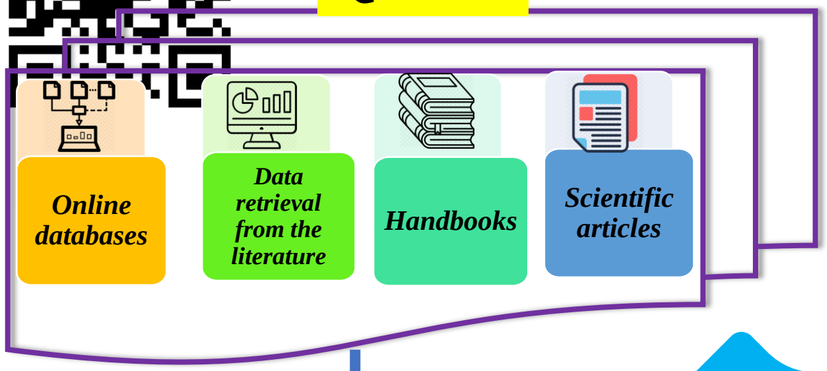
q-RASAR vs. QSAR: EFFICIENT PREDICTIONS OF ACTIVITY/PROPERTY/ TOXICITY ENDPOINTS

Kunal Roy

Drug Theoretics and Cheminformatics Lab
Division of Medicinal and Pharmaceutical Chemistry
Department of Pharmaceutical Technology
Jadavpur University, Kolkata 700 032 (India)
Email: kunalroy_in@yahoo.com
URL: http://sites.google.com/site/kunalroyindia/

QSAR



Input Data (Activity/Property/Toxicity)

Biological and Chemical curation

Descriptor calculation

- PaDEL
- Dragon
- AlvaDesc
- Ochem
- Rdkit
- One hot coding
- CBVM

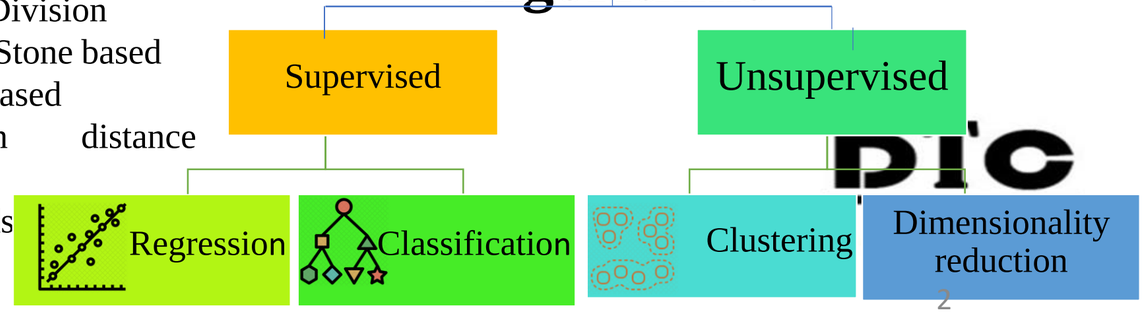
Dataset Division

- Random Division
- Kennord-Stone based
- Activity based
- Euclidiean distance based
- K-medoids

Feature Selection

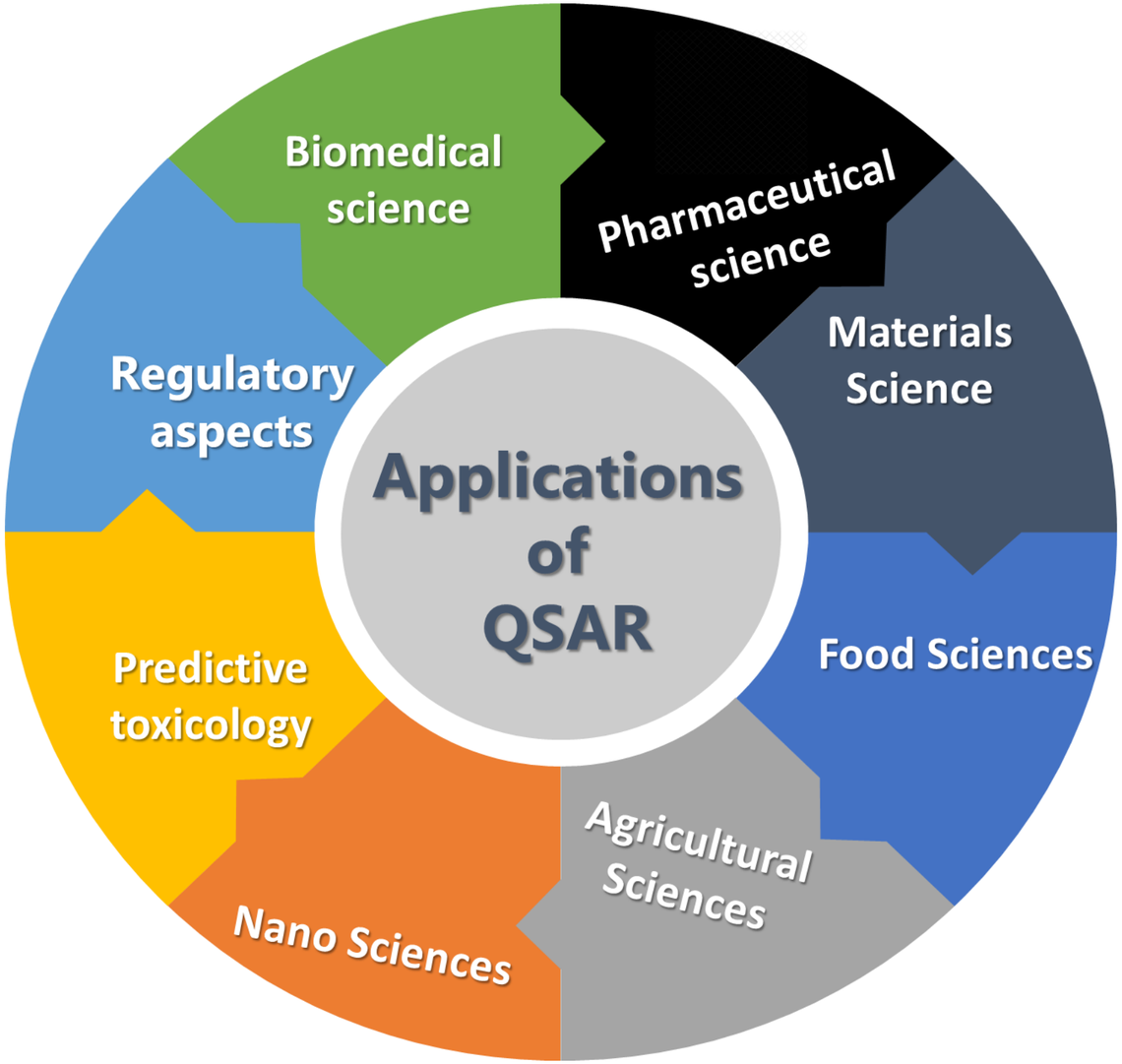
- Genetic algorithm
- Lasso feature selection
- Stepwise selection

Learning Algorithms



Model analysis

- Predictive variance
- Applicability domain
- Mechanistic interpretation
- Model quality


Biomedical science
Pharmaceutical science
Materials Science
Food Sciences
Agricultural Sciences
Nano Sciences
Predictive toxicology
Regulatory aspects
Applications of QSAR

Read-across

•Read across (RA) is a **prediction method** of unknown chemicals from the chemical analogues with known toxicity from the **same chemical category**.

•It is accepted by **REACH** and **US EPA**.

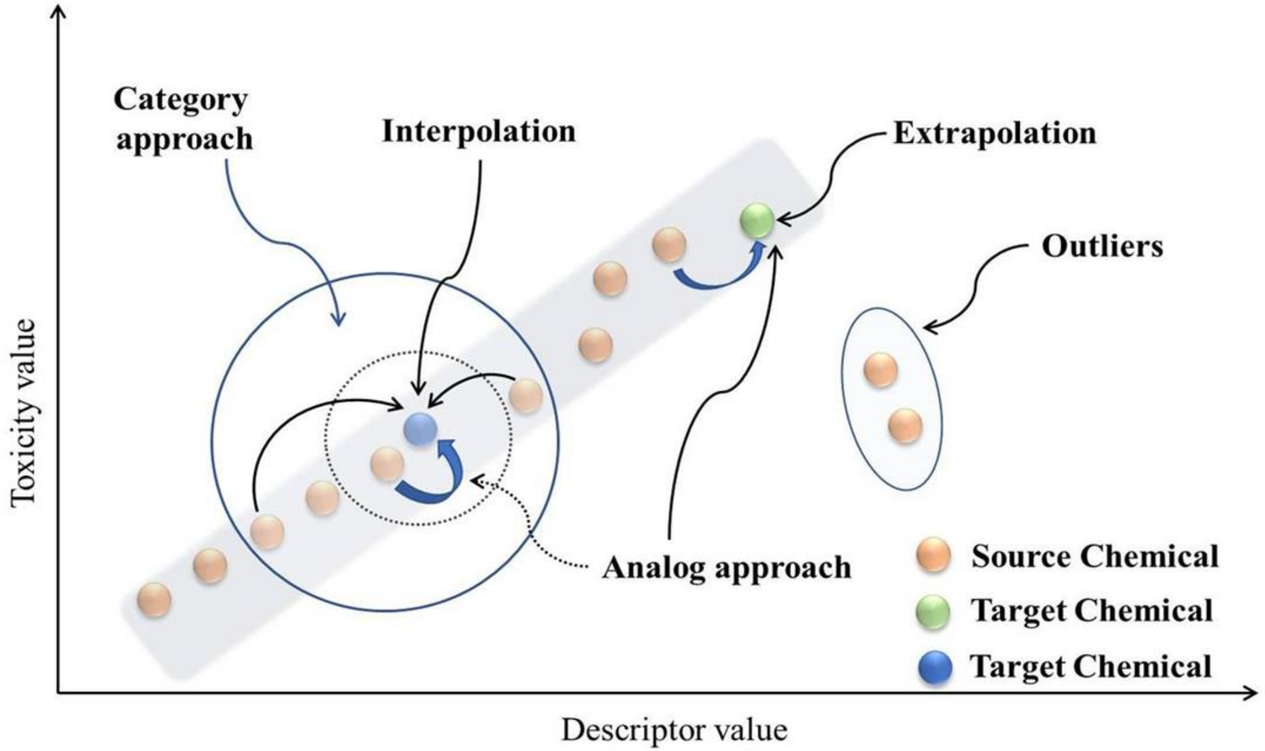
•Used for **data gap filling**.

•Defined chemical category is necessary.

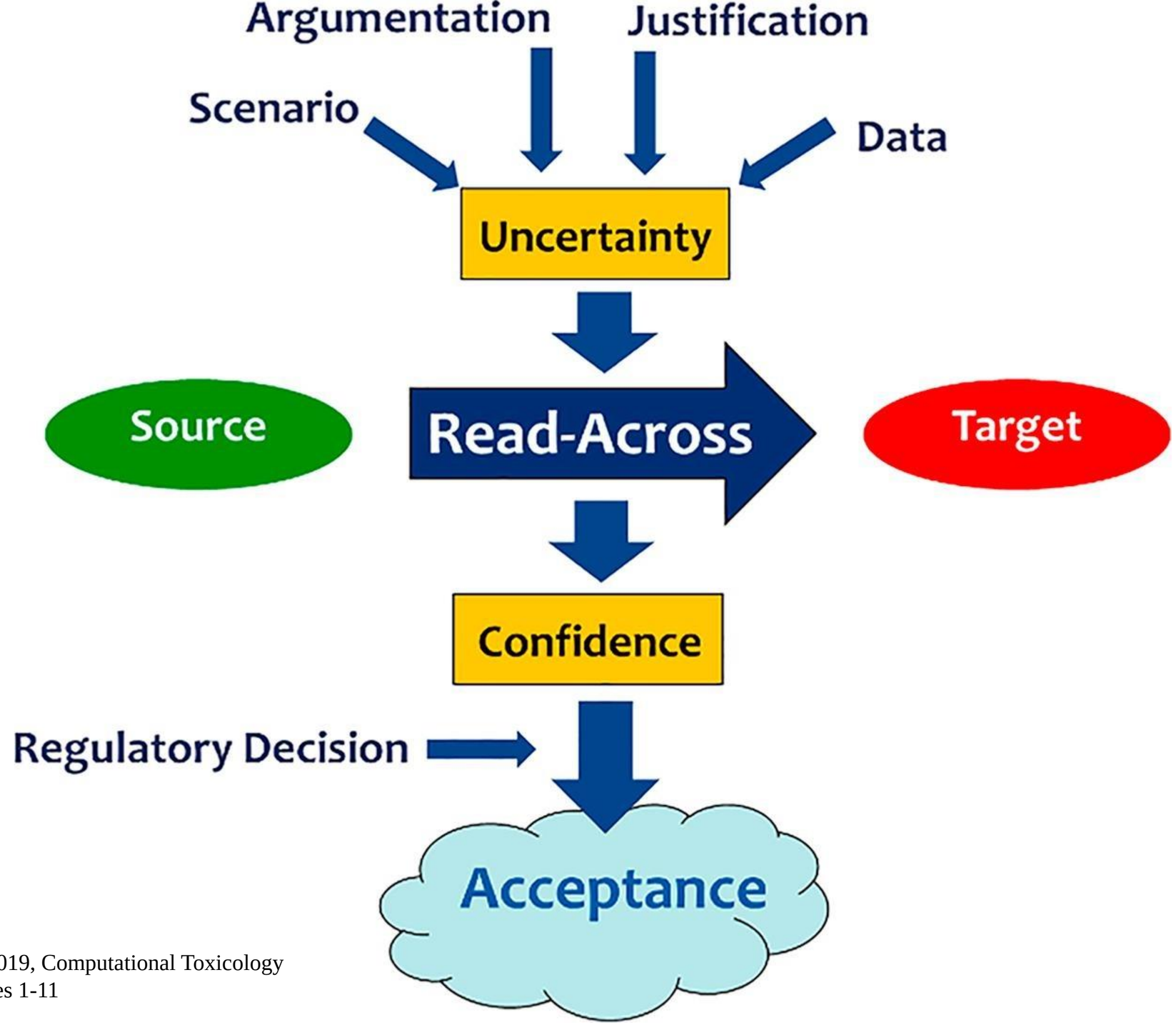
• Strategies: One→ One; One → Many Many → One; Many → Many

•**Analog** approach

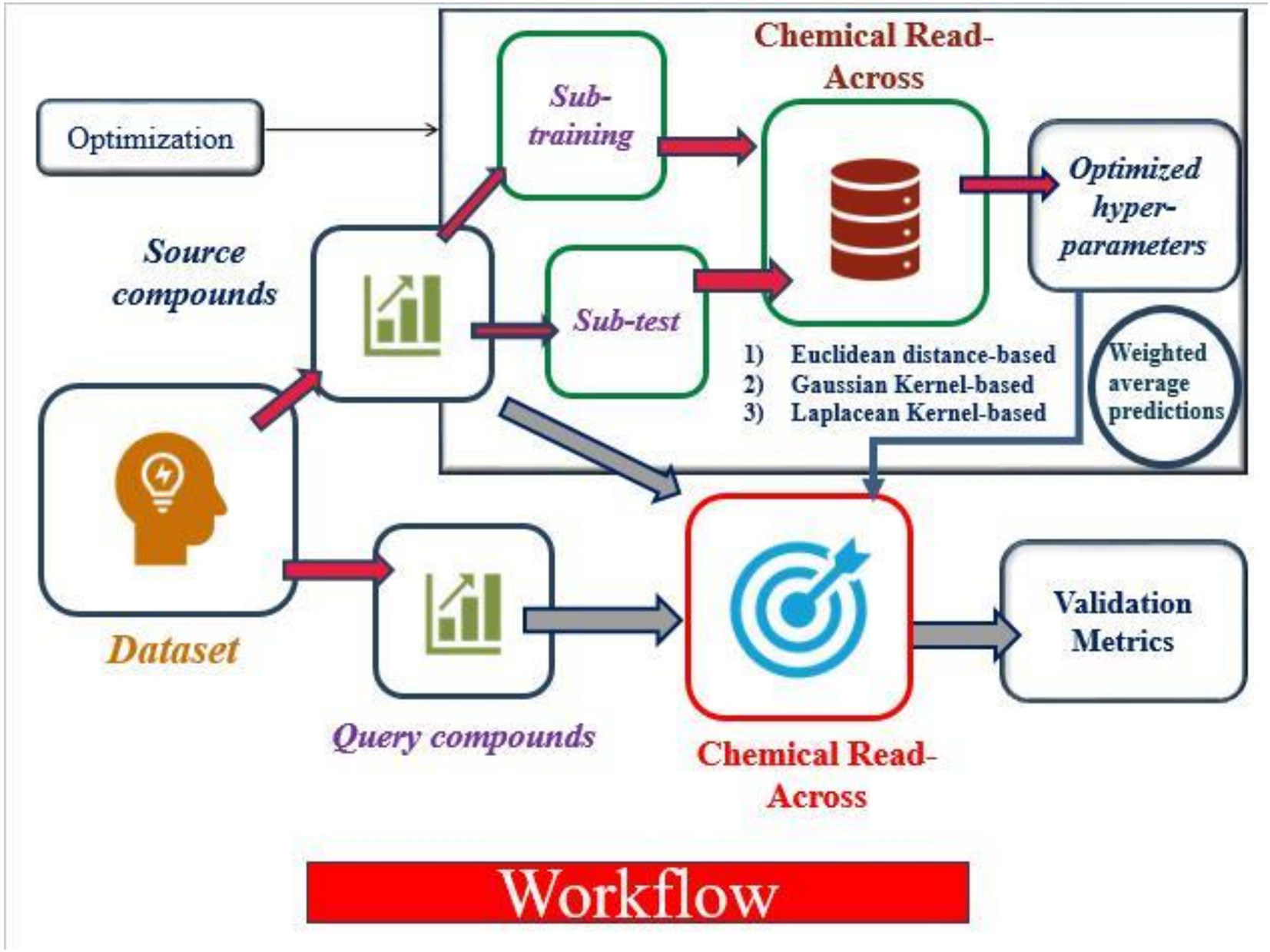
•**Category** approach



Vink, S.R. *et al.*, *Regul Toxicol Pharmacol*. **2010**, *58*, 64–71

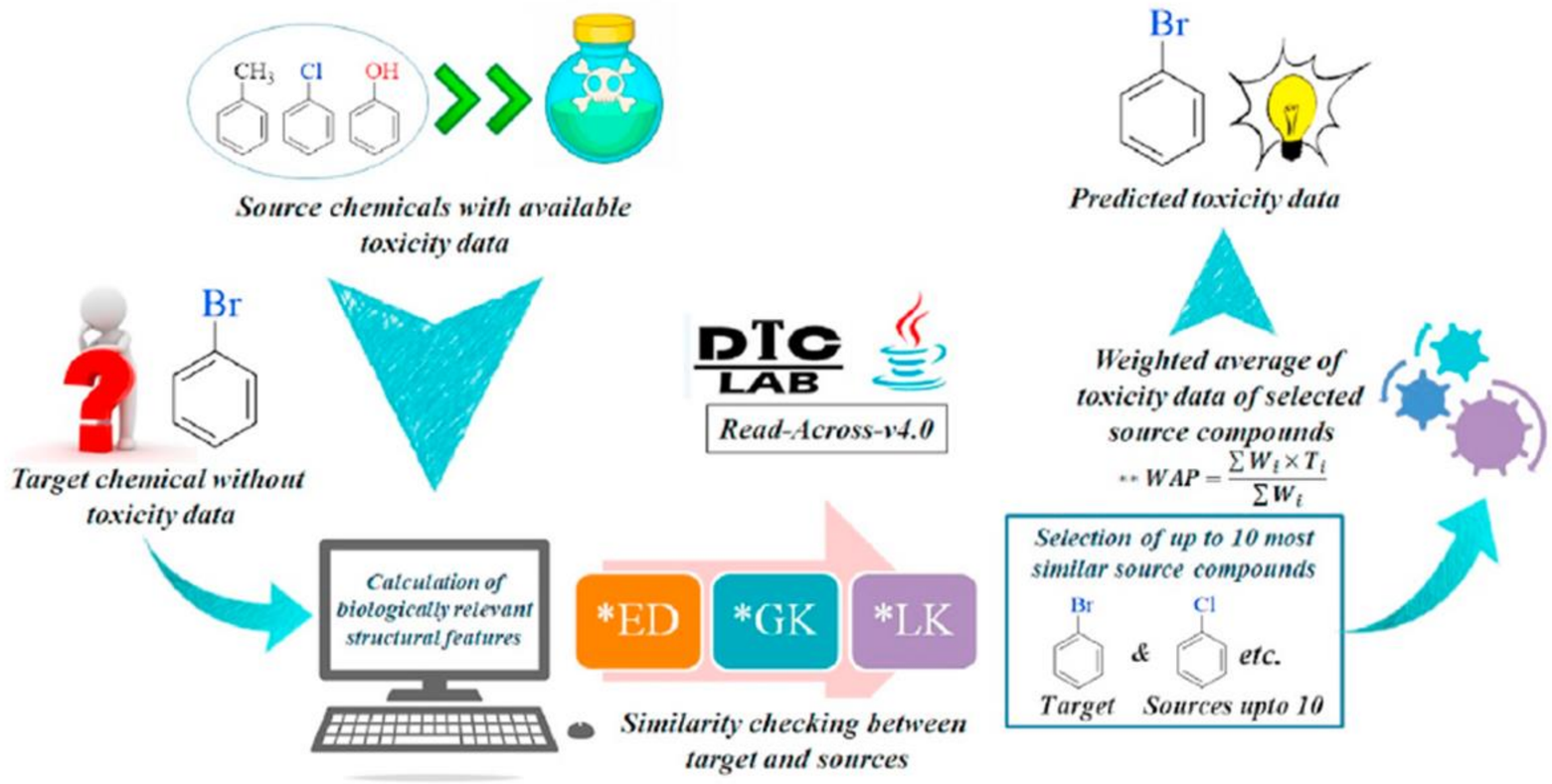


Schultz et al. 2019, Computational Toxicology Volume 9, Pages 1-11



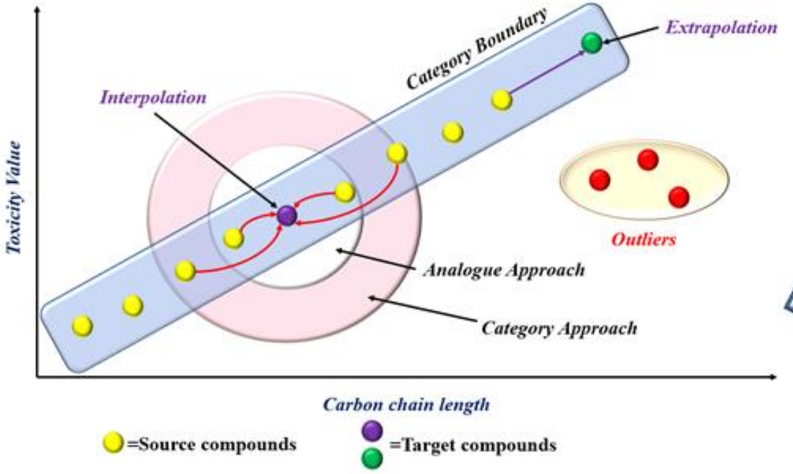
Chatterjee et al., Environ Sci Nano, 2022, 9, 189-203

Workflow for read-across predictions

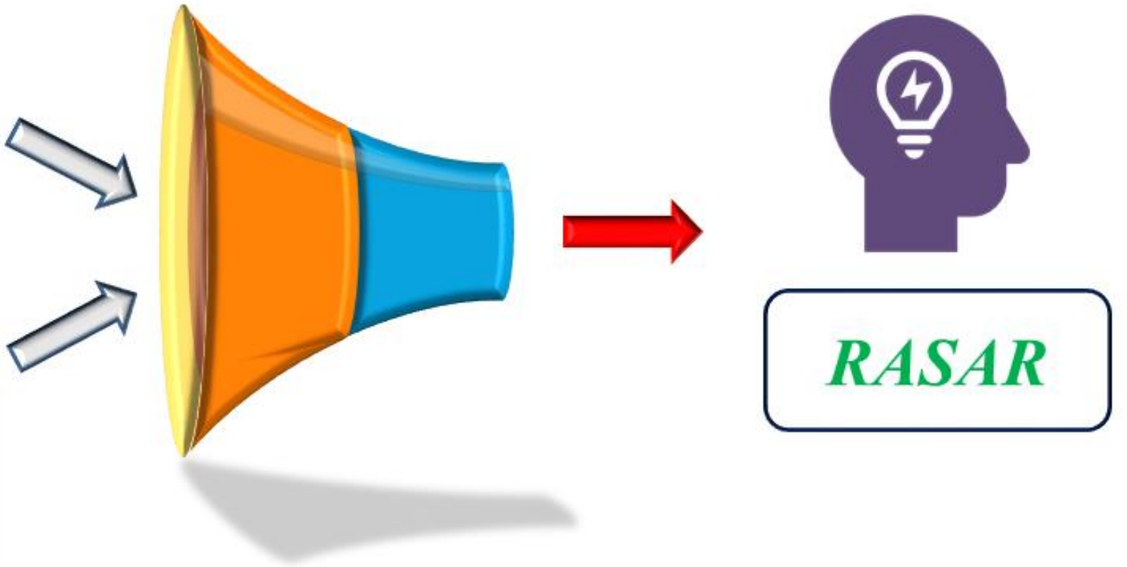


*ED: Euclidean distance-based similarity; GK: Gaussian kernel function similarity; LK: Laplacian kernel function similarity
**WAP: Weighted average predictions; W_i: weightage of i^{th} source compound (based on similarity); T_i: toxicity of i^{th} source compound

Read-Across



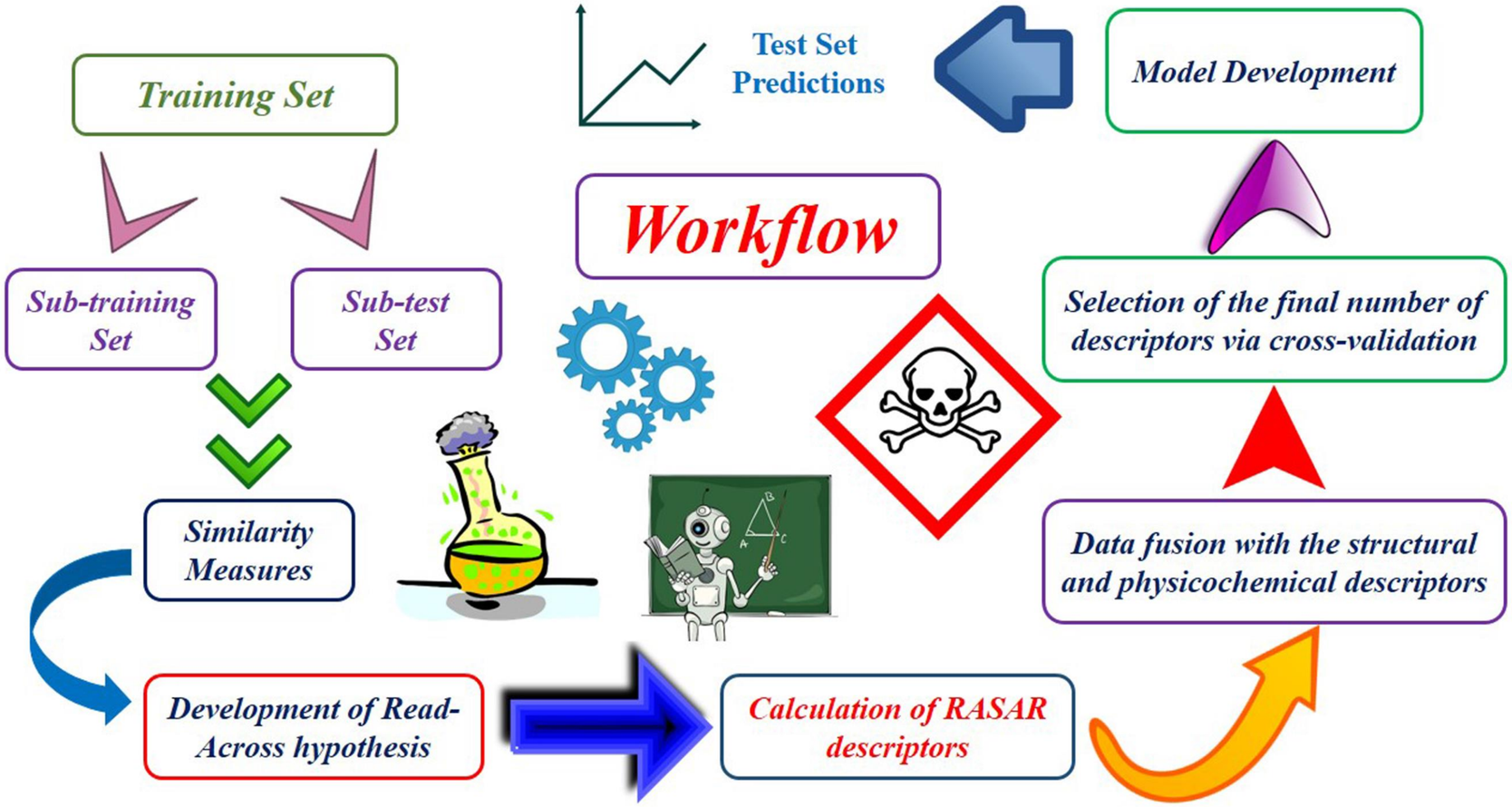
Unsupervised learning (Similarity –based)



Supervised learning (Response - dependent)

Luetchfeld et al., Toxicol Sci 165(1):198–212.

RASAR



Banerjee and Roy, Chem Res Toxicol, 2023

RASAR

Table 1 List of similarity and various error measures generated for each query compound during read-across predictions

Measure	Definition
Dispersion measures	
SD_activity	Standard deviation of the (observed) activity values of the selected close source compounds for each query compound
CV_activity	Coefficient of variation of the response
Similarity measures	
Average similarity	Mean similarity to the close source compounds for each query compound
SD_similarity	Standard deviation of the similarity values of the selected close source compounds for each query compound
MaxPos	Maximum Similarity level to the Positive close source compounds (based on source set observed mean)
MaxNeg	Maximum Similarity level to the Negative close source set compounds (based on source set observed mean)
AbsDiff	Absolute difference between MaxPos and MaxNeg
Concordance measure	
g	$g = 1 - 2 \times \lvert \text{PosFrac} - 0.5 \rvert$, where *PosFrac* is the fraction of the close source compounds belonging to the Positive Class based on the source set response mean as the threshold [11]

Banerjee and Roy, Molecular Diversity, 2022

Measure	Expression
Weighted average activity	$\overline{x_{wtd}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
SD_activity	$s_{weighted} = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \overline{x_{wtd}})^2}{\sum_{i=1}^{n} w_i}} \times \frac{n}{n-1}$
CV_activity	$CV_{activity} = \frac{s_{weighted}}{\overline{x_{wtd}}}$
ED-based similarity function	$f(ED) = 1 - d(X,Y)_{scaled}$ $d(X,Y) = \sqrt{\sum_{i=0}^{n} (X_i - Y_i)^2}$

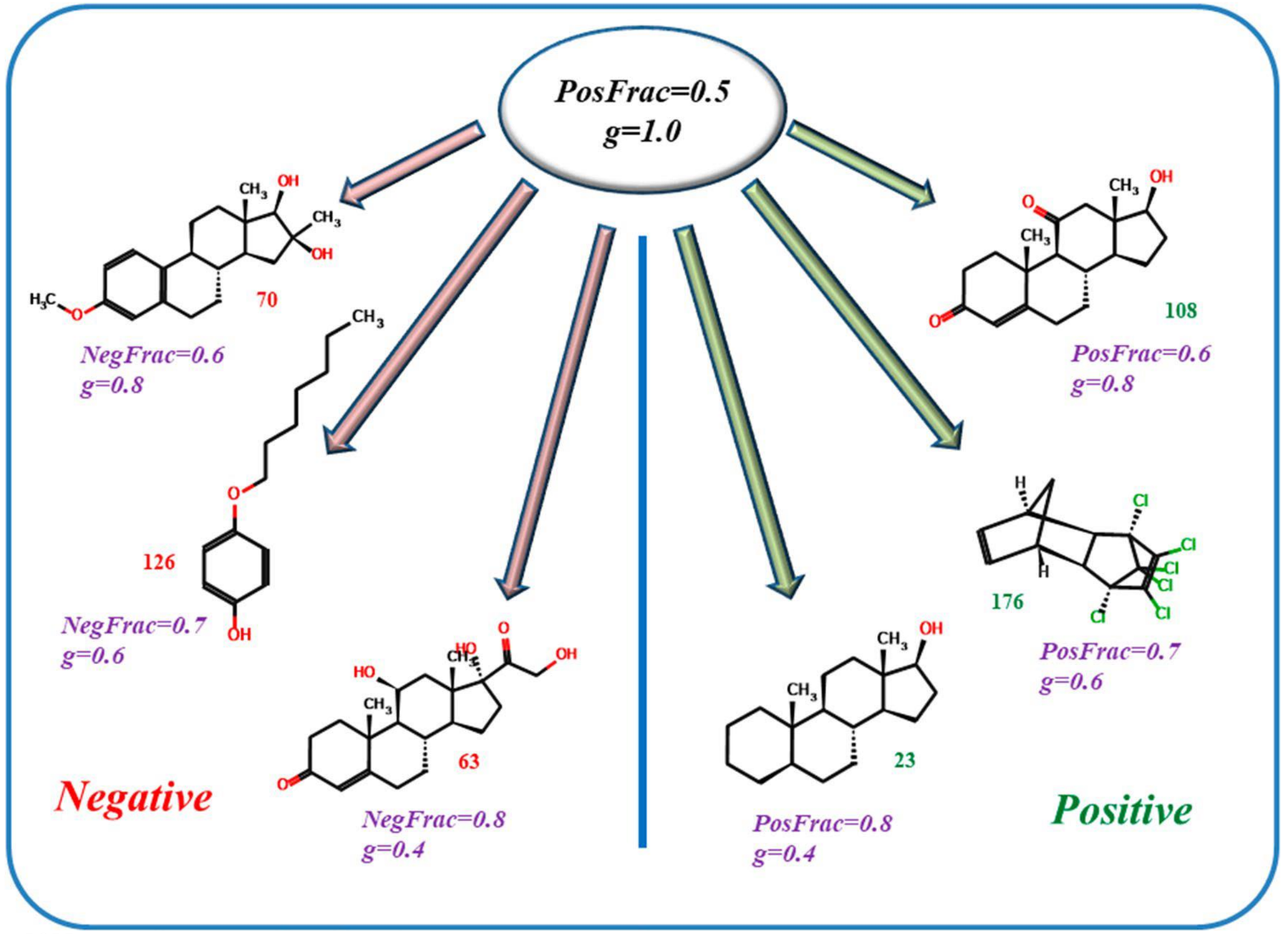
Measure	Expression
GK-based similarity function	$f(GK) = e^{-\frac{\|X_i - Y_i\|^2}{2\sigma^2}}$
LK-based similarity function	$f(LK) = e^{(-\gamma\|X-Y\|_1)}$
Average similarity	$Similarity_{average} = \frac{\sum_{i=1}^{n} f_i}{n}$
SD_similarity	$s_{similarity=} \sqrt{\frac{\sum_{i=1}^{n}(f - \overline{f})^2}{n-1}}$

Measure	Expression
MaxPos	
MaxNeg	
AbsDiff	$AbsDiff = \|MaxPos - MaxNeg\|$
Concordance measure	$g = 1 - 2 \times \|PosFrac - 1/2\|$

RASAR

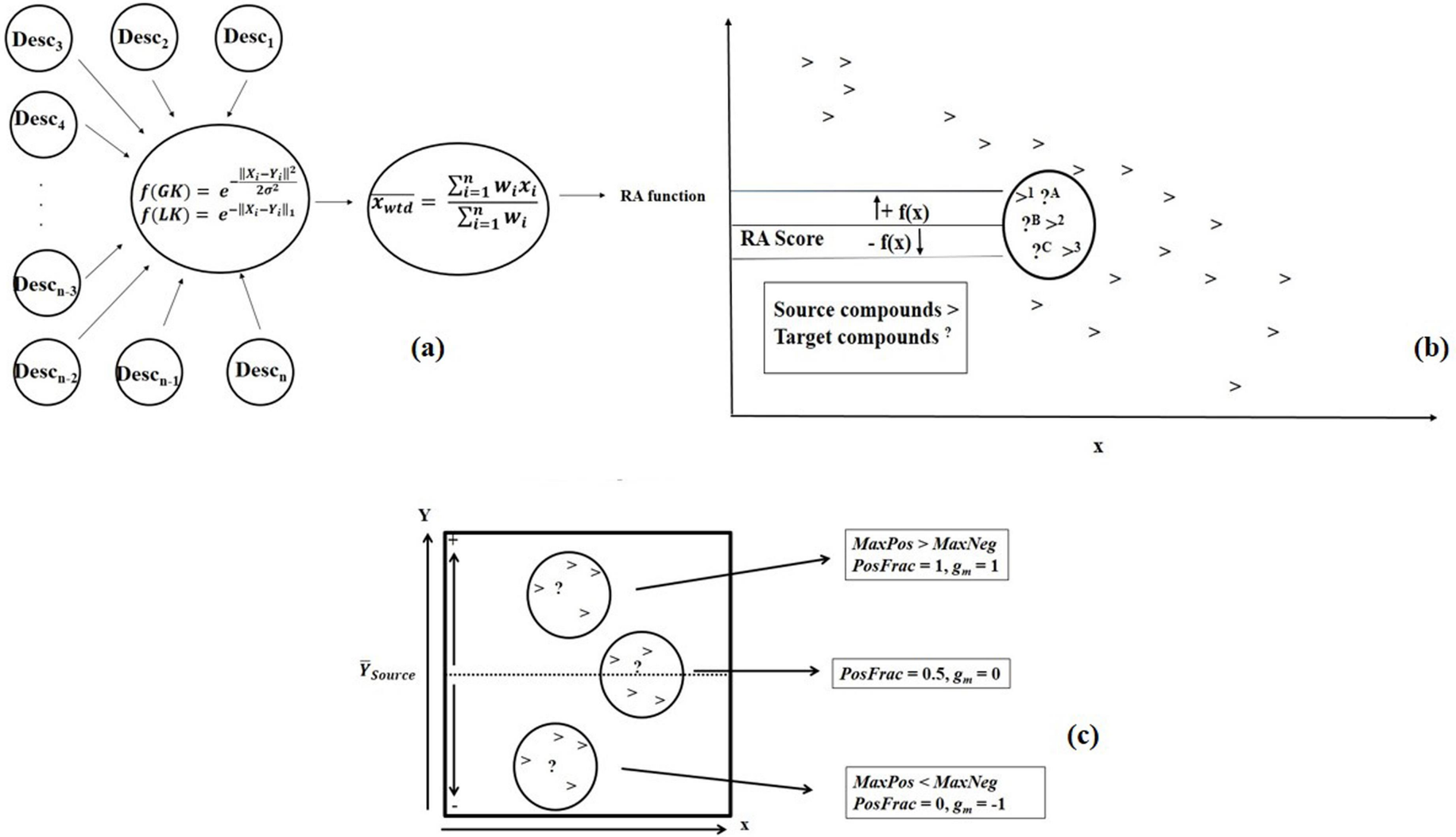
$$g = 1 - 2 \times |PosFrac - 1/2|$$

$$g_{|m} = (-1)^n \times 2|\text{PosFrac} - 0.5|$$



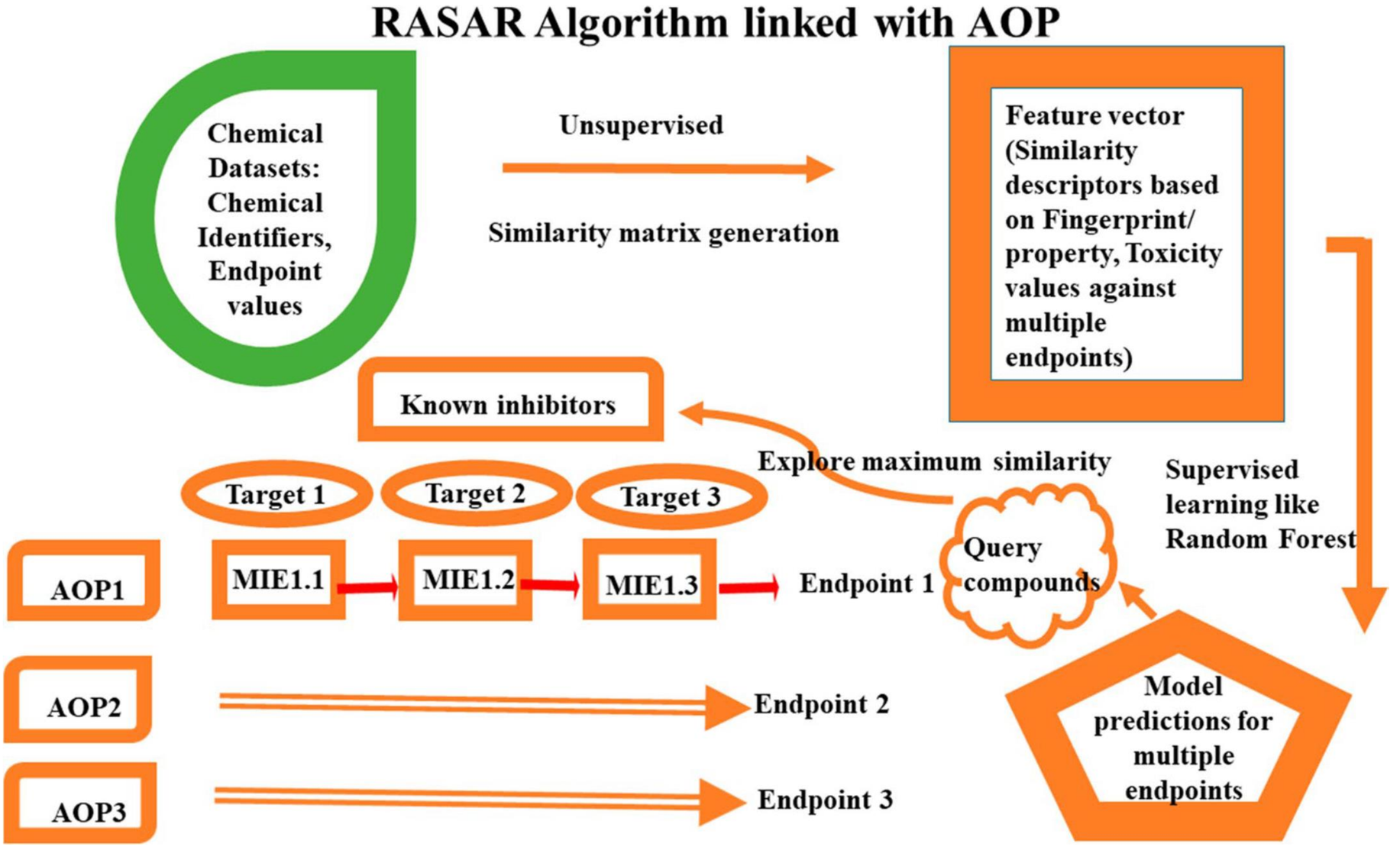
Banerjee and Roy, Molecular Diversity, 2022

RASAR



Banerjee and Roy, Chem Res Toxicol, 2023

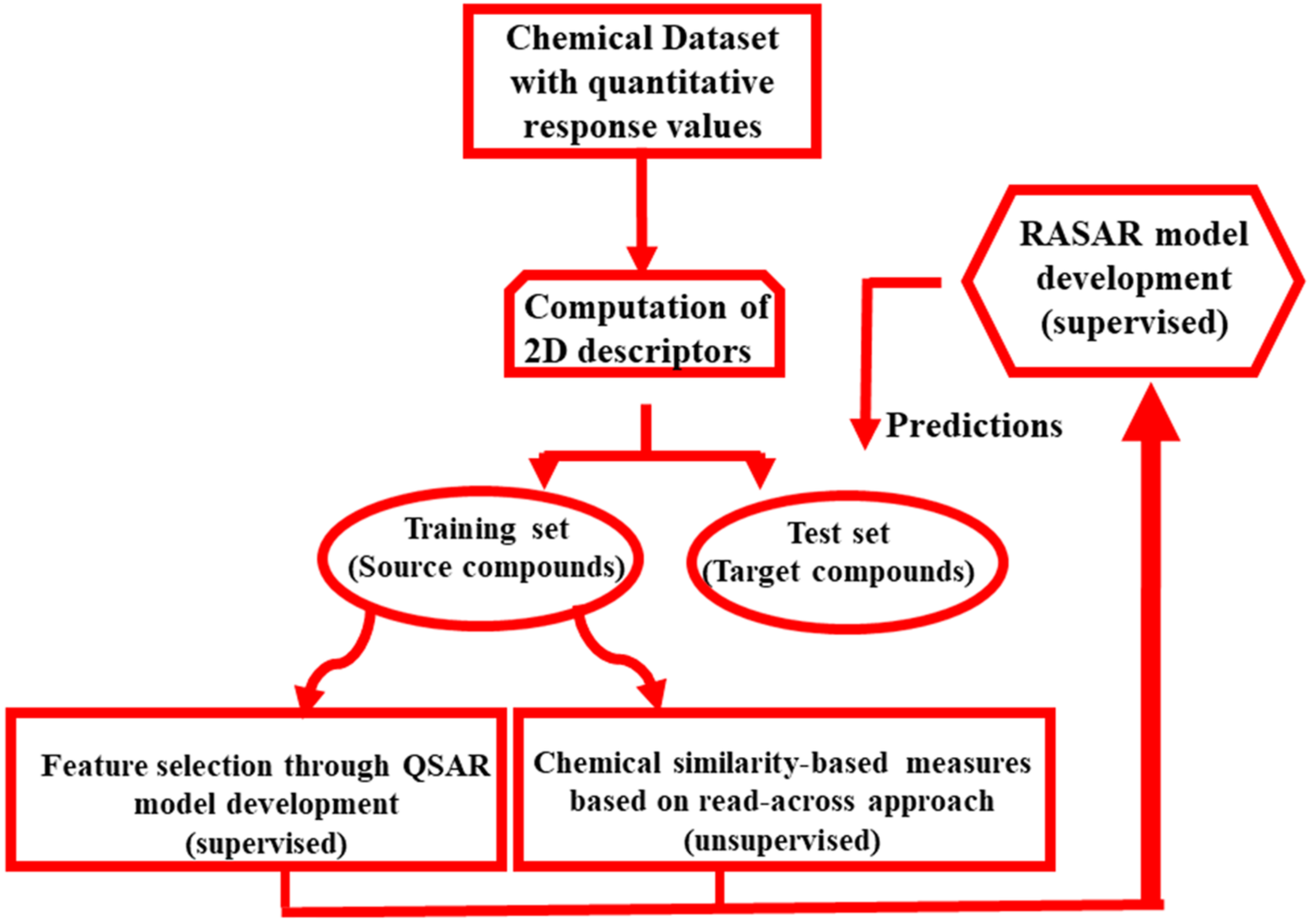
RASAR



Banerjee and Roy, Molecular Diversity, 2022

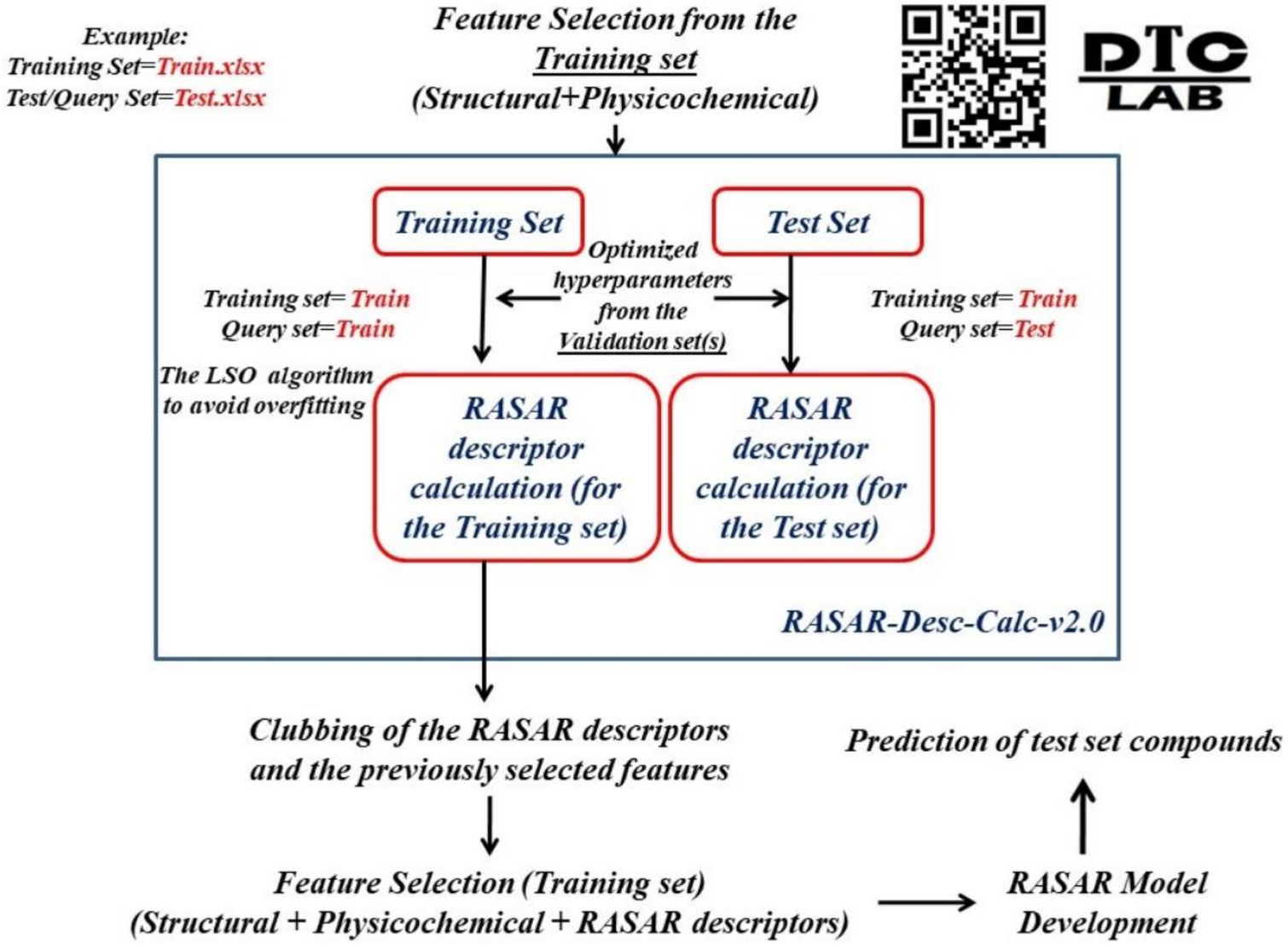
RASAR

RASAR algorithm combining QSAR and Read-across



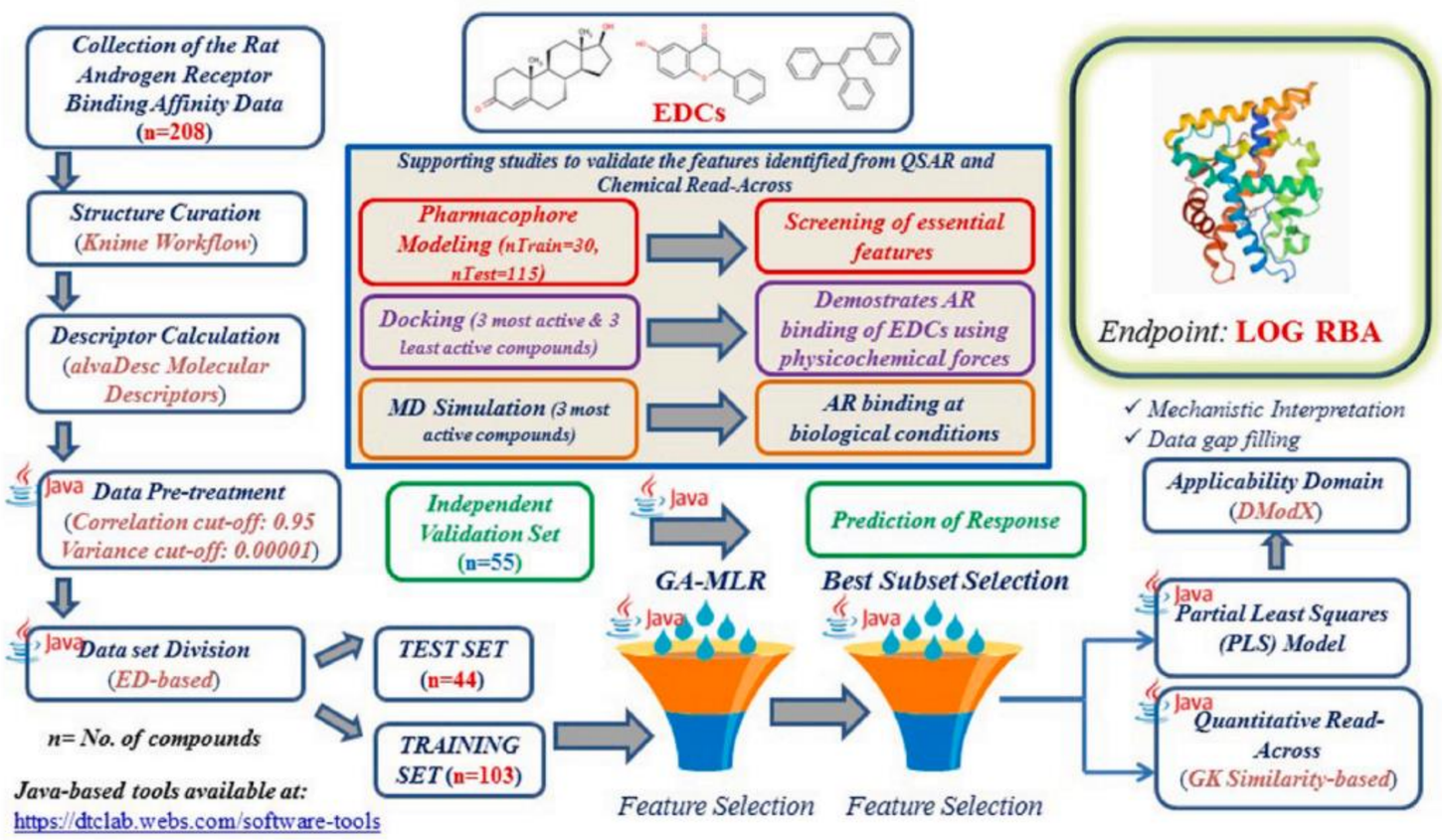
Banerjee and Roy, Molecular Diversity, 2022

RASAR Descriptor Calculation



Banerjee and Roy, Molecular Diversity, 2022

Modeling androgen receptor binding affinity



Chemosphere 309 (2022) 136579

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across

Arkaprava Banerjee [a], Priyanka De [a], Vinay Kumar [a], Supratik Kar [b,1], Kunal Roy [a,*]

$$LogRBA = -3.23 + 0.49 \times SsssCH - 0.41 \times MaxaaCH + 0.23 \times nCconj + 0.35 \times LogP99 - 0.17 \times F10[C-O] + 0.06 \times minsOH + 0.06 \times N\% + 0.67 \times F08[O-F] \quad (1)$$

$R^2_{(TRAIN)} = 0.74, Q^2_{(LOO)} = 0.68, Q^2_{F1} = 0.58, Q^2_{F2} = 0.58$

$Scaled\ average\ r^2_m(Train) = 0.57, Scaled\ average\ r^2_m(Test) = 0.50$

$Scaled\ delta\ r^2_m(Train) = 0.18, Scaled\ delta\ r^2_m(Test) = 0.07$

$MAE_{(TRAIN)} = 0.46, MAE_{(TEST)} = 0.54, n_{(Training)} = 103, n_{(Test)} = 44$

Modeling androgen receptor binding affinity

Chemosphere 309 (2022) 136579

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across

Arkaprava Banerjee [a], Priyanka De [a], Vinay Kumar [a], Supratik Kar [b,1], Kunal Roy [a,*]

We have used a data set androgen receptor binding affinity (RBA) originally collected from the Endocrine Disruptor Knowledge Base (EDKB) database (https://www.fda.gov/science-resea rch/bioinformatics-tools/endocrinedisruptor-knowledge-base), and chemical curation of the compounds was performed by the application of a KNIME workflow (https://sites.google.com/site/dtcla bdc/) taking the single.sdf file as input.

*n*Training = 102, *n*Test = 44

We have finally used the descriptors selected in the previous QSAR model as the important physicochemical measures of the compounds in addition to different similarity measures as described below for the q-RASAR analysis.

$$LogRBA = -3.23 + 0.49 \times SsssCH - 0.41 \times MaxaaCH + 0.23 \times nCconj + 0.35 \times LogP99 - 0.17 \times F10[C-O] + 0.06 \times minsOH + 0.06 \times N\% + 0.67 \times F08[O-F] \quad (1)$$

$$R^2_{(TRAIN)} = 0.74, Q^2_{(LOO)} = 0.68, Q^2_{F1} = 0.58, Q^2_{F2} = 0.58$$

$$Scaled\ average\ r^2_m(Train) = 0.57, Scaled\ average\ r^2_m(Test) = 0.50$$

$$Scaled\ delta\ r^2_m(Train) = 0.18, Scaled\ delta\ r^2_m(Test) = 0.07$$

$$MAE_{(TRAIN)} = 0.46, MAE_{(TEST)} = 0.54, n_{(Training)} = 103, n_{(Test)} = 44$$

Modeling androgen receptor binding affinity

Table 2 List of physicochemical features selected from the previously reported QSAR model [12]

Measure	Description	Comment
SsssCH	Sum of *E*-state value of tertiary carbon atoms of type >CH–	*E*-state index
MaxaaCH	Maximum *E*-state value of the carbon atom of type aaCH	*E*-state index
nCconj	Number of non-aromatic conjugated carbons (sp^2)	Constitutional descriptor
LOGP99	Wildmann-Crippen octanol–water partition coefficient	Hydrophobicity measure
F10[C–O]	Frequency of C and O at the topological distance 10	Atom pair index
minsOH	Minimum Estate of the –OH hydroxyl group	*E*-state index
N%	The percentage of nitrogen present in the molecular structure	Constitutional descriptor
F08[O–F]	The frequency of O and F atoms at the topological distance of 8	Atom pair index

RASAR: Modeling androgen receptor binding affinity

Molecular Diversity
https://doi.org/10.1007/s11030-022-10478-6

ORIGINAL ARTICLE

First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability

Arkaprava Banerjee[1] · Kunal Roy[1]

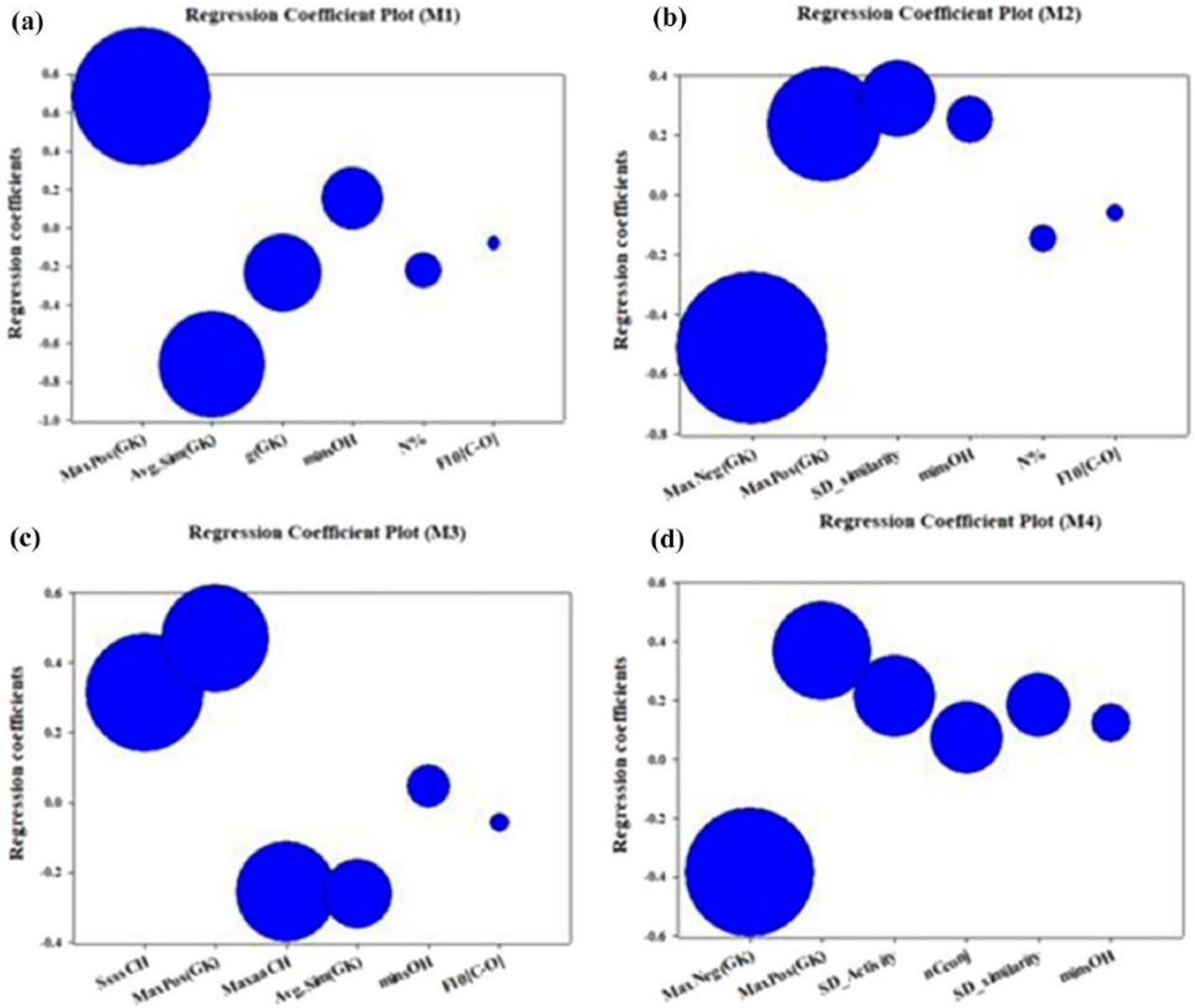
Received: 21 April 2022 / Accepted: 3 June 2022

Table 3 List of q-RASAR models

Model no.	Equation
Individual q-RASAR models	
*M*1	$\log RBA = -1.33 + 2.27MaxPos(GK) - 3.57Avg.Sim(GK) - 1.02g(GK) + 0.04minsOH - 0.14N\% - 0.06F10[C-O]$
*M*2	$\log RBA = -2.38 - 1.66MaxNeg(GK) + 0.78MaxPos(GK) + 4.32SDSimilarity(GK) + 0.06minsOH - 0.09N\% - 0.05F10[C-O]$
*M*3	$\log RBA = -1.97 + 0.35SsssCH + 1.55MaxPos(GK) - 0.34MaxaaCH - 1.31Avg.Sim(GK) + 0.01minsOH - 0.04F10[C-O]$
*M*4	$\log RBA = -2.93 - 1.25MaxNeg(GK) + 1.22MaxPos(GK) + 0.73SDActivity(GK) + 0.05nCconj + 2.47SDSimilarity(GK) + 0.03minsOH$
Pooled descriptor q-RASAR models	
*P*1 (*M*1 + *M*2)	$\log RBA = -1.71 - 1.47MaxNeg(GK) + 1.06MaxPos(GK) + 2.88SDSimilarity(GK) - 0.86Avg.Sim(GK) + 0.05minsOH - 0.41g(GK) - 0.10N\% - 0.05F10[C-O]$
*P*2 (*M*1+*M*2+*M*3)	$\log RBA = -1.76 - 1.00MaxNeg(GK) + 0.29SsssCH + 0.91MaxPos(GK) - 0.24MaxaaCH - 0.40Avg.Sim + 1.32SDSimilarity(GK) + 0.03minsOH - 0.04F10[C-O] - 0.05N\% + 0.17g(GK)$
*P*3 (*M*1+*M*2+*M*4)	$\log RBA = -2.55 - 1.13MaxNeg(GK) + 1.10MaxPos(GK) + 0.72SDActivity(GK) + 0.08nCconj - 0.48Avg.Sim(GK) + 1.81SDSimilarity(GK) + 0.03minsOH - 0.05F10[C-O] - 0.06N\% + 0.13g(GK)$

Banerjee and Roy, Molecular Diversity, 2022

RASAR: Modeling androgen receptor binding affinity



Molecular Diversity
https://doi.org/10.1007/s11030-022-10478-6

ORIGINAL ARTICLE

First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability

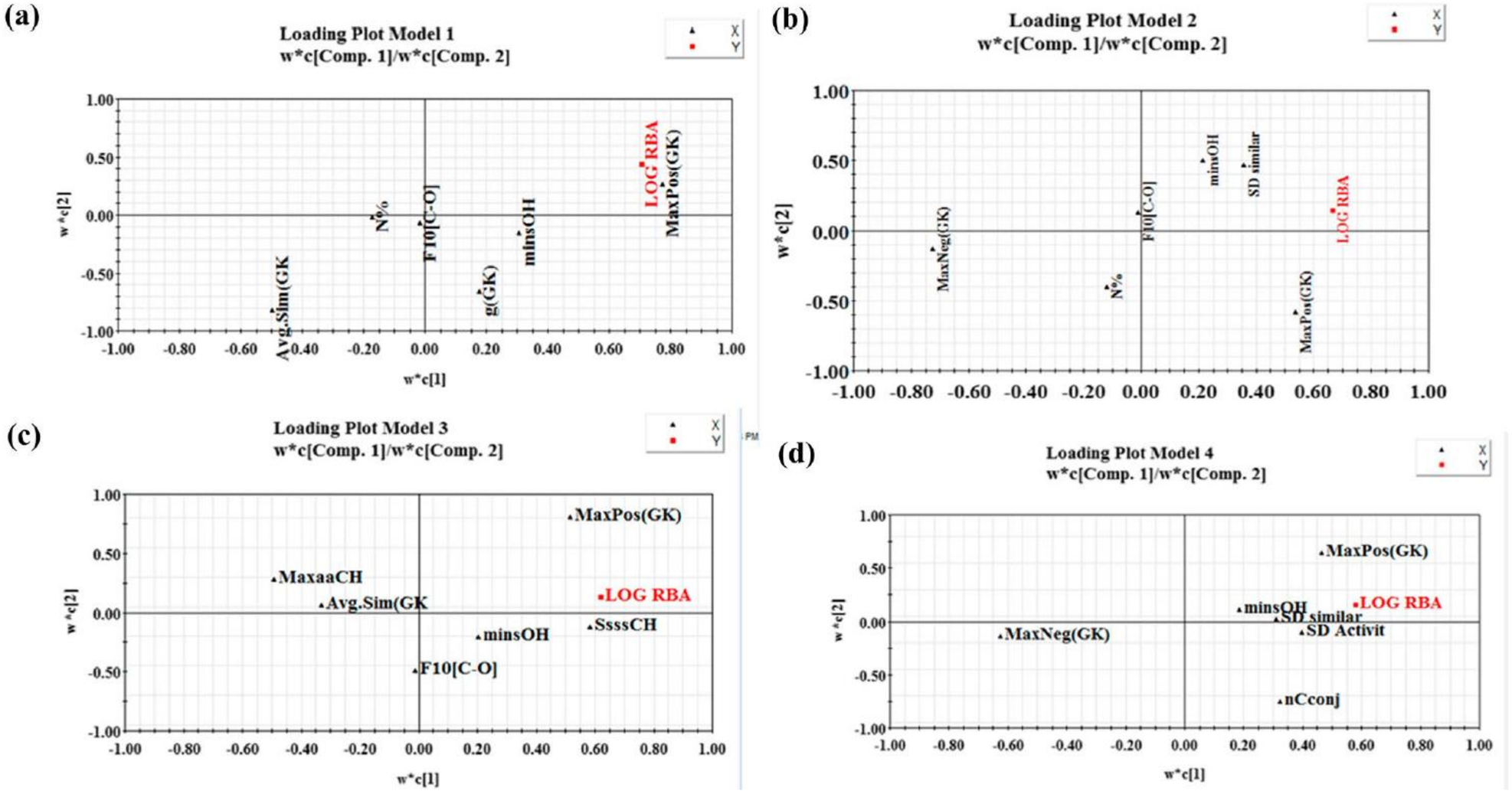
Arkaprava Banerjee[1] · Kunal Roy[1]

Received: 21 April 2022 / Accepted: 3 June 2022

DTC LAB

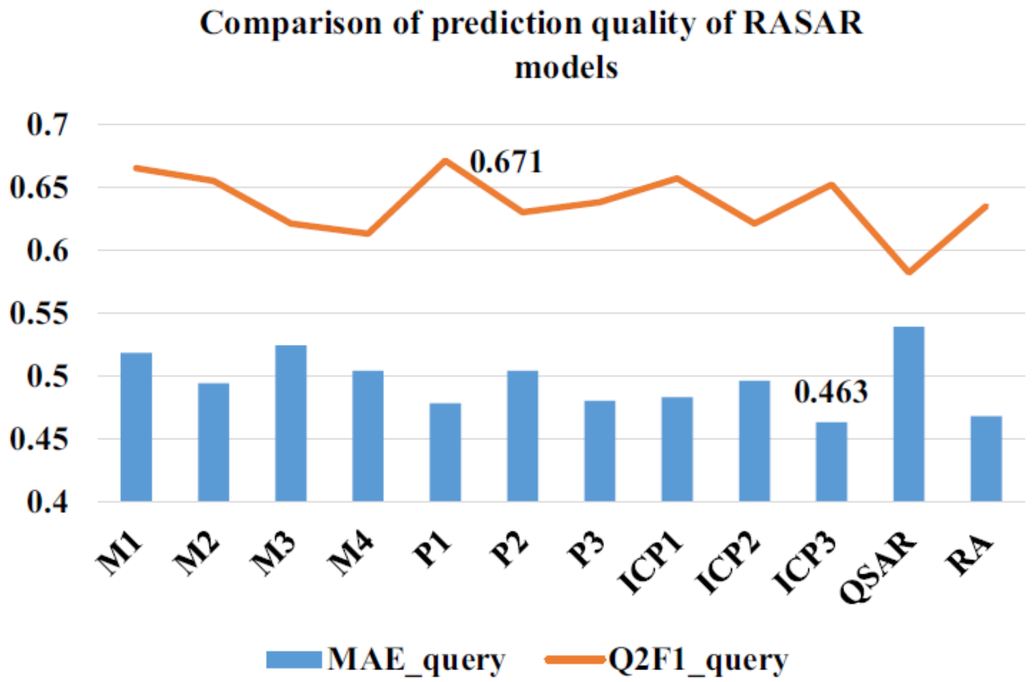
Banerjee and Roy, Molecular Diversity, 2022

RASAR: Modeling androgen receptor binding affinity



Banerjee and Roy, Molecular Diversity, 2022

RASAR: Modeling androgen receptor binding affinity



Banerjee and Roy, Molecular Diversity, 2022

RASAR: Modeling androgen receptor binding affinity

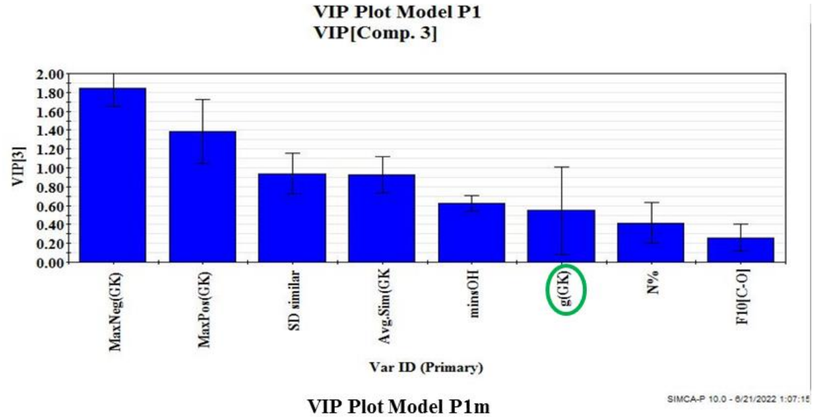
$$\log RBA = -1.21 - 1.31\text{MaxNeg(GK)} + 0.58g_m\text{(GK)} + 0.21\text{MaxPos(GK)} + 2.23\text{SD Similarity (GK)} - 0.67\text{Avg.Sim(GK)} + 0.06\min\text{sOH} - 0.10\text{N\%} - 0.13F10[C-O] \quad \text{(P1a)}$$

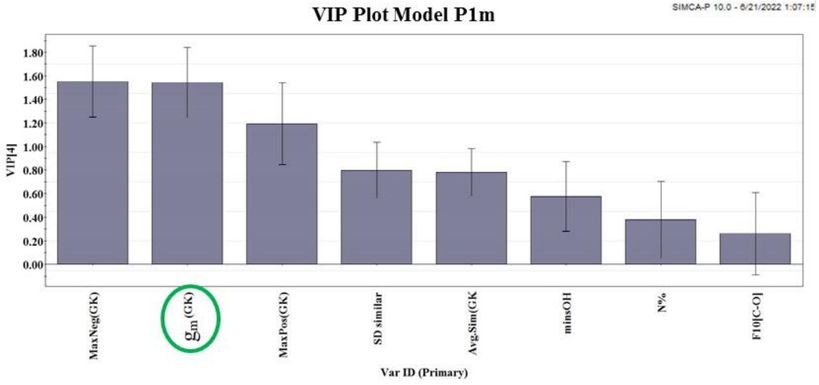
$n_{\text{Training}} = 102$ $n_{\text{Test}} = 44$ LV = 4

$R^2 = 0.753$ $Q^2_{(\text{LOO})} = 0.698$ $Q^2_{F1} = 0.674$ $Q^2_{F2} = 0.674$ $\text{MAE}_{(\text{TEST})} = 0.461$

$$g_m = (-1)^n \times 2|PosFrac - 0.5|$$

n=1 if MaxPos < MaxNeg, **n=2 if MaxPos > MaxNeg**





Banerjee and Roy, Molecular Diversity, 2022

RASAR: Modeling five toxicity endpoints

Chemical Research in Toxicology

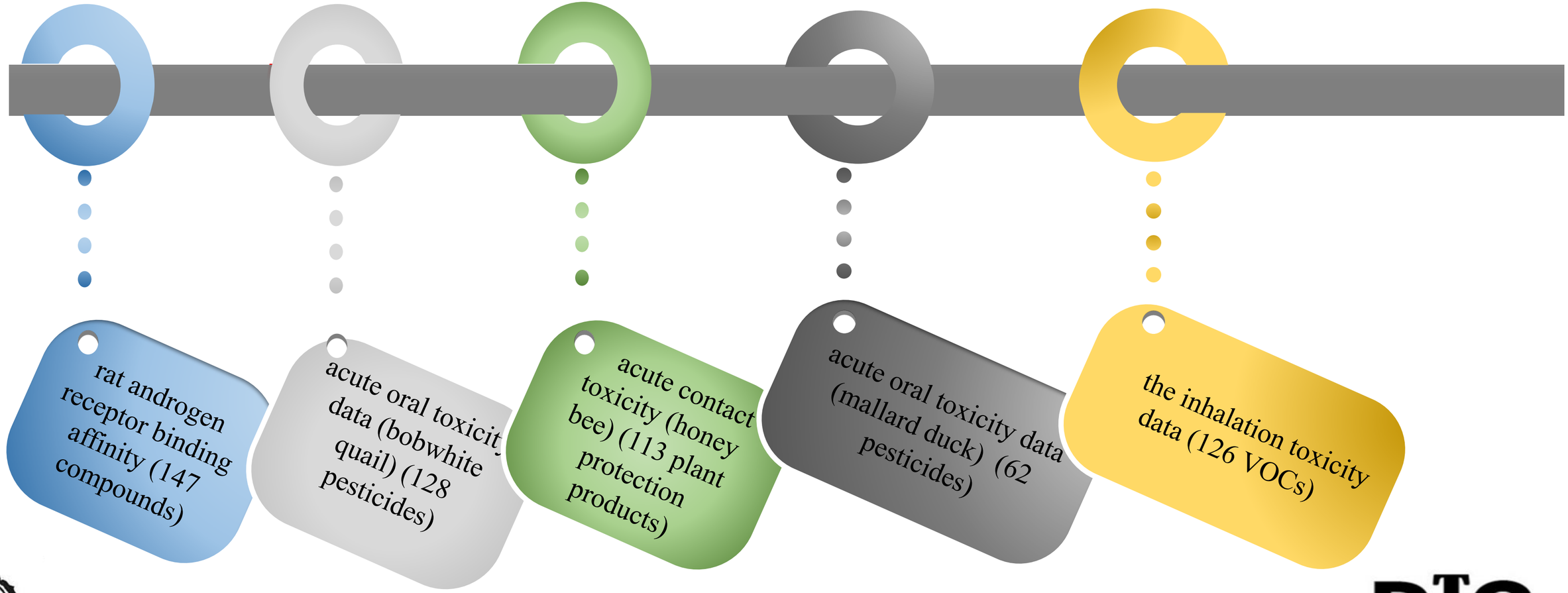
pubs.acs.org/crt Article

On Some Novel Similarity-Based Functions Used in the ML-Based q-RASAR Approach for Efficient Quantitative Predictions of Selected Toxicity End Points

Arkaprava Banerjee and Kunal Roy*

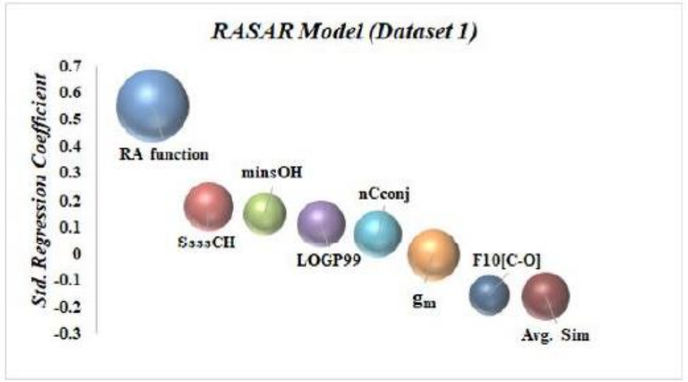
Cite This: https://doi.org/10.1021/acs.chemrestox.2c00374 Read Online

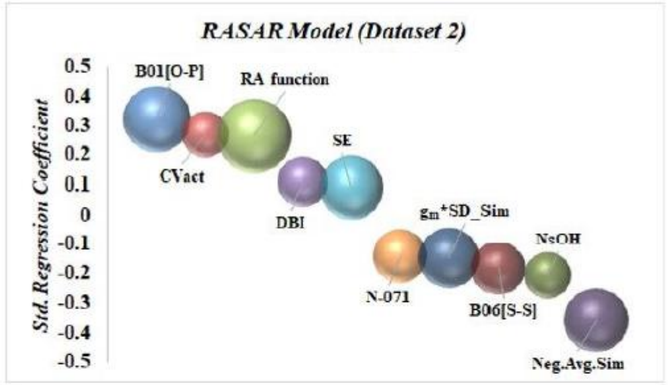
RASAR: Modeling five toxicity endpoints

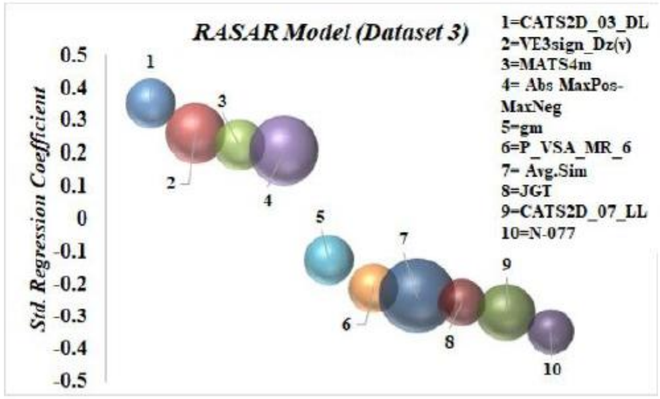


Banerjee and Roy, Chem Res Toxicol, 2023

RASAR: Modeling five toxicity endpoints

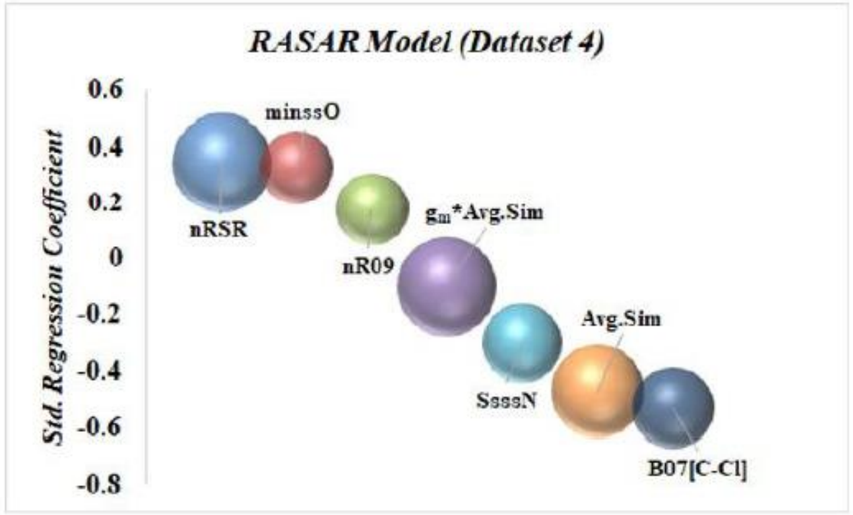


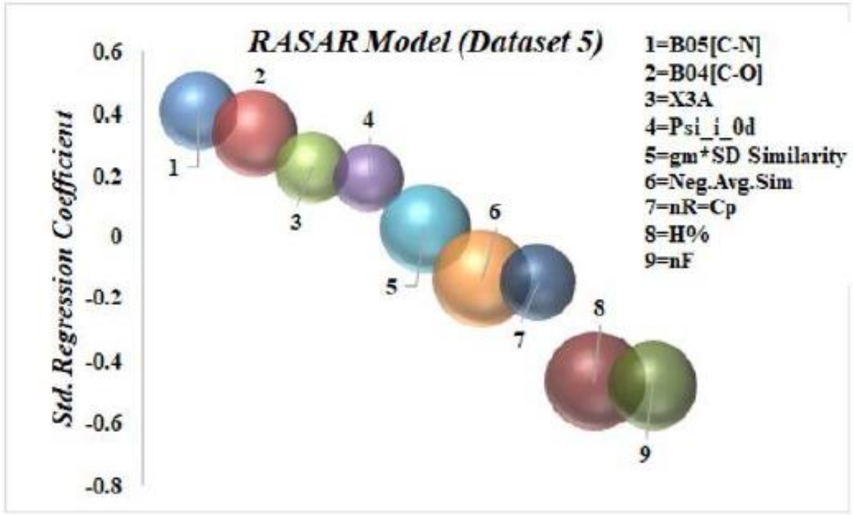




Banerjee and Roy, Chem Res Toxicol, 2023

RASAR: Modeling five toxicity endpoints





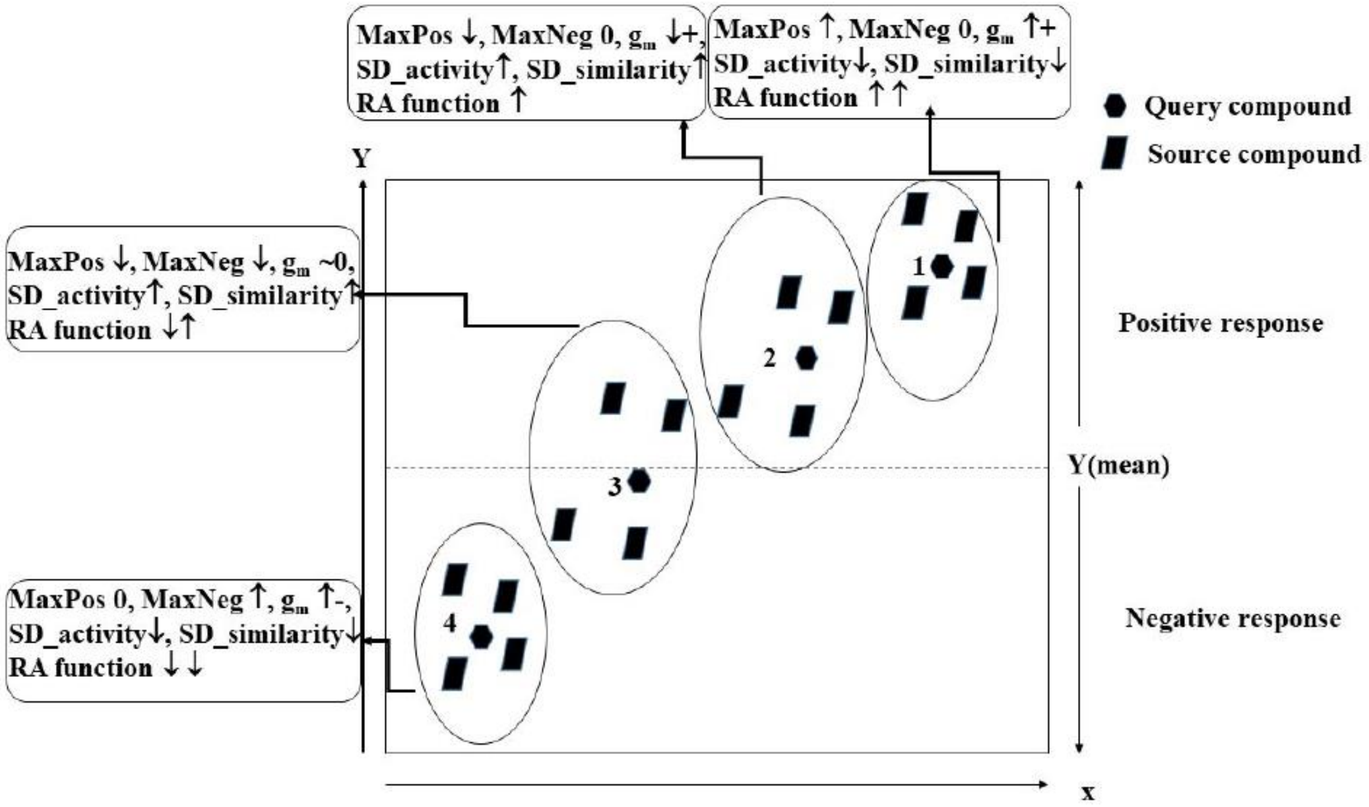
Banerjee and Roy, Chem Res Toxicol, 2023

RASAR: Modeling five toxicity endpoints

Datasets	Model	R^2	$Q^2_{(LOO)}$	Q^2_{F1}	Q^2_{F2}	$MAE_{(TEST)}$ *
Dataset 1	q-RASAR (GK)	0.71	0.63	0.70	0.70	0.44
	QSAR[13]	0.74	0.68	0.58	0.58	0.54
Dataset 2	q-RASAR (GK)	0.68	0.54	0.77	0.77	0.35
	QSAR[18]	0.66	0.58	0.65	0.65	0.46
Dataset 3	q-RASAR (LK)	0.62	0.53	0.83	0.83	0.41
	QSAR[19]	0.67	0.59	0.65	0.65	0.58
Dataset 4	q-RASAR (GK)	0.68	0.53	0.68	0.60	0.51
	QSAR[18]	0.66	0.57	0.66	0.58	0.58
Dataset 5	q-RASAR (GK)	0.73	0.64	0.74	0.74	0.45
	QSAR[20]	0.74	0.66	0.68	0.68	0.49

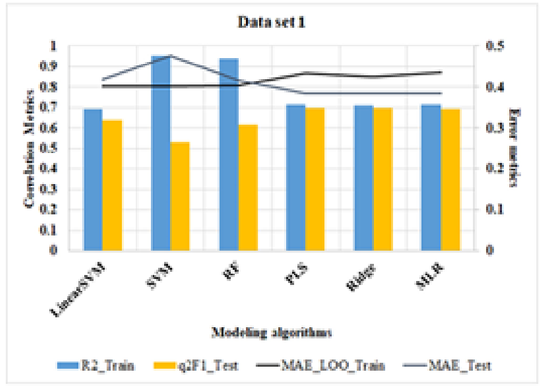
Banerjee and Roy, Chem Res Toxicol, 2023

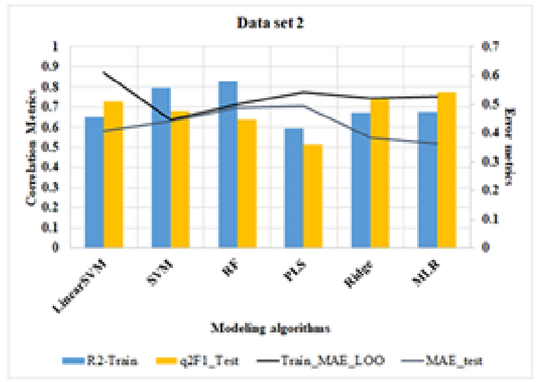
RASAR: Modeling five toxicity endpoints

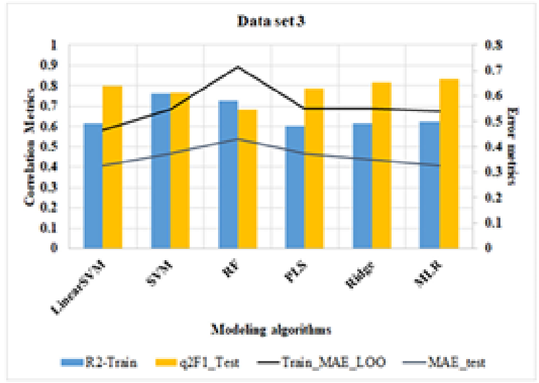


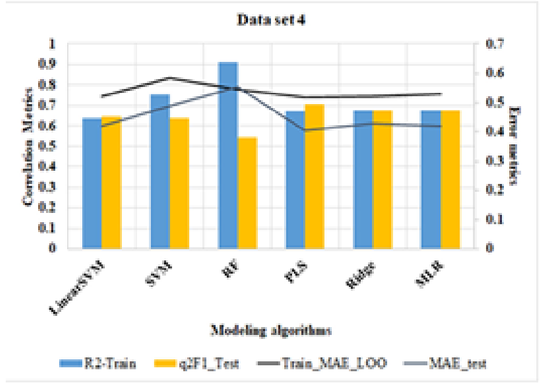
Banerjee and Roy, Chem Res Toxicol, 2023

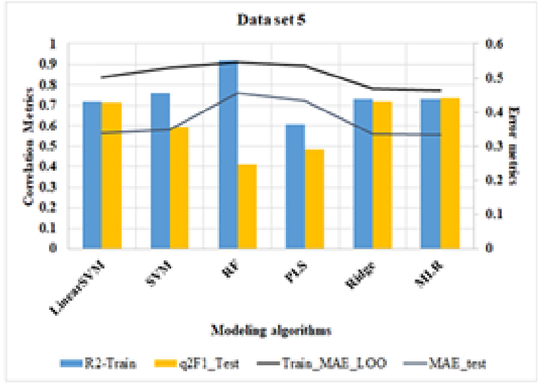
RASAR: Modeling five toxicity endpoints





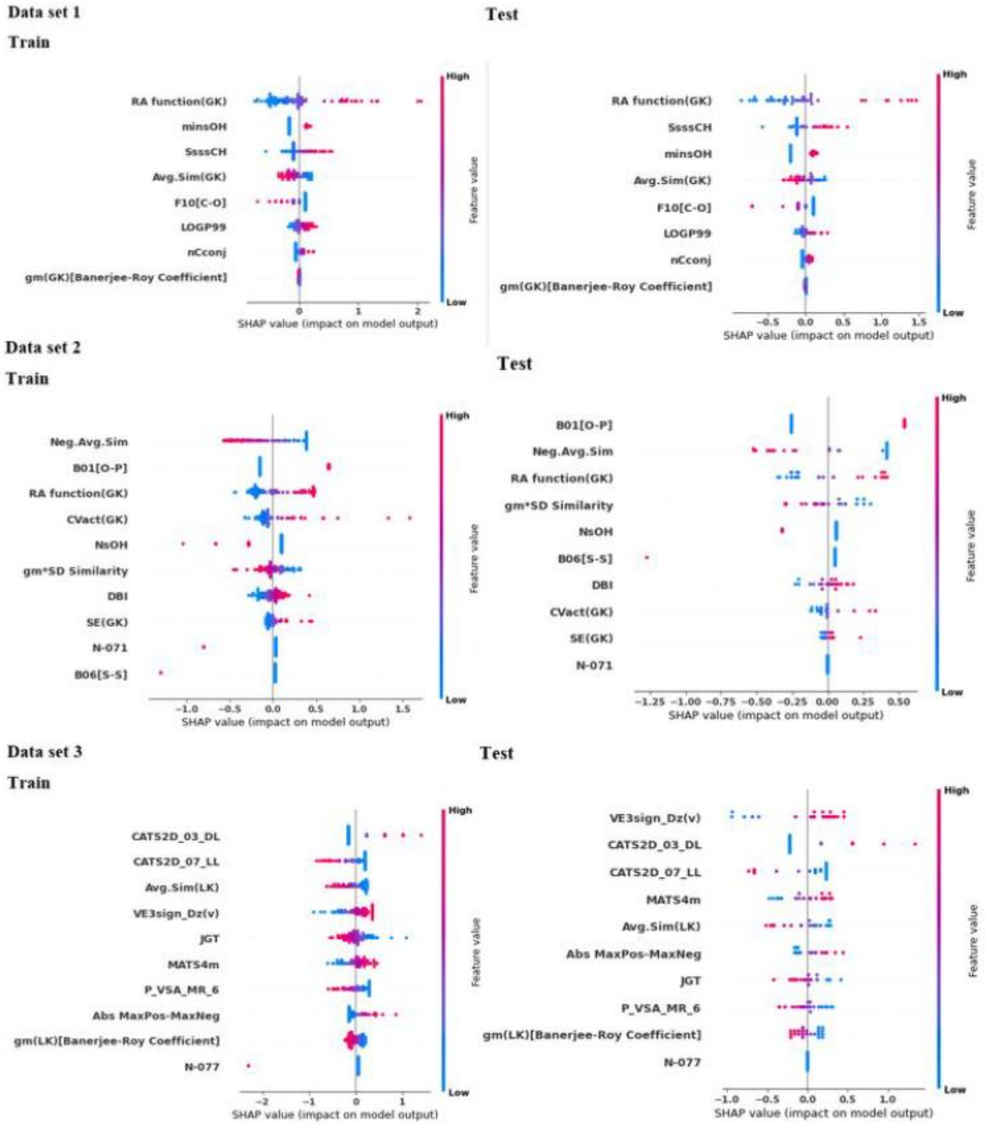






Banerjee and Roy, Chem Res Toxicol, 2023

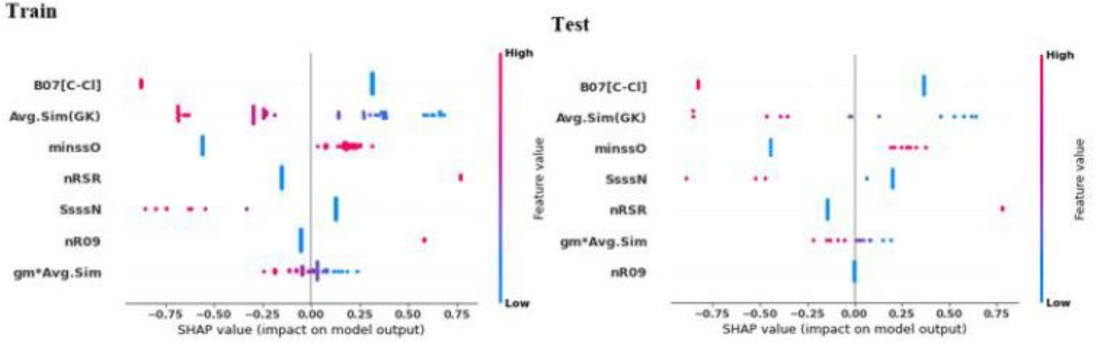
RASAR: Modeling five toxicity endpoints



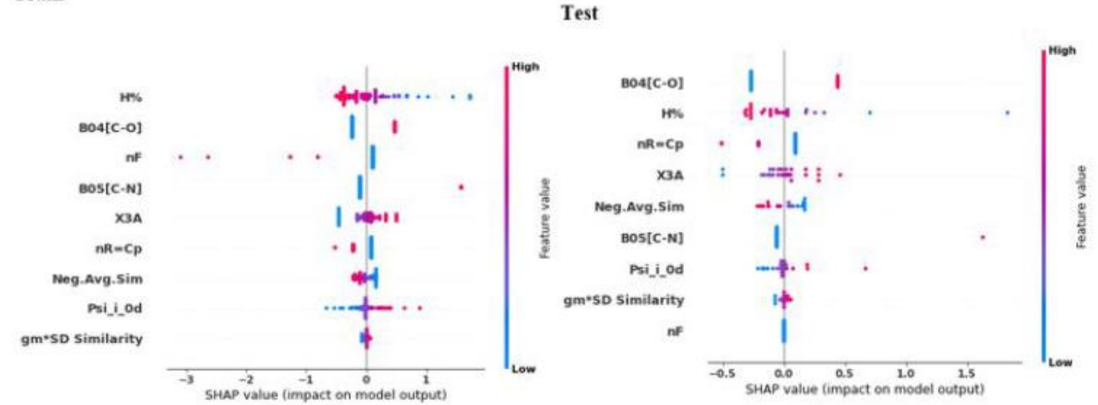
Banerjee and Roy, Chem Res Toxicol, 2023

RASAR: Modeling five toxicity endpoints

Data set 4

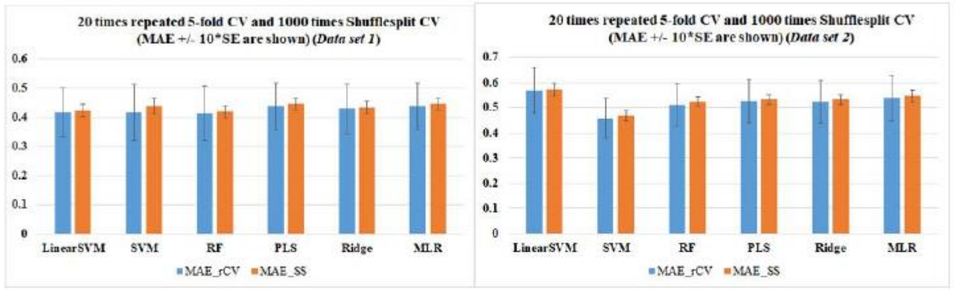


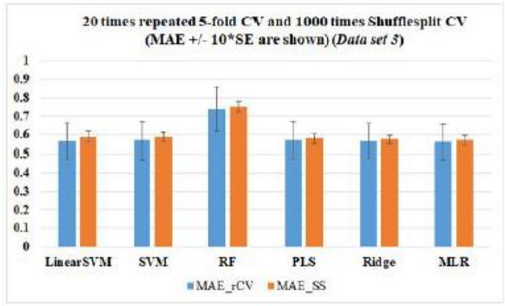
Data set 5

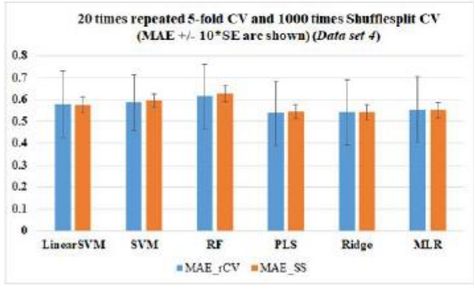


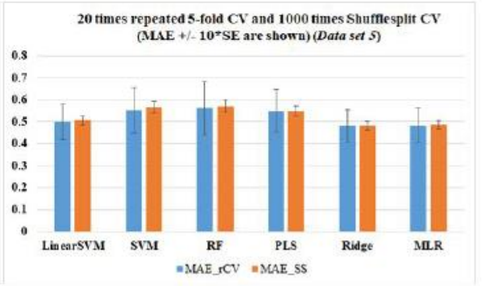
Banerjee and Roy, Chem Res Toxicol, 2023

RASAR: Modeling five toxicity endpoints





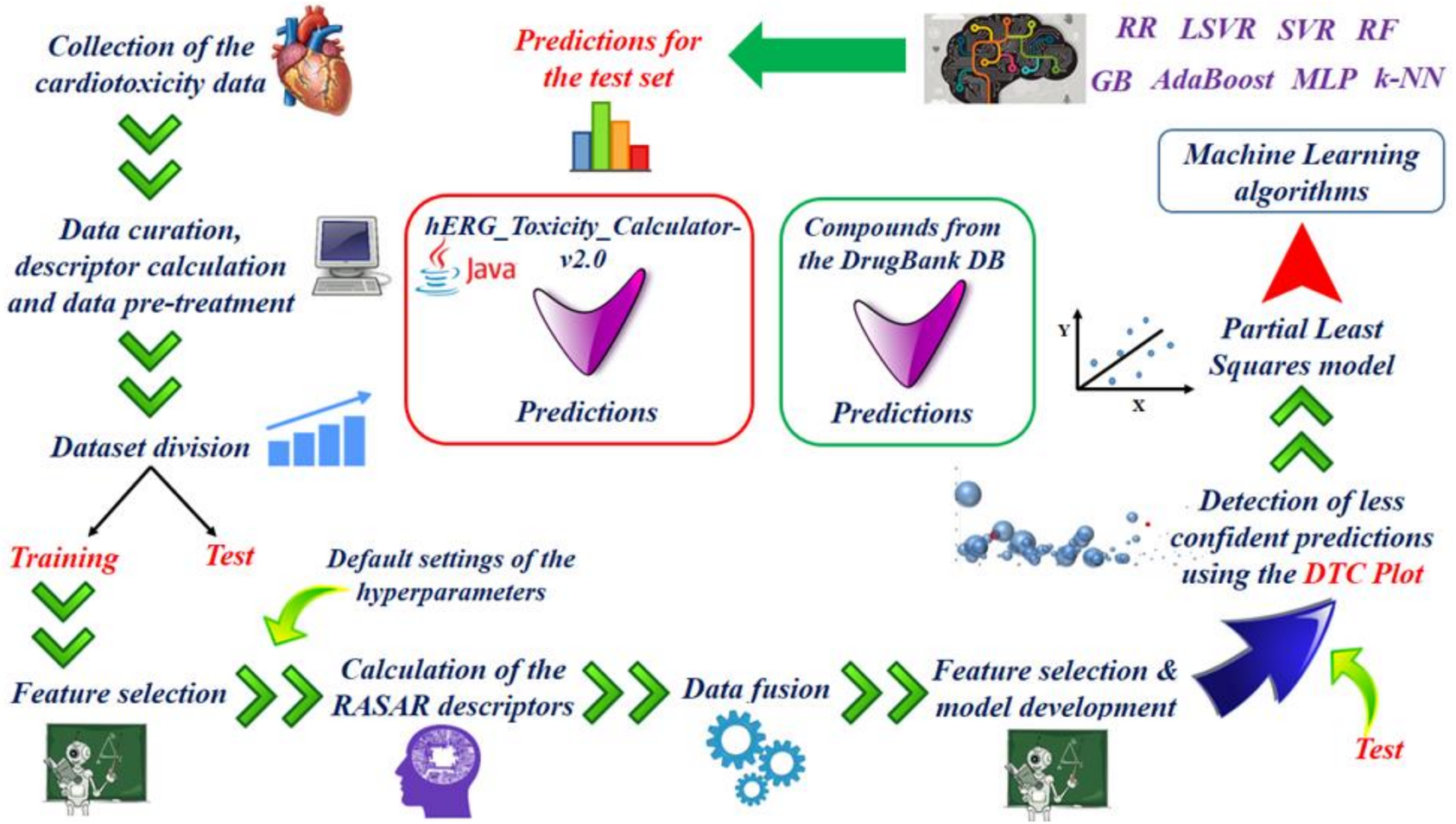




Banerjee and Roy, Chem Res Toxicol, 2023

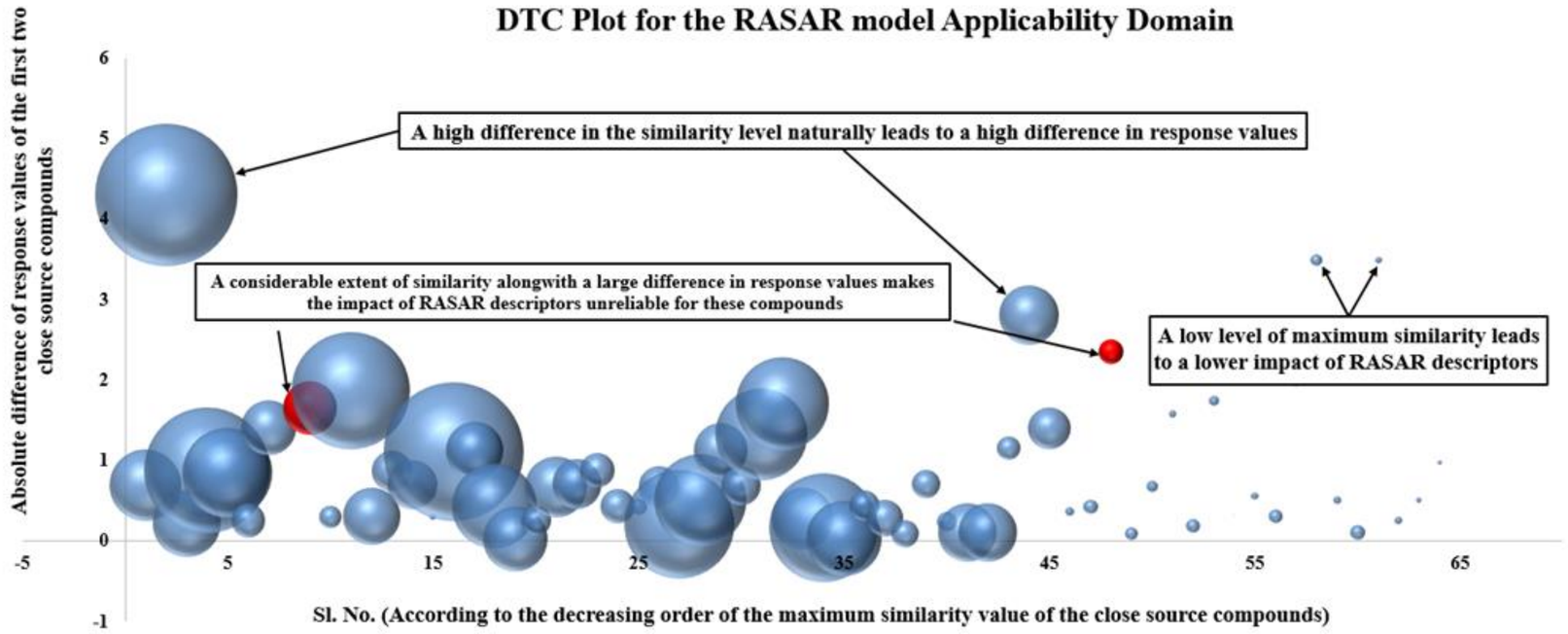
RASAR: Modeling Cardiotoxicity data

261 Compounds



Banerjee and Roy, Unpublished, 2023

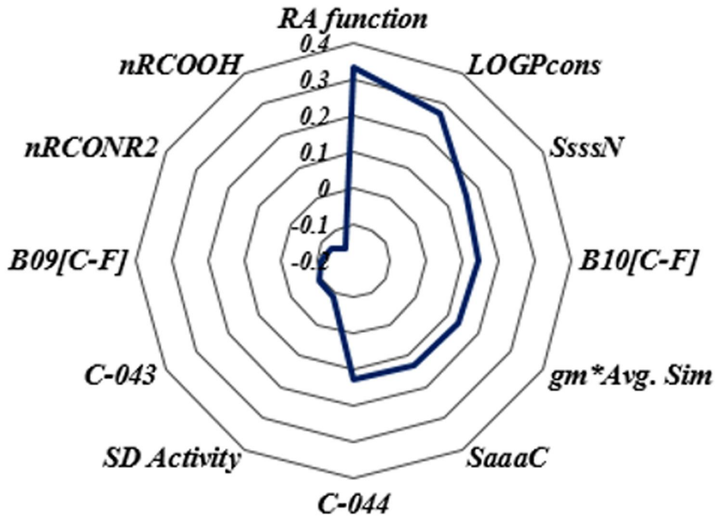
RASAR: Modeling Cardiotoxicity data



Banerjee and Roy, Unpublished, 2023

RASAR: Modeling Cardiotoxicity data

Radar Plot of the std. regression coefficients of the PLS q-RASAR model

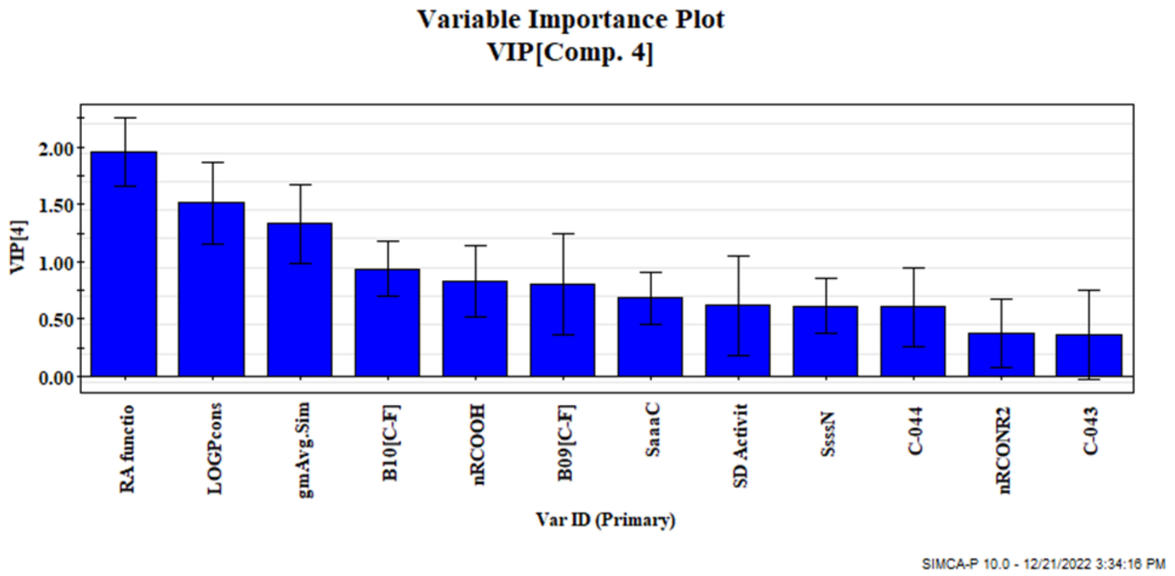


$nTrain = 196 \quad nTest = 63 \quad R^2_{Train} = 0.608 \quad Q^2_{(LOO)} = 0.546$

$Q^2_{F1} = 0.660 \quad Q^2_{F2} = 0.660 \quad MAE_{Train} = 0.581 \quad MAE_{Test} = 0.548$

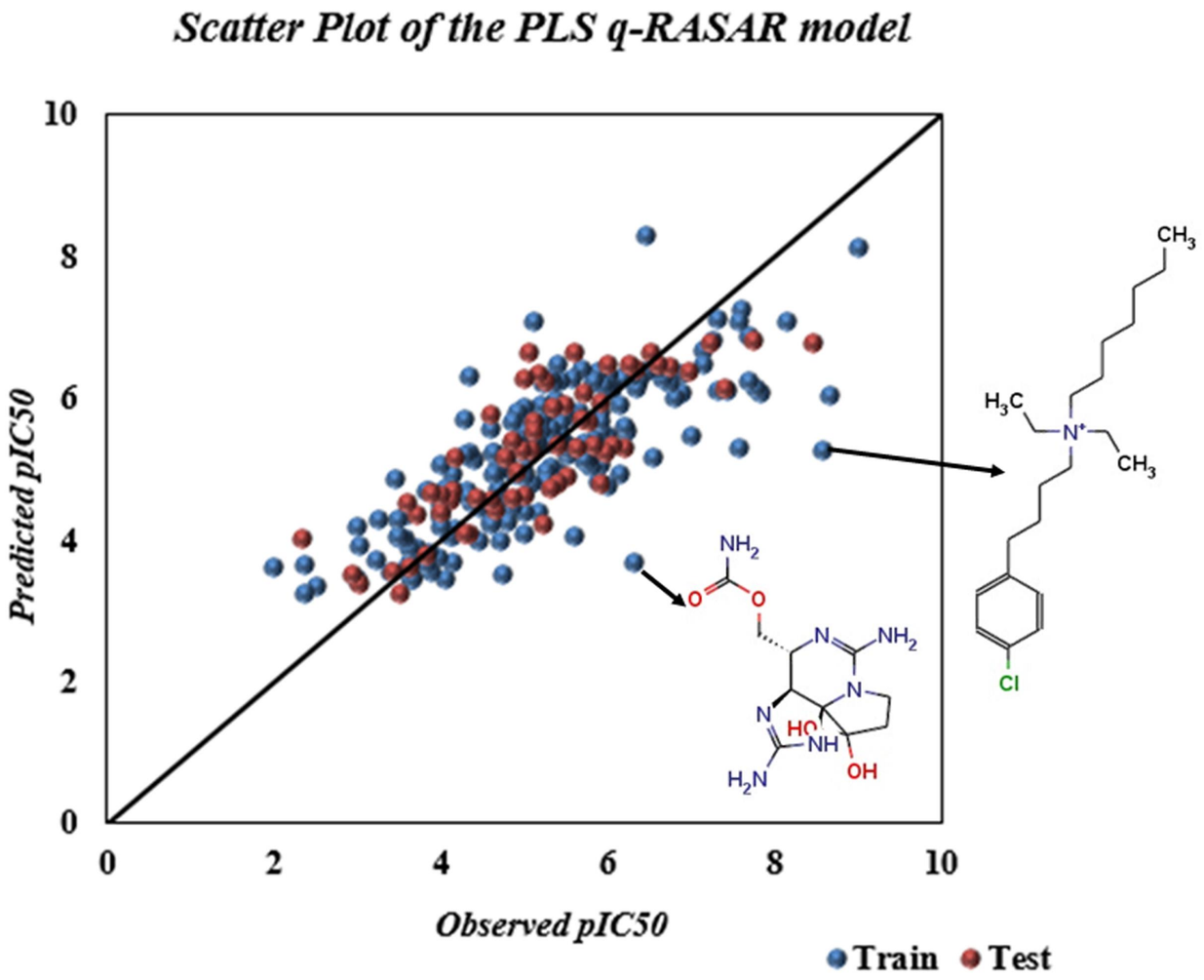
Banerjee and Roy, Unpublished, 2023

RASAR: Modeling Cardiotoxicity data



Banerjee and Roy, Unpublished, 2023

RASAR: Modeling Cardiotoxicity data



9/20/2023

Banerjee and Roy, Unpublished, 2023

RASAR: Modeling Cardiotoxicity data

Model Type	Training set statistics								Test set statistics			Optimum hyperparameters
	R^2_{Train}	MAE_{Train}	$MAE_{(LOO)}$	$MSE_{(LOO)}$	MAE ± SEM (20 times 5 fold CV)	R^2±SEM (20 times 5 fold CV)	MAE ± SEM (Shufflesplit CV, n_splits = 1000)	R^2±SEM (Shufflesplit CV, n_splits = 1000)	R^2_{Test}	MAE_{Test}	Q^2_{F1}	
PLS	0.608	0.466	0.499	0.452	0.502 ± 0.006	0.520 ± 0.013	0.508 ± 0.012	0.518 ± 0.003	0.66	**0.44**	0.66	n_components =4
RR	0.608	0.465	0.5	0.454	0.504 ± 0.006	0.517 ± 0.013	0.509 ± 0.002	0.515 ± 0.003	0.66	0.442	**0.661**	alpha = 0.5
LSVR	0.593	0.45	**0.478**	**0.433**	0.504 ± 0.006	0.513 ± 0.013	0.515 ± 0.002	0.502 ± 0.004	0.64	0.466	0.641	C=15.0, max_iter =1000000
SVR	0.676	0.378	0.503	0.479	0.513 ± 0.007	0.493 ± 0.011	0.524 ± 0.002	0.486 ± 0.003	0.63 9	0.468	0.64	degree=2, gamma='auto'

DTC LAB

Banerjee and Roy, Unpublished, 2023

RASAR: Modeling Cardiotoxicity data

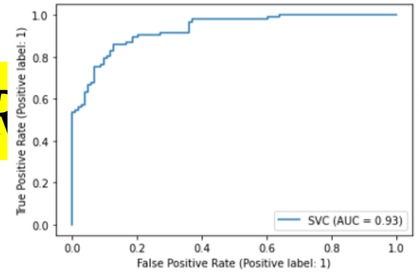
RF	0.733	0.398	0.548	0.551	0.550 ± 0.007	0.427 ± 0.013	0.554 ± 0.002	0.433 ± 0.003	0.584	0.496	0.585	max_depth=4, n_estimators=2(random_state=0
Gradboost	**0.803**	**0.337**	0.536	0.521	0.551 ± 0.007	0.422 ± 0.014	0.559 ± 0.018	0.418 ± 0.009	0.65	0.462	0.651	max_depth=2, min_samples_split=6
Adaboost	0.685	0.45	0.568	0.558	0.557 ± 0.007	0.425 ± 0.013	0.565 ± 0.002	0.424 ± 0.003	0.584	0.49	0.585	learning_rate=0.1, loss='square', r
MLP regression	0.608	0.465	0.499	0.45	0.501 ± 0.006	0.524 ± 0.012	0.505 ± 0.002	0.524 ± 0.003	0.658	0.443	0.659	activation='logistic', alpha=1, hidden_layer_sizes=(1000,), learning_rate_init=0.01, max_iter random_state=0, solver='lbfgs'
kNN regression	0.572	0.489	0.562	0.565	0.573±0.007	0.396± 0.016	0.583 ± 0.002	0.398±0.004	0.526	0.536	0.527	leaf_size=5, n_neighbors=6

550

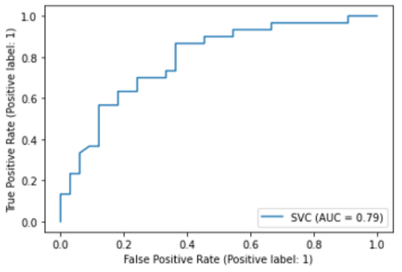
551

Banerjee and Roy, Unpublished, 2023

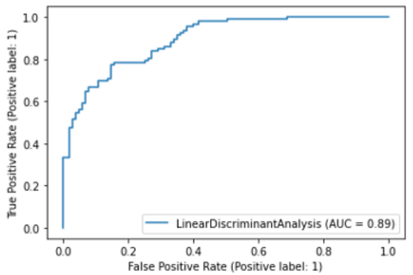
RASAR [illegible] data



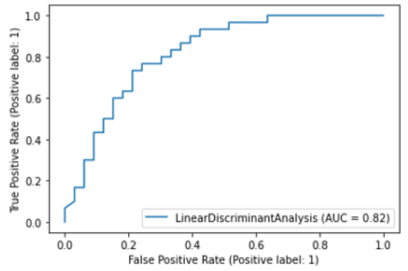
SVC TRAIN



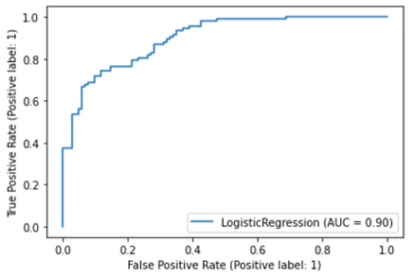
SVC TEST



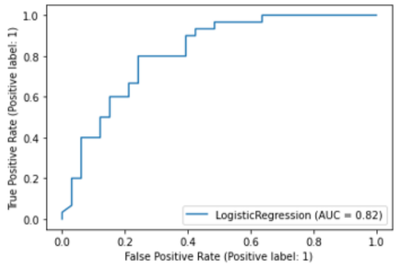
LDA TRAIN



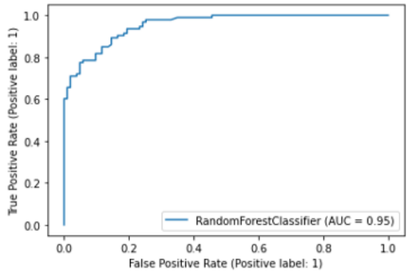
LDA TEST



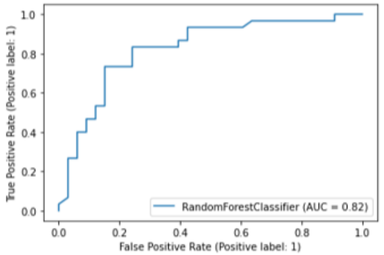
LR TRAIN



LR TEST



RFC TRAIN



RFC TEST

Banerjee and Roy, Unpublished, 2023

RASAR: Modeling Cardiotoxicity data

hERG_Toxicity Calculator v 2.0

This tool quickly provides the quantitative prediction (along with AD) of the potential cardiotoxicity induced by a compound by interacting with the hERG K^+ channel using a q-RASAR model developed by the DTC Laboratory (Banerjee and Roy, 2023, Unpublished) .

Predictions to be used for research purposes only.

Banerjee and Roy, Unpublished, 2023

DTC Lab Tools

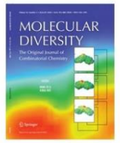
Quantitative Read-Across

V4.1

Chatterjee M, Banerjee A, De P, Gajewicz A, Roy K
Environ Sci: Nano 2021 DOI: 10.1039/D1EN00725D
Banerjee A, Roy K, *Mol Divers*, 2022, DOI: 10.1007/s11030-022-10478
Software developed by Arkaprava Banerjee (arka.banerjee16@gmail.com)
Picture Courtesy Shutterstock

RASAR Descriptor Calculator

v2.0



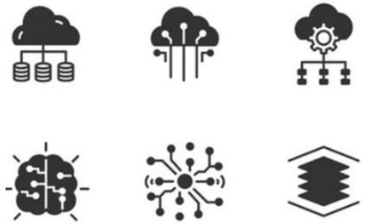
Banerjee A, Roy K, *Mol Divers*, 2022, DOI: 10.1007/s11030-022-10478-6
Banerjee A, Chatterjee M, De P, Roy K, *Chemom Intell Lab Sys*, 227, 2022, DOI: 10.1016/j.chemolab.2022.104613
Software developed by Arkaprava Banerjee (arka.banerjee16@gmail.com)

Auto RA Optimzer

Auto RA Optimizer

v1.0

Chatterjee M, Banerjee A, De P, Gajewicz A, Roy K
Environ Sci: Nano 2021 DOI: 10.1039/D1EN00725D
Software developed by Arkaprava Banerjee (arka.banerjee16@gmail.com)
Picture Courtesy Shutterstock



Machine Learning Regression

Beta version

These GUIs use scikit-learn libraries to optimize hyperparameters and develop machine learning regression models
Software developed by Souvik Pore (souvikpore123@gmail.com)
Picture Courtesy Shutterstock

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

RASAR: Modeling Property data

Received: 17 November 2022 | Revised: 5 January 2023 | Accepted: 6 January 2023

DOI: 10.1002/minf.202200261

RESEARCH ARTICLE

molecular informatics

A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites

Arkaprava Banerjee[1] | **Agnieszka Gajewicz-Skretna**[2] | **Kunal Roy**[1]

[1]Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India

[2]Laboratory of Environmental Chemoinformatics, Faculty of Chemistry, University of Gdansk,Gdansk, Poland

Correspondence
Kunal Roy, Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India.
Email: kunal.roy@jadavpuruniversity.in

Funding information
Science and Engineering Research Board (SERB), Grant/Award Number: MTR/2019/000008; Life Science Research Board, DRDO, India, Grant/Award Number: LSRB/01/15001/M/LSRB-394/SH&DD/2022

Abstract

In this study, the specific surface area of various perovskites was modeled using a novel quantitative read-across structure-property relationship (q-RASPR) approach, which clubs both Read-Across (RA) and quantitative structure-property relationship (QSPR) together. After optimization of the hyper-parameters, certain similarity-based error measures for each query compound were obtained. Clubbing some of these error-based measures with the previously selected features along with the Read-Across prediction function, a number of machine learning models were developed using Partial Least Squares (PLS), Ridge Regression (RR), Linear Support Vector Regression (LSVR), Random Forest (RF) regression, Gradient Boost (GBoost), Adaptive Boosting (Adaboost), Multiple Layer Perceptron (MLP) regression and k-Nearest Neighbor (kNN) regression. Based on the repeated cross-validation as well as external prediction quality and interpretability, the PLS model ($n_{Training} = 38$, $n_{Test} = 12$, $R^2_{Train} = 0.737$, $Q^2_{LOO} = 0.637$, $R^2_{Test} = 0.898$, $Q^2_{F1(Test)} = 0.901$) was selected as the best predictor which underscored the previously reported results. The finally selected model should efficiently predict specific surface areas of other perovskites for their use in photocatalysis. The new q-RASPR method also appears promising for the prediction of several other property endpoints of interest in materials science.

KEYWORDS
machine learning, perovskites, photocatalysis, q-RASPR, specific surface area

RASAR: Modeling Property data

Sustainable Energy & Fuels

ROYAL SOCIETY OF CHEMISTRY

PAPER

Check for updates

Cite this: *Sustainable Energy Fuels*, 2023, 7, 3412

Machine learning-based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells†

Souvik Pore, Arkaprava Banerjee and Kunal Roy *

Different computational tools are now popularly used as an alternative to experiments for predicting several property endpoints of industrial importance. Recently, read-across and quantitative structure–property relationship (QSPR) have been merged to develop a new modeling technique read-across structure–property relationship (RASPR) which appears to have much potential in predictive modeling. This approach is also promising for modeling relatively smaller data sets as the similarity-based RASPR descriptors are computed from multiple structural and physicochemical features. To understand the potential of RASPR in data gap filling, we have undertaken a case study of modeling Power Conversion Efficiency (PCE) of different classes of organic dyes used in Dye-Sensitized Solar Cells (DSSCs) for renewable energy generation. We have used a large dataset of 429 compounds covering 4 classes of organic dyes. We initially performed read-across analysis using different similarity measures with structural analogues for query compounds and calculated the weighted average predictions. Based on the read-across optimized settings, RASPR descriptors were calculated, and these were then merged with the chemical descriptors, and finally, a single partial least squares (PLS) model was developed for each of the dye classes after feature selection, followed by additional Machine Learning (ML) models. The external prediction quality of the final RASPR models superseded those of the previously developed QSPR models using the same level of chemical information. The important structural features and similarity measures contributing to the PCE have been extracted using the RASPR method which can be used to enhance the PCE values in the newly designed dyes. The RASPR method may also be efficiently applied in modeling other properties of interest in a similar manner.

Received 7th April 2023
Accepted 20th June 2023

DOI: 10.1039/d3se00457k

rsc.li/sustainable-energy

RASAR: Modeling Skin Sensitization data

Chemical Research in Toxicology

pubs.acs.org/crt Article

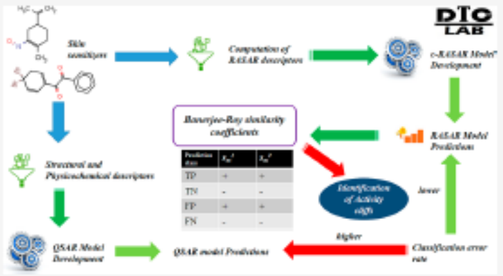
Prediction-Inspired Intelligent Training for the Development of Classification Read-across Structure–Activity Relationship (c-RASAR) Models for Organic Skin Sensitizers: Assessment of Classification Error Rate from Novel Similarity Coefficients

Arkaprava Banerjee and Kunal Roy*

Cite This: https://doi.org/10.1021/acs.chemrestox.3c00155 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The advancements in the field of cheminformatics have led to a reduction in animal testing to estimate the activity, property, and toxicity of query chemicals. Read-across structure–activity relationship (RASAR) is an emerging concept that utilizes various similarity functions derived from chemical information to develop highly predictive models. Unlike quantitative structure–activity relationship (QSAR) models, RASAR descriptors of a query compound are computed from its close congeners instead of the compound itself, thus targeting predictions in the model training phase. The objective of the present study is not to propose new QSAR models for skin sensitization but to demonstrate the enhancement in the quality of predictions of the skin-sensitizing potential of organic compounds by developing classification-based RASAR (c-RASAR) models. A diverse, previously curated data set was collected from the literature for which 2D descriptors were computed. The extracted essential features were then used to develop a classification-based linear discriminant analysis (LDA) QSAR model. Furthermore, from the read-across-based predictions, RASAR descriptors were calculated using the basic settings of the hyperparameters for the Laplacian Kernel-based optimum similarity measure. After feature selection, an LDA c-RASAR model was developed, which superseded the prediction quality of the LDA–QSAR model. Various other combinations of RASAR descriptors were also taken to develop additional c-RASAR models, all showing better prediction quality than the LDA QSAR model while using a lower number of descriptors. Various other machine learning c-RASAR models were also developed for comparison purposes. In this work, we have proposed and analyzed three new similarity metrics: g_{m_class}, s^1_m, and s^2_m. The first one is an indicator variable used to generate a simple univariate c-RASAR model with good prediction ability, while the remaining two are similarity indices used to analyze possible activity cliffs in the training and test sets and are believed to play an important role in the modelability analysis of data sets.



RASAR: Modeling Skin Sensitization data

Environmental Science Processes & Impacts

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

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

Received 26th July 2023
Accepted 4th September 2023

DOI: 10.1039/d3em00322a

RASAR: Modeling honey bee toxicity data

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Research Paper

Machine learning - based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees

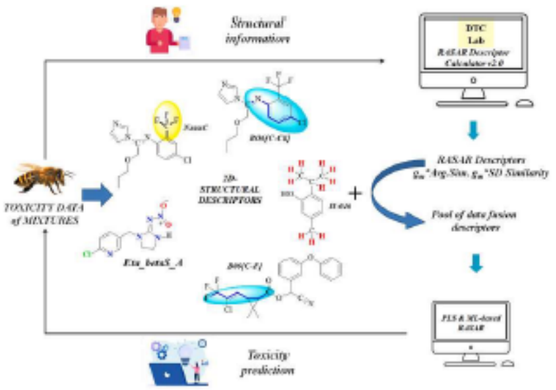
Mainak Chatterjee [a], Arkaprava Banerjee [a], Simone Tosi [b], Edoardo Carnesecchi [c], Emilio Benfenati [d], Kunal Roy [a,*]

[a] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
[b] Department of Agricultural, Forest, and Food Sciences, University of Turin, Turin, Italy
[c] Institute for Risk Assessment Sciences, Utrecht, the Netherlands
[d] Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri IRCSS, via Mario Negri 2, 20156 Milano, Italy

HIGHLIGHTS

- A novel q-RASAR model has been developed for the prediction of toxicity of organic mixtures in honey bees.
- Three different mixing rules have been used to calculate the mixture descriptors.
- The developed model has been validated following the strict OECD guidelines.
- The use of machine learning-based algorithms further enhanced the predictability of the q-RASAR model.
- The toxicity of environmentally relevant untested organic mixtures has been predicted by this new model.

GRAPHICAL ABSTRACT



RASAR: Modeling Nanotoxicity data

NANOTOXICOLOGY
https://doi.org/10.1080/17435390.2023.2186280

Taylor & Francis
Taylor & Francis Group

ARTICLE

Check for updates

Efficient predictions of cytotoxicity of TiO_2-based multi-component nanoparticles using a machine learning-based q-RASAR approach

Arkaprava Banerjee[a], Supratik Kar[b], Souvik Pore[a] and Kunal Roy[a]

[a]Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India;
[b]Chemometrics & Molecular Modeling Laboratory, Department of Chemistry, Kean University, Union, NJ, USA

ABSTRACT

The availability of experimental nanotoxicity data is in general limited which warrants both the use of *in silico* methods for data gap filling and exploring novel methods for effective modeling. Read-Across Structure-Activity Relationship (RASAR) is an emerging cheminformatic approach that combines the usefulness of a QSAR model and similarity-based Read-Across predictions. In this work, we have generated simple, interpretable, and transferable quantitative-RASAR (q-RASAR) models which can efficiently predict the cytotoxicity of TiO_2-based multi-component nanoparticles. A data set of 29 TiO_2-based nanoparticles with specific amounts of noble metal precursors was rationally divided into training and test sets, and the Read-Across-based predictions for the test set were generated. The optimized hyperparameters and the similarity approach, which yield the best predictions, were used to calculate the similarity and error-based RASAR descriptors. A data fusion of the RASAR descriptors with the chemical descriptors was done followed by the best subset feature selection. The final set of selected descriptors was used to develop the q-RASAR models, which were validated using the stringent OECD criteria. Finally, a random forest model was also developed with the selected descriptors, which could efficiently predict the cytotoxicity of TiO_2-based multi-component nanoparticles superseding previously reported models in the prediction quality thus showing the merits of the q-RASAR approach. To further evaluate the usefulness of the approach, we have applied the q-RASAR approach also to a second cytotoxicity data set of 34 heterogeneous TiO_2-based nanoparticles which further confirmed the enhancement of external prediction quality of QSAR models after incorporation of RASAR descriptors.

ARTICLE HISTORY
Received 15 December 2022
Revised 13 February 2023
Accepted 26 February 2023

KEYWORDS
QSAR; q-RASAR; random forest; machine learning; TiO_2-based nanoparticles

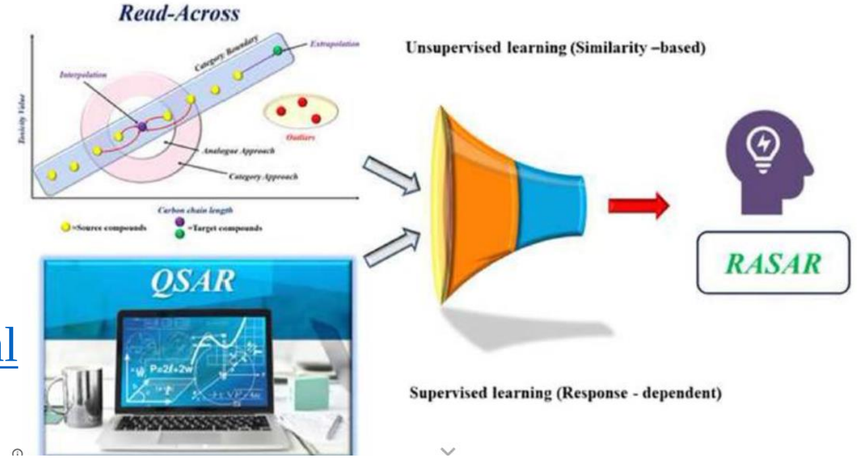
Conclusion

- These studies report the development of simple, interpretable, and reproducible q-RASAR models for various toxicity (activity/property) endpoints.
- The q-RASAR models reported here thus deliver lower prediction errors for the query sets than corresponding QSAR models, suggesting that they are the potential models of choice for efficient predictions using a given level of chemical information.
- Based on the variable importance analysis, the RASAR descriptors “RA score”, “gm” and “average similarity” appear efficient similarity-based determinants for the prediction of toxicity which warrants further extensive studies on these functions.

More about q-RASAR

https://sites.google.com/site/kunalroyindia/home/rasar

https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home#h.i79rttmog6nl



Tools developed by Arkaprava Banerjee

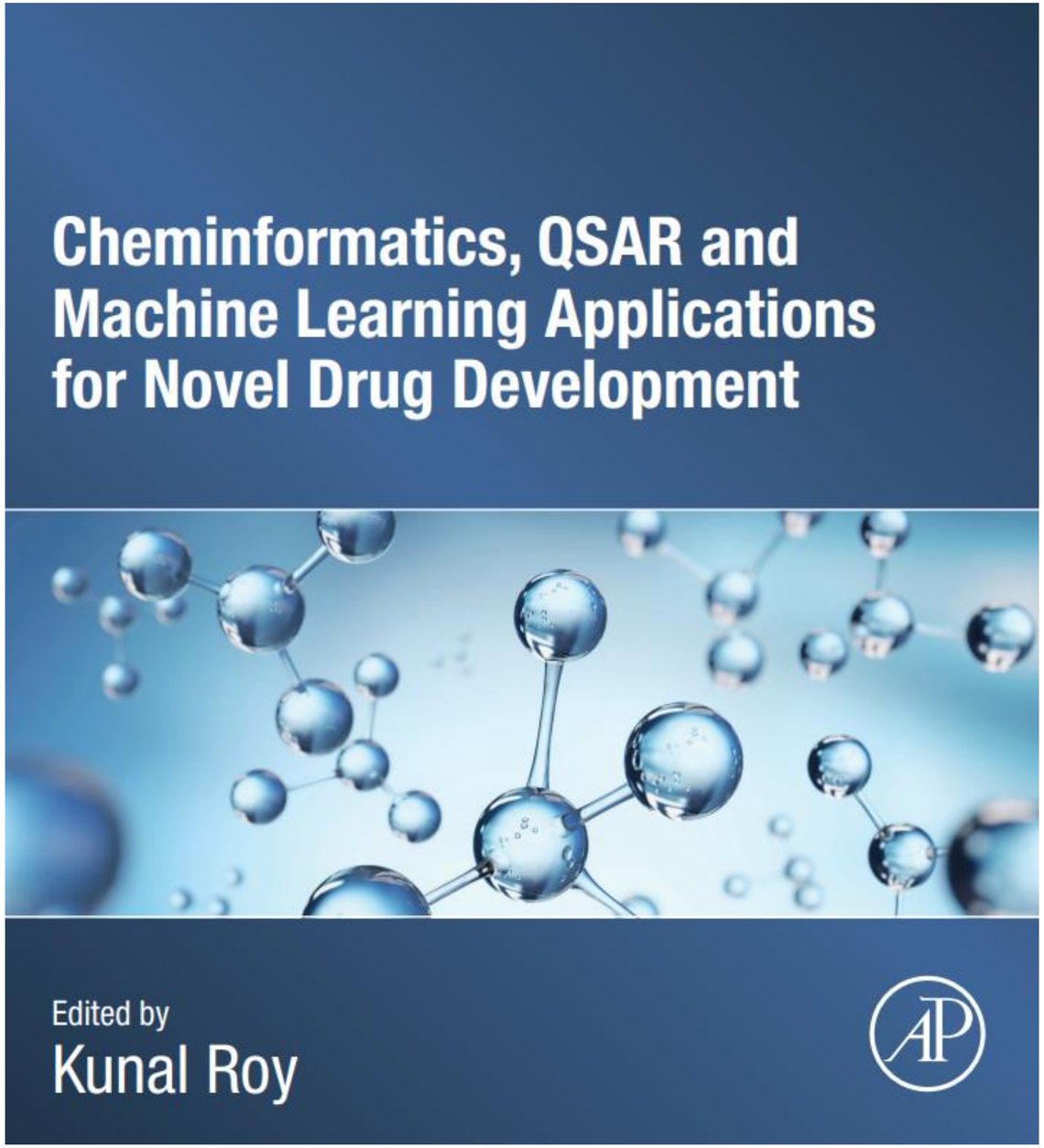
Publications with q-RASAR modeling from other laboratories

- Ecotoxicological QSAR study of fused/non-fused polycyclic aromatic hydrocarbons (FNFPAHs): Assessment and priority ranking of the acute toxicity to Pimephales promelas by QSAR and consensus modeling methods. *Science of The Total Environment*, 876, 162736 (2023)
- In silico assessment of risks associated with pesticides exposure during pregnancy. *Chemosphere*, 329, 138649 (2023)
- Data driven toxicity assessment of organic chemicals against Gammarus species using QSAR approach. *Chemosphere* 328, 138433 (2023)
- QSAR and Chemical Read-Across Analysis of 370 Potential MGMT Inactivators to Identify the Structural Features Influencing Inactivation Potency. *Pharmaceutics* 15, 2170 (2023)

Funding


SERB
DIA
Science and Engineering Research Board
Statutory Body Established through an Act of Parliament: SERB Act 2008
Government of India


सत्यमेव जयते
DRDO
रक्षा अनुसंधान एवं विकास संगठन
रक्षा मंत्रालय, भारत सरकार
DEFENCE RESEARCH &
DEVELOPMENT ORGANISATION
Ministry of Defence, Government of India
Life Sciences
Research Board



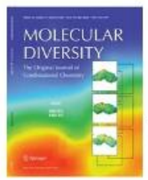
https://www.elsevier.com/books/cheminformatics-qsar-and-machine-learning-applications-for-novel-drug-development/roy/978-0-443-18638-7

DTC Lab Tools

Supplementary site

RASAR

RASAR Descriptor Calculator v2.0



Banerjee A, Roy K, *Mol Divers*, 2022, DOI: 10.1007/s11030-022-10478-6
Banerjee A, Chatterjee M, De P, Roy K, *Chemom Intell Lab Sys*, 227, 2022, DOI: 10.1016/j.chemolab.2022.104613

Software developed by Arkaprava Banerjee (arka.banerjee16@gmail.com)

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

Thank you