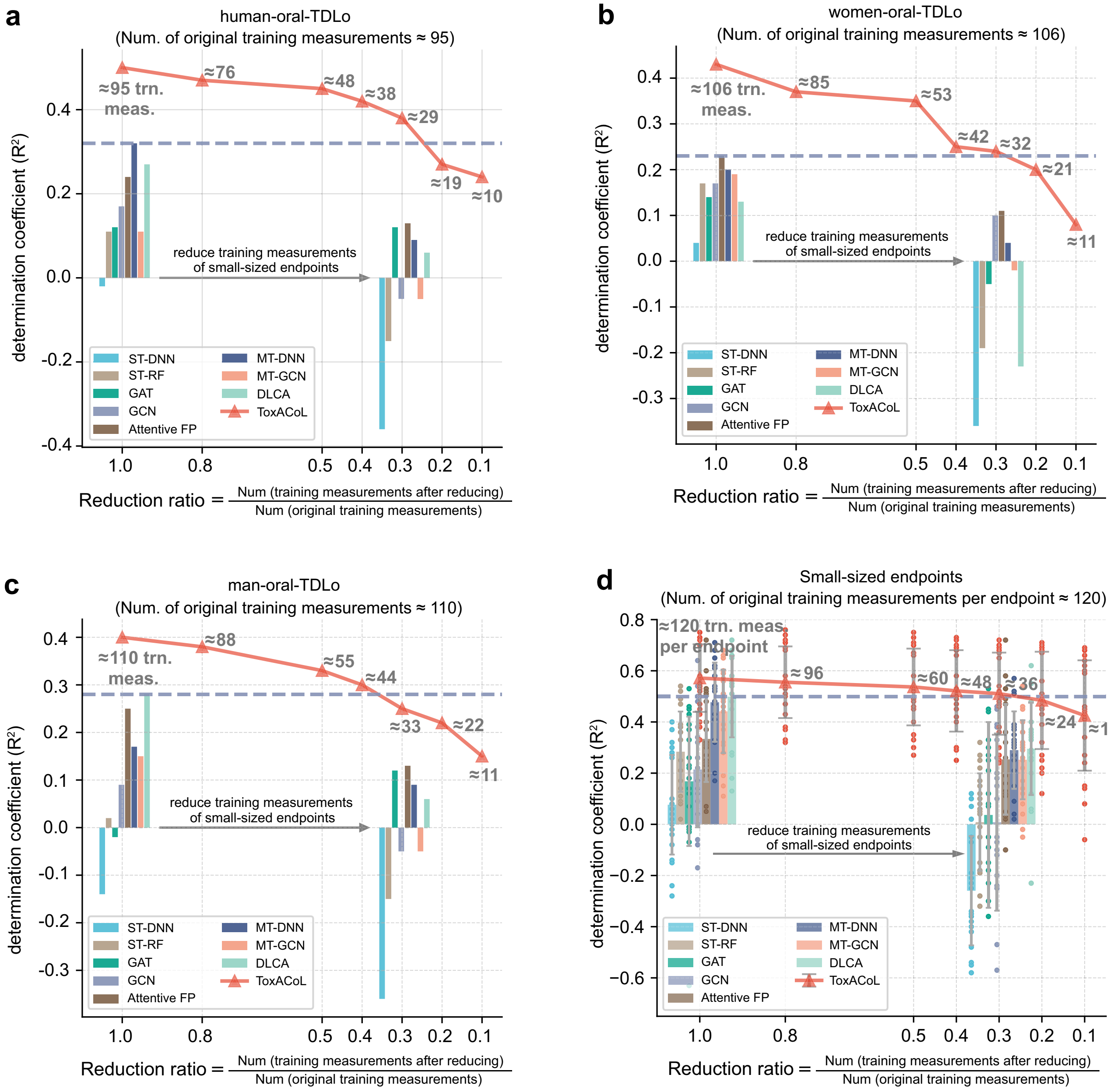


Fig. 4 | The acute toxicity evaluation performance of different methods with reducing training measurements of small-sized endpoints. a–c The performance at the three human-related endpoints, including human-oral-TDLo, women-oral-TDLo, and man-oral-TDLo, with the toxicity measurement samples used for training reduced proportionally. **d** The performance over all the 21 small-sized endpoints ($n = 21$) as the toxicity measurement samples used for training reduced proportionally, where the bar with the error bar represents the mean R^2 value with standard deviation over the 21 small-sized endpoints. Source data are provided as a Source Data file.

measurements), surpassing the previous best R^2 of 0.32 achieved by MT-DNN. Even with only 10% of the training measurements (≈10 measurements), ToxACoL can significantly outperform baseline models such as ST-DNN, ST-RF, GAT, GCN, and MT-GCN. Similar conclusions can also be observed for women-oral-TDLo (Fig. 4b) and man-oral-TDLo (Fig. 4c). At these two endpoints, ToxACoL only requires 30% (≈32 measurements) and 40% of the training samples (≈44 measurements), respectively, to surpass previously best-performing baseline models.

Statistically, we investigated the average R^2 on 21 small-sized endpoints concerning training measurement reduction (Fig. 4d). The original sample size of training data per endpoint was about 120, and ToxACoL achieved an average R^2 of 0.51 with only 30% of the training samples (≈36 measurements per endpoint), surpassing the previously best average R^2 of 0.50 achieved by DLCA. With only 10% of the training measurements (≈12 measurements per endpoint), ToxACoL also reached an average R^2 of 0.43, greatly surpassing most of the state-of-the-art methods (ST-DNN: 0.08, ST-RF: 0.28, GAT: 0.17, GCN: 0.21, Attentive FP: 0.33) and highly competitive with MTL or consensus models (MT-DNN: 0.48, MT-GCN: 0.44, DLCA: 0.50). We can also observe that when the training samples of the small-sized endpoints were reduced to 30%, the performance of other comparison methods significantly declined and falls far behind our ToxACoL. These results highlight the prospect of ToxACoL in few-shot learning and its robustness to the reduction of toxicity measurements at data-scarce endpoints. In addition, the reduction of the training measurements for data-scarce endpoints does not weaken the performance of ToxACoL on large-sized endpoints (Supplementary Fig. 4).

Table 1 | The average performance and standard deviation of various methods on the new 115-endpoint acute toxicity dataset via 5-fold cross-validation

Methods	All 115 endpoints		68 Small-sized endpoints		11 Human endpoints	
	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓
ST-DNN	0.11 ± 0.47	1.10 ± 0.38	−0.13 ± 0.45	1.28 ± 0.34	−0.50 ± 0.24	1.70 ± 0.32
ST-RF	0.21 ± 0.42	1.01 ± 0.35	0.01 ± 0.42	1.16 ± 0.32	−0.39 ± 0.25	1.60 ± 0.30
GAT	0.08 ± 0.32	0.94 ± 0.26	−0.01 ± 0.38	1.00 ± 0.30	−0.06 ± 0.24	1.23 ± 0.26
GCN	0.12 ± 0.43	0.91 ± 0.29	−0.04 ± 0.48	1.00 ± 0.32	−0.01 ± 0.19	1.19 ± 0.21
Attentive FP	0.32 ± 0.22	0.82 ± 0.24	0.26 ± 0.25	0.87 ± 0.27	0.24 ± 0.12	1.06 ± 0.24
MT-DNN	0.36 ± 0.37	0.77 ± 0.32	0.22 ± 0.40	0.88 ± 0.34	−0.22 ± 0.23	1.34 ± 0.30
MT-GCN	0.34 ± 0.35	0.80 ± 0.32	0.20 ± 0.38	0.91 ± 0.33	−0.25 ± 0.20	1.34 ± 0.31
DLCA	0.39 ± 0.32	0.76 ± 0.29	0.28 ± 0.35	0.86 ± 0.32	−0.15 ± 0.18	1.28 ± 0.29
ToxACoL	**0.51 ± 0.24**	**0.68 ± 0.23**	**0.44 ± 0.28**	**0.73 ± 0.26**	**0.28 ± 0.24**	**1.02 ± 0.26**

A small-sized endpoint contains no more than 200 available toxicity measurement data. An upward arrow next to each metric indicates that higher values represent better performance, while the upward arrow indicates the opposite. Best performance among all methods for each metric is shown in bold.

ToxACoL demonstrates robust performance in additional benchmark datasets

Next, to assess the robustness of ToxACoL's prediction, we conducted comparative experiments against other methods on two additional benchmark datasets. First, to explore the performance of the models on a broader range of acute toxicity endpoints, we collected and constructed a brand-new 115-endpoint acute toxicity dataset based on PubChem database[40]. Compared with the previous 59-endpoint acute toxicity dataset, this new dataset is more challenging for all current acute toxicity prediction models, since it has a more abundant and comprehensive number of toxicity endpoints, a larger number of small-sized endpoints with fewer available measurements, and more severe data imbalance and sparsity rate (see the "Methods" section). Similarly, the random 5-fold cross-validation was adopted.

On this more challenging dataset, the advantage of ToxACoL compared to other state-of-the-art methods is more significant (Table 1, Supplementary Figs. 5 and 6). First, the average R^2 of ToxACoL across all 115 endpoints is ~31% higher than that of DLCA (ranked second) with a lower standard deviation, indicating more balanced performance across all endpoint tasks (Supplementary Fig. 5c). Importantly, the average R^2 of ToxACoL across 68 small-sized endpoints is ~57% higher than that of the second-best DLCA, and it also outperforms all other methods on the 11 human-related endpoints. In addition, judging from the performance rankings of all methods across the 115 endpoints (Supplementary Fig. 5b, d), ToxACoL ranks first in terms of performance at the vast majority of endpoints. These results once again strongly demonstrate the superiority of ToxACoL and its powerful few-shot learning ability for acute toxicity prediction.

In addition, the U.S. Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) Acute Toxicity Workgroup developed a five-endpoint prediction project. The five acute toxicity-related endpoints include LD_{50} value, U.S. Environmental Protection Agency (U.S. EPA) hazard (four) categories, Globally Harmonized System for Classification and Labeling (GHS) hazard (five) categories, very toxic chemicals (LD_{50} ≤50 mg/kg), and nontoxic chemicals (LD_{50} ≥ 2000 mg/kg). We next test ToxACoL with the representative models on the reported LD_{50} datasets.

For the five endpoints, Mansouri et al.[5] proposed a consensus modeling framework, Collaborative Acute Toxicity Modeling Suite (CATMoS), which consolidates predictive models from all participating teams, effectively leveraging the strengths of each model. To address conflicts in results across these five endpoints, the authors proposed a WoE approach with weighted integration to generate consensus predictions. Specifically, for the LD50 endpoint (rat-oral-LD50), the WoE method refined the original predictions, significantly enhancing performance. To demonstrate the superiority of ToxACoL, we compared it with CATMoS and the improved WoE method on the LD50 dataset using the same data partitioning strategy (6398 molecules for training and 2196 molecules for evaluation). The results show that ToxACoL achieved performance improvements of 8.24% and 21.54% in R^2 metrics on the training and evaluation sets, respectively, compared to the WoE optimization proposed by Mansouri et al. (Table 2). This breakthrough highlights the significant advantages of ToxACoL in terms of prediction accuracy and generalization capability.

Table 2 | The performance comparison between ToxACoL and representative consensus models on the LD50 endpoint dataset containing 6398 molecules for training and 2196 molecules for evaluation (WoE: weight of evidence)

Methods	Training		Evaluation	
	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓
Step 1: End point (Mansouri et al.[5])	0.83	0.31	0.67	0.47
Step 2: WoE (Mansouri et al.[5])	0.85	0.30	0.65	0.49
ToxACoL	**0.92**	**0.25**	**0.79**	**0.41**

Best performance among all methods for each metric is shown in bold.

ToxACoL derives clear compound representation manifolds to various toxic endpoints

To explore the effectiveness of the molecular representations learned by ToxACoL, we visualized the top-level embeddings generated by ToxACoL in the latent space. Specifically, we randomly selected one of the 5-fold cross-validation experiments and applied the well-trained ToxACoL via the 59-endpoint dataset from TOXRIC to the test fold to extract the top-level embeddings for the testing compounds. The t-SNE algorithm[50] is used to visualize these embeddings, showcasing their local relationships and distribution patterns in low-dimensional space. The clear clustering manifolds concerning toxicity intensity were shown across different acute toxicity endpoints including both large- and small-sized endpoints (Fig. 5a). For large-sized endpoints, the model learns excellent molecular representations due to sufficient training data. Despite the limited data for small-sized endpoints, their representations still form distinct clusters and clear manifolds, demonstrating the strong predictive and representation learning capabilities of ToxACoL.

We selected data-scarce experimental species, such as cats, dogs, and guinea pigs, to perform joint embedding visualizations under different conditions within the same species (Fig. 5b). These compound embeddings of the same species can still be distributed in a well-defined manifold, forming similar clustering patterns and distributions regarding varying toxicity intensities, although they are tested under different conditions. This

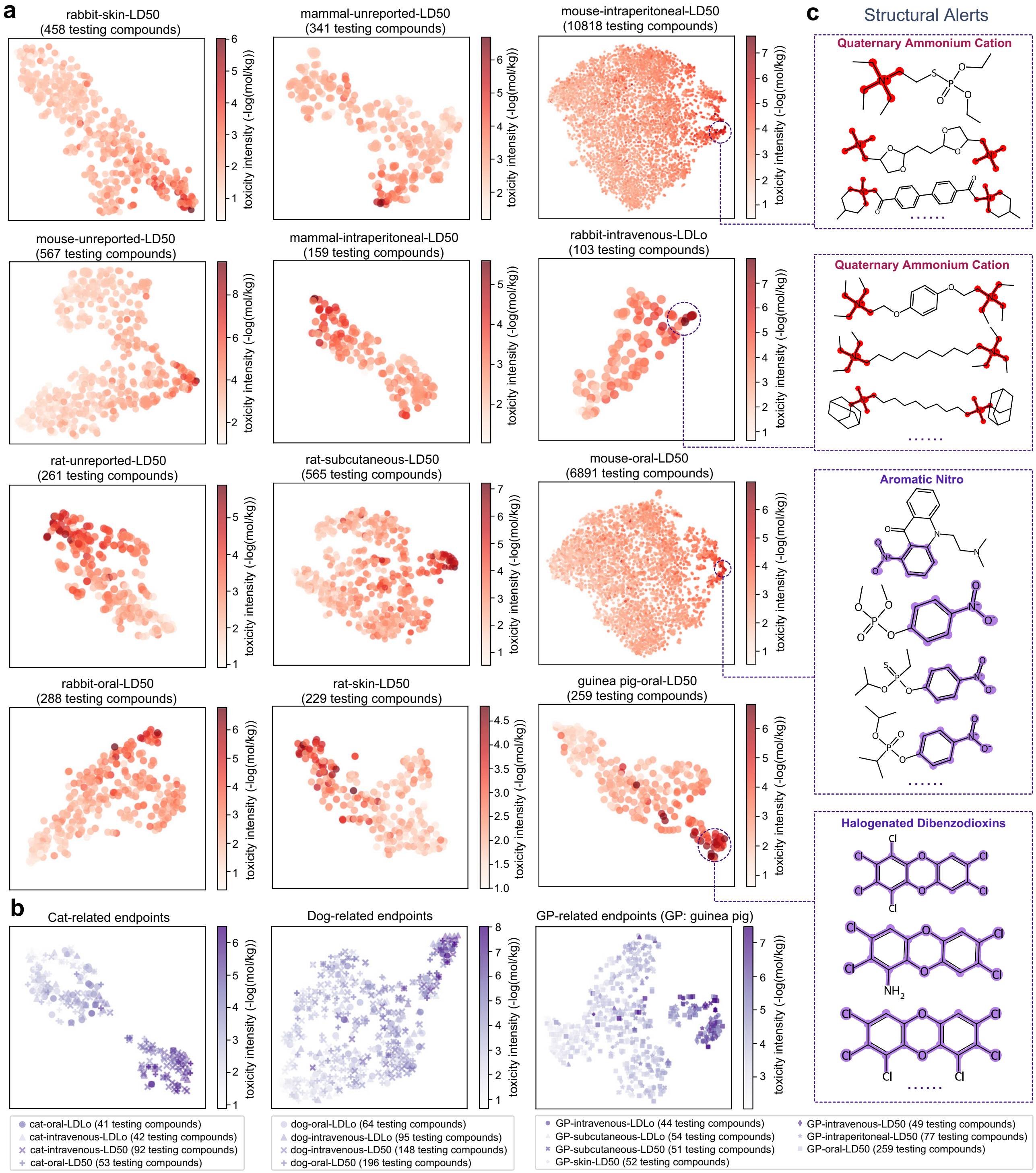


Fig. 5 | Analysis of model interpretability and practicality. a and **b** The t-SNE visualization of top-level embeddings learned by ToxACoL concerning various acute toxic endpoints. Here, ToxACoL was trained using four-fold data of the 59-endpoint acute toxicity dataset, and the displayed compounds are all from the remaining test fold. The darker the dot's color, the greater its toxicity intensity. Visualization of single endpoints (**a**) and multiple endpoints belonging to the same species (**b**). **c** Several compound examples of structural alerts discovered by ToxACoL from the high-toxicity clusters at different endpoints, including quaternary ammonium cation, aromatic nitro, and halogenated dibenzodioxin, which have been highlighted by masks.

phenomenon also conforms to our toxicology priors, that is, for a certain type of compound that has strong toxicity to the same animal, regardless of the administration route and measurement indicator, its toxicity results are generally relatively strong. This demonstrates that ToxACoL can indeed capture potential dependency relationships between endpoints and guide the learning of more scientific compound representations.

Latent space clustering patterns predicted by ToxACoL can reveal the structural alerts of high-toxicity compounds

To further investigate the structural patterns of compounds with high acute toxicity and uncover their potential mechanisms, we performed an in-depth analysis of the embedding manifold's morphology. It is well-established that a compound's structure is closely linked to its biological activity and toxicity. Certain highly reactive molecular fragments, which can be presented in the parent compound or its metabolites after biochemical transformation by human enzymes (bioactivation), have been confirmed to possess strong toxic potential[51,52]. The clustering within ToxACoL's latent space embeddings could potentially reveal key molecular fragments or functional groups shared by highly toxic compounds, which is critical for understanding the mechanisms behind acute toxicity.

For the testing compounds, we focused on clusters representing highly toxic compounds and analyzed the molecular characteristics within the same cluster in all the endpoints (Fig. 5a). The results showed that compounds in the same region typically shared molecular fragments, which may play a dominant role in influencing acute toxicity intensity. These molecular fragments are commonly referred to as structural alerts. Extensive studies show that structural alerts are effective markers for toxicity assessment and mechanistic understanding across diverse endpoints, with their presence indicating high toxic potential in compounds[53–56]. In high-toxicity clusters across multiple endpoints, we identified well-known structural alerts (Fig. 5c).

For example, in the endpoints such as mouse-intraperitoneal-LD_{50}, mouse-subcutaneous-LD_{50}, and mouse-oral-LD_{50}, we identified that ~10–25 compounds within the high-toxicity clusters contain quaternary ammonium cations, which are considered typical structural alerts. The structure of quaternary ammonium cation includes one hydrophobic hydrocarbon chain linked to a positively charged nitrogen atom, along with other alkyl groups, predominantly consisting of short-chain substituents such as methyl or benzyl groups[57]. Long-term exposure to low doses of quaternary ammonium compounds may lead to acute toxicity in aquatic organisms or humans[58–60]. In addition to the aforementioned endpoints, a considerable proportion of compounds containing quaternary ammonium cation structures, ranging from approximately 10% to 30%, was also observed in endpoints with fewer high-toxicity molecules, such as rabbit-intravenous-LDLo, mouse-oral-LDLo, and rabbit-subcutaneous-LDLo. Notably, we found that in nearly all mouse-related endpoints, as well as other mammal-related endpoints, compounds containing quaternary ammonium cation structures account for a significant proportion of high-toxicity clusters. In contrast, such molecules were nearly absent in avian species, including birds, chickens, and quails. This discrepancy may be attributed to interspecies differences in physiological structures, metabolic capacities, and excretion mechanisms.

In various mammal- and avian-related toxicity endpoints, such as mouse-oral-LD_{50}, rat-oral-LD_{50}, dog-intravenous-LDLo, and chicken-oral-LD_{50}, compounds containing nitro groups and aromatic nitro structures were found to have a high prevalence in high-toxicity clusters. It is reported that Nitro-containing compounds exhibit a broad range of toxic effects, including acute toxicity, carcinogenesis[61], and systemic impacts on various biological systems such as the reproductive, immune, nervous[62], digestive, respiratory, cardiovascular, as well as specific organs like the liver[63,64], kidney, and stomach[65]. Nitroaromatic compounds are widely used as pesticides, explosives, pharmaceuticals, and chemical intermediates in industrial synthesis[66]. They are potent toxic or carcinogenic compounds, presenting a considerable danger to the human population[66–68].

Additionally, halogenated dibenzodioxins are identified as a structural alert in the guinea pig-oral-LD_{50} endpoint. Halogenated dibenzodioxins, as a subclass of dioxins, are recognized as persistent environmental pollutants characterized by extreme toxicity[69]. Their primary emission sources include waste incineration facilities and industrial combustion processes[70]. Although the detection frequency of this structure across multiple toxicity endpoints is lower than the structural alerts discussed earlier, this discrepancy may stem from the limited samples of environmental contaminants in the current dataset. Notably, the ToxACoL successfully identified the high toxic potential of this structure through representation learning, demonstrating its capability to uncover latent hazardous properties even in sparsely populated data domains.

In summary, we believe that the latent space clustering patterns predicted by ToxACoL can help identify structural alerts responsible for acute toxicity, facilitating the analysis of structural differences across species. This offers valuable insights for further exploration into the underlying mechanisms of acute toxicity.

Exploring the extrapolation patterns from animals to humans

The preceding discussion illustrates the similarities and differences in the distribution of acute toxicity values, latent space representations, and alert structures among endpoints of different species. This suggests that interspecies extrapolation may follow specific patterns in both toxicity values and latent space distributions[71]. Furthermore, to more thoroughly assess the toxicity effects of compounds on data-scarce species such as humans, this section delves into the extrapolation patterns from experimental species to humans. We used all the acute toxicity data from TOXRIC to train the ToxACoL model and took three human-related endpoints and the compounds associated with these endpoints as the research references. In TOXRIC, the number of available compounds for the human-oral-TDLo, woman-oral-TDLo, and man-oral-TDLo endpoints were 140, 156, and 163, respectively. The trained ToxACoL was applied to predict and fill in the missing acute toxicity values for these selected reference compounds across all endpoints. Based on these filled-in toxicity values and previously existing toxicity values, we explored the toxicity responses of compounds across different species and the extrapolation patterns.

We applied PCC values to quantify the proximity of the toxicity values of all experimental species-related endpoints to those of humans (Fig. 6a and Supplementary Figs. 7–9). The endpoints with the highest PCC values associated with human-oral-TDLo are the other two human-related endpoints: woman-oral-TDLo (0.912) and man-oral-TDLo (0.903) (Fig. 6a). Similar results were observed for the other two human endpoints, which aligns with theoretical expectations and validates the reliability of the model's predictions. Beyond human endpoints, the cat-intravenous-LDLo, frog-subcutaneous-LDLo, and mouse-intravenous-LDLo showed the highest PCC values among all animal endpoints when compared to the three human endpoints. This demonstrates that these three animal endpoints have a stronger linear correlation with human endpoints, making them more suitable for transferring and extrapolating to human oral toxicity responses, indicating potentially robust extrapolation models. Additionally, we extracted the latent space representations for cat-intravenous-LDLo, human-oral-TDLO, woman-oral-TDLo, and man-oral-TDLo to examine their latent space patterns (Fig. 6b). The data distributions and trends in toxicity intensity for the human endpoints and cat-intravenous-LDLo are consistent, with highly toxic compounds being closely situated in the embedding space, further demonstrating the potential of extrapolating cat-intravenous-LDLo results to human endpoints.

Other mammalian species such as guinea pig, rabbit, and dog also show relatively high PCC values with human endpoints. In contrast, avian species like duck, chicken, and quail exhibit results that are significantly different from the human endpoints. Regarding the administration route, intravenous measurements are more concordant with human oral results, whereas other administration routes show less consistency with the target endpoints. Exploring the reasons for this phenomenon, intravenous delivers drugs directly into the bloodstream, allowing them to quickly reach effective concentrations while avoiding digestion and metabolism processes. This may make intravenous data better reflect how oral medications ultimately work in the human body. We also highlight two endpoints with poor consistency to human endpoints, quail-oral-LD_{50} and rat-subcutaneous-LD_{50} (Fig. 6a), which may be attributed to differences in species, administration routes, and measurement indicators.

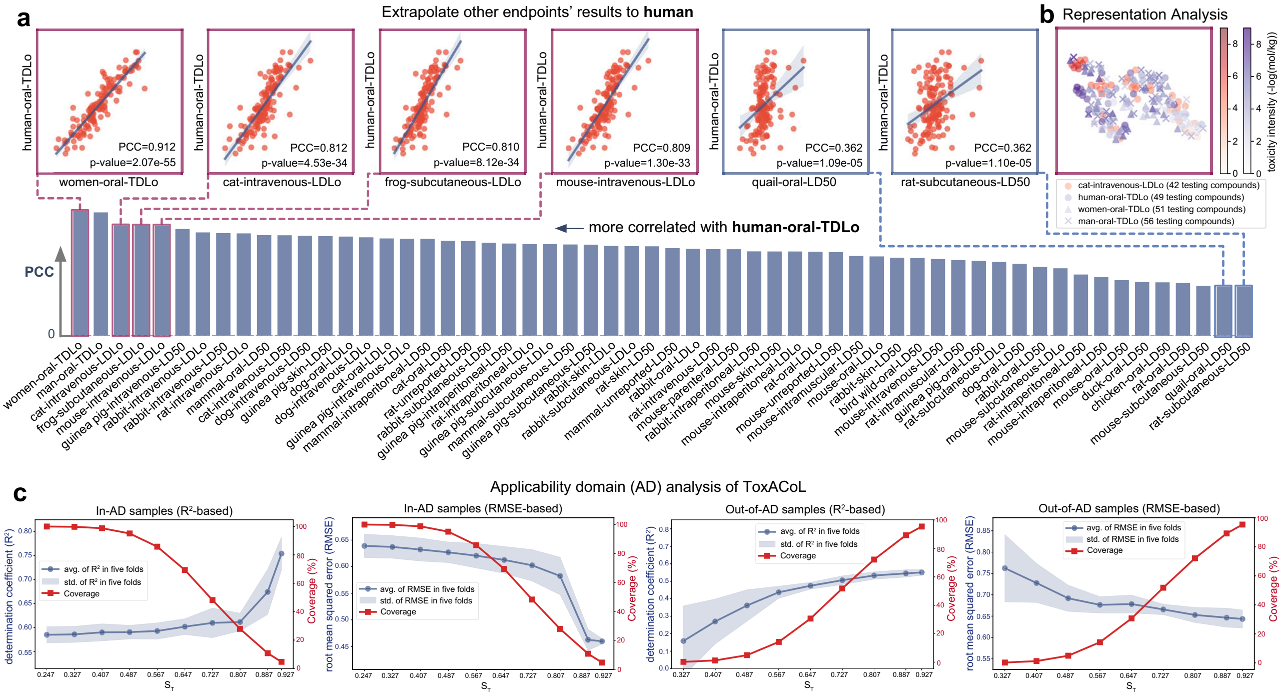


Fig. 6 | Extrapolation pattern analysis and applicability domain analysis. **a** Pearson correlation coefficient (PCC) values between human-oral-TDLo and the remaining 58 endpoints. Note that there are a total of 140 compounds in the dataset that have available toxicity measurement values at human-oral-TDLo endpoint. The missing toxicity intensity values of these 140 compounds at the other 58 endpoints were filled in by the predicted intensity values of ToxACoL. Thus, the PCC value between the two endpoints was calculated based on the two groups of toxicity intensity values of the 140 compounds concerning the two endpoints. The Pearson correlation analysis is two-sided. The center line in the correlation plots represents the regressed line and the error band denotes the confidence interval of 0.95 for linear regression. **b** Latent space representation distribution for cat-intravenous-LDLo, human-oral-TDLo, woman-oral-TDLo, and man-oral-TDLo. Here, ToxACoL was trained using four-fold data of the whole acute toxicity dataset, and the displayed compounds are all from the remaining test fold. **c** Performance metrics (R^2, RMSE) of in-AD and out-of-AD samples under varying thresholds within the AD defined in this study, averaged across 59 endpoint tasks. The *X*-axis represents the AD threshold S_T corresponding to different *Z* parameters. The left *Y*-axis (blue lines) indicates metric values, while the right *Y*-axis (red lines) denotes the proportion of extracted samples relative to the total (Coverage). Blue lines and shaded areas represent the mean and standard deviation of five-fold cross-validation results. Source data are provided as a Source Data file.

Applicability domain analysis of ToxACoL

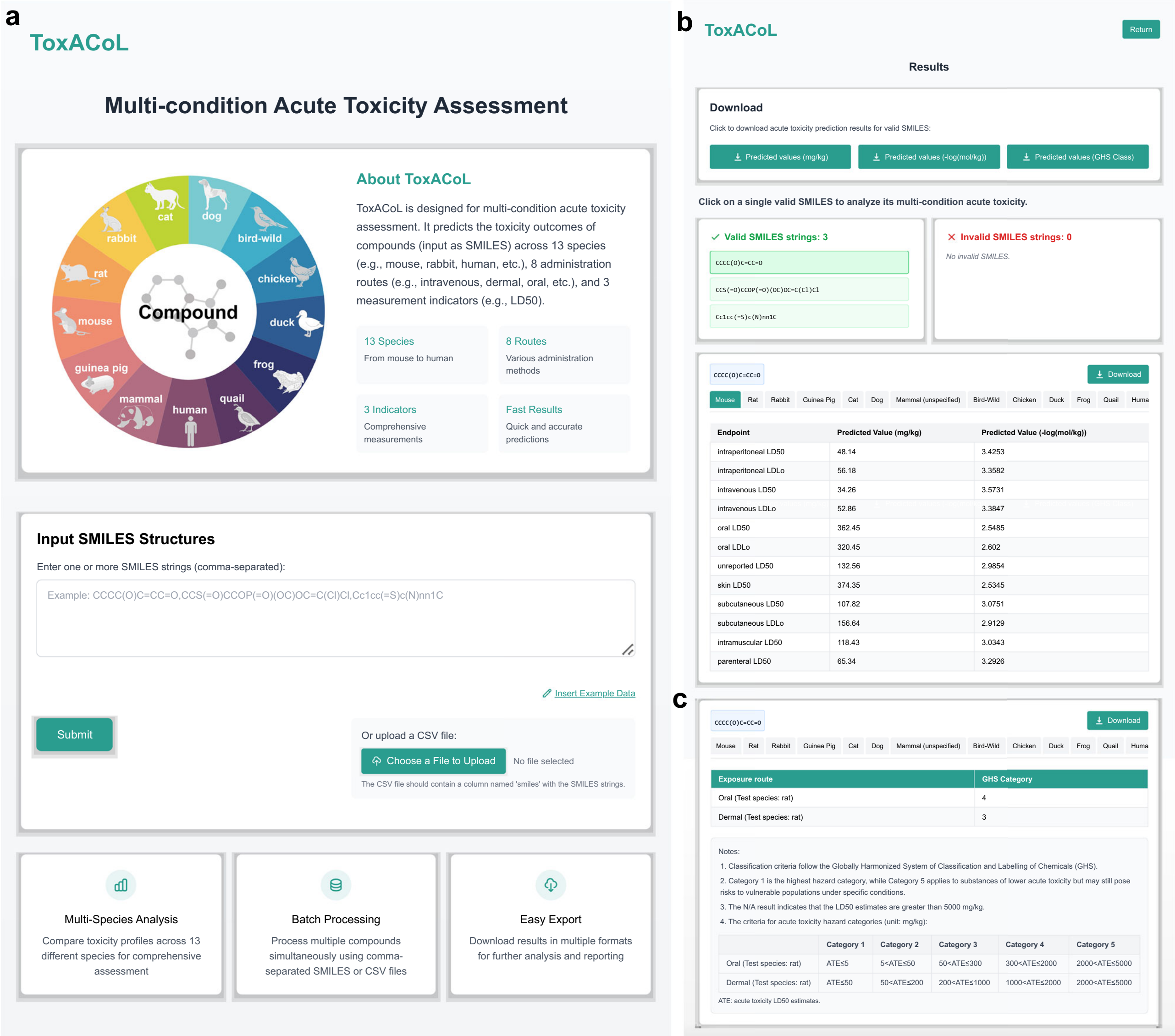
Based on Eq. (11) (see the "Methods" section), we defined the AD of ToxACoL to delineate the chemical space where the model achieves reliable predictions. We systematically evaluated the dynamic changes in prediction performance (R^2, RMSE) and coverage rates inside and outside the AD under varying parameters ($k = 10$, $Z \in [1, 9.5]$) (Fig. 6c). As the threshold increases, the coverage of in-AD samples decreases sharply from 99.96% ($Z = 1$, $S_T = 0.247$) to 4.62% ($Z = 9.5$, $S_T = 0.927$), while prediction accuracy improves significantly: R^2 rises from 0.58 to 0.75, and RMSE decreases from 0.64 to 0.46. This demonstrates that stricter AD thresholds effectively screen high-confidence predictions by narrowing the chemical space. Notably, the predictive performance for in-AD samples at all *Z*-values significantly outperforms the full-dataset test results, validating the necessity of AD demarcation for enhancing reliability. By balancing prediction accuracy and coverage, $Z = 6$ ($S_T = 0.647$) was selected as the globally optimal parameter, achieving 69.33% in-AD coverage, $R^2 = 0.60$, and RMSE = 0.61, thereby addressing both predictive robustness and practical applicability.

We further analyzed out-of-AD samples. The R^2 for out-of-AD samples increases from 0.16 to 0.55 with higher S_T thresholds (Fig. 6c). When out-of-AD molecular coverage reaches 51.82%, R^2 attains 0.51 and continues to rise, indicating ToxACoL's partial generalization capability for extrapolation beyond the AD. This phenomenon likely stems from ToxACoL's latent feature extraction capability enabled by deep representation learning, which may capture cross-chemical-space patterns to achieve reliable predictions in certain out-of-AD regions, offering useful perspectives for expanding the model's applicability boundaries.

ToxACoL online prediction web platform enables rapid assessment of multi-condition acute toxicity and GHS category

In order to enable all researchers to directly use our pre-trained ToxACoL, we have integrated ToxACoL into an online web platform (https://toxacol.bioinforai.tech/). Users can simply upload single or multiple molecular SMILES, and obtain predictions for various toxic endpoints and GHS classifications, with results downloadable in multiple formats. Notably, acute toxicity in practice is used to determine the potency of a substance (usually as GHS classes). Thus, the ToxACoL online web platform also provides the predicted GHS classification of chemicals as a toxicity endpoint. In alignment with GHS hazard classification criteria, ToxACoL converts predicted rat-oral-LD$_{50}$ and rat-skin-LD50 values into corresponding oral and dermal hazard categories. Among them, Category 1 is the highest hazard category, while Category 5 applies to substances of lower acute toxicity but may still pose risks to vulnerable populations under specific conditions. Full classification criteria are provided at ToxACoL online web.

Figure 7 illustrates the interfaces of the ToxACoL online prediction web platform. On the homepage (Fig. 7a), users can manually input or upload a CSV file containing single or multiple SMILES strings through the "Input SMILES Structures" module, followed by clicking the "Submit" button to initiate the analysis. Upon completion within seconds, the prediction results comprise two functional modules (Fig. 7b): the "Download" section for downloading and viewing all



Multi-condition Acute Toxicity Assessment

About ToxACoL

ToxACoL is designed for multi-condition acute toxicity assessment. It predicts the toxicity outcomes of compounds (input as SMILES) across 13 species (e.g., mouse, rabbit, human, etc.), 8 administration routes (e.g., intravenous, dermal, oral, etc.), and 3 measurement indicators (e.g., LD50).

13 Species
From mouse to human

8 Routes
Various administration methods

3 Indicators
Comprehensive measurements

Fast Results
Quick and accurate predictions

Input SMILES Structures

Enter one or more SMILES strings (comma-separated):

Example: CCCC(O)C=CC=O,CCS(=O)CCOP(=O)(OC)OC=C(Cl)Cl,Cc1cc(=S)c(N)nn1C

Insert Example Data

Or upload a CSV file:

No file selected

The CSV file should contain a column named 'smiles' with the SMILES strings.

Multi-Species Analysis
Compare toxicity profiles across 13 different species for comprehensive assessment

Batch Processing
Process multiple compounds simultaneously using comma-separated SMILES or CSV files

Easy Export
Download results in multiple formats for further analysis and reporting

Results

Download

Click to download acute toxicity prediction results for valid SMILES:

Click on a single valid SMILES to analyze its multi-condition acute toxicity.

✓ Valid SMILES strings: 3

```
CCCC(O)C=CC=O
CCS(=O)CCOP(=O)(OC)OC=C(Cl)Cl
Cc1cc(=S)c(N)nn1C
```

✕ Invalid SMILES strings: 0

No invalid SMILES.

`CCCC(O)C=CC=O`

Endpoint	Predicted Value (mg/kg)	Predicted Value (-log(mol/kg))
intraperitoneal LD50	48.14	3.4253
intraperitoneal LDLo	56.18	3.3582
intravenous LD50	34.26	3.5731
intravenous LDLo	52.86	3.3847
oral LD50	362.45	2.5485
oral LDLo	320.45	2.602
unreported LD50	132.56	2.9854
skin LD50	374.35	2.5345
subcutaneous LD50	107.82	3.0751
subcutaneous LDLo	156.64	2.9129
intramuscular LD50	118.43	3.0343
parenteral LD50	65.34	3.2926

`CCCC(O)C=CC=O`

Exposure route	GHS Category
Oral (Test species: rat)	4
Dermal (Test species: rat)	3

Notes:

1. Classification criteria follow the Globally Harmonized System of Classification and Labelling of Chemicals (GHS).
2. Category 1 is the highest hazard category, while Category 5 applies to substances of lower acute toxicity but may still pose risks to vulnerable populations under specific conditions.
3. The N/A result indicates that the LD50 estimates are greater than 5000 mg/kg.
4. The criteria for acute toxicity hazard categories (unit: mg/kg):

	Category 1	Category 2	Category 3	Category 4	Category 5
Oral (Test species: rat)	ATE≤5	5<ATE≤50	50<ATE≤300	300<ATE≤2000	2000<ATE≤5000
Dermal (Test species: rat)	ATE≤50	50<ATE≤200	200<ATE≤1000	1000<ATE≤2000	2000<ATE≤5000

ATE: acute toxicity LD50 estimates.

Fig. 7 | Interface of the ToxACoL online prediction web platform. a Homepage, where users can input molecular SMILES in the "Input SMILES Structures" section and click "Submit" to initiate the acute toxicity analysis. **b** The result page of acute toxicity prediction. **c** The result page of GHS classification. Credits: the icons of molecular structure and animals are sourced from https://creazilla.com/. Source data are provided as a Source Data file.

predictions, and the "Single SMILES Analysis" module for individual compound results. The "Download" module allows users to download comprehensive prediction results for all input SMILES in three formats: "Predicted values (mg/kg)" provides toxicity values in mg/kg units "Predicted values (−log(mg/kg))" offers dimensionless normalized scores, and "Predicted values (GHS class)" supplies GHS classification outcomes. Below this, the "Single SMILES Analysis" module enables users to click on any valid SMILES to view its detailed predictions across 59 toxicity endpoints and GHS classifications. Toxicity endpoints are classified by species. GHS classes are divided into two exposure routes, oral and dermal (Fig. 7c). A "Download" button within the individual SMILES result section facilitates the download of all predictions for the selected compound. We believe this platform can offer a new path for validation processes, in the hope of becoming a useful resource for regulatory applications.

Discussion

Acute toxicity assessment is commonly the initial step in evaluating the safety of compounds and plays a crucial role in determining whether a compound can proceed to chemical development and industrial application. The emergence of ML algorithms offers an excellent solution for implementing the 3Rs principle, providing an alternative to animal testing for acute toxicity. By leveraging existing animal experimental data, ML can predict acute toxicity across various conditions. However, the complexity of toxicity testing including spanning different species, administration routes, and measurement indicators, combined with imbalanced data, presents significant challenges, particularly in predicting toxicity for data-scarce species like humans.

In this study, we model the relationships between endpoints under various experimental conditions by constructing a multi-condition endpoint graph. An adjoint correlation learning paradigm is designed to simultaneously learn both sample representations and multi-condition label information. This embeds experimental condition data into the representation learning process, allowing the model to capture fine-grained toxicity patterns. Learning the relationships between species and conditions enables the model to transfer knowledge effectively when data is scarce. For example, although human experimental data is limited, the model can improve its prediction accuracy by learning the relationships between humans and other species like mice and rats. Beyond the focus of this study on acute toxicity prediction, the adjoint correlation learning framework also offers valuable insights for other multi-task learning problems with multi-label settings, rich label information, and significant data imbalance.

Based on the designed Adjoint Correlation Learning, we propose a multi-condition acute toxicity prediction model, ToxACoL. Before our work, the learning process of most QSAR models[72,73] employed for computational toxicity prediction was unidirectional from compound data to toxic endpoints. Taking the MTL paradigm as an example (Supplementary Fig. 10), it uses a deep network to extract high-level compound representations and a multi-channel linear regression layer to predict toxicity intensity at various endpoints. In ToxACoL, we propose synchronous bidirectional learning from compound data and toxic endpoints. We use graph topology to model endpoint associations and graph convolution for cross-correlation and develop the adjoint correlation mechanism to process compound and endpoint embeddings simultaneously, maintaining their interaction. Importantly, we use the top-level endpoint embeddings as toxicity regressor weights, which incorporate the learned relationships between endpoints into the final prediction step. Unlike the MTL models' independent regressors, the multi-endpoint toxicity regressors generated by ToxACoL are interdependent due to graph convolution, incorporating multi-endpoint biological priors. This enables parameter propagation, reducing training data reliance. Also, the top-level feed-forward layer's compound embeddings are endpoint-aware, facilitating a more task-focused compound representation learning. These designs are the key factors that enable the performance improvement of small-sized endpoints.

ToxACoL's effectiveness and application value are validated through various experimental scenarios. These scenarios include comprehensive multi-endpoint performance evaluation, performance improvement for scarce species endpoints, reduced training data testing, latent space manifold analysis, structural alerts clustering analysis, species extrapolation pattern exploration, AD analysis, and online prediction platform development. First, ToxACoL balanced performance across multi-condition endpoints. Due to the highly imbalanced training dataset, existing algorithms typically perform well on large-sized endpoints but poorly on small-sized ones. ToxACoL demonstrated robust handling of imbalanced multi-task datasets. Second, the performance improvement evaluation for scarce species endpoints focused on three small-sized human-related endpoints (100–200 samples), thus confirming its practical value. Results showed remarkable improvements over baselines, with R^2 performance increases of 56%, 87%, and 43% for the three endpoints, respectively. Third, in real-world applications, human and unconventional animal experimental data often do not exceed 100 samples. We further reduced training data for scarce species endpoints to observe model performance. Results indicated that ToxACoL requires only 20–30% of training data (~20–40 samples) to outperform state-of-the-art methods, demonstrating strong competitiveness in few-shot learning. Fourth, the embeddings learned by ToxACoL' revealed clear clustering manifolds based on toxicity intensity, consistent across both large and small endpoints. Fifth, further analysis of ToxACoL's learned representations revealed that compounds in the same clusters shared common substructures. High-toxicity clusters contained well-known structural alerts, suggesting that ToxACoL can assist in identifying unknown structural alerts while predicting acute toxicity for unseen compounds or endpoints, thus aiding the exploration of toxicity mechanisms. Sixth, since it is difficult to obtain experimental data for scarce species like humans, we explored interspecies extrapolation patterns by completing toxicity values across multiple endpoints. The results showed that toxicity values in cats were closer to humans than those of other species. Additionally, intravenous administration in animal experiments showed closer results to human oral than other routes. These findings may offer valuable insights for toxicity extrapolation research. Seventh, AD analysis of ToxACoL provides a detailed analysis of in-AD and out-of-AD samples and depicts the chemical space where the model achieves reliable predictions. Finally, for the convenience of researchers in using ToxACoL, we have developed an online web platform. Users can simply upload single or multiple molecular SMILES, and obtain predictions for various toxic endpoints and GHS classifications.

In future work, ToxACoL should be extended to accommodate a broader spectrum of acute toxicity tasks even other chemical-related tasks. First, Mansouri et al.[5] proposed that regulatory agencies generally require three types of acute toxicity outcomes: (1) an LD_{50} value estimate, (2) a binary outcome based on a single threshold, and (3) a multiclass scheme based on different thresholds. In this study, we have developed predictive models for LD_{50} estimation across multiple species and under various administration routes. Based on the LD_{50} estimates for oral and dermal routes, we applied GHS classification criteria to achieve GHS class-based multiclass prediction. Both prediction results are accessible on the ToxACoL online platform. While the other two binary classification scenarios should be supported, including the prediction of whether it is very toxic or nontoxic. This classification requirement can also be addressed by setting appropriate thresholds on the predictions generated by ToxACoL. Additionally, Mansouri et al. have provided ML-ready datasets for LD_{50} estimation, binary classification, and multiclass classification based on GHS and US EPA classification criteria. These datasets can be used to train models separately and then combined through weighted averaging to produce integrated prediction results, thereby improving the accuracy of chemical hazard predictions.

Second, ToxACoL has the potential to be adapted for other chemical assessment tasks beyond acute toxicity. ToxACoL is specifically designed for predictive tasks involving multi-condition labels and heterogeneous sample types. Its innovative dual-path learning paradigm, adjoint correlation learning, combined with multi-condition endpoint graph modeling capabilities, renders it particularly suitable for complex MTL architectures. Additionally, the modular design of ToxACoL demonstrates significant transfer learning potential across diverse chemical assessment scenarios. For instance, in multi-objective optimization for drug discovery, the dual-path interaction architecture can be extended to incorporate optimization targets like target inhibition potency, therapeutic indication prioritization, and pathway activation modulation as multi-condition endpoints. By constructing multi-modal endpoint graphs comprising drug–target–disease-pathway nodes, ToxACoL will systematically encode efficacy–endpoint correlations, thereby providing a computational framework to accelerate lead compound optimization. This adaptability highlights ToxACoL's capacity to transcend traditional toxicity prediction paradigms and fosters cross-domain applications in computational chemistry.

Third, these multi-endpoint-shared, high-weight structural alerts identified by ToxACoL enable early-stage toxicity flagging during virtual screening. By automatically filtering candidates containing known alerts, ToxACoL may accelerate hit-to-lead pipelines while minimizing downstream attrition risks. In future work, ToxACoL's multi-endpoint toxicity predictions can be integrated as penalty terms in optimization objectives of molecule generation, prioritizing structures with alert replacement.

Methods

Acute toxicity data

This work involves two large-scale acute toxicity datasets (59-endpoint dataset, 115-endpoint dataset) and a benchmark dataset from Mansouri et al.[5]. Unlike typical toxicity classification datasets, these acute toxicity datasets are more complex as they focus on fitting acute toxicity intensity values.

The 59-endpoint dataset is sourced from the TOXRIC database[25], a public and standardized toxicology database for compound discovery.

This dataset includes 59 various toxicity endpoints with 80,081 unique compounds represented using SMILES strings, and 122,594 usable toxicity measurements described by continuous values with a unified toxicity chemical unit: −log(mol/kg). The larger the measurement value, the stronger the toxicity intensity of the corresponding compound towards a certain endpoint. The proportion of "very toxic" (LD50 ≤50 mg/kg) molecules ranged from 3.58% (mouse-oral-LD_{50}) to 85.06% (cat-intravenous-LD_{50}) across endpoints, while "non-toxic" ($LD_{50} \geq 2000$ mg/kg) molecules varied between 2.83% (mouse-intravenous-LD_{50}) and 67.63% (rabbit-skin-LD_{50}). These classification criteria are based on studies by Mansouri et al., Tran et al., and GHS[5,74]. The 59 acute toxicity endpoints involve 15 different species including mouse, rat, rabbit, guinea pig, dog, cat, bird wild, quail, duck, chicken, frog, mammal, man, women, and human, 8 different administration routes including intraperitoneal, intravenous, oral, skin, subcutaneous, intramuscular, parenteral, and unreported, and 3 different measurement indicators including LD_{50}, LDLo, and TDLo. In this dataset, each compound only has toxicity measurement values concerning a small number of toxicity endpoints, so this dataset is very sparse with nearly 97.4% of compound-to-endpoint measurements missing (Fig. 1b, Supplementary Table 1). Meanwhile, this dataset is also data-unbalanced with some endpoints having tens of thousands of toxicity measurements available, e.g., mouse-intraperitoneal-LD_{50} has 36,295 measurements, mouse-oral-LD_{50} has 23,373 measurements, and rat-oral-LD_{50} has 10,190 measurements, etc., while some endpoints contain only around 100 measurements like mouse-intravenous-LDLo, rat-intravenous-LDLo, frog-subcutaneous-LD_{50}, and human-oral-TDLo, etc. The sparsity and unbalance of this dataset present acute toxicity evaluation as a challenging issue. Among the 59 endpoints, 21 endpoints with <200 measurements were considered small-sized endpoints, and 11 endpoints with more than 1000 measurements were treated as large-sized endpoints. Three endpoints targeting humans, human-oral-TDLo, women-oral-TDLo, and man-oral-TDLo, are typical small-sized endpoints, with only 140, 156, and 163 available toxicity measurements, respectively.

The 115-endpoint dataset is built based on PubChem database[40], a publicly available and free chemical database. First, we collected the acute toxicity effects of 120K chemicals from the PubChem database, involving information on species, administration route, measurement indicators, and corresponding toxicity values. Second, we obtained the data of a total of 636 acute toxicity endpoints. However, most of these endpoints only contained 1–3 available compounds. In order to properly conduct the design of the AI model and implement the 5-fold cross-validation, we only retained the acute toxicity endpoints with no less than 30 available samples for each endpoint. Eventually, a new acute toxicity dataset with 115 endpoints was formed. Third, similar to the operation of acute toxicity data in the TOXRIC database, we also unified the various units of all toxicity measurement values of the above new data into−log(mol/kg), so that the AI model can better regress the acute toxicity values. Overall, the "very toxic" proportions spanned 4.71% (mouse-oral-LD_{50}) to 84.38% (cat-intravenous-LD_{50}), with "non-toxic" proportions ranging from 0.68% (guinea pig-intravenous-LD_{50}) to 42.88% (rabbit-skin-LD_{50}). Compared with the previous 59-endpoint acute toxicity dataset from TOXRIC, the number of acute toxicity endpoints in this new dataset has doubled, adding more possible species (like goat, monkey, hamster, etc.), administration routes (like intracerebral, intratracheal), and measurement indicators (like LD_{10}, LD_{20}). It should be emphasized that the sample imbalance among endpoints and the data missing rate of this dataset are more severe. Its sparsity rate reaches 98.7%, and it contains 68 small-sized acute toxicity endpoints (i.e., endpoints with <200 toxicity measurement data), among which the endpoint with the fewest samples has only 30 available measurement data (Supplementary Table 2). Therefore, this dataset is more challenging for all current acute toxicity prediction models.

The benchmark dataset is collected from the study of Mansouri et al.[5]. To adapt to ToxACoL's regression framework, we exclusively utilized the LD_{50} dataset reported in this study (6398 molecules for training and 2196 molecules for evaluation), which was collected from rat oral experiments with units of mg/kg. Then, we standardized the provided toxicity values into−log(mol/kg) using the same method applied to the aforementioned two datasets, enabling compatibility with ToxACoL for experimental validation.

Design of ToxACoL

Construction of acute toxicity endpoint graph. This operation is performed on training data. Taking the experiments on the 59-endpoint dataset, for example, the graph includes 59 nodes, denoting 59 acute toxicity endpoints. The edges represent the dependency between two endpoints. For each pair of endpoints (i, j), we count the total number of compounds shared by the ith endpoint and the jth endpoint, denoted by NUM(i, j), and then calculate the PCC of the corresponding toxicity measurements between the two endpoints based on their shared compounds (if necessary), denoted by PCC(i, j). Only if their shared compounds are enough and the toxicity measurements of these shared compounds at both endpoints are highly correlated, can the two endpoints be considered dependent, and thus an edge is considered to exist between them:

$$\mathbf{A}[i,j] = \begin{cases} 1 \text{ if } i \neq j \;\&\; \mathrm{NUM}(i,j) \geq \lambda \;\&\; \mathrm{PCC}(i,j) \geq \tau \\ 0 \text{ otherwise,} \end{cases} \quad (1)$$

where λ and τ are two predefined hyperparameters (we set $\lambda = 15$ and $\tau = 0.75$). After traversing all possible edges, we can construct the adjacency matrix $\mathbf{A}$ of the acute toxicity endpoint graph.

Workflow inside adjoint correlation layer. Each adjoint correlation layer takes compound embeddings and endpoint embeddings from the previous adjoint correlation layer as inputs, and outputs the new compound embeddings and endpoint embeddings as inputs for the next adjoint correlation layer:

$$\mathbf{E}_\mathrm{c}^{(l)}, \mathbf{E}_\mathrm{e}^{(l)} = \text{Adjoint Correlation Layer}^{(l)}\left(\mathbf{E}_\mathrm{c}^{(l-1)}, \mathbf{E}_\mathrm{e}^{(l-1)}\right), \quad (2)$$

where $\mathbf{E}_\mathrm{c}^{(l)} \in \mathbb{R}^{B \times d_l}$ denotes the compound embeddings outputted by adjoint correlation layer l (B is the batch size when training), and $\mathbf{E}_\mathrm{c}^{(l-1)} \in \mathbb{R}^{B \times d_{l-1}}$ denotes the compound embeddings outputted by previous layer l–1. $\mathbf{E}_\mathrm{e}^{(l)} \in \mathbb{R}^{N \times d_l}$ denotes the N endpoint embeddings outputted by the graph convolution in adjoint correlation layer l, and $\mathbf{E}_\mathrm{e}^{(l-1)} \in \mathbb{R}^{N \times d_{l-1}}$ denotes the N endpoint embeddings of previous layer l −1. Next, we explain in detail the implementation of the adjoint correlation layer. Based on endpoint graph $\mathbf{A}$, a graph convolution layer is designed to process the endpoint embeddings in each adjoint correlation layer. In this way, the embedding of each endpoint is a mixture of the embeddings of its neighbor endpoints from the previous layer. We adopted the following operation to perform the graph convolution on endpoint embeddings:

$$\mathbf{E}_\mathrm{e}^{(l)} = \sigma(\hat{\mathbf{A}} \otimes \mathbf{E}_\mathrm{e}^{(l-1)} \otimes \mathbf{W}_\mathrm{G}^{(l)}), \quad (3)$$

where ⊗ denotes matrix multiplication, $\hat{\mathbf{A}} \in \mathbb{R}^{N \times N}$ is the normalized adjacency matrix of the N-endpoint graph, $\mathbf{W}_\mathrm{G}^{(l)} \in \mathbb{R}^{d_{l-1} \times d_l}$ is a transformation matrix, and is learnable in the training phase, $\sigma(\cdot)$ denotes a non-linear activation operation (here we used Leaky ReLU with a negative slope of 0.1). Note that $\mathbf{E}_\mathrm{e}^{(0)} \in \mathbb{R}^{N \times d_0}$ is the initial embeddings of the N endpoints and an entity encoding strategy produces it. Concretely, three attributes of the endpoint (species, administration route, and measurement indicator) were separately encoded into three one-hot subvectors and then concatenated to form the initial endpoint embedding. The 59-endpoint dataset involved 15 species, 8 routes, and 3 indicators, so the dimension of initial endpoint embeddings is $d_0 = 15 + 8 + 3 = 26$. Critically, endpoint embeddings $\mathbf{E}_{\mathrm{e}}^{(l)}$ will interact with the compound embeddings:

$$\mathbf{E}_{\mathrm{c}}^{(l)} = f_{\boldsymbol{\theta}^{(l)}}(\mathbf{E}_{\mathrm{c}}^{(l-1)}) \otimes (\mathbf{E}_{\mathrm{e}}^{(l)})^{\top} \otimes \mathbf{W}_{\mathrm{L}}^{(l)} + f_{\boldsymbol{\theta}^{(l)}}(\mathbf{E}_{\mathrm{c}}^{(l-1)}), \tag{4}$$

where $\top$ is a matrix transpose operation. $f_{\boldsymbol{\theta}^{(l)}}(\cdot) : \mathbb{R}^{d_{l-1}} \rightarrow \mathbb{R}^{d_l}$ is the feed-forward layer in adjoint correlation layer l, which consists of Linear layer, BatchNorm layer, Dropout layer, and ReLU activation, connected in series. $\boldsymbol{\theta}^{(l)}$ denotes the learnable parameters in the feed-forward layer. $\mathbf{W}_{\mathrm{L}}^{(l)} \in \mathbb{R}^{N \times d_l}$ is the learnable parameter of the Linear layer (green block in Fig. 1a) in adjoint correlation layer l, aiming at restoring the dimension of compound embeddings. The operation of $f_{\boldsymbol{\theta}^{(l)}}(\mathbf{E}_{\mathrm{c}}^{(l-1)}) \otimes (\mathbf{E}_{\mathrm{e}}^{(l)})^{\top}$ achieves the correlation calculation between compound embeddings and N endpoint embeddings, and the addition term in Eq. (4) achieves a residual connection. Note that $\mathbf{E}_{\mathrm{c}}^{(0)} \in \mathbb{R}^{B \times d'_0}$ is initial compound embedding, and we adopted the 1024-dimensional Avalon fingerprints as $\mathbf{E}_{\mathrm{c}}^{(0)} \in \mathbb{R}^{B \times 1024}$, which can be generated from the SMILES of compounds.

Top regression layer. After connecting multiple adjoint correlation layers in series, a top regression layer was devised to output the final estimation results concerning multiple endpoints. Specifically, assuming there are a total of L adjoint correlation layers (we selected $L = 4$ in our experiments), then the final predictive results $\mathbf{Y}_{\mathrm{e}} \in \mathbb{R}^{B \times N}$ of the B compounds in a mini-batch can be computed as

$$\mathbf{Y}_{\mathrm{e}} = \mathbf{E}_{\mathrm{c}}^{(L)} \otimes \mathbf{W}_{\mathrm{R}} + \mathbf{E}_{\mathrm{c}}^{(L)} \otimes \mathbf{W}_{\mathrm{T}}, \tag{5}$$

where $\mathbf{W}_{\mathrm{T}} \in \mathbb{R}^{d_{\mathrm{L}} \times N}$ denotes the weights of the top linear layer. $\mathbf{W}_{\mathrm{R}} = (\hat{\mathbf{A}} \otimes \mathbf{E}_{\mathrm{e}}^{(L-1)} \otimes \mathbf{W}_{\mathrm{G}}^{(L)})^{\top} \in \mathbb{R}^{d_{\mathrm{L}} \times N}$ denotes the N pre-nonlinear endpoint embeddings outputted by final layer L, which is treated as the N endpoint-wise regressors to fit the toxicity intensity at various endpoints.

Loss function. We assume that in a data mini-batch containing B compounds, the ground-truth toxicity intensity value of the ith compound with respect to the jth endpoint is $\mathbf{y}_{i,j}$, $1 \le i \le B, 1 \le j \le N$. Due to the extreme sparsity of the dataset, $\mathbf{y}_{i,j}$ only has available values on a few (i,j) pairs, and in most cases, it is NULL. Certainly, each compound has the ground-truth toxicity intensity value at least at one endpoint, i.e., $\sum_{1 \le j \le N} \mathrm{I}(\mathbf{y}_{i,j} \ne \mathrm{NULL}) \ge 1,\ \forall i$. So, we designed the following mini-batch loss function to train ToxACoL:

$$L(B,N) = \frac{\sum_{1 \le i \le B} \sum_{1 \le j \le N} \mathrm{I}(\mathbf{y}_{i,j} \ne \mathrm{NULL})(\mathbf{Y}_e[i,j] - \mathbf{y}_{i,j})^2}{\sum_{1 \le i \le B} \sum_{1 \le j \le N} \mathrm{I}(\mathbf{y}_{i,j} \ne \mathrm{NULL})}, \tag{6}$$

where $\mathbf{Y}_e[i,j]$ is the estimated toxicity intensity of the ith compound with regard to the jth endpoint. $\mathrm{I}(\mathbf{y}_{i,j} \ne \mathrm{NULL})$ is an indicator function, equal to 1 if $\mathbf{y}_{i,j}$ is an available value and 0 when it is missing. This loss filters out the missing compound-to-endpoint measurements and fully utilizes the existing supervision information provided by the sparse acute toxicity dataset.

Quantification and statistical analysis

For each endpoint, the determination coefficient, R^2, is used as the main evaluation metric:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\bar{y} - y_i)^2}, \tag{7}$$

where n is the total number of compounds at this endpoint, y_i and $\hat{y}_i$ denote the ground-truth toxicity value and the estimated toxicity intensity value for the ith compound, $1 \le i \le n$, respectively, while $\bar{y}$ is the average toxicity intensity value over all the n compounds at this endpoint. We also calculated the RMSE for each endpoint as an additional evaluation metric: $\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$. Unless otherwise specified, the R^2 or RMSE metric values we reported are the average results after 5-fold cross-validation.

The Friedman test and Nemenyi test[46] were employed to analyze the performance difference among various methods over all N acute toxic endpoints. The Friedman test is a nonparametric test equivalent to the analysis of variance (ANOVA), used to ascertain whether there are statistically significant differences in the means of three or more methods tested on identical endpoints. Assuming $\mathrm{Rank}_j = \frac{1}{N} \sum_{i=1}^{N} \mathrm{rank}_j^i$ denotes the average performance ranking of the jth method over N endpoints (rank_j^i is the ranking of the jth method on the ith endpoint), The null hypothesis believes that the performances of all methods are equal. The Friedman statistic in the following equation obeys the χ_F^2-distribution with (K−1) free degrees:

$$\chi_F^2 = \frac{12N}{K(K+1)} \left[\sum_{j=1}^{K} \mathrm{Rank}_j^2 - \frac{K(K+1)^2}{4} \right], \tag{8}$$

where K is the total number of methods. Furthermore, the statistic in the following equation obeys the F-distribution with free degrees of (K −1) and (K−1)(N−1):

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1) - \chi_F^2}. \tag{9}$$

Next, the Nemenyi test was used to evaluate the performance difference between pairwise methods, deeming the pairwise methods significantly different if their average ranking gap exceeds the critical difference (CD):

$$\mathrm{CD} = q_\alpha \sqrt{\frac{K(K+1)}{6N}}, \tag{10}$$

where q_α is based on the Studentized range statistic divided by $\sqrt{2}$ and α is the corresponding confidence level.

In addition, we adopted kernel density estimation[45], a powerful non-parametric statistical method used to estimate unknown probability density functions, to fit the overall performance distribution of different models on all endpoints and then compare the performance balance of different methods on all endpoints. Two-sided Wilcoxon signed-rank test[75] was used to compute the significant difference between the two methods.

Baseline models

Several state-of-the-art STL, MTL, and consensus models for acute toxicity prediction were compared with ToxACoL. ST-DNN: training a deep neural network for each endpoint and taking the Avalon fingerprints as inputs. The complexity of the single-task neural network for each endpoint varies with the sample size of the available toxicity measurement about this endpoint. ST-RF: building a random forest for each endpoint, and the total number of trees varies with the sample size concerning this endpoint. It also takes the Avalon fingerprints as inputs. GAT: combining a graph neural network with an attention mechanism that determines the relative importance of neighboring nodes. GCN: operating on the molecule graph structure of compounds using convolutional neural networks rather than on fingerprint vectors. Attentive FP: a graph architecture that represents molecules using molecular fingerprints and a graph attention mechanism. MT-DNN: a multi-channel deep neural network for all endpoints, where the anterior feature encoder in this network for all endpoints is shared but the regressors for various endpoints are independent. Its input is also the 1024-dimensional Avalon fingerprints. MT-GCN: a multi-task graph convolution network, where the input is the molecule graph structure of compounds and the molecule graph learning module is shared across all endpoints while the final regressors towards endpoints are independent. DLCA: a consensus learning architecture that averages the outputs of separate multi-task networks, including four descriptor-based networks and one descriptor-free network. The four descriptor-based networks were trained using four different types of molecular representation (Avalon, Morgan, AtomPair fingerprints, and RDKit descriptors). The one descriptor-free network was realized based on SMILES strings, consisting of a convolutional network with 1D convolutional layers and GlobalMax pooling layers followed by fully connected layers and multi-channel regression layers.

Model configuration and computational resource

The split of the training and testing sets in the dataset is completely consistent with the baseline models[17], which was randomly divided 5 times for 5-fold cross-validation. The results we reported mostly came from a ToxACoL model with four adjoint correlation layers, where the embedding dimensions are designed as $d_1 = 768$, $d_2 = 512$, $d_3 = 384$, and $d_4 = 64$, respectively. The channel number of the top regression layer was designed to match the total number of toxic endpoints. The initial compound embeddings are the 1024-dimensional Avalon fingerprints, and the initial endpoint embeddings are the 26-dimensional (59-endpoint dataset) or 35-dimensional (115-endpoint dataset) binary vectors concatenated by three one-hot subvectors corresponding to the three attributes of the endpoints. The dropout rate in feed-forward layers was maintained at 0.1. The non-linear activation function of graph convolution was set to Leaky ReLU with a negative slope of 0.1. The SGD with Nestrerov momentum[76] was used as the optimizer, where the momentum factor is 0.9, the weight decay is 5e−4 and the initial learning rate is 0.01. Our ToxACoL model was trained for 120 epochs with a batch size of 32, and the optimal model and another nine models obtained from its nine nearest epochs were saved. For inference on testing compounds, the average outputs from the 10 well-trained models were treated as the final estimation results.

The model was implemented via Python 3.10, and the main dependent packages only include Torch, Pandas, and Rdkit. When training, the codes were run on an Intel I9-13900KF processor and an NVIDIA RTX4090 GPU, with the configuration deployed on an Ubuntu 22.04 LTS system. The training takes ~15 min. During the inference phase, the toxicity of chemical compounds can be inferred entirely using the CPU environment. We conducted tests on a laptop equipped with an ordinary Intel Core i5 processor, and the model only takes about 0.8 s to predict the toxicity of a single molecule on various acute toxicity endpoints. Such low-demand computing resources can enable many researchers to use our model.

Reduction of toxicity measurements for training

For the experiments of reducing training measurements for the 59-endpoint dataset, the data reduction was made on the 21 small-sized endpoints, where the available measurements for each small-sized endpoint in the training set were randomly discarded according to a certain ratio, like 80%, 50%, 40%, 30%, 20%, and 10%. This was done by randomly replacing the ground-truth toxicity intensity of some available measurements with NULL. Technically, we can achieve it just by modifying the indicator function value of the randomly selected training measurements from $I(y_{i,j} \neq \text{NULL}) = 1$ into $I(y_{i,j} \neq \text{NULL}) = 0$ during the ToxACoL training process.

Calculation of applicability domain

According to the Organization for Economic Co-operation and Development (OECD) guidelines, QSAR modeling requires defining the AD to ensure reliable predictions. The reliability of a model's prediction depends on whether the queried compound falls within the AD defined by the model[77–79]. We quantify the AD of ToxACoL using Tanimoto similarity based on molecular fingerprints. First, we calculate the Tanimoto similarity matrix between the test set and training set using chemical Avalon fingerprints. The AD threshold S_T is calculated using the equation:

$$S_T = \gamma + Z\sigma, \tag{11}$$

where γ and σ represent the mean and the standard deviation of the Tanimoto similarities between all compounds in the training set, reflecting the overall similarity level within the training set. Z is an adjustable confidence parameter. Next, for each compound in the test set, we calculate its Tanimoto similarity to all compounds in the training set. The top k compounds from the training set with the highest similarity to the test compound are selected (k is set to 10 as a predefined value). If the average similarity of these k compounds exceeds the threshold S_T, the compound is considered within the AD otherwise, it is classified as outside the AD.

Extrapolation experiment

Different from the previous 5-fold cross-validation experiment, we used all the data in the 59-endpoint acute toxicity dataset to train our ToxACoL in the extrapolation experiment. Since we need to study the potential associations of toxicity responses between animal endpoints and human endpoints, we selected the three human-related endpoints in TOXRIC as the control benchmarks and took the compounds with existing toxicity values for these three endpoints as research objects. Among them, the number of available compounds for the human-oral-TDLo, woman-oral-TDLo, and man-oral-TDLo endpoints were 140, 156, and 163, respectively. Taking the human-oral-TDLo as an example, we used the trained ToxACoL to predict the acute toxicity of these 140 compounds at the other 58 endpoints and filled in the missing toxicity values in the dataset. After this operation, we can obtain the acute toxicity values of the same set of compounds at different endpoints. Based on this, we can conduct Pearson correlation analysis and statistical tests to explore the extrapolation relationships between different animal endpoints and the human-oral-TDLo endpoint.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The 59-endpoint acute toxicity dataset used in our study is freely available on the published data platform TOXRIC[25] at https://toxric.bioinforai.tech/home, and the 115-endpoint acute toxicity dataset is freely available on the PubChem database[40] at https://pubchem.ncbi.nlm.nih.gov/. The two datasets have also been organized by us and published at https://doi.org/10.6084/m9.figshare.27195339.v5with DOIs and citations[80]. Source data are provided with this paper.

Code availability

The code used to develop the model, perform the analyses, and generate results in this study is publicly available and has been deposited in GitHub at https://github.com/LuJiangTHU/Acute_Toxicity_FSL, under CC-BY 4.0 license. The specific version of the code associated with this publication is archived in Zenodo and is accessible via 10.5281/zenodo.15063595[81]. In addition, to enable all researchers to directly use our pre-trained ToxACoL, we have integrated our ToxACoL into an online software using the Docker containerization platform. This online software can accurately predict the acute toxicity values and the GHS classes of chemical compounds on different acute toxicity endpoints. It can be freely accessed and used at https://toxacol.bioinforai.tech/, in the hope of becoming a useful resource for regulatory applications.

References

1. Johnson, A. C., Jin, X., Nakada, N. & Sumpter, J. P. Learning from the past and considering the future of chemicals in the environment. *Science* **367**, 384–387 (2020).
2. Saiz-Lopez, A. et al. Natural short-lived halogens exert an indirect cooling effect on climate. *Nature* **618**, 967–973 (2023).
3. Lane, M. K. M. et al. Green chemistry as just chemistry. *Nat. Sustain.* **6**, 502–512 (2023).
4. Luechtefeld, T., Rowlands, C. & Hartung, T. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicol. Res.* **7**, 732–744 (2018).
5. Mansouri, K. et al. Catmos: collaborative acute toxicity modeling suite. *Environ. Health Perspect.* **129**, 047013 (2021).
6. Lorke, D. A new approach to practical acute toxicity testing. *Arch. Toxicol.* **54**, 275–287 (1983).
7. Bownik, A. Physiological endpoints in daphnid acute toxicity tests. *Sci. Total Environ.* **700**, 134400 (2020).
8. Krewski, D. et al. Toxicity testing in the 21st century: progress in the past decade and future perspectives. *Arch. Toxicol.* **94**, 1–58 (2020).
9. Wu, S. et al. Accurate prediction of rat acute oral toxicity and reference dose for thousands of polycyclic aromatic hydrocarbon derivatives based on chemometric qsar and machine learning. *Environ. Sci. Technol.* **58**, 15100–15110 (2024).
10. FDA, U. Single-dose acute toxicity testing for pharmaceuticals; revised guidance; availability notice. *Fed. Regist.* **61**, 43933–43935 (1996).
11. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
12. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
13. Chinedu, E., Arome, D. & Ameh, F. S. A new method for determining acute toxicity in animal models. *Toxicol. Int.* **20**, 224 (2013).
14. Strickland, J. et al. Status of acute systemic toxicity testing requirements and data uses by US regulatory agencies. *Regul. Toxicol. Pharmacol.* **94**, 183–196 (2018).
15. Kirchmair, J. et al. Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Discov.* **14**, 387–404 (2015).
16. Huang, R. et al. Modelling the tox21 10 k chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat. Commun.* **7**, 10425 (2016).
17. Jain, S. et al. Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods. *J. Chem. Inf. Model.* **61**, 653–663 (2021).
18. Bhattacharya, S., Zhang, Q., Carmichael, P. L., Boekelheide, K. & Andersen, M. E. Toxicity testing in the 21st century: defining new risk assessment approaches based on perturbation of intracellular toxicity pathways. *PLoS ONE* **6**, e20887 (2011).
19. Clark, J. M. The 3rs in research: a contemporary approach to replacement, reduction and refinement. *Br. J. Nutr.* **120**, S1–S7 (2018).
20. Schechtman, L. M. Implementation of the 3rs (refinement, reduction, and replacement): validation and regulatory acceptance considerations for alternative toxicological test methods. *ILAR J.* **43**, S85–S94 (2002).
21. Basile, A. O., Yahi, A. & Tatonetti, N. P. Artificial intelligence for drug toxicity and safety. *Trends Pharmacol. Sci.* **40**, 624–635 (2019).
22. Cavasotto, C. N. & Scardino, V. Machine learning toxicity prediction: latest advances by toxicity end point. *ACS Omega* **7**, 47536–47546 (2022).
23. Lin, Z. & Chou, W.-C. Machine learning and artificial intelligence in toxicological sciences. *Toxicol. Sci.* **189**, 7–19 (2022).
24. Wu, X., Zhou, Q., Mu, L. & Hu, X. Machine learning in the identification, prediction and exploration of environmental toxicology: challenges and perspectives. *J. Hazard. Mater.* **438**, 129487 (2022).
25. Wu, L. et al. Toxric: a comprehensive database of toxicological data and benchmarks. *Nucleic Acids Res.* **51**, D1432–D1445 (2023).
26. Luechtefeld, T. et al. Global analysis of publicly available safety data for 9801 substances registered under reach from 2008–2014. *Altex* **33**, 95 (2016).
27. Van Noorden, R. Software beats animal tests at predicting toxicity of chemicals. *Nature* **559**, 163–164 (2018).
28. Luechtefeld, T., Marsh, D., Rowlands, C. & Hartung, T. Machine learning of toxicological big data enables read-across structure activity relationships (rasar) outperforming animal test reproducibility. *Toxicol. Sci.* **165**, 198–212 (2018).
29. Pognan, F. et al. The evolving role of investigative toxicology in the pharmaceutical industry. *Nat. Rev. Drug Discov.* **22**, 317–335 (2023).
30. Ryu, J. Y., Jang, W. D., Jang, J. & Oh, K.-S. Predaot: a computational framework for prediction of acute oral toxicity based on multiple random forest models. *BMC Bioinform.* **24**, 66 (2023).
31. Elsayad, A. et al. Machine learning model for random forest acute oral toxicity prediction. *Glob. J. Environ. Sci. Manag.* **11**, 21–38 (2025).
32. Gao, J. et al. Multi-task multi-view and iterative error-correcting random forest for acute toxicity prediction. *Expert Syst. Appl.* **274**, 126972 (2025).
33. Xiong, Z. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**, 8749–8760 (2019).
34. Wang, G., Feng, H., Du, M., Feng, Y. & Cao, C. Multimodal representation learning via graph iso-morphism network for toxicity multitask learning. *J. Chem. Inf. Model.* **64**, 8322–8338 (2024).
35. Yang, X. et al. Multi-task aquatic toxicity prediction model based on multi-level features fusion. *J. Adv. Res.* **68**, 477–489 (2025).
36. Sosnin, S., Karlov, D., Tetko, I. V. & Fedorov, M. V. Comparative study of multitask toxicity modeling on a broad chemical space. *J. Chem. Inf. Model.* **59**, 1062–1072 (2019).
37. Sharma, B. et al. Accurate clinical toxicity prediction using multitask deep neural nets and contrastive molecular explanations. *Sci. Rep.* **13**, 4908 (2023).
38. Standley, T. et al. Which tasks should be learned together in multitask learning? In *Int. Conf. Mach. Learn.* 9120–9132 (ICML, 2020).
39. Gorelick, F. S. & Lerch, M. M. Do animal models of acute pancreatitis reproduce human disease? *Cell. Mol. Gastroenterol. Hepatol.* **4**, 251–262 (2017).
40. Kim, S. et al. Pubchem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).
41. Cohen, I. et al. Pearson correlation coefficient. In *Noise Reduction in Speech Processing* 1–4 (Springer, 2009).
42. Zakharov, A. V. et al. Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models. *J. Chem. Inf. Model.* **59**, 4613–4624 (2019).
43. Ketkar, R., Liu, Y., Wang, H. & Tian, H. A benchmark study of graph models for molecular acute toxicity prediction. *Int. J. Mol. Sci.* **24**, 11966 (2023).
44. Saravanan, K. I. et al. A deep learning based multi-model approach for predicting drug-like chemical compound's toxicity. *Methods* **226**, 164–175 (2024).
45. Chen, Y.-C. A tutorial on kernel density estimation and recent advances. *Biostat. Epidemiol.* **1**, 161–187 (2017).
46. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).
47. Chen, D. et al. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *AAAI Conf. Artificial Intell.* **34**, 3438–3445 (AAAI, 2020).
48. Keriven, N. Not too little, not too much: a theoretical analysis of graph (over) smoothing. In *Adv. Neural. Inf. Process. Syst.* **35**, 2268–2281 (NeurIPS, 2022).
49. Giraldo, J. H., Skianis, K., Bouwmans, T. & Malliaros, F. D. On the trade-off between over-smoothing and over-squashing in deep graph neural networks. In *ACM Int. Conf. Inf. Knowledge Management* 566–576 (CIKM, 2023).

50. Van der Maaten, L. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
51. Limban, C. et al. The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toricol. Rep.* **5**, 943–953 (2018).
52. Stepan, A. F. et al. Structural alert /reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the united states. *Chem. Res. Toxicol.* **24**, 1345–1410 (2011).
53. Payne, M. & Walsh, P. Structure-activity relationships for skin sensitization potential: development of structural alerts for use in knowledge-based toxicity prediction systems. *J. Chem. Inf. Comput. Sci.* **34**, 154–161 (1994).
54. Kazius, J., McGuire, R. & Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **48**, 312–320 (2005).
55. Enoch, S., Madden, J. & Cronin, M. Identification of mechanisms of toxic action for skin sensitisation using a smarts pattern based approach. *SAR QSAR Environ. Res.* **19**, 555–578 (2008).
56. Zhang, R. et al. Modeling and insights into the structural basis of chemical acute aquatic toxicity. *Ecotoxicol. Environ. Saf.* **242**, 113940 (2022).
57. Zhang, C. et al. Quaternary ammonium compounds (QACs): a review on occurrence, fate and toxicity in the environment. *Sci. Total Environ.* **518**, 352–362 (2015).
58. Mohapatra, S. et al. Quaternary ammonium compounds of emerging concern: classification, occurrence, fate, toxicity and antimicrobial resistance. *J. Hazard. Mater.* **445**, 130393 (2023).
59. Qv, X.-Y. & Jiang, J.-G. Toxicity evaluation of two typical surfactants to *Dunaliella bardawil*, an environmentally tolerant alga. *Environ. Toxicol. Chem.* **32**, 426–433 (2013).
60. Ying, G.-G. Fate, behavior and effects of surfactants and their degradation products in the environment. *Environ. Int.* **32**, 417-431 (2006).
61. Hall, B. S. & Wilkinson, S. R. Activation of benznidazole by trypanosomal type I nitroreductases results in glyoxal formation. *Antimicrob. Agents Chemother.* **56**, 115–123 (2012).
62. Rojo, G. et al. Toxic and therapeutic effects of nifurtimox and benznidazol on *Trypanosoma cruzi* ex vivo infection of human placental chorionic villi explants. *Acta Trop.* **132**, 112–118 (2014).
63. Miller, D. A. et al. Tolerance of benznidazole in a United States Chagas disease clinic. *Clin. Infect. Dis.* **60**, 1237–1240 (2015).
64. Davies, C. et al. Hepatotoxicity in mice of a novel anti-parasite drug candidate hydroxymethyl nitrofurazone: a comparison with benznidazole. *PLoS Negl. Trop. Dis.* **8**, e3231 (2014).
65. Tan, B., Liao, L., Yang, Z., Long, X. & Li, J. Toward the toxicology of some nitro-compounds. *Mod. Org. Chem. Res* **3**, 11–21 (2018).
66. Kovacic, P. & Somanathan, R. Nitroaromatic compounds: environmental toxicity, carcinogenicity, mutagenicity, therapy and mechanism. *J. Appl. Toxicol.* **34**, 810–824 (2014).
67. Clayson, D. Carcinogenic aromatic amines and related compounds. *ACS Monogr.* **173**, 366–461 (1976).
68. Purohit, V. & Basu, A. K. Mutagenicity of nitroaromatic compounds. *Chem. Res. Toxicol.* **13**, 673–692 (2000).
69. Manzetti, S., van der Spoel, E. R. & van der Spoel, D. Chemical properties, environmental fate, and degradation of seven classes of pollutants. *Chem. Res. Toxicol.* **27**, 713–737 (2014).
70. Addink, R. et al. Surface catalyzed formation of polychlorinated dibenzo-*p*-dioxins/dibenzofurans during municipal waste incineration. *Chemosphere* **20**, 1929–1934 (1990).
71. Ruelius, H. W. Extrapolation from animals to man: predictions, pitfalls and perspectives. *Xenobiotica* **17**, 255–265 (1987).
72. De, P., Kar, S., Ambure, P. & Roy, K. Prediction reliability of QSAR models: an overview of various validation tools. *Arch. Toxicol.* **96**, 1279–1295 (2022).
73. Tropsha, A., Isayev, O., Varnek, A., Schneider, G. & Cherkasov, A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discov.* **23**, 141–155 (2024).
74. Tran, T. T. V. Artificial intelligence in drug toxicity prediction: recent advances, challenges, and future perspectives. *J. Chem. Inf. Model.* **63**, 2628–2643 (2023).
75. Woolson, R. F. Wilcoxon signed-rank test. In *Wiley Encyclopedia of Clinical Trials* 1–3 (Wiley, 2007).
76. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. On the importance of initialization and momentum in deep learning. In *Int. Conf. Mach. Learn.* 1139–1147 (ICML, 2013).
77. Sahigara, F. et al. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **17**, 4791–4810 (2012).
78. Patlewicz, G. & Fitzpatrick, J. M. Current and future perspectives on the development, evaluation, and application of in silico approaches for predicting toxicity. *Chem. Res. Toxicol.* **29**, 438–451 (2016).
79. Lou, S. et al. In silico prediction of chemical acute dermal toxicity using explainable machine learning methods. *Chem. Res. Toxicol.* **37**, 513–524 (2024).
80. Lu, J. The toxicity data of compounds. figshare. Dataset. https://doi.org/10.6084/m9.figshare.27195339.v5 (2024).
81. Lu, J. LuJiangTHU/Acute_Toxicity_FSL: ToxACoL for multi-endpoint acute toxicity prediction (v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.15063595 (2025).

Acknowledgements

This work was supported by the National Key R&D Program of China (2023YFC2604400 to S.H. and 2024YFA1307700 to X.B.).

Author contributions

J.L., L.W. and S.H. conceived and designed the study. J.L. developed the methodology and code, carried out the evaluation, achieved the comparison with baseline models, and created the visualizations. L.W. built the acute toxicity dataset, analyzed the model interpretability, designed experiments for structural alerts and toxicity extrapolation, and explored the model's utility in chemical toxicology. R.L. and P.Z. provided insights into toxicity representation. M.W. collected the acute toxicity data and achieved the comparison with several graph-based models. J.Y. carried out data preprocessing and organization. J.L., L.W. and S.H. contributed to analyzing the results and writing and improving the manuscript. S.H., H.B. and X.B. supervised the study. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-60989-7.

Correspondence and requests for materials should be addressed to Hui Bai, Song He or Xiaochen Bo.

Peer review information *Nature Communications* thanks Thomas Hartung and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.