Development of a Global q-RASAR Model for the Efficient Quantitative Predictions of Skin Sensitization Potential of Diverse Organic Chemicals

Presented by:

Arkaprava Banerjee

Drug Theoretics and Cheminformatics Laboratory,
Jadavpur University, Kolkata, INDIA
https://sites.google.com/site/kunalroyindia

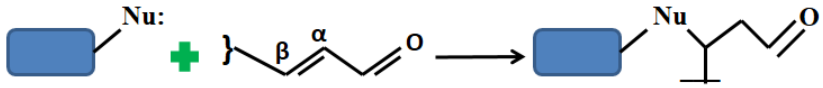
Introduction

Environmental contaminants are capable of producing a wide array of toxic effects including skin sensitization

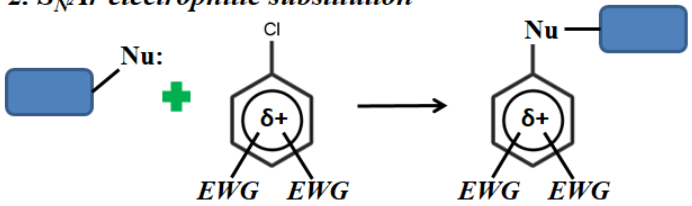
Various QSAR models have been developed but most of them are based on compounds eliciting skin sensitization by a particular reaction mechanism

This study aims at developing a “global” q-RASAR model built using diverse structures irrespective of their reaction mechanisms

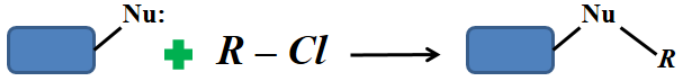
1. Michael addition reaction



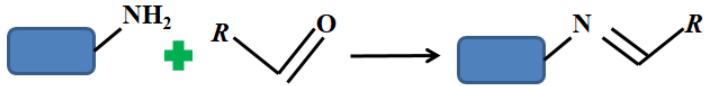
2. S_NAr electrophilic substitution



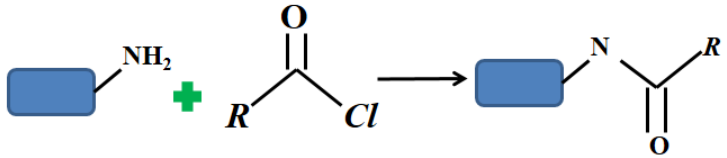
3. S_N2 electrophilic substitution

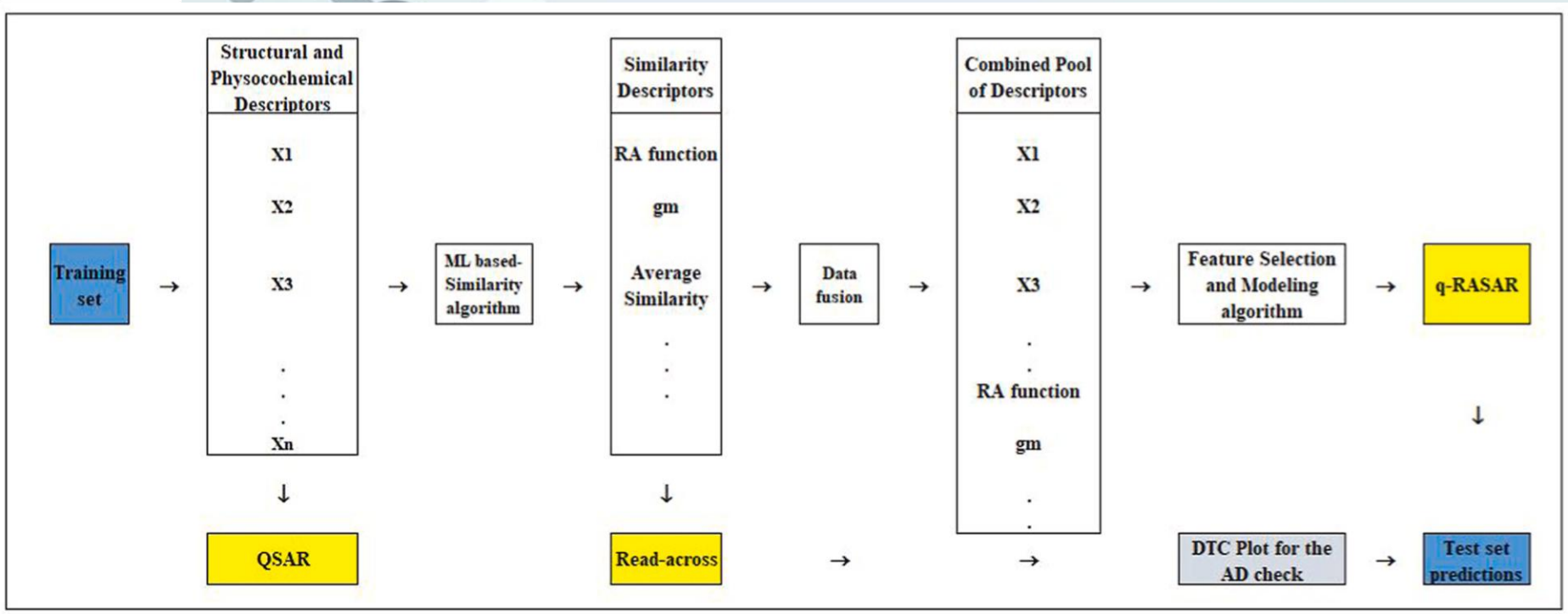


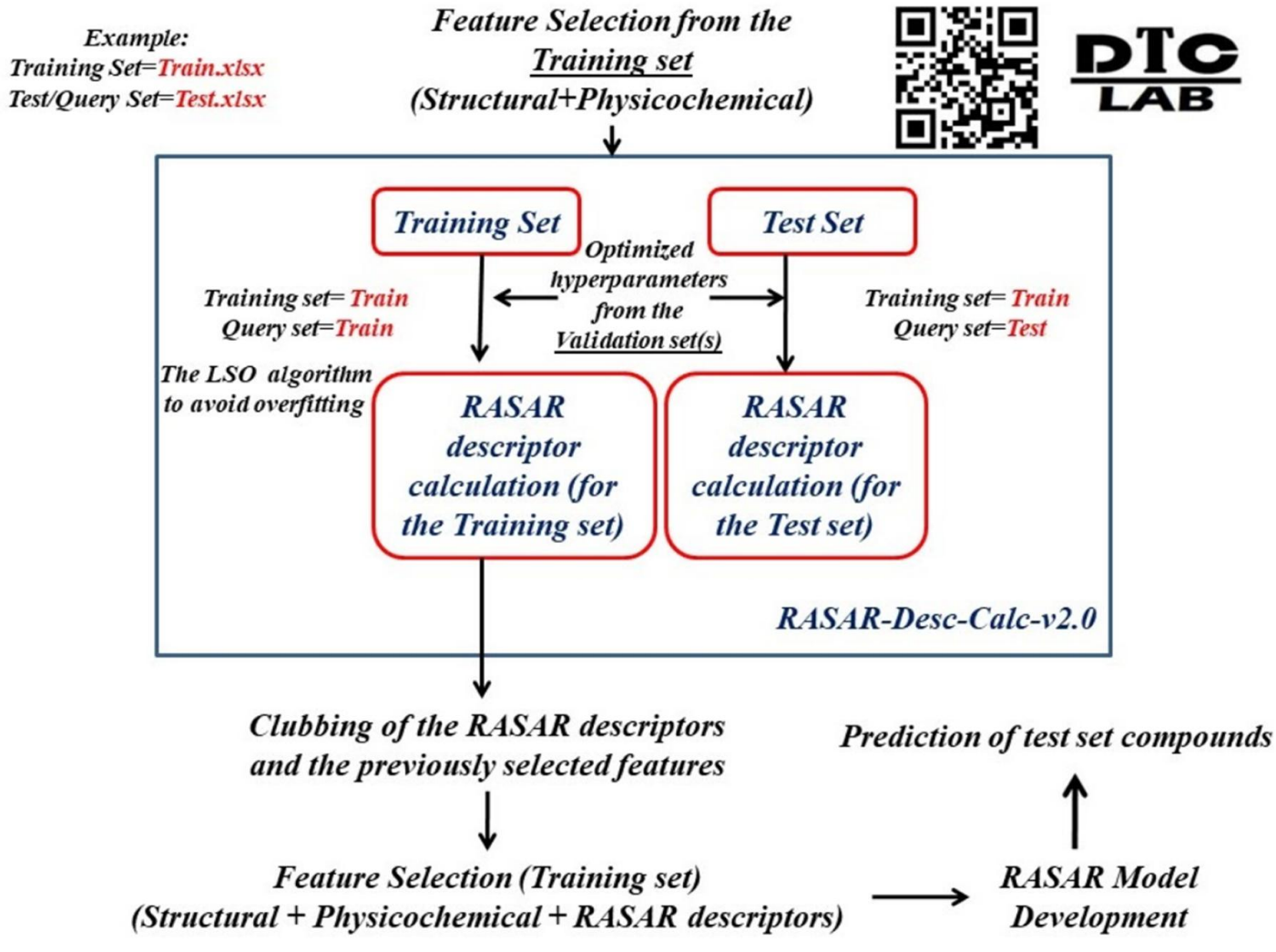
4. Schiff base formation



5. Acylating agent




Training set
Structural and Physocochemical Descriptors
X1
X2
X3
Xn
QSAR
ML based-Similarity algorithm
Similarity Descriptors
RA function
gm
Average Similarity
Read-across
Data fusion
Combined Pool of Descriptors
X1
X2
X3
RA function
gm
Feature Selection and Modeling algorithm
DTC Plot for the AD check
q-RASAR
Test set predictions

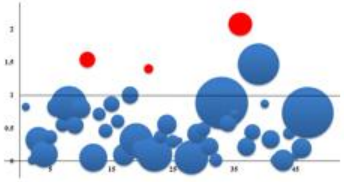

Example:
Training Set=Train.xlsx
Test/Query Set=Test.xlsx
Feature Selection from the Training set (Structural+Physicochemical)
DTC LAB
Training Set
Test Set
Optimized hyperparameters from the Validation set(s)
Training set= Train
Query set=Train
Training set= Train
Query set=Test
The LSO algorithm to avoid overfitting
RASAR descriptor calculation (for the Training set)
RASAR descriptor calculation (for the Test set)
RASAR-Desc-Calc-v2.0
Clubbing of the RASAR descriptors and the previously selected features
Feature Selection (Training set) (Structural + Physicochemical + RASAR descriptors)
RASAR Model Development
Prediction of test set compounds

Collection of skin sensitization data

RR, LSVR, SVR, RF, GB, Adaboost, XGB, kNN, MLP

PLS q-RASAR Model

Predictions



Detection of the Prediction Confidence Outliers

Curation of the data, descriptor calculation and Data Pre-Treatment

Java

Skin_Sensitizer_Calculator-v1.0

Predictions

Data fusion and Feature selection

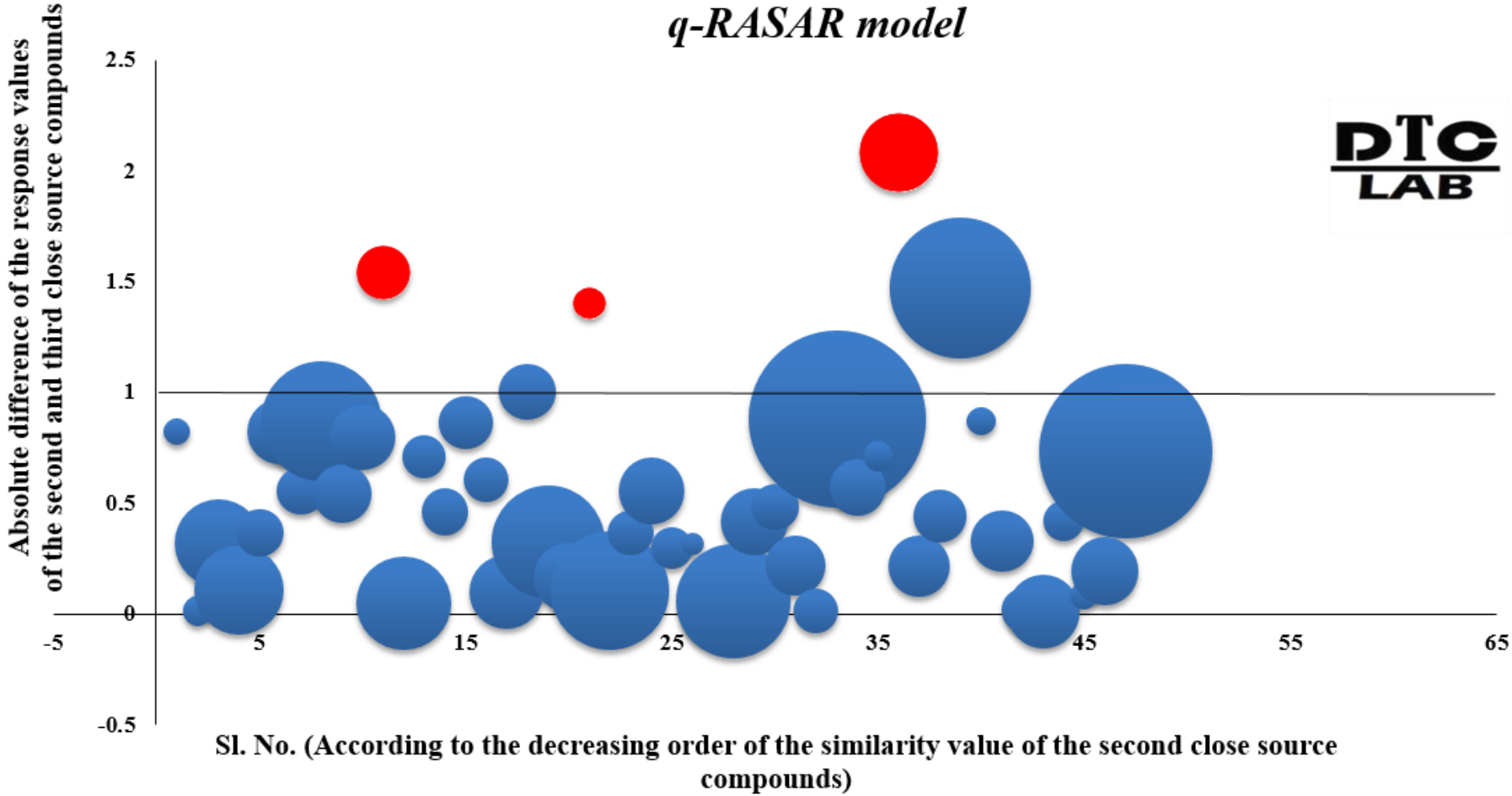
Dataset division, Feature selection and Model development

DTC_QSAR v1.0.7

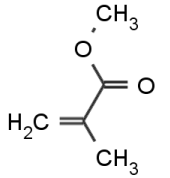
Optimization of the Read-Across hyperparameters

Computation of the RASAR descriptors

DTC Plot to determine the prediction confidence outliers of the PLS q-RASAR model

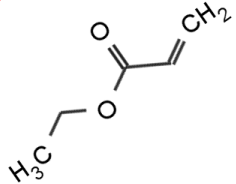


Query compound

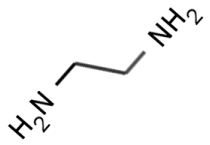


Methyl methacrylate
pEC_3 =0.046

Close source neighbors



Ethyl acrylate
pEC_3 = 0.553



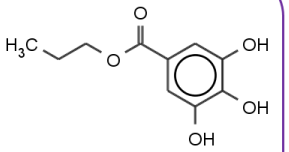
Ethylenediamine
pEC_3 = 1.436

(a)

Similarity values	*0.977*	*0.900*
△ Response	*0.883*	

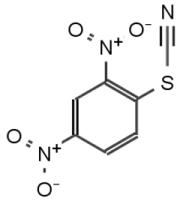
DTC Plot
- *Y-axis = |0.553-1.436|= 0.883*
- *Bubble diameter = |0.977-0.900| =0.077*

DTC Plot Outlier

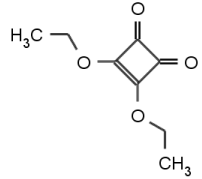


Propyl gallate (274)
pEC_3 = 2.822

Close source neighbors



2,4-Dinitrothiocyanatobenzene
pEC_3 = 3.680



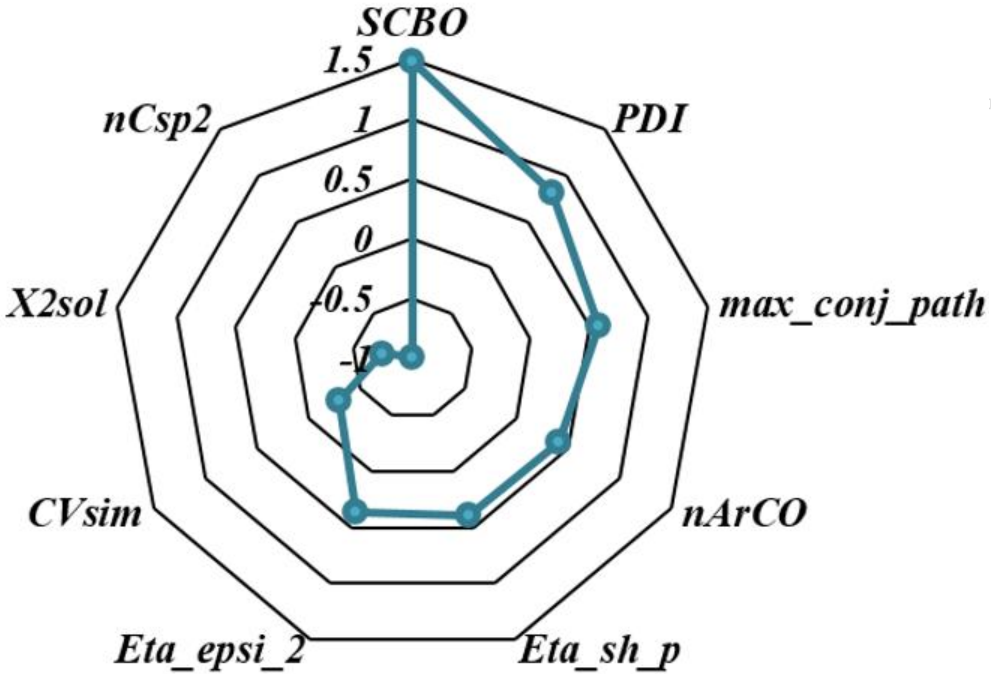
Squaric acid diethyl ester
pEC3 = 2.277

(b)

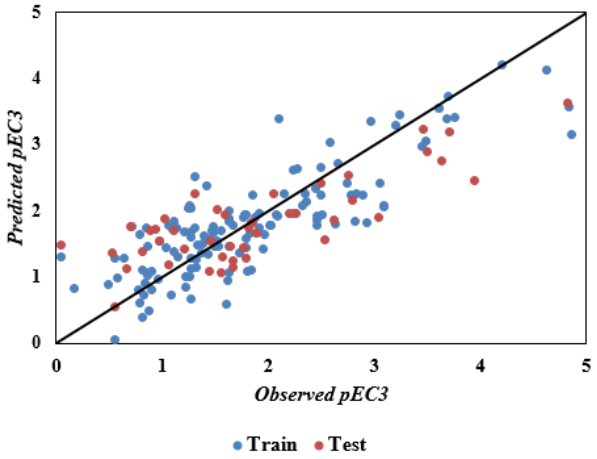
Similarity values	*0.989*	*0.986*
△Response	*1.403*	

Identification of prediction confidence outliers

Radar Plot of the PLS q-RASAR Model

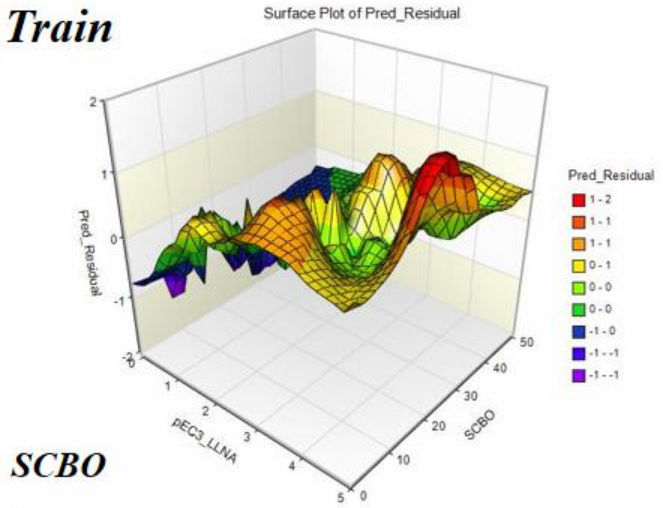


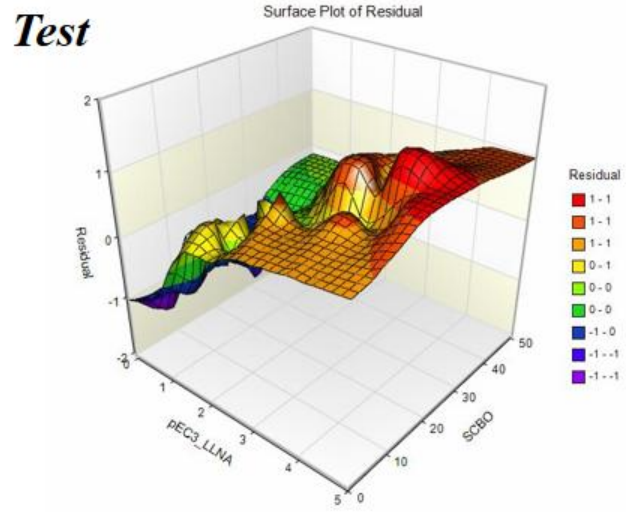
Scatter Plot of the PLS q-RASAR model

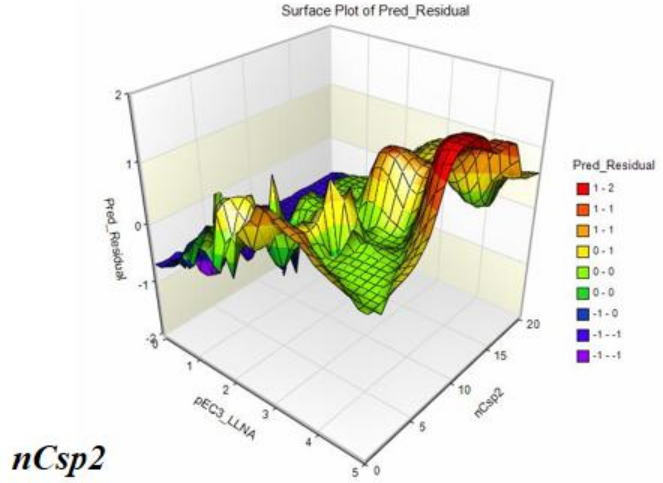


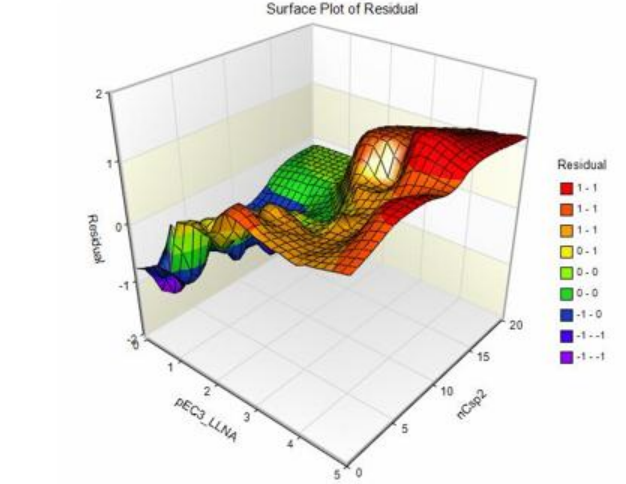
$nTrain = 133$ $R^2_{Train}=0.695$ $Q^2_{LOO}=0.649$ $MAE_{Train}=0.406$

$nTest = 44$ $Q^2_{F1}=0.607$ $Q^2_{F2}=0.606$ $MAE_{Test}=0.523$



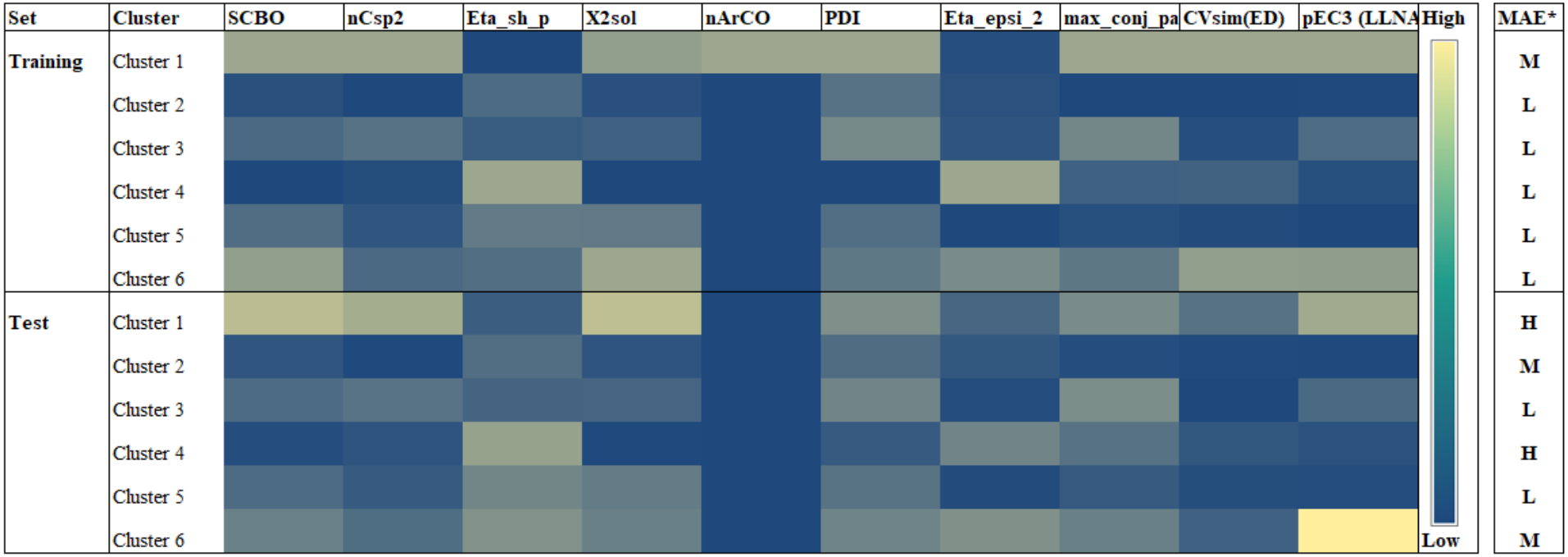




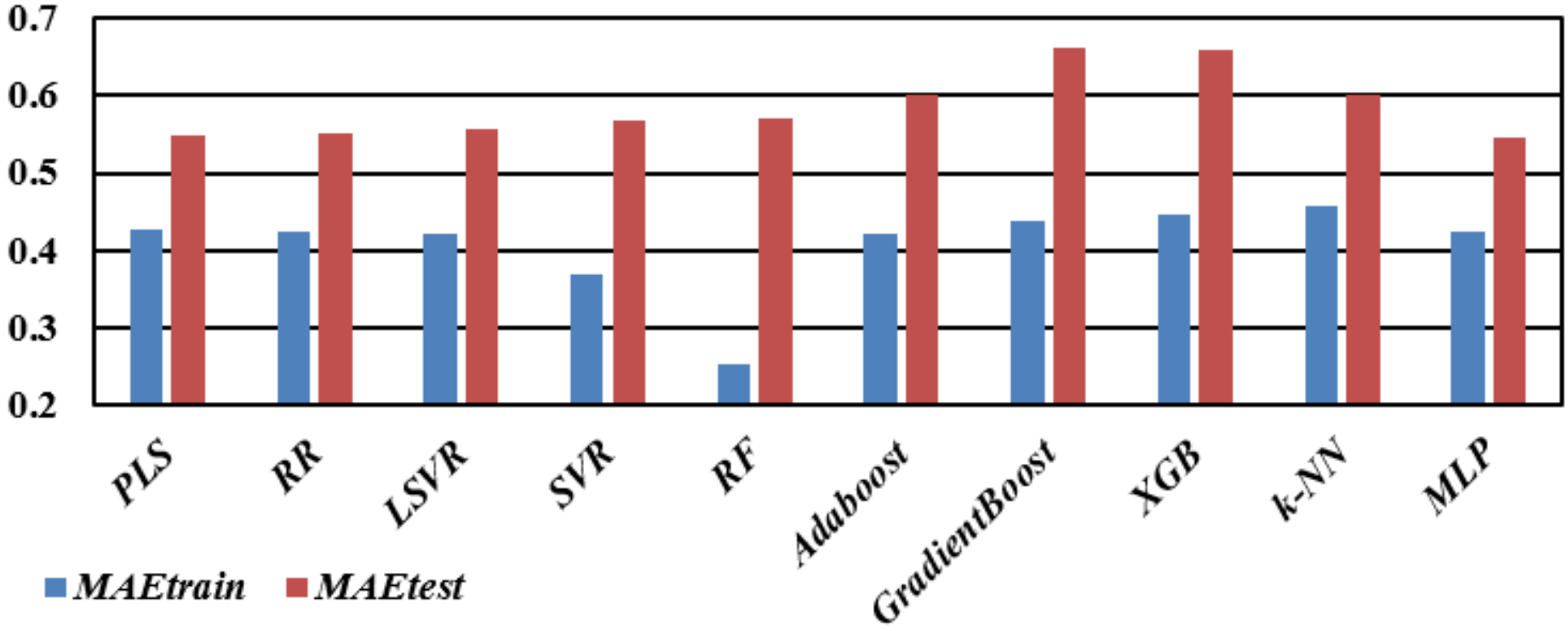


3D-Surface Plots to correlate under-prediction and over-prediction with the level of toxicity and variation of important descriptor values

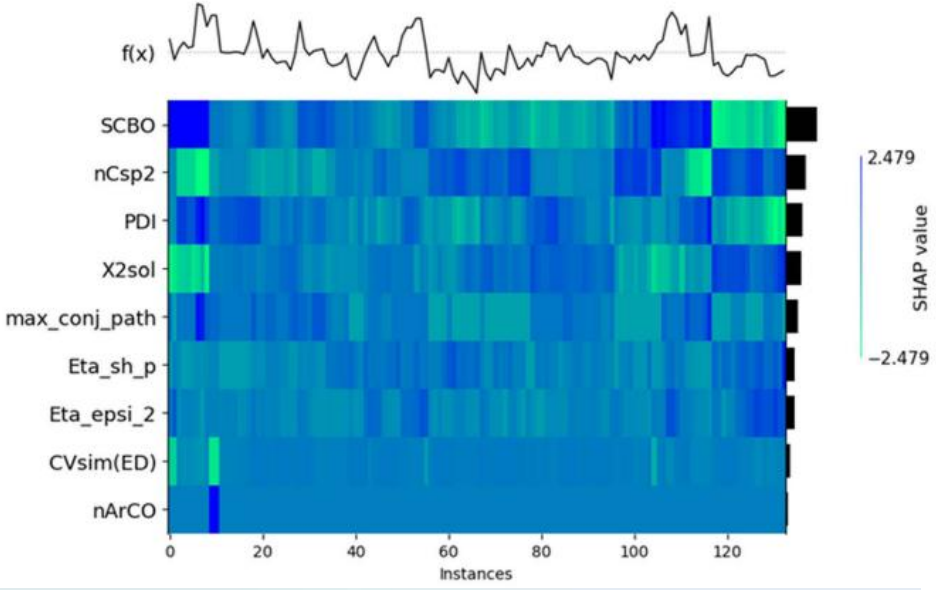
Predictions generated by the model in different clusters – an assessment of the "global" performance



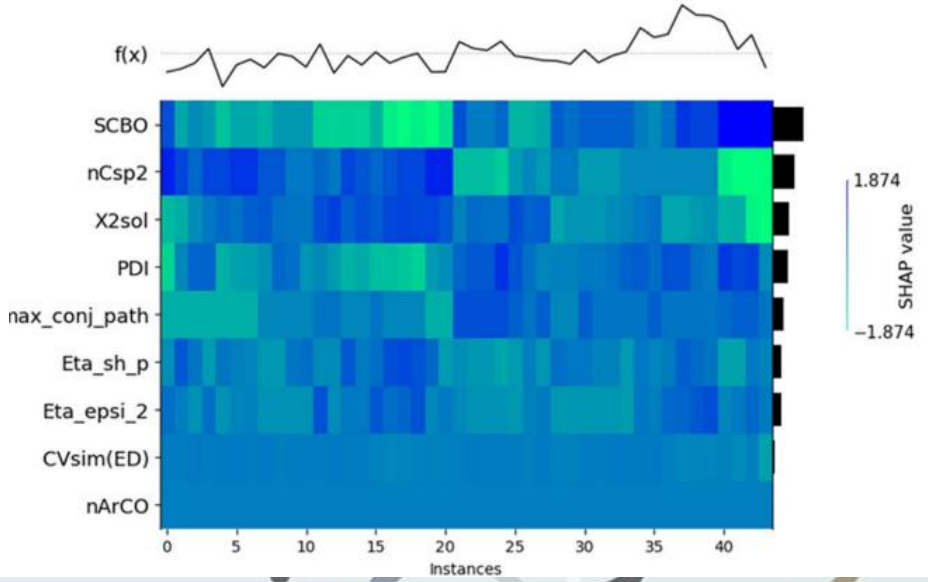
Comparison of predictivity among different ML models



SHAP Analysis to identify the important features of the Multilayer Perceptron q-RASAR Model



Training



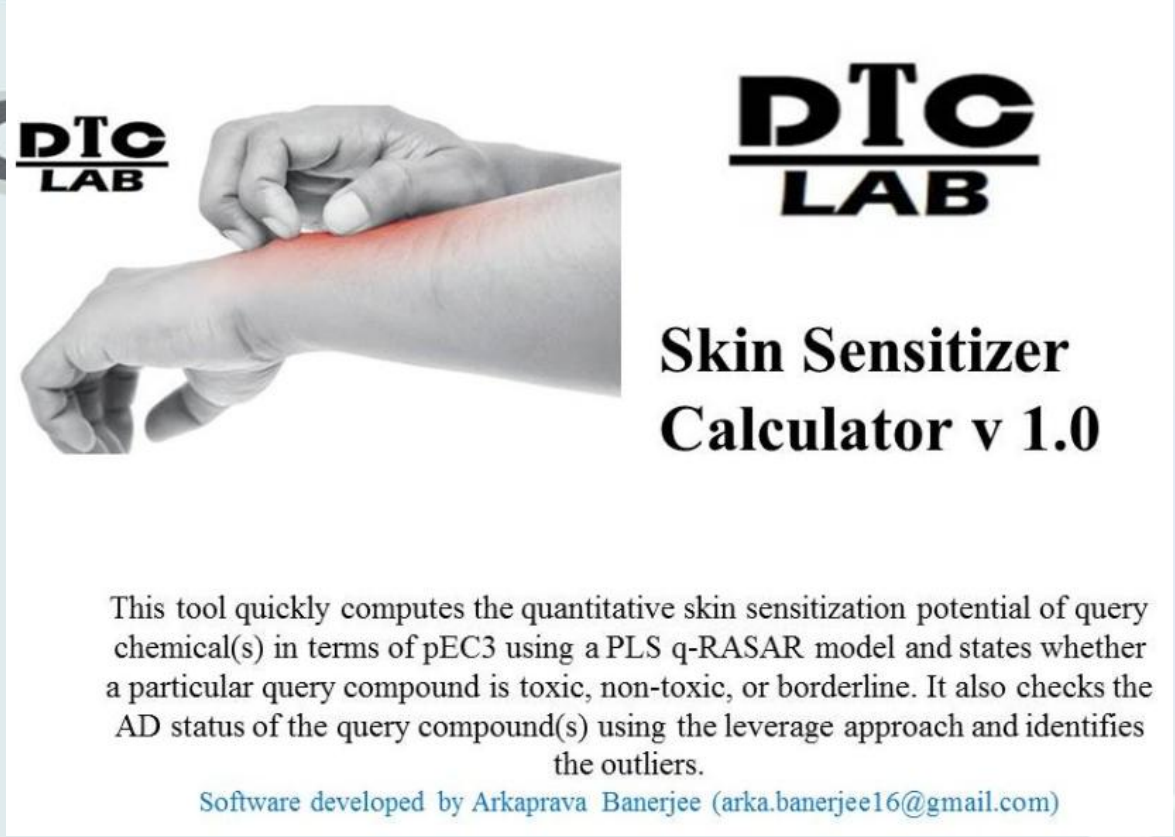
Test

Comparison with the previous QSAR models

	Training set statistics				Test set statistics				
Models	*n*Train	R_{Train}^2	Q_{Loo}^2	MAE_{Train}	*n*Test	Q_{F1}^2	Q_{F2}^2	MAE_{Test}	Comment
Nandy *et al.* 2013 (ref. 58)	37/38	0.710/0.720	0.608/0.580	—	14/13	0.689/0.580	—	—	The dataset size is small; it used 3D descriptors which requires optimization
Dearden *et al.* 2015 (ref. 60)	204	0.496	0.459	—	—	—	—	—	External validation not reported
Chayawan *et al.* 2022 (ref. 68)	—	—	—	—	—	—	—	—	Developed various QSAR models based on different reaction mechanisms
Manhas *et al.* 2022 (ref. 69)	20	0.698	0.598	—	10	0.594	—	—	The dataset size is small; performed only using Michael acceptors
Our PLS QSAR model	133	0.696	0.644	0.410	47	0.526	0.524	0.562	It used 2D descriptors; considered diverse structures; reproducible, transferable
Our PLS q-RASAR model	133	0.695	0.649	0.406	44[a]	0.607	0.606	0.523	It used 2D and similarity-based descriptors; considered diverse structures; reproducible, transferable

[a] Three compounds were omitted due to their prediction outlier nature as per the DTC plot.

Java-based expert system



Available from:

https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home

Conclusion

- ✓ Simple, reproducible, transferable, robust and predictive “global” PLS q-RASAR model has been presented
- ✓ Quality of the PLS q-RASAR model supersedes the corresponding developed PLS QSAR model and previously reported models by other research groups
- ✓ The Multilayer Perceptron q-RASAR model showed the highest predictivity (although marginal enhancement from the PLS q-RASAR model)
- ✓ A Java-based expert system has been developed for the quick and easy prediction of skin sensitizing activity

References

1. A. Nandy et al. SAR QSAR Environ. Res. 2013, 24, 1009-1023
2. J. C. Dearden et al. Chem. Res. Toxicol. 2015, 28, 1975-1986
3. A. Banerjee and K. Roy Mol. Divers. 2022, 26, 2847-2862
4. A. Banerjee and K. Roy Chemom. Intell. Lab. Syst. 2023, 237, 104829
5. A. Banerjee and K. Roy Chem. Res. Toxicol. 2023, 36, 446-464

Environmental Science Processes & Impacts

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

DTC LAB

Check for updates

Cite this: DOI: 10.1039/d3em00322a

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†

Arkaprava Banerjee and Kunal Roy*

Environmental chemicals and contaminants cause a wide array of harmful implications to terrestrial and aquatic life which ranges from skin sensitization to acute oral toxicity. The current study aims to assess the quantitative skin sensitization potential of a large set of industrial and environmental chemicals acting through different mechanisms using the novel quantitative Read-Across Structure–Activity Relationship (q-RASAR) approach. Based on the identified important set of structural and physicochemical features, Read-Across-based hyperparameters were optimized using the training set compounds followed by the calculation of similarity and error-based RASAR descriptors. Data fusion, further feature selection, and removal of prediction confidence outliers were performed to generate a partial least squares (PLS) q-RASAR model, followed by the application of various Machine Learning (ML) tools to check the quality of predictions. The PLS model was found to be the best among different models. A simple user-friendly Java-based software tool was developed based on the PLS model, which efficiently predicts the toxicity value(s) of query compound(s) along with their status of Applicability Domain (AD) in terms of leverage values. This model has been developed using structurally diverse compounds and is expected to predict efficiently and quantitatively the skin sensitization potential of environmental chemicals to estimate their occupational and health hazards.

Received 26th July 2023
Accepted 4th September 2023

DOI: 10.1039/d3em00322a

rsc.li/espi