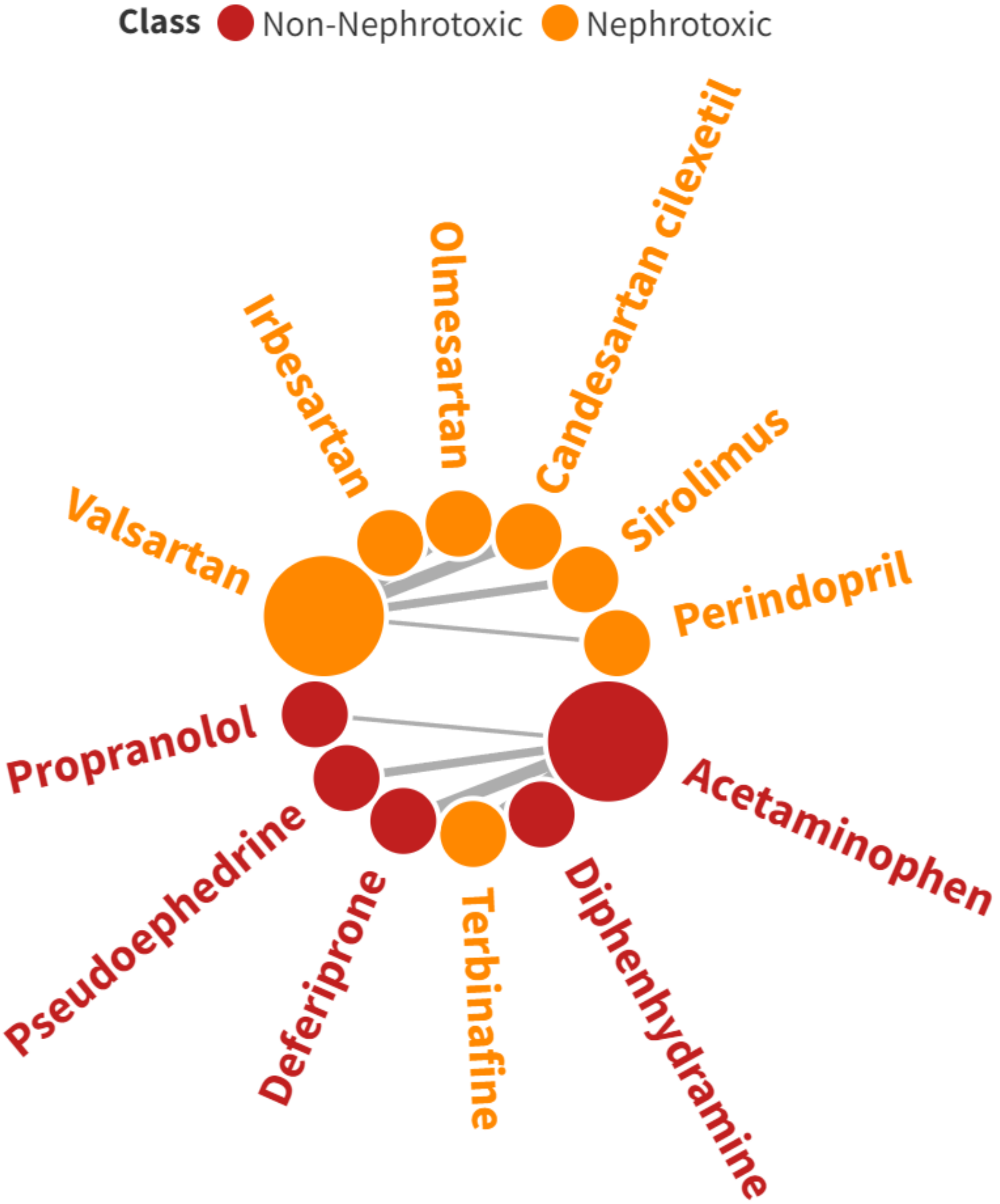
Preprint Version dated 22 August 2024
Not peer reviewed

Machine learning-assisted c-RASAR modeling of a curated set of orally active nephrotoxic drugs: Similarity-based predictions from close source neighbors

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)



Graphical Abstract

Abstract

Cheminformatics and Machine Learning (ML) have seen exponential progress in the last decade, in the field of chemical risk assessment, due to their efficiency, accuracy, and reliability. The constant evolution of New Approach Methodologies (NAM) has inspired researchers around the globe to deviate from conventional approaches and adopt or develop new, “unconventional” methods. The classification Read-Across Structure-Activity Relationship (c-RASAR) is an unconventional approach that utilizes similarity and error-based information from the nearest neighboring compounds into a Machine Learning modeling framework, resulting in enhanced predictivity. Although this technique has so far been applied to molecular descriptors, we have applied this approach in the present study on molecular fingerprints along with conventional molecular descriptors for ML-based model development from a recently reported highly curated set of orally active nephrotoxic drugs. We initially developed ML models using nine different linear and non-linear algorithms separately on molecular descriptors and MACCS fingerprints, thus generating 18 different ML QSAR models. Using the chemical spaces defined by the modeling descriptors and fingerprints, the similarity and error-based RASAR descriptors were computed, and the most discriminating RASAR descriptors were used to develop another set of 18 different ML c-RASAR models. All 36 models were cross-validated 20 times with a 5-fold cross-validation strategy, and their predictivity was checked on the test set data. A multi-criteria decision-making strategy – the Sum of Ranking Differences (SRD) approach - was adopted to identify the best-performing model based on robustness and external validation parameters. This statistical analysis suggested that the c-RASAR models had an overall good performance, while the best-performing model was also a c-RASAR model. This model was used to screen a true external set data prepared from the known nephrotoxic compounds of DrugBankDB. These results also showed that our model efficiently identifies nephrotoxic compounds. The t-SNE analyses on the descriptors, fingerprints, and the RASAR descriptor spaces inferred that the RASAR descriptors efficiently encode the chemical information, as evident from the tight and distinct clustering of the data points. Additionally, the molecular descriptors and the corresponding RASAR descriptors were used to identify potential activity cliffs using the ARKA framework.

Keywords: c-RASAR, Machine Learning, Sum of Ranking Differences (SRD), Nephrotoxicity, ARKA, t-SNE

Introduction

Kidneys, one of the most vital organs of the human body, are two bean-shaped organs responsible for filtering out toxic substances and metabolites from the blood, thus helping to excrete them from the body, resulting in detoxification. However, their efficiency is significantly reduced when certain external or internal factors prevent their proper functioning. Drug-Induced Kidney Injury (DIKI) has been a significant contributor to this issue since various drugs result in kidney damage either directly or indirectly [1]. It has been observed that out of every five drugs reaching Phase III of the clinical trial, one drug has been withdrawn due to its associated nephrotoxic effects [2]. Typically, antihypertensive classes of drugs like Diuretics, angiotensin receptor blockers, angiotensin-converting enzyme inhibitors, calcium channel blockers, and painkillers belonging to the class of cyclooxygenase inhibitors work by disrupting the renal hemodynamics and the glomerular filtration pressure [3]. Additionally, drugs like zalcitabine, cisplatin, and amphotericin B, among others, are responsible for the damage of renal mitochondrial constituents, thus aiding in the disruption of cellular energy production [4]. Other drugs like Tacrolimus, Acyclovir, and Puromycin are responsible for the decreased oxidative phosphorylation, crystal deposition in the glomerulus or the renal tubule, and formation of abnormal proteins resulting in stress to the endoplasmic reticulum, respectively. Therefore, it is essential to determine the nephrotoxic potential of drugs and drug-like molecules at an early stage of the drug discovery pipeline for a better future and to avoid the colossal expenses of developing unsuccessful drug candidates. Essentially, it takes a lot of time, labor, and cost to experimentally determine the nephrotoxic potential of drugs, and this results in the shift in paradigm towards adopting computational approaches that are fast, reliable, more efficient, and less expensive.

In silico approach is one of the go-to methods to generate fast and reliable predictions of a particular endpoint, for any query compound. With the development of the Quantitative Structure-Activity Relationship (QSAR) studies, scientists have successfully been able to correlate a molecule's structural and physicochemical features with the target endpoint [5]. Typically, this consists of a mathematical model where the structural and physicochemical features are considered a linear function of the target response. However, modern QSARs have considerably deviated from this simplicity and have started to consider non-linear relationships of the features with the target response. This is where various Machine Learning (ML) and Deep Learning (DL) algorithms have now been successfully integrated into the QSAR paradigm [6]. ML concepts are used not only in the context of model development but also for the proper and judicious identification of the essential features that have some relationship with the response values. In terms of modeling data points, the availability of various ML approaches is essentially required as they capture different linear and non-linear relationships in different data structures. With the advent of neural networks and DL, the modern world has been presented with various highly precise tools that effectively encode various hidden patterns among data. However, from a statistical point of view, we find that traditional QSAR models are not often reliable when modeling small datasets. This is because small dataset modeling warrants a considerable amount of feature space that leads to considering a larger pool of modeling descriptors, thus reducing the degree of freedom of the developed model [7]. Adherence to non-statistical approaches like Read-Across is common nowadays, especially for dealing with small datasets [8-9]. In its simplest form, Read-Across identifies close congeners of a particular query compound, and its property prediction is obtained using the experimentally known data of the close source neighbors [10]. Although this is a popular tool in predictive toxicology, its only limitation is that, in most cases, one cannot directly understand the relative contribution of the features quantitatively. To compile the advantages of both the QSAR and Read-Across approaches, Roy's group developed the quantitative Read-Across Structure-Activity Relationship (q-RASAR) approach that inducts the concepts of Read-Across into a mathematical modeling framework, using the Read-Across-derived similarity and error-based measures as descriptors [11-12]. Although the term q-RASAR is applied to modeling quantitative endpoint data, this concept has been further extended to the field of classification modeling, where the classification RASAR models are termed c-RASAR [13]. This novel chemometric technique has been shown to enhance predictivity compared to the conventional QSAR models in various previous studies, although utilizing the same amount of chemical space [14-18]. As evident from the previous studies [19-21], another important property of q-RASAR and c-RASAR models is that they can generate models using a lower number of descriptors with enhanced predictivity compared with the corresponding QSAR models. What differentiates the RASAR descriptors from the conventional QSAR descriptors of a compound is that the latter describes the property of that particular compound, while the former represents the information of its close source neighbors [22]. Moreover, the computation of the RASAR descriptors involves the standard concepts of Machine Learning (ML), where we optimize the Read-Across hyperparameters and accordingly use that setting to compute the RASAR descriptors. Therefore, even in a linear q-RASAR or c-RASAR model, the used RASAR descriptors have originally been derived through a nonlinear function, which paves the path to a novel idea where it is possible to encode non-linear relationships into a linear modeling framework [14].

A few computational modeling studies of the nephrotoxicity of chemicals and drugs have been reported previously [23-25]. However, these studies involved data sets that included organic chemicals (non-drugs), herbal medicines, and responses with conflicting reports. This means that those modeling sets did not represent highly curated data for the nephrotoxicity of drugs. Recently, Connor et al. [26] published a highly curated set of nephrotoxicity of orally active drugs which we have used here for c-RASAR model development. So far, most of the studies on applying q-RASAR and c-RASAR models have centered on using descriptors, and its application on chemical space defined by fingerprints has remained an unexplored area. Therefore, we have developed QSAR models in this study using the standard 0-2D descriptor matrix and the MACCS fingerprint. Consequently, the standard QSAR descriptors and the MACCS fingerprints defined two different molecular structure representations. After that, we developed c-RASAR models based on the two different feature matrices. We have applied various machine learning modeling algorithms to the QSAR and RASAR descriptors. The best models and the employed modeling algorithms were determined using the Sum of Ranking Differences (SRD) approach [27]. Using the best model, we have additionally screened a true external set of data and determined the generalizability of our model. Additional analyses involving the development of t-SNE plots on the four different feature spaces inferred that the RASAR descriptors more efficiently encoded the complete chemical information. Additionally, various activity cliffs were identified using the novel supervised dimensionality reduction framework – ARKA [7], and their nature has been explained using the information of their closest congeners.

Materials and Methods

Collection of the Nephrotoxicity data

A list of 317 orally active nephrotoxic drugs was assembled from the works of Connor et al. [26] and has been provided in **Supplementary Materials SI-1**. The motive of the work of Connor et al. was to create a complete, comprehensible, and curated dataset that can be used for new approach methodologies (NAMs). This study identified different orally administered drugs and their nephrotoxicity data from different literature sources. To generate a comprehensive nephrotoxicity dataset, they verified the listed nephrotoxicity data from various literature sources, including external sources like the FDA and DrugBankDB. This careful curation performed as per the strategy described in [28] was essential as it was observed that different literature sources often had contrasting nephrotoxicity data for a particular drug molecule. Additionally, some nephrotoxicity data had contrasting inferences when verified with the data from different sources like the FDA and DrugBankDB. As per the OECD principle 1 ("A defined endpoint"), the authors believe that Connor et al. did a fantastic job that can prevent the model development process from being misled in the presence of erroneous observed data.

Structural representation and chemical curation

The SMILES notations were used to draw the structures in MarvinSketch (https://chemaxon.com/marvin). The structures were manually curated to remove mixtures and inorganic components. Further curation steps involved adding explicit hydrogens and converting the ring systems to their aromatic form. The curated compounds were then saved in a single .sdf file to calculate descriptors and fingerprints.

Calculation of descriptors and fingerprints

The descriptors and fingerprints were calculated using the alvaDesc software [29]. Simple 0-2D descriptors from the classes of constitutional indices, ring descriptors, molecular properties, functional group counts, atom-centered fragments, atom-type E-state indices, 2D atom pairs, connectivity indices, and extended topochemical atom (ETA) indices were calculated. Additionally, MACCS-166-bit fingerprints were calculated for all the molecules. The descriptor matrix and fingerprint matrices were saved in two different Excel files.

Data Pre-treatment

Among the large number of computed descriptors and MACCS fingerprints, there were a lot of features that possess significant inter-correlation, noise and some missing values that are considered as "string" entities. Since these are impeding factors for the development of a statistically meaningful model, such descriptors and fingerprints were removed using the Java-based Data Pre-Treatment tool available from https://teqip.jdvu.ac.in/QSAR_Tools/.

Dataset splitting

The standard practice in developing QSAR models is to assess their performance on the training data and check how the developed models generalize with the unseen data. Following this, we split the dataset into training and test sets, where the training set was used to develop models while the test set was used to evaluate the predictive performance on unseen data.

At first, we separated the actives and inactives of our dataset. Considering the active compounds only, a *t*-distributed Stochastic neighbor embedding (t-SNE) plot [30] was developed using the pre-treated 0-2D descriptor matrix. The *t*-SNE values (t-SNE1 and t-SNE2), thus obtained, were temporarily considered as a descriptor matrix to encode non-linear relationships in our data division process. Using this temporary descriptor matrix (consisting only of t-SNE1 and t-SNE2), we have applied the Euclidean distance-based division algorithm to divide the active dataset into training and test sets, employing the Dataset division tool available from https://teqip.jdvu.ac.in/QSAR_Tools/. Next, we considered the inactive compounds only and performed the same algorithm using *t*-SNE to have another set of training and test sets. Finally, the training set of the actives and inactives were merged to obtain a complete training set. Similarly, the test set of the actives and inactives were merged to obtain a complete test set. It is to be noted that the finally obtained training and test sets were composed of the pre-treated 0-2D descriptor matrix. For the MACCS dataset, we have maintained the same training and test set data composition as obtained using the process as mentioned earlier.

Feature selection of the molecular descriptors

Among the different descriptors computed, it was essential to identify the features most likely to affect the target outcome. For this, we have employed the most discriminating feature selection technique, a.k.a. molecular spectrum analysis [31], to identify the essential features. This technique computes the absolute mean difference of the normalized values of a particular descriptor in the active and inactive classes. The descriptors that have higher absolute mean difference values in the training set are considered as important descriptors. It is to be noted that this feature selection technique is "model independent", i.e., we have not employed any modeling algorithm to screen out the essential features, which might not work well for other modeling techniques.

Development of Machine Learning QSAR models

For the development of models, we adopted an array of linear and non-linear Machine Learning (ML) modeling algorithms. We employed nine different ML models on the feature spaces defined by the molecular descriptors and MACCS fingerprints, generating 18 different ML QSAR models. The modeling techniques employed were Linear Discriminant Analysis (LDA) [32], Support Vector Machine (SVM) [33], Random Forest (RF) [34], Logistic Regression (LR) [35], Quadratic Discriminant Analysis (QDA) [36], Multilayer Perceptron (MLP) [37], Gaussian Naïve Bayes (NB) [38], Gradient Boosting (GB) [39] and Adaboost [40]. It is to be noted that the descriptor and the fingerprint matrices were standardized before the development of the ML models. The hyperparameters were optimized using GridSearchCV, adhering to a 5-fold cross-validation technique, taking accuracy as the objective function. Additionally, 20 times 5-fold cross-validation [41] of all the developed ML-based QSAR models was performed to check their robustness and identify overfitting. The developed models underwent rigorous internal and external validation to check the robustness and external predictivity on the test set (unseen) data.

Optimization of the Read-Across hyperparameters and computation of the RASAR descriptors

Once we developed the ML QSAR models using molecular descriptors and fingerprints, we used the same feature spaces to compute the similarity and error-based RASAR descriptors. However, the basic pre-requisite is to identify the optimized hyperparameter setting using Read-Across. This was done by dividing the training set into sub-training and validation sets, and Read-Across predictions for the validation set were generated using the tool Read-Across-v4.2.2 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. Different combinations of hyperparameter settings were explored, and the selection of the optimized setting was based on the prediction performance of the validation set. The selected hyperparameter settings were used to compute the RASAR descriptors for the training and test sets using RASAR-Desc-Calc-v3.0.3, available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. It is to be noted that this Read-Across optimization and computation of the RASAR descriptors using the optimized setting was done twice, first using the selected molecular descriptors and second using the MACCS fingerprints. The complete list of RASAR descriptors that was computed using the RASAR descriptor calculator tool has been listed in **Table S1** of the **Supplementary Materials SI-2**.

Feature selection of the RASAR descriptors

Like the QSAR analysis, feature selection was performed on the RASAR descriptor matrix to identify the most discriminating features. However, before this, we have deliberately removed the RASAR descriptors *SD_Activity, SE,* and *CVact*, which stands for the weighted standard deviation of the activity values of the close congeners, the corresponding standard error, and the coefficient of variation. Since we were developing classification-based models with either 0 or 1 response values, these three descriptors should be omitted [22]. We employed the same feature selection algorithm, i.e., identifying the most discriminating features, as used for the QSAR analysis which aimed towards an unbiased feature selection due to its modeling algorithm-independent nature. This procedure was carried out on the RASAR descriptor matrices generated from the molecular and fingerprint descriptor spaces.

Development of ML-based c-RASAR models

Similar to the QSAR analysis, we have employed nine different linear and non-linear ML modeling algorithms on each of the selected RASAR descriptor matrices. These modeling algorithms include Linear Discriminant Analysis, Support Vector Machine, Random Forest, Logistic Regression, Quadratic Discriminant Analysis, Multilayer Perceptron, Gaussian Naïve Bayes, Gradient Boosting and Adaboost classifiers. A total set of 18 different ML-based c-RASAR models were developed (9 models for the descriptor-based RASAR and 9 models for the fingerprint-based RASAR). It should be noted that the selected RASAR descriptor matrices were standardized before the development of the ML c-RASAR models. The hyperparameters were optimized using GridSearchCV, adhering to a 5-fold cross-validation technique, taking accuracy as the objective function. Additionally, 20 times 5-fold cross-validation of all the developed ML-based c-RASAR models was performed to check their robustness and identify overfitting. The developed models underwent rigorous internal and external validation to check the robustness and external predictivity on the test set (unseen) data.

Performance evaluation of the different ML models using the Sum of Ranking Differences (SRD) approach

This is an important aspect where judging the best-performing model is important. Among all the 36 different ML models (18 QSAR models and 18 c-RASAR models), we must identify the best model and the better modeling strategy among QSAR and c-RASAR. This was achieved by adopting the Sum of Ranking Differences (SRD) approach [27], which is a form of Multi-Criteria Decision Making (MCDM) strategy, where the best model was identified based on different external and internal validation metrics. External validation metrics like Accuracy, Balanced Accuracy, Precision, Recall, F1_score, Matthews Correlation Coefficient (MCC), Cohen's kappa (Ckappa), and AUC, while 20 times 5-fold cross-validated internal validation metrics like AccuracyCV, Balanced AccuracyCV, PrecisionCV and RecallCV were considered. Additionally, as parameters for robustness, the absolute differences of the training set Accuracy, Balanced Accuracy, Precision, and Recall, with the AccuracyCV, Balanced AccuracyCV, PrecisionCV, and RecallCV, respectively, were considered. These robustness parameters ideally equate to the lower the absolute difference, the better the model. In contrast, the other parameters are the opposites; the exact robustness parameter we considered is 1-ABS(Metric-MetricCV). All 16 different metrics formed our "Multi-Criteria," which was subjected to SRD analysis to identify the best-performing model. The SRD analysis was carried out using a software named CRRN_DNA (downloaded from http://knight.kit.bme.hu/CRRN).

Generalization of the best-performing model – Analysis of a true external set data

Screening of a true external set is essential for the proper estimation of the model's predictive performance. In this regard, we collected the list of approved drugs showing nephrotoxicity from the DrugBank Database (https://go.drugbank.com/categories/DBCAT003959). From this list, we have eliminated the drug molecules that were already a part of our training set, as well as the drugs that are organometallic in nature. The structures of the final list of 112 nephrotoxic drugs were drawn, curated and the relevant RASAR descriptors were computed. This was the true external set used for the prediction with the best-performing model.

A detailed workflow has been presented in **Figure 1**.

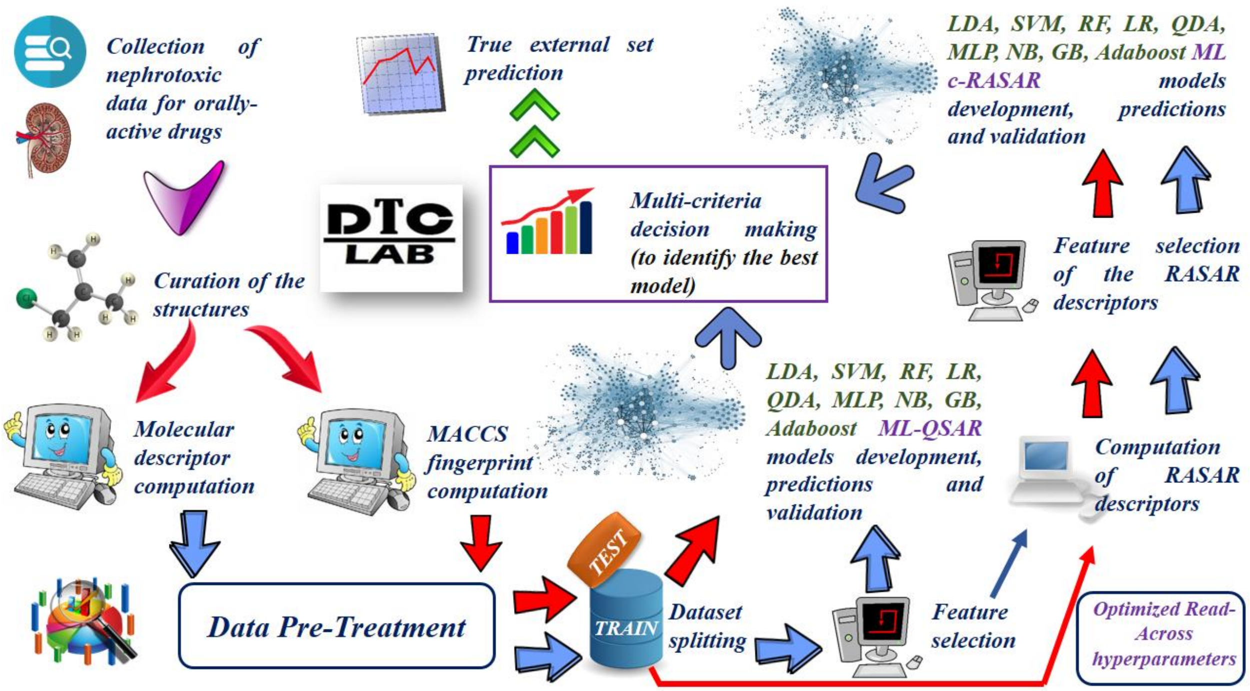

Collection of nephrotoxic data for orally-active drugs
True external set prediction
LDA, SVM, RF, LR, QDA, MLP, NB, GB, Adaboost ML c-RASAR models development, predictions and validation
DTC LAB
Multi-criteria decision making (to identify the best model)
Curation of the structures
Feature selection of the RASAR descriptors
LDA, SVM, RF, LR, QDA, MLP, NB, GB, Adaboost ML-QSAR models development, predictions and validation
Molecular descriptor computation
MACCS fingerprint computation
Computation of RASAR descriptors
TEST
TRAIN
Data Pre-Treatment
Dataset splitting
Feature selection
Optimized Read-Across hyperparameters

Figure 1. Detailed workflow of the model development procedure

Results and Discussion

Analysis of the chemical diversity of the dataset

This initial level of analysis aimed to explore the structural diversity of the compounds constituting the dataset. **Figure 2** represents a chemical diversity plot where the compounds are located according to their similarity. This plot was generated using DataWarrior (https://openmolecules.org/datawarrior/), using structural similarity information based on substructure fragment dictionary-based binary FragFp. Taking a well-known nephrotoxic compound Ibuprofen as the reference, it is evident from this plot that the dataset is highly diverse, offering a significant challenge for developing reliable mathematical models.

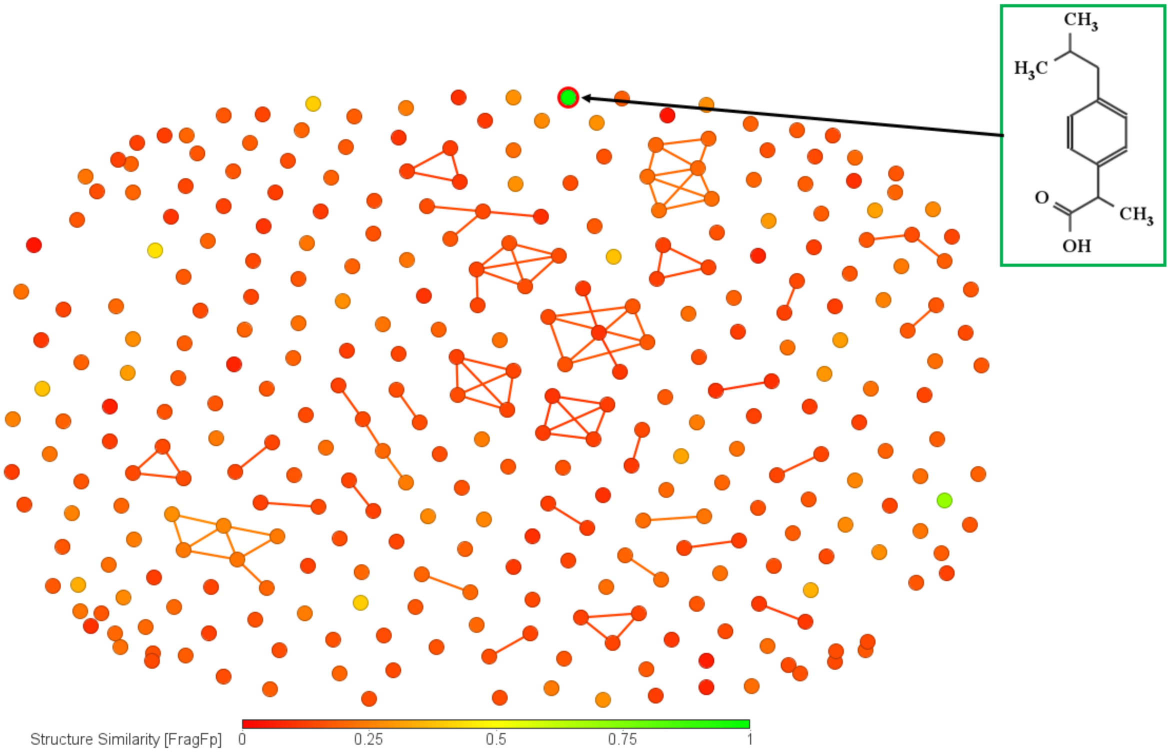


Figure 2. Chemical diversity analysis shows that the dataset compounds are highly dissimilar, taking Ibuprofen as the reference standard.

Selection of the important molecular descriptors for QSAR analysis

For the efficient selection of essential features, we identified the descriptors that have high discriminating power between the positive and the negative classes, using the most discriminating feature selection algorithm [31]. The reason for adopting this feature selection technique is that it is independent of any particular modeling algorithm, thus enabling a fair comparison between the developed ML models. We used a Java-based tool MDF_Identifier-v1.0, available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home to identify the most discriminating features. We selected those features that have absolute mean difference values > 0.05. A list of 21 descriptors falls under this category and has been presented in **Supplementary Materials SI-1**.

Selection of the important RASAR descriptors for c-RASAR analysis

The selection of the essential RASAR descriptors follows the same algorithm as the selection of the essential molecular descriptors for QSAR analysis. The RASAR descriptors with an absolute mean difference of >0.11 were selected for modeling analysis. In the case of RASAR descriptors computed from the selected molecular descriptors, five fall under this category, while six RASAR descriptors calculated from the MACCS fingerprints fall under this category. The computed RASAR descriptors used to develop c-RASAR models are presented in **Supplementary Materials SI-1**.

Results of the ML-based QSAR and c-RASAR models

The quality metrics of all the developed QSAR models (nine using molecular descriptors and nine using MACCS fingerprints) and c-RASAR models (nine using the RASAR descriptors derived from the molecular features and nine using the RASAR descriptors derived from MACCS fingerprints) have been reported in **Tables 1 and 2,** respectively. An array of different linear and non-linear ML modeling algorithms was adopted to generate classification-based QSAR models. The hyperparameters associated with the various models were optimized using GridSearchCV, adhering to a 5-fold cross-validation strategy. The models' quality was assessed using various classification-based validation metrics, and the best models were judged based on a "multi-criteria decision-making" strategy (to be discussed later in the manuscript).

Table 1. Results of the different QSAR and c-RASAR models developed from 0-2D molecular descriptors

Set	**Models**	**Acc.**	**BA**	**Precision**	**Recall**	**F1 score**	**MCC**	**Cohen's kappa**	**AUC**
QSAR (using 0-2D molecular descriptors)									
Train	**LDA**	0.686	0.687	0.721	0.682	0.701	0.372	0.371	0.74
	SVM	0.686	0.637	0.714	0.698	0.706	0.370	0.370	0.75
	RF	0.812	0.808	0.804	0.860	0.831	0.621	0.619	0.9
	LR	0.682	0.681	0.709	0.698	0.703	0.361	0.361	0.75
	QDA	0.665	0.669	0.721	0.620	0.667	0.338	0.334	0.69
	MLP	0.879	0.876	0.868	0.915	0.891	0.756	0.755	0.95

	NB	0.644	0.654	0.734	0.535	0.619	0.314	0.301	0.71
	GB	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00
	AB	0.958	0.958	0.961	0.961	0.961	0.916	0.916	0.99
Test	**LDA**	0.59	0.585	0.614	0.643	0.628	0.172	0.171	0.63
	SVM	0.641	0.685	0.659	0.690	0.674	0.275	0.275	0.65
	RF	0.628	0.627	0.659	0.643	0.651	0.254	0.253	0.67
	LR	0.577	0.571	0.600	0.643	0.621	0.144	0.144	0.64
	QDA	0.667	0.665	0.690	0.690	0.690	0.329	0.329	0.66
	MLP	0.641	0.637	0.659	0.690	0.674	0.275	0.275	0.68
	NB	0.615	0.623	0.688	0.524	0.595	0.249	0.241	0.65
	GB	0.513	0.516	0.556	0.476	0.513	0.032	0.031	0.59
	AB	0.603	0.601	0.634	0.619	0.627	0.202	0.202	0.61
c-RASAR (from 0-2D molecular descriptors)									
Train	***LDA***	0.615	0.615	0.650	0.620	0.635	0.229	0.228	0.66
	SVM	0.900	0.894	0.862	0.969	0.912	0.803	0.796	0.93
	RF	0.941	0.938	0.914	0.984	0.948	0.884	0.881	0.99
	LR	0.623	0.626	0.670	0.597	0.631	0.251	0.249	0.66
	QDA	0.603	0.608	0.660	0.543	0.596	0.216	0.212	0.65
	MLP	0.720	0.724	0.777	0.674	0.722	0.447	0.442	0.81
	NB	0.603	0.608	0.660	0.543	0.596	0.216	0.212	0.65
	GB	0.874	0.869	0.846	0.938	0.890	0.750	0.745	0.96
	AB	0.724	0.723	0.748	0.736	0.742	0.445	0.445	0.77

Test	***LDA***	0.718	0.714	0.727	0.762	0.744	0.431	0.430	0.71
	SVM	0.603	0.593	0.612	0.714	0.659	0.192	0.189	0.56
	RF	0.628	0.617	0.627	0.762	0.688	0.245	0.238	0.63
	LR	0.718	0.716	0.738	0.738	0.738	0.433	0.433	0.71
	QDA	0.679	0.683	0.730	0.643	0.684	0.364	0.361	0.7
	MLP	0.615	0.617	0.658	0.595	0.625	0.234	0.232	0.59
	NB	0.679	0.683	0.730	0.643	0.684	0.364	0.361	0.71
	GB	0.603	0.591	0.608	0.738	0.667	0.191	0.186	0.64
	AB	0.615	0.615	0.650	0.619	0.634	0.230	0.230	0.69

Acc.: Accuracy, BA: Balanced Accuracy

Table 2. Results of the different QSAR and c-RASAR models developed from MACCS fingerprints

Set	**Models**	**Acc.**	**BA**	**Precision**	**Recall**	**F1 score**	**MCC**	**Cohen's kappa**	**AUC**
			QSAR (using MACCS fingerprints)						
Train	**LDA**	0.745	0.742	0.758	0.775	0.766	0.485	0.485	0.84
	SVM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00
	RF	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00
	LR	0.745	0.734	0.718	0.868	0.786	0.490	0.477	0.79
	QDA	0.987	0.987	0.985	0.992	0.988	0.975	0.975	1.00
	MLP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00
	NB	0.623	0.625	0.667	0.605	0.634	0.249	0.248	0.69

	GB	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00
	AB	0.715	0.718	0.761	0.690	0.724	0.434	0.432	0.79
Test	**LDA**	0.615	0.621	0.676	0.548	0.605	0.243	0.238	0.66
	SVM	0.641	0.639	0.667	0.667	0.667	0.278	0.278	0.68
	RF	0.628	0.623	0.644	0.690	0.667	0.248	0.248	0.69
	LR	0.577	0.571	0.600	0.643	0.621	0.144	0.144	0.6
	QDA	0.692	0.696	0.750	0.643	0.692	0.393	0.388	0.69
	MLP	0.679	0.685	0.743	0.619	0.675	0.370	0.364	0.7
	NB	0.603	0.607	0.657	0.548	0.597	0.215	0.211	0.62
	GB	0.641	0.639	0.667	0.667	0.667	0.278	0.278	0.68
	AB	0.577	0.577	0.615	0.571	0.593	0.154	0.154	0.57
				c-RASAR (from MACCS fingerprints)					
Train	***LDA***	0.674	0.670	0.689	0.721	0.705	0.341	0.340	0.69
	SVM	0.707	0.701	0.709	0.775	0.741	0.408	0.406	0.77
	RF	0.782	0.774	0.755	0.884	0.814	0.566	0.556	0.85
	LR	0.653	0.650	0.677	0.682	0.680	0.301	0.301	0.68
	QDA	0.653	0.650	0.677	0.682	0.680	0.301	0.301	0.7
	MLP	0.665	0.653	0.654	0.806	0.722	0.323	0.312	0.72
	NB	0.653	0.650	0.677	0.682	0.680	0.301	0.301	0.68
	GB	0.774	0.767	0.755	0.860	0.804	0.546	0.540	0.86
	AB	0.678	0.679	0.717	0.667	0.691	0.356	0.355	0.72
Test	***LDA***	0.628	0.629	0.667	0.619	0.642	0.257	0.256	0.64

SVM	0.590	0.583	0.609	0.667	0.636	0.169	0.168	0.59
RF	0.641	0.641	0.675	0.643	0.659	0.281	0.281	0.65
LR	0.628	0.629	0.667	0.619	0.642	0.257	0.256	0.66
QDA	0.628	0.629	0.667	0.619	0.642	0.257	0.256	0.67
MLP	0.615	0.609	0.630	0.690	0.659	0.221	0.220	0.68
NB	0.628	0.629	0.667	0.619	0.642	0.257	0.256	0.66
GB	0.667	0.669	0.711	0.643	0.675	0.336	0.335	0.71
AB	0.667	0.675	0.750	0.571	0.649	0.354	0.342	0.69

Acc.: Accuracy, BA: Balanced Accuracy

Results for the cross-validation of all the developed models

Cross-validation is an integral aspect to judge the robustness and stability of a model and to ensure that the overall quality of models is not dependent on a certain limited number of compounds only. The purpose of cross-validation is to check whether the performance of a model is stable even with the removal of certain data points from the training set. In the present investigation, we have cross-validated all our developed QSAR and c-RASAR models to check their robustness. We have performed rigorous cross-validation by adopting 20 times 5-fold cross-validation strategy using Accuracy, Balanced Accuracy, Precision, and Recall as the objective functions. The results of cross-validation have been presented in **Table 3**. From this table, it can be observed that there has been an increase in robustness (indicated by the marginal decrease in the cross-validated objective functions) of the c-RASAR models as compared to the conventional QSAR models, indicating that the models are not overfitted. Moreover, the significantly reduced number of modeling descriptors in the c-RASAR models provides greater compliance to the statistical considerations. **Figure 3** presents a heat map of the absolute difference between the individual metric values and their cross-validated values. It can be clearly observed that the overall robustness of the MACCS QSAR models are lower, while the MACCS c-RASAR models are the most robust.

Table 3. 20 times 5-fold cross-validation statistics of the developed models (Acc: Accuracy, BA: Balanced Accuracy, Prec: Precision, Rec: Recall)

Model	**Acc.**	**BA**	**Prec.**	**Rec**	***Acc. CV***	***BA. CV***	***Prec. CV***	***Rec. CV***
20 times 5-fold CV results of QSAR models (using molecular descriptors)								
LDA_QSAR	0.686	0.687	0.721	0.682	*0.611*	*0.616*	*0.656*	*0.603*
SVM_QSAR	0.686	0.637	0.714	0.698	*0.578*	*0.583*	*0.623*	*0.577*
RF_QSAR	0.812	0.808	0.804	0.860	*0.599*	*0.598*	*0.622*	*0.682*
LR_QSAR	0.682	0.681	0.709	0.698	*0.610*	*0.613*	*0.646*	*0.631*
QDA_QSAR	0.665	0.669	0.721	0.620	*0.612*	*0.617*	*0.660*	*0.587*
MLP_QSAR	0.879	0.876	0.868	0.915	*0.603*	*0.602*	*0.633*	*0.645*
NB_QSAR	0.644	0.654	0.734	0.535	*0.611*	*0.618*	*0.685*	*0.516*
GB_QSAR	1.000	1.000	1.000	1.000	*0.592*	*0.592*	*0.624*	*0.618*
AB_QSAR	0.958	0.958	0.961	0.961	*0.600*	*0.599*	*0.633*	*0.630*
20 times 5-fold CV results of c-RASAR models (developed from molecular descriptors)								
LDA_c-RASAR	0.615	0.615	0.650	0.620	*0.599*	*0.602*	*0.638*	*0.595*
SVM_c-RASAR	0.900	0.894	0.862	0.969	*0.664*	*0.658*	*0.674*	*0.730*
RF_ c-RASAR	0.941	0.938	0.914	0.984	*0.609*	*0.608*	*0.636*	*0.656*
LR_ c-RASAR	0.623	0.626	0.670	0.597	*0.594*	*0.599*	*0.641*	*0.564*
QDA_ c-RASAR	0.603	0.608	0.660	0.543	*0.596*	*0.602*	*0.650*	*0.546*
MLP_ c-RASAR	0.720	0.724	0.777	0.674	*0.625*	*0.617*	*0.634*	*0.732*

NB_ c-RASAR	0.603	0.608	0.660	0.543	*0.590*	*0.594*	*0.638*	*0.553*
GB_ c-RASAR	0.874	0.869	0.846	0.938	*0.621*	*0.618*	*0.642*	*0.681*
AB_ c-RASAR	0.724	0.723	0.748	0.736	*0.588*	*0.586*	*0.612*	*0.693*
20 times 5-fold CV results of QSAR models (using MACCS fingerprints)								
LDA_MACCS_QSAR	0.745	0.742	0.758	0.775	*0.579*	*0.579*	*0.611*	*0.624*
SVM_MACCS_QSAR	1.000	1.000	1.000	1.000	*0.633*	*0.635*	*0.664*	*0.668*
RF_MACCS _QSAR	1.000	1.000	1.000	1.000	*0.620*	*0.619*	*0.641*	*0.701*
LR_MACCS _QSAR	0.745	0.734	0.718	0.868	*0.587*	*0.586*	*0.605*	*0.742*
QDA_MACCS_QSAR	0.987	0.987	0.985	0.992	*0.638*	*0.637*	*0.658*	*0.704*
MLP_MACCS_QSAR	1.000	1.000	1.000	1.000	*0.632*	*0.632*	*0.656*	*0.684*
NB_MACCS_QSAR	0.623	0.625	0.667	0.605	*0.535*	*0.537*	*0.575*	*0.533*
GB_MACCS_QSAR	1.000	1.000	1.000	1.000	*0.611*	*0.611*	*0.638*	*0.658*
AB_MACCS_QSAR	0.715	0.718	0.761	0.690	*0.606*	*0.611*	*0.646*	*0.623*
20 times 5-fold CV results of c-RASAR models (developed from MACCS fingerprints)								
LDA_MACCS_c-RASAR	0.674	0.670	0.689	0.721	*0.654*	*0.653*	*0.674*	*0.699*
SVM_MACCS_c-RASAR	0.707	0.701	0.709	0.775	*0.640*	*0.638*	*0.659*	*0.701*
RF_MACCS _c-RASAR	0.782	0.774	0.755	0.884	*0.665*	*0.657*	*0.667*	*0.755*
LR_MACCS _c-RASAR	0.653	0.650	0.677	0.682	*0.643*	*0.644*	*0.659*	*0.728*
QDA_MACCS_c-RASAR	0.653	0.650	0.677	0.682	*0.640*	*0.640*	*0.669*	*0.663*
MLP_MACCS_c-RASAR	0.665	0.653	0.654	0.806	*0.642*	*0.638*	*0.654*	*0.724*
NB_MACCS_c-RASAR	0.653	0.650	0.677	0.682	*0.653*	*0.651*	*0.676*	*0.681*

GB_MACCS_c-RASAR	0.774	0.767	0.755	0.860	*0.673*	*0.666*	*0.672*	*0.772*
AB_MACCS_c-RASAR	0.678	0.679	0.717	0.667	*0.665*	*0.661*	*0.674*	*0.745*

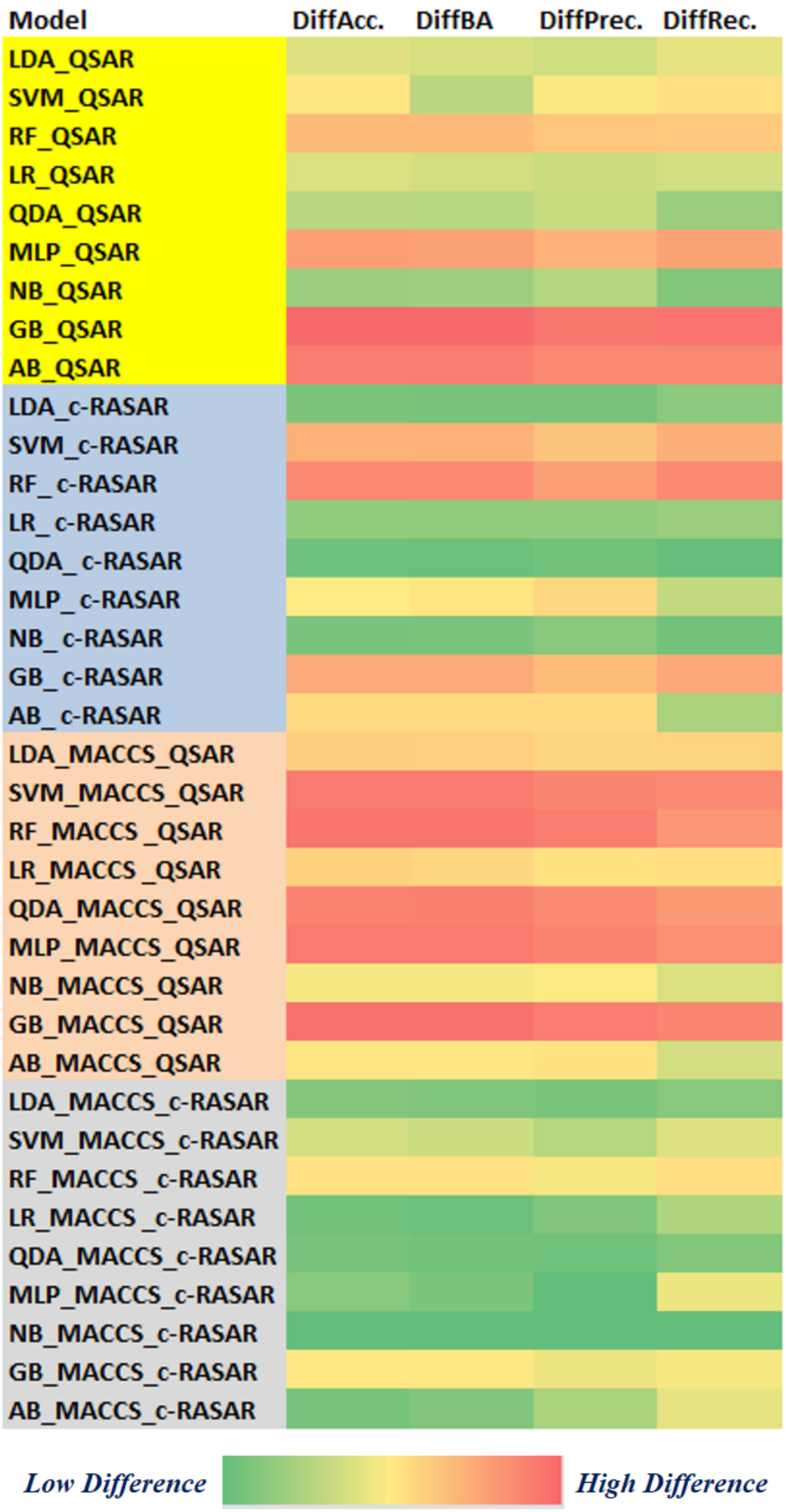


Figure 3. A heat map that pictorially demonstrates the robustness of the developed models after 20 times 5-fold cross-validation. It is observed that MACCS c-RASAR models are highly robust.

(DiffAcc= Absolute difference between Accuracy and AccuracyCV, DiffBA= Absolute difference between Balanced Accuracy and Balanced AccuracyCV, DiffPrec= Absolute difference between Precision and PrecisionCV, DiffRec= Absolute difference between Recall and RecallCV)

Identification of the best-performing model – An application of the Sum of Ranking Difference (SRD) approach

Since we have developed many mathematical models using different combinations of modeling descriptors, it is critical to identify the best-performing model directly. This is because this judgment should ideally encompass factors like robustness and predictivity. Therefore, we have adopted a multi-criteria decision-making strategy to judge the best-performing models using many objective functions. The Sum of Ranking Differences (SRD) is a well-known method to estimate the best-performing model based on multiple criteria [27]. In this approach, the data should be arranged in a matrix with the metric values in the column and models in the rows. The metric values should be scaled (for example, scaled to unit length) column-wise, and then the scaled matrix may be transposed so the comparison models appear column-wise. Then, the absolute difference between the standard reference (which may be the maximum value row-wise) and individual method ranks is deduced and summed for each technique. In this manner, the sum of ranking difference (SRD) values is calculated for each method. An SRD value closer to zero (i.e., the closer the ranking is to the reference value) signifies that the model is better. Concerning external predictivity, we have considered metrics like Accuracy, Balanced Accuracy, Precision, Recall, F1_score, MCC, Ckappa, and AUC that define the predictive performance on the test set. Additionally, metrics like AccuracyCV, Balanced AccuracyCV, PrecisionCV, and RecallCV were considered for encoding information relating to robustness. Since the difference between a metric and its cross-validated value is a measure of robustness, we have additionally considered the absolute differences in the training set Accuracy, Balanced Accuracy, Precision, and Recall, with the AccuracyCV, Balanced AccuracyCV, PrecisionCV, and RecallCV, respectively. These 16 different parameters, representing robustness and predictivity of models, were considered for SRD analysis. We have validated the method using leave-one-seventh-out cross-validation. The scaled SRD values between 0 and 100 were calculated using the software named CRRN_DNA (downloaded from http://knight.kit.bme.hu/CRRN). The results were graphically analyzed by plotting the % SRD data (Fig. X) for each modeling technique in a random environment, i.e., random ranking given to each data input for each model to generate all possible random sum of ranking differences. The SRD plot represents different modeling techniques placed in ascending order of their SRD values. The critical threshold XX1 indicates the region of randomness with $p < 0.05$ (i.e., probability of randomness less than 5%), Med means 50% randomness, and XX19 signifies 95% randomness.

Three different sets of analyses were performed. First, we intended to identify the best-performing ML models developed from molecular descriptors and their corresponding ML c-RASAR models. In the second case, we analyzed the ML models developed from the MACCS fingerprints and their corresponding ML c-RASAR models. Lastly, we took all the developed models and performed an overall comparison. As evident from **Figures 4a** and **5a**, where the analysis is between the descriptor-based ML QSAR models and their corresponding ML c-RASAR models, it can be observed that the overall performance of the ML c-RASAR models is better than the ML QSAR models. Additionally, the best-performing model appeared to be the **LDA c-RASAR model**. On the other hand, **Figures 4b** and **5b** represent the analysis between the fingerprint-based ML QSAR models and their corresponding ML c-RASAR models. Again, the c-RASAR models performed better than the corresponding QSAR models. The best-performing model in this comparison appeared to be the **Adaboost MACCS c-RASAR model**. The SRD analysis of all the developed models (36 models) has been presented in **Figures 4c** and **5c**. From this analysis comparing all the developed models, the **LDA c-RASAR model** appeared to be the best-performing model. This is quite significant where a linear model performs better than many other non-linear ML models, using different types of descriptors and fingerprints, thus demonstrating the potential of c-RASAR models fortifying previous similar observations [12, 13, 17, 42]. The models are represented in the following codes: Q1-Q9 = LDA, SVM, RF, LR, QDA, MLP, NB, GB, and AB QSAR models developed from molecular descriptors, M1-M9 = LDA, SVM, RF, LR, QDA, MLP, NB, GB, and AB QSAR models developed from MACCS fingerprints, Q1R-Q9R = LDA, SVM, RF, LR, QDA, MLP, NB, GB and AB c-RASAR models developed from molecular descriptors and M1R-M9R = LDA, SVM, RF, LR, QDA, MLP, NB, GB and AB c-RASAR models developed from MACCS fingerprints.

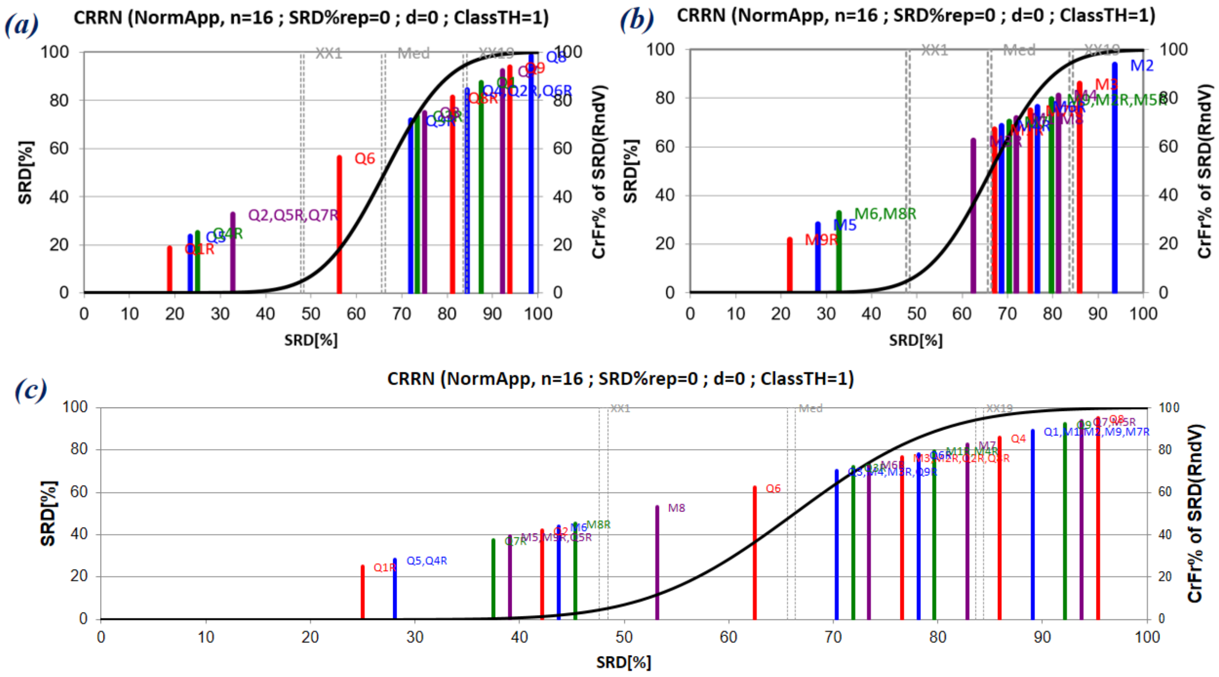


Figure 4. SRD analysis of a) the descriptor-based QSAR and c-RASAR models, b) the fingerprint-based QSAR and c-RASAR models, and c) all the developed models. The X-axis and left Y-axis represent the normalized SRD values, whose small values indicate better models. The right Y-axis represents the cumulative relative frequencies corresponding to the randomization test. (CRRN: Comparison of Ranks with Ranking Numbers)

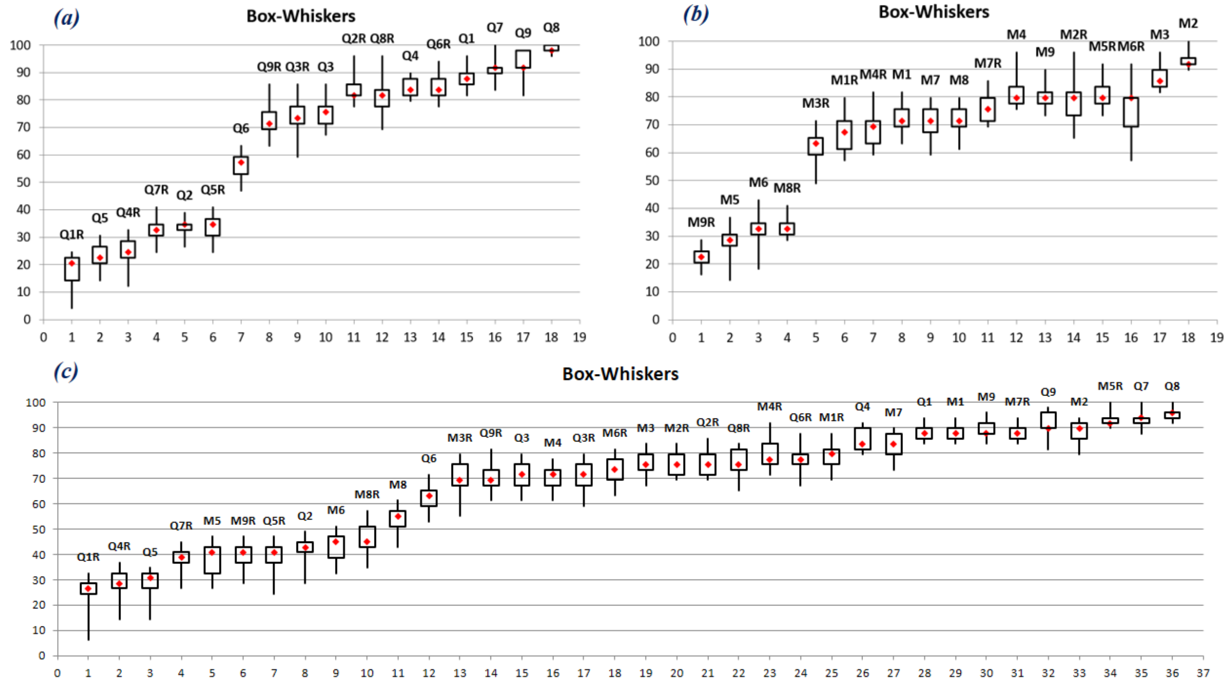


Figure 5. Leave-1/7th-out cross-validated SRD results showing the LDA q-RASAR model (Q1R) is the best model.

Interpretation of the RASAR descriptors used to develop the LDA c-RASAR model

One of the essential purposes of any modeling analysis is to interpret the modeled features and provide an idea of their contribution toward the endpoint of interest. After selecting the best-performing model (LDA c-RASAR) by statistical evaluation using the SRD approach, we used the LDA coefficients to identify the contribution of the RASAR descriptors toward the model. Since the c-RASAR models are developed using similarity and error-based RASAR descriptors, it is essential to note that the interpretation is relative, and it considers the structural characteristics of the close source congeners. The descriptor ***RA function*** is a read-across-derived function, which is a compact representation of the entire structural and physicochemical descriptor space into a single variable [43]. Since it encodes all chemical information, this descriptor contributes positively to the response. This can be observed in the case of Dabrafenib (**283**), which has a high value of *RA function* and is nephrotoxic. Similarly, Ribavirin (**265**) has a very low value of *RA function* and is observed to be non-nephrotoxic. The descriptor ***CVsim*** demonstrates the coefficient of variation of the similarity values of close source congeners for a particular target compound, and this descriptor contributes positively towards the response. This indicates the high dispersion of the similarity values of the close congeners, inferring that the dataset is highly diverse, as also previously demonstrated in **Figure 2**. This can be exemplified by Irbesartan (**197**), that have a high value of *CVsim* and is an active compound, while inactive compounds like Chlorzoxazone (**50**) have a lower *CVsim* value. The descriptor ***MaxNeg*** is the similarity value to the closest negative/inactive source compound for an individual query compound, and this descriptor contributes negatively to the response. A query compound having a higher value of *MaxNeg* justifies that it shares a high similarity to an inactive/negative compound, which increases the propensity of the query compound to be inactive. On the other hand, a compound having a lower maximum similarity to a negative compound is most likely to become an active compound. This can be exemplified with the inactive compound Theophylline (**261**), which has a high value of *MaxNeg*, while active compounds like Cefpodoxime (**167**) have a lower value of *MaxNeg*. This is shown pictorially in **Figure 6** where we also analyze the structures of the close source compounds. It is observed that the nearest inactive neighbor of Theophylline (an inactive compound) is Caffeine, which is highly similar in structure. However, the nearest inactive neighbor of Cefpodoxime (an active compound) is Terazosin, with a very low level of similarity. This proves that the highly similar compounds of Cefpodoxime are active, explaining why the *MaxNeg* expresses a negative contribution. The descriptor $\boldsymbol{s_m^1}$, a.k.a. the Banerjee-Roy similarity coefficient, is a novel concordance measure that helps identify activity cliffs [13]. However, from a modeling point of view, this descriptor makes a positive contribution. As evident from the formula mentioned in the work of Banerjee and Roy [13], a positive contribution is expected since a higher value of *MaxPos* than *MaxNeg* signifies that the query compound has a propensity to become active. This can be exemplified with active compounds like Valsartan (**11**), which has a high value of s_m^1. Similarly, inactive compounds like Labetalol (**142**) have lower values of s_m^1. The descriptor $\boldsymbol{g_m\_class}$ is a modified version of the Banerjee-Roy concordance coefficient, and this descriptor contributes positively to the response. The main property of this descriptor is that it is binary (values are either 0 or 1), and can potentially identify the propensity of a particular compound to be active or inactive. This can be observed in active compounds like Rabeprazole (102) that has a high value of g_m_class, while inactive compounds like Riboflavin (152) have a lower value of g_m_class.

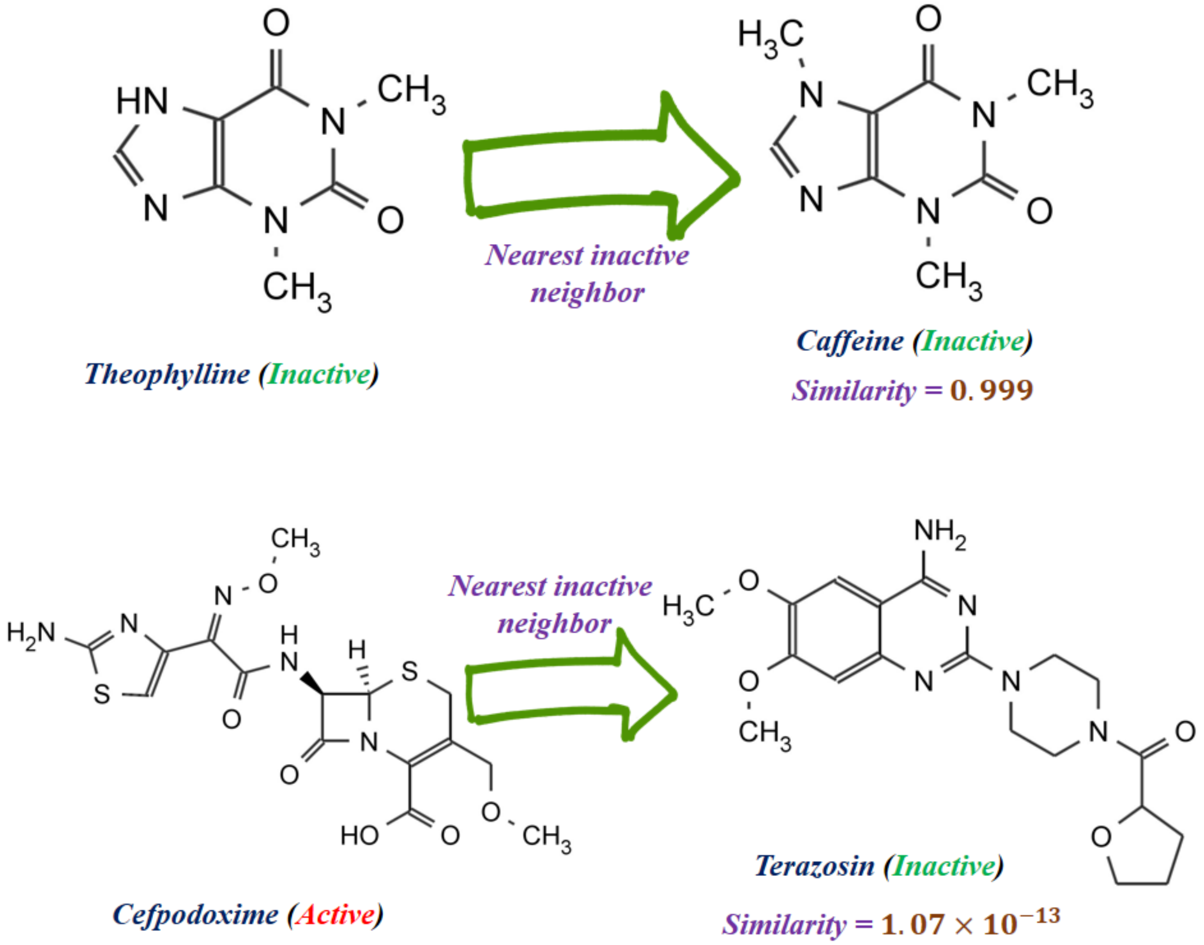


Figure 6. Analysis of the nearest negative/inactive compounds for active and inactive query compounds

Predictions of the true external set data using the LDA c-RASAR model

This is an essential aspect that further justifies the generalizability of the developed model. We identified 112 compounds labeled as nephrotoxic from the DrugBankDB. It should be noted that these 112 compounds were exclusively the compounds not present in the training set and are not organometallic. However, on analyzing the predictive performance, it was observed that out of the 112 data points, 74 compounds were correctly identified as nephrotoxic, which corresponds to a sensitivity value of 0.661. Thus, it can be concluded that the LDA c-RASAR model also generalizes well with true external data and efficiently identifies nephrotoxic compounds. The prediction results are presented in **Supplementary Materials SI-1**.

t-SNE analysis of the descriptor and fingerprint spaces

This analysis reflects not only the diversity of the dataset but also how the individual descriptor and fingerprint spaces encode the chemical information. This is analyzed by adopting non-linear dimensionality reduction techniques like the t-SNE [30]. We have subjected our different training and test sets individually encoded by different feature matrices to generate the t-SNE plots using the DataWarrior software (https://openmolecules.org/datawarrior/). **Figure 7** represents the t-SNE plots derived from the molecular descriptors of the training and test sets that were used for QSAR modeling (**Figures 7(A) and 7(B)**) and the corresponding similarity and error-based RASAR descriptors that were used for c-RASAR modeling (**Figures 7(C) and 7(D)**). From the visual representation, one can easily understand how well the RASAR descriptors encode chemical information, reflected in the tight clustering of the data points in **Figures 7(C) and 7(D)**. This highlights the underlying reasons why most of the c-RASAR models had a superior ranking in the SRD analysis, which is a reflection of the robustness and external predictivity of the models.

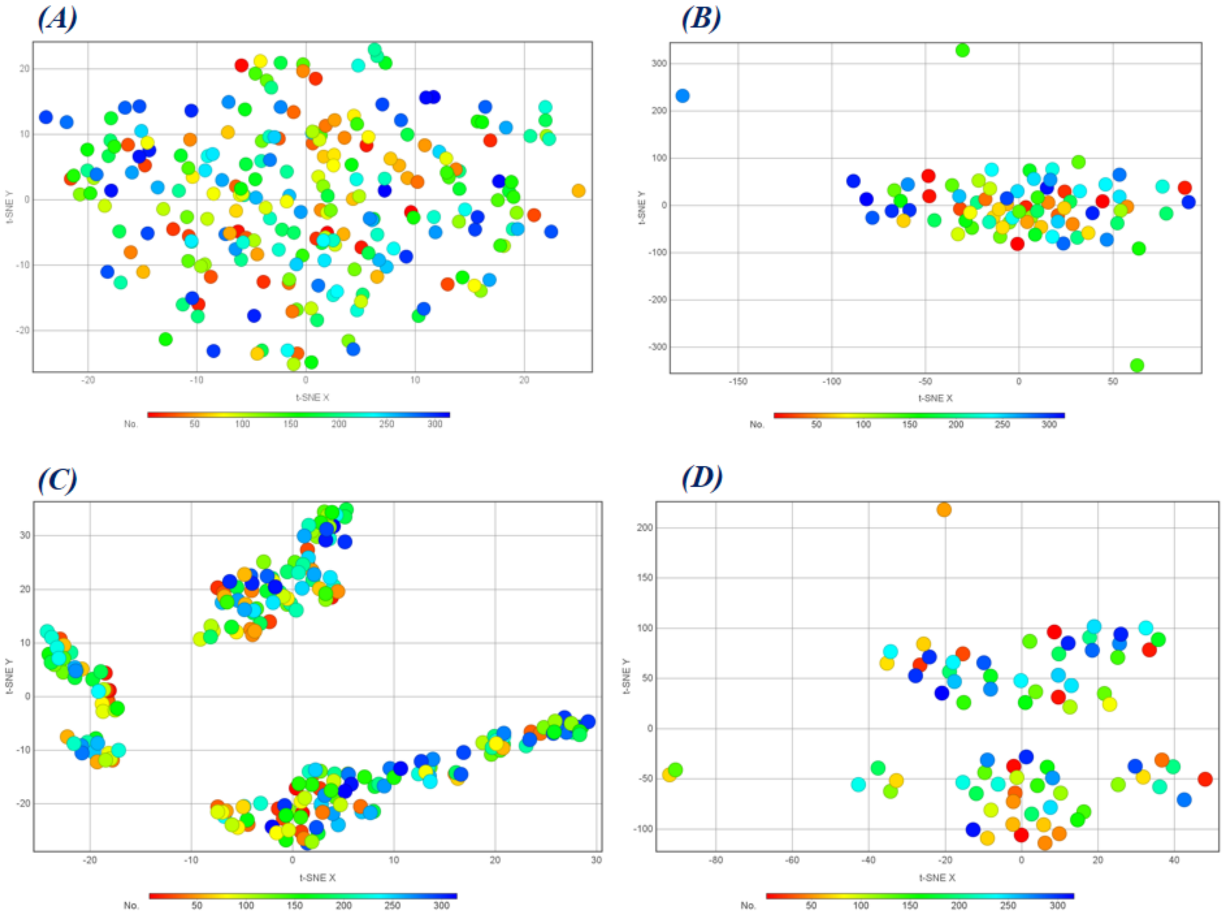


Figure 7. t-SNE plots of the (A) Training set data using the selected molecular descriptors, (B) Test set data using the selected molecular descriptors, (C) Training set data using the corresponding RASAR descriptors developed from descriptor-based feature space, and (D) Test set data using the corresponding RASAR descriptors developed from descriptor-based feature space. These plots highlight how well the RASAR descriptors encapsulate the complete chemical information, as evident from the tight clustering.

On the other hand, **Figure 8** represents the t-SNE plots of the training and test sets of the MACCS fingerprints that were used for QSAR modeling (**Figures 8(A) and 8(B)**) and the corresponding similarity and error-based RASAR descriptors that were used for c-RASAR modeling (**Figures 8(C) and 8(D)**). Similar to the previous case, the c-RASAR descriptors computed from MACCS fingerprints were observed to encode chemical information very efficiently, and one can observe from **Figure 8(C)** how it produces a near-ideal level of clustering of data points in the training set. This is also justified since, in **Figure 3,** it is observed that the MACCS c-RASAR models are the most robust among all the other approaches. Additionally, better clustering is observed in the test set constituting the RASAR descriptors compared to the MACCS fingerprints. Another important point to note is that in all the cases in this current study, the c-RASAR models are developed using only a few modeling descriptors that further justify their potential and statistical reliability.

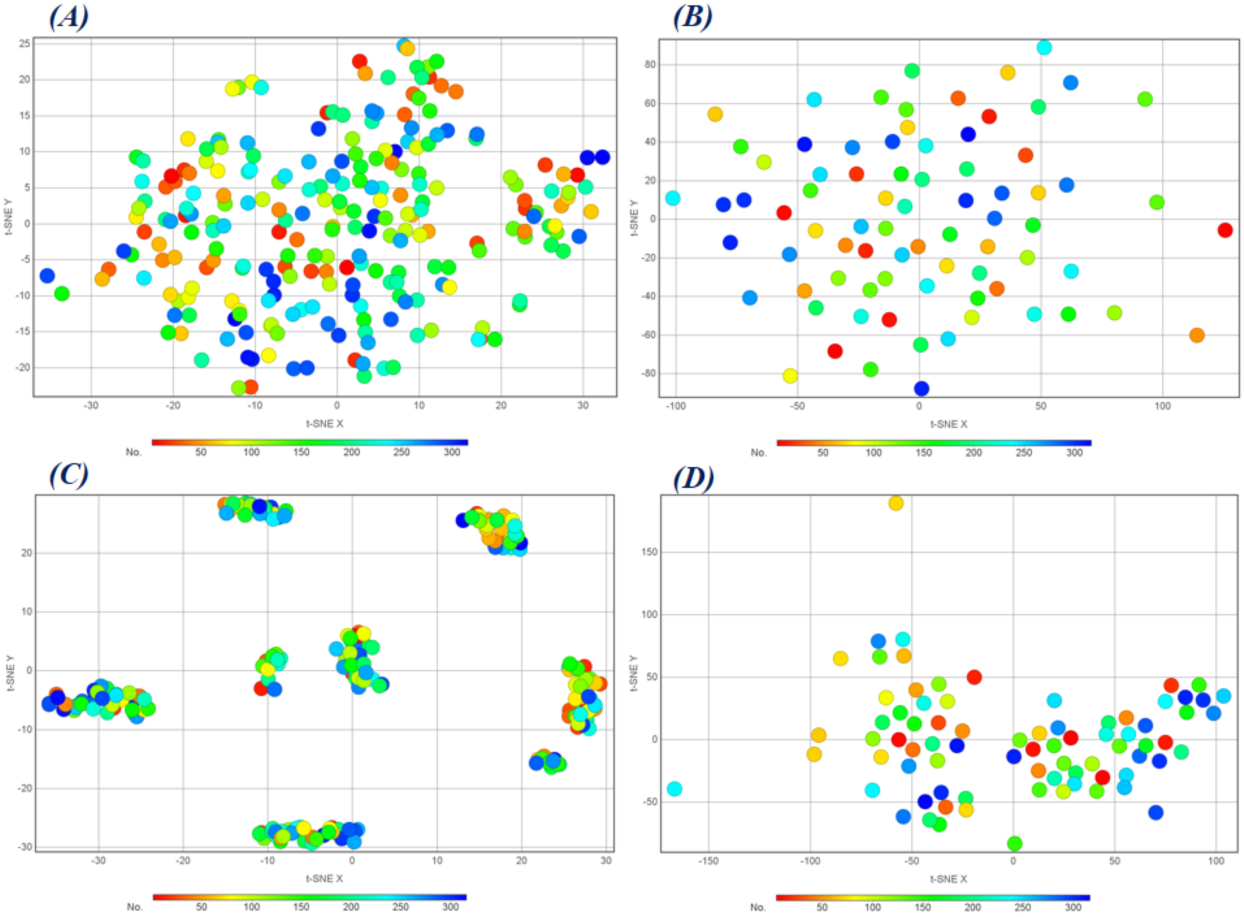


Figure 8. t-SNE plots of the (A) Training set data using the MACCS fingerprints, (B) Test set data using the MACCS fingerprints, (C) Training set data using the corresponding RASAR descriptors developed from fingerprint-based feature space, and (D) Test set data using the corresponding RASAR descriptors developed from fingerprint-based feature space. Like Figure X, these plots also highlight how well the RASAR descriptors encapsulate the complete chemical information, as evident from the tight clustering.

Analysis of the activity cliffs using a supervised algorithm for dimensionality reduction

We discussed above how the RASAR descriptors efficiently encode the chemical information through t-SNE analysis. However, one typical drawback of this approach is its unsupervised nature, which does not allow for the identification of potential activity cliffs. Therefore, we have adopted a supervised dimensionality reduction technique – the ARKA framework for identifying activity cliffs [7]. For this purpose, we have computed the ARKA descriptors, using the tool ARKAdesc-v2.0 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/arithmetic-residuals-in-k-groups-analysis-arka, for the training set that has been defined using the selected 21 molecular descriptors employed for QSAR analysis. Many less confident data points (points in the first and third quadrants) and two distinct activity cliffs (both in the fourth quadrant) were identified. More significant activity cliffs may be determined based on their distance from the origin (which is equivalent to the square root of the sum of the ARKA descriptors), provided positive compounds are located in the fourth quadrant, and negative compounds are located in the second quadrant. Additionally, many data points were in the less modelable region (the central rectangle surrounding the origin) and the borderline zone (±0.5 on either side of the axes). The corresponding ARKA_2 vs ARKA_1 plot has been shown in **Figure 9**.

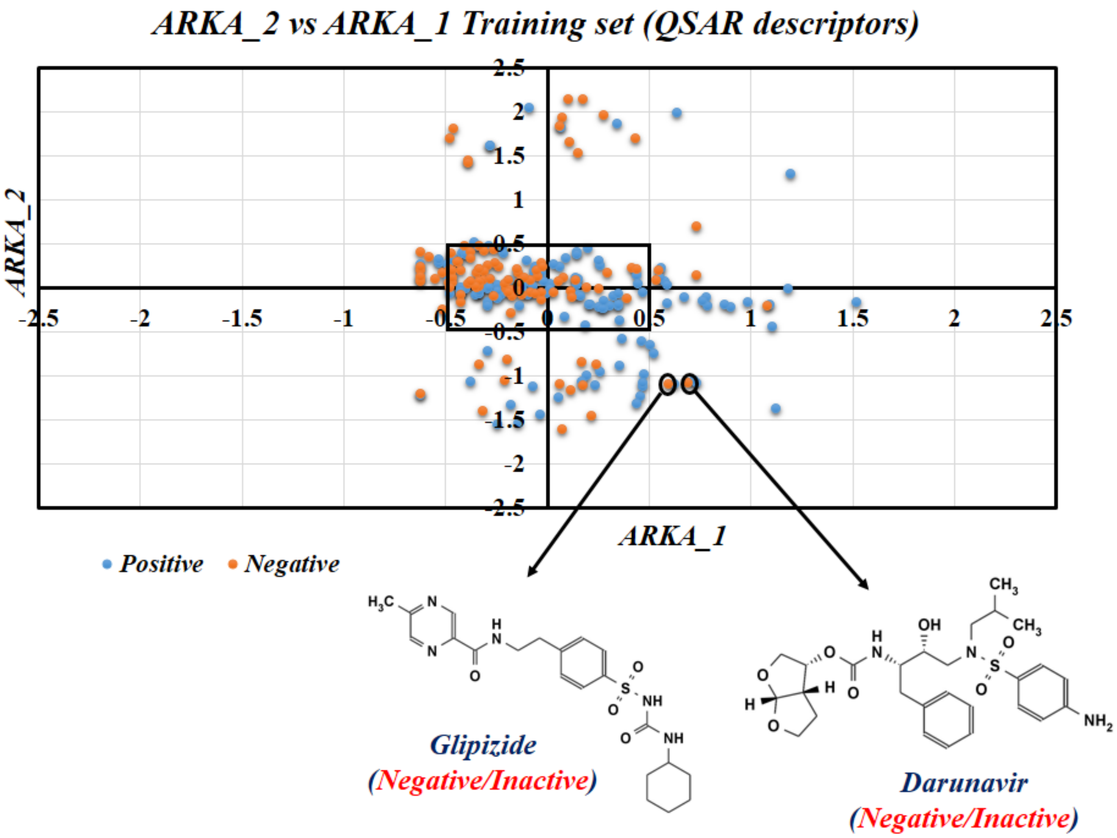


Figure 9. ARKA_2 vs ARKA_1 plot for the training set defined by molecular descriptors. Activity cliffs like Glipizide and Darunavir were identified.

The compound Glipizide, identified as an activity cliff, was reported as non-nephrotoxic. Let's analyze the Gaussian Kernel similarity of this compound with some of its close source congeners (as identified by the Read-Across algorithm of the RASAR descriptor calculator tool). All the similarity levels are very low. Moreover, out of the ten closest source neighbors of Glipizide, eight were in the nephrotoxic class, and only two were in the non-nephrotoxic class, suggesting that this compound has more structural similarity with the nephrotoxic compounds. This explains why Glipizide has been identified as an activity cliff. Darunavir, an antiretroviral drug, has also been identified as an activity cliff from the ARKA analysis. This drug was also labeled as a non-nephrotoxic compound, but all its closest ten nearest neighbors belong to the nephrotoxic class of drugs, which again suggests that this drug has more structural similarity to a nephrotoxic compound. If we observe the predictions from the QDA QSAR model (the best performing QSAR model developed from molecular descriptors), both the compounds have been mispredicted as nephrotoxic compounds, thus implying their activity cliff nature.

So far, the analysis and identification of the activity cliffs have centered on the molecular descriptors employed in QSAR modeling. However, as previously stated in the manuscript and also in [44, 45], the RASAR descriptors encode the chemical information more efficiently, ultimately reducing the number of modeling descriptors. We have computed the ARKA descriptors on the selected 5 RASAR descriptors used to develop the c-RASAR models to explore and identify additional activity cliffs, which the standard molecular descriptors could not identify. It can be observed from the ARKA_2 vs ARKA_1 plot of the training set (**Figure 10**) that a lower number of data points existed inside the central rectangular zone (as compared to **Figure 9**) ultimately infers that the modelability of the dataset has been increased on the application of the RASAR descriptors. Additionally, this plot identifies many activity cliffs (including the previously identified Glipizide and Darunavir) that belong to both the active/positive and inactive/negative classes. The regions enclosed by the two ellipses demonstrate the location of these activity cliffs.

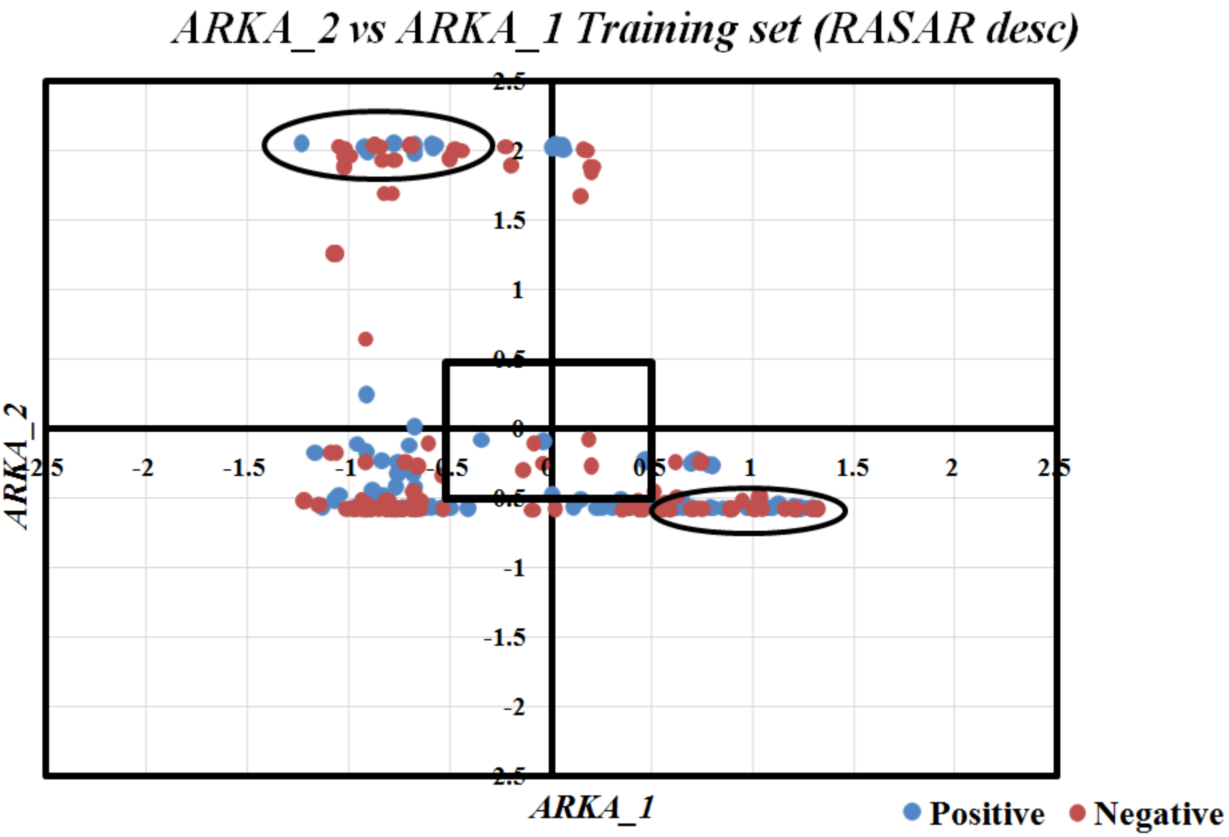
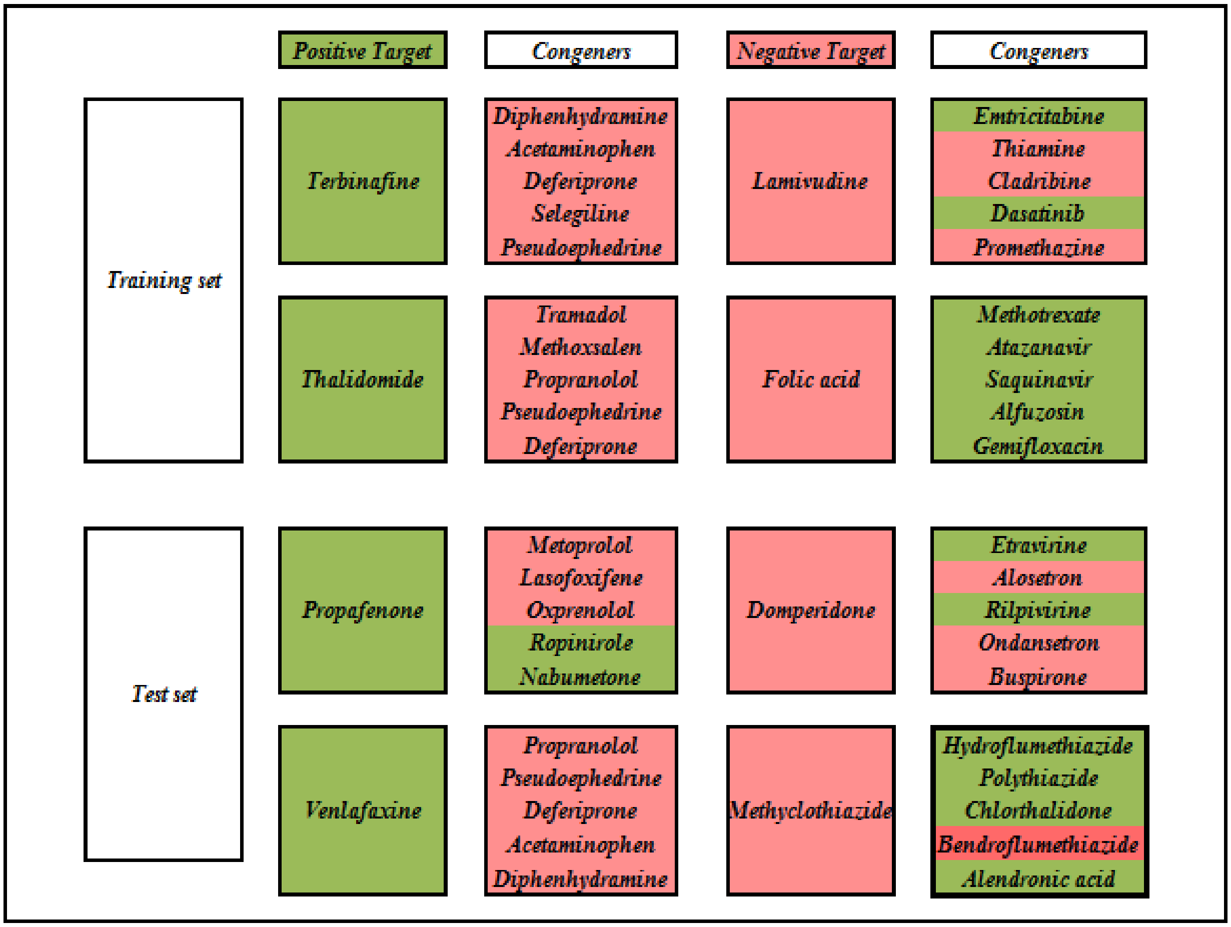


Figure 10. ARKA_2 vs ARKA_1 plot for the training set defined by the RASAR descriptors. The ellipses represent the location of the multiple activity cliffs. It is to be noted that a minimal number of compounds exist in the central rectangular region, inferring the enhancement of the modelability of the application of the RASAR descriptors.

In this case, we analyzed the two most significant activity cliffs from each positive and negative class. Initially, we have identified the compounds in the opposite quadrant (i.e., positive compounds in the second quadrant and negative compounds in the fourth quadrant). Among these compounds, we have computed the Euclidean Distance from the origin using the formula mentioned in **Equation 1**. We have identified four compounds (two each) with the highest Euclidean Distance values from the positive and negative classes, justifying that these compounds are "most confident activity cliffs". Additionally, we have explored the five nearest neighbors from our RASAR analysis and observed that most of these close congeners are from the opposite class of activity. This is pictorially represented in **Figure 11,** where we analyze the activity cliffs from the training and test sets.

$$ED = \sqrt{X^2 + Y^2} \tag{1}$$

ED is the Euclidean Distance, X is the value of ARKA_1, and Y is the value of ARKA_2. The confidence in the activity cliff nature of a compound should increase with its location from the origin, i.e., with an increase in the value of ED, provided it is located in the wrong quadrant, as mentioned above.



	🟩 Positive Target	Congeners	🟥 Negative Target	Congeners
Training set	🟩 Terbinafine	🟥 Diphenhydramine 🟥 Acetaminophen 🟥 Deferiprone 🟥 Selegiline 🟥 Pseudoephedrine	🟥 Lamivudine	🟩 Emtricitabine 🟥 Thiamine 🟥 Cladribine 🟩 Dasatinib 🟥 Promethazine
	🟩 Thalidomide	🟥 Tramadol 🟥 Methoxsalen 🟥 Propranolol 🟥 Pseudoephedrine 🟥 Deferiprone	🟥 Folic acid	🟩 Methotrexate 🟩 Atazanavir 🟩 Saquinavir 🟩 Alfuzosin 🟩 Gemifloxacin
Test set	🟩 Propafenone	🟥 Metoprolol 🟥 Lasofoxifene 🟥 Oxprenolol 🟩 Ropinirole 🟩 Nabumetone	🟥 Domperidone	🟩 Etravirine 🟥 Alosetron 🟩 Rilpivirine 🟥 Ondansetron 🟥 Buspirone
	🟩 Venlafaxine	🟥 Propranolol 🟥 Pseudoephedrine 🟥 Deferiprone 🟥 Acetaminophen 🟥 Diphenhydramine	🟥 Methyclothiazide	🟩 Hydroflumethiazide 🟩 Polythiazide 🟩 Chlorthalidone 🟥 Bendroflumethiazide 🟩 Alendronic acid

Figure 11. Activity cliffs from the positive and negative classes and their five closest neighbors. The green color indicates active/positive compounds, while the red indicates inactive/negative compounds.

Activity cliffs like Terbinafine (Positive), Thalidomide (Positive), Folic acid (Negative), and Venlafaxine (Positive) have all of their five closest neighbors in the opposite class. This infers that these compounds have structural similarities towards their opposite class, which eventually hinders the modelability of the dataset. If we consider the predictions of these compounds by our LDA c-RASAR model, it can be observed that all these compounds have been mispredicted into their opposite class. In the case of compounds like Propafenone (Positive) and Methyclothiazide (Negative), although compounds of the same class exist in the list of closest congeners, it can be observed that a higher fraction of the closest congeners belongs to the opposite class. If we consider their predictions, it was observed that Propafenone was mispredicted as inactive, while Methyclothiazide was mispredicted as active. However, in the cases of Lamivudine (Negative) and Domperidone (Negative), although a higher fraction of the closest source congeners belongs to the same class, it can be observed that the similarity levels to the closest neighbor are high but drastically decreases afterward. This infers that only one compound is close to the target compounds (Lamivudine and Domperidone). In contrast, the other close congeners are located quite far away in terms of their similarities. Therefore, the closest neighbor is the one that describes the propensity of the target compounds towards being active or inactive. If we analyze the closest neighbors of Lamivudine, it has a high similarity value of 0.971 with Emtricitabine, while the similarity level with its second closest compound (Thiamine) is 0.00002. A similar observation was obtained from Domperidone with a similarity level of 0.069 with the closest neighbor Etravirine, while the similarity level with the second closest compound (Alosetron) is only 0.001. As both Emtricitabine (Positive) and Etravirine (Positive) belong to the opposite class of Lamivudine and Domperidone, respectively, it can be concluded that this leads to the misprediction of both the compounds by our LDA c-RASAR model.

Comparison with the previous works

Gong et al. [23] and Shi et al. [24] developed multiple machine-learning models to predict the nephrotoxicity of compounds. However, the works of Connor et al. [26] involved compilation of the data from the two sources and curation adhering to the strategy proposed by Tropsha's group [28]. These authors mapped the molecules to the DrugBank database to identify the drug molecules and further verified with the Anatomical, Therapeutic and Chemical (ATC) index [46, 47] to identify "orally administered drugs" with nephrotoxicity data. These nephrotoxicity data from the two literature sources were cross-checked with sources like the FDA and DrugBankDB to obtain a final list of experimental data. This particular step was crucial since it can be observed from the works of Connor et al. that many molecules had contrasting nephrotoxicity labels in the two different sources (Gong et al. and Shi et al.). In this regard, our model stands out since we have used the fully curated dataset presented by Connor et al. to develop Machine Learning models. This increased reliability of the modeling data used, ultimately increasing the acceptability of our model and its predictions. A detailed comparison report of our work with the works of Gong et al.

and Shi et al. has been presented in **Table 4**, which justifies how this present study is better. Sun et al. [25] also predicted the nephrotoxicity of natural products and drugs. Although not specific for orally active drugs, the poor external validation results showing MCC values of 0.000 and 0.089 of the ANN and SVM models respectively justify that the models did not generalize well with the test set data. This is not the case of our LDA c-RASAR model, as its external predictivity is quite good where MCC value is 0.431. Therefore, we can infer that our LDA c-RASAR model is superior in terms of reliability and prediction quality to predict the nephrotoxicity of orally administered drugs.

Table 4. Comparison of our work with the works of Gong et al. and Shi et al.

Parameters	***Gong et al. [23]***	***Shi et al. [24]***	***Our work***
Dataset strictly focusing on drug molecules (more specifically, orally active drug molecules)	***No***	***No***	***Yes***
True external set prediction	***Yes***	***No***	***Yes***
Size of the true external set	***Lower (n=71)***	***-***	***Higher (n=112)***
Activity cliffs analysis	***No***	***No***	***Yes***
Statistical tests (using multi-criteria decision-making approaches) for the identification of best models	***No***	***No***	***Yes (by using the Sum of Ranking Differences approach)***
The presence of conflicting data labels for the same compounds reduces the reliability of the models	***Yes (e.g., Aspirin has been labelled as Nephrotoxic)***	***Yes (e.g., Aspirin has been labelled as Non-nephrotoxic)***	***No, since we developed models on the curated dataset presented by Connor et al.***
Rigorous cross-validation	***Yes***	***Yes***	***Yes***

Conclusion

Drug-induced nephrotoxicity is an area of concern since our kidneys are associated with the removal of toxic substances and metabolites from the blood. A lot of drugs that we take orally for treating specific ailments are often silently associated with producing nephrotoxicity. Since experimental identification is tedious and involves ethical complications, we have developed Machine Learning (ML) models to easily screen drugs, identifying their potential nephrotoxicity when administered orally. We have used a highly curated data set of orally active drugs for the reliability of the developed models. Simple and interpretable 0-2D molecular descriptors and MACCS fingerprints were used to develop ML models. We have also developed ML c-RASAR models on the feature spaces encoded by the selected 0-2D descriptors and MACCS fingerprints to incorporate similarity-based considerations. This resulted in the enhancement of robustness and predictivity of the c-RASAR models, justifying the more efficient and concise use of the chemical information of the close source neighbors. All the developed QSAR and c-RASAR models were subjected to statistical comparison using the Sum of Ranking Differences (SRD) approach, considering factors associated with robustness and external predictivity. This analysis suggested that the LDA c-RASAR model, developed from the feature space of the molecular descriptors, was the best-performing model among 36 different linear and non-linear ML models. Once again, this infers the successful incorporation of the non-linear information into a linear modeling framework by the RASAR descriptors, which results in linear models having enhanced performance than non-linear models. To assess the model performance on true external set data, we have used the LDA c-RASAR model to predict the nephrotoxicity of the approved drugs from the DrugBankDB. The successful results on the true external set justify the reliability of our simple model. This LDA c-RASAR model can thus be used for quick and efficient prediction of nephrotoxicity for orally active drugs. The efficient identification of activity cliffs by the ARKA analysis and the tight and distinct clustering observed in the t-SNE plots exactly points out the true potential of the c-RASAR approach in not only optimizing the utilization of the feature space but also in the identification of activity and prediction cliffs.

Acknowledgment

AB thanks the Life Sciences Research Board, DRDO, New Delhi for a senior research fellowship.

Author information

Corresponding author.

Kunal Roy, Drug Theoretics and Cheminformatics Laboratory**,** Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India. Email: kunal.roy@jadavpuruniversity.in . ORCID: https://orcid.org/0000-0003-4486-8074

First author.

Arkaprava Banerjee, Drug Theoretics and Cheminformatics Laboratory**,** Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India. ORCID: https://orcid.org/0000-0001-8468-0784

Data and Software availability

The source data used to develop the models reported in this paper are available in the **Supplementary Materials SI-1**. The RASAR descriptors and their significance has been tabulated in **Supplementary Material SI-2** The software used for the Read-Across predictions and the computation of the RASAR descriptors and ARKA descriptors is freely available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home and https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/arithmetic-residuals-in-k-groups-analysis-arka.

Supplementary Information available

Supplementary Information SI-1 contains the data set, computed descriptors for training and test sets, and prediction results for the true external set.

Supplementary Information SI-2 contains the list of RASAR descriptors.

Declaration of interest

The authors declare no competing interests.

Author contributions

AB: Data curation, Formal analysis, validation, software, Writing - Initial draft

KR: Conceptualization, Funding acquisition, Supervision, Writing - Editing

Funding

This research has been funded by the Life Sciences Research Board, DRDO, New Delhi (LSRB/01/15001/M/LSRB-394/SH&DD/2022).

References

1. Kulkarni, P. Prediction of drug-induced kidney injury in drug discovery. Drug Metabol. Rev. 2021, 53, 234-244.
2. Redfern, W.S.; Ewart, L.; Hammond, T.G.; Bialecki, R.; Kinter, L.; Lindgren, S.; Pollard, C.E.; Roberts, R.; Rolf, M.G.; Valentin, J.P. Impact and frequency of different toxicities throughout the pharmaceutical life cycle. Toxicologist. 2010, 114, 231.
3. Choudhury, D.; Ahmed, Z. Drug-associated renal dysfunction and injury. Nat. Clin. Pract. Nephrol. 2006, 2, 80-91.
4. Gai, Z.; Gui, T.; Kullak-Ublick, G.A.; Li, Y.; Visentin, M. The role of mitochondria in drug-induced kidney injury. Front. Physiol. 2020, 11, 1079.
5. Hansch, C.; Hoekman, D.; Gao, H. Comparative QSAR: Toward a Deeper Understanding of Chemicobiological Interactions. Chem. Rev. 1996, 96, 1045-1076.
6. Gini, G. QSAR methods. In: Benfenati, E.(Eds.) In Silico Methods for Predicting Drug Toxicity. 2022, Springer, NY.
7. Banerjee, A.; Roy, K. ARKA: a framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data. Environ. Sci.: Processes Impacts 2024, 26, 991-1007.

8. Gajewicz, A. What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. Nanoscale 2017, 9, 8435-8448.

9. Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz-Skretna, A.; Roy, K. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environ. Sci.: Nano 2022, 9, 189-203.

10. Manganelli, S.; Benfenati, E. Use of Read-Across tools. In: Benfenati, E.(Eds.) In Silico Methods for Predicting Drug Toxicity. 2016, Springer, NY.

11. Banerjee, A.; Roy, K. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Mol. Divers. 2022, 26, 2847-2862.

12. Banerjee, A.; Roy, K. On Some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points. Chem. Res. Toxicol. 2023, 36, 446-464.

13. Banerjee, A.; Roy, K. 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. 2023, 36, 1518-1531.

14. Banerjee, A.; Roy, K. Read-across-based intelligent learning: development of a global q-RASAR model for the efficient quantitative predictions of skin sensitization potential of diverse organic chemicals. Environ. Sci.: Processes Impacts 2023, 25, 1626-1644.

15. Wang, Y.; Wang, P.; Fan, T.; Ren, T.; Zhang, N.; Zhao, L.; Zhong, R.; Sun, G. From molecular descriptors to the developmental toxicity prediction of pesticides/veterinary drugs/bio-pesticides against zebrafish embryo: Dual computational toxicological approaches for prioritization. J. Hazard. Mater. 2024, 476, 134945.

16. Jiang, J.; Cai, W.; Chen, Z.; Liao, X.; Cai, Z. Prediction of acute toxicity for Chlorella vulgaris caused by tire wear particle-derived compounds using quantitative structure-activity relationship models. Water Res. 2024, 256, 121643.

17. Kumar, V.; Banerjee, A.; Roy, K. Breaking the barriers: Machine-learning-based c-RASAR approach for accurate blood–brain barrier permeability prediction. J. Chem. Inf. Model. 2024, 64, 4298-4309.

18. Pandey, S.K.; Roy, K. Development of a read-across-derived classification model for the predictions of mutagenicity data and its comparison with traditional QSAR models and expert systems. Toxicology 2023, 500, 153676.
19. Banerjee, A.; De, P.; Kumar, V.; Kar, S.; Roy, K. Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across. Chemosphere 2022, 309, 136579.
20. Banerjee, A.; Roy, K. Machine-learning-based similarity meets traditional QSAR: “q-RASAR” for the enhancement of the external predictivity and detection of prediction confidence outliers in an hERG toxicity dataset. Chemom. Intell. Lab. Syst. 2023, 237, 104829.
21. Varsou, D-D.; Banerjee, A.; Roy, J.; Roy, K.; Savvas, G.; Sarimveis, H.; Wyrzykowska, E.; Balicki, M.; Puzyn, T.; Melagraki, G.; Lynch, I.; Afantitis, A. The Round Robin approach applied to nanoinformatics: consensus prediction of nanomaterials zeta potential. Beil. Arch. 2024, 33. https://doi.org/10.3762/bxiv.2024.33.v1
22. Banerjee, A.; Roy, K. How to correctly develop q-RASAR models for predictive cheminformatics. Expert Opin. Drug Discov. 2024.
23. Gong, Y.; Teng, D.; Wang, Y.; Gu, Y.; Wu, Z.; Li, W.; Tang, Y.; Liu, G. In silico prediction of potential drug-induced nephrotoxicity with machine learning methods. J. Appl. Toxicol. 2022, 42, 1639-1650.
24. Shi, Y.; Hua, Y.; Wang, B.; Zhang, R.; Li, X. In silico prediction and insights into the structural basis of drug induced nephrotoxicity. Front. Pharmacol. 2022, 12, 793332.
25. Sun, Y.; Shi, S.; Li, Y.; Wang, Q. Development of quantitative structure-activity relationship models to predict potential nephrotoxic ingredients in traditional Chinese medicines. Food Chem. Toxicol. 2019, 128, 163-170.
26. Connor, S.; Li, T.; Qu, Y.; Roberts, R.A.; Tong, W. Generation of a drug-induced renal injury list to facilitate the development of new approach methodologies for nephrotoxicity. Drug Discov. Today 2024, 29, 103938.
27. Racz, A.; Bajusz, D.; Heberger, K. Multi-Level comparison of machine learning classifiers and their performance metrics. Molecules 2019, 24, 2811.

28. Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. J. Chem. Inf. Model. 2010, 50, 1189-1204.
29. Mauri, A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In: Roy, K. (Eds.) Ecotoxicological QSARs, 2020, Springer, NY.
30. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 2008, 9, 2579-2605.
31. Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F.J.; Salabert-Salvador, M.T.; Diaz-Villanueva, W.; Castro-Bleda, M.J.; Villanueva-Pareja, A. Artificial neural networks and linear discriminant analysis: A valuable combination in the selection of new antibacterial compounds. J. Chem. Inf. Comput. Sci. 2004, 44, 1031-1041.
32. Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B. Linear Discriminant Analysis. In: Robust Data Mining. SpringerBriefs in Optimization. 2013, Springer, New York, NY.
33. Lau, K.W.; Wu, Q.H. Online training of support vector classifier. Patt. Recog. 2003, 36, 1913-1920.
34. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5-32.
35. Stoltzfus, J.C. Logistic regression: A brief primer. Aca. Emer. Med. 2011, 18, 1099-1104.
36. Srivastava, S.; Gupta, M.R.; Frigyik, B.A. Bayesian quadratic discriminant analysis. J. Mach. Learn. Res. 2007, 8, 1277-1305.
37. Chaudhuri, B.B.; Bhattacharya, U. Efficient training and improved performance of multilayer perceptron in pattern classification. Neurocomputing 2000, 34, 11-27.
38. Ontivero-Ortega, M.; Lage-Castellanos, A.; Valente, G.; Goebel, R.; Valdes-Sosa, M. Fast Gaussian Naïve Bayes for searchlight classification analysis. NeuroImage 2017, 163, 471-479.
39. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. Front. Neurorobot. 2013, 7, 21.
40. Wang, R. AdaBoost for feature selection, classification and its relation with SVM, a review. Phys. Procedia 2012, 25, 800-807.
41. Fushiki, T. Estimation of prediction error by using *K*-fold cross-validation. Stat. Comput. 2011, 21, 137–146.

42. Pandey, S.K.; Banerjee, A.; Roy K. Machine learning-based q-RASPR predictions of detonation heat for nitrogen-containing compounds. Mater. Adv. 2023, 4, 5797-5807.

43. Banerjee, A.; Kar, S.; Pore, S.; Roy K. Efficient predictions of cytotoxicity of TiO2-based multi-component nanoparticles using a machine learning-based q-RASAR approach. Nanotoxicology 2023, 17, 78-93.

44. Roy, K.; Banerjee, A. q-RASAR. A Path to Predictive Cheminformatics, 2024, Springer, NY

45. Banerjee, A.; Kar, S.; Roy, K.; Patlewicz, G.; Charest, N.; Benfenati, E.; Cronin, M.T.D. Molecular similarity in chemical informatics and predictive toxicity modeling: from quantitative read-across (q-RA) to quantitative read-across structure–activity relationship (q-RASAR) with the application of machine learning. Crit. Rev. Toxicol. 2024, https://doi.org/10.1080/10408444.2024.2386260

46. World Health Organization (WHO) Anatomical therapeutic chemical (ATC) classification index with defined daily doses (DDDs). Oslo: WHO Collaborating Centre for Drug Statistics Methodology. 2000:20.

47. World Health Organization (WHO) collaborating centre for drug statistics methodology. Guidelines for ATC classification and DDD assignment. Norwegian Institute of Public Health; 2021. 2022.