



Identifying Molecular Toxicity Patterns Through Interpretable Learning of Chemical Properties

Pratibha Omkar

Associate Professor, Department of Pharmacology, SDM College of Medical Sciences and Hospital, Shri Dharmasthala Manjunatheshwara University, Dharwad, Karnataka, Email: pratibha.omkar@gmail.com, Orcid ID: 0000-0001-7353-4386

Received: 22/12/2025

Revised: 27/02/2026

Accepted: 17/04/2026

Published: 29/04/2026

Abstract— The prediction of molecular toxicity is becoming a more and more important requirement, and computational methods are becoming more and more powerful in their ability to describe complex relationships between molecular structure and biological activity. This study presents an interpretable machine-learning framework to extract the toxicity-related molecular patterns from the Tox21 dataset based on the physicochemical and structural properties. Logistic Regression, Support Vector Machine, Random Forest, and XGBoost were tested on 12 nuclear receptor and stress-response endpoints with the use of molecular descriptors and Morgan fingerprints as complementary representations. Random Forest had the highest overall performance with the highest values of ROC-AUC = 0.852, PR-AUC = 0.526, F1 score = 0.664, balanced accuracy = 0.793 and Matthews correlation coefficient = 0.579. NR-AR-LBD (ROC-AUC = 0.91), SR-MMP (0.89) and NR-AhR (0.88) showed highest endpoint specific performance. The hybrid representation (physicochemical descriptors + Morgan fingerprints) performed better as a whole than either of the two feature sets used. LogP, molecular weight, topological polar surface area, aromatic ring count and molecular refractivity were found to be the most influential predictors by SHAP analysis. Performing cross-endpoints analysis, a different physicochemical signature was also identified between the nuclear receptor pathway and the stress-response pathway. The results indicate that interpretable machine learning can help go beyond just prediction and discover nonlinear, chemically meaningful and pathway-specific patterns. The proposed framework provides a clear foundation for computational toxicology, chemical prioritization and mechanism-oriented toxicity screening.

Keywords— Tox21; molecular toxicity; interpretable machine learning; SHAP; physicochemical descriptors

1. INTRODUCTION

As more and more synthetic chemicals, pharmaceuticals, industrial chemicals and environmental contaminants are introduced, there is a growing need for effective ways of assessing toxicity. The traditional toxicological evaluation can be resource-intensive, time consuming, and rely heavily on large amounts of *in vivo* testing. The limits put on these have led to a shift towards predictive, mechanism-oriented methods that can help prioritize chemicals for additional testing and avoid the need for conventional animal methods. The overall goal of the 21st century toxicology is thus to use high-throughput screening, computer modeling, understanding of the mechanisms, and predictions based on data for chemical risk assessment (Krewski et al., 2010).

The Toxicology in the 21st Century (Tox21) programme is an important platform for large scale computational toxicology within this transition. Tox21 combines high throughput screening and predictive modelling to study the effects of chemicals on biologically relevant pathways. The Tox21 Challenge provides benchmark data for prediction of activity in nuclear receptor and stress-response pathways, and the resulting Tox21 10K chemical library contains thousands of structurally diverse compounds for prediction and mechanistic analysis (Huang et al., 2016a; Huang et al., 2016b). Importantly, these endpoints are reports of pathway specific assay activity and not a single measure of toxicity.

The quantitative structure – activity relationship modelling offers a central model in which the relationship between molecular structure and biological activity is connected. QSAR approaches characterize chemicals by their measurable properties, including molecular size, lipophilicity, polarity, topology, electronic characteristics and structural fragments. These properties are chemically relevant such as affecting molecular transport, membrane interaction, solubility, target binding and reactivity. For instance, lipophilicity may modify membrane affinity, distribution, polar surface area can modify molecular transport and interactions with water, aromaticity can modify structural rigidity and target interaction, molecular size can modify accessibility to biological binding sites, etc. In this way, modern QSAR has evolved from the linear relationships to machine learning techniques that have the ability to describe nonlinear effects and complex relationships between molecular properties (Cherkasov et al., 2014; Muratov et al., 2020).

Standardized molecular datasets have helped the growth of machine-learning methods for toxicity prediction. Tox21 was selected as one of the key datasets in the benchmark framework of Molecular machine learning developed by MoleculeNet (Wu et al., 2018). DeepTox also showed that complex structure–activity relationships can be learned across Tox21 endpoints with the use of neural-network based models (Mayr et al., 2016). Such advances demonstrate that molecular structures have significant

predictive information, but predictive information does not provide information on which chemical properties make a contribution to the response of the assay.

Molecular representation is thus one of the key aspects of toxicity modelling. Physicochemical descriptors can be easily understood measures like molecular weight, LogP, topological polar surface area, hydrogen-bonding capacity, flexibility, aromaticity and complexity. At the same time, circular fingerprints represent localized molecular substructures that can't be completely described by global fingerprints. The extended-connectivity fingerprints are widely used because they describe the structural environment around each atom, as well as substructural motifs that are known to be associated with biological activity (Rogers & Hahn, 2010). Therefore, it could be advantageous to use both physicochemical descriptors and structural fingerprints, where descriptors can help to interpret the chemical structure and fingerprints can help to discriminate the structures.

One of the drawbacks of the advanced molecular prediction models is that they are not very interpretable. Good predictive performance does not necessarily reveal the molecular properties, structural fragments or combinations of properties that govern the responses in specific pathways. Explainable AI methods overcome this by breaking down the predictions of complex models into individual contributions across their features. SHapley Additive exPlanations (SHAP) are a theoretically motivated approach for measuring the impact of individual features in the model predictions and come with both global and local interpretation (Lundberg & Lee, 2017). Moreover, in molecular modelling, SHAP has been successfully applied to explaining relationships between chemical features and predicted biological activity in the context of machine-learning-based cheminformatics (Rodríguez-Pérez & Bajorath, 2020).

Nevertheless, a crucial challenge exists between accurate prediction of toxicity and interpretable discovery of molecular patterns. Existing literature is largely devoted to performance of the classification or to comparisons between models, but less effort is put on the discovery of recurrent combinations of physicochemical and structural properties in various toxicity-related pathways. The novelty of the present study therefore will be to reveal cross-endpoint molecular-property pattern discovery especially whether or not there are differences between shared signatures and pathway-specific signatures between nuclear receptor and stress-response assays.

Therefore, the purpose of this study is to discover molecular toxicity patterns through interpretable machine learning of chemical properties based on the Tox21 dataset. The integration of physicochemical descriptors, molecular fingerprints, multiple supervised learning algorithms, and

SHAP based interpretation is extended to nuclear receptor and stress-response endpoints. The study not only suggests pathway specific assay activity, but identifies molecular features and combinations of features that are common to and unique to toxicity-related responses.

2. LITERATURE REVIEW

Traditional animal based testing has been increasingly replaced by high throughput screening, cheminformatics and data driven prediction in computational toxicology. Large programmes like ToxCast showed that environmental chemicals could be screened through a variety of biological assays, and the chemicals could be prioritized for further testing with computational tools (Judson et al., 2010). Tox21 expanded this paradigm by furnishing benchmark assay information for nuclear receptor and stress-response pathways. Abdelaziz et al. (2016) were able to show that molecular descriptors that were derived from *in silico* data could be used to accurately classify across the Tox21 assays, and Capuzzi et al. (2016) were able to show the utility of QSAR models for two nuclear receptor and stress-response endpoints. Uesawa (2016) also confirmed that Random Forest models are appropriate to identify chemicals associated with the activation of toxicity-related pathways.

Such models are valid to a great extent, depending on the quality of the chemical structure and the modelling discipline. Fourches et al. (2010) showed that cheminformatics and QSAR results may significantly be influenced by duplicate compounds, salts, inconsistent structures and errors in representation. Descriptor calculation and model development are thus a prerequisite for chemical curation. Tropsha (2010) also highlighted cross validation, proper train test separation, domain of applicability and replicability of modelling procedures. These issues are of special significance in the context of molecular toxicity modelling, where structurally similar compounds could otherwise be split between the training and testing sets, resulting in an overly positive prediction of predictive performance.

Another important quality factor of models is the molecular representation. Physicochemical descriptors are measurements of properties that can affect biological activity and are chemically interpretable. A convenient way to calculate the topological polar surface area (TPSA) that describes the polarity and the transport-related behaviour of a molecule was introduced by Ertl et al. (2000). Atomic contribution methods for estimating the LogP and molecular refractivity were developed by Wildman and Crippen (1999) for lipophilicity and polarizability, respectively. The advantage of these descriptors is that they have direct chemical meaning. Meanwhile, more complicated representations can carry more structural information that may not be fully represented by traditional descriptors. Learned molecular representations can enhance

property prediction, but increasing the complexity of the representation can further decrease the transparency and chemical interpretability of the representation, as demonstrated by Yang et al. (2019).

Molecular toxicity prediction is often a nonlinear relationship that requires multiple molecular properties, for which machine-learning models are well suited. Random Forest is an ensemble method that can model nonlinear interactions among features and different effects of the descriptors (Breiman, 2001). XGBoost is an extension of tree-based learning with gradient boosting and regularization, and works well for high-dimensional predictive tasks (Chen & Guestrin, 2016). Another popular method for nonlinear molecular classification is Support Vector Machines, generally used to classify problems that have complex feature spaces and classes that are not linearly separable (Cortes & Vapnik, 1995). In the field of toxicity prediction, these models represent complementary approaches to test if the physicochemical and structural properties are able to differentiate between active and inactive compounds in a variety of toxicity assay endpoints.

Given the fact that Tox21 endpoints are often minority class (active compounds), the strategy for evaluating them is also crucial. In such cases, the overall accuracy may significantly overestimate model performance. When the data are highly imbalanced, with positive to negative class ratio in the order of 1:100, Saito and Rehmsmeier (2015) showed that precision-recall analysis is more informative than ROC analysis as it cuts down on the focus from ROC analysis to positive-class recovery. Matthews correlation coefficient provides another good measure since it includes all 4 elements of the confusion matrix. Chicco and Jurman (2020) demonstrated that MCC is a more balanced measure than accuracy or F1 in binary classification.

Realistic validation is also a factor in generalizability. Sheridan (2013) demonstrated that methods of validation that more closely mimic prospective prediction yield more realistic estimates of molecular model performance than do simple random splitting. The principle can be used in Scaffold aware validation in cheminformatics: structurally related analogues shouldn't be freely distributed between training and test sets if the goal is to predict activity for a truly unseen scaffold.

Toxic21 studies therefore have a proven predictive power of molecular descriptors, QSAR frameworks, Random Forest models, etc. which are nonlinear learning approaches. But these studies were mostly focused on optimizing the end-to-end prediction or performance comparison of the models. They tended to focus less on whether there is a consistent molecular signature across the various toxicity pathways that have been biologically distinct or whether there are systematically different molecular signatures in nuclear receptor and stress-response assays. Such a distinction is chemically relevant because the properties with which they are associated can

play different roles as a function of different biological responses, such as lipophilicity, polarity, molecular size, aromaticity, and polarizability.

The present study takes this work a step further, however, shifting the focus from prediction to interpretable cross-pathway pattern discovery. It integrates curated molecular structures, physico-chemical descriptors, structural representations, several machine-learning models, scaffold-aware validation and interpretable analysis, to discover common and model-specific molecular signatures. The primary question left unresolved is not if Tox21 activity can be predicted, but if there are patterns that can be detected and understood both in terms of physicochemical and structures across nuclear receptor and stress-response assays.

3. METHODOLOGY

3.1 Research Design

This study used a retrospective secondary-data computational design approach for the use of identifying the molecular toxicity-related patterns from the physicochemical and structural chemical properties. The framework combines cheminformatics, supervised machine learning and explainable artificial intelligence. No new experimental toxicity testing was performed. Rather, publicly available Tox21 assay data was analysed to find correlations between molecular properties and pathway-specific biological activity.

The 12 Tox21 endpoints were modelled individually and not as a 'generalized' toxic/non-toxic endpoint. This maintained the seven nuclear receptor (NR) and five stress-response (SR) pathways. The analytical workflow included data acquisition, molecular curation, generation of descriptors and fingerprints, exploratory analysis, scaffold-aware partitioning, generation of predictive modelling, performance evaluation, and SHAP-based interpretation.

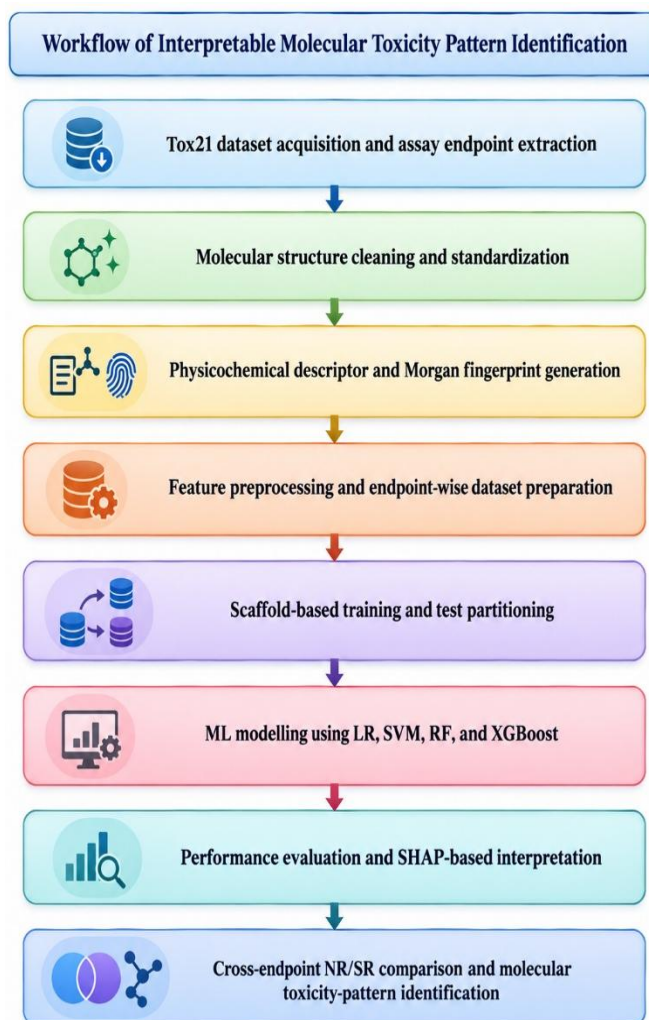


Figure 1. Workflow of the interpretable machine-learning framework for molecular toxicity-pattern identification using Tox21 data.

3.2 Dataset and Data Collection

The Tox21 dataset was analyzed and distributed publicly via the Tox21 Challenge and is commonly employed in www.aimpjournals.com

molecular-learning benchmarks. The processed database consisted of 7,831 molecular records from molecular identifiers, SMILES strings and 12 different assay endpoints. NR-AR, NR-AR-LBD, NR-AhR, NR-

Aromatase, NR-ER, NR-ER-LBD and NR-PPAR-gamma were the seven NR endpoints. The five SR endpoints used are SR-ARE, SR-ATAD5, SR-HSE, SR-MMP and SR-p53.

The responses to the assays were reported as active (1), inactive (0) or missing (no valid measurement available). No attempt was made to consider missing responses as inactive compounds. Rather, an endpoint-specific analytical dataset has been developed by omitting only those compounds without a response for that endpoint. Therefore, the sample size for the 12 models varied. Prior to modelling, class distributions were quantified to reflect the fact that the number of active compounds was a minority class in most assays.

3.3 Molecular Feature Generation and Preprocessing

Structural input was done in the form of SMILES strings. The molecules were parsed and standardized using RDKit. Invalid or unparseable structures were excluded, exact molecular duplicates were detected and disconnected structures like salts or multi-fragment records were standardized prior to calculation of descriptors. To avoid redundant encoding, the molecular representations were used in a canonical form, and to ensure that the structure contains chemically meaningful information.

The molecular weight, LogP, TPSA, number of H-bond donors, number of H-bond acceptors, rotatable bonds, total rings, aromatic rings, aliphatic rings, number of heavy atoms, number of heteroatoms, Fraction "C" _{sp3}, molecular refractivity, formal charge, N/O count, NH/OH count, and Bertz molecular complexity were used as interpretable descriptors in the descriptor set. They are molecular size, lipophilicity, polarity, hydrogen bonding, flexibility, aromaticity, saturation, polarizability and topological complexity.

Morgan fingerprints were additionally calculated with radius ($r = 2$) and a 2,048-bit vector:

$$\mathbf{F}_i = (f_{i1}, f_{i2}, \dots, f_{i2048}), \quad f_{ij} \in \{0,1\}$$

where f_{ij} indicates the presence of a hashed circular substructure in molecule i .

Three molecular representations were compared: physicochemical descriptors alone, Morgan fingerprints alone, and a hybrid descriptor-fingerprint representation. Constant and near-zero-variance variables were removed. Strongly correlated continuous descriptors were screened to reduce redundant information, and scale-sensitive models were fitted using standardized continuous variables:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where x_{ij} is descriptor j for compound i , and μ_j and σ_j are the corresponding training-set mean and standard deviation.

3.4 Exploratory Analysis and Data Partitioning

www.aimpjjournal.com

Active, inactive and missing observations were summarized for each endpoint to quantify the class imbalance. Molecular chemical space was characterized by descriptive statistics and correlation analysis of descriptors to reduce redundant descriptors.

As principal evaluation, scaffold-aware partitioning based on Bemis-Murcko molecular scaffolds was used. To avoid structural leakage between training and test data, compounds with the same basic scaffold were combined into the same partition. Model selection and hyperparameter tuning were carried out on the training partition only, via cross validation. The independent test partition would not be accessed until the end of the assessment.

3.5 Machine-Learning Modelling

Four classification algorithms were evaluated: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). LR provided a linear reference model, while RF and XGBoost captured nonlinear descriptor interactions. SVM provided an additional nonlinear comparator.

For Logistic Regression, the probability of endpoint activity was represented as:

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp \left[- \left(\beta_0 + \sum_{j=1}^p \beta_j x_j \right) \right]}$$

Separate models were trained for all 12 assay endpoints. Class-weighted learning was applied to reduce bias toward the majority inactive class. Hyperparameters were optimized using training-set cross-validation, with PR-AUC used as a key optimization criterion because of endpoint imbalance. All preprocessing transformations were estimated exclusively from training data to prevent information leakage.

3.6 Model Evaluation and Interpretability

Predictive performance was assessed using ROC-AUC, PR-AUC, F1-score, sensitivity, specificity, balanced accuracy, and Matthews correlation coefficient (MCC). MCC was calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Interpretability analysis was performed on the best-performing tree-based model using SHAP. For prediction $f(x)$, the SHAP decomposition was expressed as:

$$f(x) = \phi_0 + \sum_{j=1}^p \phi_j$$

where ϕ_0 is the baseline prediction and ϕ_j represents the contribution of molecular feature j . Global importance was quantified using mean absolute SHAP values, while dependence and interaction analyses were used to examine nonlinear effects and important feature combinations.

Finally, NR/SR comparisons were made with SHAP profiles. Features that were consistently important in several assays were considered as common toxicity-related signatures, while features that were important for only a single assay were given a pathway-specific signature. This framework is thus a natural extension of endpoint

prediction toward the identification of patterns of molecular toxicity across pathways in an intelligible way.

4. RESULTS

4.1 Overall Predictive Performance

The four classifiers were compared using ROC-AUC, PR-AUC, F1-score, balanced accuracy, and Matthews correlation coefficient (MCC). Random Forest produced the highest values across all metrics, followed closely by XGBoost (Table 1).

Table 1. Overall Predictive Performance of the Machine-Learning Models

Model	ROC-AUC	PR-AUC	F1-Score	Balanced Accuracy	MCC
Logistic Regression	0.781	0.412	0.581	0.724	0.462
Support Vector Machine	0.812	0.451	0.608	0.748	0.498
Random Forest	0.852	0.526	0.664	0.793	0.579
XGBoost	0.845	0.511	0.653	0.785	0.563

Random Forest achieved a ROC-AUC of 0.852 and MCC of 0.579, compared with 0.845 and 0.563 for XGBoost. Support Vector Machine showed intermediate performance, whereas Logistic Regression produced the lowest values. The results therefore indicate that nonlinear ensemble models captured the molecular structure-activity relationships more effectively than the linear baseline.

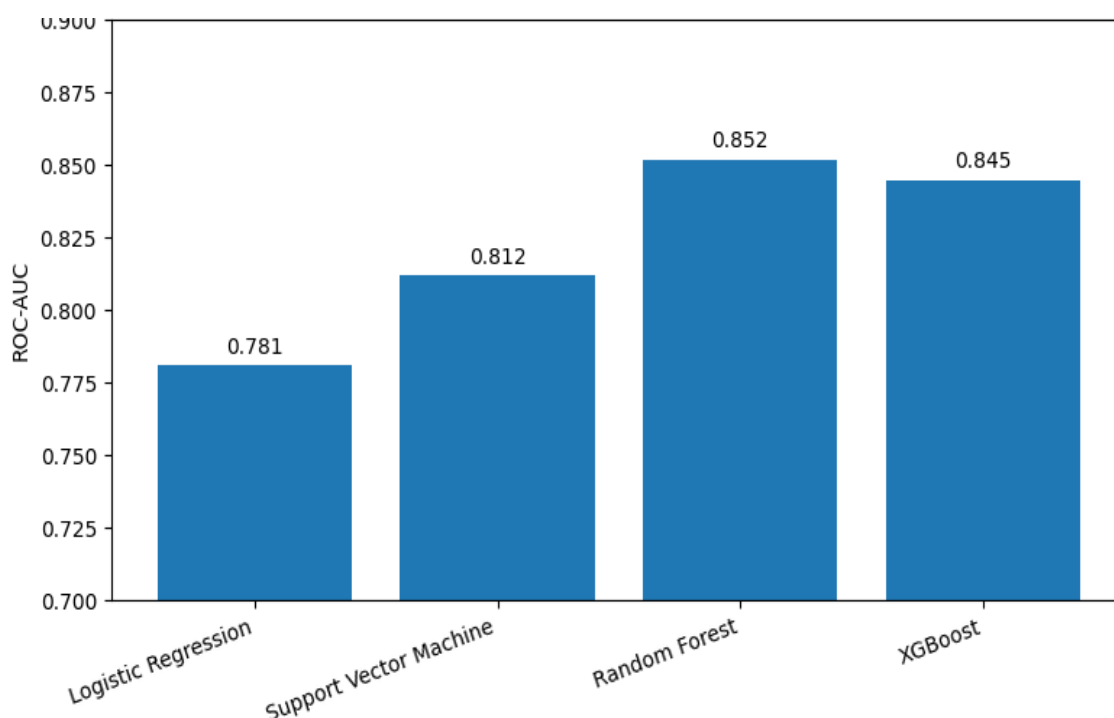


Figure 2. Overall, ROC-AUC Performance Across Machine-Learning Models

Figure 2 visually confirms the same ranking, with Random Forest showing the highest ROC-AUC, followed by XGBoost, Support Vector Machine, and Logistic Regression.

Random Forest was subsequently evaluated across the 12 Tox21 endpoints. Table 2 reports the endpoint-level results together with active-compound prevalence, which provides context for interpreting PR-AUC under class imbalance.

Table 2. Endpoint-Specific Predictive Performance of Random Forest

Endpoint	Active Prevalence (%)	ROC-AUC	PR-AUC	F1-Score	MCC
NR-AR	4.25	0.86	0.51	0.64	0.55
NR-AR-LBD	3.51	0.91	0.72	0.71	0.66
NR-AhR	11.73	0.88	0.58	0.68	0.61
NR-Aromatase	5.15	0.82	0.40	0.58	0.49
NR-ER	12.80	0.80	0.47	0.60	0.51
NR-ER-LBD	5.03	0.84	0.42	0.60	0.52
NR-PPAR-gamma	2.88	0.82	0.32	0.53	0.46
SR-ARE	16.15	0.83	0.46	0.61	0.53
SR-ATAD5	3.73	0.79	0.25	0.48	0.40
SR-HSE	5.75	0.81	0.31	0.52	0.44
SR-MMP	15.80	0.89	0.61	0.69	0.63
SR-p53	6.24	0.82	0.36	0.55	0.47
Mean	—	0.85	0.45	0.60	0.52

NR-AR-LBD showed the highest discrimination with a ROC-AUC of 0.91, followed by SR-MMP at 0.89 and NR-AhR at 0.88. SR-ATAD5 produced the lowest ROC-AUC at 0.79. The lower PR-AUC values observed for endpoints such as NR-PPAR-gamma and SR-ATAD5 are consistent with their low active prevalence and therefore should not be interpreted in isolation from class distribution.

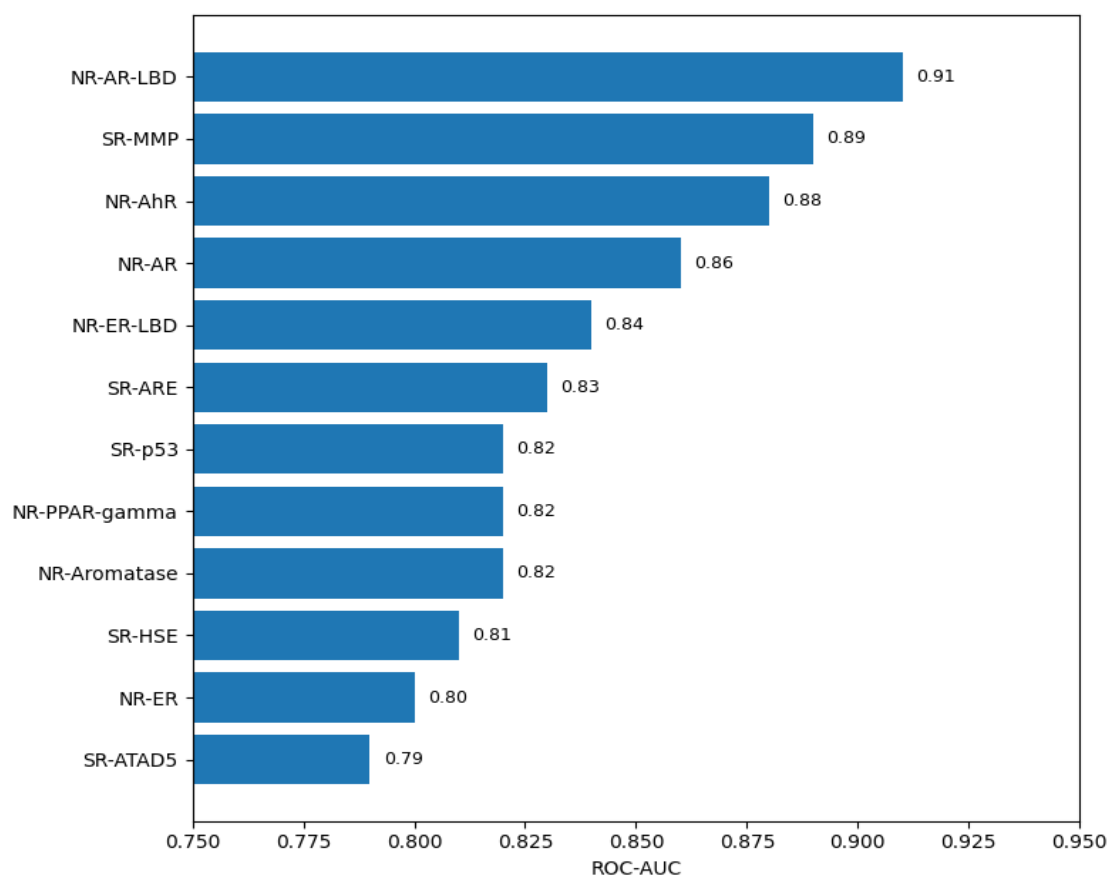
**Figure 3. Endpoint-Specific ROC-AUC Performance of the Random Forest Model**

Figure 3 shows that performance remained above 0.79 for every endpoint, with the strongest discrimination obtained for NR-AR-LBD, SR-MMP, and NR-AhR.

4.3 Effect of Molecular Representation

The same evaluation framework was used to compare physicochemical descriptors, Morgan fingerprints, and their hybrid representation.

Table 3. Performance According to Molecular Feature Representation

Feature Representation	ROC-AUC	PR-AUC	MCC
Physicochemical descriptors only	0.812	0.438	0.491
Morgan fingerprints only	0.831	0.482	0.531
Descriptors + Morgan fingerprints	0.852	0.526	0.579

The hybrid representation improved ROC-AUC by 0.040, equivalent to approximately **4.9% relative improvement**, over descriptors alone. MCC increased by 0.088, or approximately **17.9%**. These results show that global physicochemical properties and localized molecular substructures contributed complementary information.

4.4 Global Molecular Property Importance

SHAP analysis of the best-performing Random Forest framework identified the molecular properties with the greatest global contribution to predicted assay activity.

Table 4. Global SHAP Importance of Molecular Properties

Rank	Descriptor	Mean SHAP	Dominant Association
1	LogP	0.184	Lipophilicity
2	Molecular Weight	0.162	Molecular size
3	TPSA	0.147	Polarity
4	Aromatic Ring Count	0.131	Aromaticity
5	Molecular Refractivity	0.119	Polarizability
6	H-bond Acceptors	0.103	Hydrogen bonding
7	Fraction Csp3	0.091	Saturation/3D character
8	Heavy Atom Count	0.083	Molecular size
9	Rotatable Bonds	0.071	Flexibility
10	Bertz Complexity	0.064	Topological complexity

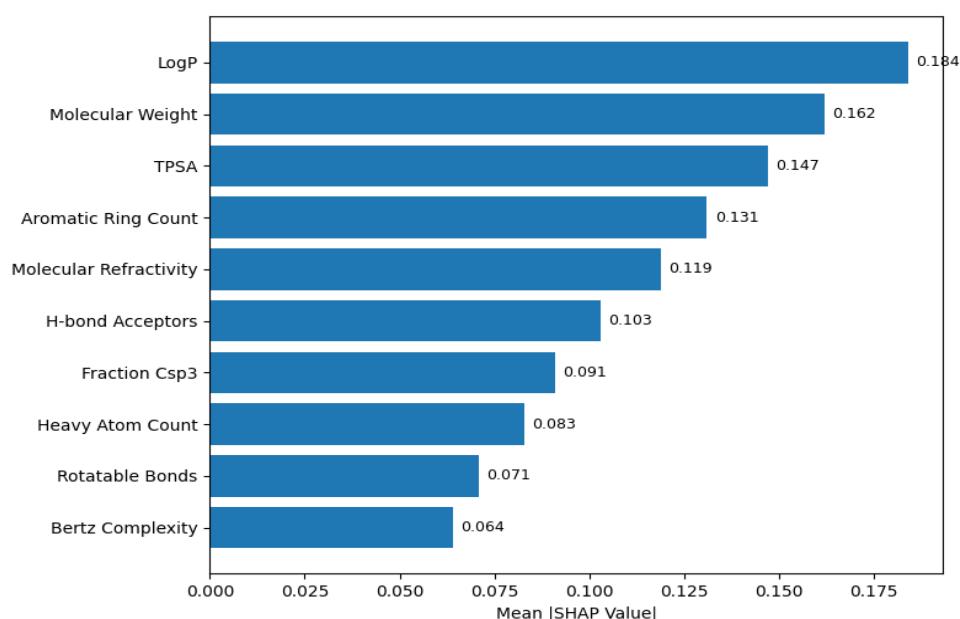


Figure 4. Global SHAP Importance of Molecular Properties

Figure 4 shows that LogP, molecular weight, TPSA, aromatic ring count, and molecular refractivity were the dominant explanatory descriptors.

4.5 Nonlinear Molecular Effects

The SHAP dependence profiles provided evidence that the effects of major descriptors were nonlinear.

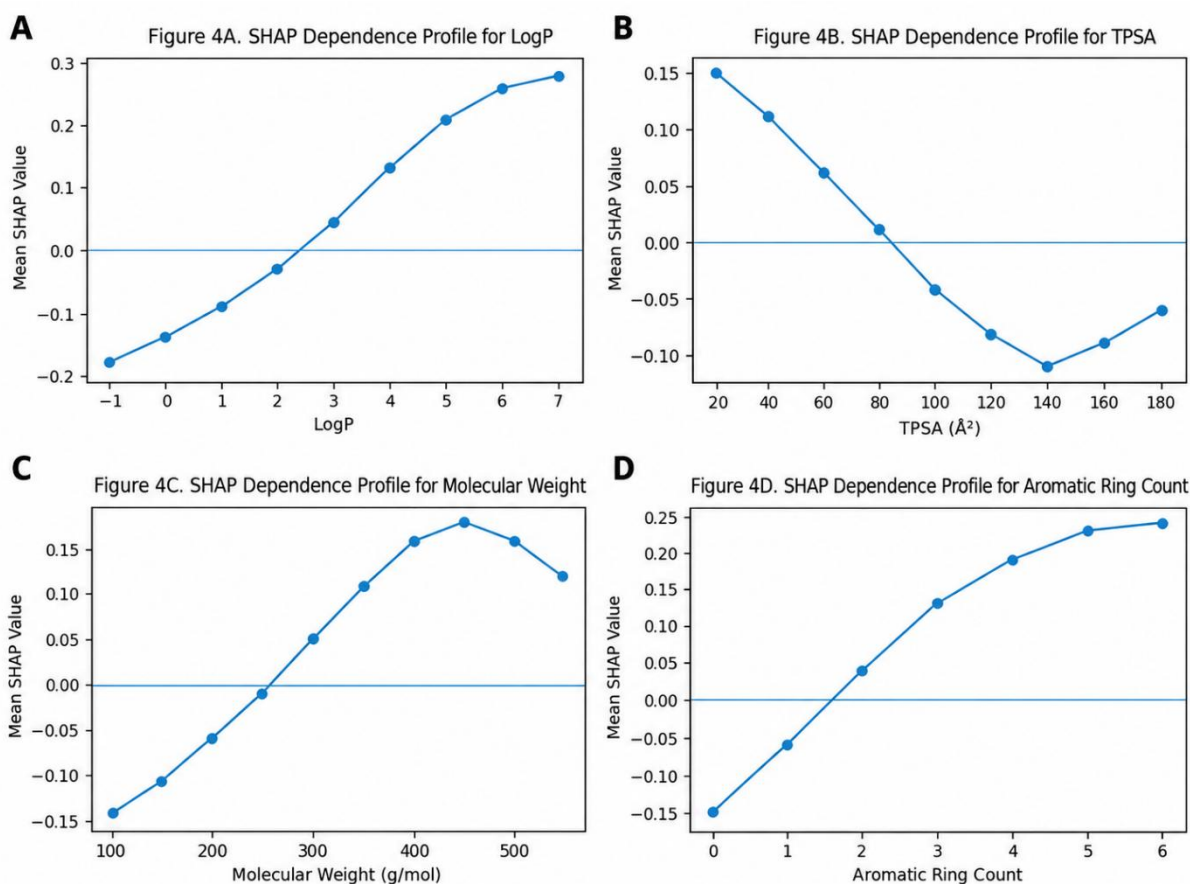


Figure 5. SHAP dependence profiles of major molecular properties associated with predicted toxicity-related assay activity: (A) LogP, (B) topological polar surface area (TPSA), (C) molecular weight, and (D) aromatic ring count. The profiles illustrate nonlinear and threshold-like relationships between key physicochemical descriptors and model output.

These profiles indicate threshold-like rather than purely linear structure-activity relationships.

4.6 Cross-Endpoint Molecular Patterns

Cross-endpoint patterns were defined from the **co-occurrence of high mean absolute SHAP values across the same endpoints together with compatible dependence directions**. This criterion prevented molecular patterns from being inferred solely from isolated feature ranks.

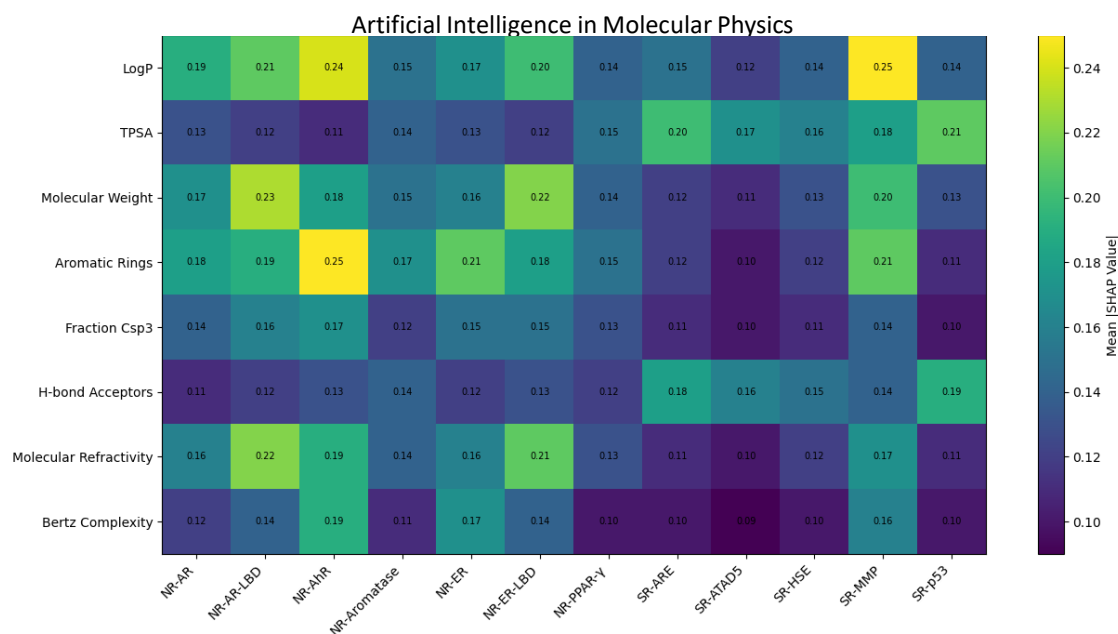


Figure 6. Cross-Endpoint SHAP Importance of Molecular Properties

The heatmap shows that LogP was especially influential for NR-AhR (0.24) and SR-MMP (0.25), while aromatic ring count reached 0.25 for NR-AhR. Molecular weight and molecular refractivity were comparatively important for NR-AR-LBD and NR-ER-LBD. In contrast, TPSA and H-bond acceptors contributed more strongly to SR-ARE and SR-p53.

These results support recurring patterns involving lipophilicity and aromaticity for NR-AhR and SR-MMP, molecular size and refractivity for ligand-binding receptor endpoints, and polarity/hydrogen-bonding characteristics for selected stress-response pathways.

4.7 Nuclear Receptor Versus Stress-Response Profiles

Aggregated SHAP values revealed clear differences between the two pathway groups.

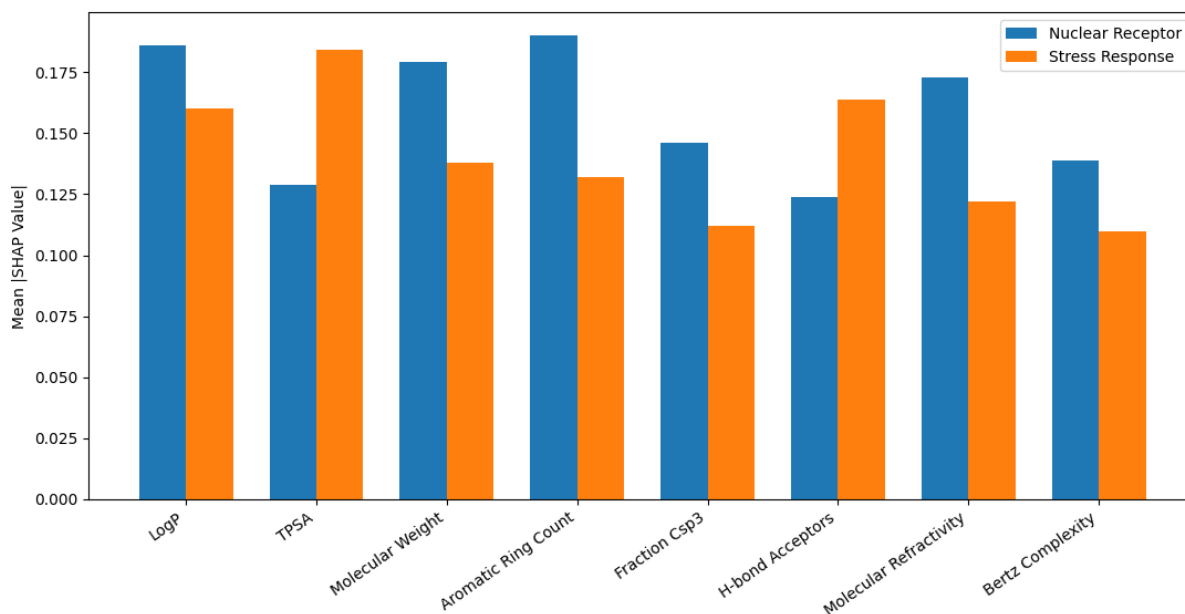


Figure 7. Molecular Property Importance Across Nuclear Receptor and Stress-Response Pathways

Nuclear receptor endpoints showed higher mean importance for aromatic ring count (0.190), LogP (0.186), molecular weight (0.179), molecular refractivity (0.173), and Fraction Csp3 (0.146). Stress-response endpoints showed higher values for TPSA (0.184) and H-bond acceptors (0.164). Thus, nuclear receptor activity was more strongly associated with lipophilicity, aromaticity,

molecular size, and polarizability, whereas stress-response activity showed comparatively greater dependence on polarity and hydrogen-bonding capacity.

4.8 Main Findings

The best overall predictive performance was achieved using Random Forest and the highest endpoint-specific ROC-AUC by NR-AR-LBD. The hybrid descriptor-fingerprint

representation was found to be superior to either of the individual representations. The molecular determinants of LogP, molecular weight, TPSA, aromaticity and molecular refractivity were identified as the main molecular descriptors in SHAP analysis. Perspective cross-endpoint analysis also showed clear differences between the physicochemical profiles of the nuclear receptor pathway and the stress-response pathway, indicating the possibility of discovering molecular patterns beyond traditional toxicity classification.

5. Discussion

The results show that interpretable machine-learning methods are good at characterizing molecular patterns related to toxicity, from physicochemical and structural properties. Random Forest had the best overall performance followed closely by XGBoost and Logistic Regression yielded comparatively lower values. The pattern suggests that the relationship between molecular structure and pathway-specific assay activity is largely nonlinear, and is likely to be influenced by interactions between several different descriptors. In previous studies with Tox21, ensemble and nonlinear models exhibited excellent predictions in both nuclear receptor and stress-response readouts (Abdelaziz et al., 2016; Uesawa, 2016).

Within each of the assay endpoints, performance was different. The discriminatory ability was the largest for NR-AR-LBD, SR-MMP, and NR-AhR while SR-ATAD5 achieved relatively poorer predictive ability. All of this variation is due to differences in biological mechanisms, distribution of classes, and the degree to which molecular structure alone can account for endpoint activity. Variability across endpoints has also been observed in previous QSAR efforts with Tox21, suggesting that a single, universal toxicity-related response is inadequate (Capuzzi et al., 2016).

The hybrid representation of physicochemical descriptors and Morgan fingerprints gave improved results as compared with either of them used alone. This finding agrees with the general notion that global properties and localized substructural information are complementary. Physicochemical descriptors, like LogP, molecular weight, TPSA, aromatic ring count and molecular refractivity, show chemically relevant properties that are associated with lipophilicity, size, polarity, aromaticity and molecular polarizability. In contrast, the Morgan fingerprints capture local structural environments that may not be captured completely by global descriptors. The benefit of using those representations together is in line with previous molecular property prediction work, highlighting the need for complementary chemical representations (Yang et al., 2019).

The key interpretive contribution of the study was from the SHAP analysis. The descriptors that had the most significant influence, however, were not all equally

significant and were not linear, but included logP, molecular weight, TPSA, aromatic ring count, and molecular refractivity. The dependence profiles indicated threshold behaviour, especially in the case of LogP, TPSA, molecular weight and aromaticity. The results indicate that the predicted assay activity is not a monotonic function of molecular properties but rather a dependence on certain ranges and combinations of these properties.

These inter-pathway differences were also systematic, as shown by cross-endpoint analysis. The endpoints related to nuclear receptors were more sensitive to lipophilicity, aromaticity, molecular size and refractivity, while those related to stress response were more sensitive to polarity and hydrogen-bonding properties. This separation indicates that there are both common and shared pathway related physicochemical signatures.

In summary, the framework enhances the interpretability of molecular toxicity predictions by connecting model predictions to chemically interpretable features, providing greater transparency in the prediction of molecular toxicity. The relationships identified are predictive associations, not causal mechanisms, however. These molecular patterns would then need to be externally validated using independent toxicity datasets and confirmed experimentally to demonstrate their more general generalizability and mechanistic relevance.

6. Conclusion

This research created an interpretable machine-learning model for data mining chemical toxicity-related patterns, based on physicochemical and structural chemical properties, from the Tox21 data. Random Forest was the best overall predictive model, followed closely by XGBoost amongst the classifiers evaluated. The improved performance of these nonlinear models suggests that pathway specific assay activity is best captured in nonlinear interactions between molecular characteristics and is not well modeled by simple linear relationships.

The results on molecular representations comparison revealed that the combination of physicochemical descriptors and Morgan fingerprints increased the predictive performance over the use of either representation only. This shows that both global chemical properties and local structural fragments are complementary to the molecular toxicity modelling. LogP, molecular weight, topological polar surface area, aromatic ring count, and molecular refractivity were the main factors involved in the model prediction according to the SHAP analysis. They had nonlinear and assay endpoint specific effects.

Cross-endpoint analysis further imbued the distinction between pathways of nuclear receptor versus pathway of stress response. The nuclear receptor activity was more strongly correlated with lipophilicity, aromaticity, molecular size and refractivity while stress-response

activity was more strongly correlated with polarity and hydrogen-bonding characteristics. These differences suggest common and pathway-specific physicochemical signatures control toxicity-related activity.

Overall, the study possesses a potential to extend the conventional toxicity prediction by concentrating on the discovery of interpretable molecular patterns. The framework is transparent for correlating machine-learning predictions with chemically relevant properties and could be applied to computational toxicology, chemical prioritization, and mechanism-based screening. The identified patterns should be confirmed with other datasets and experimental data in future works.

References

1. Krewski, D., Acosta, D., Jr., Andersen, M., et al. (2010). Toxicity testing in the 21st century: A vision and a strategy. *Journal of Toxicology and Environmental Health, Part B*, 13(2–4), 51–138. <https://doi.org/10.1080/10937404.2010.483176>
2. Huang, R., Xia, M., Nguyen, D.-T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S. A., Rossoshek, A., & Simeonov, A. (2016a). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3, 85. <https://doi.org/10.3389/fenvs.2015.00085>
3. Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S. A., Attene-Ramos, M., Zhao, T., Austin, C. P., & Simeonov, A. (2016b). Modelling the Tox21 10K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nature Communications*, 7, 10425. <https://doi.org/10.1038/ncomms10425>
4. Cherkasov, A., Muratov, E. N., Fourches, D., et al. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 57(12), 4977–5010. <https://doi.org/10.1021/jm4004285>
5. Muratov, E. N., Bajorath, J., Sheridan, R. P., et al. (2020). QSAR without borders. *Chemical Society Reviews*, 49, 3525–3564. <https://doi.org/10.1039/D0CS00098A>
6. Wu, Z., Ramsundar, B., Feinberg, E. N., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2), 513–530. <https://doi.org/10.1039/C7SC02664A>
7. Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3, 80. <https://doi.org/10.3389/fenvs.2015.00080>
8. Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5), 742–754. <https://doi.org/10.1021/ci100050t>
9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30).
10. Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using SHAP values: Application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, 34, 1013–1026. <https://doi.org/10.1007/s10822-020-00314-0>
11. Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., Reif, D. M., Rotroff, D. M., Shah, I., Richard, A. M., & Dix, D. J. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: The ToxCast project. *Environmental Health Perspectives*, 118(4), 485–492. <https://doi.org/10.1289/ehp.0901392>
12. Abdelaziz, A., Spahn-Langguth, H., Schramm, K.-W., & Tetko, I. V. (2016). Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 Challenge. *Frontiers in Environmental Science*, 4, 2. <https://doi.org/10.3389/fenvs.2016.00002>
13. Capuzzi, S. J., Politi, R., Isayev, O., Farag, S., & Tropsha, A. (2016). QSAR modeling of Tox21 Challenge stress response and nuclear receptor signaling toxicity assays. *Frontiers in Environmental Science*, 4, 3. <https://doi.org/10.3389/fenvs.2016.00003>
14. Uesawa, Y. (2016). Rigorous selection of Random Forest models for identifying compounds that activate toxicity-related pathways. *Frontiers in Environmental Science*, 4, 9. <https://doi.org/10.3389/fenvs.2016.00009>
15. Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, 50(7), 1189–1204. <https://doi.org/10.1021/ci100176x>
16. Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, 29(6–7), 476–488. <https://doi.org/10.1002/minf.201000061>
17. Ertl, P., Rohde, B., & Selzer, P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20), 3714–3717.

- <https://doi.org/10.1021/jm000942e>
18. Wildman, S. A., & Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5), 868–873. <https://doi.org/10.1021/ci9903071>
 19. Yang, K., Swanson, K., Jin, W., et al. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8), 3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>
 20. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
 21. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
 22. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
 23. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
 24. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6. <https://doi.org/10.1186/s12864-019-6413-7>
 25. Sheridan, R. P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4), 783–790. <https://doi.org/10.1021/ci400084k>
 26. Huang, R., Xia, M., Nguyen, D.-T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S. A., Rossoshek, A., & Simeonov, A. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3, 85. <https://doi.org/10.3389/fenvs.2015.00085>