communications biology

Article

https://doi.org/10.1038/s42003-024-06758-6

Deep active learning with high structural discriminability for molecular mutagenicity prediction

Huiyan Xu [1,2,3], Yanpeng Zhao [2,3], Yixin Zhang[2,3], Junshan Han[2], Peng Zan [1,4] ✉, Song He [2,4] ✉ & Xiaochen Bo [2,4] ✉

The assessment of mutagenicity is essential in drug discovery, as it may lead to cancer and germ cells damage. Although in silico methods have been proposed for mutagenicity prediction, their performance is hindered by the scarcity of labeled molecules. However, experimental mutagenicity testing can be time-consuming and costly. One solution to reduce the annotation cost is active learning, where the algorithm actively selects the most valuable molecules from a vast chemical space and presents them to the oracle (e.g., a human expert) for annotation, thereby rapidly improving the model's predictive performance with a smaller annotation cost. In this paper, we propose muTOX-AL, a deep active learning framework, which can actively explore the chemical space and identify the most valuable molecules, resulting in competitive performance with a small number of labeled samples. The experimental results show that, compared to the random sampling strategy, muTOX-AL can reduce the number of training molecules by about 57%. Additionally, muTOX-AL exhibits outstanding molecular structural discriminability, allowing it to pick molecules with high structural similarity but opposite properties.

In the field of modern drug discovery, mutagenicity has received great attention due to its high correlation with carcinogenicity and heritable variation[1–3]. Mutagenicity refers to permanent, transmissible changes in the quantity or structure of the genetic material of cells and organisms. These changes may occur in individual genes, gene clusters or entire chromosomes which can impact processes such as cell growth, differentiation, and apoptosis[4,5]. Many approved drugs have been withdrawn from the market because they have been identified as mutagens in humans or animals. A typical instance is furazolidone, it has been used to treat diarrhoea and enteritis caused by bacteria or protozoan infections, including traveler's diarrhoea, cholera and bacteremic salmonellosis. However, Nitrofurans are recognized by The Food and Drug Administration as mutagens and carcinogens, and can no longer be used since 1991[6]. Therefore, for new drugs, it is important to assess their mutagenicity which helps to evaluate their safety and potential risks, as well as provides a basis for toxicological assessment.

One of the most widely used methods for identifying the mutagenicity of chemicals is the Ames test[1,7–9]. However, due to the enormous number of molecules in chemical space, it is practically impossible to test every molecule for its mutagenicity using experimental methods. As an alternative, in silico methods have gained significant attention for predicting mutagenicity before clinical drug development trials[10–12]. Researchers have proposed some deep learning methods aimed at accurately predicting molecular mutagenicity[13–19]. However, since traditional mutagenicity prediction methods usually train the entire neural network at once using labeled data, the scarcity of labeled data limits the performance of in silico methods in prediction tasks[20,21]. Furthermore, from the perspective of data selection, not all molecules contribute significantly to the improvement of model performance. Randomly selecting new molecules from a vast chemical space for wet lab annotation often entails substantial resource consumption and may not necessarily lead to equivalent performance gains[22]. Therefore, the current challenge is to expand the labeled molecules as fast as possible from a large chemical space at a lower cost.

Active learning is a promising strategy to tackle this challenge[23], it adopts a human-in-the-loop paradigm, employing an iterative strategy of data collection, annotation, and training to assist experimentalists in guiding both data collection and model training. It involves targeted exploration within a vast chemical space, utilizing a specific set of rules to identify molecules that maximize the enhancement of model performance. By validating these molecules through wet lab experiments, active learning achieves greater improvements in model performance compared to random

[1]Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronics Engineering and Automation, Shanghai University, Shanghai, China. [2]Academy of Military Medical Sciences, Beijing, China. [3]These authors contributed equally: Huiyan Xu, Yanpeng Zhao, Yixin Zhang.[4]These authors jointly supervised this work: Peng Zan, Song He, Xiaochen Bo. ✉e-mail: zanpeng@shu.edu.cn; hes1224@163.com; boxc@bmi.ac.cn

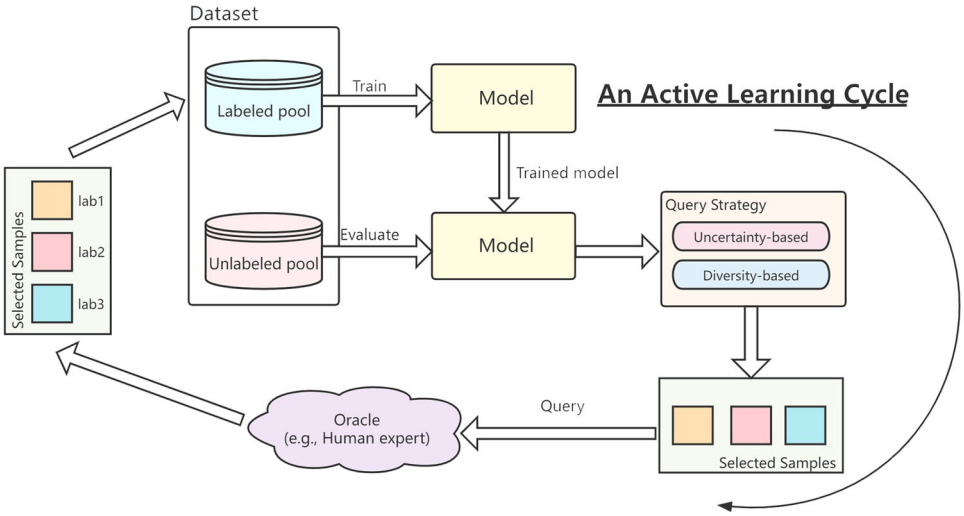

Fig. 1 | The schematic diagram of active learning.

selection strategies, all within the same experimental annotation budget. As a result, active learning reduces annotation costs and saves time through active guidance. Specifically, active learning commences with a small set of labeled data. The initial labeled data is used to train a deep learning model, and a well-designed query strategy is used to select a subset of the unlabeled data which are deemed to be the most informative ones. The molecules are submitted to a wet lab for labeling and then combined with the original labeled data to form a larger training set, which is used to retrain the model. The selection process is repeated until the desired performance is achieved or the annotation budget is exhausted. Due to the success in computer vision[24–26], active learning has gained increasing attention from researchers in drug discovery and has been applied in various fields such as guiding chemical reaction screening[22,27] and drug-drug interaction[28]. Recent reviews have also emphasized the significant role of active learning in scientific discovery, which has gradually attracted the attention of researchers[29]. It should be noted that Bayesian optimization and active learning are two different approaches. Bayesian optimization focuses on finding the optimal value, while active learning is to reduce the amount of labeled data by selecting the most informative examples to label[30].

Current research has demonstrated that, when using the same annotation cost, exploring more informative regions through active learning can achieve higher performance than random exploration[31,32]. With the query strategy guided by the active learning algorithm, we can effectively evaluate the informativeness of each molecule and selectively explore the vast chemical space, reducing the huge annotation cost. Therefore, active learning is well suited for mutagenicity prediction problems where annotation costs are high.

In this paper, we present a deep active learning pipeline called muTOX-AL (mutagenicity TOXicity-Active Learning), for molecular mutagenicity prediction. By iteratively selecting informative samples through active learning strategies, it significantly reduces the cost of data labeling and accelerates the process of drug discovery. The results show that compared to passive random exploration strategies, active learning can greatly reduce the required training sample size. Specifically, the contributions of this paper include:

1) We propose an active learning strategy for molecular mutagenicity prediction. We use uncertainty (i.e., samples that are hard to distinguish by the model) as a measure of sample informativeness and score the samples by a trained uncertainty estimation module (see "Methods" for more details).

2) Compared to traditional and state-of-the-art active learning methods, our approach achieves the same testing accuracy using fewer samples, especially compared to the random strategy, reducing about 57% of training molecules. This demonstrates the effectiveness and superiority of our method.

3) We provide an explanation for why the selected molecules are more informative via t-SNE visualization. The visualization demonstrates that these molecules are located closer to the classification boundary of the model, which further supports their value in improving model performance.

4) muTOX-AL demonstrates significant structural discriminability. On the one hand, the model prefers selecting samples with low structural similarity, preventing overfitting of simple relationships. On the other hand, the model also considers samples with similar structures but different properties, providing a more comprehensive understanding of the complex relationship between molecular structure and mutagenicity.

Results

muTOX-AL overview

muTOX-AL is developed upon a deep active learning technique. The schematic diagram of active learning is shown in Fig. 1: A small number of labeled samples are used for training, and then the trained model is used to select the most informative samples from the unlabeled pool. These samples are given to the oracle for labeling, and then added to the labeled pool. We use uncertainty as a measure of sample informativeness (see "Methods" for more details).

The whole framework of muTOX-AL, as shown in Fig. 2, consists of four parts: the feature extraction module, the backbone module, the uncertainty estimation module, and the loss calculation module. In the training phase (Phase 1), the total training set is divided into two parts: an initialed labeled pool of 200 randomly selected samples and an unlabeled pool of all remaining samples. In this case, we make all samples in the unlabeled pool "blind," i.e., the labels are not visible to the model, to simulate a real scenario with fewer labeled samples and more unlabeled samples. The molecular fingerprints and descriptors of the samples in the initial labeled pool are extracted as input features for the backbone module, which is responsible for predicting mutagenicity of the molecules. The uncertainty estimation module is used to determine the informativeness of the samples. The deeper features obtained by the hidden layer of the backbone module are extracted as input of the uncertainty estimation module. Finally, models are jointly trained by calculating the total loss of the backbone module and the uncertainty estimation module. In the active learning phase (Phase 2), the trained model calculates the uncertainty scores of all samples in the unlabeled pool. Next, the samples with the highest uncertainty scores are given to the oracle for annotation and added to the labeled pool. Here, the labels of the selected "blind" samples are revealed to the model to simulate the interaction with a real wet lab scenario. The above process is iterated until the entire label budget is exhausted or a predefined stopping condition is met.

Statistical analysis of the dataset

In this study, muTOX-AL was trained using the TOXRIC dataset, and the performance of the proposed method was verified. TOXRIC dataset was

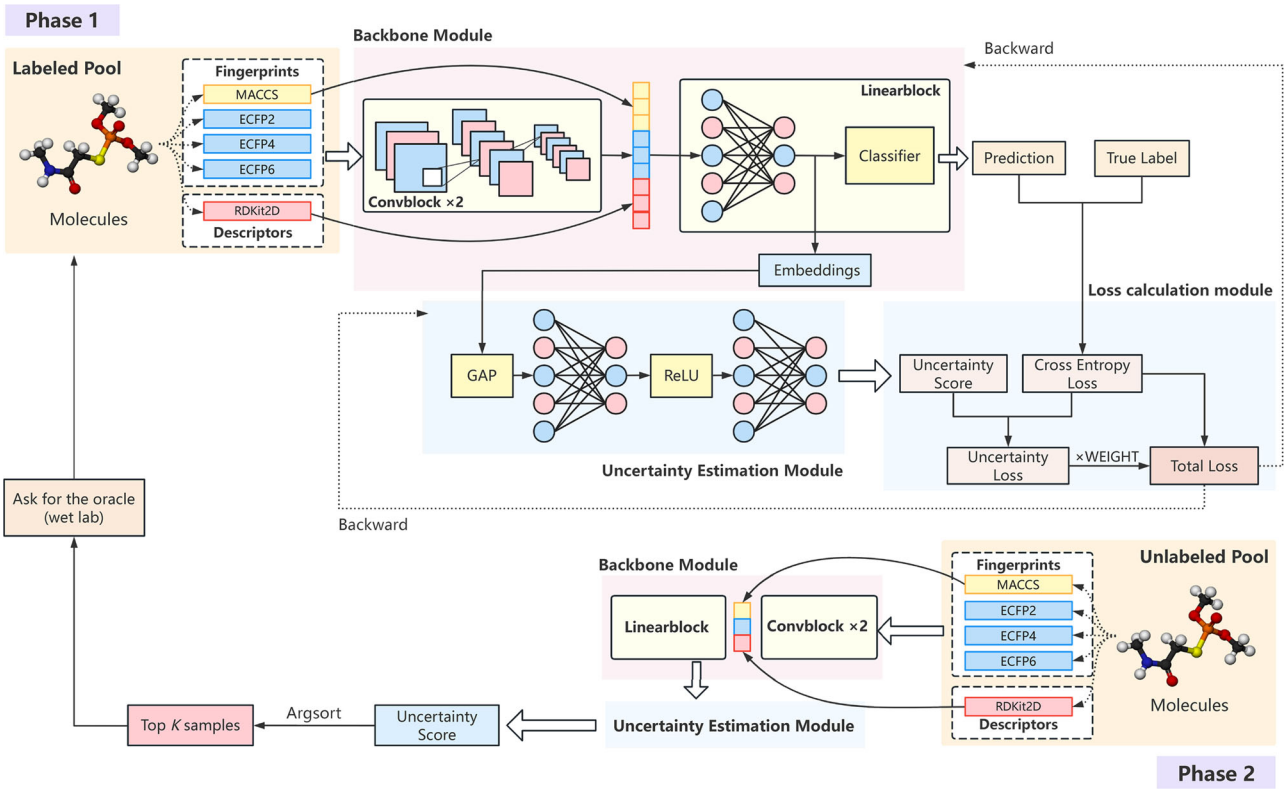

Fig. 2 | The overall structure of muTOX-AL. In Phase 1, the molecules in the labeled pool are used for training. The five features of molecules are extracted as the input for the backbone module, which is used to classify the molecules' mutagenicity. The deeper features obtained by the hidden layer of the backbone module are extracted as input of the uncertainty estimation module, which is used to estimate the uncertainty of the samples. Finally, models are jointly trained by calculating the total loss of the backbone module and the uncertainty estimation module. In Phase 2, the uncertainty estimation module calculates the uncertainty scores of all samples in the unlabeled sample pool. The samples with the highest uncertainty are selected and given to oracle for annotation, and then updates the labeled sample pool.

collected and collated from the TOXRIC website (https://toxric.bioinforai.tech/home) and includes a total of 7495 compounds[33]. We used a randomized five-fold cross-validation approach for model training and performance validation, where each fold was used in turn as a test set, while the remaining four folds were used as a training set (total training set sample size = 5988 samples). To gain insight on the distribution of mutagenicity labels as well as chemical structure domains of the samples in the dataset, we performed the following statistical analyses on the TOXRIC dataset.

First, we performed a labeling analysis on the overall data, as shown in Fig. 3A. TOXRIC dataset is balanced, the number of positive samples was 4196, accounting for 56.05%, while the number of negative samples was 3289. Overall, the prediction scores on the dataset are less affected by the class imbalance.

Second, we visualized the chemical space in which the TOXRIC dataset is embedded. We generated MACCS fingerprints for the samples in the dataset and performed principal component analyses, the results are shown in Fig. 3B. The distribution of the dataset is relatively diffuse, suggesting that the data are highly variable in the direction of the principal components and that the samples occupy a broad and dispersed chemical space.

Finally, we investigated the relationship between structural domains and mutagenicity in the TOXRIC dataset. Using locally sensitive hash forest indexing and k nearest neighbors (kNN) plots, we represent and visualize the TOXRIC dataset as a minimum spanning tree (MSP)[34], with the structural similarity distribution illustrated in Fig. 3C. This visual representation enabled the identification of clusters of structurally homogeneous molecules for both non-mutagens and mutagens. It was observed that it is challenging to directly distinguish molecular subgroups with specific substructures based on structure alone. In addition, the distribution of each fold of data in the five-fold cross-validation was analyzed (see Supplementary Fig. 1 for details). Specifically, we visualized the distribution of samples over the structural domains in the training and test sets by means of the MSP. In each fold, the molecules in the test set are uniformly distributed in different structural domains.

Comparisons between different active learning methods

To verify the effectiveness of the muTOX-AL, we compare muTOX-AL with other active learning strategies such as random strategy, margin-based active learning strategy[35], entropy-based active learning strategy[36], TOD[37] active learning strategy and Core-set active learning strategy[38]. We use some common evaluation metrics for active learning performance. Figure 4A, B shows the Accuracy and F1-score of muTOX-AL with five active learning baselines on the mutagenicity dataset, where the x-axis represents the number of labeled samples per cycle and the y-axis represents the accuracy and F1-score, respectively, other metrics can be found in Supplementary Fig. 4. From Fig. 4A, B, we can see that muTOX-AL performs best among all active learning methods, which uses only about 24% of training sets to achieve 95% of the accuracy of supervised learning using all training samples (5988 samples). The result demonstrating that our active learning strategy is able to explore more informative molecules in chemical space, which are usually more helpful for prediction tasks. The random strategy has the lowest performance because it does not make use of any information about the samples. In addition, the four remaining active learning strategies outperform the random strategy but perform slightly worse than muTOX-AL.

To further illustrate that active learning can significantly reduce labeling costs, we set the test performance threshold to 95% of the supervised learning performance using all training samples (5988 samples), comparing the sample sizes used by different active learning strategies at this threshold. For example, if the supervised learning accuracy using all samples is 84.76%, the performance threshold is 80.52%. From

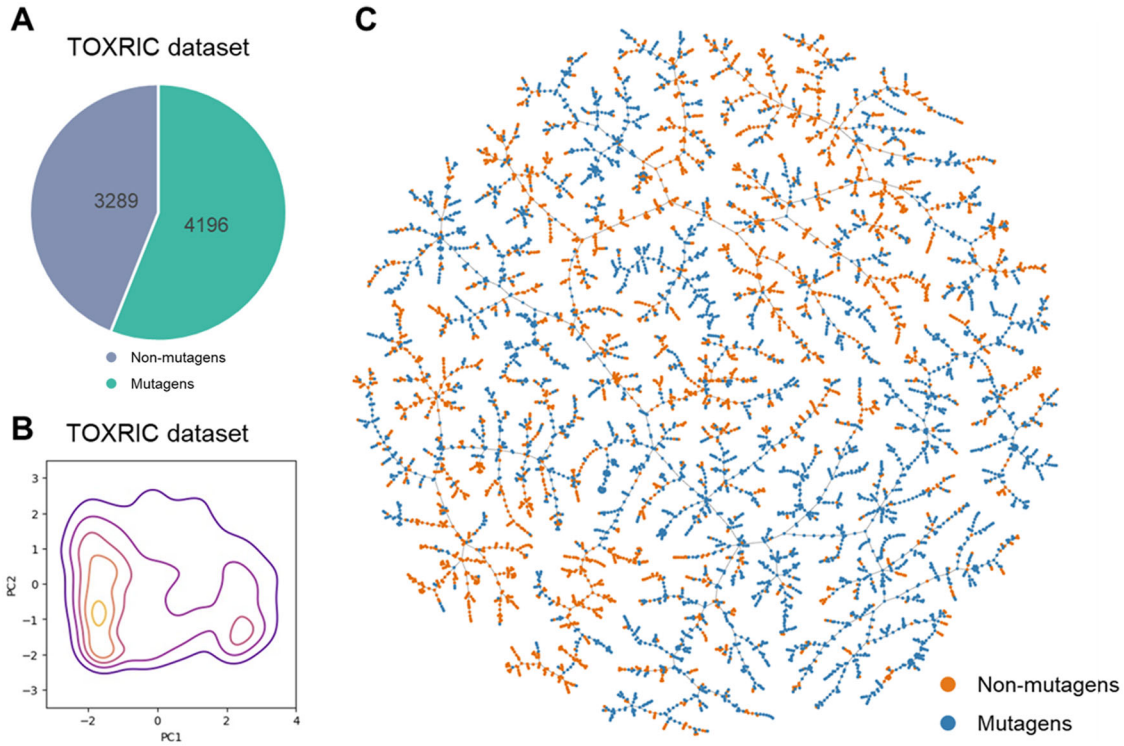

Fig. 3 | Statistical analysis of the TOXRIC dataset. A Label distribution of TOXRIC dataset. **B** Principal component analysis density for the TOXRIC dataset. **C** The TOXRIC dataset uses tree-based molecular similarity visualizations represented by LSH and MSP.

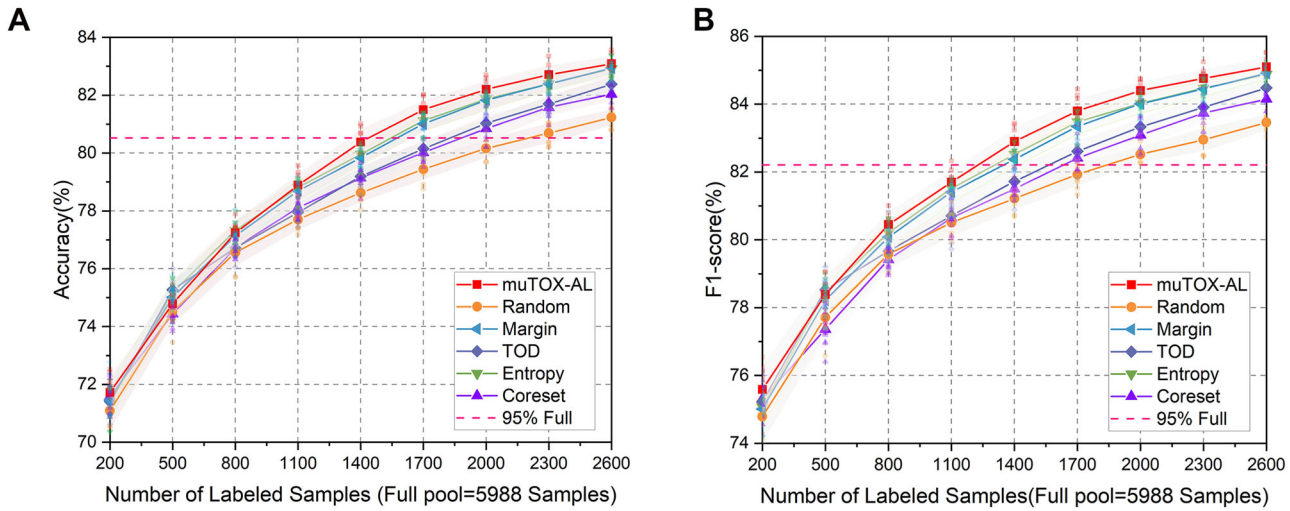

Fig. 4 | Active learning results of mutagenicity classification of muTOX-AL and five active learning baselines. A Accuracy of the six algorithms in different active learning cycles. **B** F1-score of the six algorithms in different active learning cycles. The red line represents the muTOX-AL algorithm proposed in this paper, the dark blue line represents the TOD active learning strategy, the purple line represents the Core-set active learning strategy, the blue line represents the margin-based active learning strategy, the green line represents the entropy-based active learning strategy, and the orange line represents random query strategy. The figure reports the mean and standard deviation (shaded regions) from ten repeated experiments. Source data for the curves are shown in Supplementary Data.

Table 1, we can see that our approach uses the least number of samples on most metrics when reaching the same performance threshold. The random query strategy always uses the largest number of molecules, indicating that random exploration in chemical space tends to enter relatively less informative regions and thus requires more annotation costs than active learning strategies. In particular, the last row of the table calculates the percentage of samples reduced by using the active learning strategy of this paper compared to the random query strategy. We can see that active exploration of the chemical space using the active learning strategy can reduce the training sample size by up to nearly 57%. This result is promising and demonstrates that using active learning methods can significantly reduce the number of samples that need to be labeled while maintaining competitive performance. In addition, our backbone module can achieve competitive test performance on a variety of evaluation metrics compared to several machine learning and deep learning baselines when using all training samples (see Supplementary Table 7).

In addition, to demonstrate the generalization performance of muTOX-AL on external data, we additionally collected mutagenicity data

Table 1 | Comparison of the sample sizes used to achieve 95% of the supervised learning performance in different active learning methods (i.e., performance obtained by training with all training samples in the dataset)

	Accuracy	F1-score	AUC	Precision	Recall	Specificity
muTOX-AL	**1438**	**1228**	1727	**1505**	**701**	1963
Random	2204	1842	1853	1970	1633	>2600
Margin	1574	1350	1835	1560	1039	1944
TOD	1826	1565	1868	1843	1231	2398
Entropy	1545	1308	1835	1578	1009	**1939**
Coreset	1881	1634	**1622**	1720	1602	2174
Reduced sample size (%)	34.8%	33.3%	6.8%	23.6%	57.1%	>24.5%

"Reduced sample size" indicates the percentage reduction in the number of samples in muTOX-AL compared to the random strategy. The bold text in the table indicates the best performance.

from Li et al.[21] as an external testing set (Li's Dataset), excluding samples identical to those in the TOXRIC dataset. The performance of two models was evaluated: (i) a model trained using the active learning algorithm to select the most informative 2600 TOXRIC samples (trained and saved when the "Number of Labeled Samples = 2600" in Fig. 4); (ii) a model trained using all 5988 TOXRIC samples (i.e., the four subsets of the 7485 samples) without active learning (Supplementary Table 7). The performance of the above two models on Li's dataset is shown in Supplementary Fig. 5 and Supplementary Table 8. It can be found that the performance of the model trained using only 2600 TOXRIC samples on Li's dataset is close to the performance of the model trained using all 5988 TOXRIC samples. It is encouraging and confirms that active learning can be instructive for data collection and selection. In particular, with an uncertainty sampling strategy in the presence of a large chemical space, we are able to achieve competitive performance by selecting only a small number of samples, which greatly reduces human and resources consumption due to exhaustive random screening. In terms of Precision and Specificity, the model trained with the 2600 samples selected by muTOX-AL even outperforms the model trained with all 5988 samples, which is consistent with the findings in the current research[22]. This indicates that not all samples have a positive effect on the model performance and proves that active learning is able to select the most helpful samples for model training and remove the redundant samples effectively.

The uncertainty estimation module designed in muTOX-AL is more helpful in selecting informative samples

This section attempts to explore the effect on the ability to select informative samples when the feature dimension or network depth of the uncertainty estimation module is increased. Supplementary Fig. 6 shows the schematic diagram of the two uncertainty estimation modules. Supplementary Fig. 6A displays the structure of two different hidden layer features which are used as input to the uncertainty estimation module, while Supplementary Fig. 6B presents the structure in that only a single hidden layer feature is used, but with an additional linear layer and a Rectified Linear Unit (ReLU) activation function.

Figure 5 illustrates the active learning curves for the two structures described above (source data for the curves and *p*-values calculated by independent *t*-tests are shown in Supplementary Tables 9, 10). It can be observed that the performance is slightly decreased while the two structures shown in Supplementary Fig. 6 are used. It is also demonstrated that the uncertainty estimation module designed in muTOX-AL can be more helpful in selecting informative samples. There are two possible explanations for this: (i) a simple network may be more effective in improving performance when the dataset is not large enough; (ii) increasing the complexity of the uncertainty estimation module when the backbone module network is not deep may not be conducive to the joint training of the models.

Furthermore, we conducted ablation studies on the input features of our model, with detailed results provided in Supplementary Fig. 3.

muTOX-AL can select samples that are closer to the classification boundary

To further confirm the effectiveness of our deep active learning strategy, we first use the embeddings of data points before the classifier ("Embeddings" shown in Fig. 2) to be downscaled by t-SNE visualization and display it in a two-dimensional map. Figure 6A shows the visualization of both muTOX-AL and the random query strategy.

In Fig. 6A, the boundary between positive and negative samples becomes more apparent as the cycle progresses, indicating that with the sample size increasing, the trained model can distinguish between positive and negative samples more evident. In addition, the data distribution of the proposed method is more concentrated, and the boundary is clearer than that of the random query strategy, which implies that muTOX-AL has a better prediction performance. Additionally, in each cycle, the samples selected using the random query strategy are scattered throughout the embedding space, whereas the samples selected by the proposed muTOX-AL are more concentrated on the classification boundary, which suggests that muTOX-AL can successfully select the samples with the highest uncertainty, i.e., those samples that are more difficult for the model to classify. To further quantify the distance of the sample from the classification boundary, we train a support vector machine (SVM) classifier based on the Radial Basis Function kernel function using the same embeddings mentioned above and calculate the distances of the selected samples from the SVM classification hyperplane in each cycle. As the distances calculated by SVM are signed values, we take absolute values for them, plot violin plots and use *t*-tests to obtain *p*-values for both distributions. The results are shown in Fig. 6B, with the x-axis representing muTOX-AL and random strategy and the y-axis representing the absolute values of the distances. Figure 6B demonstrates that the *p*-values are consistently less than 0.05 in different cycles, indicating that the distribution of the distances is significantly different between the two methods and that muTOX-AL picks samples with smaller distances.

In total, muTOX-AL tends to select samples closer to the classification boundary. By preferentially selecting such samples and feeding them into the model for training, this strategy allows the model to fully learn the distribution of features of the most informative samples, thus achieving competitive performance with a small size of the dataset.

muTOX-AL has high molecular structural discriminability

Generally, samples with lower molecular similarity are considered to have greater uncertainty and vice versa[39]. However, this need not always be the case, small molecules with similar structures may have opposite properties, and a typical example is chiral molecules[40,41]. For instance, thalidomide, which has both levorotatory and dextrorotatory structures, the levorotatory body is therapeutic and can reduce early pregnancy reactions, but its chiral partner (the dextrorotatory body) is mutagenic[42]. Therefore, a model is considered to have high structural discriminability when the samples selected by the model satisfy both of the following conditions: (i) low structural similarity between the selected samples; (ii) the presence of molecules with similar structures but opposite properties.

First, to demonstrate muTOX-AL is able to select samples with low structural similarity, we analyzed the structural similarity among the samples selected by each active learning cycle. Specifically, we used the Tanimoto coefficient[43] to calculate the structural similarity between the 300 samples selected in each cycle. For each cycle, a heatmap of the selected samples was plotted (Fig. 7A). To directly compare the difference in structural similarity between the selected structures from different cycles, the distribution of structural similarity is shown in the box plots (Fig. 7B).

Figure 7A, B shows that the small molecules with higher structural similarity are selected in early cycles. We argue that this phenomenon may be related to the samples selected in the initial cycles being few and the power of deep learning is not sufficiently exploited, which echoes the slightly worse

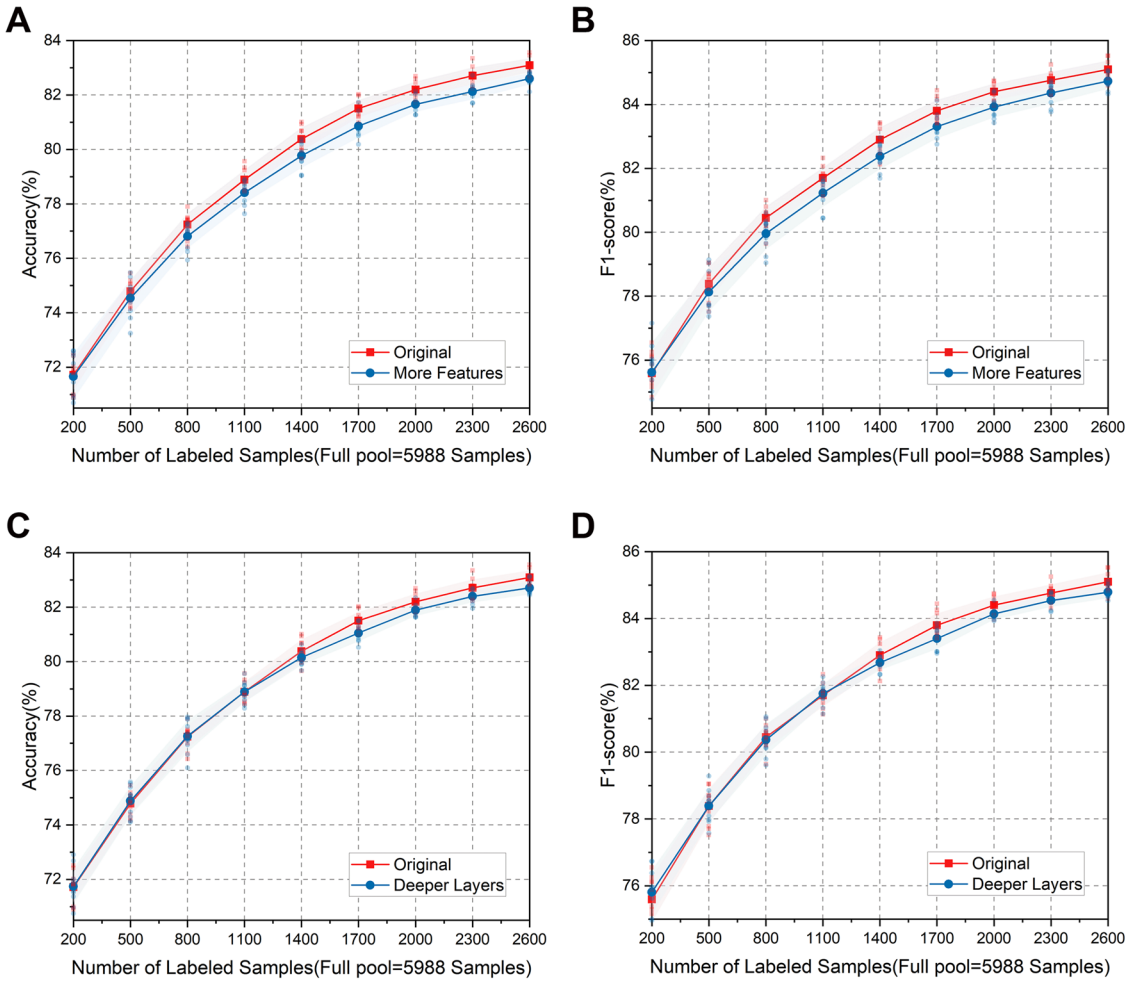

Fig. 5 | Accuracy and F1-score in different active learning cycles using two uncertainty estimation modules, where "Original" denotes the original structure in Fig. 2. A, **B** "More Features" indicates the use of two hidden layer features of the backbone module, and **C**, **D**) "Deeper Layers" indicates the addition of a linear layer and a ReLU activation function. The figure reports the mean and standard deviation (shaded regions) from ten repeated experiments. Source data for the curves and *p*-values calculated by independent *t*-tests are shown in Supplementary Tables 9, 10 and Supplementary Data.

performance of our model in the initial cycles. As the number of cycles increases, the performance of the model increases rapidly, which indicates that muTOX-AL tends to select samples with lower structural similarity, i.e., molecules with higher uncertainty.

Second, to illustrate that muTOX-AL is able to selects samples with similar structures but opposite properties. Figure 8 and Supplementary Table 11 show the sets of molecules with such properties selected by each active learning cycle. As can be seen in Fig. 8, these sets of molecules include isomers, chiral molecules, and molecules with varying numbers of substituents, etc. These small variances often result in a polarity change of the molecule, which leads to contrary properties. However, for feature extraction methods based on molecular structure, there is low feature differentiation between these structurally similar molecules; in other words, the uncertainty of such molecules is significant.

In addition, to demonstrate that the muTOX-AL can still select samples with similar structures but opposite properties in the out-of-domain case (outside of the TOXRIC dataset), we collected 50 molecules from the literature, which are not present in the TOXRIC dataset (see Supplementary Data). Subsequently, the uncertainty of each molecule was obtained using the trained muTOX-AL (the model after the 9th iteration of active learning), and these samples were ranked according to the uncertainty value. Finally, we selected the top 30% of the ranked samples to find the set of molecules with similar structures but opposite properties among them, and the results are displayed in Fig. 9. Therefore, muTOX-AL can find samples with similar structure but opposite properties, which are also the ones with high uncertainty.

Overall, although the model tends to select samples with low structural similarity, samples with similar structures but different properties are still considered, and this avoids the over-fitting of simple relationships. This is the evidence of the high molecule structural discriminability of muTOX-AL.

Discussion

This paper presents a deep active learning framework, muTOX-AL, for molecular mutagenicity prediction. In particular, our framework is straightforward but efficient, and it has high structural discriminability. In this research, we exploit multiple features of molecules and design an uncertainty estimation module to calculate the uncertainty of all samples, achieving competitive classification performance by selecting a few samples with the highest uncertainty scores. In particular, our approach is able to reduce the sample size by up to 57% compared to the random query strategy. Ablation experiments on input features and module design demonstrate the soundness of our method. In addition, we use t-SNE to visualize the projection of embeddings on a two-dimensional space and use SVM to calculate the distances of the selected samples from the classification boundary, demonstrating that muTOX-AL can select the most informative samples. Then, muTOX-AL exhibits remarkable structural discriminability by both selecting molecules with low structural similarity to prevent overfitting of simple relationships and considering samples with similar structures but different properties to provide a more comprehensive understanding of the complex relationship between molecular structure and mutagenicity.

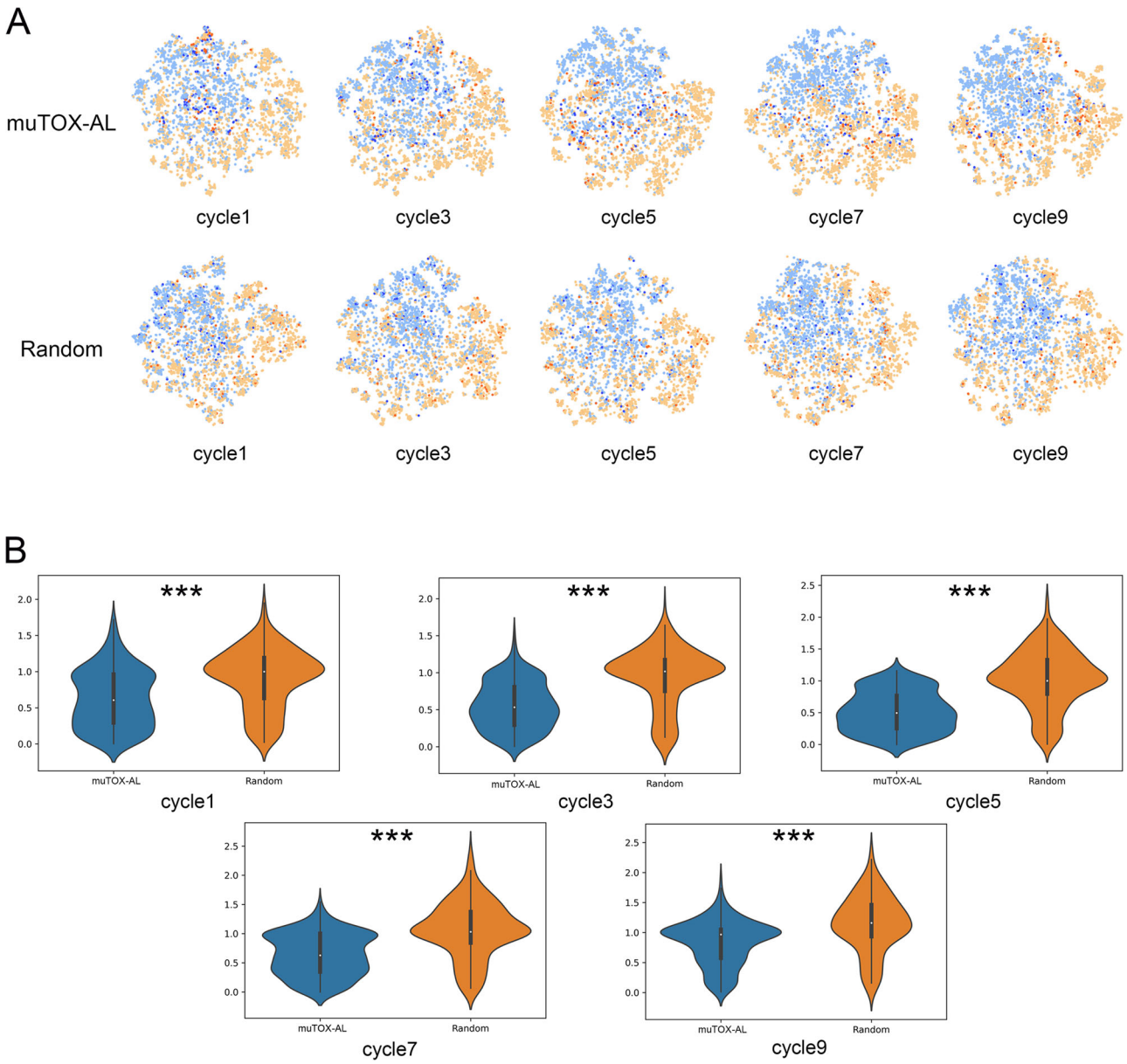

Fig. 6 | *t*-SNE visualization results of the samples selected by muTOX-AL and random sampling strategy. A t-SNE visualization for muTOX-AL and random strategy in five active learning cycles. The dark dots indicate the positive (dark yellow) and negative (dark blue) samples selected by the model after the active learning cycle. The lighter dots indicate all positive (light yellow) and negative (light blue) samples outside the selected samples in the training set. **B** Violin plot of the distance distribution from the SVM classification hyperplane of the selected samples in five active learning cycles. Asterisks indicate independent *t*-test *p*-value ***$p \le 0.001$, where cycle1 has a *p*-value of 1.32e-16, cycle3 has a *p*-value of 1.24e-15, cycle5 has a *p*-value of 1.53e-48, cycle7 has a *p*-value of 2.48e-31 and cycle9 has a *p*-value of 1.02e-18. Source data for the figure are shown in Supplementary Data.

Besides the effectiveness, the ease of use and expansibility can also make muTOX-AL more widely used in practice.

There are many promising directions to investigate as future works. First, there is an issue that the large majority of the tasks in drug development suffer, which is that the labeled data are often difficult to obtain. Active learning provides opportunities to address these questions collectively via application to the whole chain of drug discovery, such as drug combinations, drug sensitivity, and drug-target interaction prediction. Second, beyond the field of drug discovery, our method can also be extended to the task of predicting the mutagenicity of industrial chemicals. Addressing the issue of class imbalance in industrial chemical data and effectively incorporating active learning will be a promising direction. Third, better molecular representation can improve the performance of deep learning. How to combine representation learning with active learning is an area of future interest. Fourth, the generalization ability of models is one of the current research challenges in many fields. Therefore, combining active learning with advanced transfer learning strategies can better address the issue of generalization. Finally, there is an urgent need for further research to develop more suitable active learning algorithms for compounds[44]. For instance, designing active learning algorithms that consider specific structures affecting molecular properties (structural alerts). We believe that active learning can better help drug discovery and other fields in the future.

Methods

Datasets

TOXRIC dataset[33]. The raw data used in this study were the C. Xu's Ames data collection[45], which is one of the commonly used data sets for developing the prediction models. The entire database was prepared as follows[46]. Firstly, any inorganic molecules, that is, those without carbon

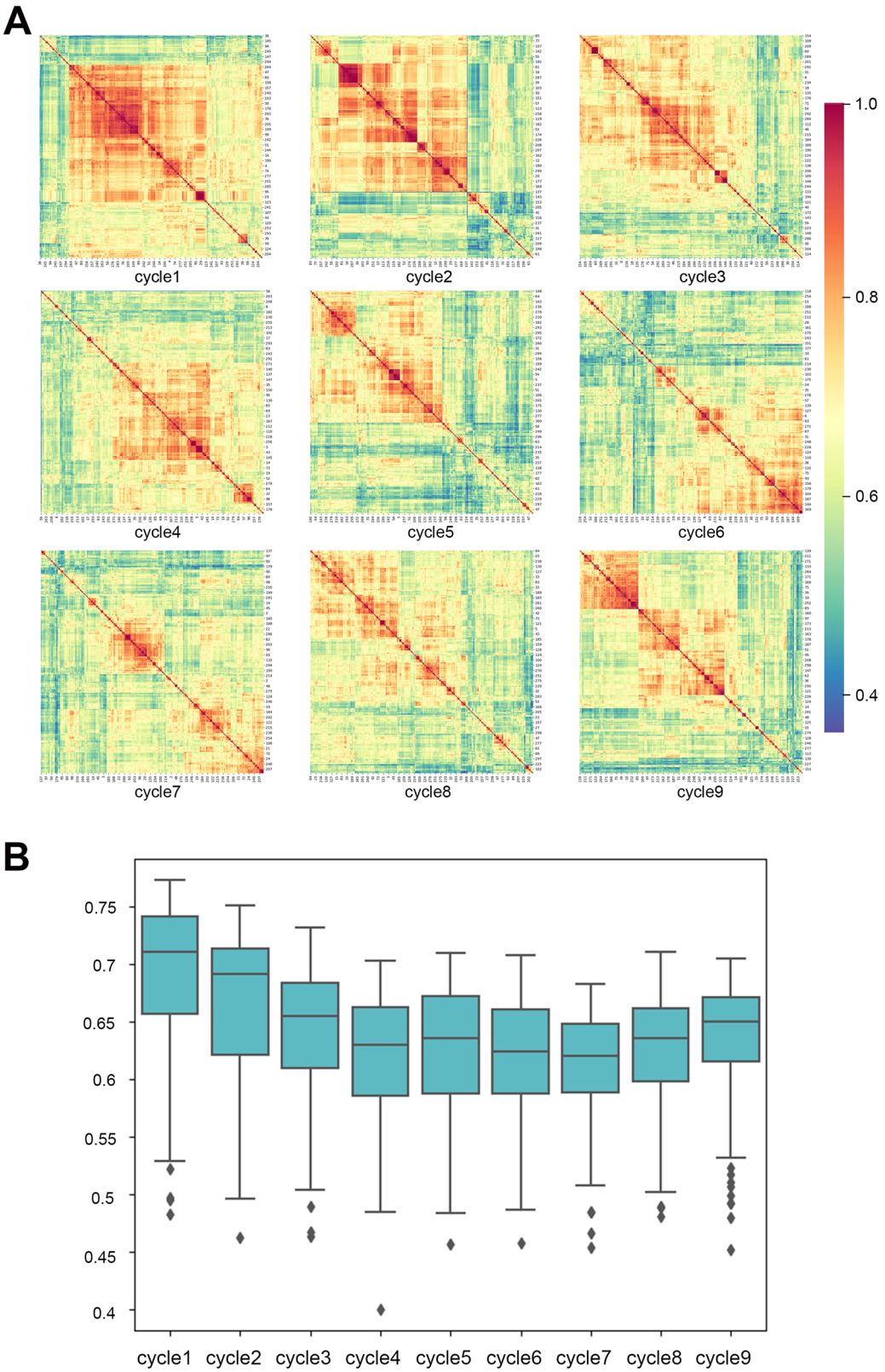

Fig. 7 | Structural similarity analysis of the samples selected by muTOX-AL. A The heatmap of structural similarities between the selected samples for each cycle. Darker blue means higher structural similarity, and darker red means lower structural similarity. **B** the distribution of structural similarity for each cycle. Source data for the figure are shown in Supplementary Data.

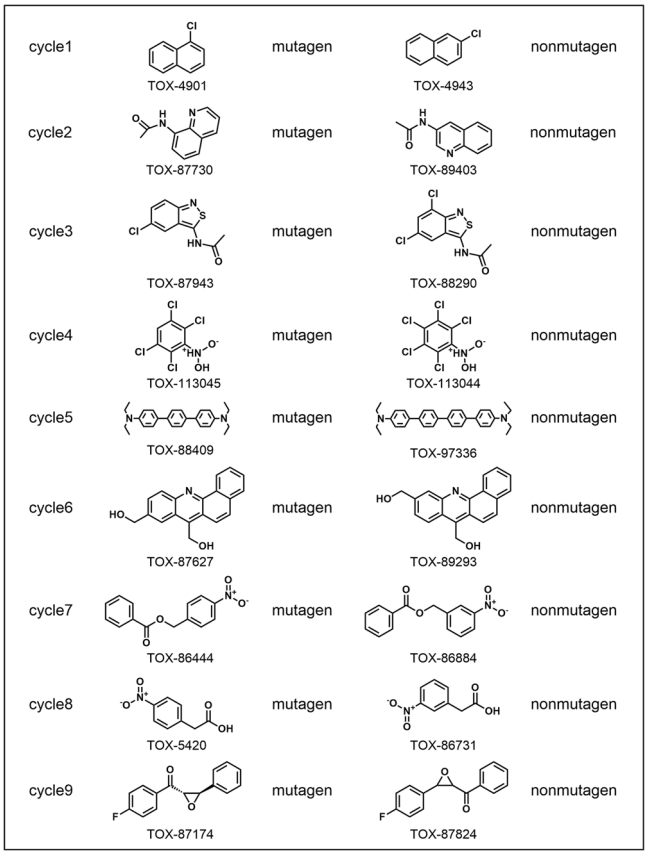

Fig. 8 | Molecules with high structural similarity and opposite mutagenicity labels. These samples were selected by muTOX-AL because of their high uncertainty.

atoms within the structure are removed. Secondly, the molecules with unspecified stereochemistry were removed. Thirdly, the molecules were standardized using the InChI key[47]. Finally, duplicates were identified and removed using the InChI key across the data collection. Ultimately, in total, 7485 compounds were used for the model building. The data sets contained 4196 mutagens and 3289 non-mutagens. The dataset is available in https://toxric.bioinforai.tech/home.

Li's dataset[21]. The dataset was constructed from data sourced from three distinguished databases: the Chemical Carcinogenesis Research Information System, the National Toxicology Program, and the Instituto Superiore di Sanita for Salmonella Typhimurium. Further refinement was achieved by removing samples that were duplicates of those found in the Ames dataset, thus establishing a new, independently verified external test set. Statistical analysis of the dataset was shown in Supplementary Fig. 2.

Molecular descriptors and fingerprint features

The fingerprint features include three sets of topological path-based features (Extended Connectivity Fingerprints with a diameter of 2, 4, and 6, ECFP2, ECFP4, and ECFP6) and one set of substructure-key SMARTS-based features MACCS.

ECFP fingerprints are generated based on the connectivity between atoms in a molecule, taking into account the bonds, hybridization states, and functional groups. ECFP fingerprints use circular fingerprints, where the radius of the circle defines the maximum distance between atoms that can be included in a particular substructure. According to different radius, ECFP is divided into ECFP2, ECFP4, and ECFP6.

MACCS fingerprints are generated based on the structural features of a molecule, such as the presence of aromatic rings, functional groups, and atom types. Each structural key in the fingerprint is assigned a binary value, where 1 indicates the presence of the key and 0 indicates the absence of the key.

RDKit2D descriptors are selected as input features to complement the fingerprint features. RDKit2D descriptors can provide information on a wide range of molecular properties, including size, shape, polarity, and flexibility. Some examples of RDKit2D descriptors include the number of atoms, the molecular weight, the number of rotatable bonds, and the number of hydrogen bond donors and acceptors.

Deep active learning strategy

Active learning aims at selecting the most informative samples from a pool of unlabeled samples in the entire sample space. Defining the amount of information in a sample is the biggest challenge in the active learning problem. To describe the deep active learning scenario proposed in this paper, an unlabeled sample pool consisting of N unlabeled molecules is assumed to be $\mathrm{U}_N = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where x_i is the feature of the molecule and y_i is the toxicity label corresponding to x_i. First, we randomly select M samples from U_N and give them to the oracle for annotation, which generates an initial pool of annotated samples $\mathcal{L}_M = \{(x_1, y_1), \ldots, (x_M, y_M)\}$. Then the five features of all samples in the initial annotated sample pool $\mathcal{L}_M$ are extracted and input to the backbone module f_b. The output of the hidden layer of the network, which can be considered the embeddings of the input features, is fed into the uncertainty estimation module f_u. The model's parameters are updated by jointly optimizing f_b and f_u according to the defined total loss.

Framework architecture

Feature extraction module. Molecular fingerprints and molecular descriptors are widely used in similarity searching and classification. Four molecular fingerprints and one molecular descriptor are used in this work. They are ECFP2, ECFP4, and ECFP6 (2048 bits), MACCS keys (MACCS, 166 bits) and RDKit2D. All the fingerprints and molecular descriptors were calculated by the RDKit python package.

Backbone module. As shown in Fig. 2, considering the higher dimensionality of extended connectivity fingerprints compared to other features, we first stitch ECFP2, ECFP4, and ECFP6 in the channel dimension to form a three-channel fusion feature, which is then fed into two convolution blocks. In each convolutional block, a 1D convolutional layer and an average pooling layer are used first to extract features and remove redundant information, thus reducing the parameters of the network. Then the ReLU activation function follows, which introduces a nonlinear element to enhance the representation ability of the network and mitigate the problems of gradient disappearance and gradient explosion. With the two convolution blocks mentioned above, it is possible to further extract features while reducing the dimensionality, which helps in the subsequent classification steps. The output of the convolution block is stitched with the lower dimensional MACCS fingerprints and RDKit2D descriptors to achieve feature fusion. The fused features are then fed into a linear block consisting of a linear layer, a ReLU layer and a Dropout layer, where the Dropout layer is used to prevent overfitting of the input. Finally, a linear classifier is used to classify the mutagenicity of the molecule.

Uncertainty estimation module. In active learning, the key issues are the criteria for measuring the informativeness of the samples and the design of the query module. For the first problem, the most commonly used measure is uncertainty-based querying, i.e., querying the samples that are most difficult for the model to classify. Uncertainty-based querying has been shown to be more applicable in classification problems with small samples[48], so we choose uncertainty as the measure of informativeness. In deep learning, the loss is often used as a measure of the difference between the predicted and true values of a model. The samples with the largest losses can usually be regarded as the samples

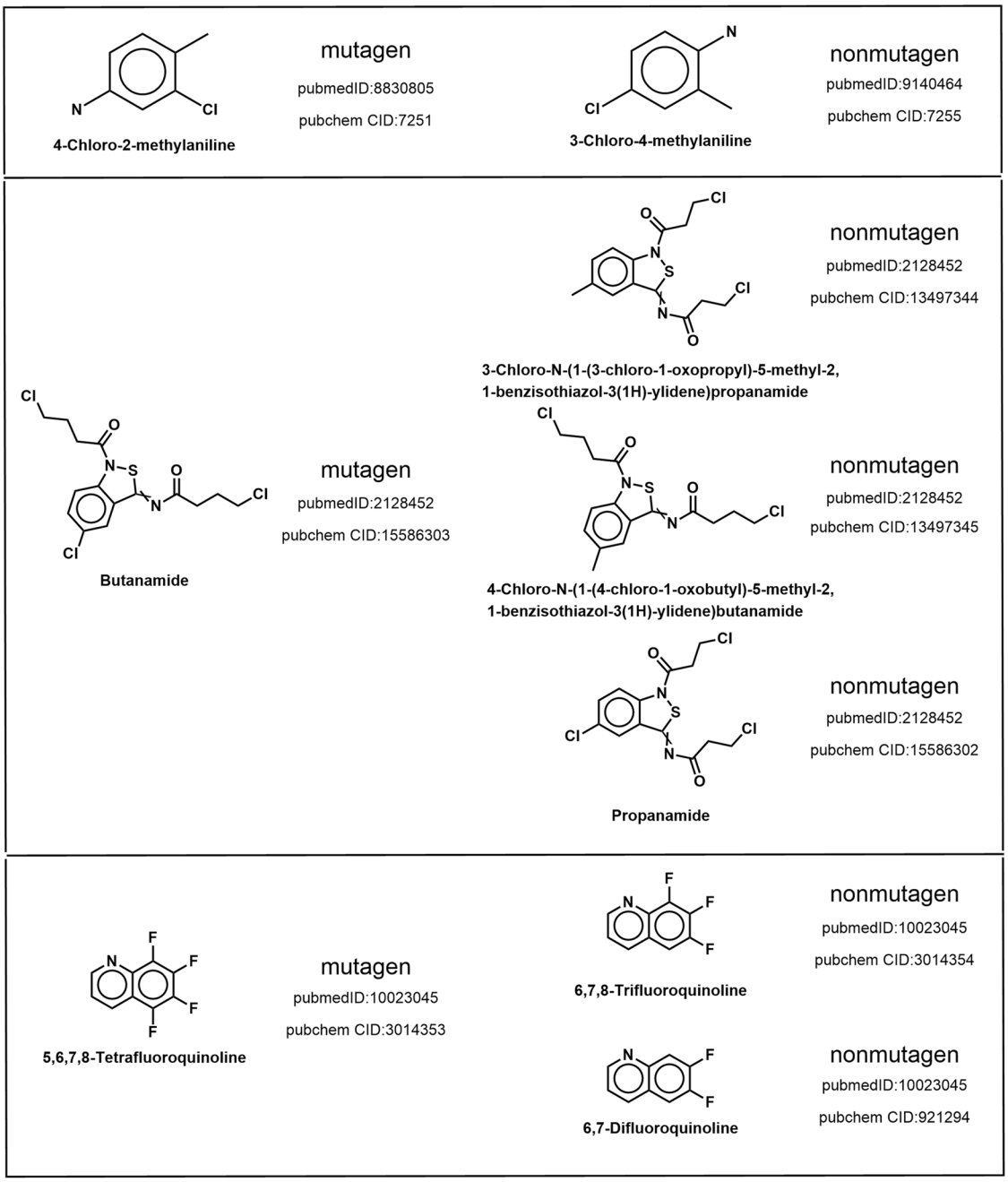

Fig. 9 | Molecules with high structural similarity and opposite mutagenicity labels reported in the literature. These samples were selected by muTOX-AL because of their high uncertainty. Literature reporting the mutagenicity of the molecules as well as the pubchem CID of the molecules are given. The detailed information can be found in Supplementary Data.

which are the hardest for the model to distinguish. Therefore, the uncertainty estimation can be converted into the loss estimation. Since the model loss values cannot be computed for samples without true labels, a module needs to be designed to estimate the loss values for unlabeled samples. By training an uncertainty estimation module using labeled samples, we can predict the loss of unlabeled samples and thus estimate their uncertainty. The uncertainty estimation module designed in this paper is shown in Fig. 2. To make good use of features extracted by the hidden layer of the backbone module, we use it as input to the uncertainty estimation module. Inspired by Yoo et al.[49], the module consists of a global average pooling layer, two linear layers and a ReLU layer, where the global average pooling layer aims to integrate feature information, and the introduction of linear and non-linear activation layers enables the network to learn better. A final linear layer maps the features into a scalar that outputs the uncertainty scores of unlabeled samples. We did not use more scaled hidden features, as this could have led to a more complex structure of the uncertainty estimation module, which would decrease the prediction performance. We confirmed this view in Results.

Loss calculation module. Having defined the structure of the backbone module f_b and the uncertainty estimation module f_u, we need to focus on how they are jointly optimized. The total loss of the modules L_{total} consists of two main components: the backbone module loss L_b and the uncertainty estimation module loss L_u, which will be described separately below.

The output of a labeled sample x after the backbone module is $\hat{y} = f_b(x)$. In the binary classification task, we usually use binary cross-

entropy loss. It is

$$L_b(\hat{y}, y) = -(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})) \quad (1)$$

We want the output of the uncertainty estimation module to be as close as possible to the binary cross-entropy loss of the sample, so the uncertainty estimation task can be considered a regression task. In usual regression tasks, the most used metric is the mean squared error (MSE) $L_u(\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}-y)^2$, but the scale of loss changes as the training progresses, so using MSE as the loss is not a sensible choice. Here, we determine the trend of uncertainty score estimation by comparing the losses of a pair of samples within a mini-batch. Assuming that the k th pair of samples (x_i, y_i) and (x_j, y_j) in the same mini-batch in the sample pool, their outputs after the uncertainty estimation module are $\hat{l}_i$ and $\hat{l}_j$, and the actual cross-entropy losses are l_i and l_j, we can define the loss for this pair of samples after the uncertainty estimation module as

$$L_u\left(\hat{l}^k_{\text{batch}}, l^k_{\text{batch}}\right) = \max(0, -\text{sign}(l_i - l_j) \cdot (\hat{l}_i - \hat{l}_j) + \xi) \quad (2)$$

where sign(·) is the sign function, margin ξ is a very small number. Equation(2) indicates that when $l_i - l_j$ and $\hat{l}_i - \hat{l}_j$ have the same sign, i.e., the loss of a pair of samples shows the same trend, the value of L_u is zero. Otherwise, the parameters of the uncertainty estimation module need to be updated by gradient descent.

Thus, given the size B of the mini-batch, the total loss of the two modules can be defined as

$$L_{\text{total}} = \frac{1}{B}\left(\sum_{(x,y)\in B} L_b(\hat{y}, y) + 2\lambda \cdot \sum_{(x^k,y^k)\in B} L_u(\hat{l}^k_{\text{batch}}, l^k_{\text{batch}})\right) \quad (3)$$

By optimizing the total loss L_{total}, we can jointly optimize the parameters of the backbone module and the uncertainty estimation module during the training process, thus estimating the uncertainty of unlabeled samples during the active learning phase. Algorithm 1 is elaborated to the algorithm logic and conceptual modeling.

Algorithm 1. The muTOX-AL framework for molecular mutagenicity prediction

Input:
unlabeled pool U
The testing set $\mathcal{T}$
The number of initialized label set M
The number of active learning cycles C
The number of samples labeled in each cycle K
The backbone module f_b, The uncertainty estimation module f_u
1: Randomly select M samples from U to gain initialized labeled set $\mathcal{L}$
2: **For** c **in** C:
3: Train the backbone module f_b and the uncertainty estimation module f_u using $\mathcal{L}$
4: Evaluate the performance on f_b using the testing set $\mathcal{T}$
5: Estimate the uncertainty of the unlabeled samples U by f_b and f_u
6: Select the top K samples with the highest uncertainty
7: Query their labels from the oracle to obtain $\mathcal{L}_K$
8: $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_K$
9: $c \leftarrow c + 1$
10: **End**

Evaluation metrics

Two commonly evaluation metrics in classification tasks are used as evaluation criteria: accuracy and F1-score. First, we define four indicators: True Positive (TP) means that the positive sample has a positive predictive value and the prediction is correct; True Negative (TN) means that the negative sample has a negative predictive value and the prediction is correct; False Positive (FP) means that the positive sample has a positive predictive value and the prediction is wrong; False Negative means that the predicted value of the negative sample is negative and the prediction is correct.

Accuracy represents the proportion of samples correctly predicted to all samples and is the most common evaluation metric in classification tasks and is defined as

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (4)$$

F1-score is defined in Eq. (5), which combines the Precision and Recall metrics.

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (5)$$

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (6)$$

$$\text{Recall} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (7)$$

Specificity refers to the proportion of TN samples that are correctly predicted as negative by the model, i.e., the probability of TN samples being correctly predicted as negative.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (8)$$

Experimental settings

All our experiments are implemented in the PyTorch framework. We set the batch size as 128 and used a 5-fold cross-validation to increase the generalizability of the experimental results. We mainly followed the training strategy in the active learning setup, following a five-fold cross-validation strategy, where all the dataset was randomly divided into five subsets, and in each fold, four of them were selected as the total training set for the model, and the remaining one subset was used to test the model's performance (the test set), which is not visible at all times. The whole active learning process is divided into nine cycles. At the beginning of the experiment (cycle = 0), we randomly select 200 samples from the unlabeled sample pool to train the initialized network and select 300 samples from the unlabeled pool in each active learning cycle. The backbone module is trained jointly with uncertainty estimation module. Separately, the backbone module is trained with 300 epochs, using an SGD optimizer with a learning rate of 5e-3, a momentum parameter of 0.9 and a weight decay parameter of 5e-4. The uncertainty estimation module is trained using an Adam optimizer with a learning rate of 8e-3. The margin in Eq. (2) is set to one. For each method, ten randomized replicate experiments are conducted using different initial labeled samples, and we report the mean of the ten experiments at the end. Detailed information on active learning training strategies can be found in the "Active learning training strategies in muTOX-AL" section of the Supplementary Information.

Active learning methods for comparison

We have compared muTOX-AL with the following five active learning methods.

Random strategy. The random strategy is the most common active learning baseline. In each active learning cycle, the K samples are selected randomly from the unlabeled pool and given to the oracle for annotation.

Margin-based active learning strategy[35]. The margin-based active learning strategy is a uncertainty-based method. It defines the

uncertainty by measuring the difference between the prediction probabilities of different categories. The K samples with the lowest margin are added to the labeled pool. The margin is defined as

$$X = \arg\min_{x\in U}(P(\hat{y}_1|x) - P(\hat{y}_2|x)) \quad (9)$$

Entropy-based active learning strategy[36]. The entropy-based active learning strategy is an uncertainty-based method. In information theory, the uncertainty of the data is higher if it has a higher entropy. Therefore, the entropy of the unlabeled samples is calculated and ranked. The K samples with the highest entropy are added to the labeled pool. The entropy is defined as

$$X = \arg\max_{x\in U} E_x = \arg\min\left\{\sum_{i=1}^{Y} P(y_i|x) \times \log P(y_i|x)\right\} \quad (10)$$

TOD active learning strategy[37]. The temporal output discrepancy TOD active learning strategy based on temporal output discrepancy is an uncertainty-based approach. It defines the uncertainty by calculating the discrepancy of model output at different active learning cycles.

Core-set active learning strategy[38]. The Core-set active learning strategy is a diversity-based approach which is also a common baseline in active learning trying to find a core set that makes the model's performance on the core set and the whole dataset as close as possible.

Machine learning-based mutagenicity prediction methods for comparison

MIL. Feeney et al.[19] propose a machine learning approach based on multi-instance learning for molecular mutagenicity prediction, particularly for metabolically activated compounds like aromatic amines. By grouping metabolites and their parent compounds under a single mutagenicity label, MIL circumvents the need for individual labels, capturing the mutagenic potential through structural considerations. MIL achieved excellent performance on the mutagenicity molecular dataset, so we used it as one of the baselines for muTOX-AL.

Enhanced_Representation_Mutagenicity. Shinada et al.[18] systematically considered and evaluated combinations of structures and molecular features that have the greatest impact on model accuracy, using various classification models (including classic machine learning and deep learning models) to assess these features. We selected Structural Representation, Molecular Descriptors, and Genotoxicity Descriptors features, with the Random Forest classifier as our evaluation baseline.

Statistics and reproducibility

The study employed five fold cross-validation with ten random repetitions, reporting the mean and standard deviation across these repetitions. The p-values reported in the study were calculated using independent t-tests.

Reporting summary

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

Data availability

The TOXRIC dataset is available in https://toxric.bioinforai.tech/home[33]. All the datasets can be found in Github: https://github.com/Felicityxuhy/muTOX-AL and Figshare: https://doi.org/10.6084/m9.figshare.26379805. All data supporting the findings of this paper are available within the paper and Supplementary Data.

Code availability

The code for this work can be found in Github: https://github.com/Felicityxuhy/muTOX-AL and Figshare: https://doi.org/10.6084/m9.figshare.26379805.

Received: 9 October 2023; Accepted: 21 August 2024;
Published online: 31 August 2024

References

1. Ames, B. N., Lee, F. D. & Durston, W. E. Improved bacterial test system for detection and classification of mutagens and carcinogens. *Proc. Natl Acad. Sci. USA* **70**, 782–786 (1973).
2. Mortelmans, K. & Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* **455**, 29–60 (2000).
3. Kramer, J. A., Sagartz, J. E. & Morris, D. L. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat. Rev. Drug Discov.* **6**, 636–649 (2007).
4. Kumar, A., Dobrovolsky, V. N., Dhawan, A. & Shanker, R. *Mutagenicity: Assays and Applications*. (Academic Press, 2017).
5. Townsend, P. A. & Grayson, M. N. Density functional theory in the prediction of mutagenicity: a perspective. *Chem. Res. Toxicol.* **34**, 179–188 (2020).
6. Timperio, A. M., Kuiper, H. A. & Zolla, L. Identification of a furazolidone metabolite responsible for the inhibition of amino oxidases. *Xenobiotica* **33**, 153–167 (2003).
7. Ames, B. N., Bartsch, H., Miller, J. A. & Gurney, E. G. Carcinogens as frameshift mutagens - metabolites and derivatives of 2-acetylaminofluorene and other aromatic amine carcinogens. *Proc. Natl Acad. Sci. USA* **69**, 3128–3132 (1972).
8. Ames, B. N., Durston, W. E., Yamasaki, E. & Lee, F. D. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. *Proc. Natl Acad. Sci. USA* **70**, 2281–2285 (1973).
9. McCann, J., Spingarn, N. E., Kobori, J. & Ames, B. N. Detection of carcinogens as mutagens: bacterial tester strains with R factor plasmids. *Proc. Natl Acad. Sci. USA* **72**, 979–983 (1975).
10. Galati, S. et al. VenomPred: a machine learning based platform for molecular toxicity predictions. *Int. J. Mol. Sci.* **23**, 2105 (2022).
11. Hillebrecht, A. et al. Comparative evaluation of in silico systems for ames test mutagenicity prediction: scope and limitations. *Chem. Res. Toxicol.* **24**, 843–854 (2011).
12. Yang, X., Zhang, Z., Li, Q. & Cai, Y. Quantitative structure–activity relationship models for genotoxicity prediction based on combination evaluation strategies for toxicological alternative experiments. *Sci. Rep.* **11**, 8030 (2021).
13. Hansen, K. et al. Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **49**, 2077–2081 (2009).
14. Chu, C. S. M., Simpson, J. D., O'Neill, P. M. & Berry, N. G. Machine learning-Predicting Ames mutagenicity of small molecules. *J. Mol. Graph. Model.* **109**, 108011 (2021).
15. Fradkin, P. et al. A graph neural network approach for molecule carcinogenicity prediction. *Bioinformatics* **38**, i84–i91 (2022).
16. Wu, Z. et al. Mining toxicity information from large amounts of toxicity data. *J. Med. Chem.* **64**, 6924–6936 (2021).
17. Martínez, M. J. et al. Multitask deep neural networks for ames mutagenicity prediction. *J. Chem. Inf. Model.* **62**, 6342–6351 (2022).
18. Shinada, N. K. et al. Optimizing machine-learning models for mutagenicity prediction through better feature selection. *Mutagenesis* **37**, 191–202 (2022).
19. Feeney, S. V., Lui, R., Guan, D. & Matthews, S. Multiple instance learning improves ames mutagenicity prediction for problematic molecular species. *Chem. Res. Toxicol.* **36**, 1227–1237 (2023).
20. Honma, M. et al. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. *Mutagenesis* **34**, 3–16 (2019).
21. Li, S. et al. MutagenPred-GCNNs: a graph convolutional neural network-based classification model for mutagenicity prediction with data-driven molecular fingerprints. *Interdiscip. Sci.* **13**, 25–33 (2021).

22. Gong, Y., Xue, D., Chuai, G., Yu, J. & Liu, Q. DeepReac+: deep active learning for quantitative modeling of organic chemical reactions. *Chem. Sci.* **12**, 14459–14472 (2021).
23. Settles, B. Active learning literature survey. *Computer Sciences Technical Report, University of Wisconsin-Madison.* http://digital.library.wisc.edu/1793/60660 (2009).
24. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *Proc. 33rd International Conference on Machine Learning*. 1050-1059 (PMLR, 2016).
25. Konyushkova, K., Sznitman, R. & Fua, P. Geometry in active learning for binary and multi-class image segmentation. *Comput. Vis. Image Undstand.* **182**, 1–16 (2019).
26. Lv, X., Duan, F., Jiang, J.-J., Fu, X. & Gan, L. Deep active learning for surface defect detection. *Sensors* **20**, 1650 (2020).
27. Eyke, N. S., Green, W. H. & Jensen, K. F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *React. Chem. Eng.* **5**, 1963–1972 (2020).
28. Xie, W. et al. Integrated random negative sampling and uncertainty sampling in active learning improve clinical drug safety drug–drug interaction information retrieval. *Front. Pharmacol.* **11**, 582470 (2021).
29. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
30. Kulichenko, M. et al. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat. Comput. Sci.* **3**, 230–239 (2023).
31. Cai, L., Wang, L., Fu, X. & Zeng, X. Active semisupervised model for improving the identification of anticancer peptides. *ACS Omega* **6**, 23998–24008 (2021).
32. Bressan, R. S., Camargo, G., Bugatti, P. H. & Saito, P. T. M. Exploring active learning based on representativeness and uncertainty for biomedical data classification. *IEEE J. Biomed. Health Inform.* **23**, 2238–2244 (2019).
33. Wu, L. et al. TOXRIC: a comprehensive database of toxicological data and benchmarks. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkac1074 (2022).
34. Probst, D. & Reymond, J. L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).
35. Scheffer, T., Decomain, C. & Wrobel, S. Active hidden Markov models for information extraction. In *Proc. International Symposium on Intelligent Data Analysis*. 309–318 (Berlin, Heidelberg, 2001).
36. Li, X. & Guo, Y. Adaptive active learning for image classification. In *Proc. 26th IEEE Conference on Computer Vision and Pattern Recognition*. 859–866 (IEEE, 2021).
37. Huang, S., Wang, T., Xiong, H., Huan, J. & Dou, D. Semi-supervised active learning with temporal output discrepancy. In *Proc. IEEE/CVF International Conference on Computer Vision*. 3427–3436 (2021).
38. Sener, O. & Savarese, S. Active learning for convolutional neural networks: a Core-Set approach. In *Proc. 6th International Conference on Learning Representations* (2018).
39. Ding, X. et al. Active learning for drug design: a case study on the plasma exposure of orally administered drugs. *J. Med. Chem.* **64**, 16838–16853 (2021).
40. Evans, C. P., Fleshner, N., Fitzpatrick, J. M. & Zlotta, A. R. An evidence-based approach to understanding the pharmacological class effect in the management of prostatic diseases. *BJU Int.* **95**, 743–749 (2005).
41. Chacko, R. et al. Data based predictive models for odor perception. *Sci. Rep.* **10**, 17136 (2020).
42. Tata, J. R. Thalidomide and induced amphibian metamorphosis. *Nature* **204**, 939–940 (1964).
43. Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).
44. Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discov. Today Technol.* **32-33**, 73–79 (2019).
45. Xu, C. et al. In silico prediction of chemical Ames mutagenicity. *J. Chem. Inf. Model.* **52**, 2840–2847 (2012).
46. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **29**, 476–488 (2010).
47. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminform.* **7**, 23 (2015).
48. Ren, P. et al. A survey of deep active learning. *ACM Comput. Surv.* **54**, 1–40 (2021).
49. Yoo, D. & Kweon, I. S. Learning loss for active learning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 93–102 (IEEE, 2019).

Acknowledgements

This work was supported by National Key R&D Program of China (No. 2023YFC2604400), National Natural Science Foundation of China (No. 62103436), Science and Technology Commission of Shanghai Municipality (No. 22xtcx00300), and Development Fund for Shanghai Talents (No. 2020010).

Author contributions

H.X., Y.Z., and S.H. conceived the study. H.X., Y.Z., Y.Z., J.H. conducted the experiments. H.X., Y.Z., and S.H. wrote the manuscript. P.Z., S.H., and X.B. supervised the work.

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/s42003-024-06758-6.

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

Peer review information *Communications Biology* thanks Slade Matthews, Giuseppina Gini, and Yuemin Bian for their contribution to the peer review of this work. Primary Handling Editors: Chien-Yu Chen and Johannes Stortz. A peer review file is available.