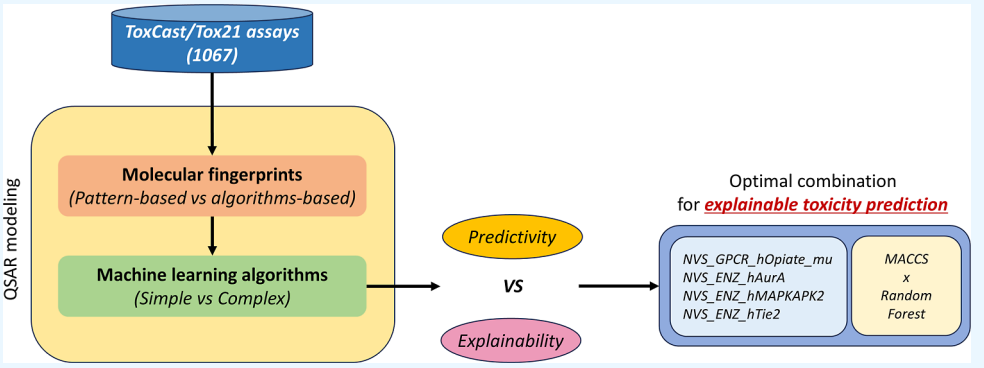
ACS OMEGA

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

Identification of Optimal Machine Learning Algorithms and Molecular Fingerprints for Explainable Toxicity Prediction Models Using ToxCast/Tox21 Bioassay Data

Donghyeon Kim, Jaeseong Jeong, and Jinhee Choi*

Cite This: ACS Omega 2024, 9, 37934−37941



ABSTRACT: Recent studies have primarily focused on introducing novel frameworks to enhance the predictive power of toxicity prediction models by refining molecular representation methods and algorithms. However, these methods are inherently complex and often pose challenges in understanding and explaining, leading to barriers in their regulatory adoption and validation. Therefore, it is necessary to select the optimal model, considering not only model performance but also interpretability. This study aimed to identify the optimal combination of molecular fingerprints (pattern-based versus algorithm-based) and machine learning algorithms (simple versus complex) for developing explainable toxicity prediction models through an comprehensive investigation of the ToxCast/Tox21 bioassay data set. For 1092 ToxCast/Tox21 assays, five molecular fingerprints (MACCS, Morgan, RDKit, Layered, and Patterned) and six algorithms (MLP, GBT, Random Forest, *k*NN, Logistic Regression, and Naïve Bayes) were used to train the models. Results showed that 35 models revealed acceptable performance (F1 score or accuracy is 0.8 or higher). Among the combinations, either MACCS or Morgan, paired with Random Forest, demonstrated robust performance compared with other molecular fingerprints and algorithms. MACCS and Random Forest are valuable, even when prioritizing interpretability. Consequently, the MACCS-Random Forest combination model based on four assays, targeting G protein-coupled receptor and kinase, were identified and they can be used to discern specific structural features or patterns in chemical compounds, offering explainable insights into toxicity-related chemical structures. This study indicates the importance of not disregarding the utilization of simple models when assessing both predictivity and interpretability within the context of chemical feature-based Tox21 data analysis.

INTRODUCTION

The recent advent of artificial intelligence (AI) is expected to catalyze fundamental changes in socioeconomic lifestyles and drive research and industrial innovations.[1] This innovation has also led to an increase in computational toxicology, allowing for the prediction of toxicity without the need to conduct tests on apical end points.[2] Particularly, advancements in text mining techniques for data collection have ushered in the era of big data in toxicology, facilitating the development of toxicity prediction models using AI.[3] Various computational techniques and databases are under development to predict toxicity based on the structures of chemicals. Toxicity prediction aims to provide information for drug development and prioritize chemicals for risk assessment by incorporating the latest findings in life science, including molecular biology and computational modeling technology.[4] With recent advances in science and technology, coupled with the utilization of existing toxicological information, the scope of

Received: May 11, 2024
Revised: July 22, 2024
Accepted: August 21, 2024
Published: August 27, 2024

https://doi.org/10.1021/acsomega.4c04474
ACS Omega 2024, 9, 37934−37941

biological tissue, exposure conditions, and variable factor analyses is expanding.[5]

The most studied method for developing a toxicity prediction model is the quantitative structure−activity relationship (QSAR) approach.[6] This approach is used to predict the toxicity of a chemical based on its structural information, under the hypothesis that the chemical structure determines its physicochemical properties. Conventionally, molecular descriptors (MD) and fingerprints (MF) are used to represent the molecular structure of a chemical during the training of a toxicity prediction model.[7] MF represents a molecular structure in the form of a bit vector based on various algorithms.[8] In our previous review of AI-based toxicity prediction models, MACCS was frequently used, followed by extended-connectivity fingerprints (ECFPs), and PubChem fingerprints.[9] Additionally, for machine learning algorithms, random forest (RF) and support vector machine (SVM) methods are widely utilized, while for deep learning, deep neural networks and artificial neural networks are commonly employed. More recently, many studies have focused on learning the structure of chemicals by employing novel methods instead of conventional MD and MF. For instance, Matsuzaka et al. developed a toxicity prediction model that learned molecular structure images using the Deep Snap technique.[10] This work is part of an effort to address the shortcomings of MD and MF, highlighting the limitations of conventional methods in achieving predictive accuracy.

However, newly developed methods are inherently complex and often pose challenges in understanding and explaining predictive results, leading to barriers to their regulatory adoption and validation. These concerns have prompted the introduction of explainable AI (XAI) models.[11] To meet the growing demand for XAI models, several explainable machine learning methods have recently been proposed. Explainable methods are largely classified into two types: intrinsic and post hoc methods.[12] Intrinsic interpretability is generally achieved by using simple models (e.g., single decision trees and generalized linear models), which is an advantage that has compelled toxicologists to continue using simple transparent models. To address the lack of interpretability of complex ML methods, post hoc interpretation can also be applied to an already trained model.[13] Several post hoc methods have been used to assess the relative importance of predictor variables. If the predictor variable analyzed at this time is utilized well, it can help analyze the relationship between the characteristics and toxicity of the chemical substance, rather than simply explaining the model's prediction results a posteriori. In this context, identifying an optimal machine learning algorithm-molecular fingerprint combination that is simple yet performs well and can retrospectively explain the structural characteristics of toxic chemicals remains an important task.

In response, this study aimed to identify optimal combinations of algorithms and molecular fingerprints using the ToxCast/Tox21 bioassay data set, which is frequently used for developing toxicity prediction models. Here, we focused on six algorithms: gradient boosting tree (GBT), RF, multilayer perceptron network (MLP), *k*-nearest neighborhood (*k*NN), logistic regression (LR), and Naïve Bayes (NB), along with five MFs: MACCS, Morgan, Layered, RDKit, and Pattern. These methods were applied to a data set comprising 1092 assays, from which several models were selected based on predictivity. We then identified the optimal combination of molecular fingerprints and algorithms, considering both predictability and explainability.

METHODS

QSAR Modeling Workflow. The study was conducted following a QSAR modeling workflow consisting of a total of nine steps including data collection, data selection, data preprocessing, MF generation, data splitting, data resampling, model training, performance evaluation and model selection, and model analysis (Figure S1). All steps were performed by using the Konstanz Information Miner (KNIME) Analytics Platform (https://www.knime.com).

ToxCast/Tox21 Data Collection and Assay Selection. ToxCast/Tox21 data were collected from the US Environmental Protection Agency (EPA)'s ToxCast and Tox21 summary file in invitroDBv3.2 (https://www.epa.gov/chemicalresearch/exploring-toxcast-data-downloadable-data). Among the data provided by ToxCast/Tox21, hit-call data labeled as 1 if the activity was confirmed (active class) and 0 if the activity was not confirmed (inactive class) were used.[14] ToxCast provides experimental results for 1473 in vitro bioassays. Since the ToxCast/Tox21 assays may or may not have a biological target depending on the assay method, only assays with defined biological target information were selected. Also, assays with at least two data points for each class (active and inactive) were selected for use in toxicity mechanism-based assessments.

Preprocessing of ToxCast/Tox21 Data. When performing in vitro assays, a burst phenomenon was observed, in which the activity of the assay suddenly increased in a narrow concentration range where cytotoxicity occurred.[15] Some of these activities might have had chemical effects on the assay target but could be false positives caused by cytotoxicity. In a previous study, the results obtained at the concentration where cytotoxicity was observed were classified using the standard *Z*-score.[15] In this study, only positive data with a *z*-score of three or higher were used as positive data. Chemical data were curated in three steps. First, chemicals with no simplified molecular-input line-entry system code representing the structure of the chemical were removed. Second, because the toxicity mechanism of inorganic compounds differs from that of organic compounds,[16] chemicals that were not organic compounds were removed. Third, the salts were converted to their corresponding largest free compound, e.g., free acid or free base. After data curation, the structural diversity of curated chemicals was calculated using the Tanimoto coefficient.[17] When calculating the Tanimoto coefficient, the results depend on the molecular fingerprint used.[18] In this study, ECFP4, which was evaluated to have a high performance among various molecular fingerprints, was used as a descriptor.[8]

Molecular Fingerprints Generation. In this study, five commonly used molecular fingerprints (MACCS, Morgan, layered, RDKit, and pattern) were used. The taxonomy of the molecular fingerprint is provided in Table S1. All molecular fingerprints were calculated using RDKit (http://www.rdkit.org/) nodes in the KNIME platform.

Model Training. For model training, a total of six algorithms including GBT,[19] RF,[20] MLP,[21] *k*NN,[22] LR,[23] and NB[24] were used to train the prediction models. Molecular fingerprints were used as variables for learning. The training and test sets were divided into 8:2 with 5-fold cross-validation. In model training, the performance decreases if the training data are imbalanced,[25] and various resampling techniques were

https://doi.org/10.1021/acsomega.4c04474
ACS Omega 2024, 9, 37934−37941

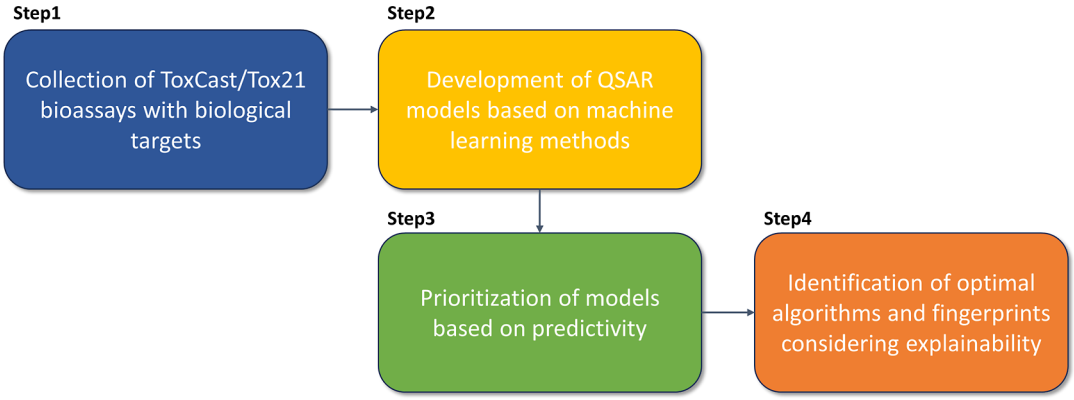


Figure 1. Workflow of the study.

Table 1. Summary of the Selected 1092 ToxCast/Tox21 Assay (invitroDBv3.2) (Adopted from the Study by Jeong et al., 2022)[a]

source	no. of assays	average no. of chemicals	average no. of active chemicals (%)	model	format	time point (h)	read out (function)
APR	62	453	4 (0.9)	HepG2	384-well plate	24, 72	signaling
ATG	240	2229	46 (2.1)	HepG2	24-well plate	24	reporter gene
BSK	146	1484	36 (2.4)	various cells	96-well plate	24	signaling
NCCT	2	317	134 (42.3)	tissue-based cell-free	384-well plate	0.5	binding
NVS	441	109	16 (14.7)	cell-free, tissue-based cell-free	48, 96, 384-well plates	0−24	enzymatic activity, binding
OT	17	1858	65 (3.5)	CHO-K1, HEK293T, HeLa	384-well plates	8, 16, 24	reporter gene, binding
TOX21	83	6542	172 (2.6)	HEK293T, MDA-kb2, MCF-7, BG1, HeLa, GH3, HepG2, HCT116, HEK293, ME-180	1536-well plates	24, 48	reporter gene
others	101	439	36 (8.2)	various	24, 96, 384-well plates	0−24	various

[a]APR: Cyprotex (formerly Apredica, LLC); ATG: Attagene, Inc.; BSK: BioSeek, Inc.; NCCT: National Center for Computational Toxicology, US EPA; NVS: NovaScreen Biosciences Corporation; OT: Odyssey Thera, Inc.; TOX21: Toxicity Testing in the 21st Century.

used to address this problem.[26] Because ToxCast/Tox21 data are extremely imbalanced,[27] SMOTE was used.[28,29] Model training was performed using base nodes provided by the KNIME platform (https://www.knime.com/).

Performance Evaluation. Accuracy, which is the most widely used metric for evaluating the performance of a model, may not be suitable for a model trained with imbalanced data. Even if prediction performance of the minor class is low, high accuracy can be achieved with only the major class predicted well.[30] If accuracy is used as the only metric for model performance evaluation, then the prediction performance of the minor class can be ignored. Therefore, in this study, the F1 score, which is a harmonized average of recall (or sensitivity) and precision, was also used. When both false positives and false negatives are low, the F1 score is high.[31] The performance of all models was confirmed by a 5-fold cross-validation performed on the KNIME platform.

RESULTS AND DISCUSSION

Study Design. Figure 1 illustrates the schematic workflow of our study, which aimed to identify the optimal machine learning algorithms and molecular fingerprints. The study consisted of four steps: data collection (**STEP1**), model fitting (**STEP2**), prioritization of models (**STEP3**), and identifying optimal combinations (**STEP4**). In **STEP1**, we compiled a set of ToxCast/Tox21 assays with biological targets. In **STEP2**, we trained machine learning models using the chemical structures and available in vitro toxicity data from the ToxCast/Tox21 database. We then prioritized models that achieved acceptable predictivity (F1 score or accuracy of 0.8 or higher) in **STEP3**. Finally, in **STEP4**, we identified an optimal combination of algorithms and fingerprints considering explainability of the models.

Assay Selection and Data Set Analysis. To develop a biological activity prediction model, 1092 assays with biological targets were selected from 1473 assays (Table 1). The source with the most abundant number of assays was NovaScreen Biosciences (NVS), with 441 assays, followed by Attagene (ATG) with 240 assays. The average number of chemicals varied from 109 to 6542, and the average number of active chemicals varied from four (0.9%) to 172 (2.6%). Except for National Center for Computational Toxicology (NCCT) and NVS, it was confirmed that the percentage of active chemicals in all assay sources was less than 10%, which was highly imbalanced. To investigate the structural diversity of chemicals in the ToxCast/Tox21 data set, molecular similarity was calculated using the Tanimoto coefficient. The results showed that most chemicals had a low structural similarity to each other (blue) (Figure S1). The mean value of all molecular similarities was 0.085 ± 0.058 and a 75% quantile value of 0.115, indicating significant structural diversity of chemicals in the data set. Because one of the purposes of the ToxCast/

https://doi.org/10.1021/acsomega.4c04474
ACS Omega 2024, 9, 37934−37941

Tox21 program is a screening of priority substances,[32] the program targets chemicals with diverse structures and has a broad range of mode of actions.[33,34] Moreover, this low similarity could be a disadvantage for the accurate prediction of active chemicals.[35] In addition, the average number of active chemicals is highly limited, which can cause a data imbalance problem. Our previous work demonstrated that an imbalanced data set in ToxCast/Tox21 assays can significantly constrain the model's predictivity. To address this issue, we employed the SMOTE method, which is identified as an optimal approach for addressing data imbalance problems.

Performance Evaluation of Models and Model Selection. A total of 30 models were trained by combining six algorithms and five MFs for each assay to identify the optimal algorithm and MF combinations for 1092 assays with different chemical structures and data point distributions (Table S2). Among the 30 combination models, the one with the best performance was selected as the representative model for the corresponding assay. Unfortunately, no model training was available for the assays with a limited number of positive data points, so only 737 out of 1092 models were acquired, and 670 out of 737 models exhibited very poor performance with an accuracy or F1 score lower than 0.5. Of the 67 models with accuracy and F1 score of 0.5 or higher, only models with at least one accuracy or F1 score greater than 0.8 were selected. Applying these criteria resulted in 35 models (Table 2) (detailed information on the selected 35 models is provided in Table S3). This cut-off criterion for the performance of predictive models trained on the ToxCast/Tox21 data set was established based on previous studies. In a study to predict hepatotoxicity using six supervised machine learning algorithms with chemical structure descriptors, the maximum accuracy was 0.84,[36] and the average accuracy was 0.88 in a study where 339 models were trained using ToxCast/Tox21 data using the MLP.[37] In a study comparing the performance of estrogen receptor activity prediction models using six algorithms, the F1 score did not exceed 0.6.[38]

Identification of Optimal Machine Learning Algorithms and Molecular Fingerprints. Among the 35 selected models, decision tree-based models were the most represented, with RF being the most prevalent at 13, followed by GBT, LR, and MLP with seven each and one NB model (Figure 2A). Since a simpler model with acceptable performance is often easier to interpret for the ToxCast/Tox21 data set, it was clearly the preferred choice due to its better explainability. RF and GBT employ ensemble methods, adding complexity, while *k*NN comprises multiple layers of interconnected nodes capable of learning complex patterns.[39] Conversely, NB, LR, and *k*NN are generally considered simpler models.[40−42] Simpler models are easier to interpret but may not capture complex relationships in the data as well as more sophisticated models, such as neural networks. Our findings, showing that RF performed well on ToxCast/Tox21 data, are consistent with those of other studies where RF was reported to be comparable to or better than SVM, *k*NN, and NB in predicting toxicity and biological activity.[43−47] Notably, the RF model has an advantage in interpreting active chemicals because, unlike other machine learning algorithms, it can identify descriptors used to predict the classification.[48] For example, Dreier et al. (2019) used the RF algorithm to develop QSAR models and identify chemical structures impairing mitochondrial membrane potential.[49] Moukheiber et al. (2021) identified protein features and pathways responsible for toxicity using RF-based

Table 2. Algorithms and Molecular Fingerprints Distribution of the 35 Selected Models[a]

	RF	GBT	LR	MLP	NB	*k*NN
MACCS (11)	NVS_GPCR_hOpiate_mu	NVS_TR_rAdoT	NVS_LGIC_h5HT3	NVS_ADME_rCYP2C6	0	0
	NVS_ENZ_hAurA		NVS_ENZ_hCSF1R	NVS_ENZ_hJak2		
	NVS_ENZ_hMAPKAPK2			CEETOX_H295R_CORTIC_noMTC_up		
	NVS_ENZ_hTie2			TOX21_TR_RXR_BLA_Agonist_Followup_ratio		
morgan (11)	NVS_ADME_rCYP3A2	NVS_ADME_hCYP2J2	NVS_ADME_rCYP2C11	TOX21_SBE_BLA_Agonist_ratio	0	0
	TOX21_GR_BLA_Agonist_ratio	NVS_ADME_rCYP1A1	NVS_ENZ_hPTPN11	CEETOX_H295R_ESTRONE_noMTC_dn		
	TOX21_PR_LUC_Followup_Agonist	NVS_GPCR_gH2				
		NVS_TR_hAdoT				
RDKit (6)	BSK_3C_MIG_down	TOX21_AR_BLA_Agonist_ratio	NVS_ADME_rCYP2A1	NVS_ENZ_hMMP13	NVS_ADME_rCYP2B1	0
			NVS_GPCR_hETB			
pattern (5)	APR_Hepat_Apoptosis_48 h_up	NVS_ENZ_hVEGFR1	0	0	0	0
	NVS_ADME_hCYP2A6					
	NVS_ADME_rCYP2A2					
	NVS_GPCR_hLTB4_BLT1					
layered (2)	CEETOX_H295R_ESTRADIOL_noMTC_up	0	NVS_GPCR_rGHB	0	0	0
total (35)	13	7	7	7	1	0

[a]RF: Random Forest, GBT: Gradient Boosting Tree, LR: Logistic Regression, MLP: Multilayered Perceptron, NB: Naïve Bayes, *k*NN: k-Nearest Neighborhood.

https://doi.org/10.1021/acsomega.4c04474
ACS Omega 2024, 9, 37934−37941

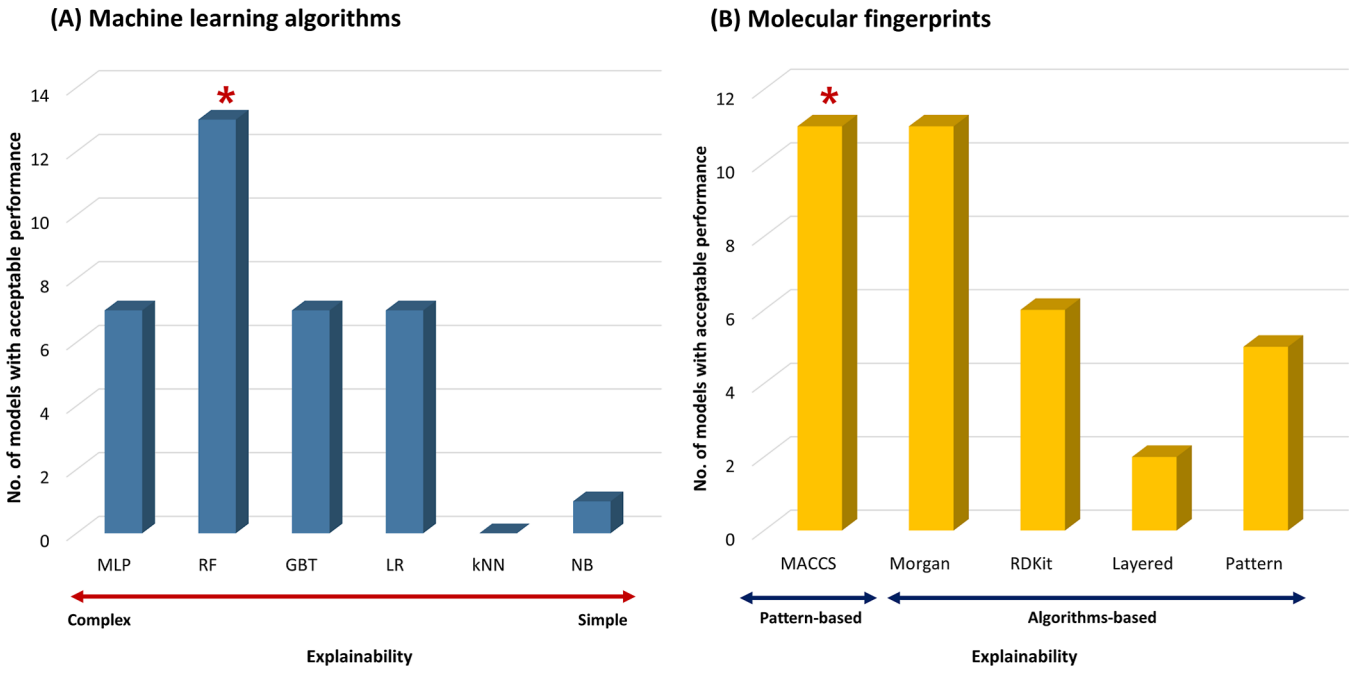


Figure 2. Number of models with acceptable performance based on (A) complexity of algorithms and (B) molecular fingerprint generation methods.

Table 3. Detailed Information on the Top-Ranked Models Which can be Utilized to Develop Explainable Toxicity Prediction Models

assay name	target family	target sub type	target symbol	no. of data	molecular fingerprint	algorithm	F1	ACC
NVS_GPCR_hOpiate_mu	GPCR	receptor	OPRM1	255	MACCS	RF	0.511	0.803
NVS_ENZ_hAurA	kinase	enzyme	AURKA	95	MACCS	RF	0.750	0.977
NVS_ENZ_hMAPKAPK2	kinase	enzyme	MAPKAPK2	102	MACCS	RF	0.500	0.938
NVS_ENZ_hTie2	kinase	receptor	TEK	136	MACCS	RF	0.500	0.967

machine learning models.[50] Therefore, RF has great potential to be used for developing robust QSAR models and identifying significant features to explain their predictive results.

Meanwhile, MACCS and Morgan fingerprints were the most represented, with 11 models each, followed by RDKit fingerprints with six, pattern fingerprints with five, and layered fingerprints with two (Figure 2B). Consistent with our results, previous studies have shown that MACCS and Morgan fingerprints are superior for model training when compared with other molecular fingerprints for the ToxCast/Tox21 data. Ciallella et al. reported that the MACCS-RF model performed the best among other combination models such as combinations of FCFP6 and ECFP6 fingerprints with *k*NN and Bernoulli NB.[51] Balabin and Judson[52] trained *k*NN models using Morgan, Indigo, Daylight, and MACCS fingerprints for hER agonist, antagonist, and binding activity data sets. Morgan fingerprints performed the best among fingerprints for the agonist and binding activity data. Banerjee et al. also reported that MACCS and ECFP4, circular topological fingerprints such as Morgan, performed better and more consistently than ToxPrint and Estate fingerprints in a toxicity database.[53] Notably, MACCS is a pattern-based fingerprint that can capture specific structural features or patterns in chemical compounds, providing explainable insights into the chemical structure related to toxicity.[54] The MACCS fingerprint contains 166 public keys that represent the most common substructures, allowing them to detect potential structural alerts via the calculation of their frequencies in a data set.[55] For example, Yang et al. (2017) used MACCS for developing QSAR models and explained the relationship between the substructure of chemicals and model accuracy.[4] This approach is also useful to define potential toxicophores which can give insight into designing greener chemicals.[56] Therefore, we suggest that the RF−MACCS combinations in machine learning models can be used to develop explainable toxicity prediction models considering both predictivity and explainability.

Consequently, four models were selected, considering both predictability and interpretability (Table 3). These models target G protein-coupled receptors (GPCRs) and kinases. GPCRs play a pivotal role in integrating extracellular signals into downstream responses, such as intracellular signaling and cell cycle regulation.[57] Kinases are enzymes that regulate the activity, reactivity, and binding ability of proteins through phosphorylation.[58] Both GPCRs and kinase are of great value for toxicological studies as they are two distinct signaling mechanisms involved in major physiological processes, consequently affecting a significant number of diseases.[59,60] Therefore, the selected models can identify potential toxicants involved in toxicity mechanisms and discern specific structural features or patterns in compounds, providing explainable insights into toxicity-related chemical structures. A follow-up case study to utilize these selected models is in progress.

This study presents a case study using the ToxCast/Tox21 assay data to address two key questions: (1) is there a trade-off between predictivity and interpretability of machine learning

https://doi.org/10.1021/acsomega.4c04474
ACS Omega 2024, 9, 37934−37941

models? and (2) what combination of molecular fingerprints and algorithms is optimal for developing explainable toxicity prediction models? The primary conclusion of our work is that for the 737 ToxCast/Tox21 bioassay data set, MACCS and RF can be considered optimal combinations considering both predictivity and explainability. Furthermore, we found that the trade-off between predictivity and interpretability is not always present, and an optimal model that simultaneously achieves predictivity and interpretability may exist, depending on the data. Wu et al. (2021) followed a similar approach to investigate the trade-off between predictivity and interpretability for machine-learning-powered predictive toxicology using the Tox21 data set.[61] While our approach shares similarities with theirs, their main focus has been on identifying machine learning algorithms, whereas our approach aims to consider both algorithms and molecular fingerprints, encompassing both intrinsic and post hoc methods. We acknowledge that optimal combinations may vary depending on the data set characteristics and thus suggest considering a comprehensive range of possible combinations when developing models.

However, several limitations exist regarding the current work. Here, we only considered chemical feature-based models, while recently, many toxicity prediction models have been trained using biological descriptors such as gene expression, toxicokinetics, and clinical data as model features.[62,63] Utilizing biological features instead of chemical-backed features may revolutionize predictive toxicology, as they can be more relevant to the toxicity end points. Therefore, a more comprehensive investigation and comparison between chemical- and biological-feature-based models are also necessary. Moreover, the current work selected the optimal model with higher interpretability through a simple comparison. Future directions may involve evaluating metrics to measure interpretability qualitatively or quantitatively on demand. This would help investigate the extent to which interpretability could be enhanced by using different approaches and whether they yield similar interpretations.

Beyond the challenges of explainability of the AI model itself, the goal of developing an explainable model is to mechanistically explain the predictive results. In this regard, the ToxCast/Tox21 data offers the feasibility to predict in vivo toxic effects, aligning with one of the primary objectives of the ToxCast/Tox21 program.[64] For instance, Liu et al. combined in vitro assay data from the ToxCast/Tox21 data with hundreds of in vivo data from ToxRefDB and developed a model with a high accuracy rate for hypertrophic, injured, and proliferative lesions.[65] Martin et al. (2011) utilized both in vivo data from ToxRefDB and unique chemicals alongside ToxCast/Tox21 assays from the ToxCast database to create a model with high accuracy.[66] Likewise, ToxCast/Tox21 data are valuable for predicting in vivo toxicity as most ToxCast/Tox21 bioassays target specific biological mechanisms to identify the toxicity mechanisms of chemicals. Therefore, it is essential to establish connections between the ToxCast/Tox21 bioassays and apical end points to facilitate the development of XAI models.

CONCLUSIONS

This study presents a case study to identify the optimal combination of molecular fingerprints (pattern-based versus algorithm-based) and machine learning algorithms (simple versus complex) for developing explainable toxicity prediction models. Through an in-depth investigation of the ToxCast/Tox21 bioassay data set, 35 models were prioritized based on their predictive power. Based on these results, the MACCS and RF combination was suggested to be optimal for developing explainable models. Therefore, when considering both predictivity and interpretability in the context of chemical-feature-based Tox21 data analysis, we recommend not overlooking the use of simple models. They offer higher interpretability while still potentially achieving similar performance compared to more complicated approaches.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c04474.

Taxonomy of fingerprints used in this study; chemical structural similarity analysis in ToxCast data set using Tanimoto coefficient; performance comparison of machine learning models on various algorithms and molecular fingerprints combination; and detailed information on the selected 35 models (XLSX)

AUTHOR INFORMATION

Corresponding Author

Jinhee Choi − *School of Environmental Engineering, University of Seoul, Seoul 02504, Republic of Korea;* orcid.org/0000-0003-3393-7505; Phone: 82-2-6490-2869; Email: jinhchoi@uos.ac.kr; Fax: 82-2-6490-2859

Authors

Donghyeon Kim − *School of Environmental Engineering, University of Seoul, Seoul 02504, Republic of Korea;* orcid.org/0000-0001-7432-2975

Jaeseong Jeong − *School of Environmental Engineering, University of Seoul, Seoul 02504, Republic of Korea;* orcid.org/0000-0002-3860-0648

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c04474

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This study was supported by the Midcareer Researcher Program (2020R1A2C3006838) through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, and by a grant from the Korean Ministry of Environment through 'Environmental Health R&D Program' (2021003310005).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Soo-yong Bae for his assistance in developing machine learning models.

REFERENCES

(1) Far, S. B.; Rad, A. I. Internet of Artificial Intelligence (IoAI): The Emergence of an Autonomous, Generative, and Fully Human-Disconnected Community. *Discover Appl. Sci.* **2024**, *6* (3), 91.

(2) Krewski, D.; Acosta, D.; Andersen, M.; Anderson, H.; Bailar, J. C.; Boekelheide, K.; Brent, R.; Charnley, G.; Cheung, V. G.; Green,

S.; Kelsey, K. T.; Kerkvliet, N. I.; Li, A. A.; McCray, L.; Meyer, O.; Patterson, R. D.; Pennie, W.; Scala, R. A.; Solomon, G. M.; Stephens, M.; Yager, J.; Zeise, L. Toxicity Testing in the 21st Century: A Vision and a Strategy. *J. Toxicol. Environ. Health, Part B* **2010**, *13*, 51−138.
(3) Kleinstreuer, N.; Hartung, T. Artificial Intelligence (AI)—It's the End of the Tox as We Know It (and I Feel Fine)*. *Arch. Toxicol.* **2024**, *98*, 735−754.
(4) 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*, 30.
(5) Schmeisser, S.; Miccoli, A.; von Bergen, M.; Berggren, E.; Braeuning, A.; Busch, W.; Desaintes, C.; Gourmelon, A.; Grafström, R.; Harrill, J.; Hartung, T.; Herzler, M.; Kass, G.; Kleinstreuer, N.; Leist, M.; Luijten, M.; Marx-Stoelting, P.; Poetz, O.; van Ravenzwaay, B.; Roggeband, R.; Rogiers, V.; Roth, A.; Sanders, P.; Thomas, R. S.; Vinggaard, A. M.; Vinken, M.; van de Water, B.; Luch, A.; Tralau, T. New Approach Methodologies in Human Regulatory Toxicology − Not If, but How and When! *Environ. Int.* **2023**, *178*, 108082.
(6) Dudek, A.; Arodz, T.; Galvez, J. Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Comb. Chem. High Throughput Screen.* **2006**, *9* (3), 213−228.
(7) Anjaneyulu, B.; Goswami, S.; Banik, P.; Chauhan, V.; Raghav, N.; Chinmay. Artificial Intelligence: Machine Learning for Chemical Sciences. *Chem. Afr.* **2024**, *7*, 3443−3459.
(8) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71* (C), 58−63.
(9) Jeong, J.; Choi, J. Artificial Intelligence-Based Toxicity Prediction of Environmental Chemicals: Future Directions for Chemical Management Applications. *Environ. Sci. Technol.* **2022**, *56* (12), 7532−7543.
(10) Matsuzaka, Y.; Uesawa, Y. Optimization of a Deep-Learning Method Based on the Classification of Images Generated by Parameterized Deep Snap a Novel Molecular-Image-Input Technique for Quantitative Structure-Activity Relationship (QSAR) Analysis. *Front. Bioeng. Biotechnol.* **2019**, *7*, 65.
(11) Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805.
(12) Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116* (44), 22071−22080.
(13) Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable Ai: A Review of Machine Learning Interpretability Methods. *Entropy* **2020**, *23*, 18−45.
(14) Filer, D. L.; Kothiya, P.; Setzer, R. W.; Judson, R. S.; Martin, M. T. Tcpl: The ToxCast Pipeline for High-Throughput Screening Data. *Bioinformatics* **2017**, *33* (4), 618−620.
(15) Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S. Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity onIn VitroAssay Activity Across a Diverse Chemical and Assay Space. *Toxicol. Sci.* **2016**, *152* (2), 323−339.
(16) LoPachin, R. M.; Gavin, T. Molecular Mechanism of Acrylamide Neurotoxicity: Lessons Learned from Organic Chemistry. *Environ. Health Perspect.* **2012**, *120* (12), 1650−1657.
(17) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015**, *7* (1), 1−13.
(18) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108−119.
(19) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56* (12), 2353−2360.
(20) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(21) Agatonovic-Kustrin, S.; Beresford, R. Basic Concepts of Artificial Neural Network (ANN) Modeling and Its Application in Pharmaceutical Research. *J. Pharm. Biomed. Anal.* **2000**, *22* (5), 717−727.
(22) Denœux, T. A K-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory. *IEEE Trans. Syst. Man Cybern.* **1995**, *25* (5), 804−813.
(23) Algamal, Z. Y.; Lee, M. H.; Al-Fakih, A. M.; Aziz, M. High-Dimensional QSAR Classification Model for Anti-Hepatitis C Virus Activity of Thiourea Derivatives Based on the Sparse Logistic Regression Model with a Bridge Penalty. *J. Chemom.* **2017**, *31* (6), 1−8.
(24) Sun, H. A Naive Bayes Classifier for Prediction of Multidrug Resistance Reversal Activity on the Basis of Atom Typing. *J. Med. Chem.* **2005**, *48* (12), 4031−4039.
(25) Abraham, A.; Elrahman, S. M. A. A Review of Class Imbalance Problem. *J. Netw. Innovat. Comput.* **2013**, *1*, 332−340.
(26) Gosain, A.; Sardana, S. Handling Class Imbalance Problem Using Oversampling Techniques: A Review. In *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*; IEEE, 2017; pp 79−85.
(27) Jeong, J.; Kim, D.; Choi, J. Application of ToxCast/Tox21 data for toxicity mechanism-based evaluation and prioritization of environmental chemicals: Perspective and limitations. *Toxicol. in Vitro* **2022**, *84*, 105451.
(28) Kovács, B.; Tinya, F.; Németh, C.; Ódor, P. Unfolding the Effects of Different Forestry Treatments on Microclimate in Oak Forests: Results of a 4-Yr Experiment. *Ecol. Appl.* **2020**, *30* (2), 321−357.
(29) Yap, B. W.; Rani, K. A.; Rahman, H. A. A.; Fong, S.; Khairudin, Z.; Abdullah, N. N. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In *Lecture Notes in Electrical Engineering*; Springer: Singapore, 2014; Vol. *285*, pp 13−22.
(30) Jeni, L. A.; Cohn, J. F.; De La Torre, F. Facing Imbalanced Data—Recommendations for the Use of Performance Metrics. In *Proceedings—2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*; ACII, 2013; pp 245−251.
(31) Powers, D. M. W. *Evaluation: from Precision*; Recall and F-Factor to ROC, Informedness, Markedness & Correlation, 2007; p 24.
(32) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The Toxcast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol. Sci.* **2007**, *95* (1), 5−12.
(33) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29* (8), 1225−1251.
(34) Jeong, J.; Kim, D.; Choi, J. Application of ToxCast/Tox21 Data for Toxicity Mechanism-Based Evaluation and Prioritization of Environmental Chemicals: Perspective and Limitations. *Toxicol. in Vitro* **2022**, *84*, 105451.
(35) Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. *Expert Opin. Drug Discovery* **2016**, *11* (2), 137−148.
(36) Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting Hepatotoxicity Using ToxCast in Vitro Bioactivity and Chemical Structure. *Chem. Res. Toxicol.* **2015**, *28* (4), 738−751.
(37) Jeong, J.; Choi, J. Development of AOP Relevant to Microplastics Based on Toxicity Mechanisms of Chemical Additives

https://doi.org/10.1021/acsomega.4c04474
ACS Omega 2024, 9, 37934−37941

Using ToxCast™ and Deep Learning Models Combined Approach. *Environ. Int.* **2020**, *137* (November 2019), 105557.
(38) Zorn, K. M.; Foil, D. H.; Lane, T. R.; Russo, D. P.; Hillwalker, W.; Feifarek, D. J.; Jones, F.; Klaren, W. D.; Brinkman, A. M.; Ekins, S. Machine Learning Models for Estrogen Receptor Bioactivity and Endocrine Disruption Prediction. *Environ. Sci. Technol.* **2020**, *54* (19), 12202−12213.
(39) More, A. S.; Rana, D. P. Review of Random Forest Classification Techniques to Resolve Data Imbalance. In *Proceedings— 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017*; IEEE, 2017; pp 72−78.
(40) Zhang, Z. Introduction to Machine Learning: K-Nearest Neighbors. *Ann. Transl. Med.* **2016**, *4* (11), 218.
(41) Bonetta, R.; Valentino, G. Machine Learning Techniques for Protein Function Prediction. In *Proteins: Structure, Function and Bioinformatics*; John Wiley and Sons Inc., 2020; pp 397−413.
(42) Zhang, Z. Naïve Bayes Classification in R. *Ann. Transl. Med.* **2016**, *4* (12), 241.
(43) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *J. Chem. Inf. Model.* **2009**, *49* (11), 2481−2488.
(44) Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of Nonlinear Qsar Models Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model.* **2009**, *49* (11), 2551−2558.
(45) Cannon, E. O.; Bender, A.; Palmer, D. S.; Mitchell, J. B. O. Chemoinformatics-Based Classification of Prohibited Substances Employed for Doping in Sport. *J. Chem. Inf. Model.* **2006**, *46* (6), 2369−2380.
(46) Low, Y.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuzmin, V.; Fourches, D.; Zhu, H.; Rusyn, I.; Tropsha, A. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chem. Res. Toxicol.* **2011**, *24* (8), 1251−1262.
(47) Wu, J.; Zhang, Q.; Wu, W.; Pang, T.; Hu, H.; Chan, W. K. B.; Ke, X.; Zhang, Y. WDL-RF: predicting bioactivities of ligand molecules acting with G protein-coupled receptors by combining weighted deep learning and random forest. *Bioinformatics* **2018**, *34* (13), 2271−2282.
(48) Strobl, C.; Boulesteix, A. L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinf.* **2008**, *9*, 307.
(49) Dreier, D. A.; Denslow, N. D.; Martyniuk, C. J. Computational in Vitro Toxicology Uncovers Chemical Structures Impairing Mitochondrial Membrane Potential. *J. Chem. Inf. Model.* **2019**, *59* (2), 702−712.
(50) Moukheiber, L.; Mangione, W.; Moukheiber, M.; Maleki, S.; Falls, Z.; Gao, M.; Samudrala, R. Identifying Protein Features and Pathways Responsible for Toxicity Using Machine Learning and Tox21: Implications for Predictive Toxicology. *Molecules* **2022**, *27* (9), 3021.
(51) Ciallella, H. L.; Russo, D. P.; Aleksunes, L. M.; Grimm, F. A.; Zhu, H. Predictive Modeling of Estrogen Receptor Agonism, Antagonism, and Binding Activities Using Machine- and Deep-Learning Approaches. *Lab. Invest.* **2021**, *101* (4), 490−502.
(52) Balabin, I. A.; Judson, R. S. Exploring Non-Linear Distance Metrics in the Structure-Activity Space: QSAR Models for Human Estrogen Receptor. *J. Cheminf.* **2018**, *10* (1), 47.
(53) Banerjee, P.; Siramshetty, V. B.; Drwal, M. N.; Preissner, R. Computational Methods for Prediction of in Vitro Effects of New Chemical Structures. *J. Cheminf.* **2016**, *8* (1), 51.
(54) Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and Applications of Chemical Fingerprint for Hit and Lead Screening. In *Drug Discovery Today*; Elsevier Ltd., 2022.
(55) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.
(56) Williams, D. P. Toxicophores: Investigations in Drug Safety. *Toxicology* **2006**, *226* (1), 1−11.
(57) Tuteja, N. Signaling through G Protein Coupled Receptors. *Plant Signal. Behav.* **2009**, *4* (10), 942−947.
(58) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298* (5600), 1912−1934.
(59) Guillien, M.; le Maire, A.; Mouhand, A.; Bernadó, P.; Bourguet, W.; Banères, J. L.; Sibille, N. IDPs and Their Complexes in GPCR and Nuclear Receptor Signaling. *Prog. Mol. Biol. Transl. Sci.* **2020**, *174*, 105−155.
(60) Zarrin, A. A.; Bao, K.; Lupardus, P.; Vucic, D. Kinase Inhibition in Autoimmunity and Inflammation. *Nat. Rev. Drug Discov.* **2021**, *20*, 39−63.
(61) Wu, L.; Huang, R.; Tetko, I. V.; Xia, Z.; Xu, J.; Tong, W. Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. *Chem. Res. Toxicol.* **2021**, *34* (2), 541−549.
(62) Ring, C.; Sipes, N. S.; Hsieh, J. H.; Carberry, C.; Koval, L. E.; Klaren, W. D.; Harris, M. A.; Auerbach, S. S.; Rager, J. E. Predictive Modeling of Biological Responses in the Rat Liver Using in Vitro Tox21 Bioactivity: Benefits from High-Throughput Toxicokinetics. *Comput. Toxicol.* **2021**, *18*, 100166.
(63) Adeluwa, T.; McGregor, B. A.; Guo, K.; Hur, J. Predicting Drug-Induced Liver Injury Using Machine Learning on a Diverse Set of Predictors. *Front. Pharmacol* **2021**, *12*, 648805.
(64) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29*, 1225−1251.
(65) Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting Hepatotoxicity Using ToxCast in Vitro Bioactivity and Chemical Structure. *Chem. Res. Toxicol.* **2015**, *28* (4), 738−751.
(66) Martin, M. T.; Knudsen, T. B.; Reif, D. M.; Houck, K. A.; Judson, R. S.; Kavlock, R. J.; Dix, D. J. Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. *Biol. Reprod.* **2011**, *85* (2), 327−339.

https://doi.org/10.1021/acsomega.4c04474
ACS Omega 2024, 9, 37934−37941