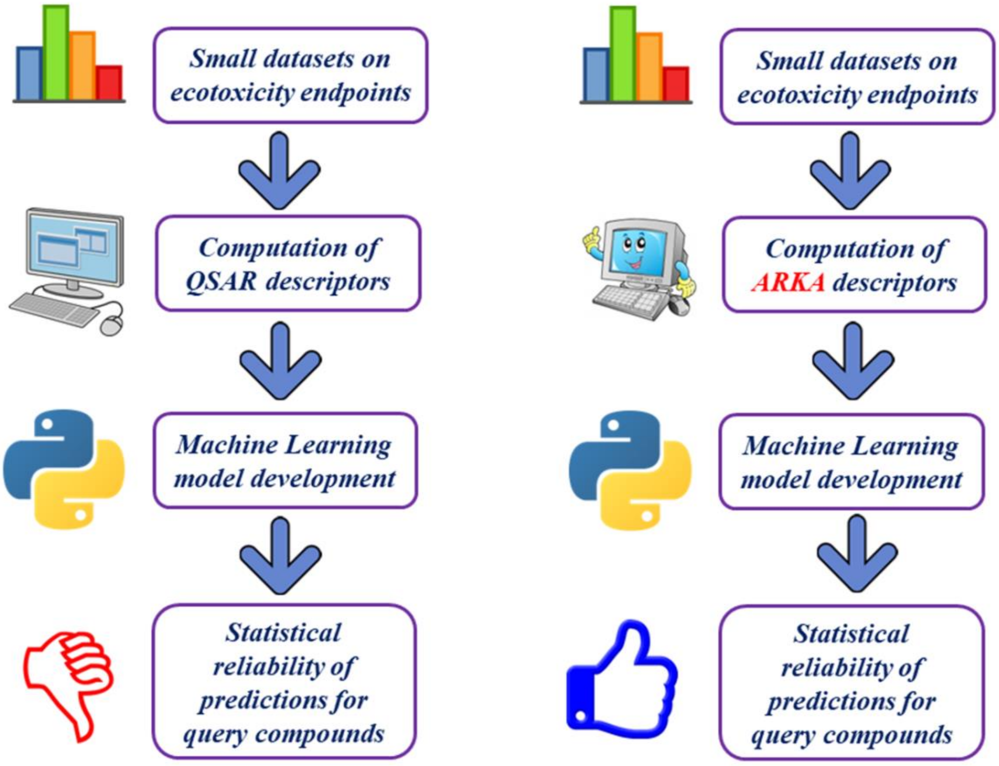
Preprint version dated 29.03.2024

ARKA: A framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data

Arkaprava Banerjee, Kunal Roy*

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

***Correspondence to: Kunal Roy (kunal.roy@jadavpuruniversity.in)**


Small datasets on ecotoxicity endpoints
Computation of QSAR descriptors
Machine Learning model development
Statistical reliability of predictions for query compounds
Small datasets on ecotoxicity endpoints
Computation of ARKA descriptors
Machine Learning model development
Statistical reliability of predictions for query compounds

Graphical Abstract

Abstract

Toxicity assessment of environmental chemicals is an integral aspect of assessing the sustainability of flora and fauna constituting the aquatic and terrestrial ecosystems. A wide variety of living organisms are constantly being exposed to these chemicals, most of which generate toxic effects. Due to the lack of experimental toxicity data of environmental chemicals, there arises a need to fill data gaps by *in silico* approaches. One of the most commonly used *in silico* approaches for toxicity assessment of small datasets is the Quantitative Structure-Activity Relationship (QSAR), which generates predictive models for the efficient prediction of query compounds. However, the predictions from these models are often erroneous for some compounds, and the reliability of the predictions from QSARs derived from small datasets is often questionable from a statistical point of view. This is due to the presence of a larger number of descriptors as compared to the number of training compounds, which reduces the degree of freedom of the developed model. To reduce the overall prediction error for a particular QSAR model, we have proposed here the computation of the novel Arithmetic Residuals in *K*-groups Analysis (ARKA) descriptors. We have reduced the number of modeling descriptors, keeping the entire chemical space and preventing the loss of chemical information. We have used here five representative environmentally relevant endpoints (skin sensitization, earthworm toxicity, milk/plasma partitioning, algal toxicity, and rodent carcinogenicity of hazardous chemicals) with graded responses to which the ARKA framework was applied for classification modeling. On comparing the performance of the models generated using conventional QSAR descriptors and the ARKA descriptors, the prediction quality of the models derived from ARKA descriptors was found much better than the models derived from QSAR descriptors signifying the potential of ARKA descriptors in ecotoxicological classification modeling of small data sets. For the ease of users, a Java-based expert system has been developed that computes the ARKA descriptors from the input of QSAR descriptors.

Keywords: ARKA descriptors, Machine Learning, Activity cliffs, Modelability, K-groups analysis

Environmental significance

The experimental data of chemical hazards for hundreds of endpoints being very limited, chemical regulatory authorities accept model-derived data for data-gap filling. However, the available sparse data for several endpoints are insufficient to develop statistically meaningful models forcing the modelers to use limited chemical features for final model development compromising the applicability domain and wide usability of the models for predictions. The problem of small data set classification modeling of ecotoxicity endpoints is addressed here by introducing the concept of Arithmetic Residuals in *K*-groups Analysis (ARKA) as a novel method of dimensionality reduction which demonstrates enhanced external prediction quality compared to the corresponding quantitative structure-activity relationship (QSAR) models.

1 Introduction

A vast array of organic molecules found in the environment can potentially induce disruption in aquatic and terrestrial ecosystems [1]. This is brought about by different structural, physicochemical, and electronic properties of such molecules that enable them to exert toxic effects on different species of flora and fauna. A growing area of environmental research is to assess the ecotoxicological risk of these harmful chemicals using non-animal alternative approaches. A real threat to the entire biodiversity is the lack of experimental data on the toxicity profile of most of the chemicals existing in the environment that have resulted in a large data gap. Entry of such substances inside the living system, either directly or by the process of biomagnification, can cause a variety of adverse effects including the disruption of the endocrine system. Therefore, the identification of these unknown "hazardous materials" is of prime importance for their safe disposal which will reduce the disruption in the ecosystem. Although extensive research is going on for the experimental toxicity assessment of these substances, they are often time-consuming and economically less viable. The limited availability of experimental ecotoxicological data warrants the need to shift toward computational methods for the quick, easy, and accurate predictions of the endpoints concerned [2, 3]. This is in line with regulatory bodies like the Organisation for Economic Co-operation and Development (OECD) [4] which encourage the use of *in silico* approaches, thus reducing time, animal suffering, expenses, and manpower associated with animal experimentations [5]. Since *in silico* approaches generate quick and accurate results, they can efficiently be used for data gap-filling [6]. This approach is also acceptable to regulations like the European Union

Registration, Evaluation, Authorisation and Restriction of Chemicals (EU REACH) [7] which accepts data generated from non-animal approaches. Since *in silico* approaches have worldwide acceptability, they can be considered useful tools to fill ecotoxicity data gaps, thus enabling the identification of toxic substances and their corresponding toxicophores.

Among the *in silico* prediction approaches, the Quantitative Structure-Activity Relationship (QSAR) [8] has been one of the go-to methods for computational model development and predictions. In its basic form, QSAR generates a simple mathematical model that correlates various structural and physicochemical features with the endpoint of interest [8]. With the advancements in this field, researchers have identified that the features may not necessarily be linearly correlated to the response values, thus introducing the concept of a non-linear relationship. To accommodate such correlations, various Machine Learning (ML) modeling algorithms have been adopted that can effectively incorporate non-linear relationships [9,10]. These models are developed using a variety of algorithms that efficiently reflect the structure-activity relationship. However, the only drawback that can be associated with the application of different ML modeling algorithms is the lack of interpretability of the features since most of the ML models have a "black box", although the recent innovations have focused on the explainability of the ML models by introducing concepts like SHAP analysis [11] and Swiss knife [12]. The limitations associated with the QSAR approach are exemplified quite often in the case of small datasets. Due to the limited number of compounds, the models should ideally be developed using a lower number of descriptors to comply with the statistical requirements, but this reduces the reliability of prediction since it compromises encoding the proper chemical information. There are two ways to deal with it: the first is to adopt non-statistical approaches like Read-Across [13, 14], and the second is to adopt dimensionality reduction techniques to reduce the size of the descriptor matrix. While the former approach primarily does not reflect the quantitative contribution of the descriptors, the latter approach adheres to the QSAR methodology. In the recent past, several studies on the applications of various *in silico* methodologies have been reported to accurately predict various ecotoxicity endpoints. Hung and Gini used a deep learning-based Quantitative Structure-Activity Relationship (QSAR) modeling to predict the mutagenicity of diverse chemicals [15]. Chatterjee et al., performed quantitative Read-Across using small datasets of nanoparticles to predict their toxicity [16]. Banerjee and Roy integrated the concept of Read-Across into a statistical modeling framework and developed quantitative Read-Across Structure-Activity Relationship (q-RASAR) models to predict the androgen receptor binding affinity of environmental chemicals [17]. Srisongkram ensembled the predictions of Read-Across and Machine Learning-based QSAR to develop a stacked model for the prediction of skin cytotoxicity [18]. A simple computational workflow for the prediction of environmentally relevant endpoints may be important for regulatory purposes from the viewpoint of transparency and easy transferability of the models, as already reported in several previous studies [19, 20].

Small dataset modeling using the Quantitative Structure-Activity Relationship (QSAR) approach has been a very challenging job since a QSAR modeling data set needs to possess sufficient data points to perfectly train itself. To address this problem, different techniques like synthetic sample generation [21], double cross-validation [22], consensus predictions [23], etc., have been used in the literature. The deficiency of sufficient data points warrants the QSAR modeler to include a higher number of features (descriptors) to establish a linear relationship between the data points. In such cases, the statistical aspect is compromised as the ultimate aim of a modeler is to develop highly predictive models using a lower number of descriptors. Moreover, the application of a higher number of descriptors coupled with ML algorithms generally tends to generate overfitted models that may not perform well on an external set of data. On the flip side of the coin, using a lower number of descriptors may not be able to develop robust and effective models since there is a loss of chemical information associated with the reduction in the number of descriptors. This calls for the development of new techniques that use a lower number of descriptors (i.e. a lower degree of freedom) while retaining the chemical information. This represents a form of dimensionality reduction technique that reduces the size of the descriptor matrix, yet retains the chemical information. While dimensionality reduction techniques like Principal Component Analysis (PCA) [24] and Partial Least Squares (PLS) [25] are already in use, we have presented here a simple form of dimensionality reduction technique – the ARKA descriptors, that effectively encode the chemical information of various descriptors in a particular form of computationally derived descriptors using an (A)rithmetic (R)esiduals in *K*-groups (A)nalysis approach. While developing models using a higher number of descriptors covers a wider chemical space as compared to models developed using a lower number of descriptors, such derived descriptors can encode the complete chemical information into a limited number of descriptors, thus not compromising the applicability domain of the developed models. We apply the suggested ARKA descriptors to classification modeling of five small data sets of environmental context which were previously analyzed using linear discriminant analysis using conventional QSAR descriptors. We aim to examine the impact of using the novel descriptors on the external predictivity of the models. Additionally, we also apply different machine-learning-based classification modeling approaches for comparison purposes.

2 Materials and Methods

2.1 Collection of environmental toxicity datasets

To evaluate the performance of the proposed novel descriptors and to check their performance on external data sets, we have taken five different environmental toxicity datasets. We have judiciously taken five sample ecotoxicity datasets, containing a limited number of data points, for which previously reported classification-based QSAR models were already reported. The purpose of selecting such data is as follows:

1. These data sets contain a limited number of data points. We are aiming here to address the problems of classification modeling of smaller data sets starting with a relatively large pool of descriptors.
2. The availability of the already-reported QSAR modeling descriptors has helped the proper comparison of the conventional QSAR models with the models developed using ARKA descriptors.

Dataset 1 represents the graded skin sensitization data of diverse organic chemicals as reported by Banerjee and Roy [26]. Dataset 2 consists of graded data for the chemical toxicity of earthworms as reported by Roy et al. [27]. Dataset 3 represents the graded data for milk/plasma concentration ratios of drugs and environmental pollutants as reported by Kar and Roy [28]. Dataset 4 consists of graded data on chemical toxicity towards *Pseudokirchneriella subcapitata* as reported by Pramanik and Roy [29]. Dataset 5 reports the graded form of rodent carcinogenicity potency data from the work of Kar et al. [30].

2.2 The algorithm for the computation of the ARKA descriptors

Paola Gramatica in one of her works stated that QSAR modeling is not “Push a Button and Find a Correlation” [31]. This is the driving force for researchers of the modern era to develop newer approaches to generating more efficient predictive ability of the models. Observing the pictorial architecture of an Artificial Neural Network published in Roy et al., 2015 [8], we thought a concept could be developed by clustering the descriptors, assigning a suitable weight, and storing as a composite value in a single descriptor specific for each cluster, giving rise to a concept of dimensionality reduction. Since one of the key motifs for this work is to stress the aspect of the simplicity of the computational approach thus allowing the broader scientific community to easily adopt the suggested strategy, we have used the same division of the training and test sets as reported by the previous authors for the computation of ARKA descriptors making the comparison of their performance with the conventional descriptors an easy task. Please note that the objective of the current work is not to develop the best model for each endpoint, but to establish the usefulness of the proposed method of dimensionality reduction in the case of small data set modeling. The basic idea behind the computation of the ARKA descriptors is to group the conventional QSAR descriptors based on a predefined criterion and then assign weightage to each descriptor in each group. We decided to explore the predictive performance of ARKA descriptors initially in a classification QSAR modeling framework and thus selected the data sets having graded response data. Although it is possible to partition the features into *K*-groups, in the present work we have restricted the value of *K* to 2 (corresponding to positive and negative classes).

Since feature selection is an integral step that is performed on the training set compounds, it is implied that the authors of the source datasets have selected the features based on the training set compounds only. Therefore, from a statistical point of view, the calculations of the ARKA descriptors should ideally be based on the training set. The first step is to normalize the training set descriptors such that the range of values for each descriptor column is from 0 to 1. This was followed by the grouping of the active and inactive class data points. The computation of the mean values of a particular descriptor in both the active and inactive classes was performed, and their difference (positive class descriptor mean – negative class descriptor mean) and absolute difference were calculated. This is the methodology of the most discriminating feature selection technique or the molecular spectrum analysis [32, 33]. It is to be noted that in this work, we have not performed additional feature selection based on the absolute mean difference values as we have already used the selected features from the previous references, and we are only considering the difference and absolute difference in mean values. Conceptually, this must be clear that this operation should be done using the normalized (scaled between 0 and 1) training set descriptor values and not using the standardized training set since the basic idea behind normalization is to bring the values of each descriptor into a same range, which is essential for the computation and comparison of the mean differences.

After the computation of the mean difference and absolute difference values of the selected features, we have assigned the descriptors to two different clusters. Cluster 1 consists of descriptors having positive difference values while cluster 2 consists of descriptors containing negative difference values. It is to be noted that defining the number of clusters depends on the modeler, but in this work, we have adhered to simplicity and uniformity and defined two clusters for all the analyses on different datasets. Once the cluster membership has been defined, it is now essential to assign weightage to each descriptor of a particular cluster. A simple weighting strategy was adopted that defines the weightage of a particular descriptor of a cluster, which has been represented in **Equation 1**.

$$Weightage\ of\ a\ descriptor = \frac{Difference\ value\ of\ a\ descriptor}{\sum Difference\ values\ of\ all\ the\ descriptors\ in\ the\ particular\ cluster}$$

(1)

Once all the descriptors in the two different clusters have been assigned the corresponding weightage, the computation of the Arithmetic Residuals in *K*-Groups Analysis (ARKA) descriptors can be easily done. The selected QSAR descriptors of the training and test sets were standardized using the Java-based tool Scale1.0 available from the DTC Lab Supplementary Website [34]. In each of the standardized training and test data sets, the descriptor ARKA_1 encodes the information for the descriptors in Cluster 1 (i.e. descriptors having positive difference values) and ARKA_2 encodes the information for the descriptors in Cluster 2 (i.e. descriptors having negative difference values). Both ARKA_1 and ARKA_2 were calculated as the weighted sum of the descriptors in their respective clusters. Considering a total of 5 contributing descriptors for a particular response, suppose descriptors x_1, x_2, and x_3 have positive difference values and are members of cluster 1, while descriptors x_4 and x_5 have negative difference values and are members of cluster 2; the corresponding mathematical expressions for the computation of the ARKA descriptors have been represented in **Equations 2 and 3**.

$$ARKA\_1 = w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3 \quad (2)$$

$$ARKA\_2 = w_4 \times x_4 + w_5 \times x_5 \quad (3)$$

In Equations 2 and 3, the terms $x_{1,\ldots\ldots}$, x_5 represent the descriptor values while $w_{1,\ldots.}$, w_5 represent the corresponding weightage values. On generalizing the formulae, the computation of ARKA descriptors has been represented in **Equation 4** where "n" represents the number of descriptors in a particular cluster.

$$ARKA\_X = \sum_{i=1}^{i=n} (w_i \times x_i) \quad (4)$$

On application of the above-mentioned concept, the computation of the ARKA descriptors for the training and test sets was performed. The workflow for the computation of ARKA descriptors has been presented in **Figure 1**

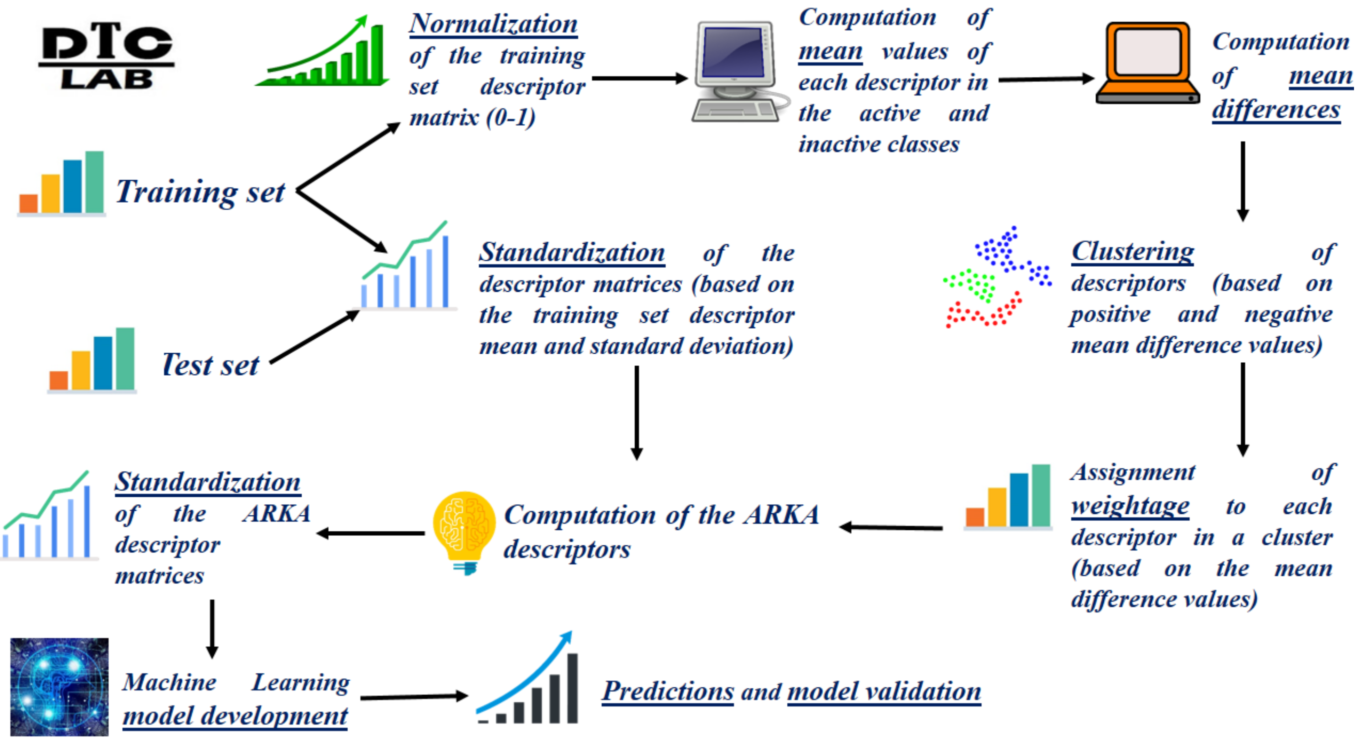

DTC LAB
Normalization of the training set descriptor matrix (0-1)
Computation of mean values of each descriptor in the active and inactive classes
Computation of mean differences
Training set
Test set
Standardization of the descriptor matrices (based on the training set descriptor mean and standard deviation)
Clustering of descriptors (based on positive and negative mean difference values)
Standardization of the ARKA descriptor matrices
Computation of the ARKA descriptors
Assignment of weightage to each descriptor in a cluster (based on the mean difference values)
Machine Learning model development
Predictions and model validation

Figure 1. Workflow for the computation of ARKA descriptors and model development.

2.3 Model development and validation

Initially, simple Linear Discriminant Analysis (LDA) models were developed separately using conventional QSAR descriptors and the ARKA descriptors by the Python-based Scikit-learn library [35] in Jupyter Notebook platform [36]. Twenty times fivefold cross-validation was performed to check the robustness of the developed LDA models. External and internal validation tests of the models were done using various standard classification-based validation metrics like Accuracy, Precision, Recall, F1 score, Matthews Correlation Coefficient (MCC), Cohen's kappa (Ckappa), and the Area Under Curve (AUC) of the receiver operating characteristic (ROC). Additional Machine Learning (ML) models like Logistic Regression (LR) [37], Support Vector Machine Classifier (SVM) [38], and Random Forest Classifier (RF) [39] were also then attempted using selected features and corresponding ARKA descriptors. In each case, the hyperparameters were optimized using a GridSearchCV approach adhering to a fivefold cross-validation strategy. The performance of these ML models was evaluated by the classification-based validation metrics stated above. The comparison of various models using standard QSAR descriptors and ARKA descriptors was based on the f1_score, MCC, Ckappa, and AUC of the test set data to compare the predictive performances. These metrics most effectively reflect the prediction performance of the developed models for both the active and inactive classes. The MCC is utilized as a measure of the quality of binary classifications. It considers true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes have different sizes. Cohen's kappa coefficient is more informative than Accuracy when working with imbalanced data [40]. However, we have additionally reported the prediction accuracy of the ARKA models (for both the training and test sets) in the Supplementary Material (*vide infra*). Among these different validation metrics, AUC can be deemed to be the most important metric as it reflects the complete classification scenario since it compares the true positive rate with the false positive rate [41]. Thus, an analysis of variance (ANOVA) [42] of the enhancement of AUC values in the ARKA models compared with corresponding QSAR models for five data sets has also been done.

2.4 Analysis of the Applicability Domain (AD) of the datasets

With every model being developed, there comes the necessity to evaluate the chemical space that the model encodes. This chemical space can be termed the applicability domain (AD), and it is believed that compounds lying outside this chemical space can generate unreliable predictions. As stated previously, the general concept is that the greater the number of descriptors, the larger the chemical space the model encodes. However, since the main focus of this work is to reduce the number of descriptors yet generate better predictive models, it is essential to analyze the AD status of the models that were developed. This calls for the computation of the Leverage values [43] for the ease of identification of the structural outliers, separately using the QSAR descriptor matrix and the ARKA descriptor matrix for the training and test sets, and then identify the number of structural outliers.

2.5 Development of the ARKA descriptor-calculating software

To make the computations more user-friendly, we have developed a Java-based ARKA descriptor calculating software: ARKA_descriptors-v1.0, which will be freely available from the DTC Lab tools supplementary website [34]. This tool takes input from the training and test set files and calculates the corresponding ARKA descriptors for the training and test sets.

The detailed modeling analysis is represented in **Figure 2**.

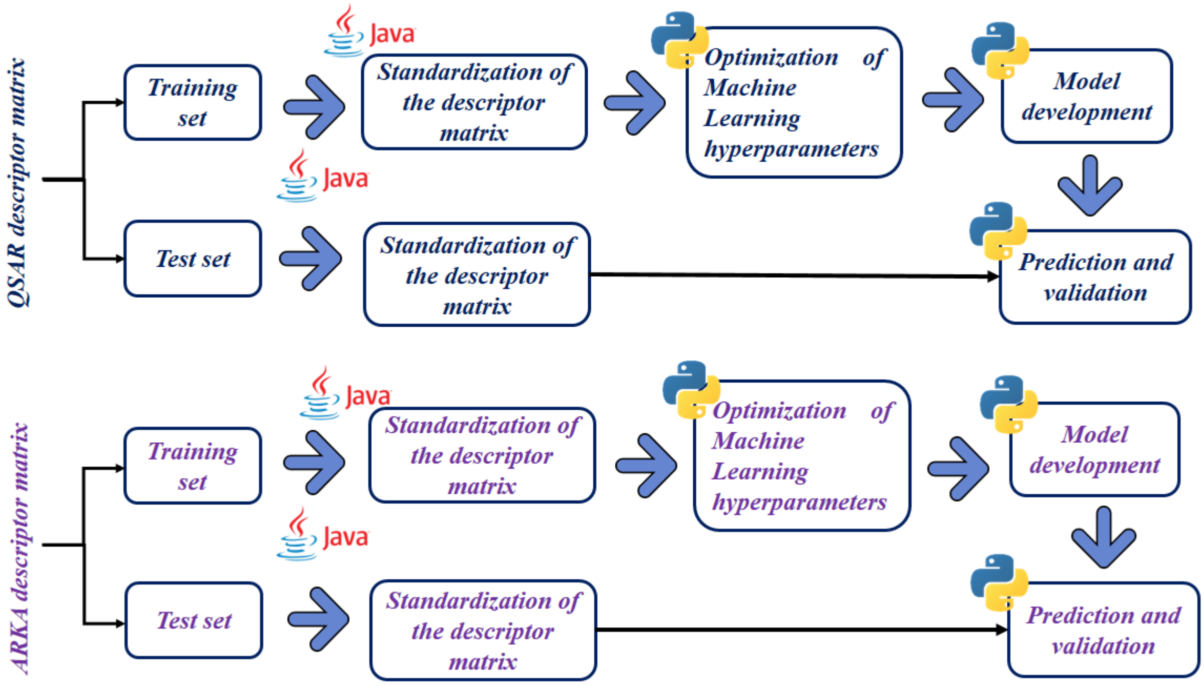

QSAR descriptor matrix
Training set
Test set
Java
Standardization of the descriptor matrix
Java
Standardization of the descriptor matrix
Optimization of Machine Learning hyperparameters
Model development
Prediction and validation
ARKA descriptor matrix
Training set
Test set
Java
Standardization of the descriptor matrix
Java
Standardization of the descriptor matrix
Optimization of Machine Learning hyperparameters
Model development
Prediction and validation

Figure 2. Detailed workflow for the modeling analysis.

2.6 Application of ARKA descriptors to chemical read-across analysis

Read-across is a similarity-based data gap-filling nonstatistical approach, which uses the endpoint information for one or more chemical(s) to predict the same endpoint for another similar chemical (based on structural features or mechanisms of action) [44]. The enhanced usage of read-across is promoted by regulatory frameworks to minimize new animal testing [45, 46]. We have applied the ARKA framework in deriving chemical read-across predictions of the endpoints considered and compared the quality of predictions using those derived from chemical read-across obtained using chemical descriptors. We have applied the Gaussian kernel–based similarity in the quantitative read-across algorithm of Chatterjee et al. [16].

3 Results and Discussion

We have used five representative ecotoxicity datasets for our experimental work to demonstrate the predictive ability of the models developed using ARKA descriptors. We selected these data sets considering that these are small to moderate-sized, and QSAR models were already developed for these models. This makes our comparison task easier. We have used the same set of features (QSAR descriptors) as used in the original analyses for comparison purposes. This concept can be used in various other datasets as well and can be extremely useful in enhancing the prediction quality of classification models developed from small data sets assessing human health risk.

3.1 Calculation of the ARKA descriptors

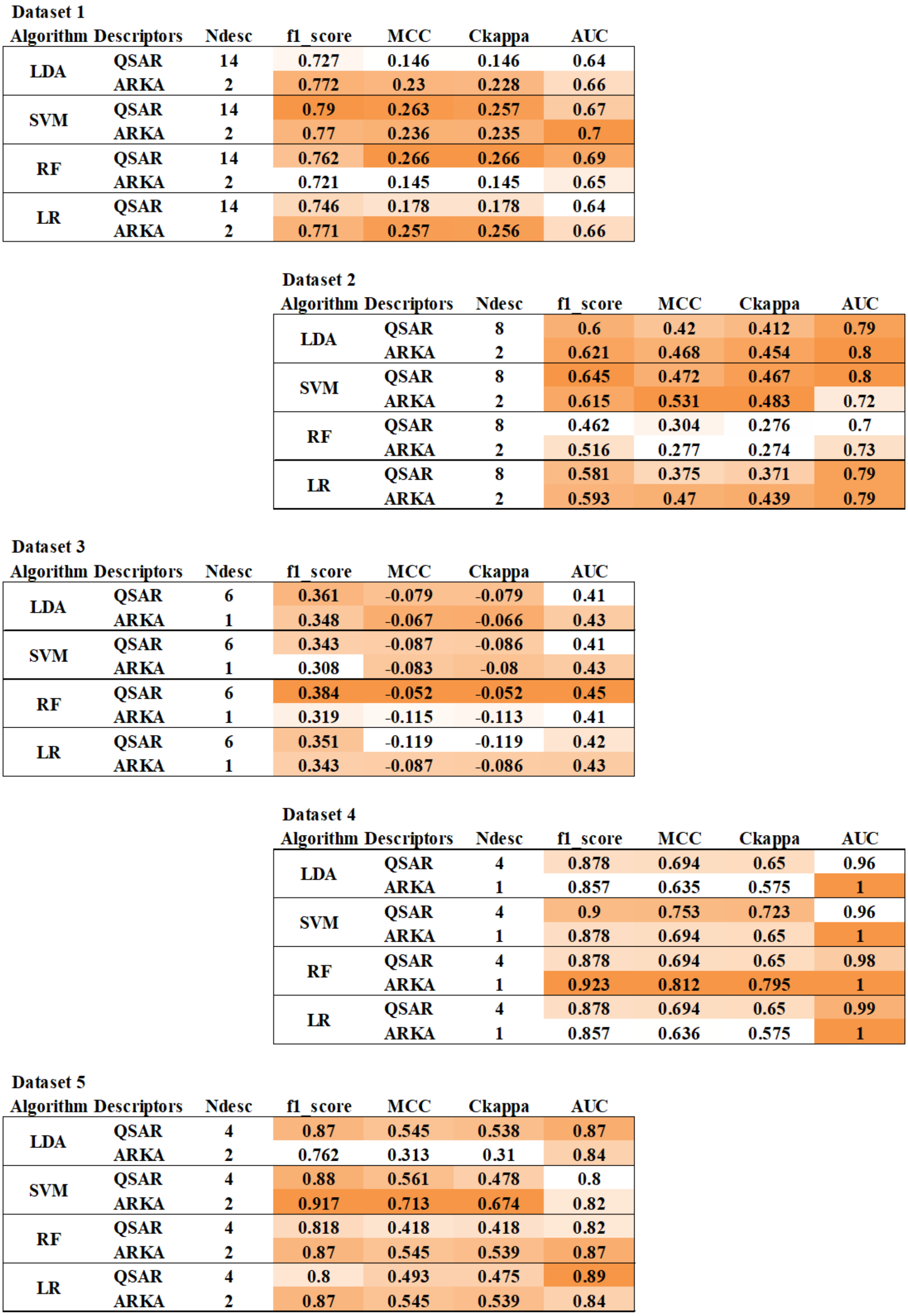
The five different ecotoxicity datasets used for the computation and analysis of the ARKA descriptors have been provided in Excel sheets of **Supplementary Materials SI-1**. The ARKA descriptors were calculated using the procedure mentioned in the Materials and Methods section. A representative example of the calculation of ARKA descriptors on Dataset 2 has been provided in **Supplementary Materials SI-2**. Note that each sheet of SI-1 contains specific calculations, and the final ARKA descriptors have been computed on the penultimate and the last sheets of the workbook.

3.2 Results of the Linear Discriminant Analysis (LDA) models

It may be noted that the objective of the present study has not been to report new predictive models for various ecotoxicological endpoints but to demonstrate the usefulness of the ARKA approach over the conventional QSAR modeling approach for external predictions of ecotoxicological data. Thus, we used the same set of QSAR descriptors and the same modeling strategy as reported in the previous studies for the ecotoxicological endpoints studied here. We like to emphasize that the performance of the ARKA descriptors will depend on the initial set of features selected for QSAR models. With a different set of selected QSAR features, the quality of ARKA models will accordingly vary. Thus, the performance of ARKA models should always be compared with the QSAR models developed with the corresponding conventional descriptors from which ARKA descriptors have been computed. The evaluation of the predictive performance of the models developed using conventional QSAR descriptors and the ARKA descriptors was initially checked using a linear modeling framework. The Linear Discriminant Analysis serves as a perfect example to correlate the performances of the models developed using conventional QSAR descriptors and the ARKA descriptors. As evident from **Figure 3**, it is observed that in most of the cases, the LDA models generated using ARKA descriptors showed enhanced predictive performance (test sets) in terms of the validation metric values than the LDA models generated using a much higher number of conventional QSAR descriptors.

3.3 Results of various Machine Learning (ML) models

Once we have compared the performance of the conventional QSAR descriptors with the ARKA descriptors based on a simple linear modeling framework, we have also compared their predictive performances when subjected to an ML modeling framework. For this, we have adopted the Logistic Regression, Support Vector Machine, and Random Forest classifiers to generate ML models using the conventional QSAR descriptors and the ARKA descriptors, separately. In this manuscript, we have aimed to address the current ecotoxicity data gap of “hazardous materials” found in the environment. Since experimental toxicity testing has certain limitations, adherence to *in silico* approaches for data gap-filling is encouraged in regulatory settings. However, from a statistical point-of-view, models developed on small datasets should ideally encode a lower number of descriptors, capture the entire chemical space, and be sufficiently predictive, resulting in the reliability of the developed models. Therefore, this work aims to reduce the number of modeling descriptors (dimensionality reduction) while preventing the loss of chemical information. The success of the models developed using the ARKA descriptors is evident from the enhanced prediction quality (for test set compounds) as compared to the corresponding conventional QSAR models, which is represented in **Figure 3**. Please note that while validating the models (QSAR or ARKA), the predictions for the test set compounds have been compared to the experimental observations while the models were developed solely from the training set compounds. Apart from the enhancement of the prediction quality metrics (like f1-score, MCC, Cohen's kappa, and AUC-ROC) in the majority of the cases, the use of ARKA descriptors was further supported by the reduced number of structural outliers as compared to the conventional QSAR descriptors, which proves that the entire chemical space has been preserved. From **Figure 3**, it is clear that in most cases, the predictive performance (test sets) of a particular ML model is better when the ARKA descriptors have been used and not when the conventional QSAR descriptors have been used. The optimized hyperparameter setting used to develop the ML models has been represented in **Table S1 in Supplementary Materials SI-3**. We have additionally reported now the prediction accuracy of the ARKA models (for both the training and test sets) in the **Supplementary Materials SI-3 (Table S2).** Although in some cases the statistical quality of models (training sets) is somewhat inferior (but well acceptable statistically) for ARKA models compared to QSAR models, we are interested here in exploring the enhancement of predictive performance of the models on the test sets using ARKA descriptors, and hence only test set statistics are reported here (**Figure 3**).



Dataset 1

Algorithm	Descriptors	Ndesc	f1_score	MCC	Ckappa	AUC
LDA	QSAR	14	0.727	0.146	0.146	0.64
	ARKA	2	0.772	0.23	0.228	0.66
SVM	QSAR	14	0.79	0.263	0.257	0.67
	ARKA	2	0.77	0.236	0.235	0.7
RF	QSAR	14	0.762	0.266	0.266	0.69
	ARKA	2	0.721	0.145	0.145	0.65
LR	QSAR	14	0.746	0.178	0.178	0.64
	ARKA	2	0.771	0.257	0.256	0.66

Dataset 2

Algorithm	Descriptors	Ndesc	f1_score	MCC	Ckappa	AUC
LDA	QSAR	8	0.6	0.42	0.412	0.79
	ARKA	2	0.621	0.468	0.454	0.8
SVM	QSAR	8	0.645	0.472	0.467	0.8
	ARKA	2	0.615	0.531	0.483	0.72
RF	QSAR	8	0.462	0.304	0.276	0.7
	ARKA	2	0.516	0.277	0.274	0.73
LR	QSAR	8	0.581	0.375	0.371	0.79
	ARKA	2	0.593	0.47	0.439	0.79

Dataset 3

Algorithm	Descriptors	Ndesc	f1_score	MCC	Ckappa	AUC
LDA	QSAR	6	0.361	-0.079	-0.079	0.41
	ARKA	1	0.348	-0.067	-0.066	0.43
SVM	QSAR	6	0.343	-0.087	-0.086	0.41
	ARKA	1	0.308	-0.083	-0.08	0.43
RF	QSAR	6	0.384	-0.052	-0.052	0.45
	ARKA	1	0.319	-0.115	-0.113	0.41
LR	QSAR	6	0.351	-0.119	-0.119	0.42
	ARKA	1	0.343	-0.087	-0.086	0.43

Dataset 4

Algorithm	Descriptors	Ndesc	f1_score	MCC	Ckappa	AUC
LDA	QSAR	4	0.878	0.694	0.65	0.96
	ARKA	1	0.857	0.635	0.575	1
SVM	QSAR	4	0.9	0.753	0.723	0.96
	ARKA	1	0.878	0.694	0.65	1
RF	QSAR	4	0.878	0.694	0.65	0.98
	ARKA	1	0.923	0.812	0.795	1
LR	QSAR	4	0.878	0.694	0.65	0.99
	ARKA	1	0.857	0.636	0.575	1

Dataset 5

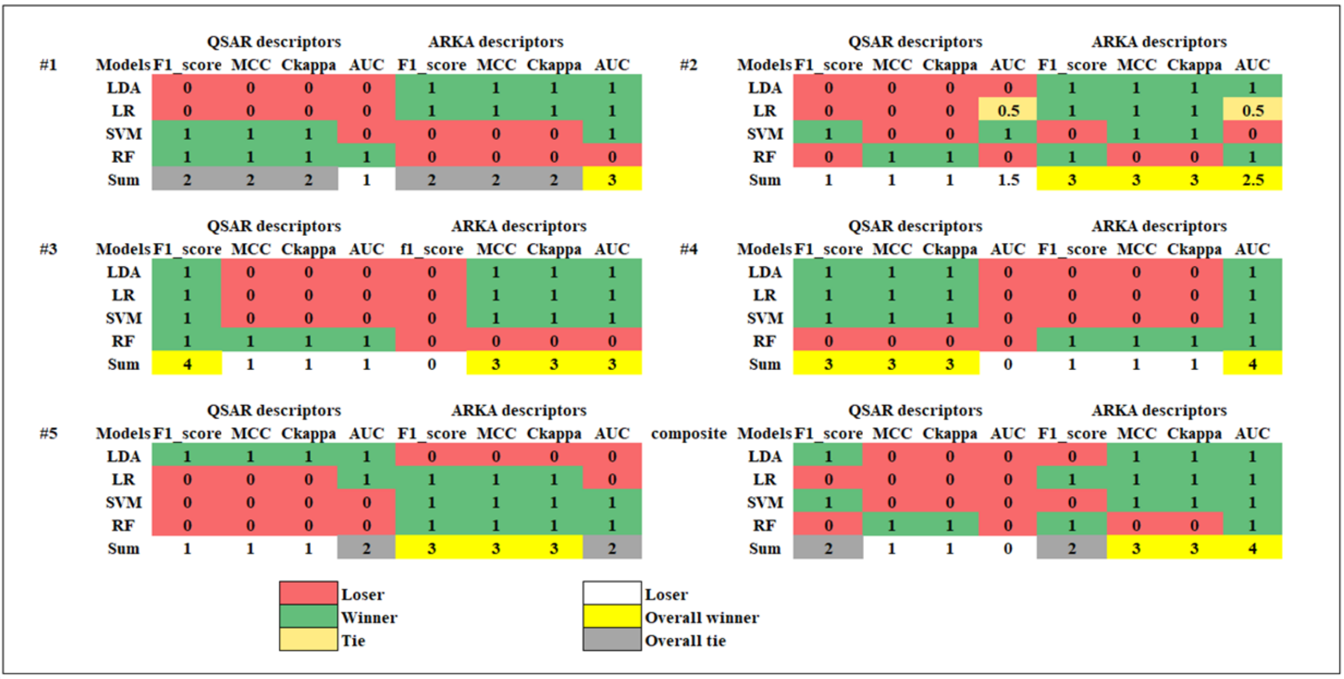
Algorithm	Descriptors	Ndesc	f1_score	MCC	Ckappa	AUC
LDA	QSAR	4	0.87	0.545	0.538	0.87
	ARKA	2	0.762	0.313	0.31	0.84
SVM	QSAR	4	0.88	0.561	0.478	0.8
	ARKA	2	0.917	0.713	0.674	0.82
RF	QSAR	4	0.818	0.418	0.418	0.82
	ARKA	2	0.87	0.545	0.539	0.87
LR	QSAR	4	0.8	0.493	0.475	0.89
	ARKA	2	0.87	0.545	0.539	0.84

Figure 3. Results of the external prediction quality (heatmap of quality metrics) of different models developed using conventional QSAR descriptors and ARKA descriptors (Ndesc = the number of descriptors).

3.4 Evaluation of the predictive performance of the models developed using conventional QSAR descriptors and ARKA descriptors

A comprehensive evaluation of the predictive performance of the different models using the two different classes of descriptors has been performed by a voting approach. In this approach, a modeling algorithm was taken one at a time and its predictive performance using the QSAR descriptors and ARKA descriptors was evaluated using four different essential external validation metrics, namely f1_score, MCC, Ckappa, and AUC. The purpose of considering these metrics is that they are capable of providing the complete picture of the classifiability of the models by considering the correct and incorrect predictions of the actives and the inactives in an unbiased manner. Moreover, these metrics especially AUC can handle imbalanced data and do not provide results that are biased to a particular class (actives or inactives). Moreover, AUC is a very suitable metric for comparing different models considering that this is derived from consideration of multiple threshold values [47]. In this study, for each endpoint, we have compared a model developed using conventional QSAR descriptors and another model using ARKA descriptors, using the same machine learning algorithm and the same dataset. The model with a higher value of a particular metric has been assigned a value of 1 for that metric. while the other model with a lower value of the metric has been assigned 0. If the values of a particular metric are equal for both models, an equal value of 0.5 each has been assigned. This was done for all the different external validation metrics stated above, using all the different modeling algorithms (LDA, LR, SVM, and RF), for all five different datasets as shown in **Figure 4**. A sum was calculated demonstrating the count of winner votes, specific for a particular validation metric, for both the QSAR and ARKA descriptors, and the validation metric with a higher count in QSAR or ARKA was considered the winner. An overall evaluation was done to get a comprehensive idea of the performance of the QSAR descriptors and the

ARKA descriptors. In this case (represented as “composite” in **Figure 4**), the count of voted winners for a particular validation metric in a particular modeling algorithm was taken into consideration, and the final voting is based on the sum of the voted winners among the QSAR and ARKA descriptors. For example, in the case of f1_score for the LDA model, we find that Datasets 1 and 2 have voted winners for the ARKA descriptors, while Datasets 3-5 have voted winners for the QSAR descriptors. Since a greater number of winners is observed using the QSAR descriptors, it received a composite vote of 1 while the ARKA descriptors received a composite vote 0. Similar to the individual datasets, a sum was computed to count the voted winners. From the final analysis in the composite set, it was found that the AUC, MCC, and Ckappa of the models derived from the ARKA descriptors were clear winners, i.e., these models showed enhanced performance with the ARKA descriptors, while the F1_score showed equal performance for both the QSAR and ARKA descriptors. The complete picture of this analysis has been represented in the form of a heat map in **Figure 4**.



		QSAR descriptors				ARKA descriptors			
#1	Models	F1_score	MCC	Ckappa	AUC	F1_score	MCC	Ckappa	AUC
	LDA	0	0	0	0	1	1	1	1
	LR	0	0	0	0	1	1	1	1
	SVM	1	1	1	0	0	0	0	1
	RF	1	1	1	1	0	0	0	0
	Sum	2	2	2	1	2	2	2	3

		QSAR descriptors				ARKA descriptors			
#2	Models	F1_score	MCC	Ckappa	AUC	F1_score	MCC	Ckappa	AUC
	LDA	0	0	0	0	1	1	1	1
	LR	0	0	0	0.5	1	1	1	0.5
	SVM	1	0	0	1	0	1	1	0
	RF	0	1	1	0	1	0	0	1
	Sum	1	1	1	1.5	3	3	3	2.5

		QSAR descriptors				ARKA descriptors			
#3	Models	F1_score	MCC	Ckappa	AUC	f1_score	MCC	Ckappa	AUC
	LDA	1	0	0	0	0	1	1	1
	LR	1	0	0	0	0	1	1	1
	SVM	1	0	0	0	0	1	1	1
	RF	1	1	1	1	0	0	0	0
	Sum	4	1	1	1	0	3	3	3

		QSAR descriptors				ARKA descriptors			
#4	Models	F1_score	MCC	Ckappa	AUC	F1_score	MCC	Ckappa	AUC
	LDA	1	1	1	0	0	0	0	1
	LR	1	1	1	0	0	0	0	1
	SVM	1	1	1	0	0	0	0	1
	RF	0	0	0	0	1	1	1	1
	Sum	3	3	3	0	1	1	1	4

		QSAR descriptors				ARKA descriptors			
#5	Models	F1_score	MCC	Ckappa	AUC	F1_score	MCC	Ckappa	AUC
	LDA	1	1	1	1	0	0	0	0
	LR	0	0	0	1	1	1	1	0
	SVM	0	0	0	0	1	1	1	1
	RF	0	0	0	0	1	1	1	1
	Sum	1	1	1	2	3	3	3	2

		QSAR descriptors				ARKA descriptors			
composite	Models	F1_score	MCC	Ckappa	AUC	F1_score	MCC	Ckappa	AUC
	LDA	1	0	0	0	0	1	1	1
	LR	0	0	0	0	1	1	1	1
	SVM	1	0	0	0	0	1	1	1
	RF	0	1	1	0	1	0	0	1
	Sum	2	1	1	0	2	3	3	4

🟥	Loser	⬜	Loser
🟩	Winner	🟨	Overall winner
🟨	Tie	⬛	Overall tie

Figure 4. Heat map demonstrating the voting results and how the models developed using ARKA descriptors showed enhanced predictive performance. 1 indicates a winner model for a particular metric, 0 indicates a loser model for a particular metric and 0.5 indicates a tie.

We have now additionally compared the enhancement of model quality for different data sets and different modeling algorithms taking AUC-ROC as the objective function and performed an analysis of variance of the change of AUC-ROC values due to the factors of data sets and modeling algorithms. The results show there has been indeed an enhancement of prediction quality on using ARKA descriptors in most of the cases **(Table 1).** The analysis of variance (ANOVA) results of the enhancement of AUC values in ARKA models compared with corresponding QSAR models for five data sets showed that the variations in the enhancement values were due to neither the structures of the datasets, nor the Machine Learning algorithms employed. This enhancement in the quality has been brought about by the ARKA descriptors, signifying their importance

Table 1. The enhancement of AUC-ROC (test set) due to the use of ARKA descriptors compared to QSAR descriptors

Data set	LDA	SVM	RF	LR	ANOVA results
1	**0.02**	**0.03**	-0.04	**0.02**	Row (data set) effect (insignificant):
2	**0.01**	-0.08	**0.03**	0	F1(df 4,12) = 0.14 (p = 0.965)
3	**0.02**	**0.02**	-0.04	**0.01**	Column (Modeling method) effect
4	**0.04**	**0.04**	**0.02**	**0.01**	(insignificant):
5	-0.03	**0.02**	**0.05**	-0.05	F2 (df 3,12) = 0.45 (p = 0.722)

3.5 Analysis of the Applicability Domain (AD) of the models developed from QSAR descriptors and ARKA descriptors

For the proper evaluation of a particular model, it becomes imperative to consider the applicability domain, i.e., the chemical space that the model encodes. The general idea is that a model generated using a larger number of descriptors encodes a wider chemical space as compared to a model generated using a lower number of descriptors. This is due to a loss of chemical information when the number of descriptors is low. However, in our novel approach, although the number of descriptors was reduced, this approach ensured that the entire chemical space was covered. However, for the complete picture and the identification of the number of structural outliers, we have computed the Leverages of the QSAR and ARKA descriptor matrices for all five different datasets. On analysis of the number of outliers, it was observed that the number of outliers computed using the ARKA descriptors was not only lower in all five different test sets but also lower in four of the five different training sets as well (**Table S3 in Supplementary materials SI-3**). This suggests that while modeling with the individual QSAR descriptors, there is some loss of information, but while computing ARKA descriptors, such information is retained.

3.6 Establishing a generalized relationship between ARKA_1 and ARKA_2 with the observed activity values – the identification of potential activity cliffs and less confident data points

As mentioned in the previous sections, the descriptors having a higher normalized mean value in the active class of compounds of the training set than in the inactive class are encoded into the ARKA_1 descriptor, and the other descriptors are encoded in ARKA_2. To estimate their relationships with the active and inactive class of compounds, we have generated violin plots representing ARKA_1 vs Activity and ARKA_2 vs Activity for both the training and test sets in all 5 different datasets (the results of Dataset 1 are given in **Figure 5** and the results of other data sets in **Figure S1** of Supplementary Materials **SI-3**.). As a general observation from **Figure 5** and **Figure S1**, the median values of the ARKA_1 descriptor in the active class are higher than the inactive class for both the training and test sets. Similarly, the median values of the ARKA_2 descriptor are higher in the inactive class than the active class for both the training and test sets, thus justifying the probable outcome. Also, from a representative example of Dataset 2 (**Figure S2 in supplementary Materials SI-3**), it is observed that in most of the active compounds, the values of ARKA_1 are positive and those of ARKA_2 are negative. Similarly, for most of the inactive compounds, the values of ARKA_1 are negative and those of ARKA_2 are positive. Similar observations may also be checked with other data sets (for example, the results from Dataset 1 are given in **Figures S3 and S4** of **Supplementary Materials SI-3**).

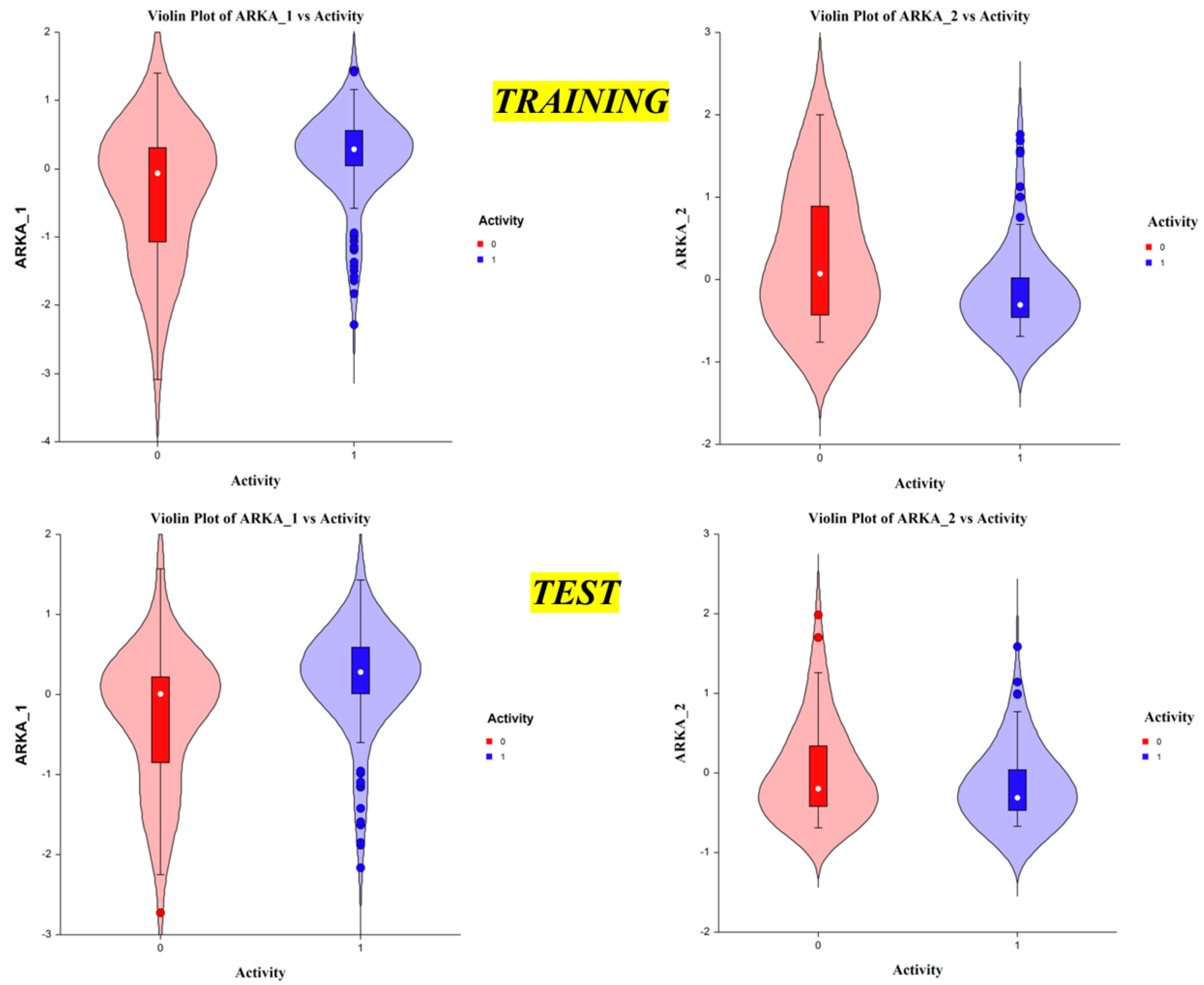


Figure 5. Violin Plots of ARKA_1 vs Activity and ARKA_2 vs Activity for the training and test sets (Dataset 1).

3.7 Dataset modelability from ARKA descriptors

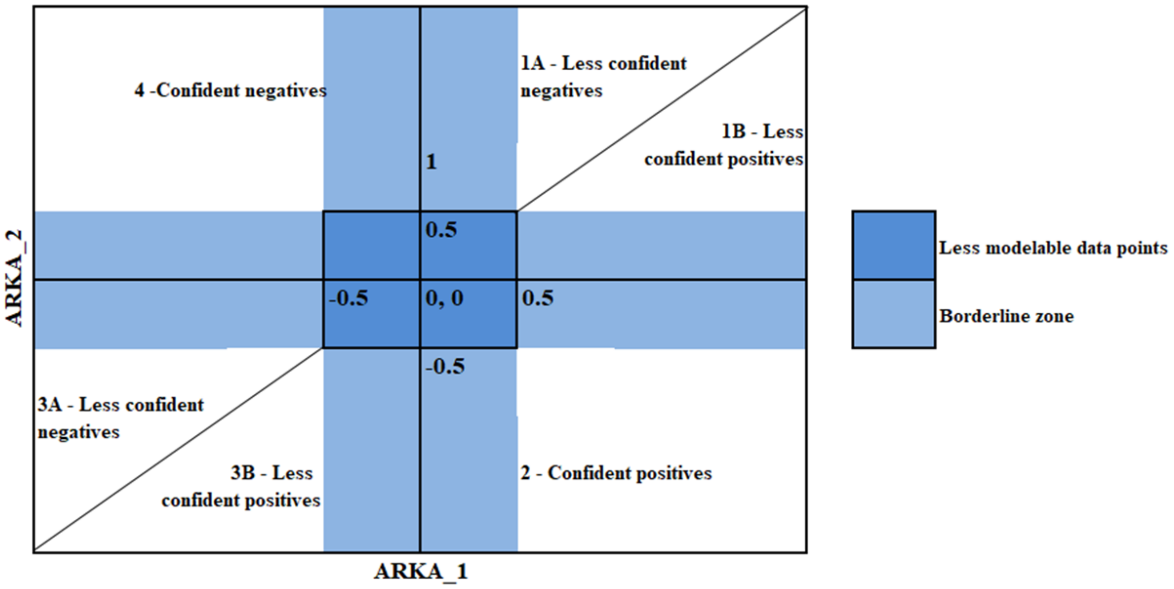

4 -Confident negatives
1A - Less confident negatives
1B - Less confident positives
1
0.5
-0.5
0, 0
0.5
-0.5
3A - Less confident negatives
3B - Less confident positives
2 - Confident positives
ARKA_1
ARKA_2
Less modelable data points
Borderline zone

Figure 6. Generalized interpretation to identify activity cliffs, borderline compounds, less modelable data points, and less confident data points based on their positions in the ARKA_2 vs ARKA_1 plot. Note that this interpretation is a "**model independent**" process, where simply the values of ARKA_1 and ARKA_2 are efficient enough to identify the nature of the data points.

A scatter plot of ARKA_2 vs. ARKA_1 descriptors for the training/test set compounds may indicate the modelability/model performance of the data set. Using knowledge from the data distribution of the violin plots, it can be inferred that the active compounds would most likely be present in the second quadrant of the scatter plots (where ARKA_1 is positive and ARKA_2 is negative), and the inactive compounds would most likely be present in the fourth quadrant (where ARKA_1 is negative and ARKA_2 is positive). Generally, based on the analysis of the studied data sets, for a confident classification of a data point into the positive and negative classes, it is found that the absolute values of both ARKA_1 and ARKA_2 descriptors should be more than 0.5 and their absolute difference is expected to be more than 0.75. Any point very near to the X or Y axis will have low decidability for classification purposes. As per **Figure 6**, the 2nd and 4th quadrants signify confident positives and negatives, respectively. If a negative (or inactive) compound is found in the 2nd quadrant or a positive (active) compound in the 4th quadrant, these may be potential activity cliffs reducing the modelability of the data set (in the case of the training set) or potential prediction cliffs indicative of poor prediction performance (in the case of the test set). Again, the compounds falling in the 1st and 3rd quadrants represent less confident data points for classifiability. Any misclassified compound falling in these quadrants as shown in **Figure 6** may be less confident activity/prediction cliffs.

We illustrate this (**Figure 7**) with Dataset 1 where the scatter plot of ARKA_2 vs. ARKA_1 shows five positive activity cliffs (compounds **260, 278, 320, 362,** and **370**) and one negative activity cliff (compound **93**) in the training set, and on comparing the definition of activity cliffs based on Banerjee-Roy similarity coefficients s_m^1 and s_m^2 (Banerjee and Roy, 2023), it was found that all the six compounds identified as activity cliffs from the training set, as evidenced from the plot of **Figure 7**, complying with both the approaches (i.e., ARKA descriptors and Banerjee-Roy similarity coefficients). In the case of the test set, one active compound (compound **349**) and two negative compounds (compounds **74** and **94**) were identified as the potential prediction cliffs (poor prediction performance potential). It is interesting that compound **74** was identified as a poor prediction potential data point also based on Banerjee-Roy similarity coefficients s_m^1 and s_m^2 (Banerjee and Roy, 2023). It is also observed that s_m^1 and s_m^2 identify a greater number of activity cliffs as compared to ARKA_1 and ARKA_2. This is because the s_m^1 and s_m^2 method detects additional activity cliffs that fall in either 1st or 3rd quadrants or near the axes thus showing less confidence in their activity cliff behavior since the ARKA criteria for determining activity cliffs are more strict than that of Banerjee-Roy coefficients. A similar observation was made in Dataset 2 (**Figure S5** of

Supplementary Materials SI-3). Thus, the ARKA descriptors may be used in detecting serious activity/prediction cliffs and understanding the modelability of a data set. This may be exercised even in case of regression modeling problems considering the whole data set response mean as the threshold and then classifying the response values as positives or negatives. For initial modelability analysis (when the contributing features are not known), one may proceed with computing the ARKA descriptors starting with the (whole) pretreated pool of descriptors while for the final model development, one should use only the selected features. In the case of Dataset 3, the scatter plot of ARKA_2 vs. ARKA_1 for the training set (**Figure S6 in Supplementary Materials SI-3**) shows poor modelability of the data set (the negative data points in the fourth quadrant are near the Y-axis) while the plot for the test set (**Figure S6**) shows poor quality of projections in the fourth quadrant suggesting poor information content of the QSAR descriptors for predictions of the endpoint (consistent with the poor model performance as per **Figure 3**). This is evident just from the ARKA descriptor scatter plot even without performing any modeling.

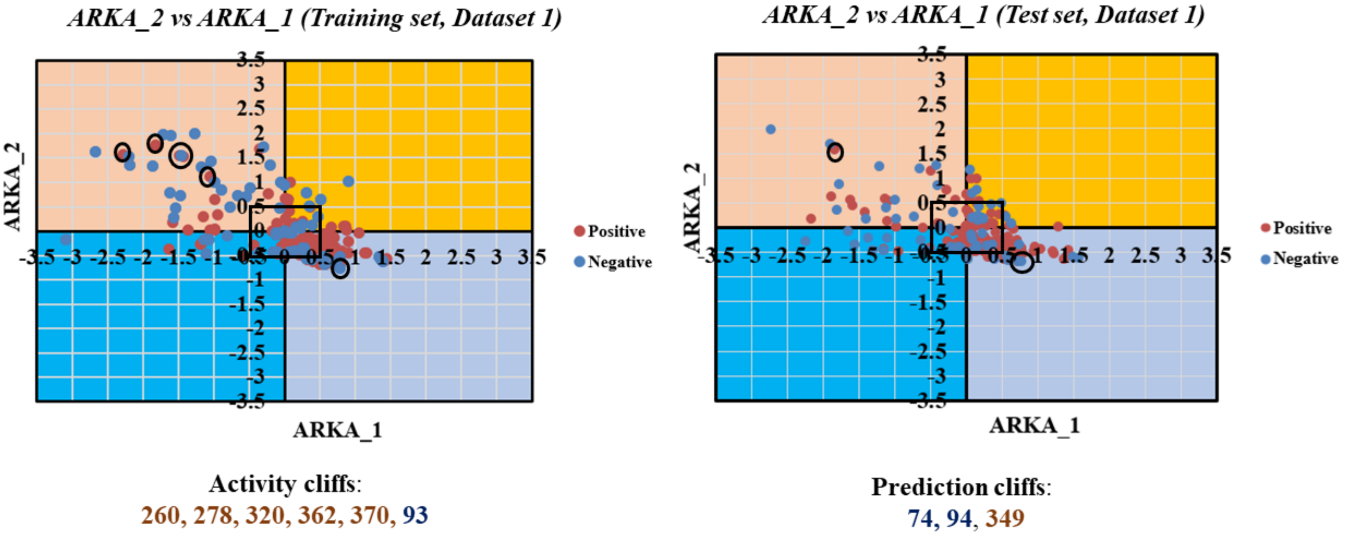


Figure 7. Representative scatter plots of ARKA_2 vs ARKA_1 for the training and test sets of Dataset 1.

3.8 Analysis of the conventional QSAR descriptors encoded in ARKA_1 and ARKA_2

This section of the manuscript will detail the various conventional QSAR descriptors that are encoded within the ARKA descriptors in the present study.

Case study 1: MDF analysis of descriptors contributing to skin sensitization potential of diverse organic chemicals

The dataset used for this study (Dataset 1) reports the skin sensitization data of harmful organic chemicals based on a local lymph node assay (LLNA) of murine species. The basic chemical theory behind skin sensitization is that the skin proteins act as nucleophiles and the sensitizers act as electrophiles [48]. As evident from our clustering analysis, two descriptors contributed to ARKA_1 while 12 descriptors contributed to ARKA_2. The descriptors contributing to ARKA_1 have a positive difference value signifying a higher mean value in the positive compounds, while the ones that contributed to ARKA_2 have a negative difference value signifying a higher mean value in the negative compounds. The descriptors gmin and minsssCH represent the minimum atom E-state value in a molecule and the minimum E-state value of a tertiary carbon atom, respectively. These descriptors have a near-equal contribution to ARKA_1, as evident from the weightage values (**Figure 9**). As these descriptors refer to the electronic environment of a molecule, they have a significant contribution towards electrophilic properties in the molecules, making them active skin sensitizers. On the other hand, the descriptors B06[C-N], B04[N-O], B10[C-O] depict the presence or absence of atom pairs C…N, N…O, and C…O at the topological distances of 6, 4, and 10 respectively, and they have higher contributions than most of the other descriptors contributing to ARKA_2. As evident from these 2D-atom pair descriptors, it is evident that they indicate the presence of heteroatoms rich in electrons, thereby reducing the electrophilic properties of the compounds **(Figure 8).**

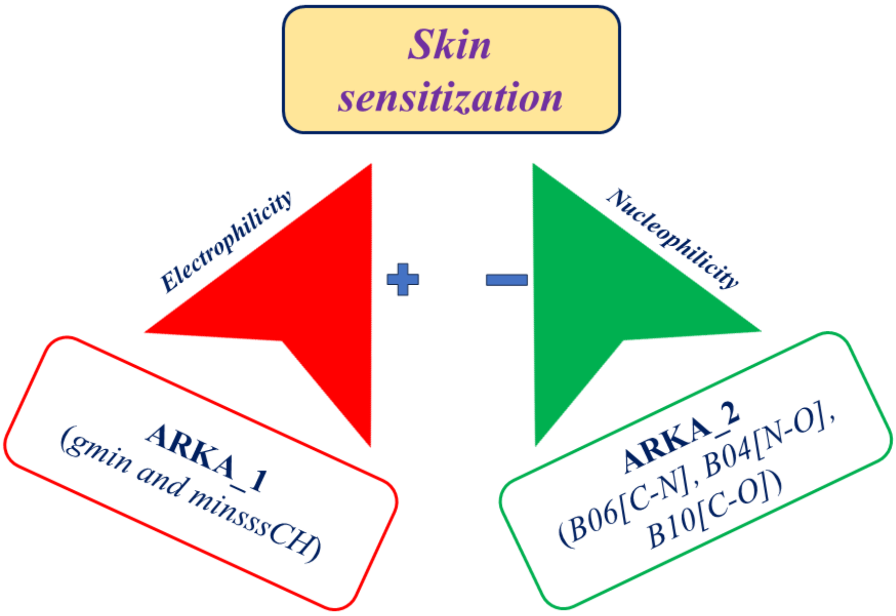


Figure 8. Analysis of the descriptors potentiating and inhibiting skin-sensitizing properties of a molecule.

Case study 2: MDF analysis of descriptors contributing to the chemical toxicity to earthworms

The dataset used for this study (Dataset 2) reports the chemical toxicity of diverse organic chemicals to earthworms. As evident from our clustering analysis, the descriptors B03[O-O] (presence or absence of O…O at the topological distance 3), ETA_Psi_1 (hydrogen bonding propensity and/or polar surface area), B02[O-S] (presence or absence of O…S at the topological distance 2) and X3A (lower degree of branching) have a positive difference value and contribute to ARKA_1. Similarly, descriptors like B09[C-C] (presence or absence of C…C at the topological distance 9), MLOGP^2 (squared Moriguchi o/w partition coefficient), ETA_Beta (depicting electron richness), and MLOGP (Moriguchi o/w partition coefficient) contribute to ARKA_2. Among the descriptors contributing to ARKA_1, it is observed that the descriptor B03[O-O] (higher electronegativity) has a significantly higher contribution than the others (**Figure 9**). Additionally, the descriptors ETA_Psi_1, and B02[O-S] have similar contributions and X3A has the least contribution. Similarly, among the descriptors contributing to ARKA_2, B09[C-C] (molecular size and hydrophobicity) contributes the maximum, $MLOGP^2$ and ETA_Beta have near-equal contributions, and MLOGP has the least contribution.

Case study 3: MDF analysis of descriptors contributing to the milk/plasma concentration ratios of drugs and environmental pollutants

The dataset used for this study (Dataset 3) reports the data for milk/plasma concentration ratios of drugs and environmental pollutants. The cluster analysis suggests that the descriptors ETA_EtaP_B_RC (indicating branching), nCrs (number of sp3 hybridized secondary carbons present in a ring), and S_tsC (electronic environment of an acetylenic carbon atom) contribute to the ARKA_1 descriptor (positive class). Similarly, the descriptors nAB (depicting the number of aromatic bonds present in the compound), nRCONHR (depicting the number of secondary aliphatic amides present in the compound), and Jurs-DPSA-1 (depicting the difference in the partial positive solvent-accessible surface area and the partial negative solvent-accessible surface area) contribute to the ARKA_2 descriptor (negative class). As evident from the contributions of the descriptors constituting ARKA_1, nCrs has the highest contribution while ETA_EtaP_B_RC has the lowest contribution (**Figure 9**). Similarly, in the case of ARKA_2, the descriptors nAB and nRCONHR have similar and highest contributions to ARKA_2.

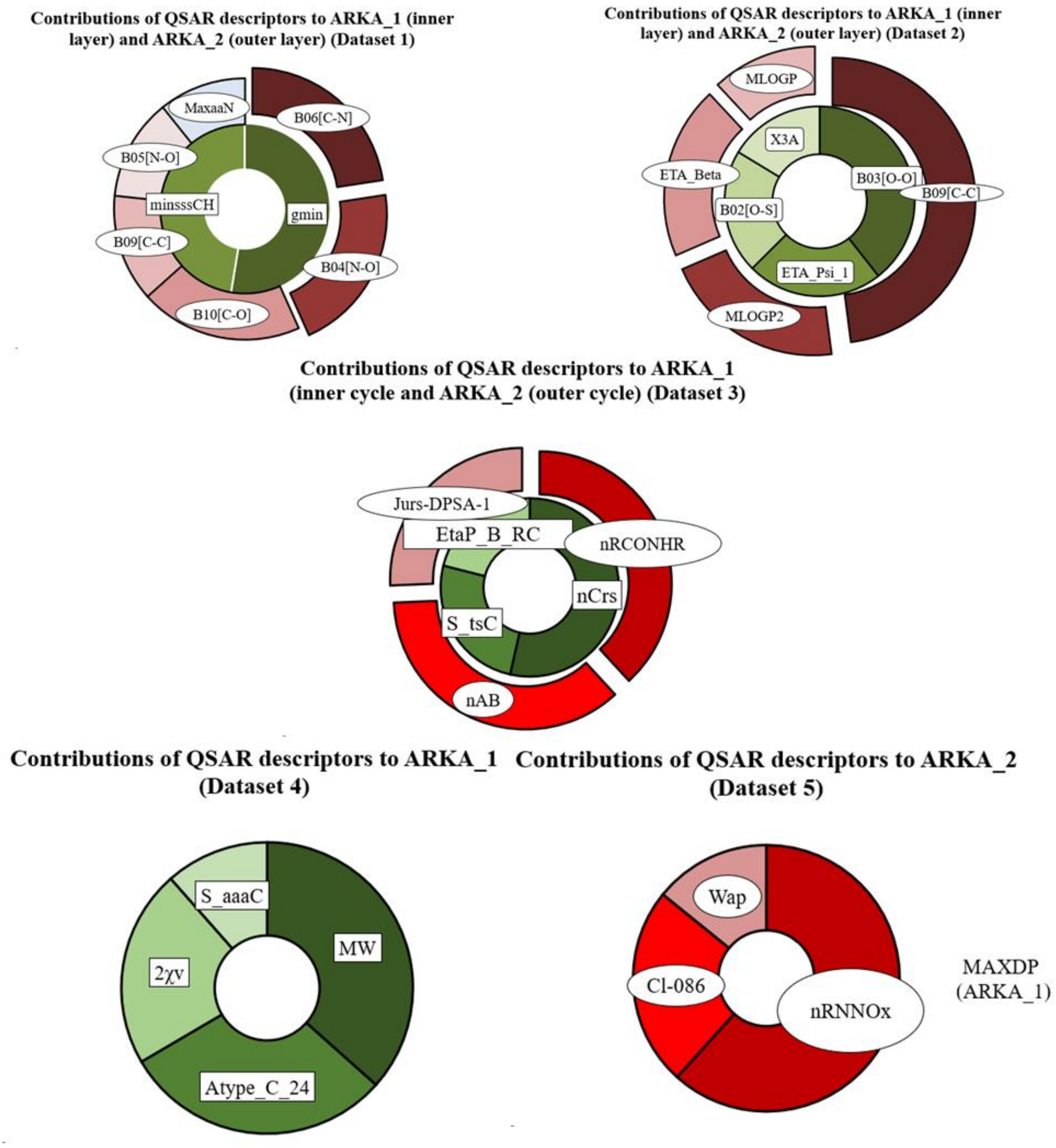
Case study 4: MDF analysis of descriptors contributing to the toxicity towards P. subcapitata

The dataset used for this study (Dataset 4) reports the toxicity of organic chemicals towards *P. subcapitata*. Like in the previous cases, the features were clustered based on the difference values. Since all the features had a positive difference value, this was the only set that used only one ARKA descriptor (ARKA_1) to generate models. Among the different features, MW (denotes the molecular weight of the compound) had the highest contribution towards ARKA_1, which was followed by the contributions of Atype_C_24 (representing fragments containing secondary carbon atom), $^{2}\chi^{v}$ (denoting size and shape) and S_aaaC (representing fused ring system) (**Figure 9**).

Case study 5: MDF analysis of descriptors contributing to rodent carcinogenicity potential

The dataset used for this study (Dataset 5) reports the rodent carcinogenic potency. On assigning clusters to the descriptors, it was observed that the descriptor MAXDP (a measure reflecting the electrophilicity of a molecule) contributes to ARKA_1. Similarly, the descriptors Wap (denoting the Wiener index, i.e., the edge count through the shortest path between all pairs of non-hydrogen atoms), nRNNOx (representing the number of N-nitroso groups that are aliphatic), and Cl-086 (depicting the presence of Cl atoms attached to an sp3 hybridized carbon atom) contribute to ARKA_2. As evident from the weightage values, among the three descriptors contributing to ARKA_2, the descriptor nRNNOx is observed to have the highest contribution in the computation of the ARKA_2 descriptor (**Figure 9**).

The details of the descriptors contributing to ARKA_1 and ARKA_2, for all the five different datasets, have been represented in **Figure 9**.



Contributions of QSAR descriptors to ARKA_1 (inner layer) and ARKA_2 (outer layer) (Dataset 1)

Contributions of QSAR descriptors to ARKA_1 (inner layer) and ARKA_2 (outer layer) (Dataset 2)

Contributions of QSAR descriptors to ARKA_1 (inner cycle and ARKA_2 (outer cycle) (Dataset 3)

Contributions of QSAR descriptors to ARKA_1 (Dataset 4)

Contributions of QSAR descriptors to ARKA_2 (Dataset 5)

Figure 9. QSAR descriptors that are encoded in ARKA_1 and ARKA_2 for all five different datasets. Note that the model for dataset 4 is derived from only one ARKA descriptor (ARKA_1).

3.9 Application of the ARKA framework to chemical read-across predictions

We have applied the ARKA descriptors for chemical-similarity-based read-across classification analysis for all the considered endpoints and compared the results with the predictions obtained from conventional QSAR descriptors (**Table 2**). The application of the Gaussian kernel-based similarity [16] showed that the ARKA framework outperforms the conventional QSAR descriptors in the external prediction quality for most of the data sets. This warrants further studies on the application of the ARKA framework in similarity-based cheminformatics studies.

Table 2. Effects of ARKA descriptors on the chemical read-across-based external predictions using the Gaussian kernel function for five data sets (Ndesc = the number of descriptors, MCC = Matthews Correlation Coefficient, Ckappa = Cohen's kappa)*

Dataset	Descriptors	Ndesc	f1_score	MCC	Ckappa	AUC
1	QSAR	14	**0.729**	0.21	0.209	**0.66**
	ARKA	**2**	0.699	**0.235**	**0.227**	**0.66**
2	QSAR	8	0.6	0.42	0.412	0.78
	ARKA	**2**	**0.645**	**0.472**	**0.467**	**0.79**
3	QSAR	6	0.361	-0.079	-0.079	0.43
	ARKA	**2**	**0.375**	**-0.144**	**-0.143**	**0.49**
4	QSAR	4	0.9	0.753	0.723	0.95
	ARKA	**1**	**0.923**	**0.812**	**0.795**	**1**
5	QSAR	4	**0.917**	**0.713**	**0.673**	**0.96**
	ARKA	**2**	**0.917**	**0.713**	**0.673**	0.95

*The winner metric values are shown in bold.

3.10 Limitations and Future Prospects

With every new method being developed, the associated limitations show the avenues for future prospects. This analysis is essential for the mitigation and generation of further possibilities that ultimately lead to progress in science. In this particular method of dimensionality reduction, we have stressed the modeling of small datasets, considering mainly the environmental aspects where the toxicity and ecotoxicity data are limited, amplifying the need for data gap-filling. Primarily, the two ARKA descriptors (ARKA_1 and ARKA_2) should be employed for modeling small datasets that do not have a very large number of modeling descriptors. When larger datasets with a considerably higher number of QSAR features are involved, the computation of only two classes of ARKA descriptors becomes somewhat of an oversimplification, since there is a high chance that the information of the larger pool of descriptors may not be efficiently encoded in just two classes of ARKA descriptors. From a general conscience, this calls for the need to develop a greater number of ARKA descriptors by dividing the original QSAR descriptor pool into *K*-groups ("*K*-groups analysis"), instead of just two groups. By generating a somewhat greater number of ARKA descriptors, a lower number of features get encoded into a single ARKA descriptor, which reduces noise and redundancy. However, it may differ from case to case depending on the complexity of a data set. For example, within the results of the reported five datasets, Dataset 1 had the highest number of compounds (n = 471) and the highest number of descriptors (Ndesc =14), and it was observed that the models developed using two-descriptors (ARKA_1 and ARKA_2) had better predictive performance than the QSAR models developed using 14 descriptors (Data set 1). In this work, we have worked on classification-based models and compared the predictive performance of the models generated using conventional QSAR descriptors and ARKA descriptors. However, this method of dimensionality reduction can also be explored when the response values are quantitative, leading to the development of regression-based models. To develop the regression-based models, the scheme for the computation of the descriptors may remain the same and these descriptors may be submitted to a regression-based modeling algorithm. However, in the initial step where we have grouped the active and inactive classes of the training compounds, this grouping should be done taking a certain value as the threshold (preferably the experimental response mean of the training compounds) in case of regression modeling, since this will contain quantitative data for the endpoint values. In addition, the correlation (*r*) of a particular descriptor with the response can also be possibly used as a substitute for the residual value and weightage may be assigned based on the *r-value* of a particular descriptor (with the training set response). Additionally, the computation and assignment of the weightage to each descriptor can be customized. In this work, we have chosen a simple arithmetic weighing strategy but other modelers may also use a customized weighing strategy based on their choice. One can apply different weighting summation schemes as applied in the computation of mixture descriptors from multi-component chemical mixtures using various algebraic expressions like Quadratic mixture descriptors, Logarithmic mixture descriptors, etc. [49].

The ARKA descriptors can also play an important role in the similarity assessment of chemicals for regulatory decision-making. The plot of ARKA_1 vs ARKA_2 can not only identify potential activity cliffs but can also help one to understand the similar types of chemicals that are grouped in a cluster – a basic form of Read-Across. Additionally, it is also possible to identify the chemical nature and possible adverse outcome pathways (AOPs) of the close congeners using the concepts of Read-Across [50, 51] and quantitative read-across structure-activity relationship (q-RASAR) [26, 52]. This approach can not only be used in assessing environmental/ecotoxicity endpoints but can also be extended to other fields like drug discovery [53].

4 Conclusion

As a thumb rule for any statistical modeling analysis for small datasets, there should be a minimal number of descriptors used for modeling. This enhances the degree of freedom of the developed model and increases the statistical reliability. In this particular work, we have used the same amount of chemical information from the descriptors used in the previously reported ecotoxicological QSAR models and encoded them in such a way that has significantly lowered the number of modeling descriptors – ARKA descriptors (a form of dimensionality reduction technique). On retraining the model using ARKA descriptors, it was observed that the models generated using ARKA descriptors had better predictivity (for the test sets) as compared to the previously published QSAR models, which was evident from various classification-based statistical validation metrics (Mathews correlation coefficient, Cohen's kappa, f1 score, and AUC-ROC) that provide an unbiased result in evaluating the classification ability of a model into the actives and the inactives even for imbalanced data sets. To ensure that this observation was not limited to a particular modeling algorithm, we have trained various additional Machine Learning (ML) models using the QSAR descriptors and ARKA descriptors separately, the comparative prediction results of which strengthen our inference. From the modeling exercise on five diverse ecotoxicity data, we observe that in most of the cases, models developed using ARKA descriptors outperform the predictive ability of the models developed using conventional QSAR descriptors. Additionally, the two ARKA descriptors can potentially identify activity cliffs, less confident data points, and less modelable data points; the results obtained in this study comply with the previously reported method of the detection of activity cliffs utilizing the concept of chemical similarity [26].

Therefore, we infer that the models generated using ARKA descriptors can quickly and efficiently identify toxic environmental chemicals with enhanced predictivity, thus leading to increased reliability of the predictions. However, there is room for further development of the approach by its applications in regression-based and/or read-across approaches, classification modeling of larger ecotoxicity data sets, and exploring other customized ways of weighing strategies in deriving ARKA descriptors. Lastly, we like to infer that although this work shows that the predictive performance of the models developed using ARKA descriptors supersedes the predictive performance of models developed using conventional QSAR descriptors in the majority of the studied cases, all computational models are not 100% accurate since every model is associated with some errors and uncertainties in predictions. We feel that experimental testing is confirmatory of any kind of developed hypothesis and there should not be over-dependence on any kind of computational models.

Acknowledgment

AB thanks Life Sciences Research Board, Defence Research and Development Organisation, Govt. of India, for a Senior Research Fellowship

Conflict of interest

Declared none.

Funding

Life Sciences Research Board, Defence Research and Development Organisation, Govt. of India

Author contributions

AB – Conceptualization, Computation, Validation, Writing - Initial draft, Software

KR – Conceptualization, Supervision, Writing –Editing, Funding

References

1. Khan, K., Roy, K., 2022. Ecotoxicological risk assessment of organic compounds against various aquatic and terrestrial species: application of interspecies i-QSTTR and species sensitivity distribution techniques. Green Chem. 24, 2160-2178. https://doi.org/10.1039/D1GC04320J.
2. Fjodorova, N., Novich, M., Vrachko, M., Smirnov, V., Kharchevnikova, N., Zholdakova, Z., Novikov, S., Skvortsova, N., Filimonov, D., Poroikov, V., Benfenati, E., 2008. Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. J. Environ. Sci. Health Part C 26, 201-236. https://doi.org/10.1080/10590500802135578.
3. Khan, K., Baderna, B., Cappelli, C., Toma, C., Lombardo, A., Roy, K., Benfenati, E., 2019. Ecotoxicological QSAR modeling of organic compounds against fish: Application of fragment based descriptors in feature analysis. Aquat. Toxicol. 212, 162-174. https://doi.org/10.1016/j.aquatox.2019.05.011.
4. OECD: https://www.oecd.org/about/. Accessed on 18th March 2024.
5. Piir, G., Kahn, I., Garcia-Sosa, AT., Sild, S., Ahte, P., Maran, U., 2018. Best practices for QSAR model reporting: Physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints. Environ. Health Pers. 126, 126001. https://doi.org/10.1289/EHP3264.
6. Banerjee, A., De, P., Kumar, V., Kar, S., Roy, K., 2022. Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across. Chemosphere 309, 136579. https://doi.org/10.1016/j.chemosphere.2022.136579.
7. EU REACH: https://echa.europa.eu/it/regulations/reach/legislation. Accessed on 18th March 2024.

8. Roy, K., Kar, S., Das, RN., 2015. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press. NY. https://doi.org/10.1016/C2022-0-00080-5.

9. Mansouri, K., Cariello, NF., Korotcov, A., Tkachenko, V., Grulke, CM., Sprankle, CS., Allen, D., Casey, WM., Kleinstreuer, NC., Williams, AJ., 2019. Open-source QSAR models for pKa prediction using multiple machine learning approaches. J. Cheminform. 11, 60. https://doi.org/10.1186/s13321-019-0384-1.

10. Gini, G., Zanoli, F., Machine Learning and Deep Learning Methods in Ecotoxicological QSAR Modeling. In: Roy, K., (Eds.) Ecotoxicological QSARs. Springer, NY, 111-149. https://doi.org/10.1007/978-1-0716-0150-1_6.

11. Rodriguez-Perez, R., Bajorath, J., 2020. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. J. Med. Chem. 63, 8761-8777. https://doi.org/10.1021/acs.jmedchem.9b01101.

12. Karpov, P., Godin, G., Tetko, IV., 2020. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J. Cheminformatics 12. https://doi.org/10.1186/s13321-020-00423-w.

13. Manganelli, S., Benfenati, E., 2016. Use of Read-Across Tools. In: Benfenati, E. (eds) In Silico Methods for Predicting Drug Toxicity. Methods in Molecular Biology, vol 1425. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-3609-0_13.

14. Ball, N., Cronin, MTD., Shen, J., Blackburn, K., Booth, ED., Bouhifd, M., Donley, E., Egnash, L., Hastings, C., Juberg, DR., Kleensang, A., Kleinstreuer, N., Kroese, ED., Lee, AC, Luechtefeld, T., Maertens, A., Marty, S., Naciff, JM, Palmer, J., Pamies, D., Penman, M., Richarz, AN., Russo, DP., Stuard, SB., Patlewicz, G., van Ravenzwaay,

B., Wu, S., Zhu, H., Hartung, T., 2016. Toward Good Read-Across Practice (GRAP) guidance. ALTEX 33, 149-166. https://doi.org/10.14573/altex.1601251.

15. Hung, C., Gini, G., 2021. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. Mol. Divers. 25, 1283-1299. https://doi.org/10.1007/s11030-021-10250-2.
16. Chatterjee, M., Banerjee, A., De, P., Gajewicz-Skretna, A., Roy, K., 2022. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environ. Sci.: Nano 9, 189-203. https://doi.org/10.1039/D1EN00725D.
17. Banerjee, A., Roy, K., 2022. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Mol. Divers. 26, 2847-2862. https://doi.org/10.1007/s11030-022-10478-6.
18. Srisongkram, T., 2023. Ensemble quantitative read-across structure–activity relationship algorithm for predicting skin cytotoxicity. Chem. Res. Toxicol. 36, 1961-1972. https://doi.org/10.1021/acs.chemrestox.3c00238.
19. Keshavarz, MH., Gharagheizi, F., Shokrolahi, A., Zakinejad, S.,2012. Accurate prediction of the toxicity of benzoic acid compounds in mice via oral without using any computer codes. J. Hazard. Mater. 237-238, 79-101. https://doi.org/10.1016/j.jhazmat.2012.07.048.
20. Jafari, M., Keshavarz, MH., Salek, H., 2019. A simple method for assessing chemical toxicity of ionic liquids on Vibrio fischeri through the structure of cations with specific anions. Ecotox. Environ. Safety 182, 109429. https://doi.org/10.1016/j.ecoenv.2019.109429.
21. Sivakumar, J., Ramamurthy, K., Radhakrishnan, M., Won, D., 2022. Synthetic sampling from small datasets: A modified mega-trend diffusion approach using k-

nearest neighbors. Know. Based. Syst. 236, 107687. https://doi.org/10.1016/j.knosys.2021.107687.

22. Nath, A., De, P., Roy, K., 2021. In silico modelling of acute toxicity of 1, 2, 4-triazole antifungal agents towards zebrafish (Danio rerio) embryos: Application of the Small Dataset Modeller tool. Toxicol. in Vitro 75, 105205. https://doi.org/10.1016/j.tiv.2021.105205.
23. Khan, K., Kumar, V., Colombo, E., Lombardo, A., Benfenati, E., Roy, K., 2022. Intelligent consensus predictions of bioconcentration factor of pharmaceuticals using 2D and fragment-based descriptors. Environ. Int. 170, 107625. https://doi.org/10.1016/j.envint.2022.107625.
24. Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemom. Intell. Lab. Syst. 2, 37-52. https://doi.org/10.1016/0169-7439(87)80084-9.
25. Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1.
26. Banerjee, A., Roy, K., 2023. Prediction-inspired intelligent training for the development of classification read-across structure–activity relationship (c-RASAR) models for organic skin sensitizers: Assessment of classification error rate from novel similarity coefficients. Chem. Res. Toxicol. 36, 1518-1531. https://doi.org/10.1021/acs.chemrestox.3c00155.
27. Roy, J., Ojha, PK., Carnesecchi, E., Lombardo, A., Roy, K., Benfenati, E., 2020. First report on a classification-based QSAR model for chemical toxicity to earthworm. J. Hazard. Mater. 386, 121660. https://doi.org/10.1016/j.jhazmat.2019.121660.
28. Kar, S., Roy, K., 2013. Prediction of milk/plasma concentration ratios of drugs and environmental pollutants using in silico tools: Classification and regression based

QSARs and pharmacophore mapping. Mol. Inform. 32, 693-705. https://doi.org/10.1002/minf.201300018.

29. Pramanik S., Roy, K., 2014. Predictive modeling of chemical toxicity towards Pseudokirchneriella subcapitata using regression and classification based approaches. Ecotox. Environ. Safety 101, 184-190. https://doi.org/10.1016/j.ecoenv.2013.12.030.
30. Kar, S., Deeb, O., Roy, K., 2012. Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor. Ecotox. Environ. Safety 82, 85-95. https://doi.org/10.1016/j.ecoenv.2012.05.013.
31. Gramatica, P., Cassani, S., Roy, PP., Kovarich, S., Yap, CW., Papa, E., 2012. QSAR modeling is not “push a button and find a correlation”: A case study of toxicity of (benzo-)triazoles on algae. Mol. Inform. 31, 817-835. https://doi.org/10.1002/minf.201200075.
32. Murcia-Soler, M., Perez-Gimenez, F., Garcia-March, FJ., Salabert-Salvador, MT., Diaz-Villanueva, W., Medina-Casamayor, P., 2003. Discrimination and selection of new potential antibacterial compounds using simple topological descriptors. J. Mol. Graph. Model. 21, 375-390. https://doi.org/10.1016/S1093-3263(02)00184-5.
33. Das, RN., Roy, K., 2014. Predictive modeling studies for the ecotoxicity of ionic liquids towards the green algae Scenedesmus vacuolatus. Chemosphere 104, 170-176. https://doi.org/10.1016/j.chemosphere.2013.11.002.
34. DTC Lab tools Supplementary Website: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. Accessed on 18th March 2024.
35. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825-2830.

36. Kluyver, T., Ragan-Kelly, B., Perez, F., Granger, BE., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, JB., Grout, J., Corlay, S., Ivanov, P., 2016. Jupyter Notebooks-a publishing format for reproducible computational workflows. In Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing; Loizides,F., Schmidt, B., Eds., IOS Press, 87-90.
37. Stoltzfus, JC., 2011. Logistic Regression: A brief primer. Aca. Emergen. Med. 18, 1099-1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x.
38. Lau, KW., Wu, QH., 2003. Online training of support vector classifier. Pattern Recognit. 36, 1913−1920. https://doi.org/10.1016/S0031-3203(03)00038-4.
39. Pal, M., 2003. Random forest classifier for remote sensing classification. Int. J. Rem. Sens. 26, 217-222. https://doi.org/10.1080/01431160412331269698.
40. De Diego, I.M., Redondo, A.R., Fernández, R.R., Navarro, J., Moguerza, JM., 2022. General performance score for classification problems. Appl. Intell. 52, 12049–12063. https://doi.org/10.1007/s10489-021-03041-7.
41. Nahm, FS., 2022. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J. Anes. 75, 25-36. https://doi.org/10.4097/kja.21209.
42. Snedecord GW, Cochran WG, Statistical Methods, Wiley-Blackwell, NJ, 8th edition, 1989.
43. Gramatica, P., Giani, E., Papa, E., 2007. Statistical external validation and consensus modeling: A QSPR case study for Koc prediction. J. Mol. Graph. Model. 25, 755-766. https://doi.org/10.1016/j.jmgm.2006.06.005.
44. OECD Grouping of Chemicals: Chemical Categories and Read-Across: https://www.oecd.org/chemicalsafety/risk-assessment/groupingofchemicalschemicalcategoriesandread-

across.htm/#:~:text=In%20the%20read%2Dacross%20approach,same%20mode%20or%20mechanisms%20of. Accessed on 18th March 2024.

45. Kovarich S, Ceriani L, Gatnik MF, Bassan A, Pavan M, 2019. Filling Data Gaps by Read-across: A Mini Review on its Application, Developments and Challenges. Mol. Inform. 38, 1800121, https://doi.org/10.1002/minf.201800121.

46. Patlewicz G, Chemical Categories and Read-across, EUR 21898 EN, EUROPEAN COMMISSION DIRECTORATE GENERAL JOINT RESEARCH CENTRE, 2005, https://publications.jrc.ec.europa.eu/repository/bitstream/JRC31792/Chemical%20Categories%20and%20Read%20across_Dec.pdf.

47. Ling, CX., Huang, J., Zhang, H., 2003. AUC: A better measure than accuracy in comparing learning algorithms. In: Xiang, Y., Chaib-draa, B., (Eds.) Advances in Artificial Intelligence. Canadian AI 2003. Lecture notes in computer science, 2671. Springer. 329-341. https://doi.org/10.1007/3-540-44886-1_25.

48. Enoch, SJ., Cronin, MTD., Schultz, TW., Madden, JC., 2008. Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via michael addition. Chem. Res. Toxicol. 21, 513-520. https://doi.org/10.1021/tx700322g.

49. Saldana, DA., Starck, L., Mougin, P., Rousseau, B., Creton, B., 2013. Prediction of flash points for fuel mixtures using machine learning and a novel equation. Energy Fuels 27, 3811-3820. https://doi.org/10.1021/ef4005362.

50. Lizarraga, LE., Suter, GW., Lambert, JC., Patlewicz, G., Zhao, JQ., Dean, JL., Kaiser, P., 2023. Advancing the science of a read-across framework for evaluation of data-poor chemicals incorporating systematic and new approach methods. Regulat. Toxicol. Pharmacol. 137, 105293. https://doi.org/10.1016/j.yrtph.2022.105293.

51. Spinu, N., Cronin, MTD., Enoch, SJ., Madden, JC., Worth, AP., 2020. Quantitative adverse outcome pathway (qAOP) models for toxicity prediction. Arch. Toxicol. 94, 1497-1510. https://doi.org/10.1007/s00204-020-02774-7.

52. Banerjee, A., Roy, K., 2023. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points. 36, 446-464. https://doi.org/10.1021/acs.chemrestox.2c00374.

53. Patlewicz, G., Fitzpatrick, JM., 2016. Current and future perspectives on the development, evaluation, and application of in silico approaches for predicting toxicity. Chem. Res. Toxicol. 29, 438-451. https://doi.org/10.1021/acs.chemrestox.5b00388.