

RESEARCH

Open Access



PredAOT: a computational framework for prediction of acute oral toxicity based on multiple random forest models

Jae Yong Ryu^{1,2*}, Woo Dae Jang³, Jidon Jang³ and Kwang-Seok Oh^{3,4*}

*Correspondence:

jyryu@duksung.ac.kr; ksoh@krikt.re.kr

¹ Department of Biotechnology, Duksung Women's University, 33 Samyang-Ro 144-Gil, Dobong-gu, Seoul 01369, Republic of Korea

² Center for Research and Development, Oncocross Ltd., Seoul, Republic of Korea

³ Data Convergence Drug Research Center, Korea Research Institute of Chemical Technology, 141, Gajeong-ro, Yuseong-gu, Daejeon 34114, Republic of Korea

⁴ Department of Medicinal and Pharmaceutical Chemistry, University of Science and Technology, 176 Gajeong-Ro, Yuseong-gu, Daejeon 34129, Republic of Korea

Abstract

Background: Acute oral toxicity of drug candidates can lead to drug development failure; thus, predicting the acute oral toxicity of small compounds is important for successful drug development. However, evaluation of the acute oral toxicity of small compounds considered in the early stages of drug discovery is limited because of cost and time. Here, we developed a computational framework, PredAOT, that predicts the acute oral toxicity of small compounds in mice and rats.

Methods: PredAOT is based on multiple random forest models for the accurate prediction of acute oral toxicity. A total of 6226 and 6238 compounds evaluated in mice and rats, respectively, were used to train the models.

Results: PredAOT has the advantage of predicting acute oral toxicity in mice and rats simultaneously, and its prediction performance is similar to or better than that of existing tools.

Conclusion: PredAOT will be a useful tool for the quick and accurate prediction of the acute oral toxicity of small compounds in mice and rats during drug development.

Keywords: Acute oral toxicity, Machine learning, Random forest, Drug discovery

Background

After evaluating the effectiveness of drugs in the process of discovery and development of new drugs, preclinical tests were performed to evaluate their toxicity in animals prior to clinical trials [1]. Preclinical toxicity testing can predict toxic responses in humans, determine safe doses in clinical trials, and monitor toxicity-related symptoms and target organs in patients [2]. After having confirmed the safety of drugs by evaluating various toxicities, such as acute toxicity, nephrotoxicity, cardiotoxicity, reproductive toxicity, and genotoxicity, clinical trials can be performed.

Acute toxicity is defined as deleterious toxicological effects of a chemical from single or multiple exposures over a short duration (usually < 24 h) [3]. Studies on acute toxicity have examined various routes of exposure (e.g., oral, dermal, and inhalation) using rodents, such as mice and rats, to assess lethal doses. During the drug development



process, acute toxicity is generally evaluated using acute oral toxicity (AOT) tests that assess acute toxic reactions and the lethal dose 50 (LD₅₀) after a single oral administration to rodents [4, 5]. The LD₅₀ is defined as the dose of the test substance that can kill 50% of animals within 24 h of exposure.

AOT tests in animals are conducted after confirming the efficacy of a drug [5]. Therefore, if drug development fails owing to AOT at this stage, economic loss of development costs incurs. However, as dozens to thousands of compounds are considered drug candidates in the early stages of development, performing AOT tests on all compounds is limited by time and cost issues. Therefore, it is important to evaluate the AOT of such compounds in the early stages of drug development.

To address these issues, various prediction models have been developed based on the results of AOT tests of thousands of compounds [6–8]. In addition, various software programs that predict the AOT of small compounds based on machine learning models are currently available [7–10]. However, several aspects need to be improved for the development of an AOT prediction model. First, it is necessary to develop a computational framework that can predict the AOT in both mice and rats. AOT evaluation of small compounds is still being performed using either mice or rats. Second, it is necessary to further improve the prediction performance of models. One strategy to improve the performance of a model is to build it so that it can properly consider the distribution of the data. For example, using data composed of skewed LD₅₀ values for model training may reduce prediction performance [11].

In the present study, we propose a new computational framework, PredAOT, that predicts the AOT for a given compound in mice and rats. PredAOT is based on multiple random forest models for AOT prediction. For the development of PredAOT, we used a total of 6,226 and 6,238 compounds whose AOT was evaluated in mice and rats, respectively. Moreover, we compared the prediction performance of PredAOT with that of other existing tools. PredAOT is a useful tool for predicting the acute oral toxicity of small compounds during drug development.

Results and discussion

Development of the PredAOT framework

To develop a computational framework (i.e., PredAOT) for accurate prediction of AOT, we first collected data on the AOT of compounds (i.e., LD₅₀) reported for mice and rats. The AOT data for mice for 6226 compounds were obtained from the OCHEM database [7]. The AOT data for rats for 6238 compounds were obtained from the literature [6].

According to the GHS Classification, the AOT level of compounds can be divided into five categories [12] (Table 1). For example, compounds in Category 1 are toxic, while Category 5 compounds are less likely to be toxic. Additionally, we found that the distribution of LD₅₀ values for the compounds was skewed towards categories 4 and 5 in both mice and rats. This data imbalance can adversely affect model training, such as overfitting. To address this issue, we decided to classify the AOT of compounds into two categories (i.e., “toxic” and “less or non-toxic”) instead of five categories (Table 2); compounds with LD₅₀ ≤ 300 mg/kg and compounds with LD₅₀ > 300 mg/kg. Notably, although a compound is classified as less or non-toxic, it

Table 1 Number of compounds in datasets for each acute oral toxicity category

Toxicity category	LD ₅₀ (mg/kg)	Hazard statement	Mouse	Rat
Category 1	< 5	Fatal	67	173
Category 2	5–50	Fatal	282	490
Category 3	50–300	Toxic	1183	1103
Category 4	300–2000	Harmful	3413	2560
Category 5	2000–5000	May be harmful	1281	1912

Table 2 Number of compounds in the “less or non-toxic” and “toxic” datasets

	Less or non-toxic	Toxic
Mouse	4,694	1,532
Rat	4,472	1,766

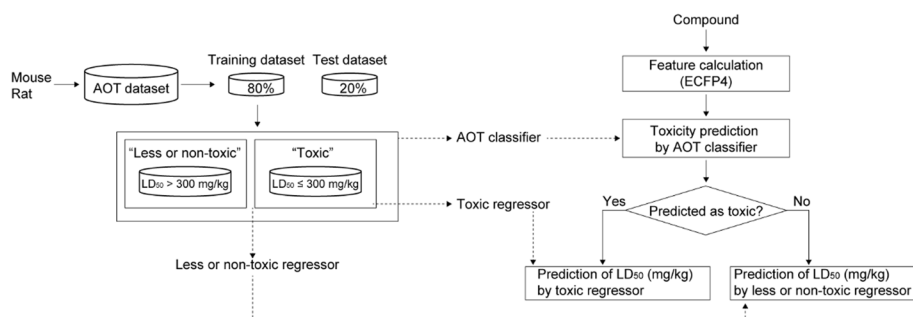


Fig. 1 Overall scheme of PredAOT. PredAOT uses the chemical structure as an input. Thereafter, the molecular fingerprint (i.e., ECFP4) is used as an input feature for models in PredAOT. PredAOT is composed of one classification model (i.e., “AOT classifier”) and two regression models (i.e., “toxic regressor” and “less or non-toxic regressor”). The AOT classifier is used for prediction as “toxic” ($LD_{50} \leq 300$ mg/kg) or “less or non-toxic” ($LD_{50} > 300$ mg/kg) for a given compound. If a compound is predicted to be toxic, PredAOT predicts the LD_{50} of the compound using the toxic regressor trained with compounds with $LD_{50} \leq 300$ mg/kg. If a compound is predicted to be less or non-toxic, PredAOT predicts the LD_{50} of the compound using the less or non-toxic regressor trained with compounds with $LD_{50} > 300$ mg/kg. All these procedures are equally applied to the AOT prediction process in mice and rats

does not necessarily have to involve no AOT at all. During the model development, LD_{50} values were transformed to \log_{10} transformed LD_{50} values.

We then used the datasets to train a binary classification model, called “AOT classifier,” that predicts AOT as toxic (i.e., $LD_{50} \leq 300$ mg/kg) or less or non-toxic (i.e., $LD_{50} > 300$ mg/kg) for a given compound. In addition, as shown in Fig. 1, toxic and less or non-toxic datasets were used to train two regression models called “toxic regressor” and “less or non-toxic regressor”, respectively. In particular, the “toxic regressor” was trained with the toxic dataset (i.e., $LD_{50} \leq 300$ mg/kg) and the “less or non-toxic regressor” was trained with the less or non-toxic dataset (i.e., $LD_{50} > 300$ mg/kg).

To summarize the AOT prediction process in the PredAOT, “AOT classifier” first predicts AOT as “toxic” or “less or non-toxic” for a given compound. Thereafter, if a compound is predicted as toxic, the “toxic regressor” predicts the LD_{50} of the compound; otherwise, the “less or non-toxic regressor” predicts the LD_{50} of the

compound. All procedures were equally applied to the AOT prediction process in mice and rats.

Optimization and evaluation of AOT prediction models

To build an optimal AOT prediction model, we constructed and evaluated six different machine learning models: a message passing neural network (MPNN) based on graph neural networks, MPNN with molecular fingerprints, MPNN with molecular descriptors, random forest (RF), support vector machine (SVM), and artificial neural network (ANN) models based on molecular fingerprints (see [Materials and Methods](#)). The training dataset was used for model training to determine the model with the best prediction performance. A test dataset was used to assess the performance of the final model.

First, we optimized the binary classification model (i.e., “AOT classifier”) used to predict the AOT of a given compound in mice and rats. To this end, we evaluated various hyperparameters using a grid search technique with fivefold cross-validation, and used accuracy as the model performance metric. RF showed the highest accuracy for mice (0.8672) and rats (0.8377; Fig. 2). To further improve the prediction performance of the RF model, we applied an oversampling approach to manage imbalanced data (Table 2). Specifically, we used the synthetic minority oversampling technique (SMOTE), which is a popular algorithm used to generate artificial data [13]. In doing so, RF with SMOTE showed better prediction performance than RF without SMOTE: accuracies of 0.9586 and 0.9335 in mice and rats, respectively (Fig. 3). In addition, we evaluated the prediction performance using the test dataset. The RF model with SMOTE showed an area under the receiver operating characteristic (AUROC) of 0.7778, Matthew’s correlation coefficient (MCC) of 0.5514, positive predictive value (PPV) of 0.6627, and negative predictive value (NPV) of 0.8845 in mice, and an AUROC of 0.7442, MCC of 0.4929, PPV of 0.6435, and NPV of 0.8539 in rats (Table 3). Based on these results, we used RF models with SMOTE in both mice and rats in the PredAOT framework.

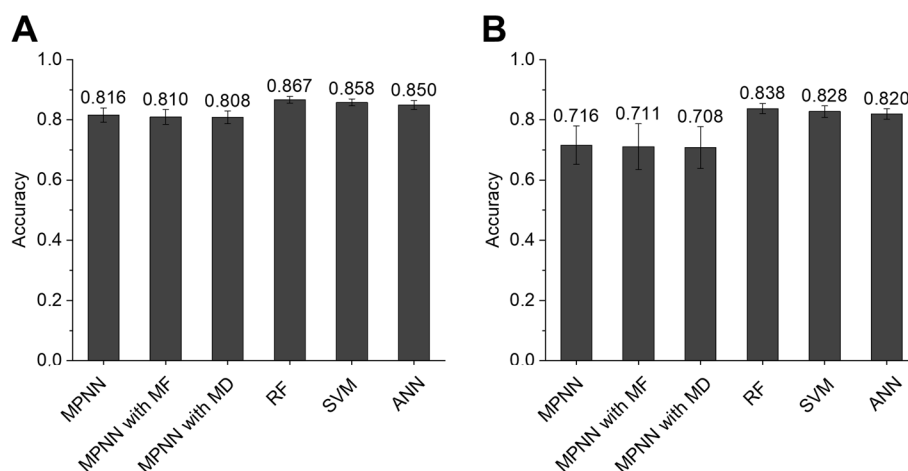


Fig. 2 Cross-validation prediction performances of the MPNN, MPNN with MF, MPNN with MD, RF, SVM, and ANN models for the AOT classifier using mouse (A) and rat (B) datasets. Each performance metric value was calculated by five-fold cross-validation. MPNN, message passing neural network; MF, molecular fingerprint; MD, molecular descriptor; RF, random forest; SVM, support vector machine; ANN, artificial neural network

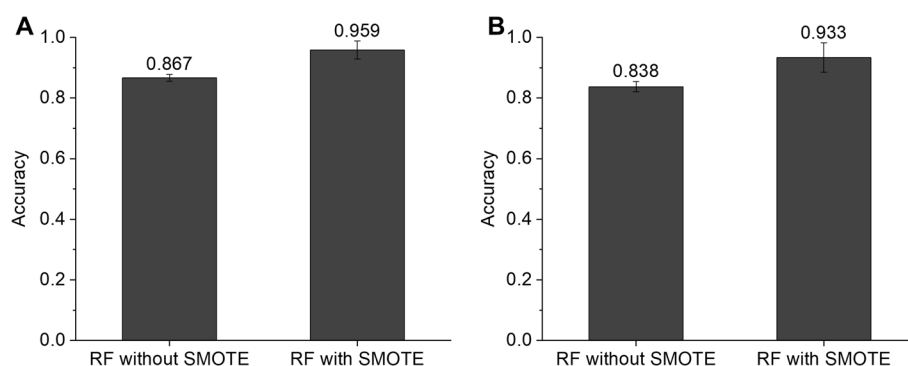


Fig. 3 Cross-validation prediction performances of RF models with and without SMOTE using mouse (A) and rat (B) datasets. MPNN, message passing neural network; RF, random forest; SVM, support vector machine; ANN, artificial neural network

Table 3 Prediction performance of classification models in PredAOT on the test dataset

	AUROC	MCC	PPV	NPV
Mouse	0.7778	0.5514	0.6627	0.8845
Rat	0.7442	0.4929	0.6435	0.8539

AUROC, area under receiver operating characteristic; MCC, Matthew's correlation coefficient; PPV, positive predictive value; NPV, negative predictive value

As shown in Fig. 1, the AOT of the compound was first predicted to be “toxic” or “less or non-toxic” using an AOT classifier in mice and rats. The LD₅₀ value (mg/kg) was subsequently quantitatively predicted using one of the regression models (i.e., “toxic regressor” or “less or non-toxic regressor”) according to the prediction result of the AOT classifier. Here, we optimized both regressors (i.e., “toxic regressor” or “less or non-toxic regressor”). The prediction performance of these regressors was evaluated through five-fold cross-validation using the training dataset. The root-mean-square error (RMSE) was used as the performance metric for the regression models. Consequently, RF showed the lowest RMSE, i.e., the best performance, in the “toxic regressor” and “less or non-toxic regressor” in both mice and rats. The RF model for both regressors showed an RMSE of 0.2999 and 0.3767 in mice, respectively (Figs. 4A, 5A) and 0.3919 and 0.4984 in rats, respectively (Figs. 4B, 5B). Thereafter, we evaluated the prediction performance using the test dataset. In mice, the toxic regressor showed an RMSE of 0.3806 and an R² of 0.3557 on the test dataset (Table 4), whereas the less or non-toxic regressor showed an RMSE of 0.2923 and an R² of 0.3881. In rats, the toxic regressor showed an RMSE of 0.5323 and an R² of 0.3065 on the test dataset, whereas the less or non-toxic regressor showed an RMSE of 0.3863 and an R² of 0.2702.

Prediction performance of PredAOT compared with other prediction models

We compared the prediction performance of PredAOT with that of other existing tools. First, we compared the prediction performance of PredAOT in mice. Thereafter, we compared the prediction performance of PredAOT with OCHEM Predictor using our test dataset [7]. The OCHEM Predictor predicted the LD₅₀ (mg/kg) result in the same manner as PredAOT, such that predicted values could be directly

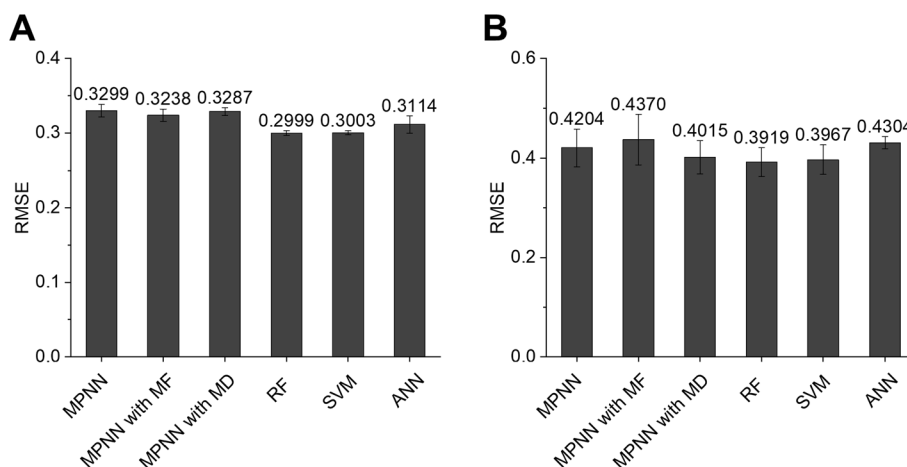


Fig. 4 Cross-validation prediction performances of the MPNN, MPNN with MF, MPNN with MD, RF, SVM, and ANN models for toxic regressors using mouse (A) and rat (B) toxic datasets (i.e., LD₅₀ ≤ 300 mg/kg). MPNN, message passing neural network; MF, molecular fingerprint; MD, molecular descriptor; RF, random forest; SVM, support vector machine; ANN, artificial neural network

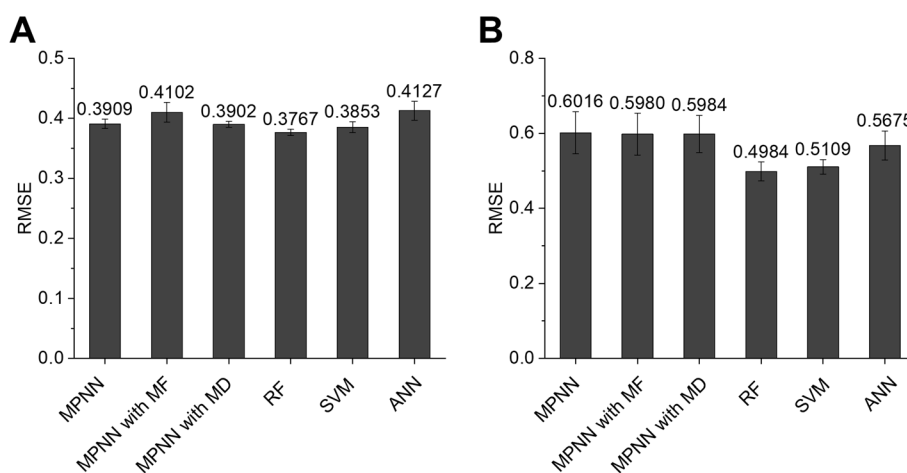


Fig. 5 Cross-validation prediction performances of the MPNN, MPNN with MF, MPNN with MD, RF, SVM and ANN models for the non-toxic regressor using mouse (A) and rat (B) less or non-toxic datasets (i.e., LD₅₀ > 300 mg/kg). MPNN, message passing neural network; MF, molecular fingerprint; MD, molecular descriptor; RF, random forest; SVM, support vector machine; ANN, artificial neural network

Table 4 Prediction performance of regression models in PredAOT on the test dataset. RMSE and R² values are calculated using log₁₀ transformed LD₅₀ values

	Regression model	RMSE	R ²
Mouse	Less or non-toxic	0.2923	0.3881
	Toxic	0.3806	0.3557
Rat	Less or non-toxic	0.3863	0.2702
	Toxic	0.5323	0.3065

RMSE, root mean squared error; R², R-squared

Table 5 Comparison of RMSE values predicted by PredAOT with OCHEM Predictor on the mouse test dataset

Toxicity category	OCHEM Predictor	PredAOT
Category 1	193.11	82.13
Category 2	273.24	294.64
Category 3	805.42	289.55
Category 4	513.90	533.00
Category 5	1645.57	1645.25

RMSE, root mean squared error

Table 6 Comparative performance evaluation of PredAOT with BESTox and aiQSAR on the rat test dataset

	Pearson's r	Spearman's r
BESTox	0.7170	0.6978
aiQSAR	-0.8989	-0.8918
PredAOT	0.7984	0.7639

compared. We compared the prediction performance for each toxicity group in Table 1. PredAOT yields relatively low RMSE values (i.e., improved performance) in groups with relatively little training data (e.g., Categories 1 and 2) (Table 5). Notably, the performance comparison indicates that the test dataset may have been used as training data in OCHEM Predictor.

Second, we compared the prediction performance of PredAOT in rats with that of BESTox and aiQSAR using our test dataset [9, 10]. The prediction results of BESTox and aiQSAR were not directly comparable with the prediction results of PredAOT because the unit for LD₅₀ was not mg/kg. Therefore, we compared the prediction results using two correlation coefficients: Pearson's correlation coefficient (Pearson's r) and Spearman's rank correlation coefficient (Spearman's r). PredAOT showed the highest Pearson's r of 0.7984 and Spearman's r of 0.7340 compared to BESTox and aiQSAR (Table 6).

Although we did not compare LD₅₀ values directly, PredAOT performed comparably or better than existing tools. PredAOT does not outperform other existing tools; however, it can be a useful tool for predicting AOT. In addition, it has the advantage of being able to predict AOT in mice and rats simultaneously.

Conclusions

In the present study, we developed a computational framework called PredAOT, which predicts the AOT of a given compound in mice and rats. PredAOT first classifies the given compound as "toxic" or "less or non-toxic", and then further qualitatively predicts the LD₅₀ value using a regression model. PredAOT is trained with information on the AOT of 6,226 and 6,238 compounds in mice and rats, respectively. PredAOT has the advantage of predicting AOT in mice and rats simultaneously, and its prediction performance is similar to or better than that of existing tools. The web

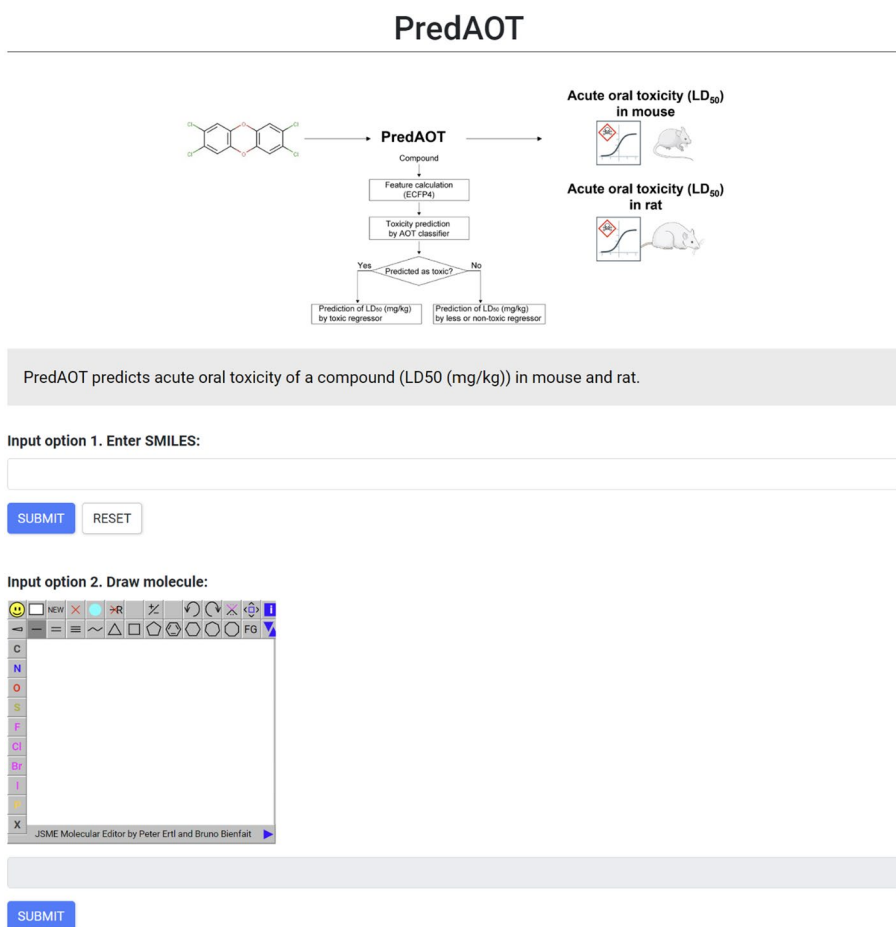


Fig. 6 User interface of the PredAOT web server

server for implementing PredAOT is available at <https://predaot.netlify.app/> (Fig. 6). PredAOT will be a useful tool for the quick and accurate prediction of the AOT of small compounds in mice and rats for successful drug development.

Methods

Data preparation

Information on acute oral toxicity (AOT) of 6,226 compounds determined in mice was obtained from the OCHEM database [7], and information on AOT of 6,238 compounds determined in rats was obtained from the literature [6]. We defined compounds with lethal dose 50 (LD₅₀) values ≤ 300 mg/kg as “toxic” and compounds with LD₅₀ values > 300 mg/kg as “less or non-toxic”. The dataset was divided into training (80%) and test (20%) datasets. The training dataset was used for hyperparameter optimization, and the test dataset was used for model evaluation (Fig. 1). During the model training and evaluation, we used log₁₀ transformed LD₅₀ values.

Preparation of molecular features

The structures of the compounds were presented in the simplified molecular-input line-entry system (SMILES) format [14]. To train message-passing neural networks

(MPNNs), we used the Chemprop Python package [15]. The RDKit Python package was used to calculate the molecular fingerprint. Extended connectivity fingerprints with a maximum diameter parameter of 4 (ECFP4) were used [16].

Optimization of machine learning algorithms

In this study, we tested six different machine learning (ML) algorithms, including MPNN, MPNN with molecular fingerprints, MPNN with molecular descriptors, random forest (RF), support vector machine (SVM), and artificial neural network (ANN) models, to build both classification and regression models for predicting AOT in mice and rats. Here, the RF, SVM, and ANN models were trained using molecular fingerprints as input features. MPNN learns directly from a molecular graph to predict molecular properties [15]. ANN is an ML algorithm inspired by the biological neuronal network of the human brain [17]. The ANN structure consisted of an input layer, hidden layer(s), and an output layer. The ANN learns non-linear relationships from the data. RF is an ensemble learning algorithm that constructs multiple decision trees [18]. The ANN algorithm was implemented using the Keras package (version 2.2.5) with TensorFlow backend (version 2.0.0) [19]. The RF and SVM algorithms were implemented using the *scikit-learn* Python package [20].

To build the optimal model with the best prediction performance, hyperparameter optimization was performed. For the classification model, we selected the optimal hyperparameter that showed the highest accuracy (ACC) using the grid-search cross-validation method. In addition, for the regression model, we selected the optimal hyperparameter that showed the lowest root mean square error (RMSE) using the grid search cross-validation method.

Five metrics were used to evaluate the performance of the classification model: ACC, area under the receiver operating characteristic (AUROC), Matthew's correlation coefficient (MCC), positive predictive value (PPV), and negative predictive value (NPV). In addition, to evaluate the regression model performance, two performance metrics were used: the RMSE and R-squared value (R^2).

Abbreviations

ACC	Highest accuracy
ANN	Artificial neural network
AOT	Acute oral toxicity
AUROC	Area under the receiver operating characteristic
LD ₅₀	Lethal dose 50
ML	Machine learning
MCC	Matthew's correlation coefficient
MPNN	Message passing neural network
NPV	Negative predictive value
PPV	Positive predictive value
RF	Random forest
RMSE	Root-mean-square error
SMOTE	Synthetic minority oversampling technique
SMILES	Simplified molecular-input line-entry system
SVM	Support vector machine

Acknowledgements

Not applicable.

Author contributions

J.Y.R. and K.S.O. designed the research; J.Y.R. developed the prediction model; J.Y.R., J.D.J., and W.D.J. analyzed the data; All authors have read and approved the manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MST) (nos. NRF-2020R1C1C1003218 and 2022M3J4A1091452). This work was also funded by the Korea Research Institute of Chemical Technology (no. SI2231-30-0221082086810001).

Availability of data and materials

The data and source code are available at <https://github.com/CSB-L/PredAOT>, and the PredAOT web server is available at <https://predaot.netlify.app>. The datasets supporting the conclusions of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 15 October 2022 Accepted: 6 February 2023

Published online: 24 February 2023

References

1. Kramer JA, Sagartz JE, Morris DL. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat Rev Drug Discov.* 2007;6(8):636–49.
2. Holt JD, Nuttall JP. Preclinical safety evaluation. *Curr Top Microbiol Immunol.* 2014;383:55–78.
3. OECD: Test No. 425: acute oral toxicity: up-and-down procedure, OECD guidelines for the testing of chemicals, section 4. In: 2008; Paris. OECD Publishing.
4. Balls M. Why modification of the LD50 test will not be enough. *Lab Anim.* 1991;25(3):198–206.
5. Walum E. Acute oral toxicity. *Environ Health Perspect.* 1998;106(Suppl 2):497–503.
6. Gadaleta D, Vukovic K, Toma C, Lavado GJ, Karmaus AL, Mansouri K, Kleinstreuer NC, Benfenati E, Roncaglioni A. SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. *J Cheminform.* 2019;11(1):58.
7. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, et al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des.* 2011;25(6):533–54.
8. Garcia-Jacas CR, Marrero-Ponce Y, Cortes-Guzman F, Suarez-Lezcano J, Martinez-Rios FO, Garcia-Gonzalez LA, Pupo-Merino M, Martinez-Mayorga K. Enhancing acute oral toxicity predictions by using consensus modeling and algebraic form-based 0D-to-2D molecular encodes. *Chem Res Toxicol.* 2019;32(6):1178–92.
9. Vukovic K, Gadaleta D, Benfenati E. Methodology of aiQSAR: a group-specific approach to QSAR modelling. *J Cheminform.* 2019;11(1):27.
10. Chen J, Cheong H-H, Siu SWI. BESTox: a convolutional neural network regression model based on binary-encoded SMILES for acute oral toxicity prediction of chemical compounds. *Proc Int Conf Algorithms Comput Biol* 2020:155–166.
11. Green JA. Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. *Health Psychol Behav Med.* 2021;9(1):436–55.
12. Administration UNAAr. Code of federal regulations, protection of environment, title 40, sec. 156.62. In: 2006.
13. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* 2013;14:106.
14. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988, 28:31–36.
15. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, et al. Analyzing learned molecular representations for property prediction. *J Chem Inf Model.* 2019;59(8):3370–88.
16. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model.* 2010;50(5):742–54.
17. Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol.* 2016;12(7):878.
18. Vincenzi F, Targa M, Corciulo C, Gessi S, Merighi S, Setti S, Cadossi R, Goldring MB, Borea PA, Varani K. Pulsed electromagnetic fields increased the anti-inflammatory effect of A(2)A and A(3) adenosine receptors in human T/C-28a2 chondrocytes and hFOB 1.19 osteoblasts. *PLoS ONE* 2013, 8(5):e65561.
19. Abadi M, Barham P, Chen JM, Chen ZF, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. TensorFlow: a system for large-scale machine learning. *Proceedings of Osd1'16: 12th Usenix Symposium on Operating Systems Design and Implementation* 2016:265–283.
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* 2011;12:2825–30.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.