Fundamental Toxicological Sciences (Fundam. Toxicol. Sci.)
Vol.11, No.6, 279-288, 2024

The Japanese Society of Toxicology

Fundamental Toxicological Sciences
URL : http://www.fundtoxicolsci.org/index_e.html

Original Article

Constructing a graph neural network-based artificial intelligence model to predict drug-induced phospholipidosis potential

Yoshinobu Igarashi[1], Aki Hasegawa[2], Shigeyuki Matsumoto[2], Hiroaki Iwata[2,3], Ryosuke Kojima[2], Yasushi Okuno[2] and Hiroshi Yamada[1]

[1]Toxicogenomics Informatics Project, National Institutes of Biomedical Innovation, Health and Nutrition, Kento Innovation Park, NK Building, 3-17 Senrioka Shinmachi, Settsu-shi, Osaka 566-0002, Japan
[2]Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, 53 Kawahara-cho, shogoin Sakyo-ku, Kyoto 606-8397, Japan
[3]Present address: Department of Biological Regulation, Faculty of Medicine, Tottori University, 86 Nishi-cho, Yonago-shi, Tottori 683-8503, Japan

(Received October 24, 2024; Accepted October 31, 2024)

ABSTRACT — Drug-induced phospholipidosis (DIPL) is linked to various toxicities, including hepatotoxicity, making it a critical screening factor in the early stages of drug discovery. Several models based on chemical structures have been constructed to predict compounds with DIPL potential. However, most of these models only classify results as inducers or non-inducers, without identifying the specific substructures responsible for positive outcomes. To address this limitation, we constructed an artificial intelligence (AI) model to predict compounds with DIPL potential and visualize structural alerts. The proposed model was constructed using kMoL, an open-source software library that employs a graph neural network approach to learn from chemical structural data. We employed the bagging method, resulting in a model with a high predictive performance. The model attained an F1 score of 0.796 on the external test set. In addition, we used the integrated gradient method to visualize the substructures that contributed to positive predictions. When applying the method to compounds that experimentally conducted structure-activity relationship investigations, the proposed AI model accurately predicted DIPL potential, demonstrating its practical utility in early-stage drug discovery. By predicting DIPL based on the chemical structure of compounds, the proposed model can aid in the screening for DIPL, potentially improving the safety profile of new drug candidates.

Key words: Bagging method, Deep learning, Drug-induced phospholipidosis, Explainable artificial intelligence, Graph neural networks

INTRODUCTION

Drug-induced phospholipidosis (DIPL) frequently poses a significant challenge in drug development because of its substantial impact on the safety profile of pharmaceutical compounds. DIPL is characterized by the abnormal accumulation of phospholipids, predominantly within lysosomes, potentially resulting in toxicity in major organs such as the liver, kidneys, and lungs, thereby compromising the long-term administration of the drug. Consequently, compounds with a high risk of DIPL are often terminated early in clinical trials. The spectrum of drugs

Correspondence: Hiroshi Yamada (E-mail: h-yamada@nibiohn.go.jp)

capable of inducing phospholipidosis is remarkably diverse and includes antibiotics, antidepressants, antipsychotics, antimalarials, and antiarrhythmics (Anderson and Borlak, 2006).

Several studies have investigated the mechanisms underlying DIPL (Hostetler and Matsuzawa, 1981; Mingeot-Leclercq *et al.*, 1995; Sawada *et al*., 2005; Lowe *et al*., 2012). It is primarily believed that cationic amphiphilic drugs (CADs) induce DIPL by binding to phospholipids within lysosomes, thereby inhibiting their metabolism (Halliwell, 1997). Specifically, certain drugs readily pass through the lysosomal membrane and bind to phospholipids to form complexes (Joshi *et al*., 1988). These complexes accumulate reversibly within the lysosomes and are observed as lamellar bodies. In addition, CADs inhibit phospholipase activity, prevent phospholipid breakdown, and result in abnormal accumulation (Hostetler and Matsuzawa, 1981). Phospholipase interacts with negatively charged phospholipids in lysosomes, and its activity increases with increasing membrane negative charge. CADs bind to these membrane lipids, neutralizing their negative charge, and inhibiting phospholipase activity (Mingeot-Leclercq *et al*., 1995). Furthermore, there are reports that CADs can expand lysosomal volume (Funk and Krise, 2012). Although these complex interactions are being clarified, they do not explain DIPL, and a clear mechanism remains to be determined.

Current methods for screening DIPL potential are widely employed in *in vitro* assays. These assays primarily utilize cell culture systems, including human hepatocytes (Shahane *et al*., 2014) and animal-derived cell lines (Kasahara *et al*., 2006), enabling the efficient screening of potential DIPL risks associated with drug candidates. In these tests, cells are exposed to various concentrations of the drug for a specified period (typically 24–72 hr), and the accumulation of intracellular phospholipids is measured to assess the risk of DIPL. Although definitive diagnosis involves observing the formation of lamellar bodies within cells using electron microscopy, this method is time-consuming and labor-intensive. Consequently, several *in vitro* assays employ fluorescent staining techniques. For instance, fluorescent dyes, such as LipidTOX Green or Nile Red, are used in conjunction with flow cytometry or fluorescence microscopy to quantitatively evaluate phospholipid accumulation. This approach allows for early and efficient detection of the potential of a drug to induce phospholipidosis. Furthermore, recent developments in high-content screening technologies have enabled simultaneous evaluation of several compounds, thereby enhancing the efficiency of DIPL risk assessment in the early stages of drug discovery.

In silico methods play a crucial role in evaluating DIPL potential. Notably, quantitative structure-activity relationship (QSAR) models (Orogo *et al*., 2012; Kruhlak *et al*., 2008), classifications based on physicochemical properties (Ploemen *et al.*, 2004; Tomizawa *et al*., 2006; Pelletier *et al*., 2007; Hanumegowda *et al*., 2010), and model construction using molecular descriptors (Sun *et al*., 2012; Przybylak and Cronin, 2011) have gained widespread recognition. These models initially relied on small datasets and limited physicochemical properties for classification. Recent advancements have led to the use of large-scale datasets and sophisticated machine learning algorithms such as random forests and support vector machines, enabling a more precise analysis of the relationship between compounds and phospholipidosis (Lowe *et al*., 2010; Lowe *et al*., 2012; Nath and Sahu, 2019). This progression has significantly improved our understanding of complex structure-activity relationship (SAR). However, a more accurate model for predicting phospholipidosis is yet to be developed. Therefore, we constructed a predictive model for phospholipidosis using artificial intelligence (AI).

In recent years, AI techniques based on deep learning have been applied in toxicology research. One notable method is a graph neural network (GNN), which learns the latent representations of graph nodes (Kipf and Welling, 2016; Wu *et al*., 2019; Morris *et al*., 2019; Li *et al*., 2020; Brody *et al.*, 2021). Traditional models for predicting toxicity use molecular descriptors derived from a chemical structure representation format, such as the simplified molecular input line entry system (SMILES), which represents chemical structures with alphanumeric strings. These conventional models rely on different methods to calculate the molecular descriptors, potentially introducing biases that depend on the chosen method. In contrast, GNNs can directly learn from the chemical structure without converting it into other forms, such as descriptors, thereby reducing the biases related to descriptor calculation. Moreover, the integrated gradient (IG) technique (Sundararajan *et al*., 2017) can be used to visualize the basis for prediction, addressing a key challenge in AI-based predictive model development. The kMoL software package (https://github.com/elix-tech/kmol), which integrates both GNN and IG functions, was employed to construct the proposed AI model for predicting DIPL. This software was based on the kGCN framework from a previous study (Kojima *et al*., 2020) and was developed by Kojima and Okuno of Kyoto University and Elix Corporation.

One of the challenges in developing toxicity prediction models is achieving a high predictive performance. In this

study, the accuracy was initially insufficient. Therefore, we employed the bagging method to enhance the predictive capabilities of the model (Breiman, 1996). The bagging method is an ensemble learning technique in which multiple models are trained on different resampled subsets of the data and their predictions are aggregated through averaging or majority voting. This approach resulted in significant improvements in model stability and robustness, reduced overfitting, and enhanced the ability of the model to generalize. This study demonstrates that the bagging method effectively boosts the performance of the AI model in DIPL potential prediction.

We created an AI model based on a GNN to predict the DIPL potential using kMoL. The proposed model not only identified toxicity but also visualized substructures potentially linked to DIPL and assessed their impact using the IG method. This AI model was developed through a collaborative research project between the industry and academia. This project aimed to create cutting-edge AI systems to enhance early-stage drug discovery by leveraging recent significant advancements in AI. The developed model is incorporated into the AI system of the project for practical use. In this paper, we present our findings and discuss the benefits of the proposed AI model.

MATERIALS AND METHODS

Data preparation

The data sources and number of compounds used in the construction of the AI model for predicting DIPL potential are listed in Table 1 (Shahane *et al*., 2014; Fusani *et al*., 2017; Przybylak *et al*., 2014). Compounds that induced phospholipid accumulation were labeled as positive, whereas those that did not induce phospholipid accumulation were labeled as negative. When a molecule had multiple fragments, the 'parent' molecule was used, and SMILES with multiple labels were excluded to ensure each SMILES was uniquely labeled. SMILES were standardized using the MELLODDY-TUNER pipeline standardization tool (Heyndrickx *et al*., 2024). After removing duplicate data, 690 positive compound data points and 1,556 negative compound data points were collected. Additionally, 13 compounds were obtained as more recent data from two literature sources (Sakai *et al*., 2020; Hu *et al*., 2023). Of these, the data from the former literature consists of three compounds, including a lead compound that initially demonstrated phospholipidosis-inducing potential, along with two derivatives developed through SAR optimization. The data from the latter literature comprise molecular pairs, known as matched molecular pairs (MMPs), in which only a single structural segment is modified. From this group, 10 compounds, including five pairs confirmed to have the potential to induce phospholipidosis, were selected. These 13 compounds were collected after model construction, standardized using the same methodology, and used in this study.

Visualization of the chemical space

SMILES data for pharmaceuticals were sourced from https://open.fda.gov/apis/drug/ndc/download/. A total of 56,651 records from the "HUMAN OTC DRUG" category and 52,198 from the "HUMAN PRESCRIPTION DRUG" category were used to extract information from the "active_ingredients" field. After deduplication, 2,826 distinct active ingredient entries were obtained. Drug SMILES were retrieved via PubChem and standardized using the ChEMBL structure pipeline. Chemical space

Table 1. Source of the data sets for training and evaluation.

Article	Author	Content
Detection of Phospholipidosis Induction: A Cell-Based Assay in High-Throughput and High-Content Format	Shahane S. A., *et al*.	Cell-based phospholipidosis assay using LipidTOX Red reagent Positive compound data: 108 Negative compound data: 0
Predicting the Risk of Phospholipidosis with in Silico Models and an Image-Based in Vitro Screen	Fusani L., *et al*.	HTS (high-throughput screening) Positive compound data: 497 Negative compound data: 1,223
How Does the Quality of Phospholipidosis Data Influence the Predictivity of Structural Alerts?	Przybylak K. R., *et al.*	Data from databases and literature Positive compound data: 442 Negative compound data: 810
Total amount		Positive compound data: 690 Negative compound data: 1,556 Total: 2,246

visualizations were generated using ChemPlot (Cihan Sorkun *et al*., 2022) and uniform manifold approximation and projection (UMAP) methods (McInnes *et al*., 2018). UMAP enables high-dimensional data to be projected onto lower dimensions, facilitating the visualization of relationships among the data.

Computational resources for model construction

The model was constructed on a desktop computer equipped with an AMD EPYC 7542 (2.90GHz) and 128 GB RAM. For model training, we utilized NVIDIA RTX A6000 and Quadro RTX 8000 GPUs under CUDA 11.7 and PyTorch 1.13.1. Training, including a hyperparameter search for a single model, required an average of 3 hr, and the total process consumed approximately 25 hr.

Deep learning software package

The AI models were constructed using kMoL version 1.1.9 (https://github.com/elix-tech/kmol) and ChemPlot 1.2.0 (https://chemplot.readthedocs.io/en/latest/).

AI model construction

The AI model was constructed using a GNN approach with two methods: one used the entire dataset, while the other applied a bagging technique by splitting the dataset into subsets. In both the cases, the same external test data were first separated from the dataset. For the method without splitting, the remaining data were divided into training, validation, and internal test sets in an 8:1:1 ratio through random sampling. Next, the hyperparameters were tuned using the training and validation sets with 10-fold cross-validation. A hyperparameter search for optimization was performed using the tree-structured Parzen estimator algorithm implemented in Optuna (Akiba *et al*., 2019). The AdaBelief (Zhuang *et al*., 2020) optimizer was used for model training. In the splitting method, this process was repeated seven times with different random seeds, resulting in seven distinct sub-AI models. Finally, each sub-AI model and the model constructed through bagging were evaluated using the F1 score of the prediction for each external dataset (Fig. 1). This procedure is described in detail below.

1. Data partitioning: From all data, 50 positive and 50 negative data points were randomly sampled and assigned to the external test set. The remaining data were used to train each AI model.
2. Creation of sub-datasets: The sub-datasets were designed to minimize data overlap between subsets while ensuring that more than seven subsets were generated from 640 positive and 1506 negative samples, with an adjusted proportion of positive to negative samples. Consequently, seven subsets were created, each containing 200 positive and 308 negative samples, maintaining a consistent proportion for training.
3. Sub-dataset split: Each training dataset was split

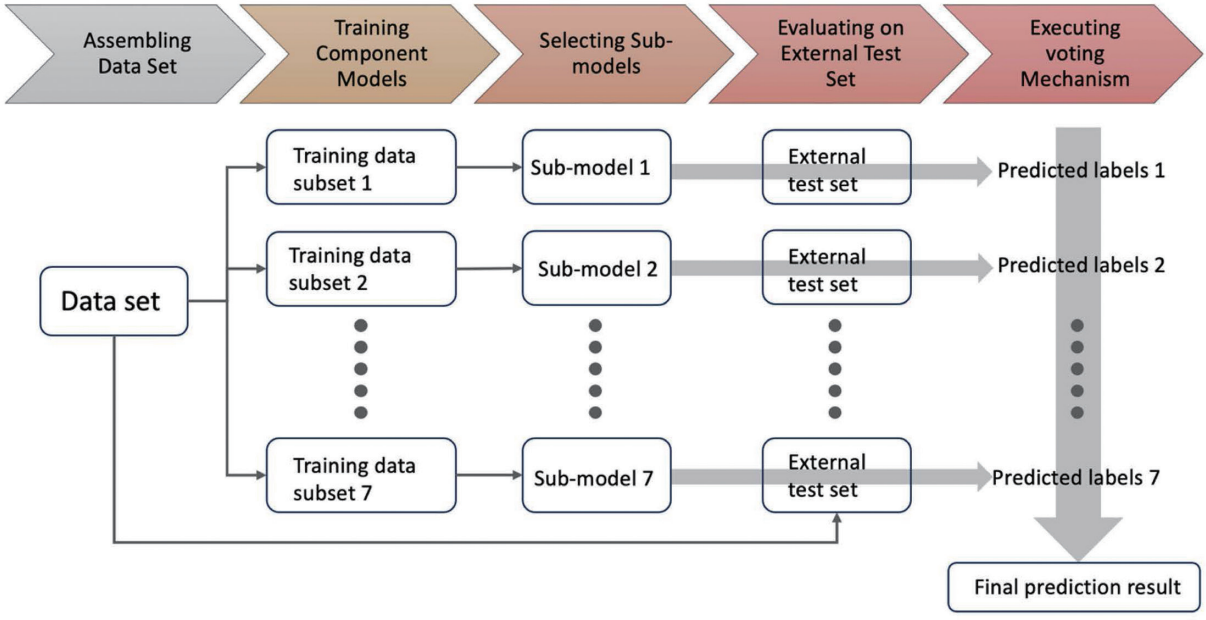

Fig. 1. Schema of data partitioning and AI model construction. Detailed information is provided in the Methods section.

into training, validation, and internal test sets in a ratio of 8:1:1.
4. Sub-AI model training: The models were trained using the training sets and their hyperparameters were tuned using 10-fold cross-validation.
5. Validation of the sub-AI models: The discriminant performance of the sub-models was evaluated using a validation set.
6. Models were selected for internal testing using the F1 score as an indicator and the top 10 models were selected to increase the F1 score.
7. Internal testing: An internal test set was evaluated on 10 selected models.
8. Determination of the best sub-AI model: The model with the highest F1 score was selected for internal testing.
9. Final validation of the AI model: The discriminant performance of the AI model, consisting of seven sub-models, was evaluated using an external test set.

Evaluation of predictive performance

The metrics used in the present study were as follows.

$$\text{F1 score} = \frac{\text{precision} \times \text{recall}}{1/2 \times (\text{precision} + \text{recall})}$$

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

$$\text{Recall (true positive rate)} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{Accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}$$

RESULTS

Comparing chemical spaces of compounds for AI model and pharmaceuticals

We assessed the chemical space of the compounds within the dataset for the AI model construction and com-

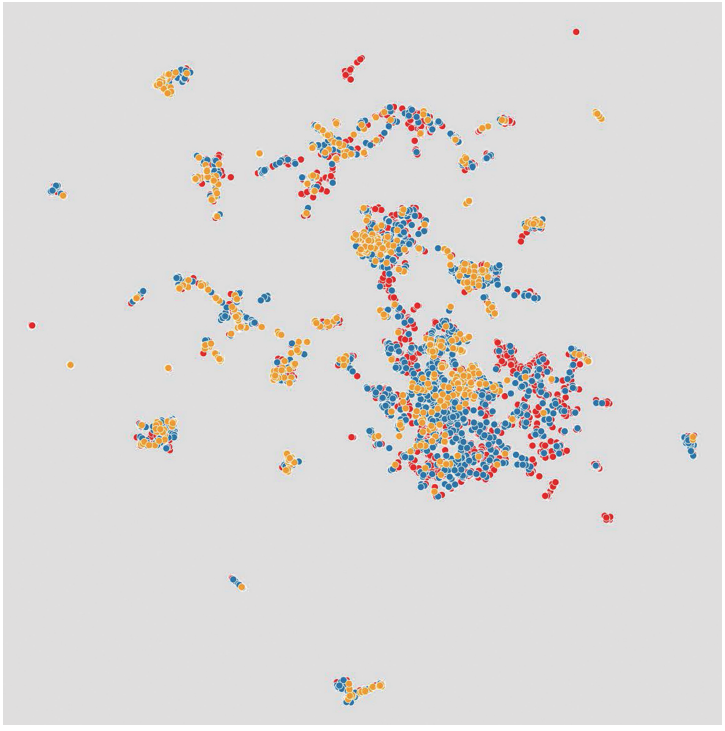

Fig. 2. Comparison of chemical spaces of compounds in the study dataset and pharmaceuticals. Orange represents compounds positive for drug-induced phospholipidosis, blue indicates compounds negative for drug-induced phospholipidosis, and red indicates pharmaceuticals.

Table 2. Performance of sub-models through the model construction phase and the performance of the bagging model.

	External test set				Internal test set				Training set			
Sub-model number	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall
1	0.747	0.770	0.680	0.829	0.744	0.784	0.696	0.800	0.829	0.863	0.810	0.850
2	0.692	0.680	0.720	0.667	0.783	0.804	0.692	0.900	0.844	0.863	0.760	0.950
3	0.792	0.800	0.760	0.826	0.722	0.804	0.813	0.650	0.833	0.882	0.938	0.750
4	0.772	0.770	0.780	0.765	0.667	0.706	0.600	0.750	0.905	0.922	0.864	0.950
5	0.643	0.700	0.540	0.794	0.737	0.804	0.778	0.700	0.865	0.902	0.941	0.800
6	0.722	0.730	0.700	0.745	0.667	0.706	0.600	0.750	0.810	0.843	0.773	0.850
7	0.660	0.670	0.640	0.681	0.789	0.843	0.833	0.750	0.829	0.863	0.810	0.850
Bagging	0.796	0.800	0.780	0.813	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
No bagging	0.736	0.770	0.640	0.865	0.639	0.795	0.672	0.610	0.760	0.856	0.754	0.766

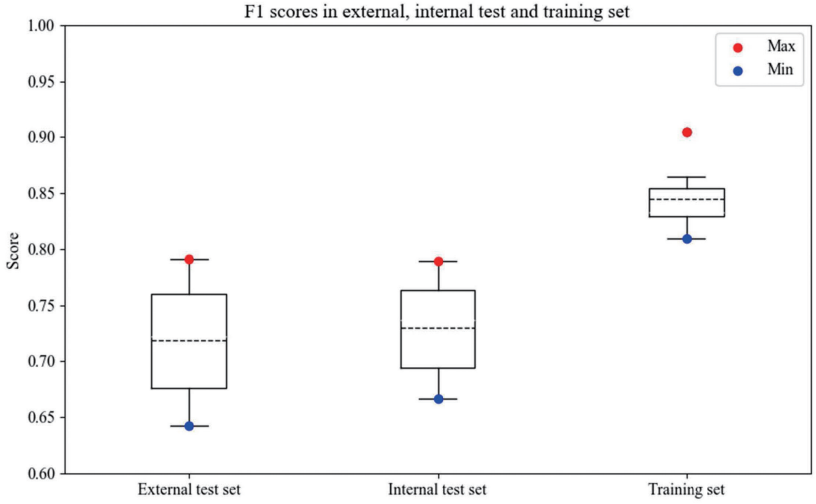

Fig. 3. F1 Score distributions of sub-AI models in training, internal, and external test sets. The average F1 score for the external test, internal test, and the training set is 0.718, 0.730, and 0.845, respectively.

pared it with that of pharmaceuticals currently marketed in the United States. Figure 2 shows the chemical spaces for both dataset compounds and pharmaceuticals. The chemical space of the dataset largely overlapped with that of pharmaceuticals.

Performance of AI models

A model constructed using the no-bagging method and evaluated on the external test set yielded an F1 score of 0.736, which did not indicate adequate prediction performance (Table 2). The AI model constructed using the bagging method achieved an F1 score of 0.796 on the external test set (Table 2). Each of the seven sub-models was evaluated on an external test set, yielding an average F1 score of 0.718 with a standard deviation of 0.056 (Fig. 3). The best-performing sub-model achieved an F1 score of 0.792 (Table 2). The predictive performance of the AI model was improved by applying the bagging method, highlighting the effectiveness of the ensemble approach.

Prediction of SAR examples and MMPS

Using the AI model, the lead compound SUN13837 in the SAR examples (Sakai *et al*., 2020) was predicted to be positive, with five out of seven sub-models indicating positivity (Table 3). In contrast, compounds 6 and 17, which were designed to reduce the risk of phospholipidosis, were correctly predicted as negative. The three compounds, SUN13837 and its derivatives, are visualized in Fig. 4. Thus, the proposed AI model demonstrates its

Table 3. Prediction results of 13 compounds of SAR examples and MMPs data.

Compound name	Label	Voting number	Predicted label	Category	Reference
SUN13837	pos	5	pos	**tp**	Sakai *et al.*, 2020
Compound_6	neg	3	neg	**tn**	
Compound_17	neg	3	neg	**tn**	
MS023	pos	6	pos	**tp**	Hu *et al.*, 2023
MS094	neg	5	pos	fp	
PPTN	pos	6	pos	**tp**	
PPTN-NC	neg	7	pos	fp	
TP-472	pos	4	pos	**tp**	
TP-472N	neg	5	pos	fp	
TP-064	pos	7	pos	**tp**	
TP-064N	pos	7	pos	**tp**	
SGC3027N	pos	4	pos	**tp**	
SGC3027	neg	3	neg	**tn**	

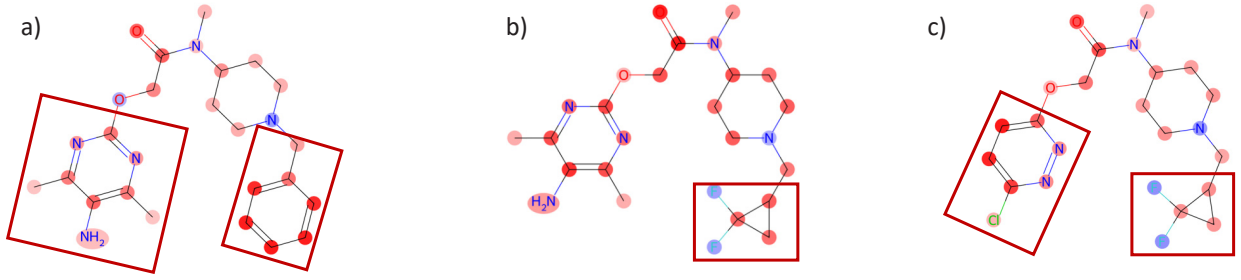

Fig. 4. Visualization results of a) SUN13837: phospholipidosis inducer, b) compound 6: phospholipidosis non-inducer, and c) compound 17: phospholipidosis non-inducer. All of them are visualized using sub-model 3, which correctly predicted all of them. The red rectangles from the left indicate: a) amino-dimethyl-pyrimidine moiety, benzyl group, b) difluoro-cyclopropyl group, and c) chloro-pyridazine moiety, difluoro-cyclopropyl group. In the structure, red indicates substructures contributing to a positive decision. The more intense the color, the higher the contribution.

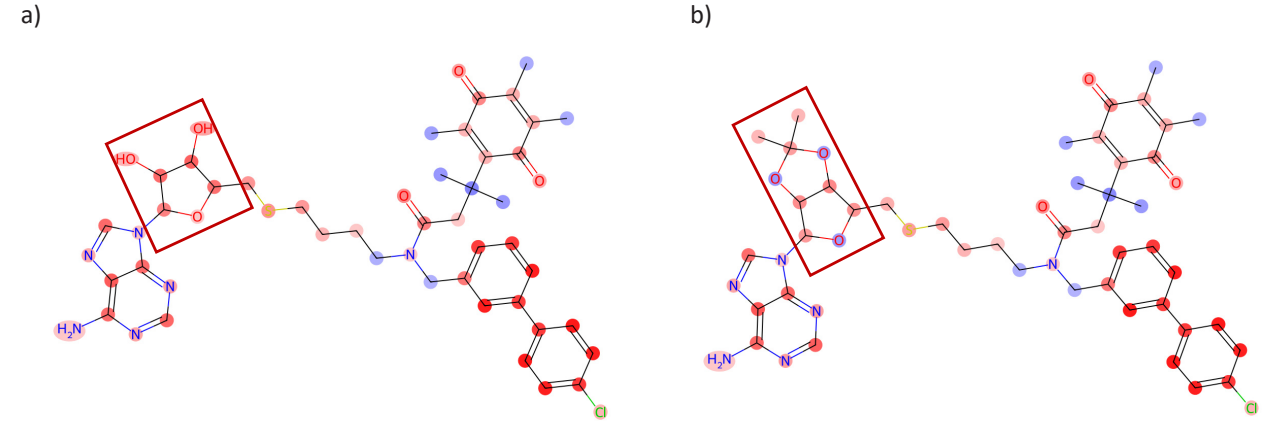

Fig. 5. Visualization results of a) SGC3027N: phospholipidosis inducer and b) SGC3027: phospholipidosis non-inducer. The red rectangles highlight the regions of the chemical structures that differ from one another.

robustness and applicability in a practical structural optimization context.

The next example is MMPs, which are pairs of chemical compounds that differ by a single structural modification (Griffen *et al*., 2011). Five pairs were used in this trial, including one positive pair. Of these five pairs, two were correctly predicted by both the compounds, whereas the remaining three were correctly predicted by only one compound in each pair. One of the MMP pairs shown in Fig. 5 is SGC3027N and SGC3027; *in vitro* experiments have shown that SGC3027N is a phospholipidosis inducer, whereas SGC3027 is a phospholipidosis non-inducer (Hu *et al*., 2023). Although these two compounds have similar physicochemical properties, including pKa and cLogP, only SGC3027N exhibited this induction capability. When similar compounds were predicted using this prediction tool, SGC3027N was predicted to be positive by a majority of four votes and SGC3027 was predicted to be negative by less than half (three votes). Both predictions accurately matched the experimental results.

The prediction results for the 13 compounds from these two sets are summarized in Table 3. The number of compounds correctly predicted as positive and negative, consistent with the experimental results, were seven out of seven and three out of six, respectively. Based on these results, the evaluation metric F1 score for the 13 compounds was calculated as 0.800.

DISCUSSION

The proposed AI model was constructed using an ensemble learning approach, in which the dataset was divided into subsets. In this method, the final decision is made by a majority vote of the multiple sub-models. For model development, the training set was split into seven subsets to ensure minimal overlap. This approach enabled the construction of a DIPL potential prediction AI model with a high discriminative performance, achieving an F1 score of 0.796. In contrast, the model without the bagging method evaluated on the same external test set yielded an F1 score of 0.736 (Table 2). The results demonstrated the effectiveness of ensemble learning in improving the predictive accuracy of the AI model.

The evaluation metrics for the AI model using 13 compounds related to the SAR examples and matched molecular pairs yielded an F1 score of 0.800. These 13 data points possess certain challenging features for prediction. For instance, the first dataset includes SUN13837, whose hydrolyzed metabolites induce phospholipidosis. Remarkably, the AI model successfully predicted the potential of the metabolites to induce phospholipidosis based on the chemical structure of the parent compound. Thus, even without predicting the chemical structures of the metabolites, the AI model has shown the potential to predict the induction of phospholipidosis by the metabolites.

Figure 4 presents visualizations of how the AI model evaluated SUN13837 and its derivatives, compounds 6 and 17. Compound 6 is a derivative in which the benzyl group of SUN13837, which originally exhibited DIPL-inducing activity, was replaced with a difluorocyclopropyl group, resulting in the loss of DIPL-inducing activity. Compound 17 was developed by substituting the amino-dimethyl-pyrimidine moiety in compound 6 with a chloropyridazine moiety, thereby improving its pharmacokinetic properties while maintaining the loss of DIPL-inducing activity (Sakai *et al.*, 2020). The AI model correctly predicted that SUN13837 is DIPL-positive, whereas compounds 6 and 17 are both DIPL-negative. The visualizations in Fig. 4 confirm that the AI recognized the benzyl group of SUN13837 as a strong contributor to DIPL activity. In compound 6, AI also identified fluorine atoms in the substituted difluorocyclopropyl group, contributing to the loss of DIPL activity. For compound 17, the AI model determined that substitution with a chloropyridazine moiety did not result in significant changes in its contribution to DIPL activity. Therefore, the proposed AI model demonstrated its ability to accurately show structural alerts associated with DIPL.

In the second dataset, MMPs consisted of pairs of chemical compounds that differed only by a single structural segment modification. In addition, the pairs were not only structurally similar but also similar in their pKa (a measure of ionization) and cLogP values (a measure of hydrophobicity). Therefore, even if the properties embedded within the chemical structures can be calculated, predictions must be made based primarily on structural differences. In this trial, five pairs were used (one positive pair); however, only two of the five pairs were correctly predicted. Figure 5 illustrates one of the correctly predicted pairs. When focusing on the differences between the two compounds, the positive compound, SGC3027N, contributed positively, as indicated by the red regions, whereas the negative compound, SGC3027, contributed more strongly in the negative direction, as shown by the blue regions. Although there is room for improvement in terms of predicting subtle structural changes, a certain degree of prediction performance has been confirmed.

We aimed to develop an AI model suitable for drug safety evaluation. Consequently, the predictive capability of the proposed AI model encompassed the chemical space pertinent to pharmaceuticals. A comparative anal-

ysis demonstrated that the chemical space of our dataset used to construct the AI model largely overlapped with the pharmaceutical chemical space (Fig. 2). However, as the pharmaceutical chemical space continues to expand, it is essential to update models using new datasets. Additionally, the predictive performance of the proposed AI models might not be optimal when applied to detailed SAR studies of specific chemical structures. Further training with datasets containing similar compounds may be required to enhance the predictive accuracy in such cases.

We developed an AI model to predict DIPL, achieving high predictive accuracy by integrating a GNN with bagging methods. DIPL potential is a critical parameter in the drug discovery phase, making the proposed AI model a valuable tool for the selection and prioritization of drug candidates. In addition, the proposed model is notable for its ability to visualize structural alerts and predict DIPL. This feature is particularly beneficial for modifying chemical structures to reduce the risk of inducing phospholipidosis.

ACKNOWLEDGMENT

This study was supported by the Japan Agency for Medical Research and Development (AMED) under Grant Number JP24nk0101111.

Conflict of interest---- The authors declare that there is no conflict of interest.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019): Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining), pp. 2623-2631.

Anderson, N. and Borlak, J. (2006): Drug-induced phospholipidosis. FEBS Lett., **580**, 5533-5540.

Breiman, L. (1996): Bagging predictors. Mach. Learn., **24**, 123-140.

Brody, S., Alon, U. and Yahav, E. (2021): How attentive are graph attention networks? arXiv preprint arXiv:2105.14491.

Cihan Sorkun, M., Mullaj, D., Koelman, J.V. and Er, S. (2022): ChemPlot, a python library for chemical space visualization. Chem. Methods, **2**, e202200005.

Funk, R.S. and Krise, J.P. (2012): Cationic amphiphilic drugs cause a marked expansion of apparent lysosomal volume: implications for an intracellular distribution-based drug interaction. Mol. Pharm., **9**, 1384-1395.

Fusani, L., Brown, M., Chen, H., Ahlberg, E. and Noeske, T. (2017): Predicting the Risk of Phospholipidosis with in Silico Models and an Image-Based in Vitro Screen. Mol. Pharm., **14**, 4346-4352.

Griffen, E., Leach, A.G., Robb, G.R. and Warner, D.J. (2011): Matched molecular pairs as a medicinal chemistry tool. J. Med. Chem., **54**, 7739-7750.

Halliwell, W.H. (1997): Cationic amphiphilic drug-induced phospholipidosis. Toxicol. Pathol., **25**, 53-60.

Hanumegowda, U.M., Wenke, G., Regueiro-Ren, A., Yordanova, R., Corradi, J.P. and Adams, S.P. (2010): Phospholipidosis as a function of basicity, lipophilicity, and volume of distribution of compounds. Chem. Res. Toxicol., **23**, 749-755.

Heyndrickx, W., Mervin, L., Morawietz, T., *et al*. (2024): MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information. J. Chem. Inf. Model., **64**, 2331-2344.

Hostetler, K.Y. and Matsuzawa, Y. (1981): Studies on the mechanism of drug-induced lipidosis. Cationic amphiphilic drug inhibition of lysosomal phospholipases A and C. Biochem. Pharmacol., **30**, 1121-1126.

Hu, H., Tjaden, A., Knapp, S., Antolin, A.A. and Muller, S. (2023): A machine learning and live-cell imaging tool kit uncovers small molecules induced phospholipidosis. Cell Chem Biol, **30**, 1634-1651 e1636.

Joshi, U.M., Kodavanti, P.R., Coudert, B., Dwyer, T.M. and Mehendale, H.M. (1988): Types of interaction of amphiphilic drugs with phospholipid vesicles. J. Pharmacol. Exp. Ther., **246**, 150-157.

Kasahara, T., Tomita, K., Murano, H., *et al*. (2006): Establishment of an *in vitro* high-throughput screening assay for detecting phospholipidosis-inducing potential. Toxicol. Sci., **90**, 133-141.

Kipf, T.N. and Welling, M. (2016): Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Kojima, R., Ishida, S., Ohta, M., Iwata, H., Honma, T. and Okuno, Y. (2020): kGCN: a graph-based deep learning framework for chemical structures. J. Cheminform., **12**, 32.

Kruhlak, N.L., Choi, S.S., Contrera, J.F., *et al*. (2008): Development of a phospholipidosis database and predictive quantitative structure-activity relationship (QSAR) models. Toxicol. Mech. Methods, **18**, 217-227.

Li, G., Xiong, C., Thabet, A. and Ghanem, B. (2020): Deepergcn: All you need to train deeper gcns. arXiv preprint arXiv:2006.07739.

Lowe, R., Glen, R.C. and Mitchell, J.B. (2010): Predicting phospholipidosis using machine learning. Mol. Pharm., **7**, 1708-1714.

Lowe, R., Mussa, H.Y., Nigsch, F., Glen, R.C. and Mitchell, J.B. (2012): Predicting the mechanism of phospholipidosis. J. Cheminform., **4**, 2.

McInnes, L., Healy, J. and Melville, J. (2018): Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mingeot-Leclercq, M.P., Brasseur, R. and Schanck, A. (1995): Molecular parameters involved in aminoglycoside nephrotoxicity. J. Toxicol. Environ. Health, **44**, 263-300.

Morris, C., Ritzert, M., Fey, M., *et al*. (2019): Weisfeiler and leman go neural: Higher-order graph neural networks. In: Proceedings of the AAAI conference on artificial intelligence), pp. 4602-4609.

Nath, A. and Sahu, G.K. (2019): Exploiting ensemble learning to improve prediction of phospholipidosis inducing potential. J. Theor. Biol., **479**, 37-47.

Orogo, A.M., Choi, S.S., Minnier, B.L. and Kruhlak, N.L. (2012): Construction and Consensus Performance of (Q)SAR Models for Predicting Phospholipidosis Using a Dataset of 743 Compounds. Mol. Inform., **31**, 725-739.

Pelletier, D.J., Gehlhaar, D., Tilloy-Ellul, A., Johnson, T.O. and Greene, N. (2007): Evaluation of a published in silico model and construction of a novel Bayesian model for predicting phospholipidosis inducing potential. J. Chem. Inf. Model., **47**, 1196-1205.

Ploemen, J.P., Kelder, J., Hafmans, T., *et al*. (2004): Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential: a case study with structurally related piperazines. Exp. Toxicol. Pathol., **55**, 347-355.

Przybylak, K.R., Alzahrani, A.R. and Cronin, M.T. (2014): How does the quality of phospholipidosis data influence the predictivity of structural alerts? J. Chem. Inf. Model., **54**, 2224-2232.

Przybylak, K.R. and Cronin, M.T. (2011): In Silico Studies of the Relationship Between Chemical Structure and Drug Induced Phospholipidosis. Mol. Inform., **30**, 415-429.

Sakai, H., Inoue, H., Murata, K., *et al*. (2020): Fibroblast growth factor receptor modulators employing diamines with reduced phospholipidosis-inducing potential. Bioorg. Med. Chem., **28**, 115562.

Sawada, H., Takami, K. and Asahi, S. (2005): A toxicogenomic approach to drug-induced phospholipidosis: analysis of its induction mechanism and establishment of a novel *in vitro* screening system. Toxicol. Sci., **83**, 282-292.

Shahane, S.A., Huang, R., Gerhold, D., Baxa, U., Austin, C.P. and Xia, M. (2014): Detection of phospholipidosis induction: a cell-based assay in high-throughput and high-content format. J. Biomol. Screen., **19**, 66-76.

Sun, H., Shahane, S., Xia, M., Austin, C.P. and Huang, R. (2012): Structure based model for the prediction of phospholipidosis induction potential of small molecules. J. Chem. Inf. Model., **52**, 1798-1805.

Sundararajan, M., Taly, A. and Yan, Q. (2017): Axiomatic attribution for deep networks. In: International conference on machine learning), pp. 3319-3328, PMLR.

Tomizawa, K., Sugano, K., Yamada, H. and Horii, I. (2006): Physicochemical and cell-based approach for early screening of phospholipidosis-inducing potential. J. Toxicol. Sci., **31**, 315-324.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T. and Weinberger, K. (2019): Simplifying graph convolutional networks. In: International conference on machine learning), pp. 6861-6871, PMLR.

Zhuang, J., Tang, T., Ding, Y., *et al*. (2020): Adabelief optimizer: adapting stepsizes by the belief in observed gradients. Adv. Neural Inf. Process. Syst., **33**, 18795-18806.