IJRDO Journal of Applied Science
ISSN:2455-6653
Review Article

EXPLAINABLE MACHINE LEARNING FOR PREDICTING MOLECULAR TOXICITY FROM PHYSICOCHEMICAL AND STRUCTURAL PROPERTIES

Sneha S Kulkarni

Assistant Professor, Department of Chemistry, KLS Vishwanathrao Deshpande Institute of Technology Haliyal, Uttar Kannada, karnataka, India.E-mail ID: snehakulkarni721@gmail.com https://orcid.org/0000-0003-0753-1237

Corresponding Author: Sneha S Kulkarni, Email: snehakulkarni721@gmail.com

Abstract

The prediction of molecular toxicity is becoming more and more critical for effective chemical safety assessment, drug development and environmental risk assessment. In this study, an explainable machine learning model for predicting molecular toxicity was developed using physicochemical descriptors and structural properties of Tox21 dataset. There were 7,831 molecules and 12 binary toxicity endpoints in the dataset, with 7,823 molecules remaining after structure validation for the molecules. Morgan molecular fingerprints were created along with physicochemical descriptors such as molecular weight, LogP, topological polar surface area, hydrogen-bond properties, rotatable bonds, heavy-atom count, ring properties, and Fraction Csp3. The accuracy, precision, recall, specificity, F1 score, balanced accuracy, ROC-AUC and PR-AUC were used to evaluate the following logistic regression, random forest, XGBoost and Support Vector Machine models. Random Forest had the best mean ROC-AUC (0.839) and PR-AUC (0.480) values while XGBoost had the best balanced accuracy (0.738) and F1 score (0.402). Generally, active molecules were larger, more lipophilic and more aromatic than inactive compounds for the representative SR-ARE endpoint. The molecular weight, LogP, Fraction Csp3, TPSA, number of heavy atoms and aromatic properties were found to be influential predictors using SHAP analysis. In general, this work shows that explainable machine learning offers accurate and interpretable molecular toxicity predictions to enable compound prioritization and computational chemical-safety assessment.

Key Words: Molecular toxicity; Explainable machine learning; Tox21; Physicochemical descriptors; SHAP

DOI: 10.69980/as.v12i2.6805
Received:11-01-2026 | Revised:08-03-2026 | Accepted:11-04-2026 | Published:30-04-2026
Copyright: © 2026 The Author(s). This work is licensed under a CC BY-NC-ND 4.0 International License.

Introduction

Molecular toxicity testing is a key aspect of chemical safety testing, drug research, environmental surveillance and human-health safety. There are thousands of synthetic and naturally occurring compounds whose possible adverse biological effects need to be considered, but due to the large number of chemicals and their structural diversity, experimental testing of them is challenging. The latest developments in data-driven modelling have demonstrated that molecular representations are able to encapsulate complex chemical motifs related to toxicological responses and enable faster chemical risk screening (Gao et al. 2022). Graph-based methods have also shown that molecular topology can be easily used to enhance toxicity prediction by modeling atoms as nodes and chemical bonds as edges (D. Jiang et al. 2021).

Conventional toxicity testing usually uses in vitro tests, animal testing, and other lab-based tests. Biological variability, assay-specific uncertainty, and variability in testing conditions also influence experimental results. The same sources of uncertainty and bias may affect computational quantitative structure-activity relationship models in cases where datasets or domains of applicability are poorly defined (Cronin, Richarz, and Schultz 2019). In turn, there is an increasing necessity to use alternative predictive methods that assist in more rapid and systematic chemical prioritization.

Surveys of chemical safety applications have shown that the increasing number of machine learning and deep learning applications in toxicity classification, hazard identification, and chemical risk assessment (Jiao et al. 2020). It has also been demonstrated with semi-supervised graph convolutional networks that labelled and unlabelled compound information can be used to enhance chemical toxicity modelling (Chen et al. 2021). More specific deep-learning models, such as DeepDILI, have shown the usefulness of learned molecular representations to predict drug-induced liver injury (Li et al. 2021) and multi-task learning concurrently utilize the knowledge shared among multiple toxicity endpoints (Yuan Li et al. 2023).

Major important descriptors are molecular weight, lipophilicity expressed as LogP, topological polar surface area, hydrogen-bond acceptors and donors, rotatable bonds, aromaticity, and ring-related. Mordred and other tools can be used to systematically compute large numbers of molecular descriptors of chemical structures, which then be modelled (Moriwaki et al. 2018). Along with traditional descriptors, molecular fingerprints and graph structures encode structural fragments and connectivity patterns. Comparisons of the descriptors-based and graph-based models have revealed that both types of features complement each other to predict molecular properties (J. Jiang, Wang, and Wei 2021). Massive assessments have also demonstrated the high performance of graph neural networks in molecular supervised learning (Hwang et al. 2020) and pharmacophore-integrated graph approaches enhance prediction and chemical understanding (Kong et al. 2022).

Although complex machine-learning models have predictive benefits, their internal decision processes are frequently subject to criticism as being hard to interpret. This is especially significant in the field of toxicology where scientists and regulators must be able to comprehend the reasons why a compound is considered potentially toxic. This limitation can be overcome with explainable artificial intelligence that recognizes influential molecular features and explains their contribution to single predictions. SHAP-based approaches also give local and global explanations about tree-based machine-learning models (Lundberg et al. 2020). Explainability is also becoming a concept to be considered as crucial to trustworthy artificial intelligence in chemical and drug-discovery applications (Jiménez-Luna, Grisoni, and Schneider 2020). Recent toxicity investigations have shown that explainable machine-learning techniques discover chemical properties related to acute dermal toxicity and enhance interpretation of model results (Lou et al. 2024).

Examples of geometric graph-learning-based methods include GGL-Tox, which demonstrates how structure-based methods have high predictive power (D. Jiang et al. 2021) or packages like Chemprop, which make use of machine learning to predict properties of molecules based on their structure (Heid et al. 2024). Nonetheless, models

properties of molecules based on their structure (Heid et al. 2024). Nonetheless, models integrating competitive predictive capabilities with easily comprehensible chemically significant physicochemical and structural properties are still in demand.
Based on this, this study test explainable machine-learning models to predict molecular toxicity with physicochemical and structural properties. These are aimed at obtaining useful molecular descriptors and structural properties, comparing various classification models, assessing their predictive accuracy across the toxicity endpoints, determining which molecular properties are most predictive, and using explainable artificial intelligence models to understand the role of individual features in predicting toxicity. The framework is supposed to facilitate correct, interpretable, and useful computational toxicity evaluation.

Materials and Methods

Study Design and Dataset

This is a quantitative computational study in which models for toxicity prediction that can be explained were developed from the publicly available Tox21 molecular toxicity dataset. There were 7,831 unique molecules in the dataset and 12 binary toxicity endpoints along with two other variables (mol_id, SMILES representations). Molecular preprocessing, descriptor and fingerprint generation, exploratory analysis, model development, performance evaluation and explainability analysis were all part of the analytical workflow (Chakraborty 2024).

Toxicity Endpoints

The 12 endpoints comprised seven nuclear receptor assays: NR-AR, NR-AR-LBD, NR-AhR, NR-Aromatase, NR-ER, NR-ER-LBD and NR-PPAR-gamma; and five stress-response assays: SR-ARE, SR-ATAD5, SR-HSE, SR-MMP and SR-p53. Each endpoint was modelled as an independent binary classification (1 = active, 0 = inactive) event. Molecules that did not have a toxicity label were only excluded from the endpoint analysis for which the label was missing.

Data Preprocessing

The data set was checked for missing values, duplicate data entries, duplicate molecular structures, and invalid outcome values. No duplicate molecular records or SMILES representations were found. Before SMILES strings were converted to molecular objects and evaluated for structural validity for feature generation. To maximize the number of molecular data that can be retained without needlessly removing compounds, missing toxicity outcome data were treated separately by each endpoint.

Physicochemical and Structural Feature Generation

Validated SMILES structures were used to calculate the physicochemical descriptors using RDKit. These key properties were molecular weight, LogP, topological polar surface area (TPSA), hydrogen-bond donors and acceptors, rotatable bonds, heavy-atom count, total and aromatic ring characteristics, and Fraction Csp3. Moreover, molecular structural information was also modeled via Morgan/extended-connectivity fingerprints, which were used to model local atomic environments and molecular substructures as fixed-length feature vectors. The models were able to integrate both global and structural properties as descriptors and fingerprints were combined.

Exploratory Data Analysis

Major physicochemical descriptors were plotted graphically and values calculated by descriptive statistics. To find any relationship or potential multicollinearity among the continuous molecular properties, correlation analysis was performed. The toxicity class distributions were calculated for each endpoint separately. Class imbalance was taken into account in the development and evaluation of models since the number of active compounds was less in each assay. All continuous predictors

were standardized, using parameters estimated only from the training data, if necessary.

Machine-Learning Model Development

Supervised classification algorithms were compared: Logistic Regression, Random Forest, Gradient Boosting/XGBoost, Support Vector Machine. Logistic Regression was used as an interpretable baseline while the other algorithms enabled the detection of nonlinear relationships and interactions between molecular characteristics. Using the combined set of physicochemical and structural features, separate models were built for each toxicity end-point.

Model Training and Validation

Labelled compounds were split into training and independent test sets for each endpoint in a stratified fashion to maintain the toxicity class proportions. Stratified cross-validation was used to optimize the models within the training data. The hyperparameters were suitably tuned and class weighting or similar imbalance-specific methods were used if necessary. To reduce information leakage, all procedures of preprocessing and model selection were done on training data only.

Performance Evaluation

The model performance was evaluated based on accuracy, precision, recall/sensitivity, specificity, F1-score, balanced accuracy, ROC-AUC and PR-AUC. Due to the imbalance between active and inactive compounds, special focus was given to F1-score, balanced accuracy, ROC-AUC, PR-AUC. Confusion matrices were also created, along with ROC curves. The comparative performance of the algorithms was used to find the best model for toxicity prediction.

Explainability and Computational Analysis

The best-performing compatible model was then used to calculate the importance of the molecular features in the toxicity predictions using SHapley Additive exPlanations (SHAP). Global feature importance and SHAP summary analyses were used to explore which physicochemical and structural characteristics were influential and if these characteristics led to an increase or decrease in predicted toxicity. Analysis was done in Python with RDKit, pandas, NumPy, scikit-learn, XGBoost, SHAP and Matplotlib. This framework allowed for concurrent evaluation of predictive performance and interpretable molecular determinants of toxicity.

Results

Dataset and Molecular Characteristics

The original Tox21 dataset comprised 7,831 molecular records and 12 binary toxicity endpoints. After SMILES validation, 7,823 molecules were successfully converted to valid molecular structures, and eight records with invalid structural encodings were not analyzed using descriptors. Labelled compounds were not the same at all endpoints due to missing endpoint-specific toxicity outcomes. The most labelled molecules were found in NR-AR (n = 7,258) while the least were found in SR-MMP (n = 5,804). There was significant class imbalance across assays, with positive toxicity prevalence from 2.89% for NR-PPAR-gamma to 16.17% for SR-ARE. Inactive compounds significantly outnumbered active compounds for each endpoint, which lends credence to the use of imbalance-sensitive performance measures (see Table 1).

Table 1. Dataset characteristics and distribution of toxicity endpoints

Toxicity endpoint	Labelled molecules	Missing labels	Active, n	Inactive, n	Active (%)
NR-AR	7,258	565	308	6,950	4.24
NR-AR-LBD	6,751	1,072	237	6,514	3.51
NR-AhR	6,542	1,281	768	5,774	11.74
NR- Aromatase	5,815	2,008	300	5,515	5.16
NR-ER	6,186	1,637	791	5,395	12.79
NR-ER-LBD	6,948	875	349	6,599	5.02
NR-PPAR-gamma	6,443	1,380	186	6,257	2.89
SR-ARE	5,825	1,998	942	4,883	16.17
SR-ATAD5	7,065	758	264	6,801	3.74
SR-HSE	6,460	1,363	372	6,088	5.76
SR-MMP	5,804	2,019	918	4,886	15.82
SR-p53	6,767	1,056	423	6,344	6.25

Physicochemical Characteristics of Molecules

The chemically analysed molecules were quite varied in terms of their physicochemical properties. The mean molecular weight was 276.14 Da ± 164.73 Da and the median molecular weight was 240.30 Da. The mean logP was 2.37 ± 2.30, and there was considerable variability in molecular lipophilicity. The mean TPSA was59.47 ± 57.80 Å². The molecules had an average number of 1.23 hydrogen-bond donors, 3.56 hydrogen-bond acceptors, 4.30 rotatable bonds, and 18.57 heavy atoms. The mean total ring count was 1.77 with 1.10 aromatic rings per molecule and a mean Fraction Csp3 of 0.46. The molecular weight, LogP, TPSA, hydrogen-bonding properties, number of rotatable bonds, number of rings, and Fraction Csp3 distributions of all 7,823 valid molecular structures are shown in Figure 1.

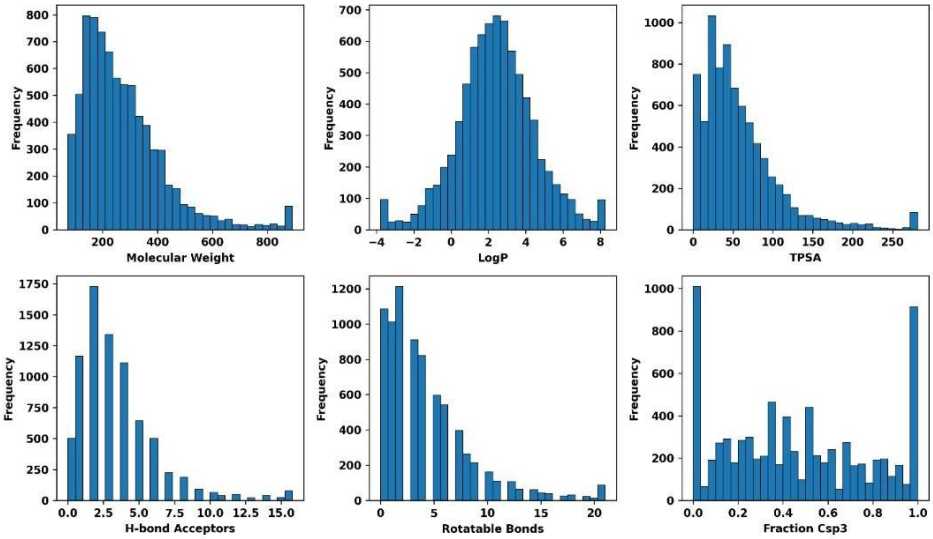


Figure 1. Distribution of major physicochemical molecular descriptors

The distributions were generally right-skewed for molecular weight, TPSA, and rotatable bonds because a relatively small number of structurally large and complex molecules produced high upper-range values.

Relationships Among Molecular Properties

There were many strong correlations between several physicochemical descriptors. Heavy-atom count exhibited a strong correlation with molecular weight (r = 0.959), and TPSA was correlated strongly with the number of hydrogen-bond acceptors (r = 0.908) and hydrogen-bond donors (r = 0.828). There was also a positive correlation between heavy atom count and ring count (r = 0.723). LogP, on the other hand, showed a negative correlation with TPSA (r = -0.375) as well as hydrogen-bond donor count (r = -0.423).

between heavy atom count and ring count (r = 0.723). LogP, on the other hand, showed a negative correlation with TPSA (r = -0.375) as well as hydrogen-bond donor count (r = -0.423). The number of aromatic rings was found to be significantly inversely correlated with Fraction Csp3 (r = -0.609). As seen from Table 2, the most compact correlation groups were the molecular size-related and polarity-related descriptors while Fraction Csp3 exhibited an inverse correlation with aromaticity.

Table 2. Correlation matrix of selected molecular descriptors

Descriptor	MolWt	LogP	TPSA	HBA	Heavy atoms	Aromatic rings	Fraction Csp3
MolWt	1.000	0.275	0.695	0.745	0.959	0.435	0.017
LogP	0.275	1.000	-0.375	-0.263	0.279	0.318	-0.036
TPSA	0.695	-0.375	1.000	0.908	0.709	0.239	-0.034
HBA	0.745	-0.263	0.908	1.000	0.766	0.282	0.033
Heavy atoms	0.959	0.279	0.709	0.766	1.000	0.482	0.018
Aromatic rings	0.435	0.318	0.239	0.282	0.482	1.000	-0.609
Fraction Csp3	0.017	-0.036	-0.034	0.033	0.018	-0.609	1.000

Molecular Properties According to Toxicity Status

SR-ARE was taken as a representative endpoint for detailed comparison as it had the highest percentage of active molecules. The mean molecular weight of active SR-ARE compounds was higher compared to inactive ones (297.89 vs. 234.74 Da), as well as their mean Lipophilicity LogP (3.00 vs. 2.04). The active compounds also exhibited higher TPSA, heavy-atom count, total ring count and aromatic ring count. However, Fraction Csp3 was lower in the case of active compounds (0.341) as compared to inactive molecules (0.485), indicating the presence of more aromatic and less saturated structural character in active compounds. As can be seen in Figure 2, the active SR-ARE molecules had generally higher G, L, A, and Ring values and lower Fraction Csp3 values than the inactive molecules.

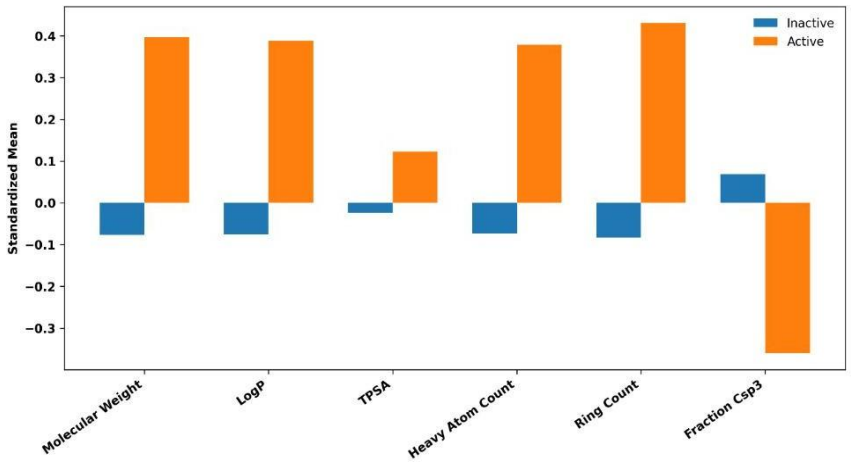


Figure 2. Comparison of key molecular properties between active and inactive SR-ARE compounds

Machine-Learning Model Performance

Performance of the models in all 12 toxicity endpoints was summarized. Random Forest has the best mean ROC-AUC of 0.839 and PR-AUC of 0.480 and thus the best overall ranking

capability. XGBoost had the largest mean F1-score (0.402) and balanced accuracy (0.738), and much higher recall than Random Forest. Random Forest demonstrated high mean accuracy (0.934) and specificity (0.996) but relatively low recall (0.178) which are indicative of a conservative prediction of the minority active class. Support Vector Machine and Logistic Regression showed equal overall performance. Random Forest gave the highest overall discrimination as seen in Table 3 and XGBoost gave a better trade-off between sensitivity and specificity.

Table 3. Mean performance of machine-learning models across the 12 toxicity endpoints

Model	Accuracy	Precision	Recall	Specificity	F1-score	Balanced accuracy	ROC-AUC	PR-AUC
Logistic Regression	0.764	0.189	0.667	0.769	0.285	0.718	0.779	0.279
Random Forest	0.934	0.747	0.178	0.996	0.271	0.587	0.839	0.480
XGBoost	0.873	0.313	0.585	0.892	0.402	0.738	0.820	0.437
Support Vector Machine	0.762	0.187	0.661	0.767	0.282	0.714	0.778	0.276

Discrimination Performance of the Best Models

For the representative SR-ARE toxicity endpoint, XGBoost achieved the highest ROC-AUC (0.853), closely followed by Random Forest (ROC-AUC = 0.850). Support Vector Machine achieved an ROC-AUC of 0.823, while Logistic Regression showed the lowest discrimination performance (ROC-AUC = 0.766). These findings indicate that the ensemble-based XGBoost and Random Forest models provided stronger discrimination between active and inactive SR-ARE compounds than Logistic Regression and Support Vector Machine. As illustrated in Figure 3, the ROC curves of XGBoost and Random Forest were consistently positioned above those of the other evaluated models across much of the false-positive-rate range. The relatively similar ROC-AUC values obtained by XGBoost and Random Forest further demonstrate that both ensemble approaches were effective for distinguishing toxicity classes for the representative SR-ARE endpoint.

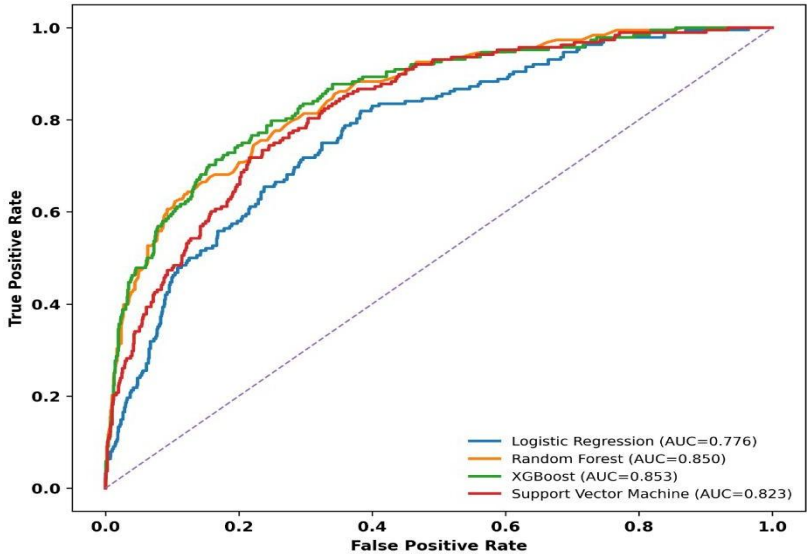


Figure 3. ROC curves of machine-learning models for the SR-ARE toxicity endpoint

Explainability and Important Molecular Features

Global physicochemical descriptors were found to be very important for the Random Forest model in the case of SR-ARE. Molecular weight, followed by LogP, Fraction Csp3, TPSA, number of heavy atoms, and number of rotatable bonds, were the most important features. Some Morgan fingerprint bits were also important predictors, with both global physicochemical properties and local molecular substructures playing a role in prediction. Table 4 reveals that the molecular size, lipophilicity, saturation, polarity and aromatic structural characteristics were more important than the individual bits of the fingerprint in the representative model.

Table 4. Most important molecular features contributing to SR-ARE toxicity prediction

Rank	Molecular feature	Feature importance
1	Molecular weight	0.0673
2	LogP	0.0639
3	Fraction Csp3	0.0475
4	TPSA	0.0429
5	Heavy-atom count	0.0389
6	Rotatable bonds	0.0257
7	Hydrogen-bond acceptors	0.0234
8	Ring count	0.0222
9	Aromatic ring count	0.0210
10	Hydrogen-bond donors	0.0185
11	Morgan fingerprint bit 33	0.0106
12	Morgan fingerprint bit 107	0.0102

SHAP-Based Interpretation of Toxicity Prediction

The direction of the features' effect in the Random Forest model was also further elucidated using SHAP analysis. The mean absolute SHAP value was the highest for molecular weight, followed by LogP, Fraction Csp3, number of heavy atoms, number of aromatic rings, and TPSA. There was a general trend of an increase in predicted SR-ARE activity as molecular weight, LogP, number of heavy atoms, number of rings, and number of aromatic rings increased with positive SHAP contributions. However, Fraction Csp3 had a strong negative correlation with SHAP values, meaning that molecules with a higher three-dimensional saturated carbon fraction were less likely to receive a high predicted toxicity score. As seen in Figure 4, generally an increase in molecular weight, lipophilicity, aromaticity, and ring complexity favored predictions to the active toxicity class, while an increase in Fraction Csp3 favored predictions to the inactive class.

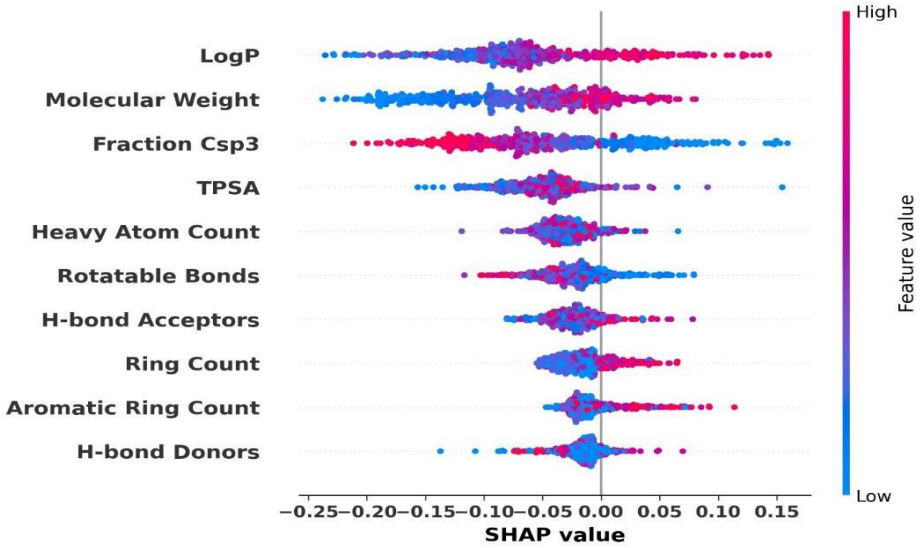


Figure 4. SHAP summary plot of molecular features influencing SR-ARE toxicity predictions

Overall, the explainability findings were consistent with the descriptive comparisons between active and inactive molecules and demonstrated that interpretable physicochemical properties, together with structural fingerprint information, contributed meaningfully to molecular toxicity prediction.

Discussion

The present study showed that molecular toxicity can be effectively predicted by using the structural fingerprint information along with physicochemical descriptors. Overall, Random Forest had the highest discrimination across the evaluated endpoints, while XGBoost had the best balance between sensitivity and specificity. Explainability analysis additionally revealed that MW, LogP, TPSA, heavy Atom Count, aromatic ring features, and Fraction Csp3 were among the most predictive factors. The overall significance of these results is that traditional QSAR approaches can be combined with current machine-learning-based toxicity prediction methods (Muratov et al. 2020). They also align with previous findings that complex representation-learning methods lead to better predictions in molecular toxicity tasks (Peng et al. 2020).

Generally, higher molecular weight and LogP were correlated with higher predicted toxicity, while Fraction Csp3 was inversely related to the representative endpoint. Such patterns are biologically realistic because they involve lipophilicity and polarity and these properties influence membrane permeability, distribution and interactions with biological macromolecules. The critical role of molecular descriptors in machine-learning-based toxicity assessment has also been highlighted in previous reviews (Tang et al. 2018). In addition, combinations of physicochemical descriptors are effective at distinguishing toxicological outcomes in environmental toxicology via QSAR (Ai et al. 2019).

By capturing local atomic environments and recurring substructures, Morgan fingerprints brought extra structural information beyond global physicochemical descriptors. The additional compounds were able to enhance the models discrimination among compounds of similar overall physicochemical profiles with different structural arrangements. This result is consistent with other findings that learned molecular representations are able to represent important molecular properties for predicting molecular properties (Yang et al. 2019). Graph-attention methods have also shown that local molecular context and structural connectivity significantly boost predictive modelling in drug discovery (Xiong et al. 2020). More recent contrastive graph-learning methods further support the importance of rich molecular representations in chemical prediction tasks (Wang et al. 2022).

Random Forest achieved the best results in ROC-AUC and PR-AUC, likely due to its ability to capture nonlinear relationships and interactions among the various molecular descriptors and fingerprint variables used. XGBoost, on the other hand, had superior balanced accuracy and F1 score, which reflect better minority-class recognition. This difference is significant because conventional accuracy is not sufficient for comparing models (Johnson and Khoshgoftaar 2019) on toxicity datasets that are usually highly imbalanced. Ensemble and nonlinear models perform well in line with previous works in the machine learning field for predicting drug toxicity (Zhang et al. 2018). Benchmark studies also have indicated that the level of model effectiveness is highly dependent on the molecular representation, the characteristics of the endpoints, and the evaluation approach (Wu et al. 2018).

The SHAP analysis showed that higher molecular weight, greater lipophilicity, heavier atoms, more rings, and more aromaticity tended to move the predictions towards the active toxicity class, while higher Fraction Csp3 moved them away. These explanations provide chemically meaningful insights into model behaviour and increase transparency in molecular risk prediction. This interpretability is especially crucial in high-stakes scientific use cases where using black-box models exclusively hinders trust and real-world deployment (Rudin 2019). Explainable AI is known to be crucial for understanding, validating, and communicating complex model decisions (Arrieta et al. 2020).

The results reported herein agree with other computational toxicology studies that have shown that structural and assay-based information provide strong toxicity prediction. For instance, the

chemical structure, combined with in vitro assay data, can be used to improve the prediction of drug-induced injury to the liver and cardiotoxicity (Ye et al. 2022). Likewise, molecular benchmark studies have shown that proper representation and optimized learning algorithms are essential to obtain reliable predictive performance (Wu et al. 2018).

The prepared framework can be used in early chemical screening, pharmaceutical development, environmental hazard assessment, and prioritization of compounds for experimental testing. Explainable predictions can be used to aid in identifying molecular properties of toxicity and to aid in more targeted follow-up experiments. This computational prioritization limits the number of lab tests and enhances the efficiency of chemical-safety evaluations.

Some restrictions must be noted. Toxicity endpoints were unbalanced and had missing labels, potentially affecting model stability. The analysis was also based on the descriptors derived from SMILES and was limited to Tox21 outcomes for each assay. The use of external datasets, further descriptors, deep-learning graph architectures, and future validation should be assessed. Generalizability and relevance are likely to be further enhanced by wider adoption of QSAR, explainable AI and mechanistic toxicology.

Conclusion

In this study, explainable machine-learning models were developed to accurately predict molecular toxicity, combining physicochemical descriptors with structural fingerprint data. The Tox21 dataset was analyzed to identify notable molecular features that significantly helped discriminate between toxicity classes: molecular weight, LogP, topological polar surface area, number of heavy atoms, aromatic ring features, and Fraction Csp3. Overall, the best discriminatory performance was obtained by the Random Forest algorithm, whereas the highest balance between sensitivity and specificity was obtained by the XGBoost algorithm, demonstrating that Random Forest and XGBoost are valuable ensemble methods in imbalanced toxicological datasets evaluated in terms of ROC-AUC and PR-AUC. The addition of Morgan fingerprints further enriched the structural representation as it captured the local molecular substructures which were not fully described by the global physicochemical characteristics. Explainability analysis with SHAP gave further insight into the direction and relative importance of the molecular features, and demonstrated that the larger the molecule, the more lipophilic, the more aromatic, the more complex the ring system, the more likely it was to be predicted to be more toxic in the active toxicity class, while greater Fraction Csp3 seemed to decrease predicted activity. In general, the study underscores the potential of interpretable machine learning in the context of chemical hazard screening and prioritization. But it is important to take into account the class imbalance, the fact that toxicity labels are sometimes missing, the different results from different assays, and the lack of external validation of the results when interpreting the results. Independent datasets and advanced molecular representations should be included in future studies to boost generalizability and predictive robustness.

References

1. Ai, Haixin, Xuewei Wu, Li Zhang, Mengyuan Qi, Ying Zhao, Qi Zhao, Jian Zhao, and Hongsheng Liu. 2019. “QSAR Modelling Study of the Bioconcentration Factor and Toxicity of Organic Compounds to Aquatic Organisms Using Machine Learning and Ensemble Methods.” Ecotoxicology and Environmental Safety 179:71–78.
2. Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.” Information Fusion 58:82–115.
3. Chakraborty, Soumyadeep. 2024. “Tox21 Dataset.” https://www.kaggle.com/datasets/epicskills/tox21-dataset.
4. Chen, Jiarui, Yain-Whar Si, Chon-Wai Un, and Shirley W. I. Siu. 2021. “Chemical Toxicity Prediction Based on Semi-Supervised Learning and Graph Convolutional Neural Network.” Journal of Cheminformatics 13(1):93. doi:10.1186/s13321-021-

00570-8.
5. Cronin, Mark TD, Andrea-Nicole Richarz, and Terry W. Schultz. 2019. “Identification and Description of the Uncertainty, Variability, Bias and Influence in Quantitative Structure-Activity Relationships (QSARs) for Toxicity Prediction.” Regulatory Toxicology and Pharmacology 106:90–104.
6. Gao, Feng, Wei Zhang, Andrea A. Baccarelli, and Yike Shen. 2022. “Predicting Chemical Ecotoxicity by Learning Latent Space Chemical Representations.” Environment International 163:107224.
7. Heid, Esther, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. 2024. “Chemprop: A Machine Learning Package for Chemical Property Prediction.” Journal of Chemical Information and Modeling 64(1):9–17.
8. Hwang, Doyeong, Soojung Yang, Yongchan Kwon, Kyung Hoon Lee, Grace Lee, Hanseok Jo, Seyeol Yoon, and Seongok Ryu. 2020. “Comprehensive Study on Molecular Supervised Learning with Graph Neural Networks.” Journal of Chemical Information and Modeling 60(12):5936–45.
9. Jiang, Dejun, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. 2021. “Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models.” Journal of Cheminformatics 13(1):12. doi:10.1186/s13321-020-00479-8.
10. Jiang, Jian, Rui Wang, and Guo-Wei Wei. 2021. “GGL-Tox: Geometric Graph Learning for Toxicity Prediction.” Journal of Chemical Information and Modeling 61(4):1691–1700.
11. Jiao, Zeren, Pingfan Hu, Hongfei Xu, and Qingsheng Wang. 2020. “Machine Learning and Deep Learning in Chemical Healthand Safety: A Systematic Review of Techniques and Applications.” ACS Chemical Health & Safety 27(6):316–34. doi:10.1021/acs.chas.0c00075.
12. Jiménez-Luna, José, Francesca Grisoni, and Gisbert Schneider. 2020. “Drug Discovery with Explainable Artificial Intelligence.” Nature Machine Intelligence 2(10):573–84.
13. Johnson, Justin M., and Taghi M. Khoshgoftaar. 2019. “Survey on Deep Learning with Class Imbalance.” Journal of Big Data 6(1):27. doi:10.1186/s40537-019-0192-5.
14. Kong, Yue, Xiaoman Zhao, Ruizi Liu, Zhenwu Yang, Hongyan Yin, Bowen Zhao, Jinling Wang, Bingjie Qin, and Aixia Yan. 2022. “Integrating Concept of Pharmacophore with Graph Neural Networks for Chemical Property Prediction and Interpretation.” Journal of Cheminformatics 14(1):52. doi:10.1186/s13321-022-00634-3.
15. Li, Ting, Weida Tong, Ruth Roberts, Zhichao Liu, and Shraddha Thakkar. 2021. “DeepDILI: Deep Learning-Powered Drug-Induced Liver Injury Prediction Using Model-Level Representation.” Chemical Research in Toxicology 34(2):550–65.
16. Lou, Shang, Zhuohang Yu, Zejun Huang, Haoqiang Wang, Fei Pan, Weihua Li, Guixia Liu, and Yun Tang. 2024. “In Silico Prediction of Chemical Acute Dermal Toxicity Using Explainable Machine Learning Methods.” Chemical Research in Toxicology 37(3):513–24. doi:10.1021/acs.chemrestox.4c00012.
17. Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. “From Local Explanations to Global Understanding with Explainable AI for Trees.” Nature Machine Intelligence 2(1):56–67.
18. Moriwaki, Hirotomo, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. 2018. “Mordred: A Molecular Descriptor Calculator.” Journal of Cheminformatics 10(1):4. doi:10.1186/s13321-018-0258-y.
19. Muratov, Eugene N., Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, and Adrian Roitberg. 2020. “QSAR without Borders.” Chemical Society Reviews 49(11):3525–64.

20. Peng, Yuzhong, Ziqiao Zhang, Qizhi Jiang, Jihong Guan, and Shuigeng Zhou. 2020. "TOP: A Deep Mixture Representation Learning Method for Boosting Molecular Toxicity Prediction." Methods 179:55–64.
21. Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." Nature Machine Intelligence 1(5):206–15.
22. Tang, Weihao, Jingwen Chen, Zhongyu Wang, Hongbin Xie, and Huixiao Hong. 2018. "Deep Learning for Predicting Toxicity of Chemicals: A Mini Review." Journal of Environmental Science and Health, Part C 36(4):252–71. doi:10.1080/10590501.2018.1537563.
23. Wang, Yuyang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022.