ACS OMEGA

Open Access

This article is licensed under CC-BY-NC-ND 4.0

http://pubs.acs.org/journal/acsodf

Article

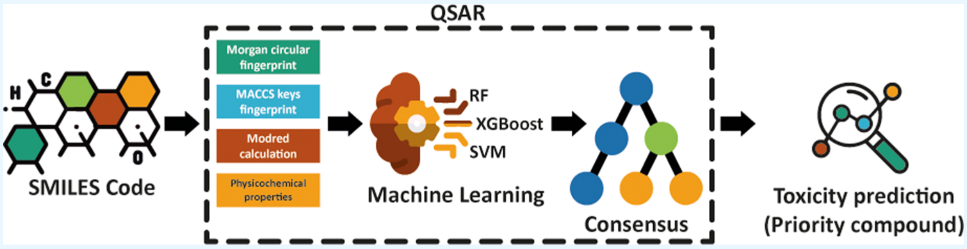
QSAR Classification Modeling Using Machine Learning with a Consensus-Based Approach for Multivariate Chemical Hazard End Points

Yunendah Nur Fuadah, Muhammad Adnan Pramudito, Lulu Firdaus, Frederique J. Vanheusden, and Ki Moo Lim*

Cite This: *ACS Omega* 2024, 9, 50796−50808

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: This study introduces an innovative computational approach using hybrid machine learning models to predict toxicity across eight critical end points: cardiac toxicity, inhalation toxicity, dermal toxicity, oral toxicity, skin irritation, skin sensitization, eye irritation, and respiratory irritation. Leveraging advanced cheminformatics tools, we extracted relevant features from curated data sets, incorporating a range of descriptors such as Morgan circular fingerprints, MACCS keys, Mordred calculation descriptors, and physicochemical properties. The consensus model was developed by selecting the best-performing classifier—Random Forest (RF), eXtreme Gradient Boosting (XGBoost), or Support Vector Machines (SVM)—for each descriptor, optimizing predictive accuracy and robustness across the end points. The model obtained strong predictive performance, with area under the curve (AUC) scores ranging from 0.78 to 0.90. This framework offers a reliable, ethical, and effective in silico approach to chemical safety assessment, underscoring the potential of advanced computational methods to support both regulatory and research applications in toxicity prediction.

INTRODUCTION

Toxicity can manifest in various forms, be quantified through measures including the lethal dose (LD50), or be described qualitatively in terms such as low, moderate, or high toxicity.[1,2] These assessments consider multiple factors, including the route and frequency of exposure, chemical properties, and biological characteristics of the subject.[3,4] Toxicity assessment is essential for understanding chemicals' potential risks to human health and the environment.[5] Traditionally, it has relied heavily on animal testing to identify harmful effects.[6] However, ethical considerations, high costs, and limited relevance to human biology raised questions about this approach.[7,8]

For example, the European Commission has highlighted the significant costs associated with toxicity testing under the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation, with expenses for a single chemical reaching millions of euros for long-term tests.[9,10] This has led to a growing shift toward more sophisticated methods that can reduce the dependence on animal testing while improving the accuracy and relevance of toxicity predictions. Regulatory and scientific communities, such as the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM 2018), have increasingly embraced New Approach Methods (NAMs), which emphasize the reduction, refinement, and replacement of animal testing.[11−13] The financial and ethical challenges have prompted the 2018 ICCVAM strategic roadmap and the U.S. Environmental Protection Agency's plan to phase out mammalian studies by 2035 while underscoring the urgent need for reliable alternative methods.[13]

Among these alternatives, in silico toxicology has gained significant attention as a computational approach that complements and enhances traditional testing methods.[14,15] This field encompasses a range of techniques, including databases for chemical and toxicity data, molecular descriptor generation, and predictive modeling, such as Quantitative Structure−Activity Relationship (QSAR) models.[16−18] These

Received: October 14, 2024
Revised: November 15, 2024
Accepted: December 3, 2024
Published: December 12, 2024

computational tools analyze and predict chemical toxicity before substances are synthesized, offering a cost-effective and efficient approach to hazard identification.[19] Using in silico methods, researchers and regulatory bodies can more accurately predict toxicological outcomes, reduce the need for animal testing, and ensure that chemical safety assessments are ethical and reliable. Various research groups[20−22] have developed reliable QSAR models to predict the skin sensitization potential of chemicals. Alves et al. utilized molecule descriptors, Quantitative Neighborhoods of Atoms (QNA), and "biological" descriptors.[20] Their models demonstrated a 71% correct classification rate (CCR) for human skin sensitization prediction. Kang et al. developed machine learning models to predict skin irritation and corrosion of liquid chemicals. Their study calculated 34 physicochemical descriptors from chemical structures, which were curated and analyzed using 22 descriptors for model construction. The XGBoost model demonstrated an accuracy of 0.73, which was the highest compared to the other classifier models.[23]

A study by Lou et al. predicting chemical acute dermal toxicity used a machine learning model with molecular fingerprint, molecular descriptor, and molecular graph as features.[24] Their proposed model obtained an Area Under the Curve (AUC) score of 0.63 and 0.764 for the external validation data sets of rats and rabbits, respectively. The RespiraTox study developed a QSAR method to predict respiratory irritants, significantly contributing to the reduction in animal testing. By comparing several machine learning approaches, the study found that Gradient-Boosted Decision Trees (GBTs) obtained the highest area under the curve (AUC) of 0.71.[25]

Schieferdecker et al. proposed oral toxicity prediction based on fingerprints and molecular descriptors with a consensus classification model, including GBT, multilayer perceptron (MLP), and graph attention network.[26] They evaluated the classification performance of the majority voting classifier, stacking ensemble with logistic regression (LR), and MLP as a meta classifier. Their majority voting classifier obtained the highest performance with an AUC score of 0.72 compared to the LR and MLP stacking classifiers, which provided 0.70 and 0.71 AUC scores, respectively.

Several studies developed QSAR modeling for predicting hERG channel blocker-induced cardiac toxicity.[27−32] The DMFGAM model, a novel deep learning approach, integrates multiple molecular fingerprints and graph features to predict hERG channel blockers accurately.[33] Utilizing a data set of 10,355 compounds, standardized using the research and development kit (RDKit) and molecular validation and standardization (MolVS), the study calculated molecular fingerprints and represented each molecule as a graph. A simplified molecular input line entry system (SMILES) graph attention network (SGAT) processed these features using a multihead attention mechanism and then fed the combined features into a fully connected neural network for classification. Performance evaluation revealed that DMFGAM provided an accuracy of 0.82, an AUC of 0.89, a specificity of 0.78, and a sensitivity of 0.85, outperforming classical methods and highlighting its potential in early-stage drug discovery for assessing cardiotoxicity risks associated with hERG channel blocking.

The STopTox study established a comprehensive set of QSAR models which was designed to predict toxicity for six critical end points, including three topical end points (skin sensitization, skin irritation, and eye irritation) and three systemic end points (acute oral toxicity, acute inhalation toxicity, and acute dermal toxicity).[34] The study gathered and refined the most extensive publicly accessible data sets, building and validating the models following the Organization for Economic Cooperation and Development (OECD) QSAR guidelines with compounds excluded from the training data sets. The proposed QSAR model utilizing Morgan circular fingerprints, molecular access system keys (MACCS keys), and modified calculation descriptors with the random forest classifier exhibited high internal accuracy through cross-validation and attained an external correct classification rate between 70 and 77%. A study by Chushak et al. also developed QSAR modeling with "six-pack" end points, similar to STopTox. Their study reported accuracy with a range of 0.72−0.78.[35]

Despite their potential, QSAR models face challenges due to their limited ability to consistently provide reliable assessments, necessitating continuous refinement and validation. While previous studies have developed QSAR models for predicting chemical toxicity, they often faced limitations in specificity, sensitivity, and generalization across diverse chemical structures due to their reliance on limited feature representations. These models frequently failed to capture the full complexity of molecular interactions, resulting in less reliable predictions. Our study addresses these issues by employing a hybrid machine learning approach that integrates various fingerprint and molecular descriptors with physicochemical properties, leveraging the strengths of multiple machine learning models to enhance prediction accuracy and robustness. Additionally, our study expands the scope by classifying a broader range of end points compared to previous models, ensuring that the models are validated and robust for many end points.

We curated data sets from reputable sources, including STopTox, RespiraTox, and DMFGAM,[25,33,34] ensuring high-quality input data. The STopTox database, for instance, is sourced from REACH regulations and managed by the European Chemicals Agency (ECHA) with OECD support. Our methodology leverages cheminformatics tools to convert molecular structures, represented by SMILES codes, into detailed molecular representations. We calculate various molecular descriptors, including Morgan circular fingerprints, MACCS keys, modified calculations, and physicochemical properties using RDKit and the Chemistry Development Kit (CDK).

We employ a hybrid machine learning approach that integrates the best-performing model for each descriptor across three machine learning algorithms: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM). Each model undergoes optimization through grid search with 5-fold cross-validation to identify the optimal hyperparameters, ensuring robust predictive performance. By combining the strengths of these models, we establish a consensus prediction method that selects the most accurate predictions based on descriptor-specific performance. This research provides a comprehensive framework for the in silico prediction of acute systemic and topical toxicity, respiratory irritation, and cardiotoxicity. Leveraging advanced cheminformatics tools and integrating diverse molecular descriptors, this study aims to enhance the accuracy and reliability of chemical hazard assessments, ultimately supporting safer and more efficient regulatory evaluations.

■ RESULTS AND DISCUSSION

This study aimed to evaluate the performance of QSAR classification models for multivariate chemical hazard end points (Table 1) using different feature sets (Table 2): fingerprint and molecular descriptors (Morgan, MACCS keys, and Mordred), physicochemical properties derived from RDKit and CDK tools, and a combination of the best prediction result from each descriptor. The machine learning models applied were Random Forest (RF), XGBoost, and Support Vector Machine (SVM), each trained individually on these descriptor sets. The best-performing model for each set was identified based on various classification metrics, and a final prediction was made using a consensus approach, which involved an equal-weighted average of the prediction scores from each model (Figure 1).

Table 1. Number of Compounds for Each End Point

end points	number of compounds	source
inhalation toxicity	335 nontoxic; 342 toxic	STopTox
cardiac toxicity	4479 hERG blockers; 4596 non hERG blockers	DMFGAM
dermal toxicity	315 nontoxic; 314 toxic	STopTox
oral toxicity	3600 nontoxic; 4596 toxic	STopTox
respiratory irritation	1890 nonirritation; 1777 irritation	RespiraTox
skin irritation	276 nonirritation; 275 irritation	STopTox
eye irritation	1139 nonirritation; 1146 nonirritation	STopTox
skin sensitization	514 non sensitizer; 479 sensitizer	STopTox

In addition, this study examines the comparative effectiveness of individual machine learning models versus a consensus model in predicting toxicity end points. By comparing models such as RF, XGBoost, and SVM to a consensus model that integrates predictions from these individual models, the study aims to identify which approach delivers better predictive accuracy. The analysis primarily uses the AUC scores, F1 score, sensitivity, and specificity as evaluation metrics to assess model performance across different toxicity scenarios.

Tables 3–5 summarize the performance of the QSAR classification models based on fingerprint and molecular descriptors (Table 3), physicochemical properties (Table 4), and a consensus approach combining the best prediction outcomes from each descriptor (Table 5). The performance metrics include accuracy, AUC, sensitivity, specificity, F1 score, positive predictive value (PPV), negative predictive value (NPV), and CCR for different toxicity end points. The performance of QSAR classification modeling for cardiac toxicity shows only slight differences when using various feature sets. Utilizing fingerprint and molecular descriptors alone (Table 3) results in an accuracy of 0.82, an AUC of 0.90, a sensitivity of 0.86, and a specificity of 0.76, demonstrating a robust model with a good balance between the detection of positive and negative cases. The positive predictive value (PPV) is 0.82, the negative predictive value (NPV) is 0.81, the F1 score is 0.84, and the correct classification rate (CCR) is 0.81. However, when using physicochemical properties alone (Table 4), the model shows a slightly lower performance, with an accuracy of 0.75, an AUC of 0.83, a sensitivity of 0.83, and a lower specificity of 0.65. The F1 score, PPV, NPV, and CCR for this model are 0.79, 0.75, 0.74, and 0.74, respectively.

Importantly, when combining the best prediction outcomes from each descriptor with a consensus approach (Table 5), the

Table 2. List of Descriptors Used in This Study

descriptors	description	features characteristics	dimension
Morgan fingerprint	circular fingerprints based on the Morgan algorithm, which encodes molecular structure as a binary vector by considering atom environments within a certain radius	captures information about the molecular structure, including atom connectivity and substructures	1024 bit vector
MACCs keys fingerprint	a predefined set of structural keys that represent the presence or absence of particular substructures within a molecule	a predefined set of structural keys that represent the presence or absence of particular substructures within a molecule	166 bit vector
Modred calculation	descriptors calculated using the Mordred software, which provides a wide range of two-dimensional (2D) and three-dimensional (3D) molecular descriptors derived from chemical structure	descriptors calculated using the Mordred software, which provides a wide range of molecular descriptors derived from chemical structure	vary depending on the number of specific end points data set
physicochemical properties (RDKit and CDK)	calculated properties using RDKit and CDK libraries, covering various physicochemical aspects such as log *P*, hydrogen bond donors/acceptors, and more	provides insight into molecular properties relevant to solubility, permeability, and reactivity	56 features

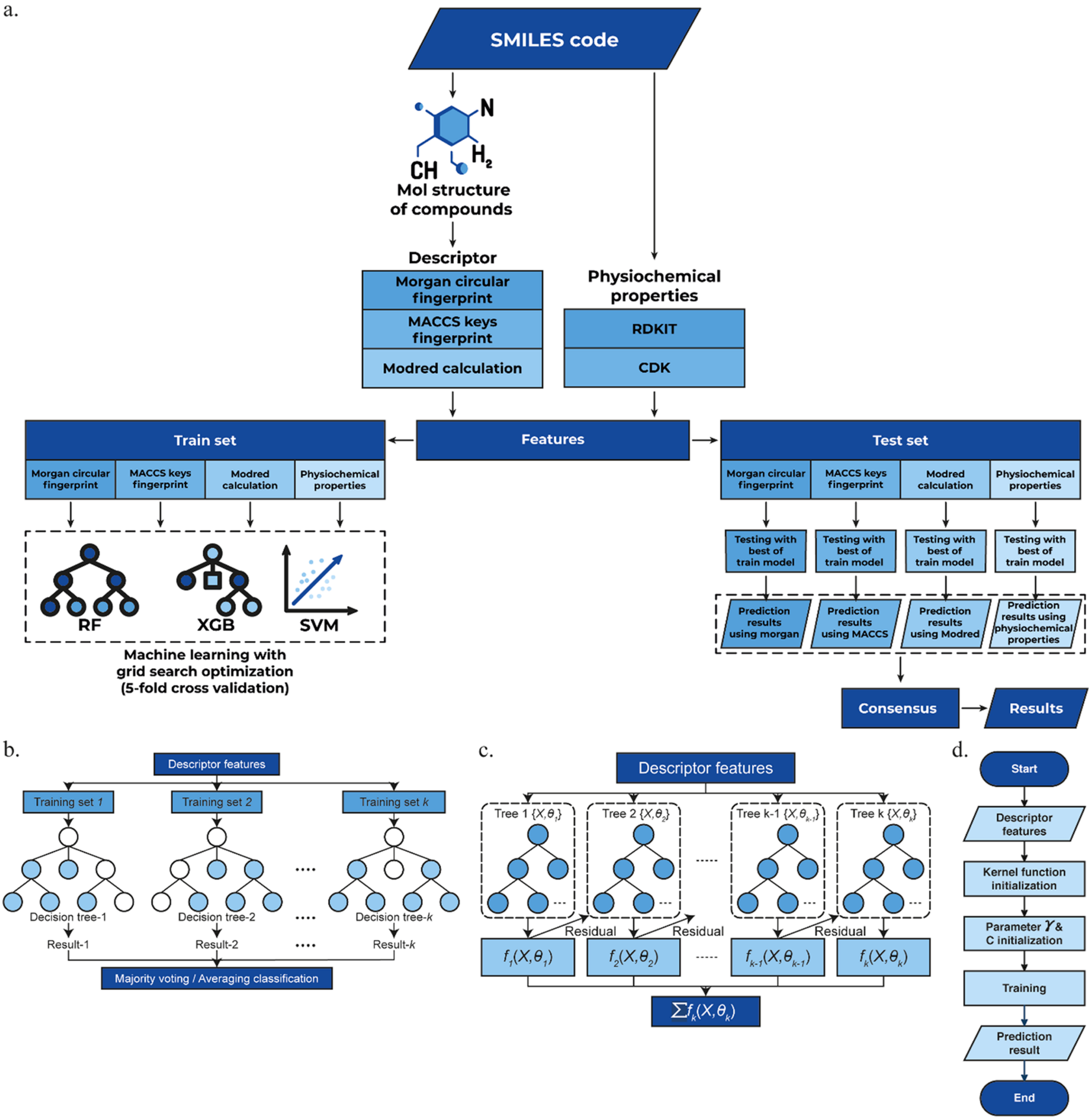

a.
SMILES code
N
H_2
CH
Mol structure of compounds
Descriptor
Morgan circular fingerprint
MACCS keys fingerprint
Modred calculation
Physiochemical properties
RDKIT
CDK
Features
Train set
Morgan circular fingerprint
MACCS keys fingerprint
Modred calculation
Physiochemical properties
RF
XGB
SVM
Machine learning with grid search optimization (5-fold cross validation)
Test set
Morgan circular fingerprint
MACCS keys fingerprint
Modred calculation
Physiochemical properties
Testing with best of train model
Testing with best of train model
Testing with best of train model
Testing with best of train model
Prediction results using morgan
Prediction results using MACCS
Prediction results using Modred
Prediction results using physiochemical properties
Consensus
Results
b.
Descriptor features
Training set 1
Training set 2
Training set k
Decision tree-1
Decision tree-2
Decision tree-k
Result-1
Result-2
Result-k
Majority voting / Averaging classification
c.
Descriptor features
Tree 1 $\{X,\theta_1\}$
Tree 2 $\{X,\theta_2\}$
Tree k-1 $\{X,\theta_{k-1}\}$
Tree k $\{X,\theta_k\}$
Residual
Residual
Residual
$f_1(X,\theta_1)$
$f_2(X,\theta_2)$
$f_{k-1}(X,\theta_{k-1})$
$f_k(X,\theta_k)$
$\sum f_k(X,\theta_k)$
d.
Start
Descriptor features
Kernel function initialization
Parameter γ & C initialization
Training
Prediction result
End

Figure 1. Proposed schematic of QSAR classification modeling based on descriptors with machine learning. (a) Whole schematic of QSAR modeling with consensus approach. (b) Topology of random forest. (c) Topology of XGBoost. (d) Schematic of support vector machine.

Table 3. Performance Result of QSAR Classification Modeling Using Fingerprint and Molecular Descriptor

	classification performance metric							
end points	accuracy	AUC	sensitivity	specificity	F1 score	PPV	NPV	CCR
cardiac toxicity	0.82	0.90	0.86	0.76	0.84	0.82	0.81	0.81
inhalation toxicity	0.77	0.81	0.76	0.78	0.74	0.73	0.81	0.77
dermal toxicity	0.75	0.84	0.71	0.81	0.76	0.82	0.70	0.76
oral toxicity	0.78	0.86	0.84	0.71	0.81	0.78	0.78	0.78
skin irritation	0.86	0.89	0.92	0.81	0.85	0.79	0.93	0.86
skin sensitization	0.70	0.79	0.66	0.75	0.69	0.72	0.69	0.70
eye irritation	0.73	0.80	0.77	0.70	0.75	0.73	0.74	0.73
respiratory irritation	0.72	0.78	0.71	0.73	0.70	0.69	0.74	0.72

Table 4. Performance Result of QSAR Classification Modeling Using Physicochemical Properties

	classification performance metric							
end points	accuracy	AUC	sensitivity	specificity	F1 score	PPV	NPV	CCR
cardiac toxicity	0.75	0.83	0.83	0.65	0.79	0.75	0.74	0.74
inhalation toxicity	0.73	0.82	0.78	0.69	0.71	0.66	0.80	0.73
dermal toxicity	0.69	0.79	0.65	0.74	0.70	0.75	0.64	0.69
oral toxicity	0.72	0.79	0.73	0.69	0.73	0.71	0.72	0.71
skin irritation	0.81	0.88	0.83	0.79	0.79	0.75	0.86	0.81
skin sensitization	0.68	0.74	0.66	0.70	0.67	0.68	0.67	0.68
eye irritation	0.72	0.79	0.73	0.69	0.73	0.71	0.72	0.71
respiratory irritation	0.68	0.75	0.72	0.65	0.68	0.64	0.73	0.69

Table 5. Performance Results of the Proposed QSAR Classification Model, Achieved through a Consensus Approach That Combines the Prediction Outcomes from the Best-Performing Machine Learning Model for Each Individual Descriptor (Fingerprint Descriptors and Physicochemical Properties)

	classification performance metric							
end points	accuracy	AUC	sensitivity	specificity	F1 score	PPV	NPV	CCR
cardiac toxicity	0.82	0.90	0.89	0.75	0.85	0.82	0.83	0.82
inhalation toxicity	0.80	0.83	0.73	0.86	0.76	0.80	0.80	0.79
dermal toxicity	0.78	0.84	0.75	0.81	0.79	0.83	0.73	0.78
oral toxicity	0.79	0.87	0.86	0.70	0.82	0.78	0.80	0.78
skin irritation	0.86	0.90	0.90	0.83	0.84	0.80	0.91	0.86
skin sensitization	0.72	0.78	0.70	0.74	0.71	0.73	0.71	0.72
eye irritation	0.74	0.81	0.77	0.71	0.75	0.73	0.75	0.74
respiratory irritation	0.72	0.78	0.72	0.72	0.71	0.69	0.75	0.72

model's performance achieves an accuracy of 0.82, an AUC of 0.90, and a sensitivity of 0.89, with a balanced specificity of 0.75. The F1 score, PPV, NPV, and CCR for the combined model are 0.85, 0.82, 0.83, and 0.82, respectively. This indicates that the consensus prediction across all descriptors provides better overall predictive performance, particularly by improving sensitivity while maintaining strong accuracy, AUC, and F1 score. These results suggest that integrating multiple descriptor types enhances the model's effectiveness in predicting cardiac toxicity.

For inhalation toxicity, using fingerprint and molecular descriptors (Table 3) achieves an accuracy and Correct Classification Rate (CCR) of 0.77, with an AUC of 0.81. The model demonstrates balanced sensitivity and specificity values of 0.76 and 0.78, respectively, with an F1 score of 0.74. The Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are 0.73 and 0.81, respectively, indicating a relatively good performance. In contrast, using physicochemical properties alone (Table 4) results in slightly lower performance, with an accuracy and CCR of 0.73, an AUC of 0.82, and higher sensitivity at 0.78 but lower specificity at 0.69. The F1 score is 0.71, with PPV and NPV values at 0.66 and 0.80, respectively. When conducting a consensus of prediction result from each descriptor (Table 5), the model improves, achieving an accuracy and CCR of 0.80, an AUC of 0.83, with a notable increase in specificity to 0.86, while maintaining a reasonable sensitivity of 0.73. The F1 score increases to 0.76, with PPV and NPV values at 0.80, demonstrating that a consensus approach offers a more reliable prediction for inhalation toxicity.

Predicting dermal toxicity using fingerprint and molecular descriptors (Table 3) results in an accuracy and Correct Classification Rate (CCR) of 0.75 and an AUC of 0.84, with a sensitivity of 0.71 and a specificity of 0.81, indicating a reasonably balanced model. The F1 score is 0.76, and the positive predictive value (PPV) and negative predictive value (NPV) are 0.82 and 0.70, respectively. In contrast, using physicochemical properties alone (Table 4) yields a lower accuracy and CCR of 0.69 and an AUC of 0.79, with a lower sensitivity of 0.65 but still a decent specificity of 0.74. The F1 score for this model is 0.70, with PPV and NPV values at 0.75 and 0.64, respectively. When conducting a consensus of prediction result from each descriptor (Table 5), the accuracy and CCR are enhanced to 0.78 while maintaining the same AUC of 0.84, with an improved sensitivity of 0.75 and a similar specificity of 0.81. The F1 score increases to 0.79, with PPV and NPV values of 0.83 and 0.73, respectively. This suggests that a consensus approach provides a better predictive performance for dermal toxicity, particularly in terms of sensitivity.

For oral toxicity, using fingerprints and molecular descriptors alone (Table 3) achieves an accuracy and correct classification rate (CCR) of 0.78 and an AUC of 0.86, with a high sensitivity of 0.84 and a lower specificity of 0.71. The F1 score is 0.81, and the positive predictive value (PPV) and negative predictive value (NPV) are both 0.78. In contrast, using physicochemical properties alone (Table 4) results in a lower accuracy and CCR of 0.72 and an AUC of 0.79, with a balanced sensitivity of 0.73 and specificity of 0.69. The F1 score is 0.73, with PPV and NPV values at 0.71 and 0.72, respectively. When combining both prediction results from each descriptor with a consensus approach (Table 5), the model's performance improves, with accuracy and CCR increasing to 0.79 and the AUC to 0.87. The model also shows a high sensitivity of 0.86 and a similar specificity of 0.70. The F1 score increases to 0.82, with PPV and NPV values of 0.78 and 0.80, respectively, suggesting that a consensus model is more effective in predicting oral toxicity.

The prediction performance for skin irritation is notably strong across all models. Using fingerprint and molecular

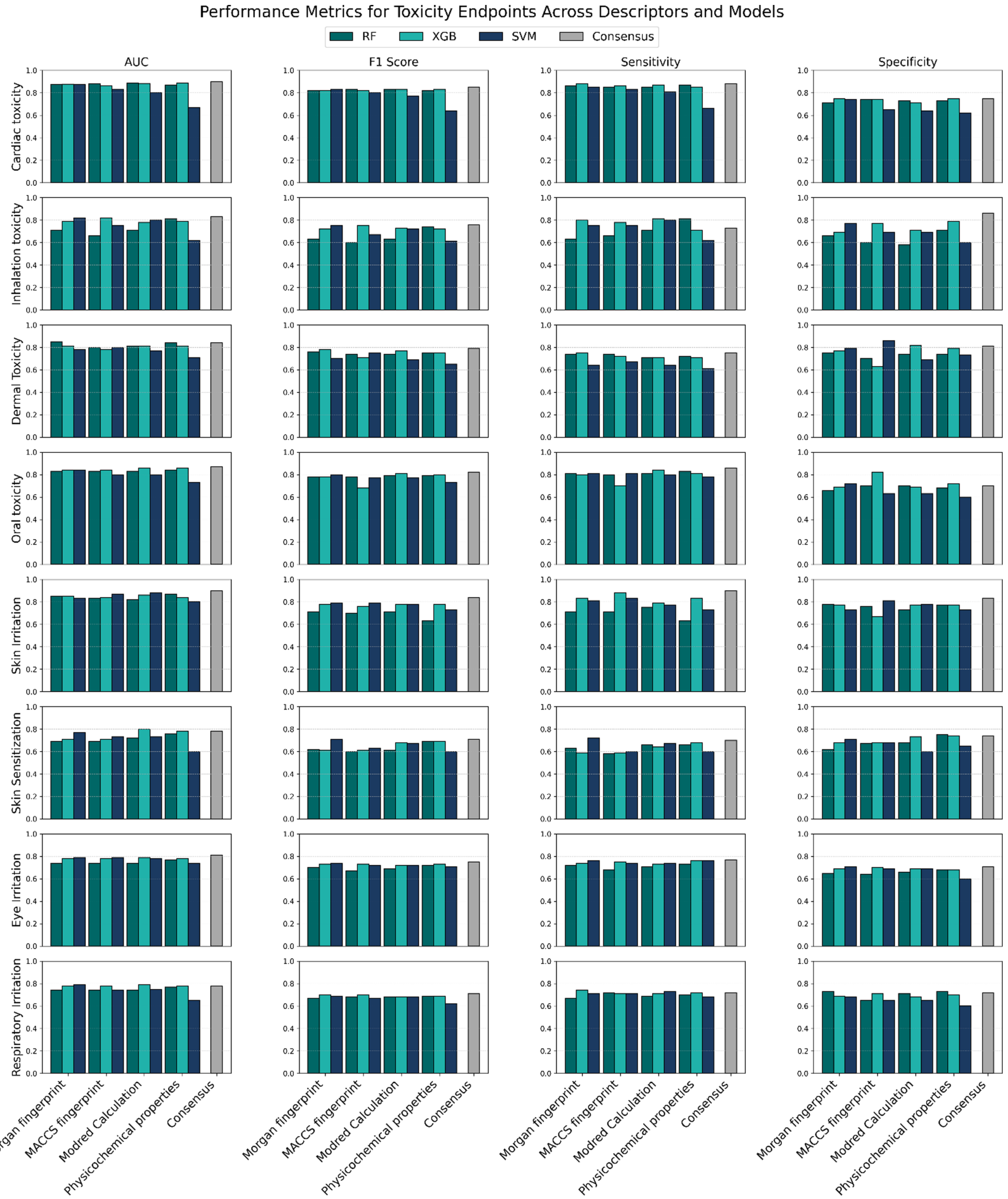


Figure 2. Evaluation metrics, including area under the curve (AUC), F1 score, sensitivity, and specificity, for various toxicity end points using multiple molecular descriptors and machine learning models. Each panel corresponds to a specific toxicity end point, showing the performance of three models—Random Forest (RF), XGBoost (XGB), and Support Vector Machine (SVM). The "Consensus" bar represents an aggregated score, derived by combining the best predictions from individual descriptors across models.

descriptors alone (Table 3) provides an accuracy and correct classification rate (CCR) of 0.86, with an AUC of 0.89, a high sensitivity of 0.92, and a specificity of 0.81. The F1 score is 0.85, with a positive predictive value (PPV) of 0.79 and a negative predictive value (NPV) of 0.93. In contrast, using physicochemical properties alone (Table 4) yields a slightly lower performance, with an accuracy and CCR of 0.81 and an AUC of 0.88, alongside a balanced sensitivity of 0.83 and specificity of 0.79. The F1 score for this model is 0.79, with PPV and NPV at 0.75 and 0.86, respectively. When combining both prediction results from each descriptor with a consensus approach (Table 5), the model maintains an accuracy and CCR of 0.86 but with a slightly higher AUC of 0.90, along with a high sensitivity of 0.90 and an improved specificity of 0.83. The F1 score is 0.84, with PPV and NPV values at 0.80 and 0.91, respectively. This indicates that a consensus approach yields the most balanced and robust model for predicting skin irritation.

For skin sensitization, using fingerprint and molecular descriptors (Table 3) results in an accuracy and correct classification rate (CCR) of 0.70, with an AUC of 0.79, a sensitivity of 0.66, and a specificity of 0.75. The F1 score is 0.69, with a positive predictive value (PPV) and negative predictive value (NPV) of 0.72 and 0.69, respectively. In comparison, using physicochemical properties alone (Table 4) shows a similar accuracy and CCR of 0.68 but a lower AUC of 0.74, with balanced sensitivity and specificity of 0.66 and 0.70, respectively. The F1 score for this model is 0.67, with PPV and NPV values of 0.68 and 0.67. The consensus prediction result from each descriptor (Table 5) slightly improves the accuracy and CCR to 0.72 and the AUC to 0.78, with a sensitivity of 0.70 and a specificity of 0.74. The F1 score increases to 0.71, with PPV and NPV values at 0.73 and 0.71, suggesting that the consensus approach provides a marginally better prediction for skin sensitization.

The prediction of eye irritation using fingerprint and molecular descriptors (Table 3) results in an accuracy and correct classification rate (CCR) of 0.73 and an AUC of 0.80, with a balanced sensitivity of 0.77 and a specificity of 0.70. The F1 score is 0.75, and the positive predictive value (PPV) and negative predictive value (NPV) are 0.73 and 0.74, respectively. Using physicochemical properties alone (Table 4) yields a similar performance, with an accuracy and CCR of 0.72 and an AUC of 0.79, along with a balanced sensitivity of 0.73 and specificity of 0.69. The F1 score is 0.73, with PPV and NPV at 0.71 and 0.72, respectively. The consensus prediction result from each descriptor (Table 5) slightly improves accuracy and CCR to 0.74 and AUC to 0.81, while maintaining a similar sensitivity of 0.77 and specificity of 0.71. The F1 score remains 0.75, with PPV and NPV at 0.73 and 0.75, respectively, indicating a marginal benefit of the consensus approach of all for predicting eye irritation.

For respiratory irritation, using fingerprint and molecular descriptors (Table 3) gives an accuracy and correct classification rate (CCR) of 0.72 and an AUC of 0.78, with a balanced sensitivity of 0.71 and specificity of 0.73. The F1 score is 0.70, with a positive predictive value (PPV) of 0.69 and a negative predictive value (NPV) of 0.74. In contrast, using physicochemical properties alone (Table 4) results in a lower accuracy and CCR of 0.68 and an AUC of 0.75, with a balanced sensitivity of 0.72 but a lower specificity of 0.65. The F1 score for this model is 0.68, with PPV and NPV values of 0.64 and 0.73, respectively. The consensus of all descriptors with the best machine learning model (Table 5) maintains the same accuracy and CCR of 0.72 and an AUC of 0.78, with a balanced sensitivity of 0.72 and specificity of 0.72. The F1 score remains 0.71, with PPV and NPV both at 0.69 and 0.75, respectively. This suggests that while the consensus model does not significantly improve performance, it maintains a balanced and consistent prediction for respiratory irritation.

Figure 2 highlights the effectiveness of the consensus approach in toxicity prediction, which enhances performance compared to previous studies by combining the best predictions from various descriptor-model pairs. This method improves AUC, F1 score, sensitivity, and specificity, creating a balanced, robust outcome that surpasses individual descriptors or models alone. The consensus approach shows greater stability and accuracy, particularly for end points where single descriptors or models may yield inconsistent results, making it a valuable method for reliable and generalizable toxicology predictions.

The performance of evaluation metrics shows that Random Forest (RF) and Extreme Gradient Boosting (XGB) consistently demonstrate superior performance in specific toxicity end points. These models excel due to their robust algorithms that effectively manage complex, nonlinear relationships within the data.[36] This capability allows them to capture detailed structural information, leading to more accurate predictions in those end points. While competitive, support vector machine (SVM) models generally show slightly lower AUC values than RF and XGB. This may be attributed to SVM's reliance on kernel functions, which might not capture all of the nuances of the data as effectively as the ensemble methods.[37] In contrast, the consensus model integrates predictions from multiple models and performs more consistently across a wide range of toxicity end points. This indicates that the consensus approach effectively leverages the strengths of individual models, resulting in more reliable and robust predictions overall.

Additionally, we evaluate the performance of models using individual descriptors and consensus predictions combining all descriptors' performance. We evaluate the model performance using individual descriptors, such as fingerprints and molecular descriptors, along with physicochemical properties and compare these results with consensus predictions that integrate both types of descriptors. The superior performance of fingerprint descriptors over physicochemical properties can be attributed to the detailed structural information that they capture. Fingerprint descriptors are designed to represent specific substructures and patterns within molecules, which are often critical for biological activity and toxicity. This detailed representation allows machine learning models to make more accurate predictions based on the presence or absence of specific chemical features. In contrast, while valuable, physicochemical properties provide a more general overview of molecular behavior. They include molecular weight, log P, and hydrogen bond donors/acceptors, among other properties which are essential but might not capture the nuanced structural features that directly influence toxicity.[38] Combining the prediction score with consensus approach when utilizing fingerprint descriptors (Morgan, MACCS keys, and Mordred) with physicochemical properties derived from RDKit and CDK tools provides a holistic view of the molecules, significantly enhancing the predictive power of QSAR models for various toxicity end points (as shown in Tables 3−5). This integrated approach allows the model to utilize both detailed structural information and general physicochemical behavior, leading to more accurate and reliable predictions. For several end points, such as inhalation toxicity, dermal toxicity, oral toxicity, and skin irritation, combining the predictions from each descriptor with the best machine learning model notably improves accuracy, sensitivity, and specificity compared to using fingerprint and molecular descriptors alone. This improvement

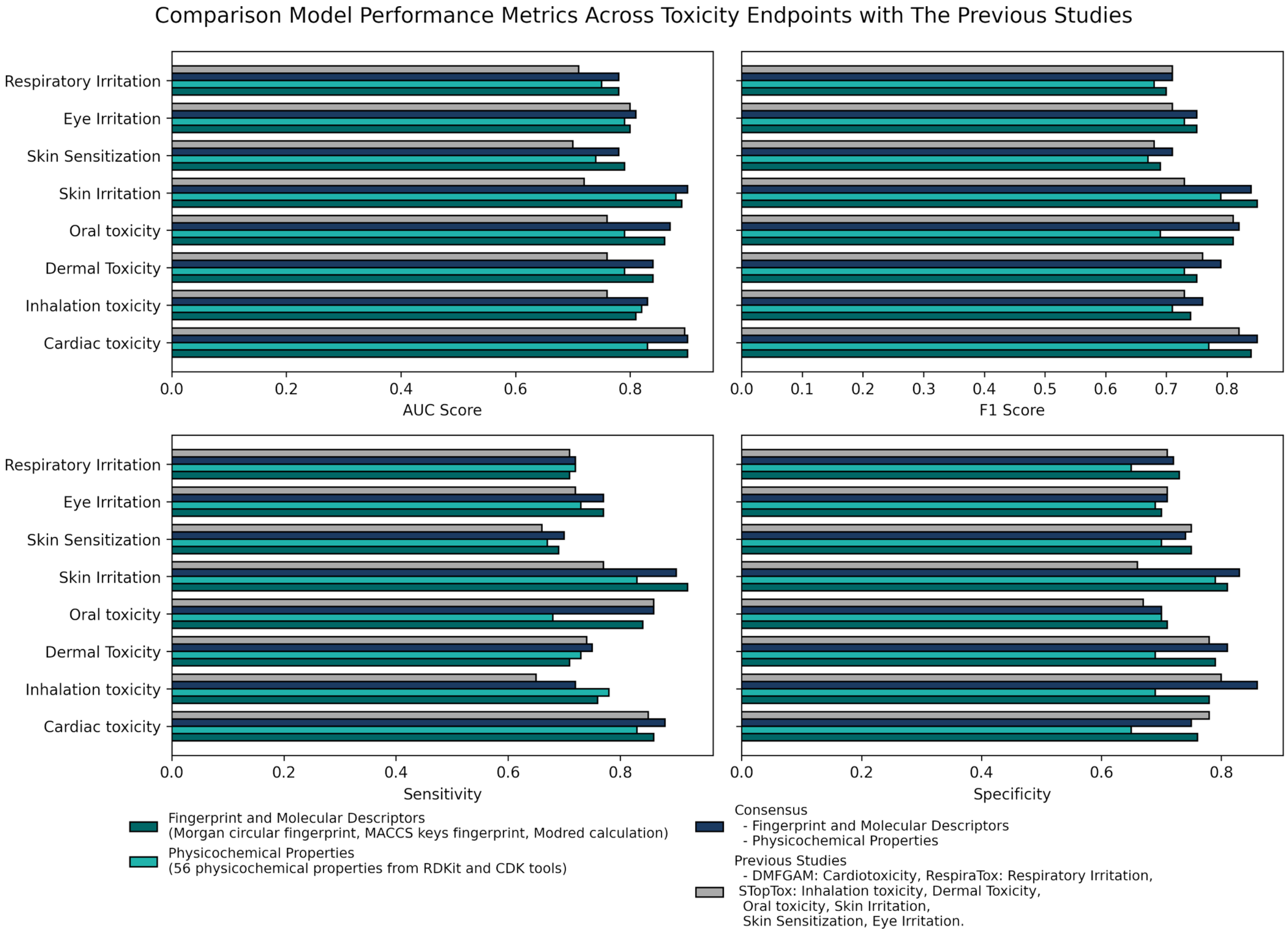


Figure 3. Comparison of evaluation metrics—AUC, F1 score, sensitivity, and specificity—for various toxicity end points, showcasing the performance of the proposed consensus model using multiple machine learning algorithms. The feature sets include fingerprint and molecular descriptors, physicochemical properties, and a consensus of all descriptors. Results are also compared against previous studies (STopTox, RespiraTox, and DMFGAM) to provide a comprehensive view of the predictive capabilities and the effectiveness of the consensus of all descriptors across different toxicity end points.

suggests that integrating diverse data types captures a more comprehensive representation of the chemical and biological properties relevant to these toxicities, resulting in more robust predictions.

However, for other end points, such as cardiac toxicity, skin sensitization, eye irritation, and respiratory irritation, the predictive performance is not significantly different between using fingerprint and molecular descriptors alone and consensus all descriptors. In these cases, the fingerprint and molecular descriptors are already highly effective, and adding physicochemical properties provides only marginal enhancements. This indicates that while the combined approach generally improves predictive accuracy, the extent of improvement can vary depending on the specific toxicity end point being modeled.

The consensus approach, which integrates the prediction scores from multiple models using different descriptors, demonstrated superior performance compared to individual descriptors and machine learning models alone. By employing an equal weight average for the prediction scores, the consensus method effectively balanced the contributions of different models and descriptor sets, mitigating the weaknesses of any single model or descriptor set. This approach leveraged the strengths of each model and descriptor set, resulting in more robust and accurate predictions. Figure 2 shows that the consensus models consistently outperformed the best individual models across various toxicity end points. This robustness is crucial, indicating that the consensus approach can handle variability and inconsistencies better than individual models.

Figure 3 presents the AUC, F1 score, sensitivity, and specificity for various toxicity end points using a consensus model that combines predictions from different machine learning algorithms across various feature sets, including fingerprint and molecular descriptors, physicochemical properties, and a combination of all descriptors. The consensus model's performance is compared against previous studies such as STopTox, RespiraTox, and DMFGAM. The results show that leveraging multiple feature sets through a consensus approach consistently enhances predictive accuracy and robustness, outperforming or comparable to previous studies. This underscores the value of incorporating diverse types of features in toxicity assessments rather than relying on a single type of descriptor.

For instance, in the case of cardiac toxicity, our consensus model obtained an area under the curve (AUC) of 0.90. In terms of additional metrics, our consensus model demonstrates a sensitivity of 0.88, a specificity of 0.75, and an accuracy of 0.822. This shows that our model is highly sensitive, effectively identifying true positives in cases of toxicity. While specificity is slightly lower at 0.75 compared to DMFGAM model's 0.78, our model maintains competitive specificity, balancing the reduction in false positives with high sensitivity. The accuracy of our model at 0.822 also slightly surpasses that of DMFGAM model (0.817), emphasizing its reliable overall performance.

Although this is only a slight improvement over the previous study's DMFGAM model, which had an AUC of 0.895, it is important to note that these models are based on fundamentally different methodologies. The DMFGAM model employs a sophisticated deep learning approach that integrates multiple molecular fingerprints, and graph features to predict hERG channel blockers, which are indicators of cardiac toxicity, with high accuracy.[33] Each molecule is represented as a graph, and these graphs are analyzed using the SMILES Graph Attention Network (SGAT), a complex deep learning architecture that utilizes a multihead attention mechanism designed to identify complex relationships within the molecular structure. The combined features are fed into a fully connected neural network, allowing for highly nuanced predictions.

In contrast, our study employs a consensus model that integrates several machine learning algorithms, including Random Forest (RF), eXtreme Gradient Boosting (XGB), and Support Vector Machines (SVM), using a set of molecular descriptors. While these machine learning methods are generally considered less complex than deep learning models such as DMFGAM, our consensus approach leverages the strengths of each algorithm. Combining the predictions from multiple models, we achieve a more robust and reliable outcome, as evidenced by the higher AUC for cardiac toxicity. The advantage of our approach lies in its simplicity, interpretability, and ability to provide stable predictions without requiring extensive computational resources typically associated with deep learning models.

Furthermore, our model consistently outperforms previous studies across a range of toxicity end points by achieving a well-balanced performance in sensitivity, specificity, AUC, and F1 score. For end points such as inhalation toxicity, dermal toxicity, and skin sensitization, the model's higher specificity compared to sensitivity makes it particularly effective at correctly identifying nontoxic or nonsensitizing cases, reducing the occurrence of false positives. This balance is crucial in applications where accurately identifying nonhazardous substances is essential. The model's strong AUC and F1 scores across these end points highlight its reliability in achieving both accuracy and stability, capturing toxicity complexities more effectively than StopTox.

Conversely, for skin irritation, eye irritation, and oral toxicity, the model demonstrates sensitivity higher than specificity, excelling at detecting cases that could pose risks. This high sensitivity ensures that true toxic or irritant cases are not missed, which is a priority for safety evaluations. At the same time, balanced specificity reduces the likelihood of misclassifying safe substances as toxic, supported by strong F1 scores that indicate consistent and accurate detection of true cases. The model's elevated AUC scores across these end points underscore its strong discriminative power, enhancing prediction accuracy compared to StopTox's approach.

For respiratory irritation, where balanced sensitivity and specificity are crucial, our model performs consistently, reflecting an ability to reliably distinguish between irritant and nonirritant cases. This balance, along with robust AUC and F1 scores, reinforces the model's versatility and adaptability, establishing it as a reliable tool in toxicological assessments across diverse end points and a significant improvement over RespiraTox.

In conclusion, the integration of fingerprint descriptors and physicochemical properties, coupled with a consensus approach, significantly enhances the performance of QSAR classification models. This study provides a robust framework for improving toxicity predictions, demonstrating that the consensus method is superior to that of individual descriptors and machine learning models alone. The higher overall enhanced performance metrics validate the effectiveness of our approach, offering valuable insights for future QSAR modeling efforts.

While the results are promising, the study has several certain constraints. First, the models were evaluated using a specific set of descriptors and toxicity end points; expanding these to include additional descriptors and a wider range of end points could yield a more comprehensive understanding of model performance. Second, the consensus approach used an equal weighting of model predictions, which may not be optimal in all contexts. Future research could explore dynamic weighting strategies to further enhance the prediction accuracy.

In addition, incorporating explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP), would allow for the identification of significant molecular features, Improving the clarity and understanding of the model's predictions. By pinpointing which descriptors contribute most to each end point, XAI methods could help clarify the relationships between molecular structure and toxicity, offering a more transparent model framework. Finally, applying these models to larger, more diverse data sets could help validate their utility in real-world scenarios, providing a stronger foundation for their adoption in QSAR modeling.

■ MATERIALS AND METHODS

Data Set Curation: Collection and Preparation. In this study, we curated data sets from previous studies, including STopTox, RespiraTox, and DMFGAM.[34] The STopTox database was mainly gathered from Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH).[34] REACH is a regulation of the European Union established to enhance the protection of human health and the environment from the hazards associated with chemicals.[39] Furthermore, there is European Chemicals Agency (ECHA) as the regulatory agency responsible for implementing REACH. It manages the regulation's technical, scientific, and administrative aspects, including the collection, evaluation, dissemination, and restriction of chemical data.[40] Moreover, the Organization for Economic Cooperation and Development (OECD) provides guidelines and frameworks to support the implementation of REACH.[41] These include internationally harmonized test methods and assessment strategies to ensure consistent and reliable data on the properties of the chemicals.

For collecting six-pack toxicity end points, including inhalation toxicity, dermal toxicity, oral toxicity, skin irritation, skin sensitization, eye irritation, the STopTox procedure consisted of entailing meticulous data cleaning, standardization, and unit harmonization across various data sets.[25,33,34] To transform the data into binary toxicity categories, we applied the classification criteria outlined by the Globally Harmonized

System of Classification and Labeling of Chemicals (GHS) standards. Acute systemic end points falling under GHS classes 1−4 were labeled as "toxic", whereas class 5 was labeled "not classified". Each end point followed specific GHS criteria: for example, in skin irritation, classes 1−3 were identified as skin irritants; in eye irritation, classes 1−2B as causing eye irritation, and class 1 as a sensitizer for skin sensitization. This extensive data curation process aligned both chemical and biological data with standardized protocols, eliminating inconsistencies to enhance data quality.[25,33,34]

The STopTox approach excluded the data that did not adhere to Organization for Economic Cooperation and Development (OECD) guidelines, lacking multiconcentration testing, and could not be assigned to GHS categories. Measurements that differed from standard protocols for the six-pack end points were treated specifically: standardized toxicity measurements such as LD50 (lethal dose, 50%) for systemic end points, the effective concentration at the third percentile (EC3) was applied for measuring skin sensitization, considered mean score of erythema and edema along with reversibility data for skin irritation and considered corneal and iritis gradings with reversibility details for eye irritation, all in alignment with the GHS procedure.[34]

After biological data were curated, chemical structures were refined by removing mixtures, inorganic compounds, and organometallic substances. Salts were neutralized, specific chemotypes standardized, and duplicate records addressed as follows: (a) for duplicates with the same binary outcome, only one record was retained; (b) for duplicates with a majority binary outcome and one exception, the majority outcome record was kept; and (c) for duplicates with conflicting binary outcomes, all records were excluded.[25,33,34]

Furthermore, respiratory irritation assessment used RespiraTox database which was collected from Fraunhofer RepDose, ChemIDplus (https://chem.nlm.nih.gov/chemidplus/), Chemicalbook (https://www.chemicalbook.com), European Chemicals Agency Chemical Database (ECHA CHEM), and Hazardous Substances Data Bank (HSDB) databases.[25] Only high-quality acute toxicological studies from ECHA CHEM, specifically those conducted on rodents or dogs, were selected. Lower-quality studies and those based on the weight of evidence rather than direct experimental data were excluded.

Meanwhile, the DMFGAM for cardiotoxicity database was collected from the CHEMBL v29 database (https://www.ebi.ac.uk/chembl)[33] and various literature-derived data.[42−45] Each compound's chemical structure was standardized using Python packages RDKit and Molecular Validation and Standardization (MolVS), which consisted of extracting the largest fragment, removing explicit hydrogens, adjusting ionization states, and determining stereochemistry.[33] The half-maximal inhibitory concentration (IC50) was used as a measure of activity, with compounds classified as hERG blockers if their IC50 was below 10 μM and nonblockers if IC50 was 10 μM or higher.[33] To enhance classification, IC50 values were converted to pIC50, a negative logarithmic scale, as it provides a clearer representation of the inhibitory activity. Compounds with pIC50 values below 5 were classified as hERG blockers, while those with pIC50 values of 5 or above were labeled as nonblockers. Following filtering, the final number of compounds retained for each end point with the source of the data set is presented in Table 1.

Descriptor Calculation. Each compound was represented by a SMILES code, which was converted into a molecular structure for further processing with cheminformatics tools. From these structures, a diverse set of molecular descriptors and physicochemical properties were computed, enhancing the model's capability to generalize across multiple toxicity end points by capturing a broad spectrum of molecular features. Instead of tailoring specific descriptors to individual end point, this approach leverages a broad spectrum of molecular features, improving flexibility and predictive accuracy across diverse toxicological profiles. The fingerprint and molecular descriptors with physicochemical properties, as detailed in Table 2 were independently evaluated, utilizing the highest-performing model for each descriptor contributing to the consensus prediction.

The Morgan circular fingerprint is a form of molecular fingerprint that was computed to capture the circular neighborhoods of atoms within each molecule. This method generates a binary- or integer-based fingerprint by analyzing substructures centered around each atom within a defined radius. By encoding the presence or absence of specific atom-centered substructures, Morgan fingerprints are highly effective in identifying essential molecular patterns relevant to chemical behavior. This makes them particularly useful for tasks such as similarity searching, clustering, and structure−activity relationship (SAR) studies, where understanding substructural motifs is crucial for predicting toxicity.

We also calculated the Molecular ACCess System (MACCS) keys, which provide a binary representation of the presence or absence of 166 predefined structural features commonly found in chemical compounds. Each key corresponds to a specific molecular substructure with its presence or absence encoded as a binary value (1 or 0). MACCS keys simplify the identification and comparison of structural motifs across molecules, offering a streamlined method for compound screening, classification, and database searching. By providing structural alerts linked to specific toxicity risks, MACCS keys allow the model to detect potentially hazardous compounds based on established structural patterns.

In addition, Mordred descriptors were calculated to generate an extensive set of chemical descriptors that encompass a wide range of molecular properties, including topological, geometric, electronic, and hybrid characteristics. These descriptors offer a detailed and varied representation of the molecular structure, capturing intricate details essential for in-depth molecular analysis, QSAR modeling, and various machine learning applications in cheminformatics. The inclusion of Mordred descriptors provides the model with a nuanced view of the molecule's structural and electronic features, enhancing a more thorough analysis of molecular interactions that may influence toxicity.

Furthermore, physicochemical properties were computed using RDKit and Chemistry Development Kit (CDK), both open-source cheminformatics tools. RDKit provided a range of properties, such as the count of aromatic atoms, log P (partition coefficient), and the number of hydrogen bond donors and acceptors, which are critical for predicting a compound's behavior in biological systems. CDK complemented RDKit by calculating additional properties, including atomic partial charges, labute accessible surface area, and the topological polar surface area. These physicochemical properties contribute insights into the compound's chemical behavior, solubility, and bioavailability, which are essential for assessing interactions within biological environments.

By combining all of these descriptor types across multiple end points, the model benefits from a comprehensive molecular representation. This integrative approach allows the model to capture both detailed structural attributes and general chemical behaviors, improving flexibility and predictive accuracy across various toxicity end points. This broad-spectrum use of descriptors ultimately supports a more robust prediction of toxicological profiles, making the model well suited to evaluate complex chemical and biological relationships across multiple end points.

Data Set Splitting. The data set was subsequently divided into 80% of the data set as a training set and 20% as a test set. The training set was employed to build and optimize our machine learning models, and the test set was provided to evaluate the performance of these models. This splitting ensures that the models are assessed on data that they have not encountered during training, providing a more accurate measure of their predictive capabilities.

Machine Learning Model. This study employs descriptors as input to the hybrid machine learning models to predict the conditions of chemical compounds, utilizing three distinct machine learning models: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM) are presented in Figure 1a. Each model undergoes optimization through grid search using 5-fold cross-validation approach to determine the best hyperparameters, ensuring that the models are fine-tuned for optimal performance and mitigating the risk of overfitting by accounting for variability within the training data. Subsequently, the models are fitted on the training data set, which includes specific descriptors and physicochemical properties critical to the prediction task.

Random Forest (RF) as shown in Figure 1b is a learning technique that builds several decision trees throughout the training process and merges their results to obtain a more accurate and stable prediction.[46,47] Key hyperparameters for RF include the number of trees in the forest (n_estimators) set to 300 as the best number of trees in the forest. Furthermore, the optimal number of features is using all of the features as input to the RF. Optimizing these hyperparameters ensures that the RF model can capture the complexity of the data.

eXtreme Gradient Boosting (XGBoost) as illustrated in Figure 1c is a powerful gradient boosting technique that enhances model accuracy by sequentially building models to correct the errors of previous models.[47,48] Important hyperparameters for XGBoost include (the learning rate = 0.1), which influences the contribution of each tree to the model; the maximum tree depth (set to 5), which defines the model's complexity; and the number of estimators (n_estimators = 200), specifying the boosting rounds to enhance prediction accuracy and generalization.

Support Vector Machine (SVM) as presented in Figure 1d is a robust classification method that identifies the optimal hyperplane that most effectively separates the data into classes.[49,50] The primary hyperparameters for SVM include the penalty parameter (C = 10), which balances the trade-off between minimizing training error and reducing testing error; the kernel type (e.g., linear, polynomial, radial basis function), which specifies the transformation applied to project the data into a higher-dimensional space; radial basis function kernel was selected to captured nonlinear patterns in the data, and the kernel coefficient (γ) was set to "scale" which determines the influence of a single training example. Optimizing these hyperparameters ensures that the SVM model can effectively capture the underlying patterns in the data.

Once the best hyperparameters are selected, each model is fitted using the entire training data, ensuring comprehensive learning from the data's relevant features. This training process incorporates descriptors, including Morgan circular fingerprints, MACCS keys fingerprints, modified calculations, and physicochemical properties derived from RDKit and CDK. The trained models, tailored to these specific descriptors, are then evaluated on a separate test data set to provide an unbiased performance estimate. This evaluation step is crucial as it assesses the models' ability to generalize to new, unseen data, thereby validating their predictive accuracy and robustness.

Consensus and Model Evaluation. Incorporating a consensus approach, where prediction scores from each descriptor are combined using a weighted average of the best machine learning models, enhances the system's predictive performance and robustness.[51] By ensuring that each model's contribution is unbiased, this method leverages the strengths of all models and descriptors, leading to a more accurate and reliable final prediction than relying on any single model.

Mathematically, if P_{Morgan}, P_{MACCSKey}, P_{Modred}, and $P_{\text{physicochemical properties}}$ represent the predictions from the best models for Morgan circular fingerprints, MACCS keys fingerprints, Modred calculation, and physicochemical properties, respectively, the consensus prediction $P_{\text{consensus}}$ is calculated using eq 1.

$$P_{\text{consensus}} = \frac{1}{4}(P_{\text{Morgan}} + P_{\text{MACCSKey}} + P_{\text{Modred}} + P_{\text{physicochemical properties}}) \tag{1}$$

Evaluating the consensus model involves several key performance metrics to ensure a comprehensive assessment of its predictive capability, including accuracy, sensitivity, specificity, PPV, NPV, and CCR, which are calculated using eq 2—eq 7 and AUC score.[34] In AUC, the curve represents the Receiver Operating Characteristic (ROC) curve, which displays the true positive rate (sensitivity) versus the false positive rate (1 − specificity) across different threshold levels.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{positive predictive value (PPV)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{negative predictive value (NPV)} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{6}$$

$$\text{correct classification rate (CCR)} = \frac{\text{sensitivity} + \text{specificity}}{2} \tag{7}$$

Software and Package Information. The QSAR modeling framework was developed using Python, with a range of libraries and tools to ensure a comprehensive descriptor calculation and machine learning implementation.

Chemical descriptors were generated using RDKit (version 2023.9.5) and CDK (version 0.1.0 released on Jan 16th 2024), providing a robust set of molecular features for model input. The machine learning models were implemented using TensorFlow (version 2.7.0), enabling the efficient training and tuning of complex algorithms. In addition, essential Python libraries such as scikit-learn (version 1.3.0) for model evaluation, NumPy (version 2.7.0) for numerical operations, and Pandas (version 2.2.1) for data handling were used throughout the modeling process. These specific package versions ensure reproducibility and allow others to replicate the computational environment used in this study.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c09356.

> SMILES codes for eight end points; training and testing data sets; code for feature generation (Morgan circular fingerprints, MACCS keys fingerprint, Mordred calculations, physicochemical properties); machine learning code for model training (random forest, XGBoost, SVM); and code for testing and consensus performance evaluation (PDF)

AUTHOR INFORMATION

Corresponding Author

Ki Moo Lim − *Computational Medicine Lab, Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea; Computational Medicine Lab, Department of Medical IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea; Meta Heart Co., Ltd., Gumi 39253, Republic of Korea;* orcid.org/0000-0001-6729-8129; Email: kmlim@kumoh.ac.kr

Authors

Yunendah Nur Fuadah − *Computational Medicine Lab, Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea; School of Electrical Engineering, Telkom University, Bandung 40257, Indonesia*

Muhammad Adnan Pramudito − *Computational Medicine Lab, Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea*

Lulu Firdaus − *Computational Medicine Lab, Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea*

Frederique J. Vanheusden − *Department of Engineering, School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, U.K.;* orcid.org/0000-0003-2369-6189

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c09356

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Food and Drug Safety (22213MFDS3922 and 24212MFDS266), the National Research Foundation of Korea (NRF) under the Basic Science Research Program (2022R1A2C2006326), the Ministry of Science and ICT (MSIT), Korea, under the Grand Information Technology Research center support program (IITP-2022-2020-0-01612) supervised by the Institute for Information & communications Technology Planning & Evaluation (IITP), and Directorate of Research and Community Service Telkom University.

REFERENCES

(1) Hamm, J.; Sullivan, K.; Clippinger, A. J.; et al. Alternative approaches for identifying acute systemic toxicity: Moving from research to regulatory testing. *Toxicol. In Vitro* **2017**, *41*, 245−259.

(2) Erhirhie, E. O.; Ihekwereme, C. P.; Ilodigwe, E. E. Advances in acute toxicity testing: Strengths, weaknesses and regulatory acceptance. *Interdiscip. Toxicol.* **2018**, *11*, 5−12.

(3) Wang, Y.; Ning, Z. H.; Tai, H. W.; et al. Relationship between lethal toxicity in oral administration and injection to mice: Effect of exposure routes. *Regul. Toxicol. Pharmacol.* **2015**, *71*, 205−212.

(4) Lane, T. R.; Harris, J.; Urbina, F.; Ekins, S. Comparing LD50/LC50Machine Learning Models for Multiple Species. *ACS Chem. Health Saf.* **2023**, *30*, 83−97.

(5) Gallegos Saliner, A.; Worth, A. P. Testing Strategies for the Prediction of Skin and Eye Irritation and Corrosion for Regulatory Purposes. http://www.jrc.cec.eu.int.

(6) Lawrence, T.; National Research Council, *Animals as Sentinels of Environmental Health Hazards*; National Academies Press: Washington, DC1991. https://doi.org/10.17226/1351

(7) Seok, J.; Warren, H. S.; Cuenca, A. G.; et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 3507−3512.

(8) Bracken, M. B. Why animal studies are often poor predictors of human reactions to exposure. *J. R. Soc. Med.* **2009**, *102*, 120−122.

(9) U.S. EPA. Directive to Prioritize Efforts to Reduce Animal Testing, 2019.

(10) U.S. EPA. Cost Estimates of Studies Required for Pesticide Registration, 2019.

(11) Hubrecht, R. C.; Carter, E. The 3Rs and humane experimental technique: Implementing change. *Animals* **2019**, *9*, No. 754.

(12) Flecknell, P. Replacement, Reduction and Refinement, 2002.

(13) A Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States, 2018. https://ntp.niehs.nih.gov/pubhealth/evalatm/natl-strategy/index.html.

(14) Myatt, G. J.; Ahlberg, E.; Akahori, Y.; et al. In silico toxicology protocols. *Regul. Toxicol. Pharmacol.* **2018**, *96*, 1−17.

(15) Yang, H.; Sun, L.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front. Chem.* **2018**, *6*, No. 30.

(16) Mao, J.et al. Comprehensive Strategies of Machine-learning-based Quantitative Structure-Activity Relationship Models.

(17) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.

(18) Lapenna, S.; Fuart-Gatnik, M.; Worth, A. *Review of QSAR Models and Software Tools for Predicting Acute and Chronic Systemic Toxicity*; Publications Office, 2010.

(19) Raies, A. B.; Bajic, V. B. In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2016**, *6*, 147−172.

(20) Alves, V. M.; Capuzzi, S. J.; Braga, R. C.; et al. A Perspective and a New Integrated Computational Strategy for Skin Sensitization Assessment. *ACS Sustainable Chem. Eng.* **2018**, *6*, 2845−2859.

(21) Alves, V. M.; Capuzzi, S. J.; Muratov, E. N.; et al. QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. *Green Chem.* **2016**, *18*, 6501−6515.

(22) Borba, J. V. B.; Braga, R. C.; Alves, V. M.; et al. Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem. Res. Toxicol.* **2021**, *34*, 258−267.

(23) Kang, Y.; Kim, M. G.; Lim, K. M. Machine-learning based prediction models for assessing skin irritation and corrosion potential of liquid chemicals using physicochemical properties by XGBoost. *Toxicol. Res.* **2023**, *39*, 295−305.

(24) Lou, S.; Yu, Z.; Huang, Z.; et al. In Silico Prediction of Chemical Acute Dermal Toxicity Using Explainable Machine Learning Methods. *Chem. Res. Toxicol.* **2024**, *37*, 513−524.

(25) Wehr, M. M.; Sarang, S. S.; Rooseboom, M.; et al. RespiraTox – Development of a QSAR model to predict human respiratory irritants. *Regul. Toxicol. Pharmacol.* **2022**, *128*, No. 105089.

(26) Schieferdecker, S.; Rottach, F.; Vock, E. In Silico Prediction of Oral Acute Rodent Toxicity Using Consensus Machine Learning. *J. Chem. Inf. Model.* **2024**, *64*, 3114−3122.

(27) Kim, H.; Park, M.; Lee, I.; Nam, H. BayeshERG: a robust, reliable and interpretable deep learning model for predicting hERG channel blockers. *Briefings Bioinf.* **2022**, *23*, No. bbac211.

(28) Ryu, J. Y.; Lee, M. Y.; Lee, J. H.; Lee, B. H.; Oh, K. S. DeepHIT: A deep learning framework for prediction of hERG-induced cardiotoxicity. *Bioinformatics* **2020**, *36*, 3049−3055.

(29) Cai, C.; Guo, P.; Zhou, Y.; et al. Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *J. Chem. Inf. Model.* **2019**, *59*, 1073−1084.

(30) Ye, L.; Ngan, D. K.; Xu, T.; et al. Prediction of drug-induced liver injury and cardiotoxicity using chemical structure and in vitro assay data. *Toxicol. Appl. Pharmacol.* **2022**, *454*, No. 116250.

(31) Chen, Y.; Yu, X.; Li, W.; Tang, Y.; Liu, G. In silico prediction of hERG blockers using machine learning and deep learning approaches. *J. Appl. Toxicol.* **2023**, *43*, 1462−1475.

(32) Lee, H. M.; Yu, M. S.; Kazmi, S. R.; et al. Computational determination of hERG-related cardiotoxicity of drug candidates. *BMC Bioinf.* **2019**, *20*, No. 250.

(33) Wang, T.; Sun, J.; Zhao, Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism. *Comput. Biol. Med.* **2023**, *153*, No. 106464.

(34) Borba, J. V. B.; Alves, V. M.; Braga, R. C.; et al. STopTox: An in Silico Alternative to Animal Testing for Acute Systemic and Topical Toxicity. *Environ. Health Perspect.* **2022**, *130* (2), No. 27012.

(35) Chushak, Y.; Gearhart, J. M.; Clewell, R. A. Structural alerts and Machine learning modeling of "Six-pack" toxicity as alternative to animal testing. *Comput. Toxicol.* **2023**, *27*, No. 100280.

(36) Boldini, D.; Grisoni, F.; Kuhn, D.; Friedrich, L.; Sieber, S. A. Practical guidelines for the use of gradient boosting for molecular property prediction. *J. Cheminf.* **2023**, *15*, No. 73.

(37) Hastie, T.; Tibshirani, R.; Friedman, J. Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction.

(38) Meier, R. J. A way towards reliable predictive methods for the prediction of physicochemical properties of chemicals using the group contribution and other methods. *Appl. Sci.* **2019**, *9*, No. 1700.

(39) REACH-The New, 2006. https://pubs.acs.org/sharingguidelines.

(40) *Guidance on Requirements for Substances in Articles*; ECHA, 2017.

(41) OECD. *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*; OECD Publishing, 2023.

(42) Liu, M.; Zhang, L.; Li, S.; et al. Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints. *Toxicol. Lett.* **2020**, *332*, 88−96.

(43) Konda, L. S. K.; Keerthi Praba, S.; Kristam, R. hERG liability classification models using machine learning techniques. *Comput. Toxicol.* **2019**, *12*, No. 100089.

(44) Munawar, S.; Vandenberg, J. I.; Jabeen, I. Molecular docking guided grid-independent descriptor analysis to probe the impact ofwater molecules on conformational changes of hERG inhibitors in drug trapping phenomenon. *Int. J. Mol. Sci.* **2019**, *20*, No. 3385.

(45) Negami, T.; Araki, M.; Okuno, Y.; Terada, T. Calculation of absolute binding free energies between the hERG channel and structurally diverse drugs. *Sci. Rep.* **2019**, *9*, No. 16586.

(46) Zhang, Y.; Liu, J.; Shen, W. A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Appl. Sci.* **2022**, *12*, No. 8654.

(47) Fuadah, Y. N.; Qauli, A. I.; Marcellinus, A.; Pramudito, M. A.; Lim, K. M. Machine learning approach to evaluate TdP risk of drugs using cardiac electrophysiological model including inter-individual variability. *Front. Physiol.* **2023**, *14*, No. 1266084.

(48) Montomoli, J.; Romeo, L.; Moccia, S.; et al. Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *J. Intensive Med.* **2021**, *1*, 110−116.

(49) Fuadah, Y. N.; Lim, K. M. Optimal Classification of Atrial Fibrillation and Congestive Heart Failure Using Machine Learning. *Front. Physiol.* **2022**, *12*, No. 761013.

(50) Tanabe, K.; Lučić, B.; Amić, D.; et al. Prediction of carcinogenicity for diverse chemicals based on substructure grouping and SVM modeling. *Mol. Diversity* **2010**, *14*, 789−802.

(51) Shahhosseini, M.; Hu, G.; Pham, H. Optimizing Ensemble Weights and Hyperparameters of Machine Learning Models for Regression Problems.