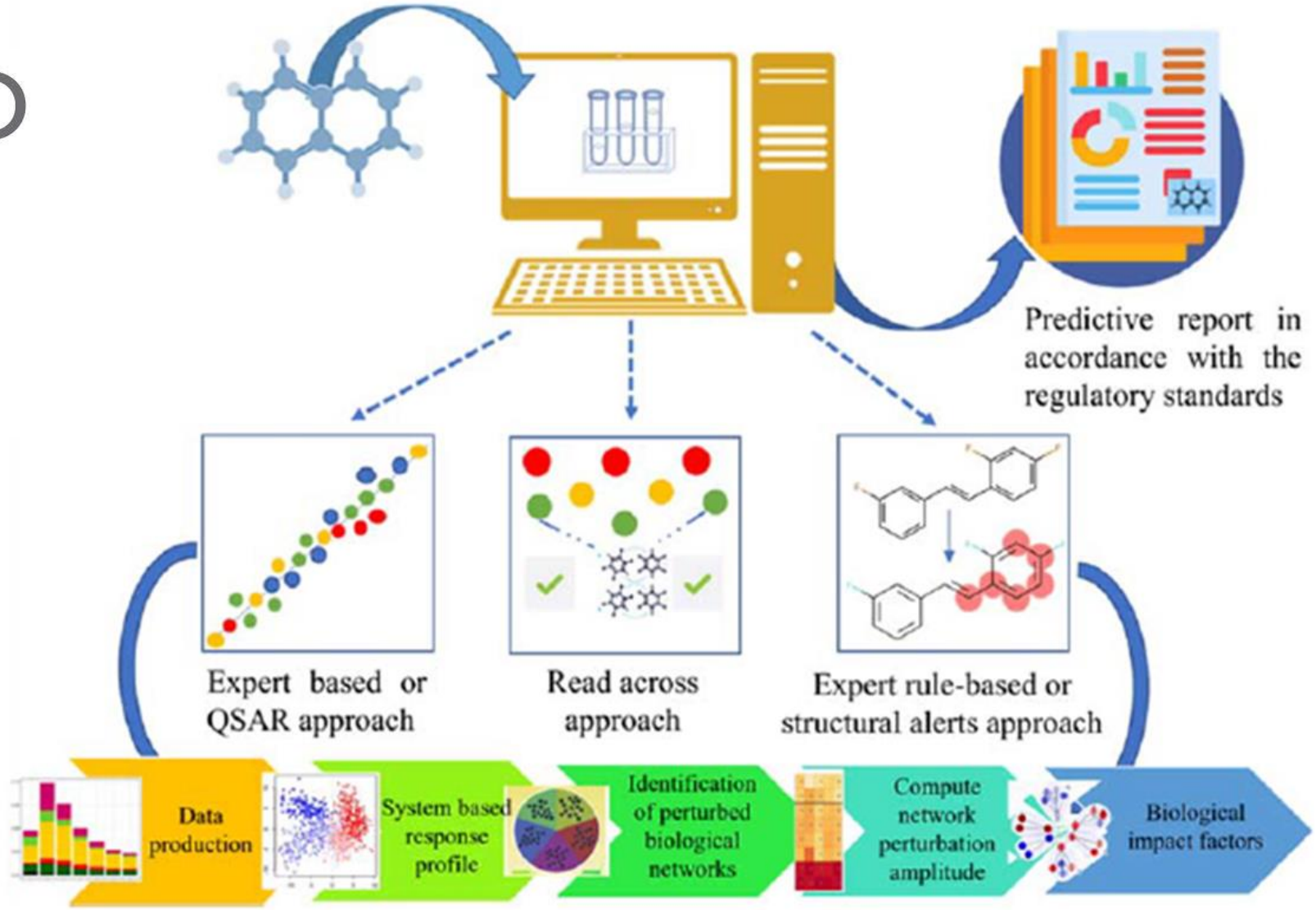
An improved q-RASAR modeling framework for environmental toxicity endpoints

*Arkaprava Banerjee and Kunal Roy**

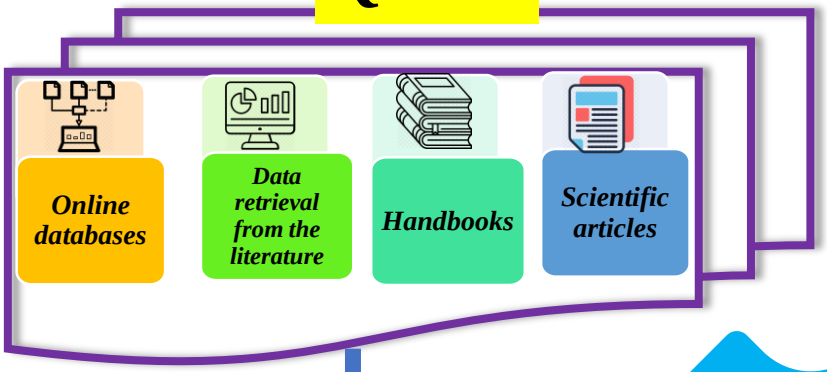
Drug Theoretics and Cheminformatics Laboratory, Jadavpur University, Kolkata, INDIA

Data gap filling approaches



Singh et al., QSAR in Safety Evaluation and Risk Assessment, Elsevier, https://doi.org/10.1016/B978-0-443-15339-6.00026-6

QSAR



DTC LAB

Biological and Chemical curation

Input Data (Activity/Property/Toxicity)

Descriptor calculation

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

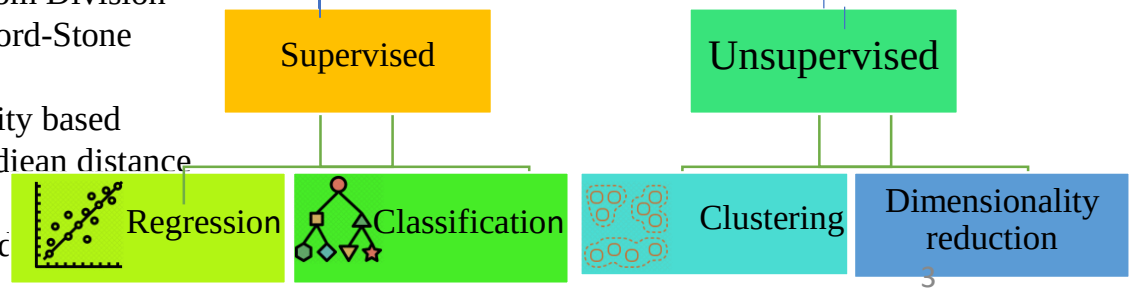
Dataset Division

- Random Division
- Kennord-Stone based
- Activity based
- Euclidean distance based
- K-mec

Feature Selection

- Genetic algorithm
- Lasso feature selection
- Stepwise selection

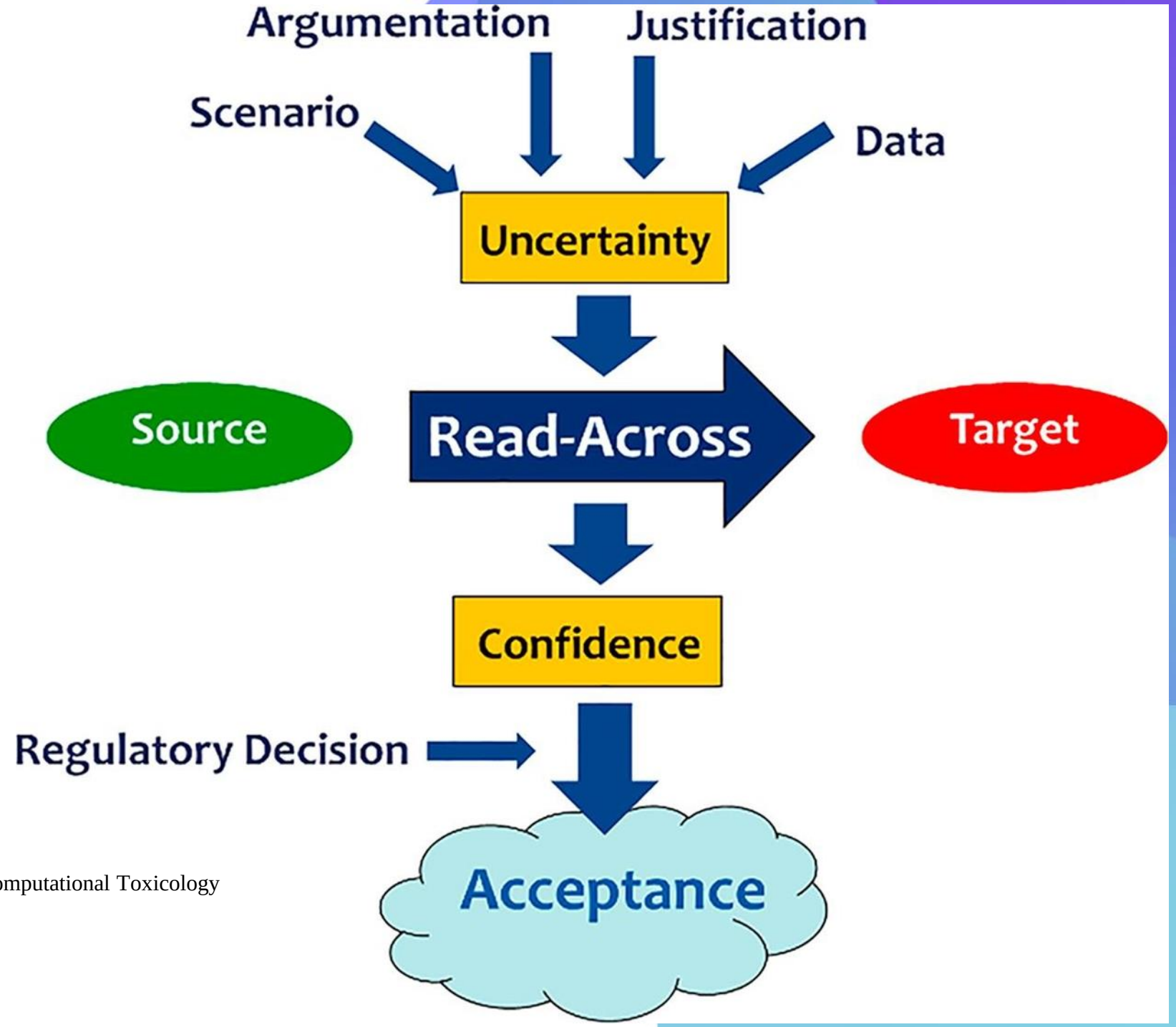
Learning Algorithms



Model analysis

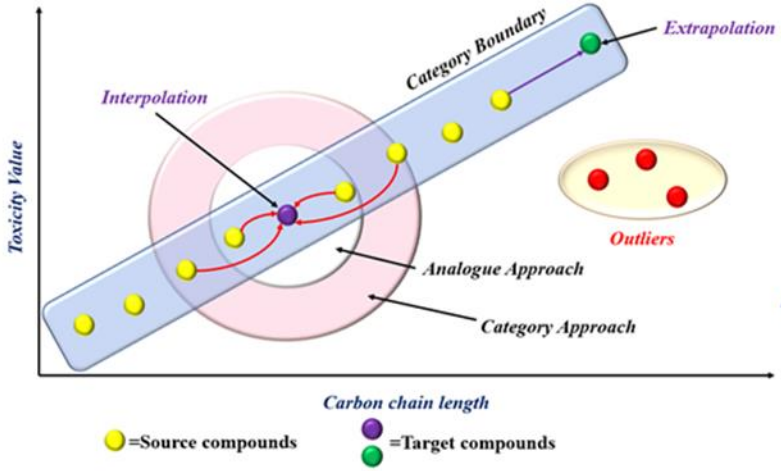
- Predictive variance
- Applicability domain
- Mechanistic interpretation
- Model quality

Roy, Kar and Das, Understanding QSAR.. (Elsevier), 2015

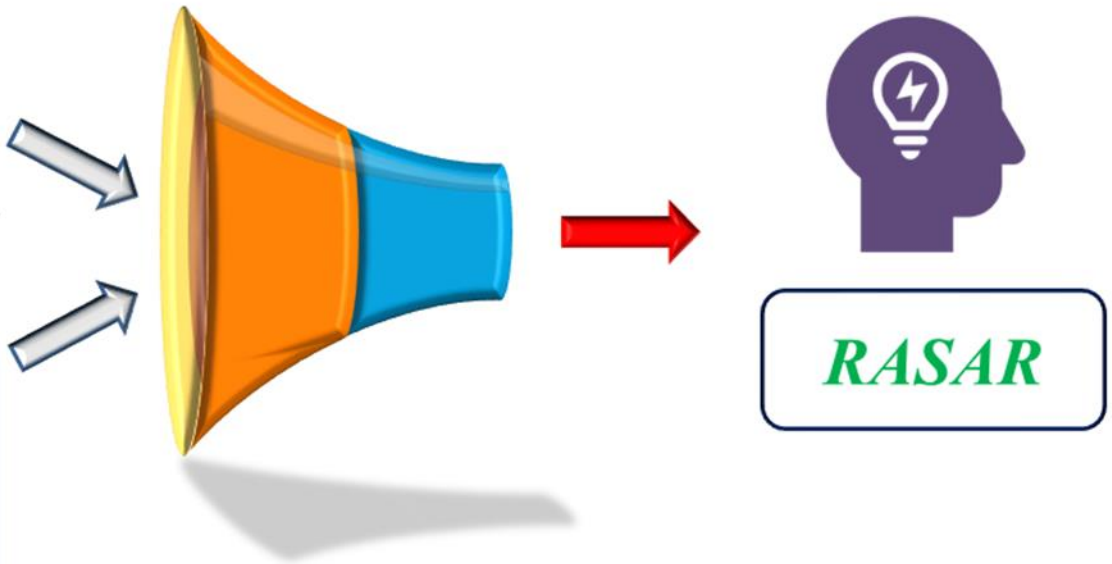


Schultz et al. 2019, Computational Toxicology
Volume 9, Pages 1-11

Read-Across



Unsupervised learning (Similarity –based)

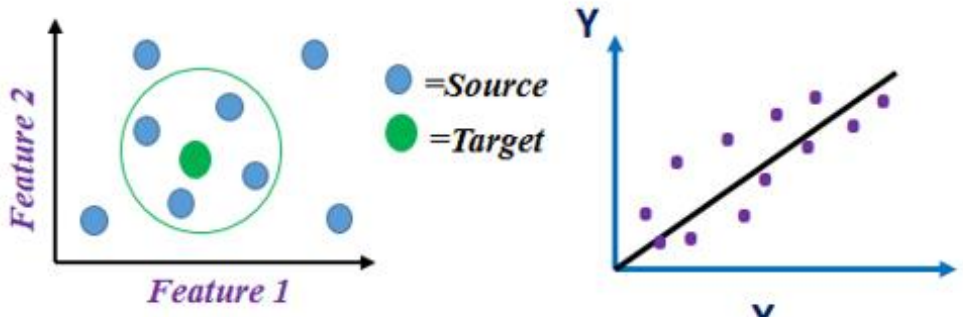


RASAR

Supervised learning (Response - dependent)

Banerjee and Roy, Mol Divers, 202

What is q-RASAR?



Read-Across-based Similarity hypothesis

QSAR

$$Y = f(Similarity, Chemical\ Features)$$

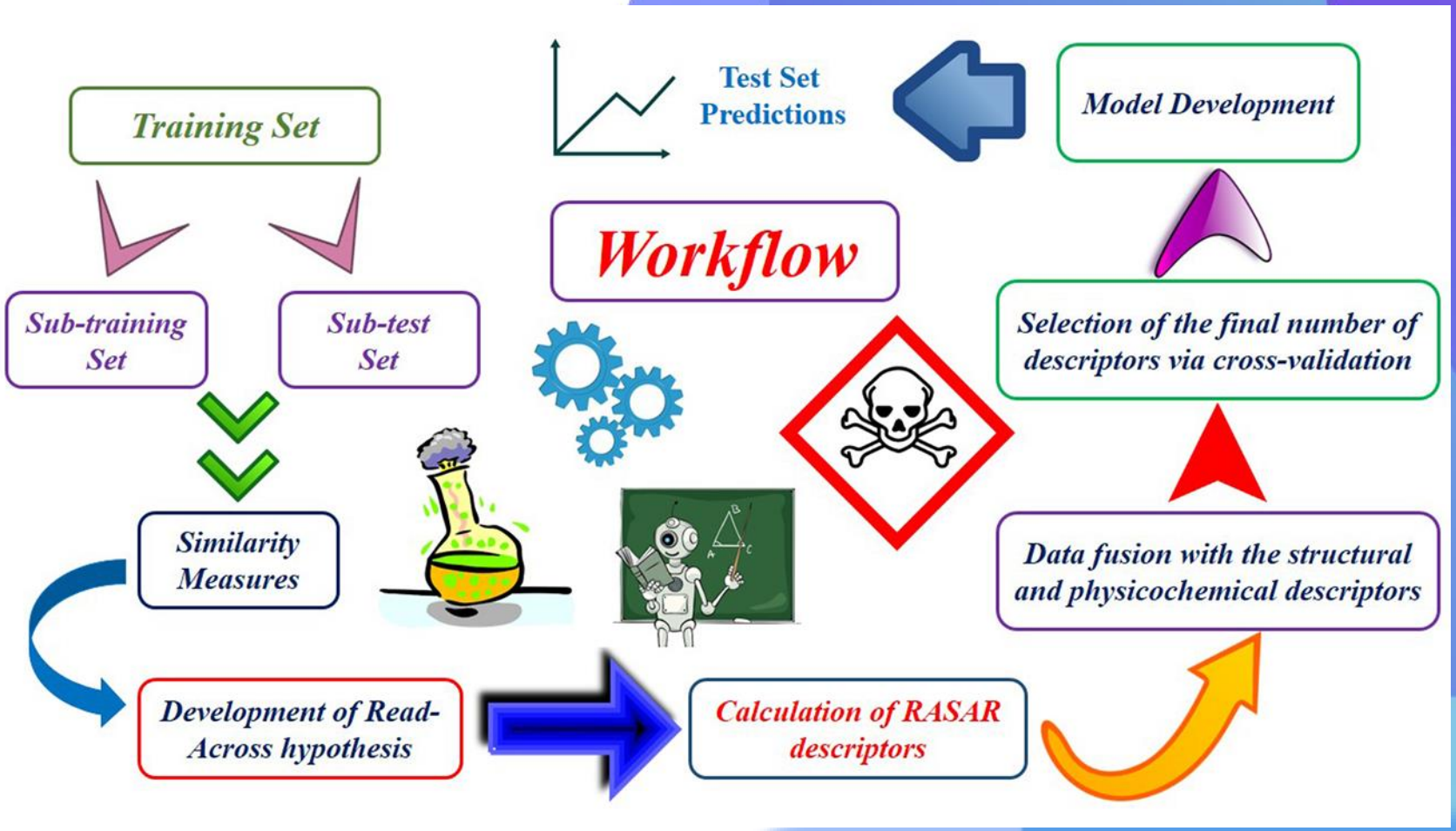
q-RASAR

- ✓ *Quantitative Read-Across Structure-Activity Relationship (q-RASAR)*
- ✓ *Derived from the concepts of QSAR and Read-Across*
- ✓ *Similarity and error-based measures as descriptors*

Banerjee, A.; Roy, K. Mol. Divers. 2022, 26, 2847-2862

Banerjee, A.; Roy, K. Chem. Res. Toxicol. 2023, 36, 446-464

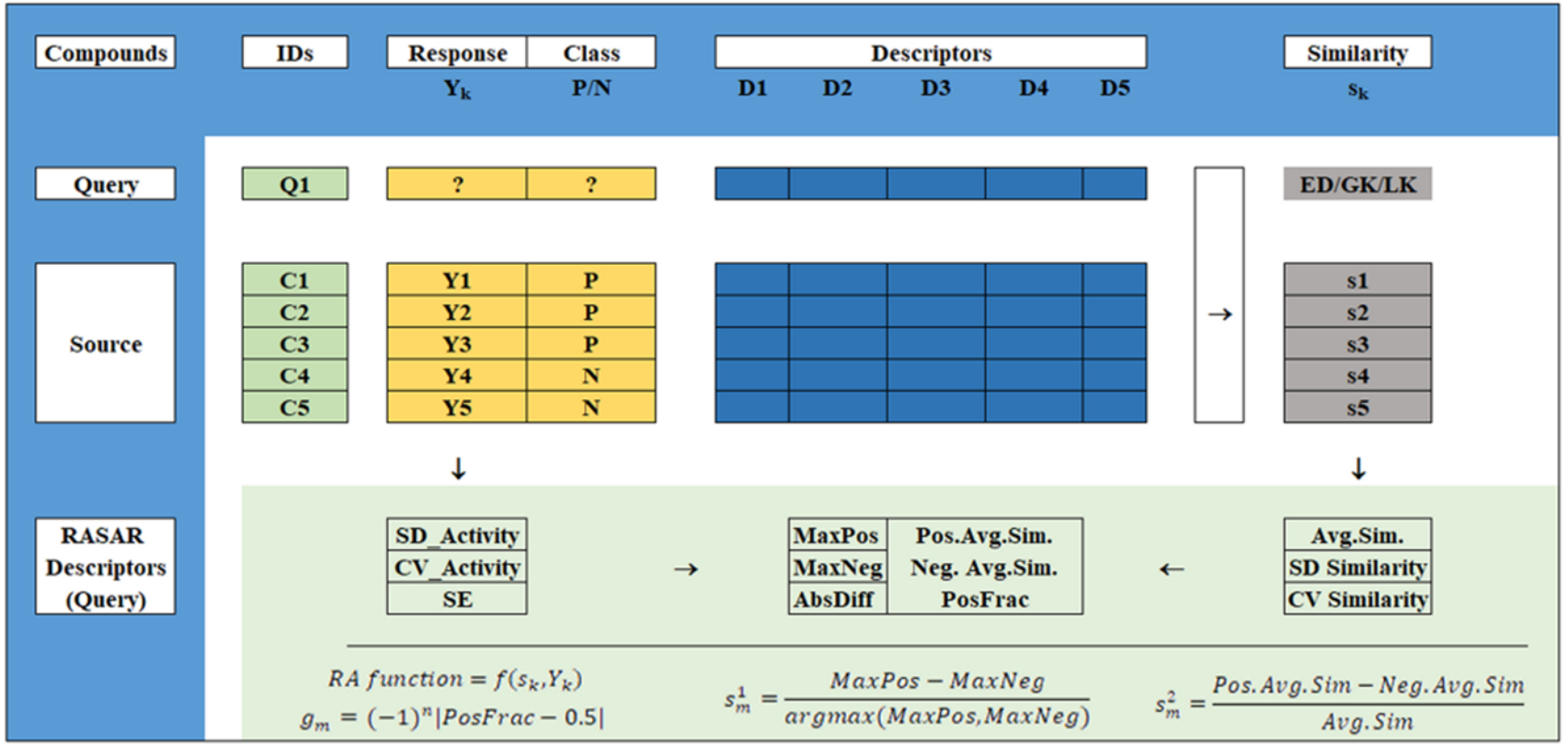
Workflow of q-RASAR



Banerjee, A.; Roy, K. Mol. Divers. 2022, 26, 2847-2862

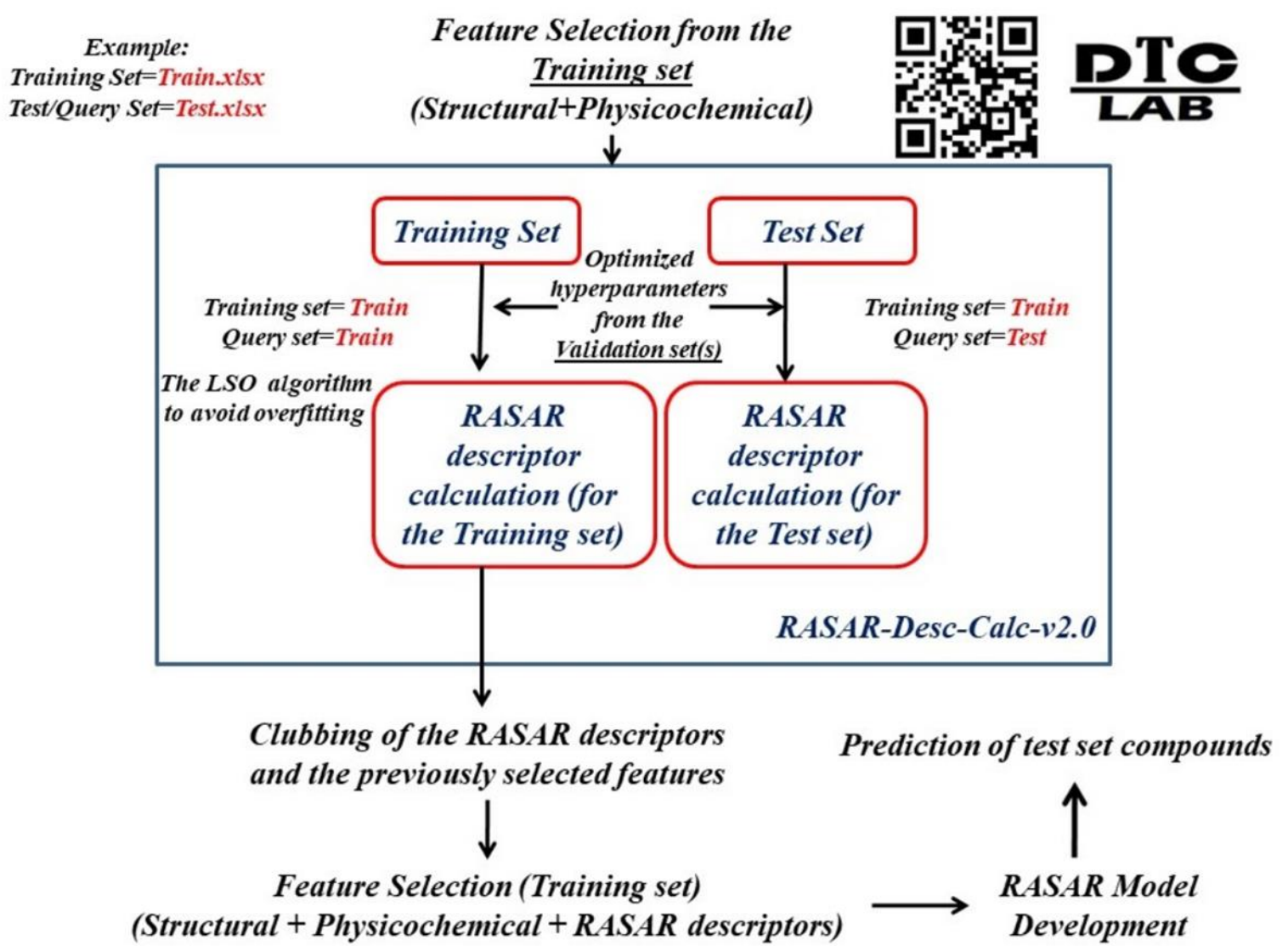
Banerjee, A.; Roy, K. Chem. Res. Toxicol. 2023, 36, 446-464

q-RASAR Descriptors



Banerjee, A.; Roy, K. Mol. Divers. 2022, 26, 2847-2862

Banerjee, A.; Roy, K. Chem. Res. Toxicol. 2023, 36, 446-464



Banerjee, A.; Roy, K. Mol. Divers. 2022, 26, 2847-2862

Banerjee, A.; Roy, K. Chem. Res. Toxicol. 2023, 36, 446-464

QSARs and their limitations

No.	X	Y
1	.	.
2	.	.
3	.	.
4	.	.
5	.	.
6	.	.
7	.	.
8	.	.

$$Contribution_X = f(X, Y)_{(1-8)}$$

Role of the ARKA descriptors

No.	X	Y
1	.	1
2	.	≥ 0.75
3	.	< 0.75
4	.	≥ 0.5
5	.	< 0.5
6	.	≥ 0.25
7	.	< 0.25
8	.	0

ARKA = Arithmetic Residuals in K-groups Analysis

$$Contribution_{X(1,2)} = f(X,Y)_{(1,2)}$$
$$Contribution_{X(3,4)} = f(X,Y)_{(3,4)}$$
$$Contribution_{X(5,6)} = f(X,Y)_{(5,6)}$$
$$Contribution_{X(7,8)} = f(X,Y)_{(7,8)}$$

Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2024, 26, 991-1007

Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2025.

Grouping strategy in Multiclass ARKA

No.	X	Y
1	.	1
2	.	≥ 0.75
3	.	< 0.75
4	.	≥ 0.5
5	.	< 0.5
6	.	≥ 0.25
7	.	< 0.25
8	.	0

Iteration #1

= Positive class

= Negative class

Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2025

Grouping strategy in Multiclass ARKA

No.	X	Y
1	.	1
2	.	≥ 0.75
3	.	< 0.75
4	.	≥ 0.5
5	.	< 0.5
6	.	≥ 0.25
7	.	< 0.25
8	.	0

Iteration #2

= Positive class

= Negative class

Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2025

Grouping strategy in Multiclass ARKA

No.	X	Y
1	.	1
2	.	≥ 0.75
3	.	< 0.75
4	.	≥ 0.5
5	.	< 0.5
6	.	≥ 0.25
7	.	< 0.25
8	.	0

Iteration #3

= Positive class

= Negative class

Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2025

Grouping strategy in Multiclass ARKA

No.	X	Y
1	.	1
2	.	≥ 0.75
3	.	< 0.75
4	.	≥ 0.5
5	.	< 0.5
6	.	≥ 0.25
7	.	< 0.25
8	.	0

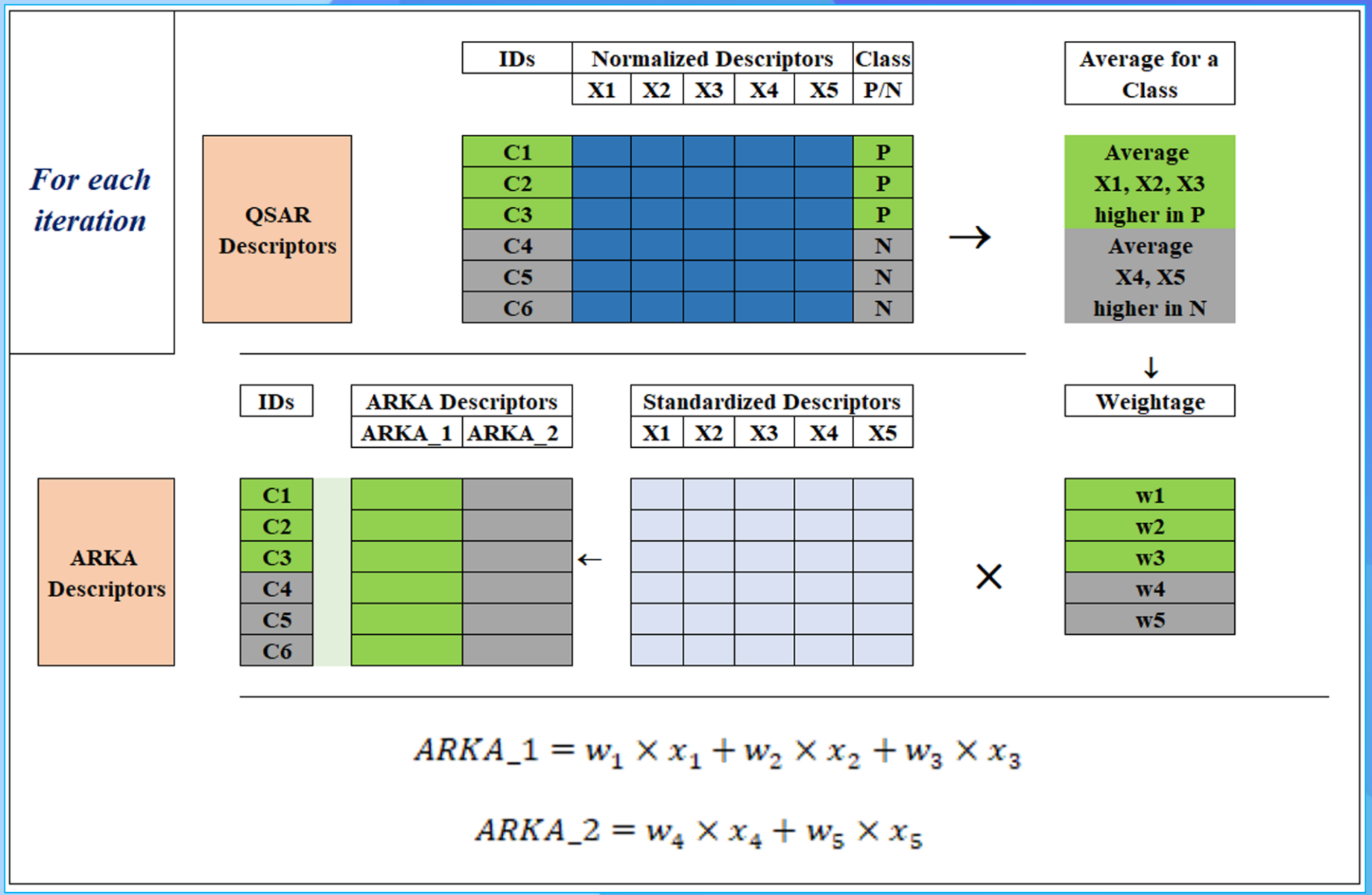
Iteration #4

= Positive class

= Negative class

Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2025

Algorithm

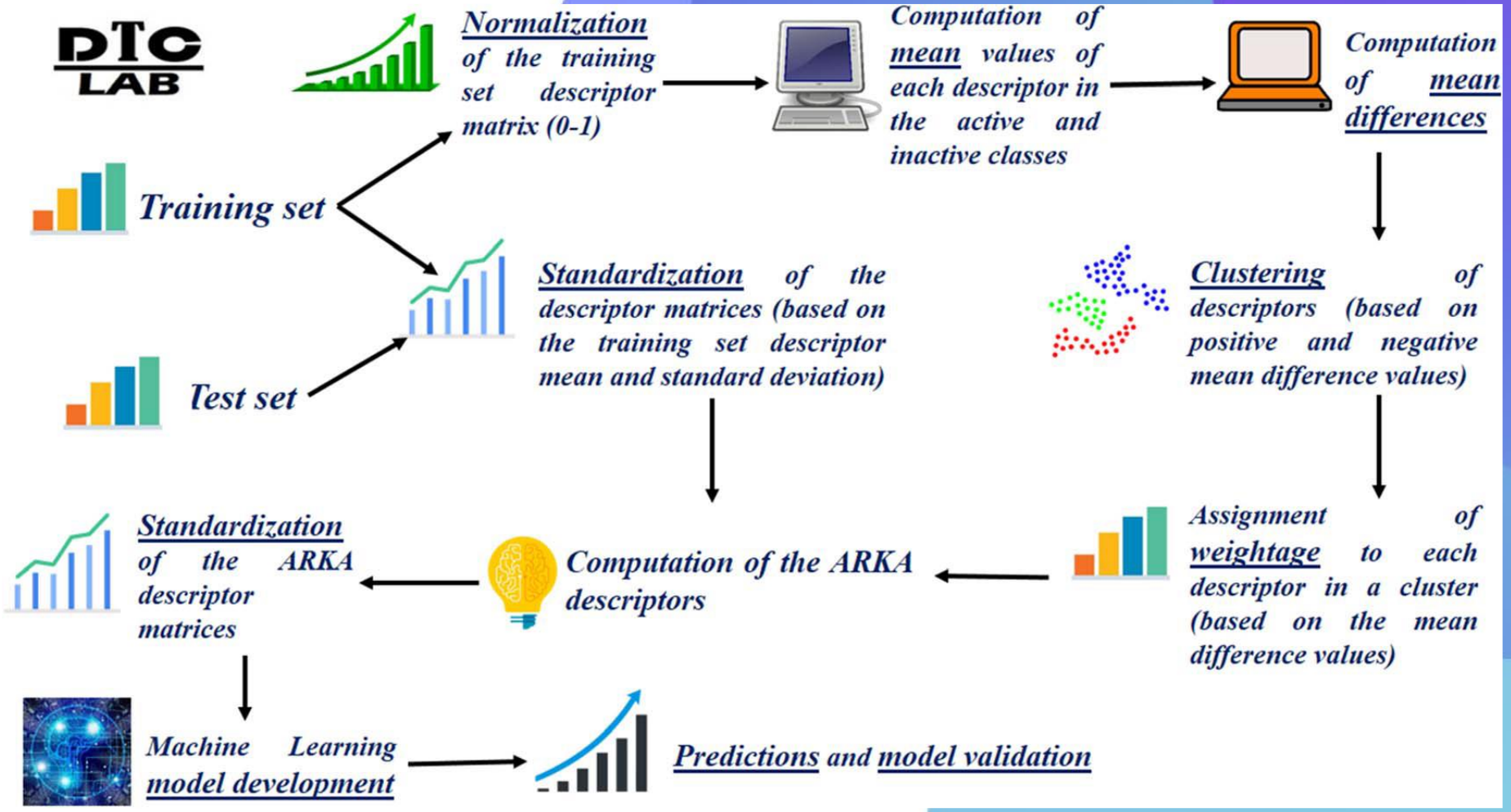


$$ARKA\_1 = w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3$$

$$ARKA\_2 = w_4 \times x_4 + w_5 \times x_5$$

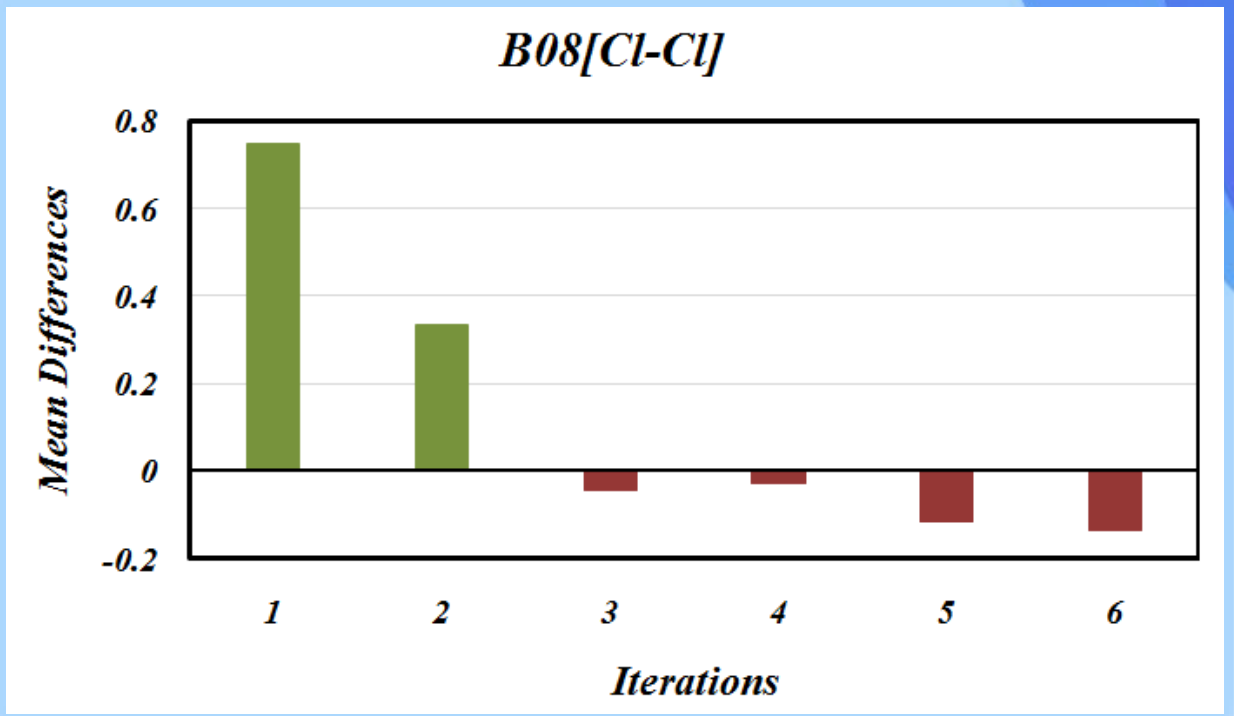
Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2025

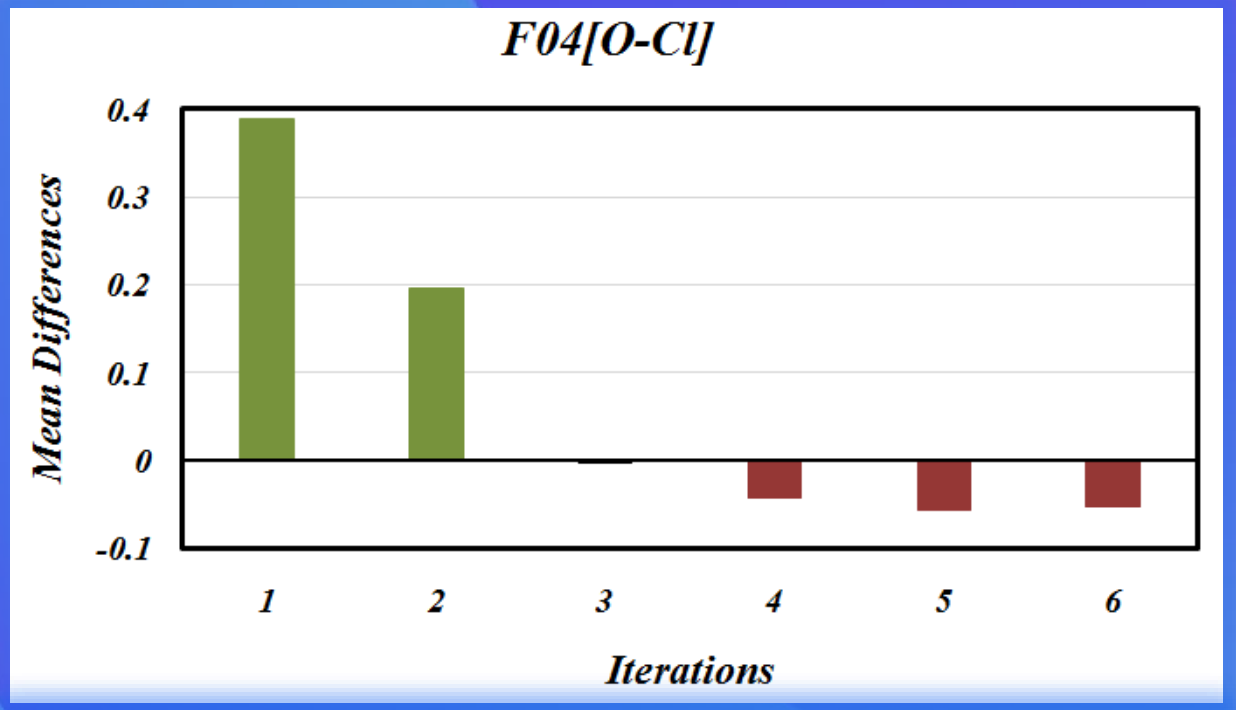
Algorithm

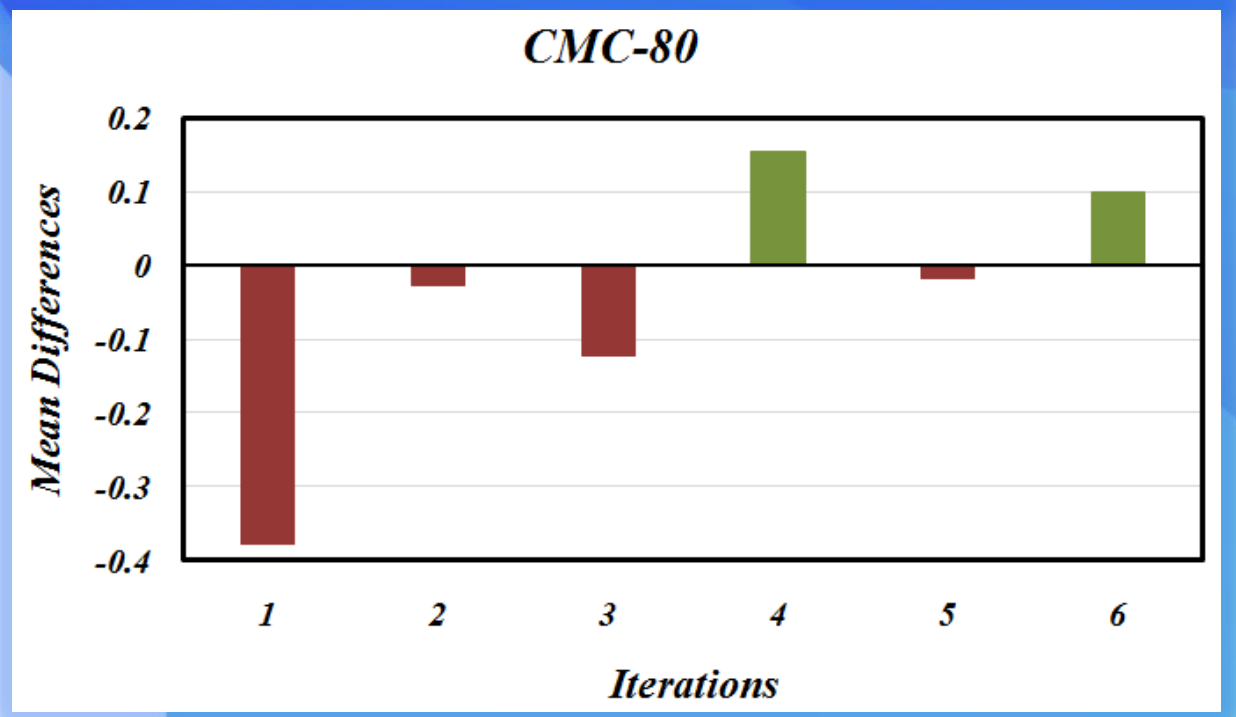


Banerjee, A.; Roy, K. Environ. Sci.: Processes Impacts 2025

Contributions of descriptors across different iterations

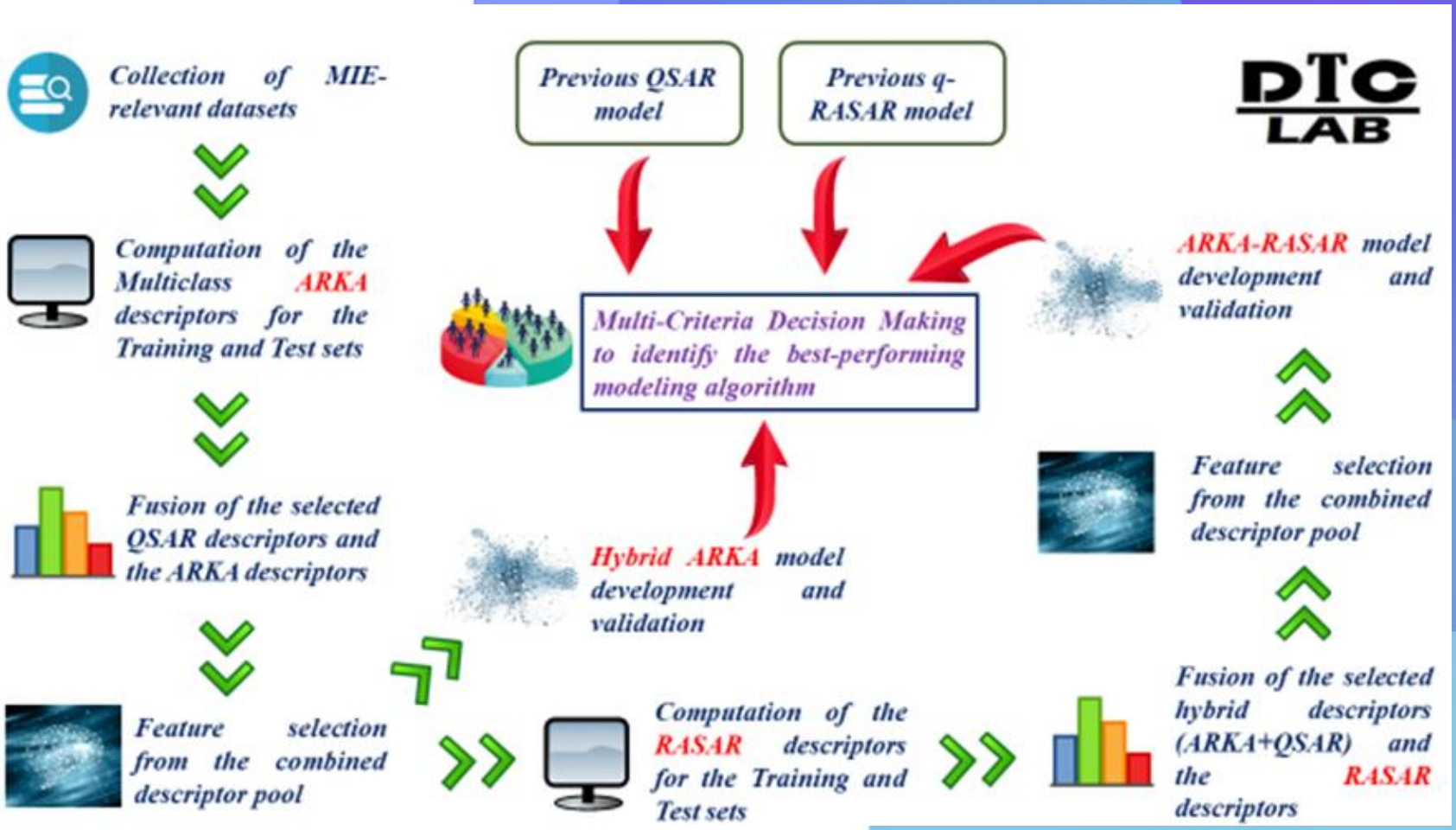






Banerjee A, Roy K. J Hazard Mater 2025

Workflow of improved q-RASAR



Banerjee A, Roy K. J Hazard Mater 2025

Skin sensitization potential of industrial and environmental chemicals

Banerjee and Roy, Environ Sci: Processes Impacts 2023, 25, 1626-1644

hERG K^+ channel inhibition of chemicals

Banerjee and Roy, Chemom Intell Lab Syst 2023, 237, 104829

Aquatic toxicity of pesticides against *Lepomis sp.*

Ghosh et al., Aquat Toxicol 2023, 265, 106776

Androgen receptor binding affinity of endocrine disruptors in rats

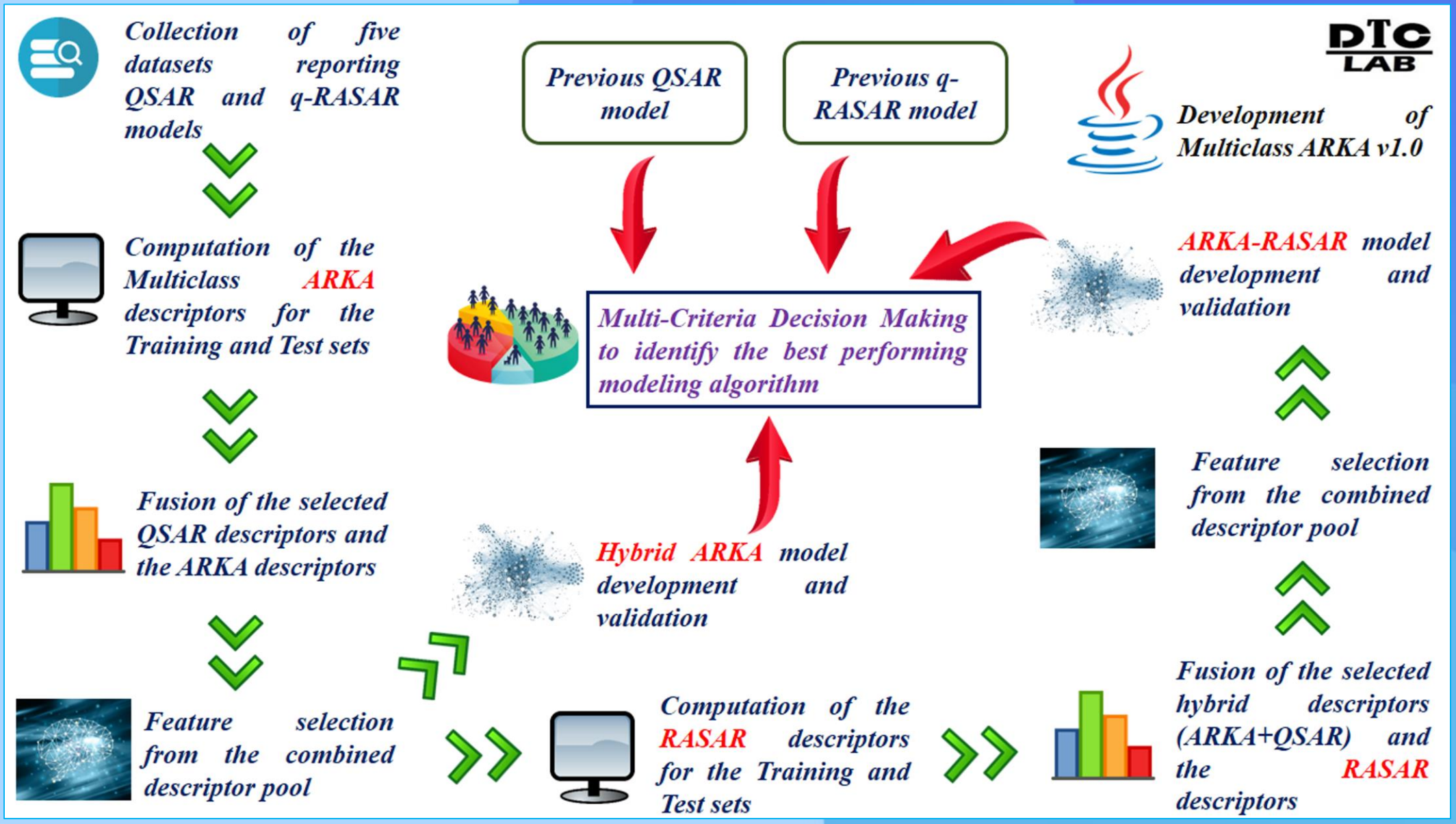
Banerjee et al., Chemosphere 2022, 309, 136579

Aquatic toxicity of pesticides against Rainbow trout

Ghosh et al., Aquat Toxicol 2023, 265, 106776

Datasets used in the current study

Workflow


Collection of five datasets reporting QSAR and q-RASAR models
Computation of the Multiclass ARKA descriptors for the Training and Test sets
Fusion of the selected QSAR descriptors and the ARKA descriptors
Feature selection from the combined descriptor pool
Computation of the RASAR descriptors for the Training and Test sets
Fusion of the selected hybrid descriptors (ARKA+QSAR) and the RASAR descriptors
Feature selection from the combined descriptor pool
ARKA-RASAR model development and validation
Development of Multiclass ARKA v1.0
Previous QSAR model
Previous q-RASAR model
Multi-Criteria Decision Making to identify the best performing modeling algorithm
Hybrid ARKA model development and validation
DTC LAB

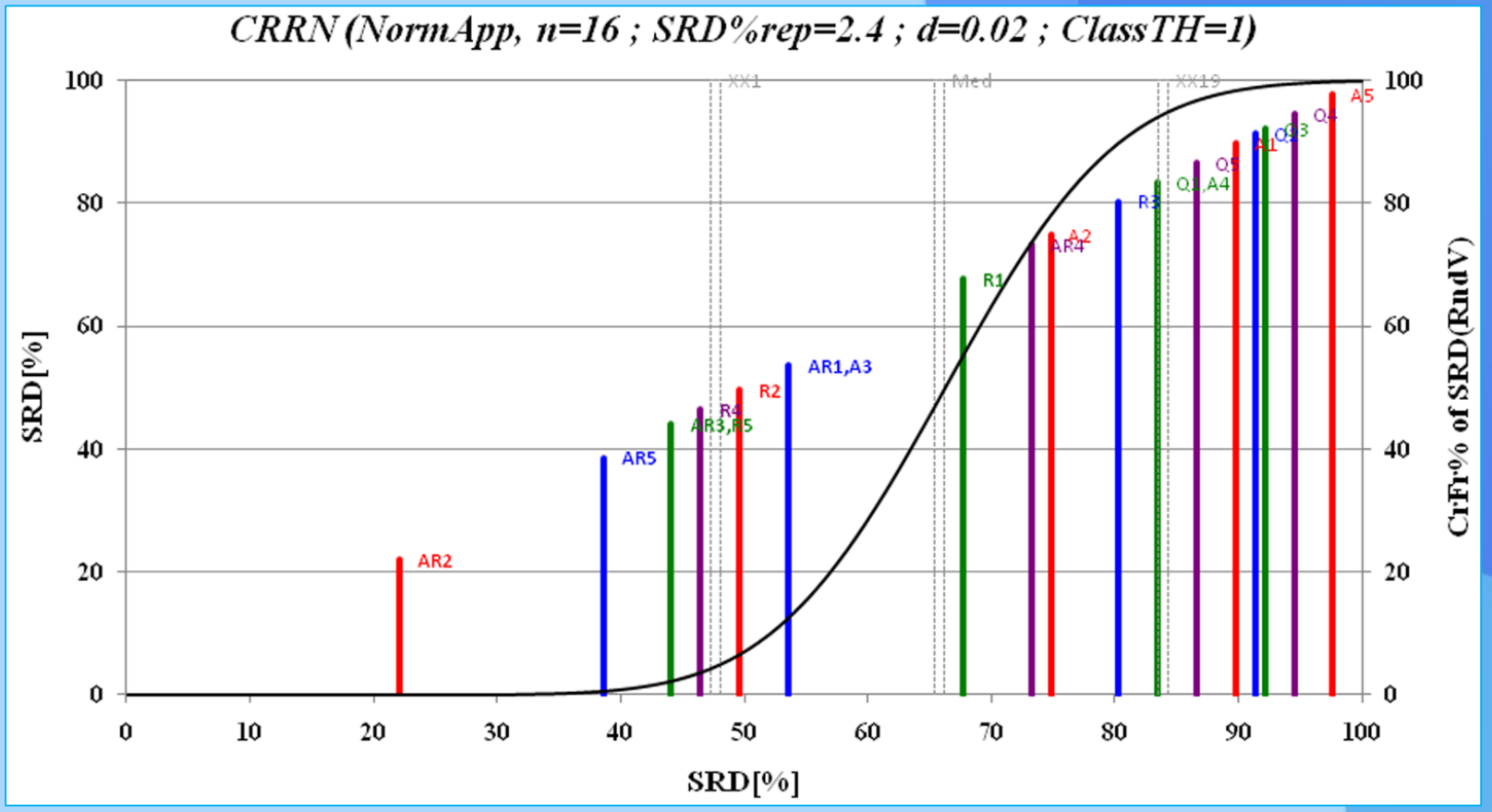
Results

Table 1 Model statistics of the previous QSAR, previous q-RASAR, hybrid ARKA and ARKA-RASAR models[a]

			Training set					Test set				
Dataset	Model	n_{Desc}	n_{Train}	R^2	Q_{LOO}^2	MAE_{train}	MAE_{LOO}	n_{Test}	Q_{F1}^2	Q_{F2}^2	MAE_{test}	Model specifications
1	PLS (LV = 3)	8	103	0.737	0.68	0.456	0.497	44	0.582	0.582	0.539	Previous QSAR
	Univariate	1	103	0.675	0.657	0.434	0.444	44	0.633	0.633	0.483	Previous q-RASAR
	PLS (LV = 4)	5	103	0.728	0.685	0.472	0.507	44	0.609	0.609	0.527	Hybrid ARKA
	MLR	4	103	0.724	0.684	0.461	0.488	44	0.675	0.675	0.472	ARKA-RASAR
2	PLS (LV = 3)	15	196	0.635	0.549	0.555	0.607	65	0.485	0.484	0.695	Previous QSAR (without removal)
	PLS (LV = 3)	15	196	0.635	0.549	0.555	0.607	63	0.575	0.574	0.642	Previous QSAR
	PLS (LV = 4)	12	196	0.608	0.546	0.581	0.623	63	0.66	0.66	0.548	Previous q-RASAR
	MLR	7	196	0.624	0.591	0.559	0.583	65	0.487	0.486	0.687	Hybrid ARKA (without removal)
	MLR	7	196	0.624	0.591	0.559	0.583	63	0.578	0.577	0.634	Hybrid ARKA
	PLS (LV = 10)	12	196	0.663	0.623	0.534	0.568	63	0.67	0.669	0.557	ARKA-RASAR
	PLS (LV = 10)	12	196	0.663	0.623	0.534	0.568	65	0.574	0.574	0.612	ARKA-RASAR (without removal)
3	PLS (LV = 9)	10	133	0.696	0.644	0.41	0.445	47	0.526	0.524	0.562	Previous QSAR (without removal)
	PLS (LV = 9)	10	133	0.696	0.644	0.41	0.445	44	0.586	0.585	0.528	Previous QSAR
	PLS (LV = 8)	9	133	0.695	0.649	0.406	0.436	44	0.607	0.606	0.523	Previous q-RASAR
	MLR	7	133	0.703	0.672	0.399	0.421	47	0.561	0.559	0.55	Hybrid ARKA (without removal)
	MLR	7	133	0.703	0.672	0.399	0.421	44	0.602	0.601	0.526	Hybrid ARKA
	MLR	8	133	0.708	0.673	0.395	0.422	44	0.607	0.606	0.524	ARKA-RASAR
	MLR	8	133	0.708	0.673	0.395	0.422	47	0.563	0.561	0.55	ARKA-RASAR (without removal)
4	PLS (LV = 3)	8	102	0.679	0.62	0.763	0.839	34	0.703	0.658	0.678	Previous QSAR
	PLS (LV = 3)	8	102	0.67	0.598	0.767	0.845	34	0.74	0.701	0.644	Previous q-RASAR
	PLS (LV = 5)	6	102	0.684	0.647	0.756	0.804	34	0.7	0.655	0.697	Hybrid ARKA
	PLS (LV = 5)	8	102	0.706	0.649	0.742	0.805	34	0.708	0.664	0.682	ARKA-RASAR
5	PLS (LV = 5)	8	537	0.534	0.515	0.783	0.797	178	0.551	0.541	0.752	Previous QSAR
	PLS (LV = 4)	8	537	0.52	0.504	0.79	0.804	178	0.588	0.579	0.715	Previous q-RASAR
	MLR	4	537	0.52	0.51	0.795	0.802	178	0.561	0.552	0.749	Hybrid ARKA
	PLS (LV = 4)	7	537	0.527	0.513	0.791	0.801	178	0.58	0.571	0.729	ARKA-RASAR

[a] **BOLD TEXT** indicates the overall best-performing model in a dataset considering internal and external validation statistics.

Multi-Criteria Decision Making



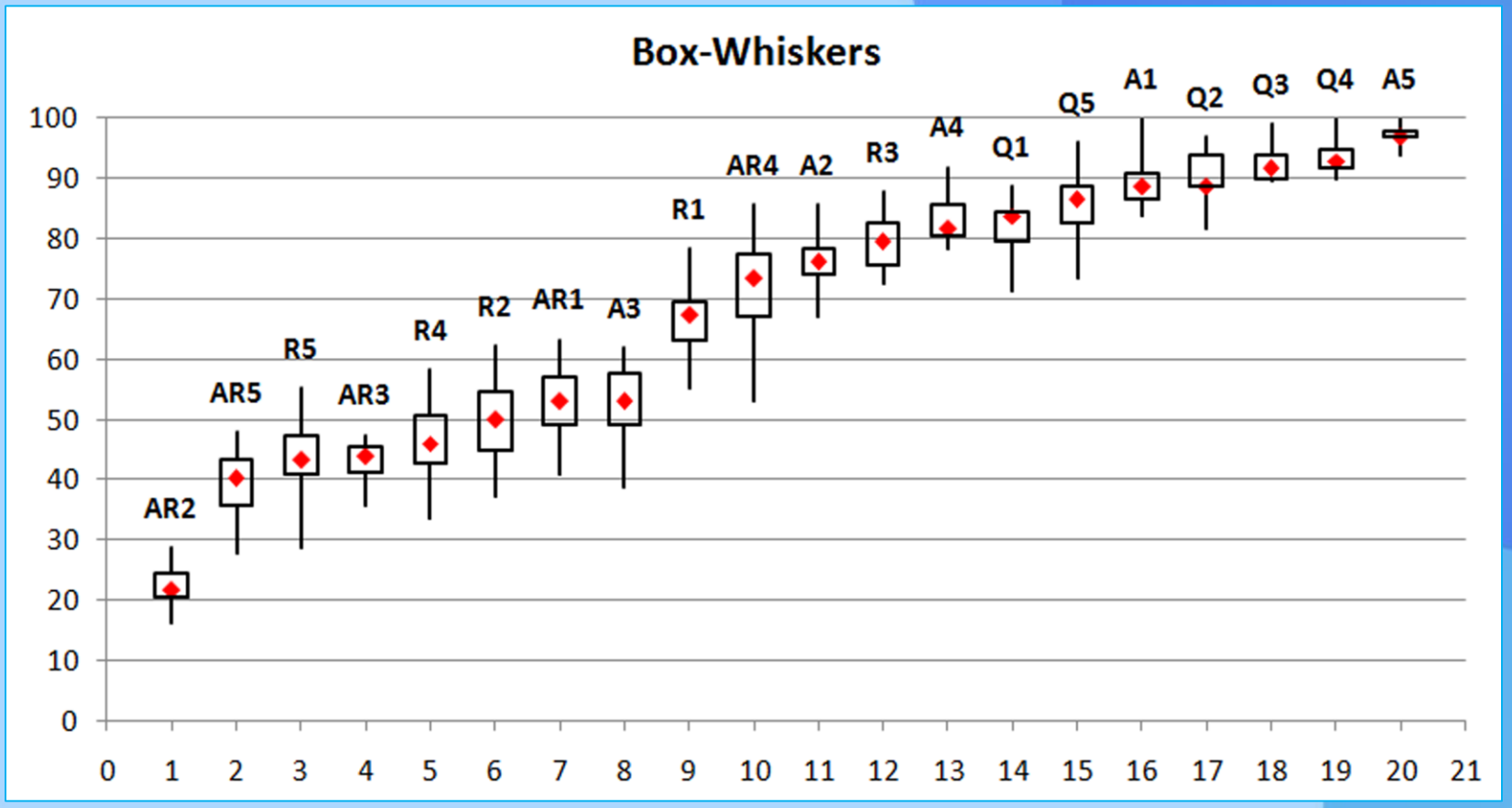
SRD Plot of all the developed models

The Sum of Ranking Differences (SRD)

- *Lower the SRD, better is the model.*
- *AR = ARKA-RASAR*
  A = Hybrid ARKA
  R = q-RASAR
  Q = QSAR
- *ARKA-RASAR models are the best*

Heberger K. TrAC Trends Anal Chem 2010, 29, 101-109

Multi-Criteria Decision Making



Leave-1/7th-Out Cross-Validated SRD results, showing the ARKA-RASAR models are the best (X-axis represents the models, Y-axis represents the SRD-CV)

The Sum of Ranking Differences (SRD)

Lower the SRD-CV, better is the model.

AR = ARKA-RASAR
A = Hybrid ARKA
R = q-RASAR
Q = QSAR

ARKA-RASAR models are the best

Sziklai et al., Cent. Eur. J. Oper. Res. 2024

ANOVA – To ensure unbiased observations

Source	DF	SS	MS	*F*	*p*	Inference
Factor 2	4	28 177	7044	351.52	0.000	The datasets are significantly different from each other
Factor 1	3	321 411	107 137	5346.33	0.000	The results from the modeling algorithms are significantly different from each other
Interaction	12	163 826	13 652	681.27	0.000	Both the factors are inter-dependent

[a] DF = degree of freedom, SS = sum of squares, MS = mean squares.

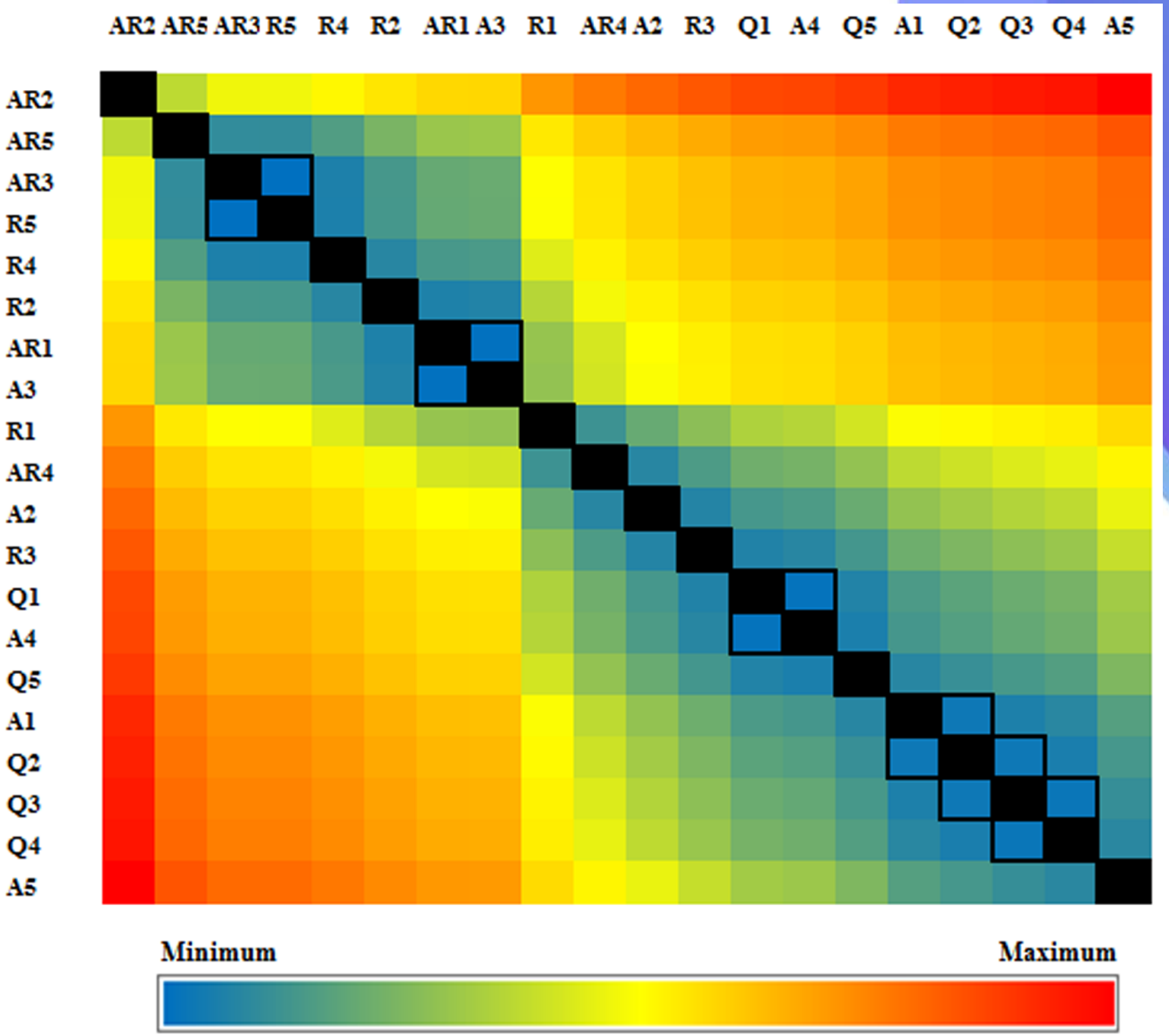
- ✓ ***Factor 1***: ***Between the four different modeling approaches (QSAR, q-RASAR, Hybrid ARKA, ARKA-RASAR)***
- ✓ ***Factor 2***: ***Between the five different datasets***

Snedecor and Cochran, Statistical Methods, 8th Edition, Wiley-Blackwell

Inferences

- ❑ ***Bias due to the datasets is absent***
- ❑ ***Modeling algorithms are not similar***

One-way ANOVA analysis



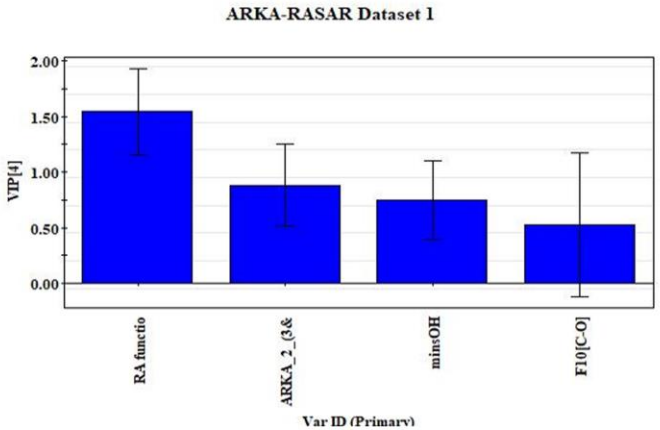
Aim

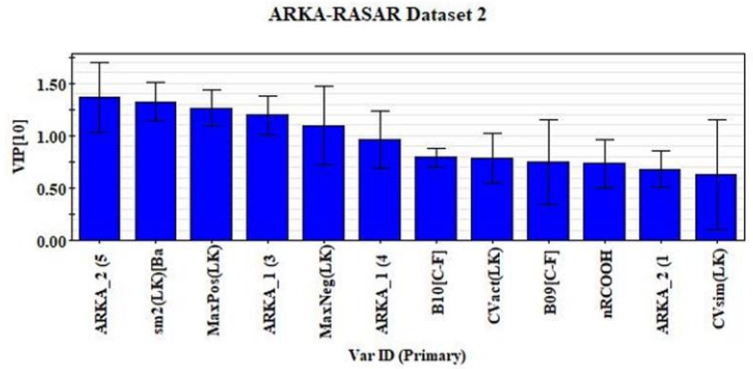
- ❑ *To show that the results obtained from most of the models in the five datasets were significantly different from each other*

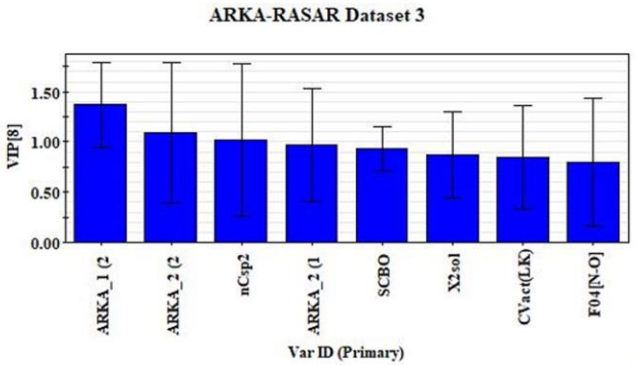
- ✓ *Least Significant Difference (LSD) procedure coupled with One-way ANOVA, using Fisher's test (95% CI) of the SRD values*

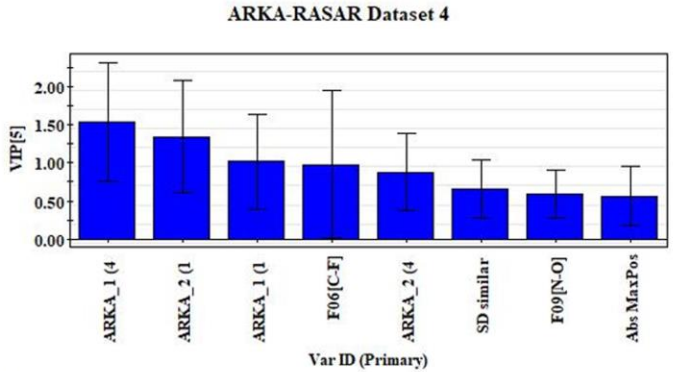
Bolton S, Statistics, in: Remington JP. Remington: the Science and Practice of Pharmacy (ed. D. Troy), Lippincott Williams & Wilkins, Baltimore, 2006.

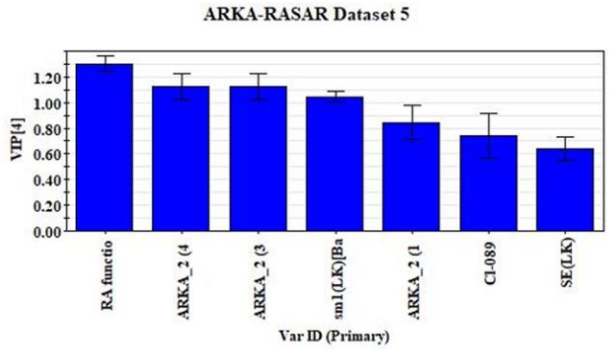
Variable Importance Plots











Inferences

- ***ARKA descriptors appear to be the most important.***
- ***This is followed by the RASAR descriptors computed on the Hybrid ARKA feature matrix***
- ***The conventional QSAR descriptors appear to have low importance***

Wu et al., Introduction to SIMCA-P and its application. In: Handbook of partial least squares: concepts, methods and applications 2009, 757-774, Springer, Heidelberg.

Real world effectiveness in data gap filling

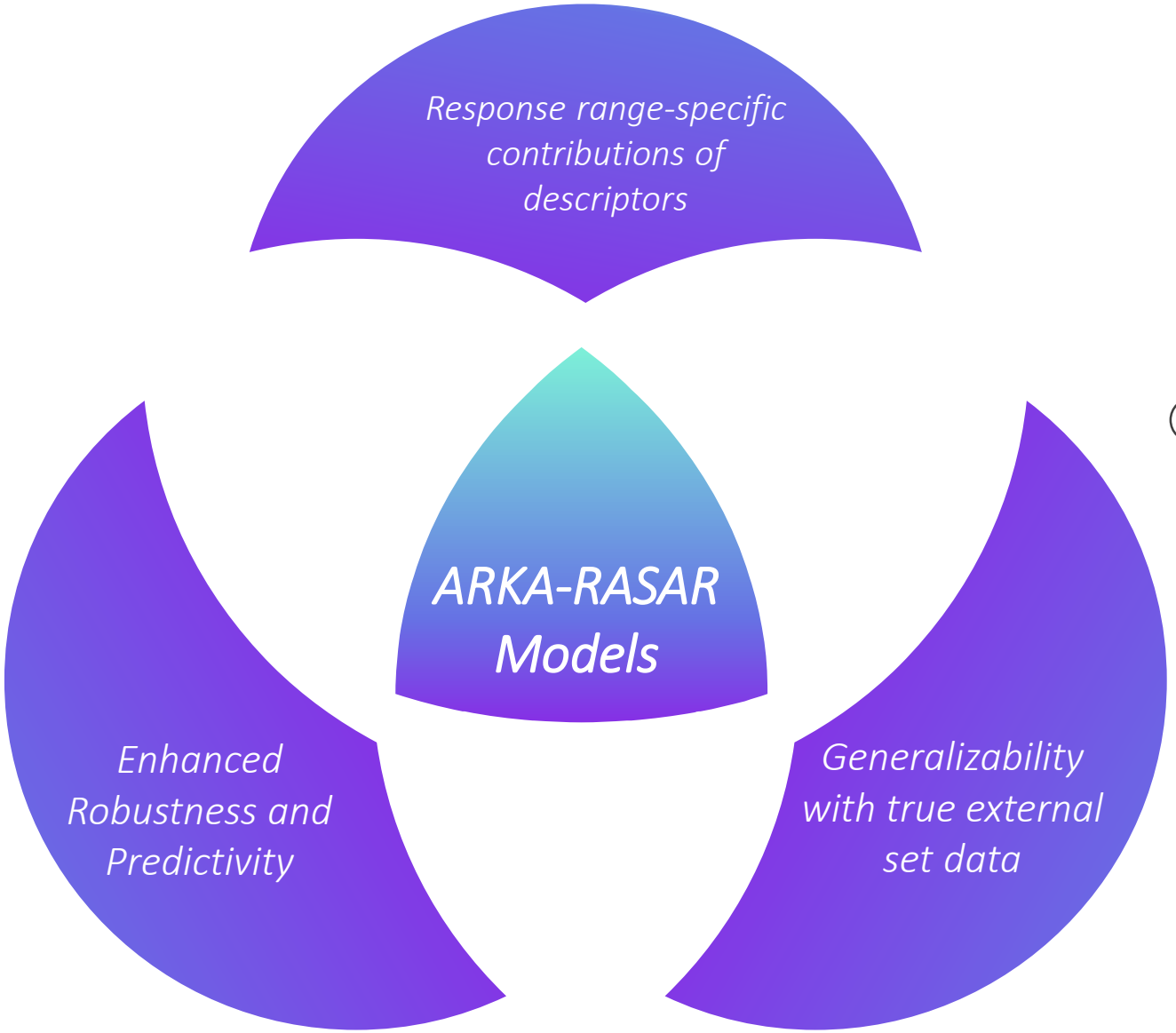
Sl	Metabolite	SMILES code	Expt. Fish LC_{50} (mg L^{-1})	Expt. GHS class	ECOSAR LC_{50} (mg L^{-1})	ECOSAR GHS class
1	2-Ethyl-4,5,6,7-tetrahydro-4-oxo-6-(2,4,6-trimethylphenyl)benzoxazole	CCc1nc2c(o1)CC(CC2=O)c3c(cc(cc3C)C)C	>0.61	1	0.36	1
2	Acifluorfen	Clc2cc(ccc2Oc1cc(C(=O)O)c([N+]([O-])=O)cc1)C(F)(F)F	54	3	33.16	3
3	Chlordecone	ClC54C(=O)C1(Cl)C2(Cl)C5(Cl)C3(Cl)C4(Cl)C1(Cl)C2(Cl)C3(Cl)Cl	0.02	1	0.99	1
4	Ethion	S=P(SCSP(=S)(OCC)OCC)(OCC)OCC	0.5	1	0.07	1
5	Fipronil sulfide	c1c(cc(c(c1Cl)n2c(c(c(n2)C#N)SC(F)(F)F)N)Cl)C(F)(F)F	0.03	1	0.031	1
6	Ioxynil	Ic1cc(C#N)cc(I)c1O	8.5	2	2.14	2
7	Triadimefon	CC(C)(C)C(=O)C(N1C=NC=N1)OC2=CC=C(C=C2)Cl	4.08	2	13.76	3

Burden et al., Regulat Toxicol Pharmacol 2016, 80, 241-246

(A curated set of 150 pesticide metabolites)

ARKA-RASAR models of Datasets 4 and 5

(Datasets for fish toxicity)

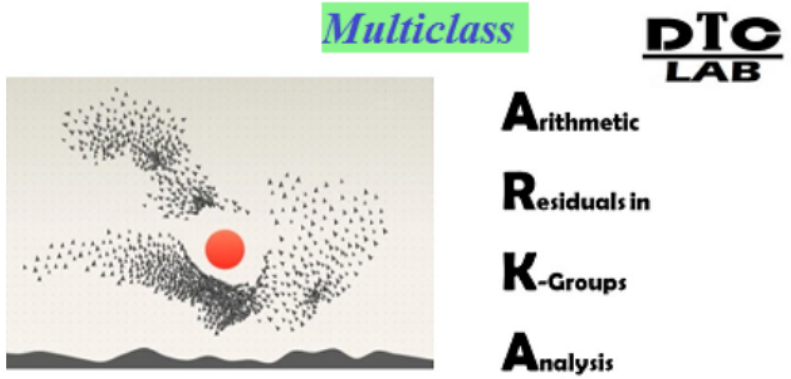


Conclusions

- *ARKA-RASAR models are robust and highly predictive*
- *Capture response range-specific contributions of descriptors*
- *Show effective generalizability with true external set data*

A Java-based tool to compute multiclass ARKA descriptors

MultiClass ARKA



MultiClass ARKA

This tool calculates multiple ARKA descriptors based on the user's requirements to develop regression-based QSAR models. This considers different contributions of the relevant features to different response ranges of the training set within a particular regression model.

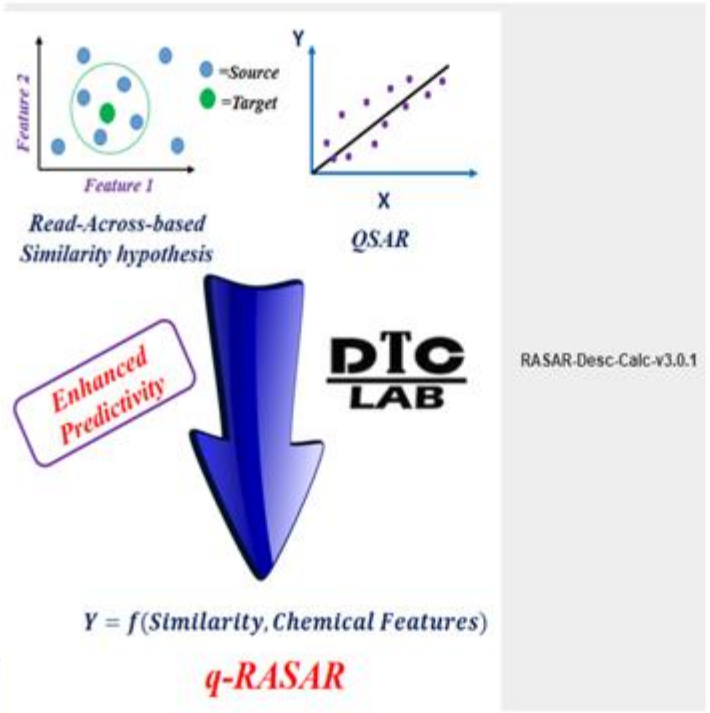
Download link (Uploaded on 19.12.2024; Unrestricted from April 03, 2025)

Reference: Banerjee A, Roy K, The multiclass ARKA framework for developing improved q-RASAR models for environmental toxicity endpoints. *Environ Sci Process Impacts*, 2025, https://doi.org/10.1039/D5EM00068H

To use this tool, please fill in https://forms.gle/1r3TTy7RmZCQvqBt5 and sign the License agreement form

https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/arithmetic-residuals-in-k-groups-analysis-arka

A Java-based tool to compute RASAR descriptors

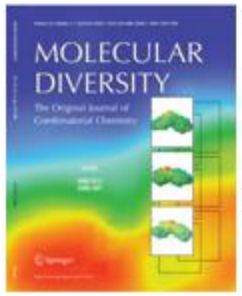


RASAR-Desc-Calc-v3.0.1

DTC LAB

RASAR Descriptor Calculator

v3.0.3



Banerjee A, Roy K, *Mol Divers*, 26, 2022, 2847-2862, DOI: 10.1007/s11030-022-10478-6
Banerjee A, Roy K, Chem Res Toxicol, 36, 2023, 446-464, DOI: 10.1021/acs.chemrestox.2c00374

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

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

Environmental Science
Processes & Impacts

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5em00068h

The multiclass ARKA framework for developing improved q-RASAR models for environmental toxicity endpoints†

Arkaprava Banerjee * and Kunal Roy *

The continuous quest for the quick, accurate, and efficient methods for filling the gaps in the toxicity data of commercial chemicals is the need of the hour. Thus, it has become essential to develop simple and improved modeling strategies that aim to generate more accurate predictions. Recently, quantitative Read-Across Structure–Activity Relationship (q-RASAR) modeling has been reported to enhance the external predictivity of QSAR models. However, the cross-validation metrics of some q-RASAR models show compromised values compared to those of the corresponding QSAR models. We report here an improved q-RASAR workflow coupled with the Arithmetic Residuals in *K*-groups Analysis (ARKA) framework. This improved workflow (ARKA-RASAR) considers two important aspects: the contribution of different QSAR descriptors to different experimental response ranges, and the identification of similarity among close congeners based on both the selected QSAR descriptors and the contribution of different QSAR descriptors to different experimental response ranges. A simple, free, and user-friendly Java-based tool, Multiclass ARKA-v1.0, has been developed to compute the multiclass ARKA descriptors. In this study, five different toxicity datasets previously used for the development of QSAR and q-RASAR models were considered. We developed hybrid ARKA models that consist of a combination of QSAR descriptors and ARKA descriptors. These hybrid feature spaces were used to compute RASAR descriptors and develop ARKA-RASAR models. We used the same modeling strategies used to develop the previously reported QSAR and q-RASAR models for a fair comparison. Additionally, these modeling algorithms are straightforward, reproducible, and transferable. A multi-criteria decision-making statistical approach, the Sum of Ranking Differences (SRD), indicated that the ARKA-RASAR models are the best-performing models, considering training, test, and cross-validation statistics. The least significant difference procedure ensured that the SRD values were significantly different for most models, presenting an unbiased workflow. True external validation using a set of pesticide metabolites and predicting their early-stage acute fish toxicity using relevant ARKA-RASAR models was also carried out and yielded encouraging results. The promising results and the ease of computation of ARKA and RASAR descriptors using our tools suggest that the ARKA-RASAR modeling framework may be a potential choice for developing highly robust and predictive models for filling the gaps in environmental toxicity data.

Received 27th January 2025
Accepted 2nd April 2025

DOI: 10.1039/d5em00068h

rsc.li/espi

Environmental significance

Due to limited availability of quantitative environmental toxicity data for existing and newer chemicals, computational-model-derived data provides an alternative approach for filling gaps in the data. However, developing meaningful statistical models using limited quantitative environmental toxicity data is quite challenging. The problem of small data set classification modeling of ecotoxicity endpoints was previously addressed by introducing the concept of Arithmetic Residuals in *K*-groups Analysis (ARKA) as a novel method of supervised dimensionality reduction. Here, a multiclass-ARKA framework is introduced for developing robust and predictive regression-based quantitative read-across-structure–activity relationship (q-RASAR) models to deal with limited quantitative environmental toxicity data.

Volume 27
Number 5
May 2025
Pages 1187–1484

Environmental Science
Processes & Impacts

rsc.li/espi

ISSN 2050-7887

ROYAL SOCIETY OF CHEMISTRY

PAPER
Arkaprava Banerjee and Kunal Roy
The multiclass ARKA framework for developing improved q-RASAR models for environmental toxicity endpoints

The first article on ARKA

Environmental Science Processes & Impacts

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Environ. Sci.: Processes Impacts*, 2024, **26**, 991

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

Arkaprava Banerjee and Kunal Roy *

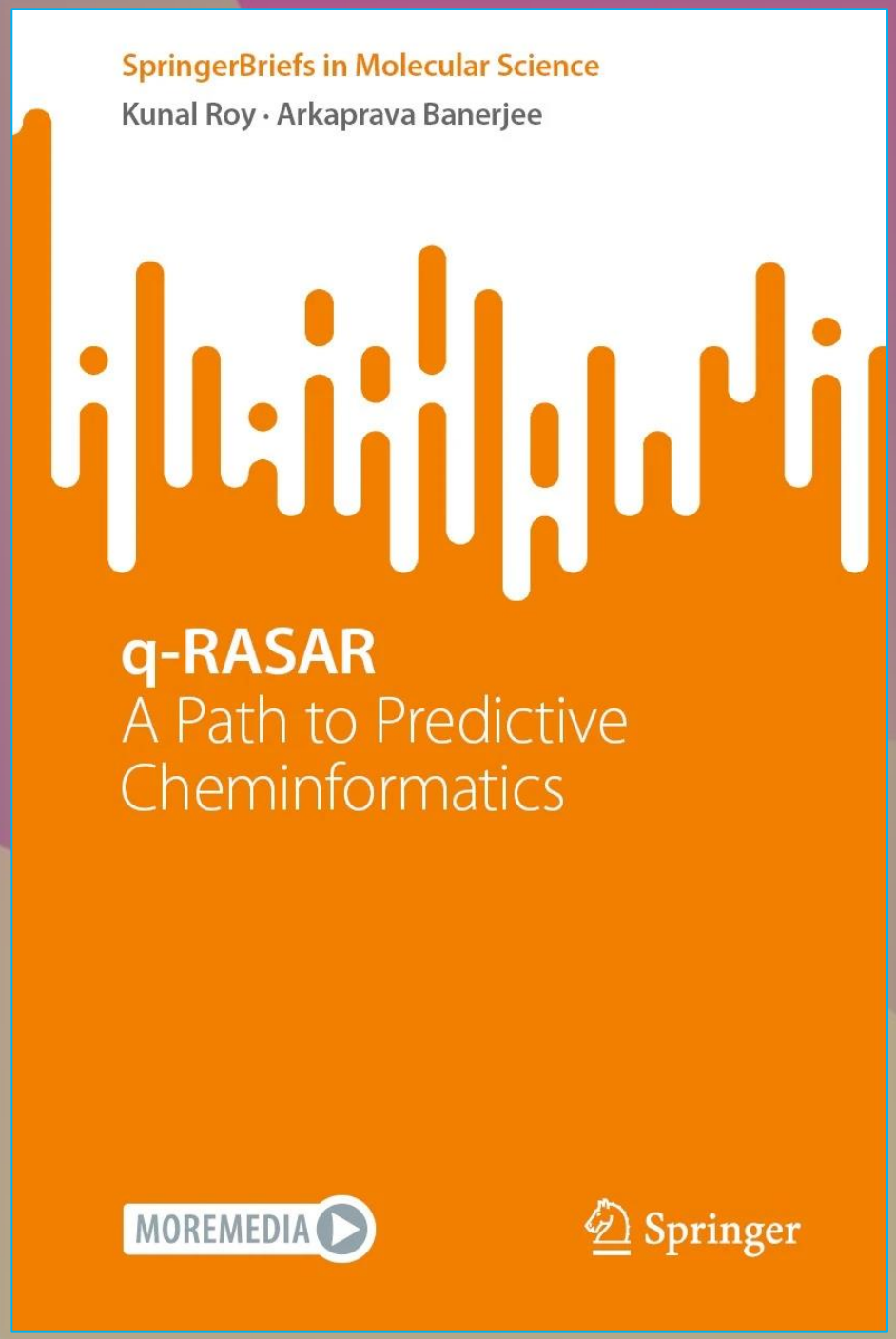
https://doi.org/10.1039/D4EM00173G

- *Identification of Activity Cliffs*
- *Machine Learning-based Classification modeling*

The book on q-RASAR

- *Introduces the reader to a novel cheminformatic workflow*
- *Presents the genesis and model development*
- *Includes practical examples and software tools*

https://link.springer.com/book/10.1007/978-3-031-52057-0



The book on activity cliffs

- *Provides an introduction to the concepts of activity cliffs*
- *Details the impact of activity cliffs on the modelability of data sets and the prediction quality of QSAR models*
- *Analyzes dataset modelability, defining the applicability domain of QSAR models and identifying activity cliffs*

https://link.springer.com/book/9783032100801

Users of q-RASAR across the globe

9 countries, 26 institutes (SCOPUS)

Meet The Team

DTC LAB

যাদবপুর বিশ্ববিদ্যালয় Jadavpur University

Acknowledgement

DTC LAB

Science and Engineering Research Board

Statutory Body Established through an Act of Parliament: SERB Act 2008

Government of India

DTC LAB

Thank You !