ON A SIMPLE FRAMEWORK OF DIMENSIONALITY REDUCTION FOR CLASSIFICATION MODELING OF SPARSE ENVIRONMENTAL TOXICITY DATA

Kunal Roy
Arkaprava Banerjee
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/

ON A SIMPLE FRAMEWORK OF DIMENSIONALITY REDUCTION FOR CLASSIFICATION MODELING OF SPARSE ENVIRONMENTAL TOXICITY DATA

This talk is a part of the Outreach Activity under the Scientific Social Responsibility (SSR) program of the ANRF Project CRG/2023/000202 on Food Informatics (PI: Prof. Kunal Roy)

विज्ञान एवं प्रौद्योगिकी विभाग
DEPARTMENT OF
SCIENCE & TECHNOLOGY
सत्यमेव जयते

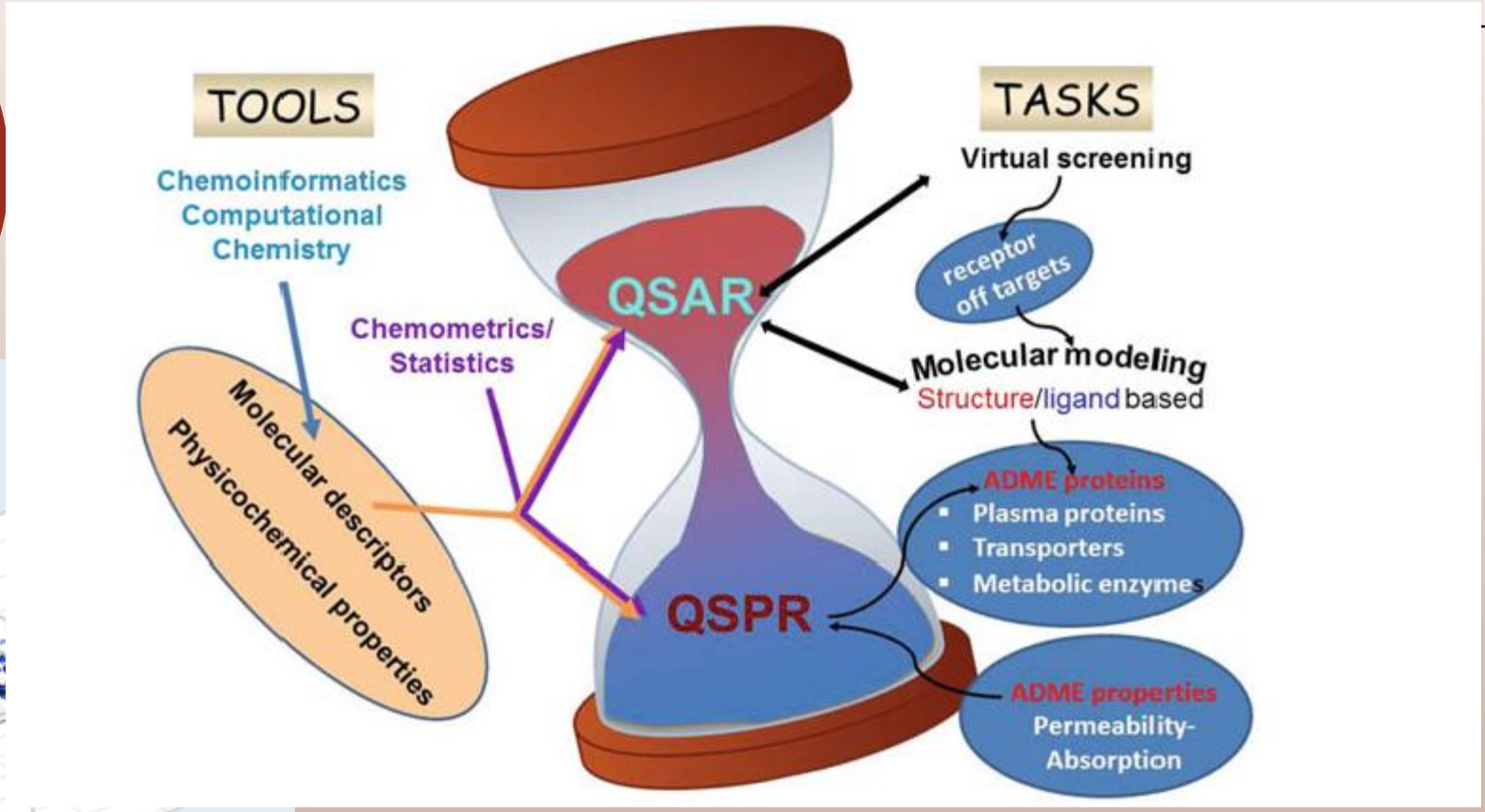
Anusandhan National Research Foundation (ANRF), DST, New Delhi

QSAR (Quantitative Structure-Activity Relationship)

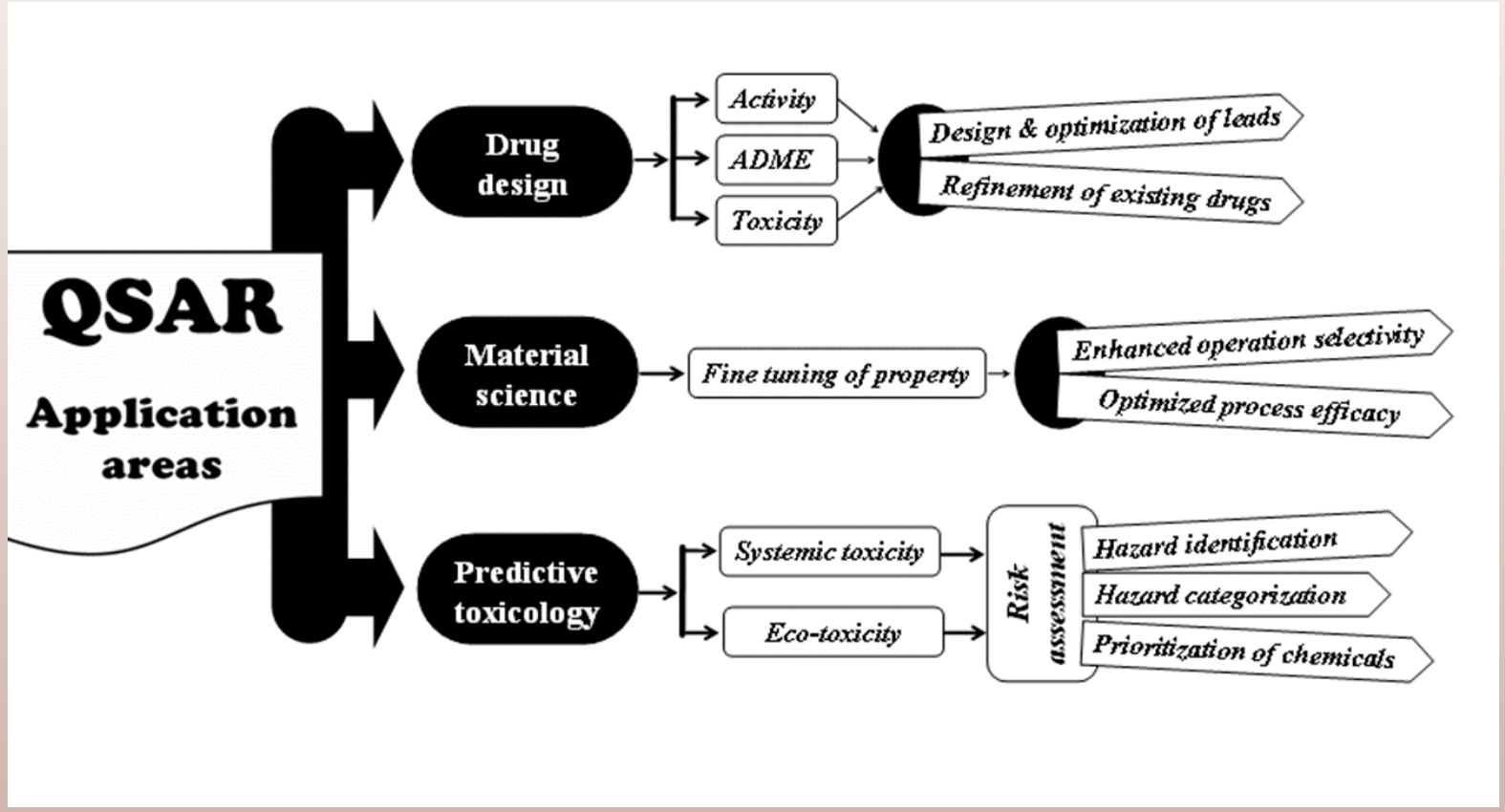
- **QSAR deals with development of predictive models correlating <u>biological activity</u> (including therapeutic and toxic) of chemicals (drugs/toxicants/environmental pollutants) with <u>descriptors</u> representative of molecular structure and/or property by application of <u>statistical tools</u>.**

- ***BA = f (chemical structure or property)***
= f (descriptors)

Yang G F, Huang X, *Curr Pharm Des*, 2006, **12**, 4601-4612

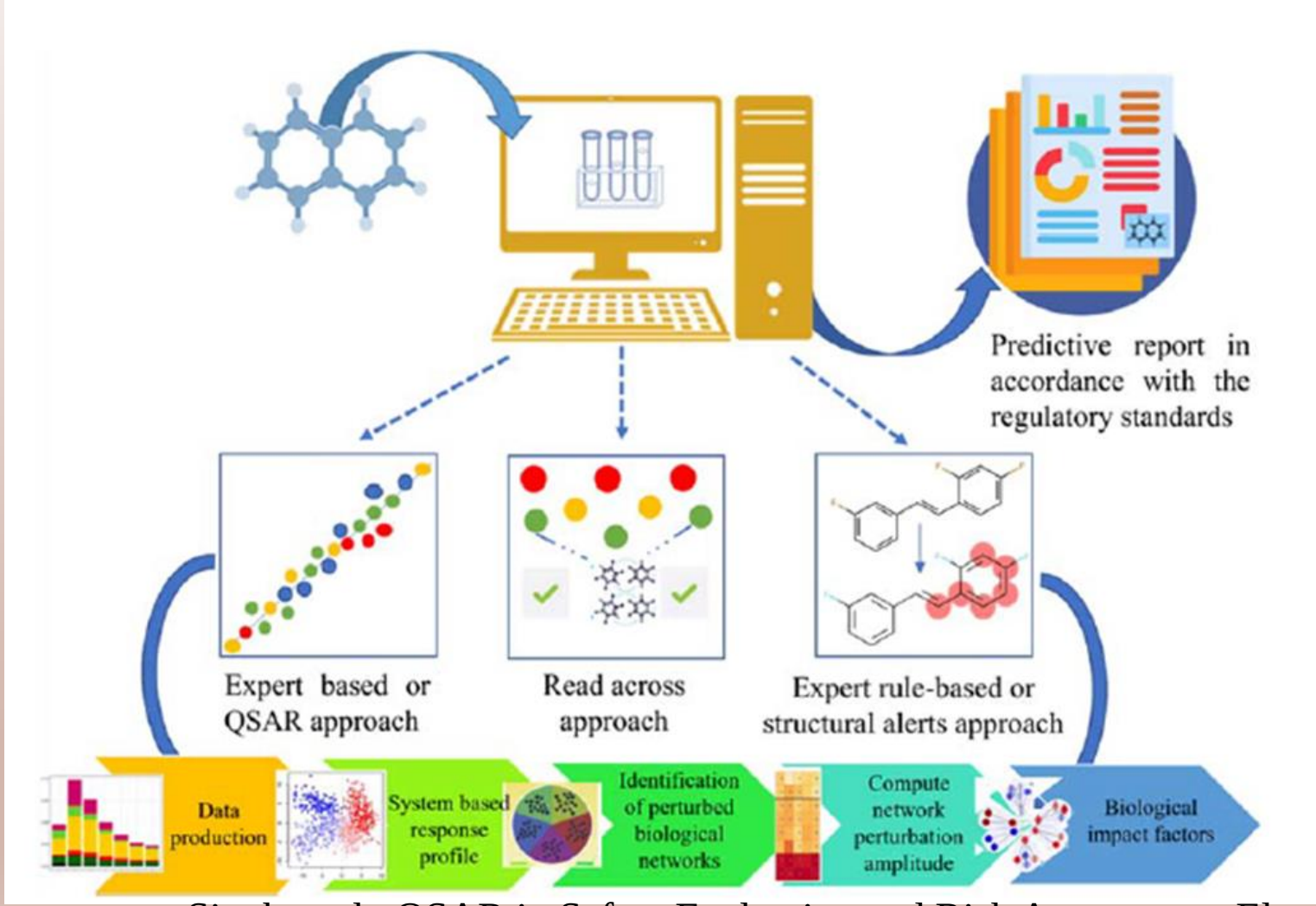


Lambrinidis et al., 2017, DOI 10.1007/978-3-319-56850-8_9



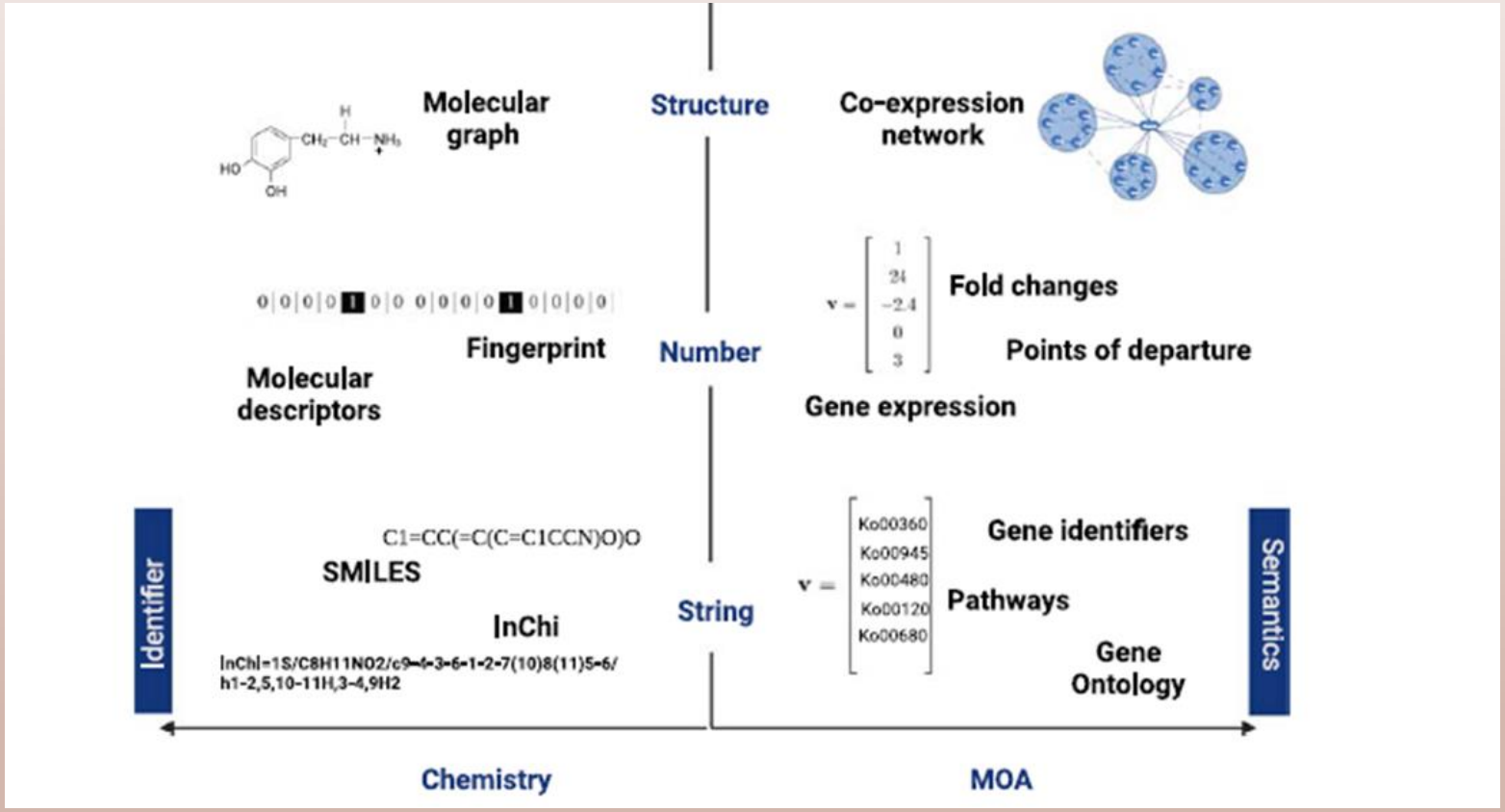
Yang G F, Huang X, *Curr Pharm Des*, 2006, **12**, 4601-4612
Mazzatorta P, Benfenati E, Lorenzini P, Vighi M, *J Chem Inf Comput Sci*, 2004, **44**, 105-112.

Data gap filling approaches



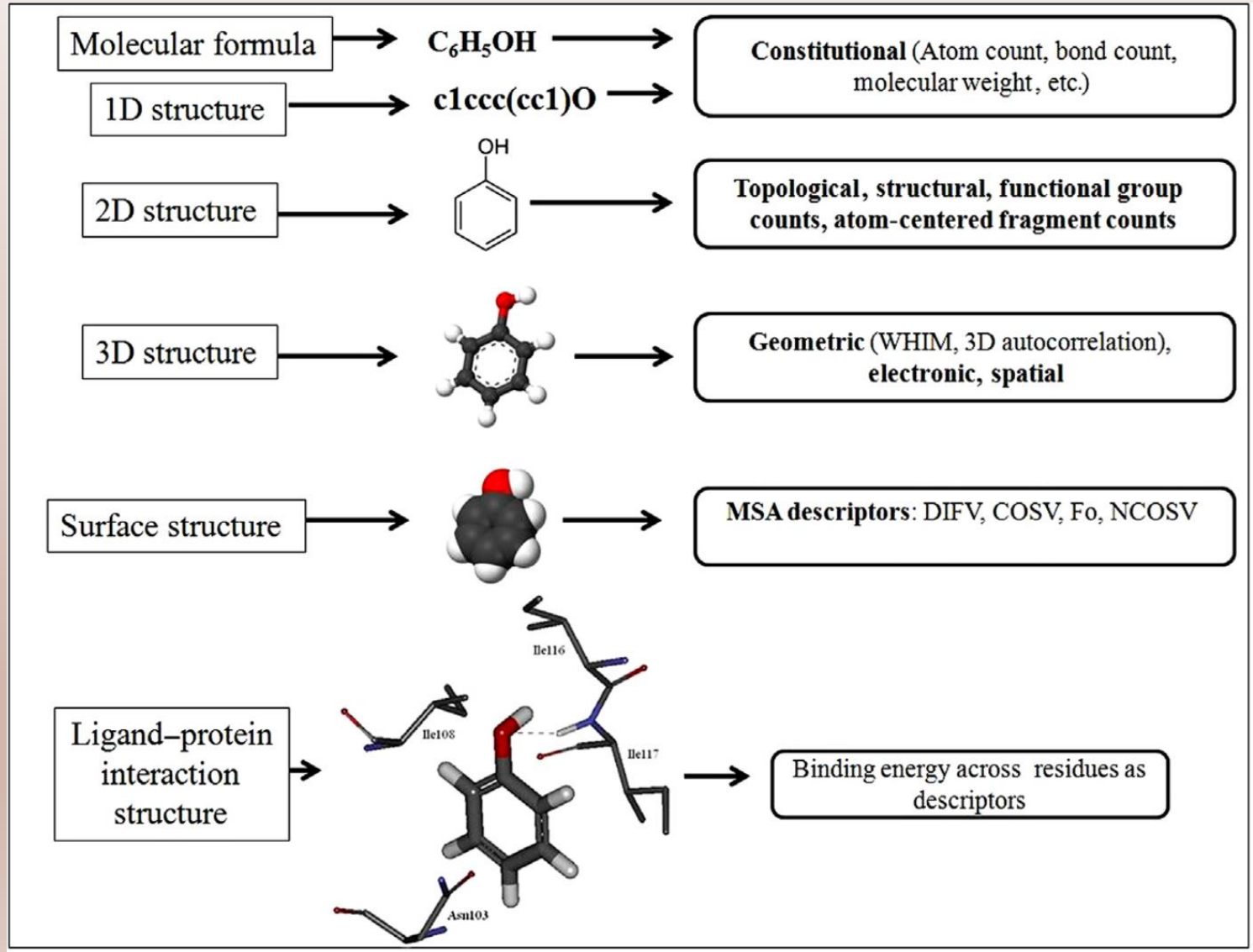
Singh et al., QSAR in Safety Evaluation and Risk Assessment, Elsevier, https://doi.org/10.1016/B978-0-443-15339-6.00026-6

Molecular Structure Representation



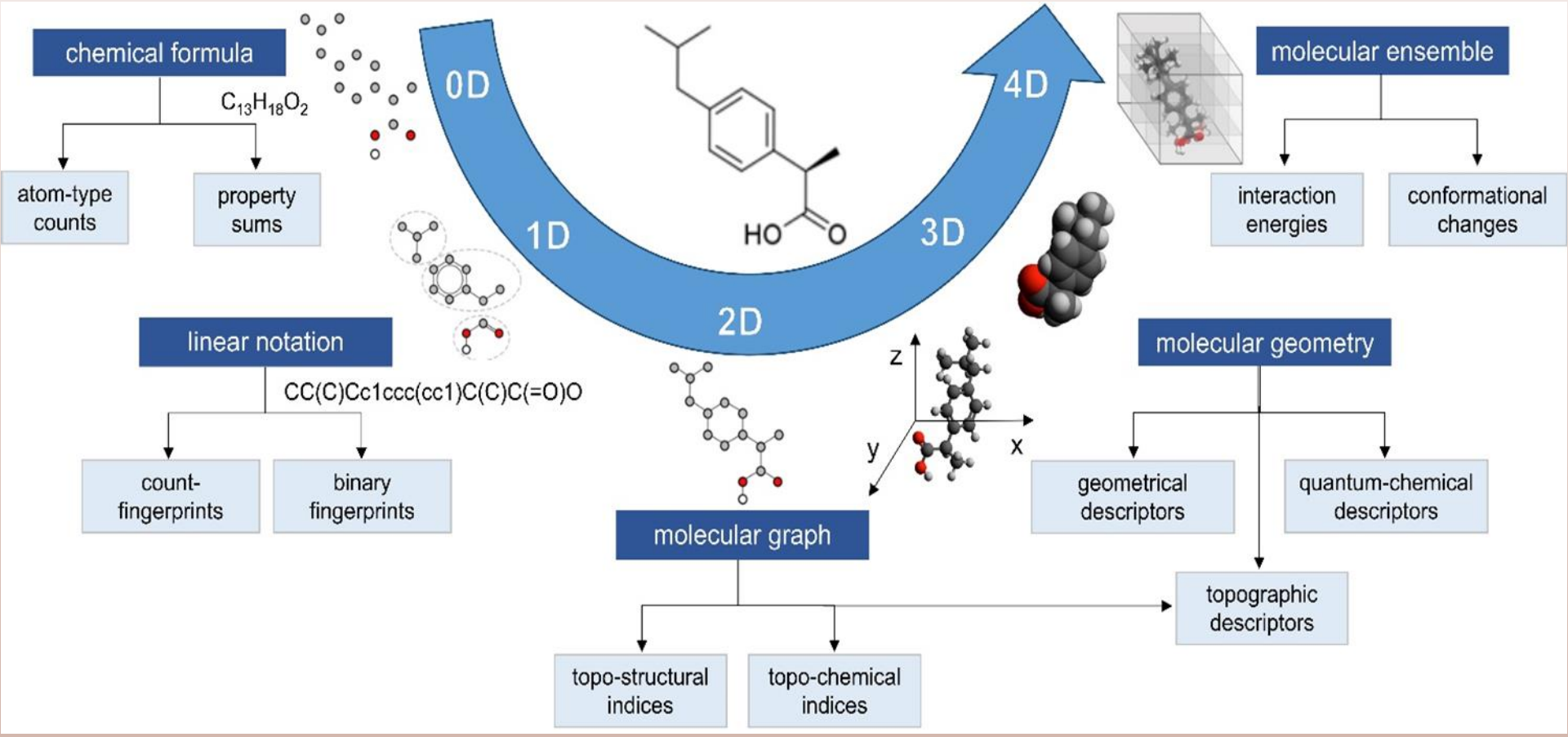
Serra et al., QSAR in Safety Evaluation and Risk Assessment, Elsevier, https://doi.org/10.1016/B978-0-443-15339-6.00011-4

Molecular Descriptors

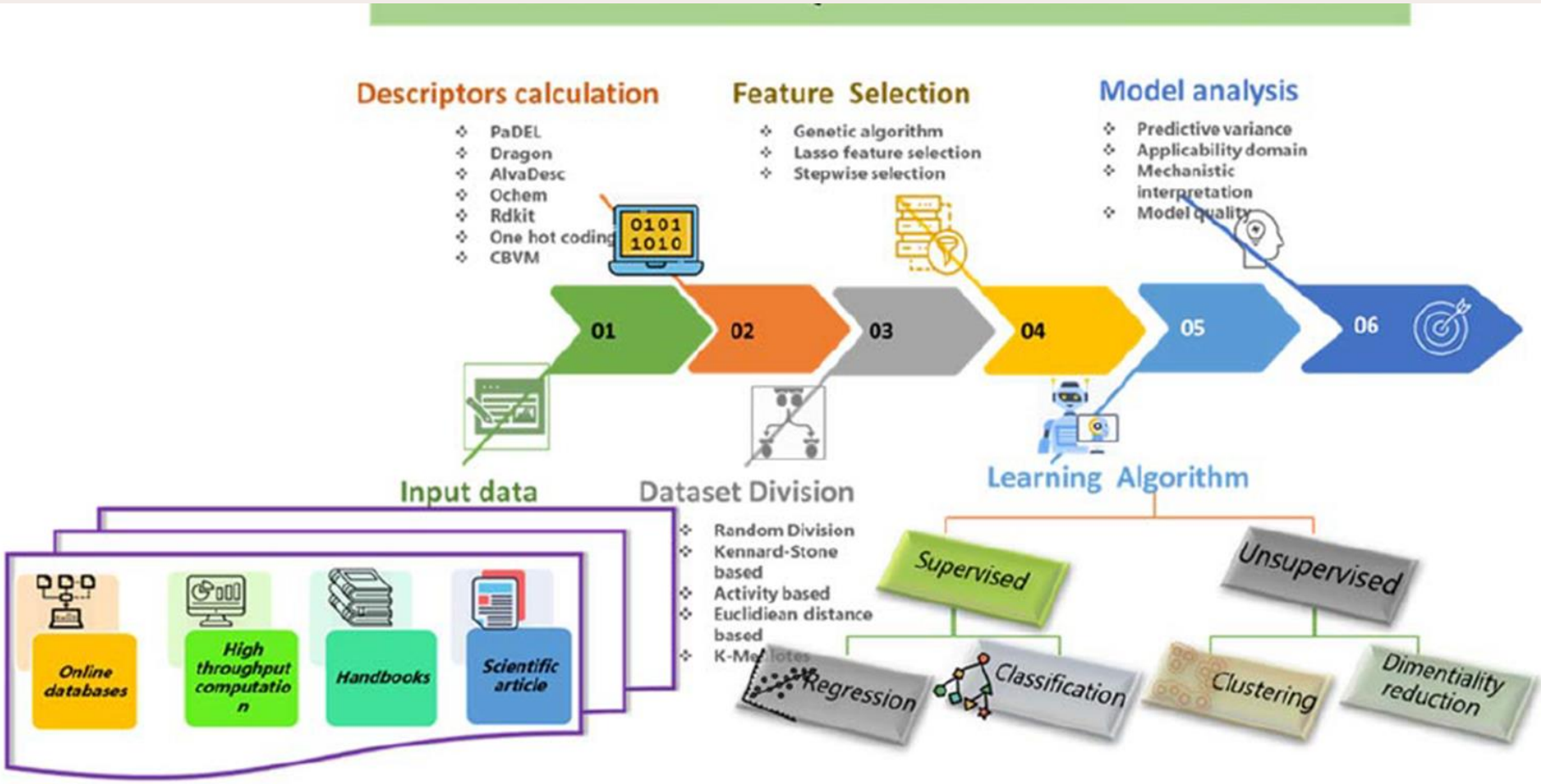


Roy, Kar and Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. ISBN: 978-0-12-801505-6, DOI: http://dx.doi.org/10.1016/B978-0-12-801505-6.00001-6

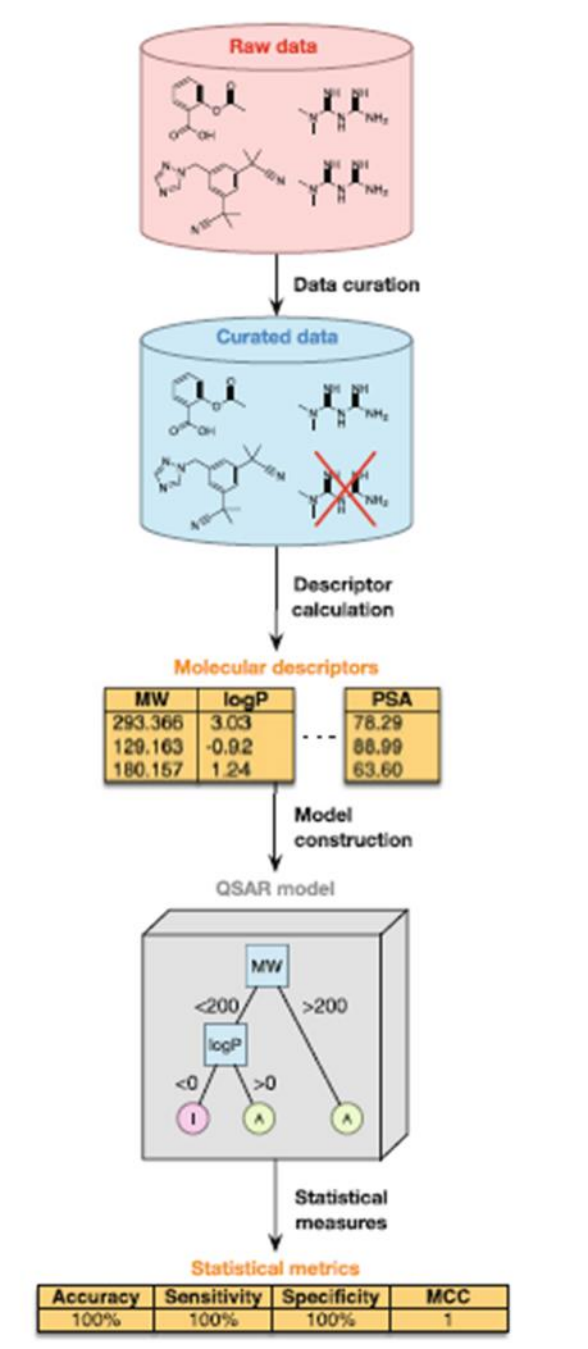
Molecular Descriptors



Consonni et al., Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development (K. Roy ed.),Elsevier Inc. https://doi.org/10.1016/B978-0-443-18638-7.00022-0



Khan et al., QSAR in Safety Evaluation and Risk Assessment, Elsevier, https://doi.org/10.1016/B978-0-443-15339-6.00035-7

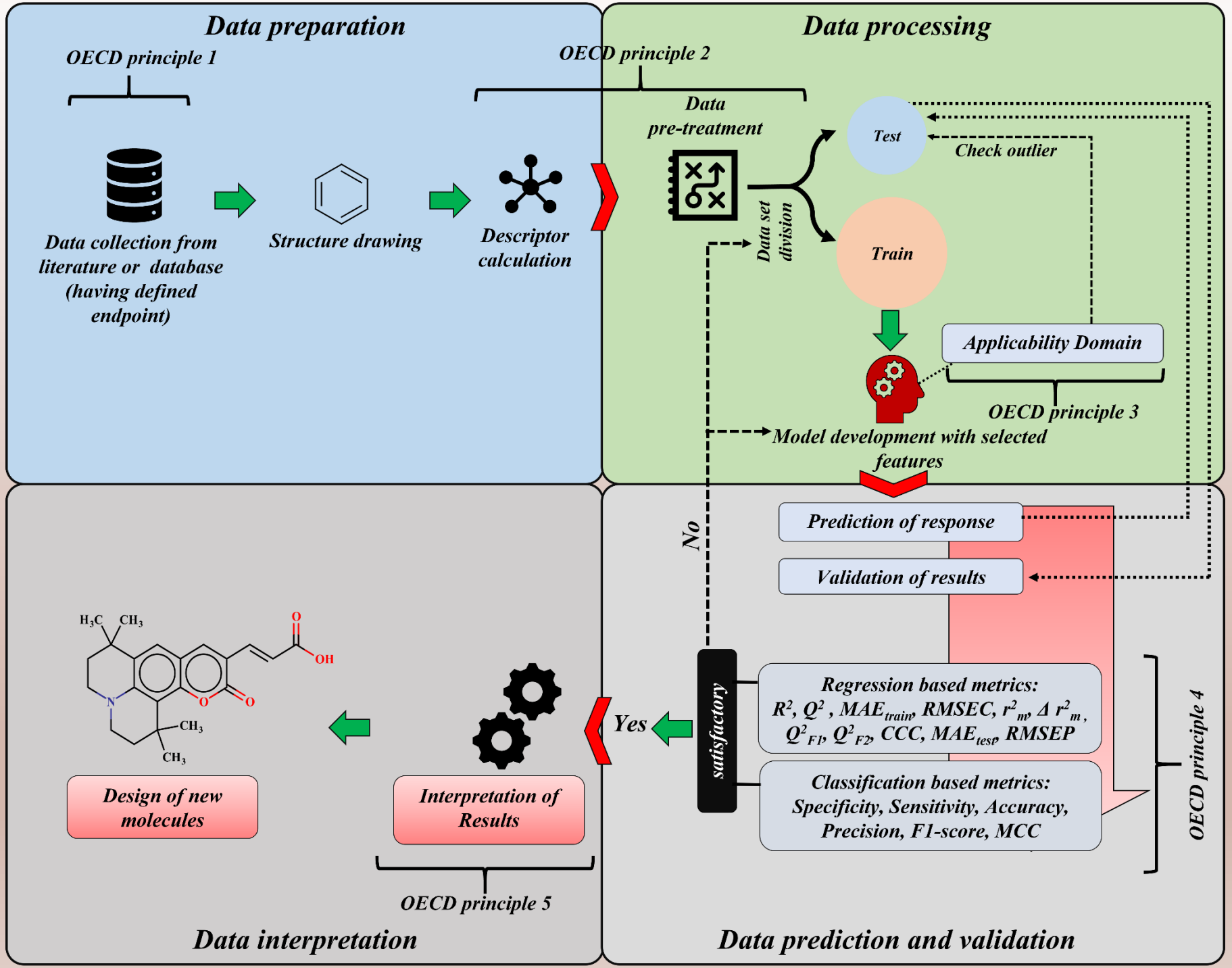


Shoombuatong et al., K. Roy (ed.), Advances in QSAR Modeling, Challenges and Advances in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_1

OECD Guidelines for QSAR model development

- **a defined endpoint;**
- **an unambiguous algorithm;**
- **a defined domain of applicability;**
- **appropriate measures of goodness of fit, robustness and predictivity;**
- **a mechanistic interpretation, if possible.**

Dearden JC, Cronin MTD, Kaiser KLE, *SAR QSAR Environ Res*, **2009**, *20*, 241-266.


Data preparation
OECD principle 1
Data collection from literature or database (having defined endpoint)
Structure drawing
Descriptor calculation
OECD principle 2
Data processing
Data pre-treatment
Data set division
Test
Check outlier
Train
Applicability Domain
OECD principle 3
Model development with selected features
Prediction of response
Validation of results
No
satisfactory
Yes
Regression based metrics:
R^2, Q^2 , MAE_{train}, RMSEC, r^2_m, Δr^2_m , Q^2_{F1}, Q^2_{F2}, CCC, MAE_{test}, RMSEP
Classification based metrics:
Specificity, Sensitivity, Accuracy, Precision, F1-score, MCC
OECD principle 4
Interpretation of Results
OECD principle 5
Design of new molecules
Data interpretation
Data prediction and validation

Chemometric tools for model development

- **Regression based methods**
  - **Method of least squares**
  - **Partial least squares**
- **Classification based methods**
  - **Discriminant analysis**
  - **Logistic regression**
- **Machine learning methods**
  - **Artificial neural network**
  - **Support vector machine**

Requirements for Endpoint data to be modeled (for QSAR)

- **The response to be modeled should be "dose for fixed response" type; e.g., IC_{50}, EC_{50} etc.**
- **The concentration should be measured in a molar unit**
- **The molar concentration (C) should be converted to a log basis; e.g., log1/C or pC**
- **There should be a span of at least 4-5 log units in the response data**
- **There should be sufficient data points present (the ratio of number of data points to number of descriptors should be at least 5:1 for multiple linear regression).**

Metrics for judging quality of QSAR models

Metrics defining statistical quality of the classification based QSAR models		
Sl. No.	**Mathematical definition**	
20	$Sensitivity = \frac{TP}{TP + FN}$	*Internal and external validation metrics*
21	$Specificity = \frac{TN}{TN + FP}$	
22	$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$	
23	$Precision = \frac{TP}{TP + FP}$	
24	$F\text{ - }measure = \frac{2}{1/Precision + 1/Sensitivity}$	
25	$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$	
26	$G\text{-}means = \sqrt{Sensitivity \times Specificity}$	
27	$Cohen's\ \kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$ $P_r(a) = \frac{(TP + TN)}{(TP + FP + FN + TN)}$ $P_r(e) = \frac{\{(TP + FP) \times (TP + FN)\} + \{(TN + FP) \times (TN + FN)\}}{(TP + FN + FP + TN)^2}$	

Small data set modeling

- Small dataset modeling using the QSAR approach has been a very challenging job since a QSAR modeling data set needs to possess sufficient data points to perfectly train itself. To address this problem, different techniques like synthetic sample generation, double cross-validation, consensus predictions, etc., have been used in the literature.
- In spite of the deficiency of sufficient data points, the QSAR modeler may be required to include a higher number of features (descriptors) to encode all available chemical functionalities. In such cases, the statistical aspect is compromised as the ultimate aim of a modeler is to develop highly predictive models using a lower number of descriptors.

Banerjee and Roy, *Environ Sci: Process Impacts*, 2024.

Small data set modeling

- Moreover, the application of a higher number of descriptors coupled with ML algorithms generally tends to generate overfitted models that may not perform well on an external set of data.
- On the flip side of the coin, using a lower number of descriptors may not be able to develop robust and effective models since there is a loss of chemical information associated with the reduction in the number of descriptors.
- This calls for the development of new techniques that use a lower number of descriptors (i.e. a lower degree of freedom) while retaining the chemical information.

Banerjee and Roy, Environ Sci: Process Impacts, 2024
Gramatica P, *QSAR Comb Sci*, **2007**, *26*, 694-701.

Environmental Science Processes & Impacts

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d4em00173g

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*

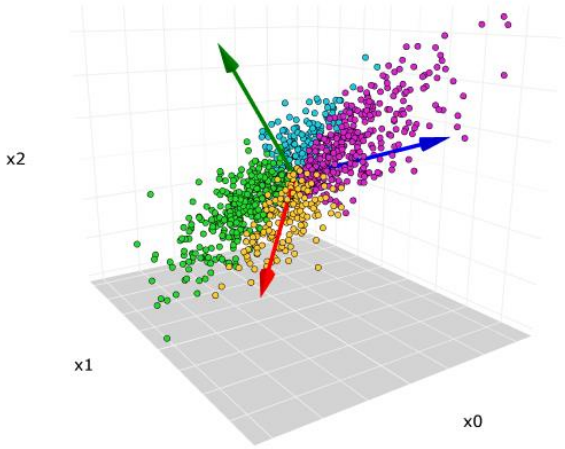
Due to the lack of experimental toxicity data for environmental chemicals, there arises a need to fill data gaps by *in silico* approaches. One of the most commonly used *in silico* approaches for toxicity assessment of small datasets is the Quantitative Structure–Activity Relationship (QSAR), which generates predictive models for the efficient prediction of query compounds. However, the reliability of the predictions from QSARs derived from small datasets is often questionable from a statistical point of view. This is due to the presence of a larger number of descriptors as compared to the number of training compounds, which reduces the degree of freedom of the developed model. To reduce the overall prediction error for a particular QSAR model, we have proposed here the computation of the novel Arithmetic Residuals in *K*-groups Analysis (ARKA) descriptors. We have reduced the number of modeling descriptors in a supervised manner by partitioning them into *K* classes ($K = 2$ here) depending on the higher mean normalized values of the descriptors to a particular response class, thus preventing the loss of chemical information. A scatter plot of the data points using the values of two ARKA descriptors (ARKA_2 *vs.* ARKA_1) can potentially identify activity cliffs, less confident data points, and less modelable data points. We have used here five representative environmentally relevant endpoints (skin sensitization, earthworm toxicity, milk/plasma partitioning, algal toxicity, and rodent carcinogenicity of hazardous chemicals) with graded responses to which the ARKA framework was applied for classification modeling. On comparing the performance of the models generated using conventional QSAR descriptors and the ARKA descriptors, the prediction quality of the models derived from ARKA descriptors was found, based on multiple graded-data validation metrics-derived decision criteria, much better than the models derived from QSAR descriptors signifying the potential of ARKA descriptors in ecotoxicological classification modeling of small data sets. Additionally, this holds true for the Read-Across approach as well, since the Read-Across predictions using ARKA descriptors supersede the predictions generated from QSAR descriptors. For the ease of users, a Java-based expert system has been developed that computes the ARKA descriptors from the input of QSAR descriptors.

Received 29th March 2024
Accepted 5th May 2024

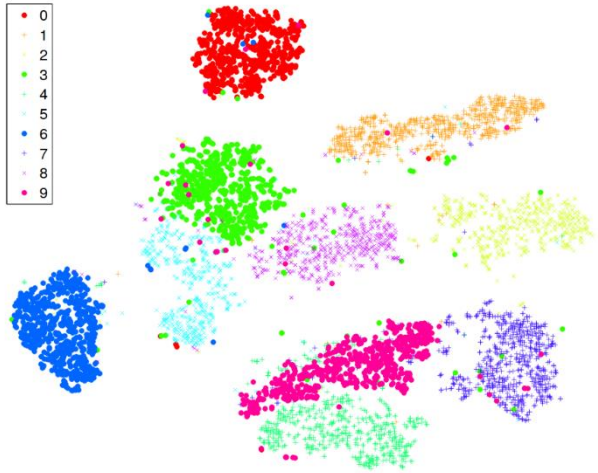
DOI: 10.1039/d4em00173g

rsc.li/espi

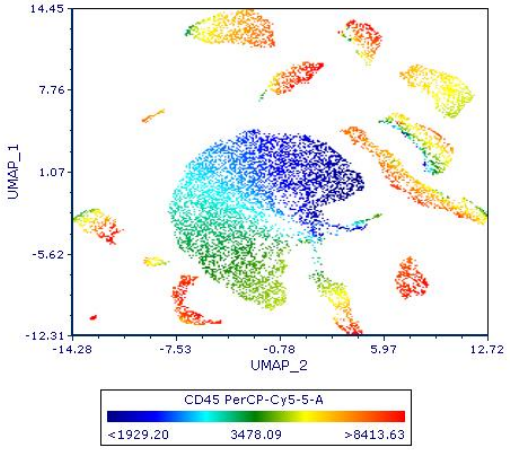
Dimensionality reduction methods



PCA

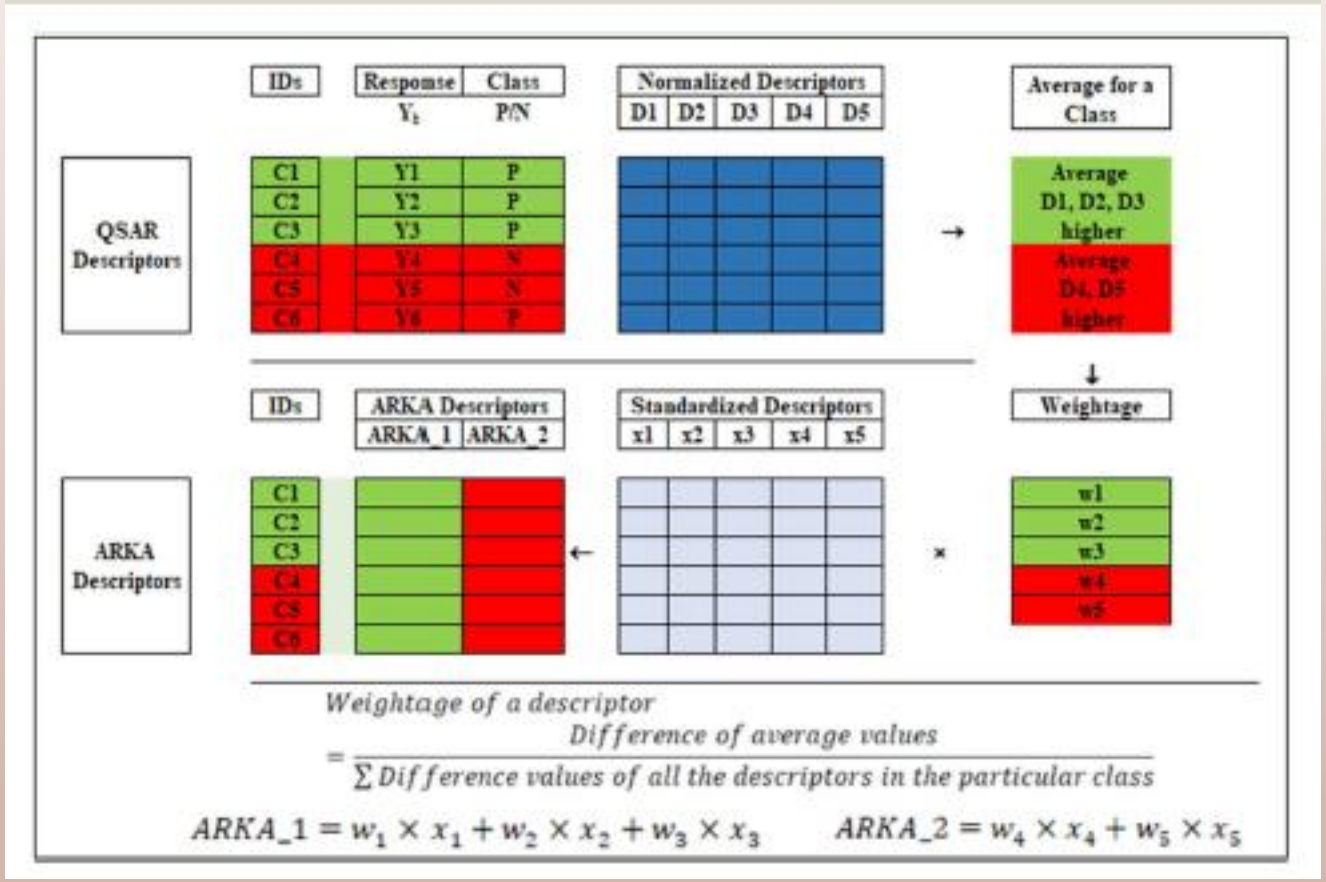


t-SNE

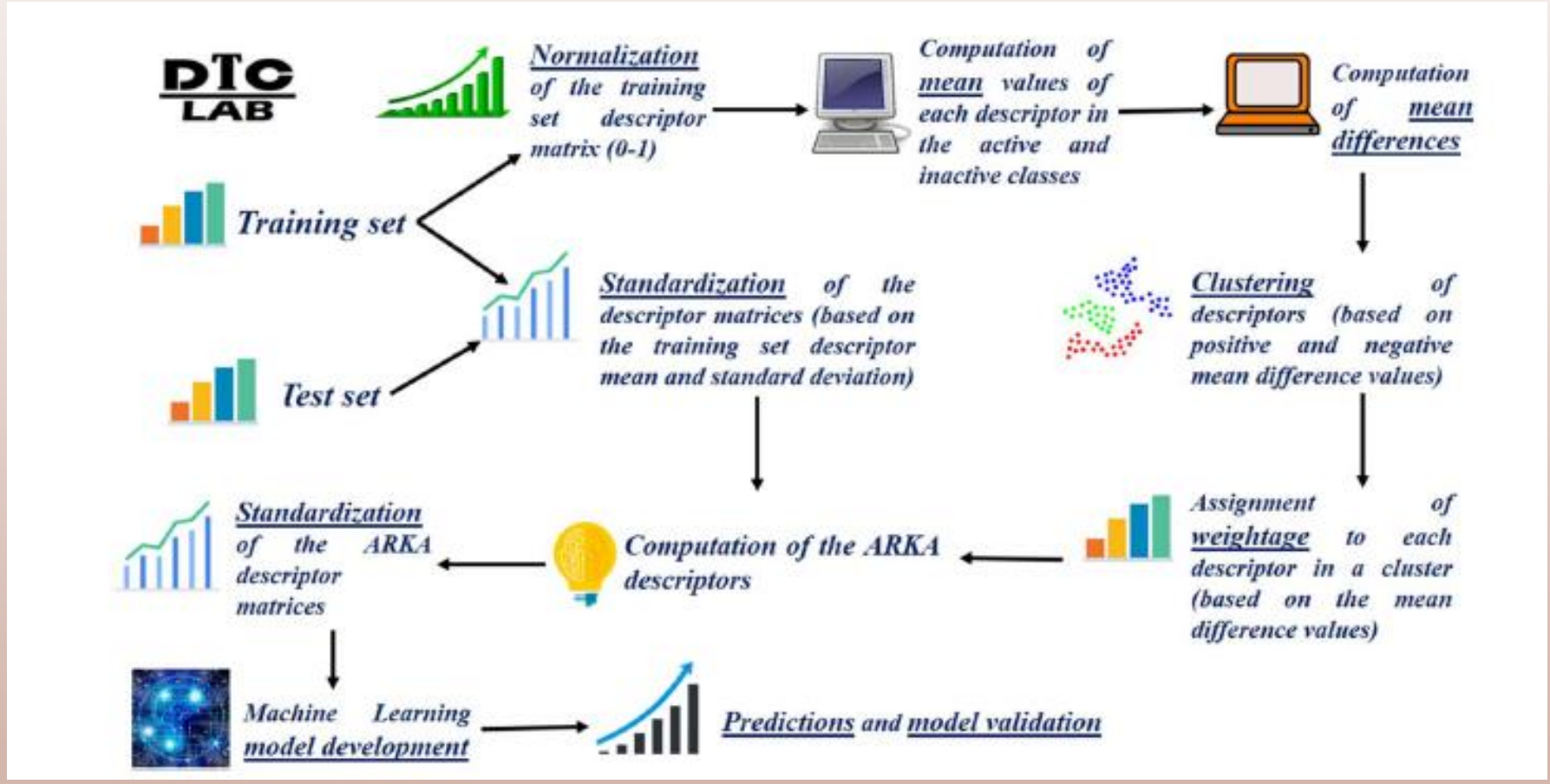


UMAP

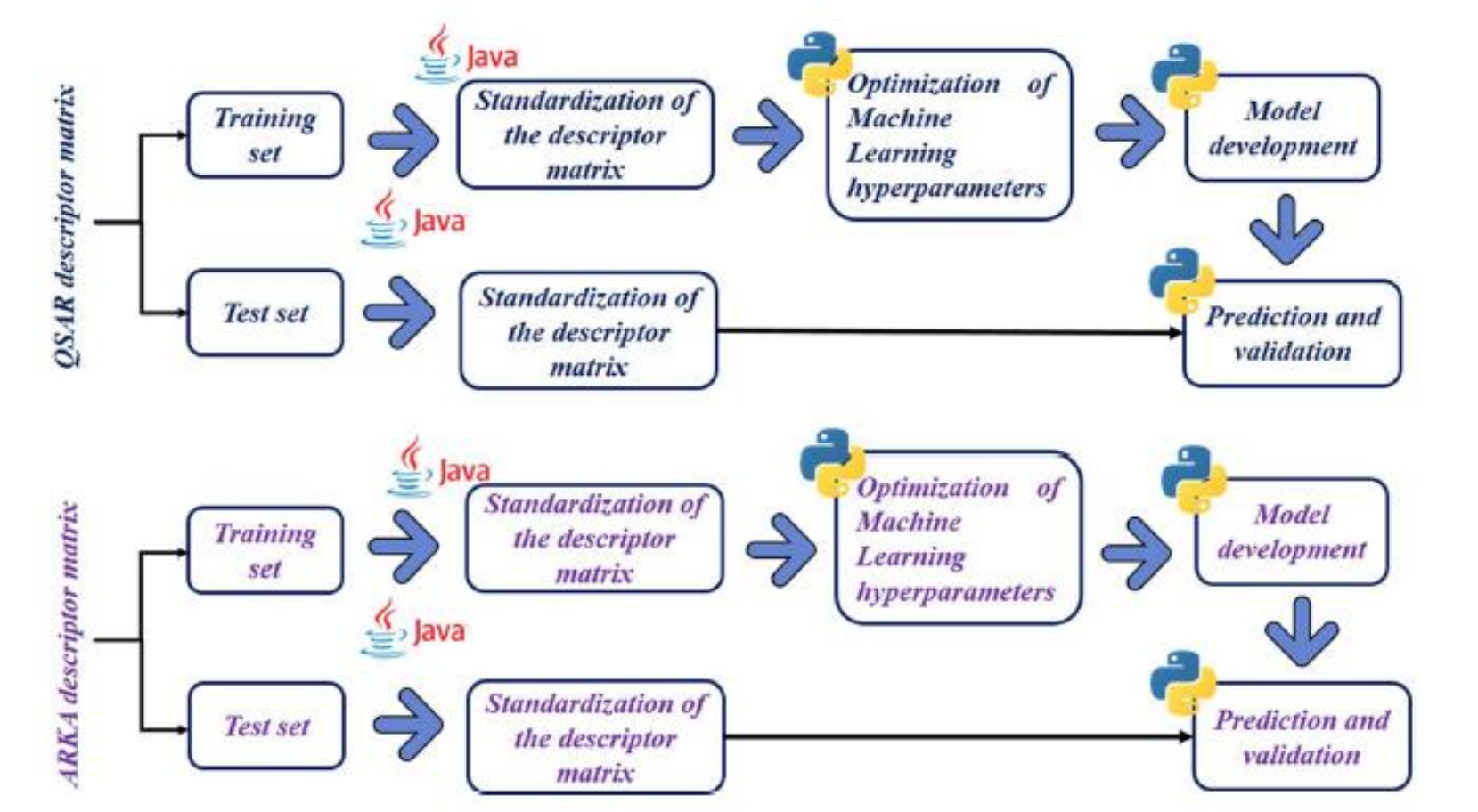
ARKA: Computation



ARKA: Workflow



ARKA: modeling workflow



Data sets

1. ***Skin sensitization potential:*** *Banerjee A, Roy K. Chem. Res. Toxicol. 2023, 36, 1518-1531*

2. ***Earthworm toxicity:*** *Roy J, Ojha PK, Carnesecchi E, Lombardo A, Roy K, Benfenati E. J. Hazard. Mater. 2020, 386, 121660*

3. ***Milk/Plasma concentration ratio of drugs and environmental pollutants:*** *Kar S, Roy K. Mol. Inform. 2013, 32, 693-705*

4. ***Toxicity towards Algae****: Pramanik S, Roy K. Ecotox. Environ. Safety 2014, 101, 184-190*

5. ***Rodent carcinogenicity****: Kar S, Deeb O, Roy K. Ecotox. Environ. Safety 2012, 82, 85-95*

Results of the model performance on the test set data

Dataset 1

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

- ✓ *LDA: Linear Discriminant Analysis*
- ✓ *SVM: Support Vector Machine*
- ✓ *RF: Random Forest*
- ✓ *LR: Logistic Regression*

Dataset 2

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

Results of the model performance on the test set data

Dataset 3

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

- ✓ *LDA: Linear Discriminant Analysis*
- ✓ *SVM: Support Vector Machine*
- ✓ *RF: Random Forest*
- ✓ *LR: Logistic Regression*

Dataset 4

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

Results of the model performance on the test set data

Dataset 5

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

- ✓ *LDA: Linear Discriminant Analysis*
- ✓ *SVM: Support Vector Machine*
- ✓ *RF: Random Forest*
- ✓ *LR: Logistic Regression*

Voting of the predictive performance on the test set data

#1

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

#2

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

#3

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

#4

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

#5

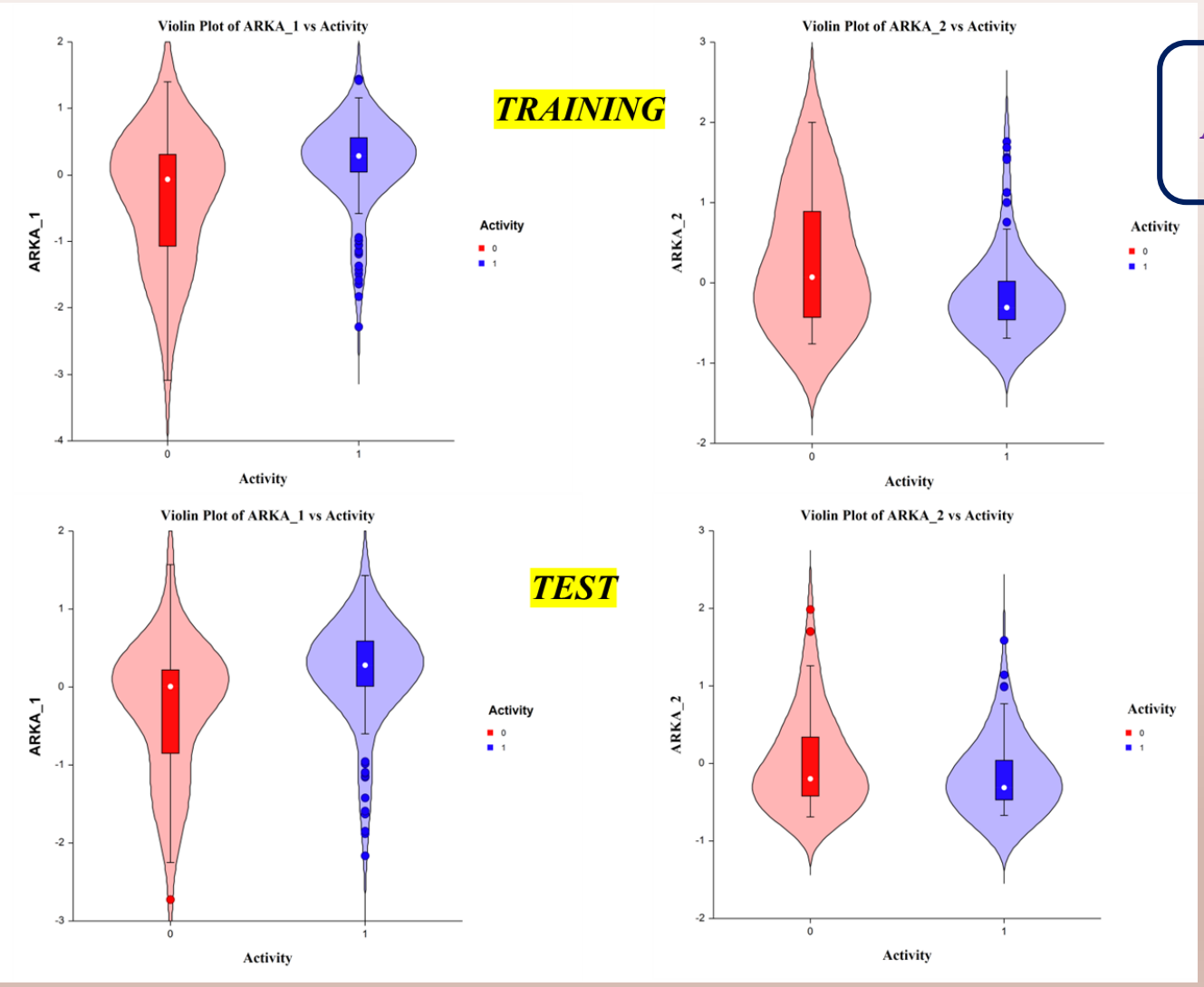
	QSAR descriptors				ARKA descriptors			
Models	F1_score	MCC	Ckappa	AUC	F1_score	MCC	Ckappa	AUC
LDA	1	1	1	1	0	0	0	0
LR	0	0	0	1	1	1	1	0
SVM	0	0	0	0	1	1	1	1
RF	0	0	0	0	1	1	1	1
Sum	1	1	1	2	3	3	3	2

composite

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

Loser
Winner
Tie

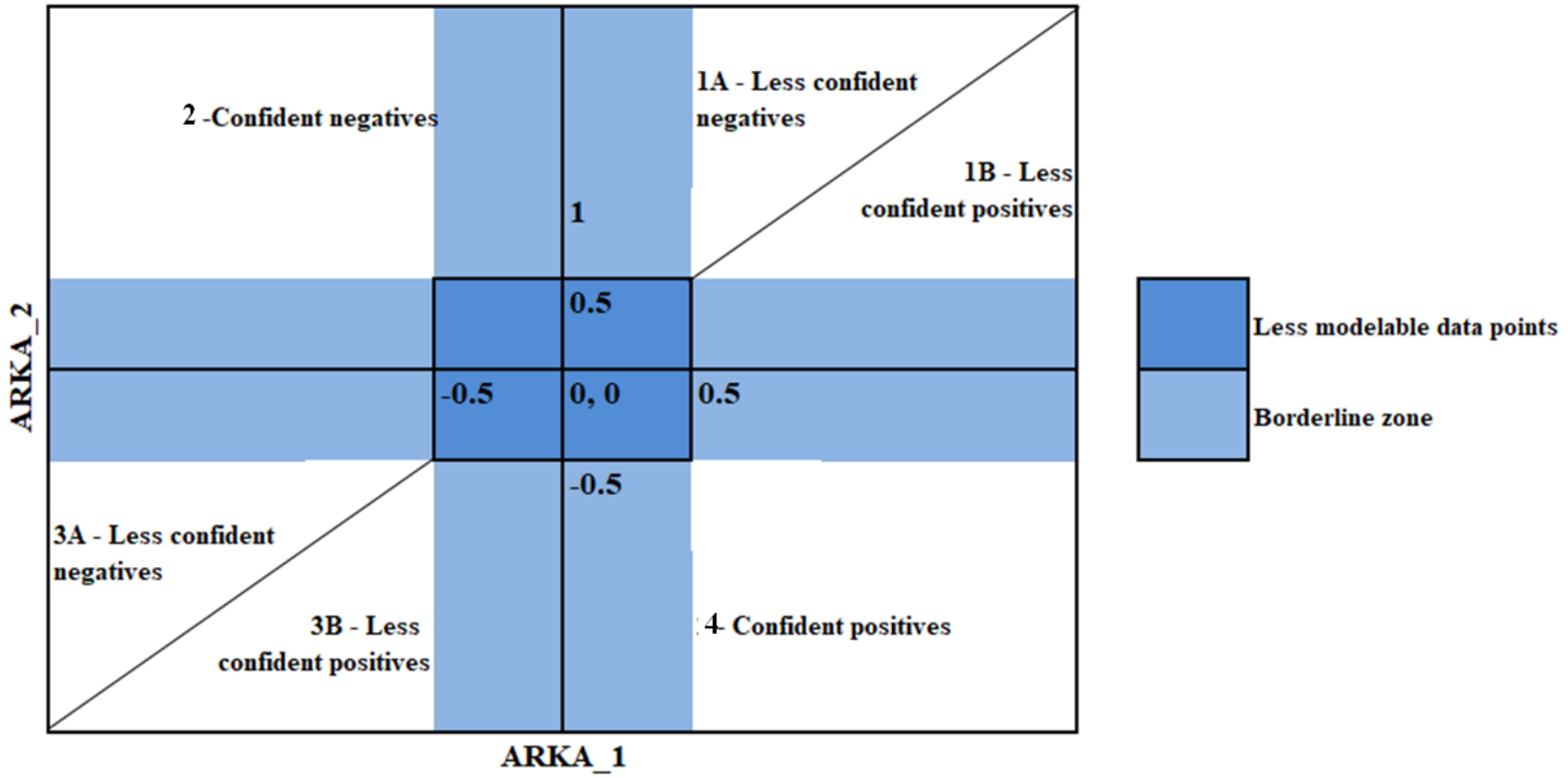
Loser
Overall winner
Overall tie



Values of ARKA_1 and ARKA_2 in the active and inactive classes

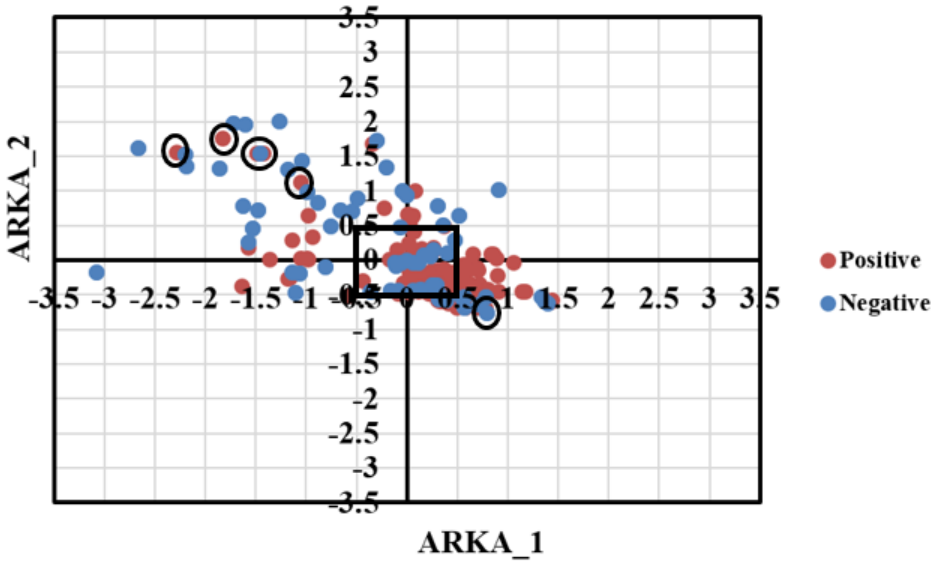
- ✓ ***Representative example of Dataset 1***
- ✓ ***Median values of ARKA_1 is higher in the active class***
- ✓ ***Median values of ARKA_2 is higher in the inactive class***

Analysis of – Modelability, Activity cliffs and less confident data points



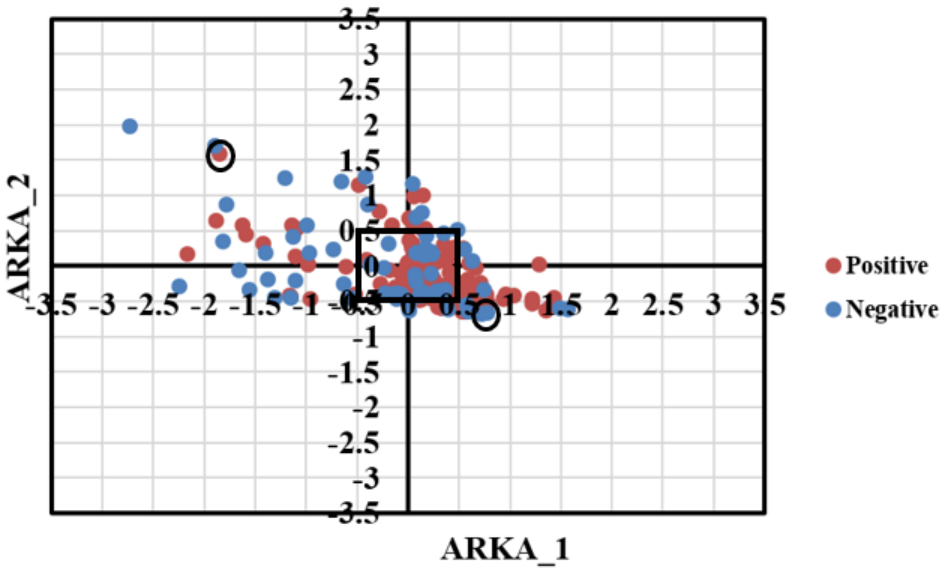
Identification of activity and prediction cliffs

ARKA_2 vs ARKA_1 (Training set, Dataset 1)



Activity cliffs: 260, 378, 320, 362, 370, 93

ARKA_2 vs ARKA_1 (Test set, Dataset 1)



Prediction cliffs: 74, 94, 349

✓ *Representative example of Dataset 1*

Analysis of the chemical Read-Across predictions

Table 2 Effects of ARKA descriptors on the chemical Read-Across-based external predictions using the Gaussian kernel function for five data sets (N_{desc} = the number of descriptors, MCC = Matthews correlation coefficient, C_{kappa} = Cohen's kappa)[a]

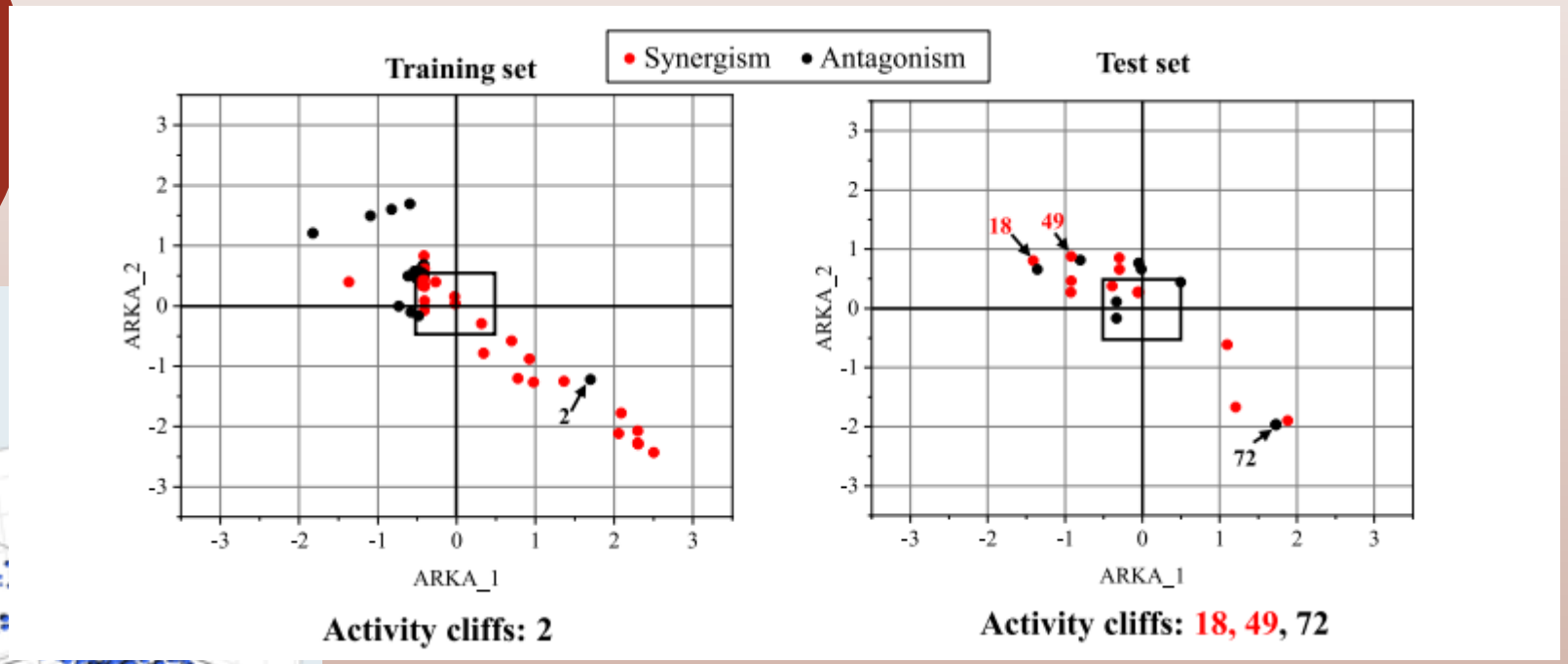
Dataset	Descriptors	N_{desc}	F1_score	MCC	C_{kappa}	AUC
1	QSAR	14	**0.729**	0.21	0.209	**0.66**
	ARKA	2	0.699	**0.235**	**0.227**	**0.66**
2	QSAR	8	0.6	0.42	0.412	0.78
	ARKA	2	**0.645**	**0.472**	**0.467**	**0.79**
3	QSAR	6	0.361	−0.079	−0.079	0.43
	ARKA	2	**0.375**	**−0.144**	**−0.143**	**0.49**
4	QSAR	4	0.9	0.753	0.723	0.95
	ARKA	1	**0.923**	**0.812**	**0.795**	**1**
5	QSAR	4	**0.917**	**0.713**	**0.673**	**0.96**
	ARKA	2	**0.917**	**0.713**	**0.673**	0.95

[a] The winner metric values are shown in bold.

✓ ***Default settings of the hyperparameters***

✓ ***Gaussian Kernel similarity-based predictions***

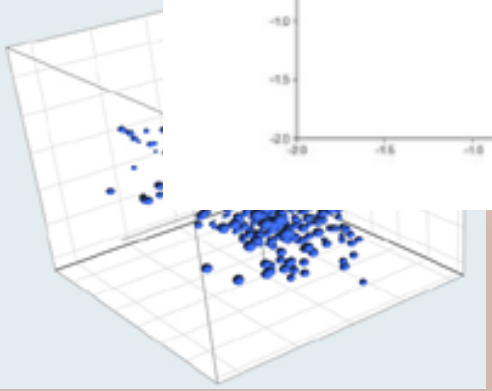
Identification of activity cliffs

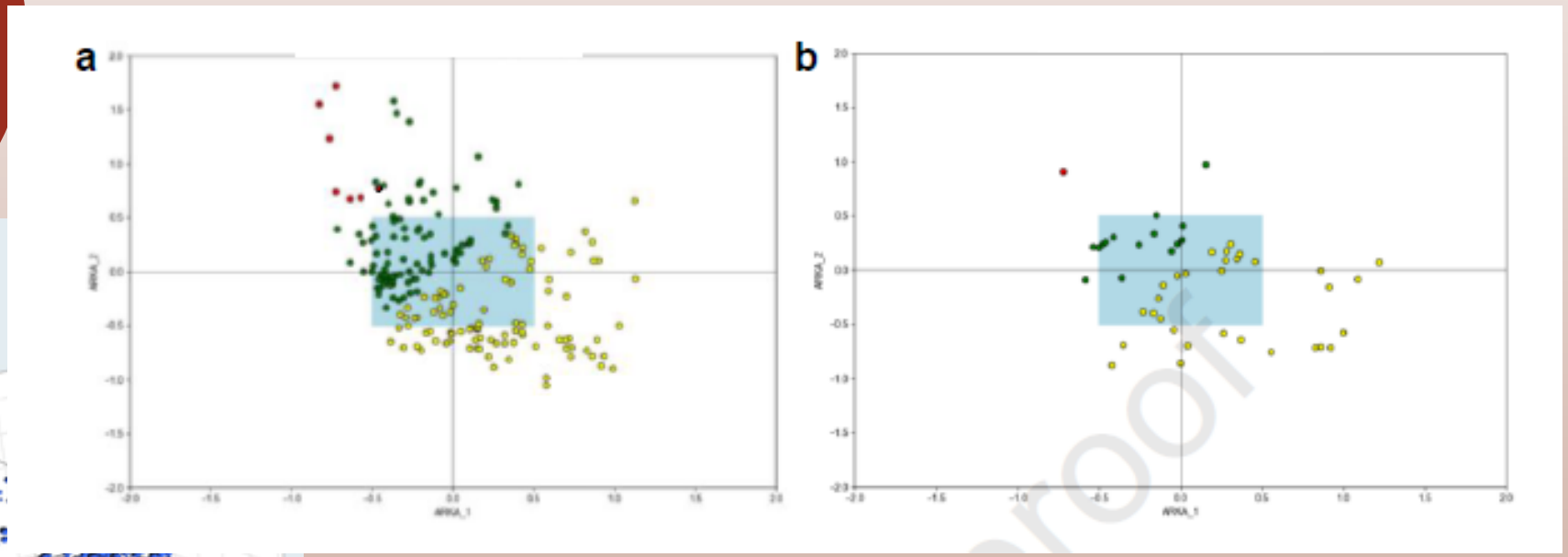


Toxicities of binary mixtures of antibiotics and fungicides against *Auxenochlorella pyrenoidosa*.

Qin et al., Environmental Pollution 360, 2024, 124565

Identification of activity cliffs





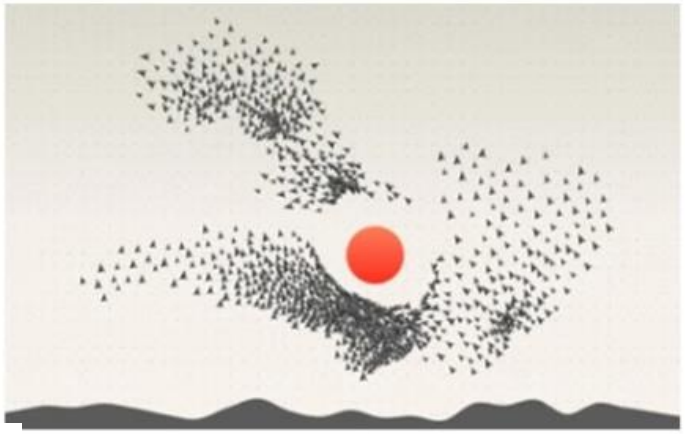
Ionic liquid toxicity.

Shan et al., Green Chemical Engineering, 2024

Conclusion

- The **ARKA** framework, a supervised dimensionality reduction technique conceptualized and developed by the DTC Laboratory, can potentially identify activity cliffs, less confident and less modelable data points and should be useful for the classification modeling of small data sets.
- There is room for further development of the approach by its applications in regression-based and/or Read-Across approaches, classification modeling of larger ecotoxicity data sets, and exploring other customized ways of weighing strategies in deriving **ARKA** descriptors.

Development of a Java-based ARKA descriptor calculating tool



DTC LAB

Arithmetic

Residuals in

K-Groups

Analysis

V 1.0

This program calculates ARKA descriptors for the dimensionality reduction of QSAR descriptor matrix for classification modeling of small data sets

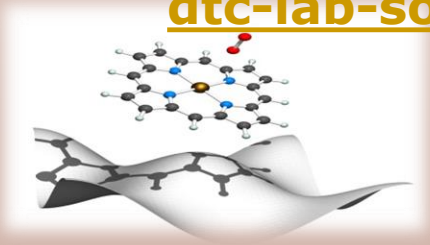
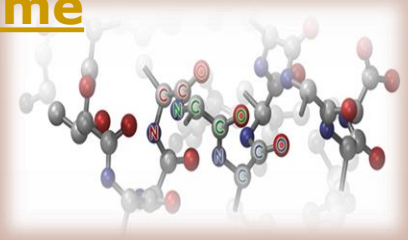
Software developed by Arkaprava Banerjee (arka.banerjee16@gmail.com)

Picture Courtesy Shutterstock

The Drug Theoretics and Cheminformatics (DTC) Laboratory

http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/

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


Acknowledgements

विज्ञान एवं प्रौद्योगिकी विभाग
DEPARTMENT OF
SCIENCE & TECHNOLOGY

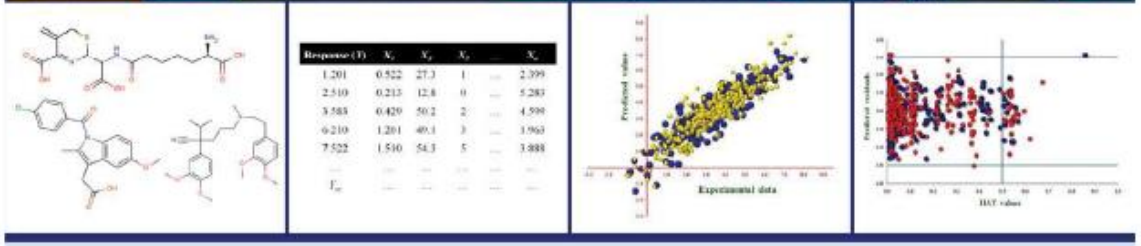
Anusandhan National
Research Foundation
(ANRF), DST, New Delhi

DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology

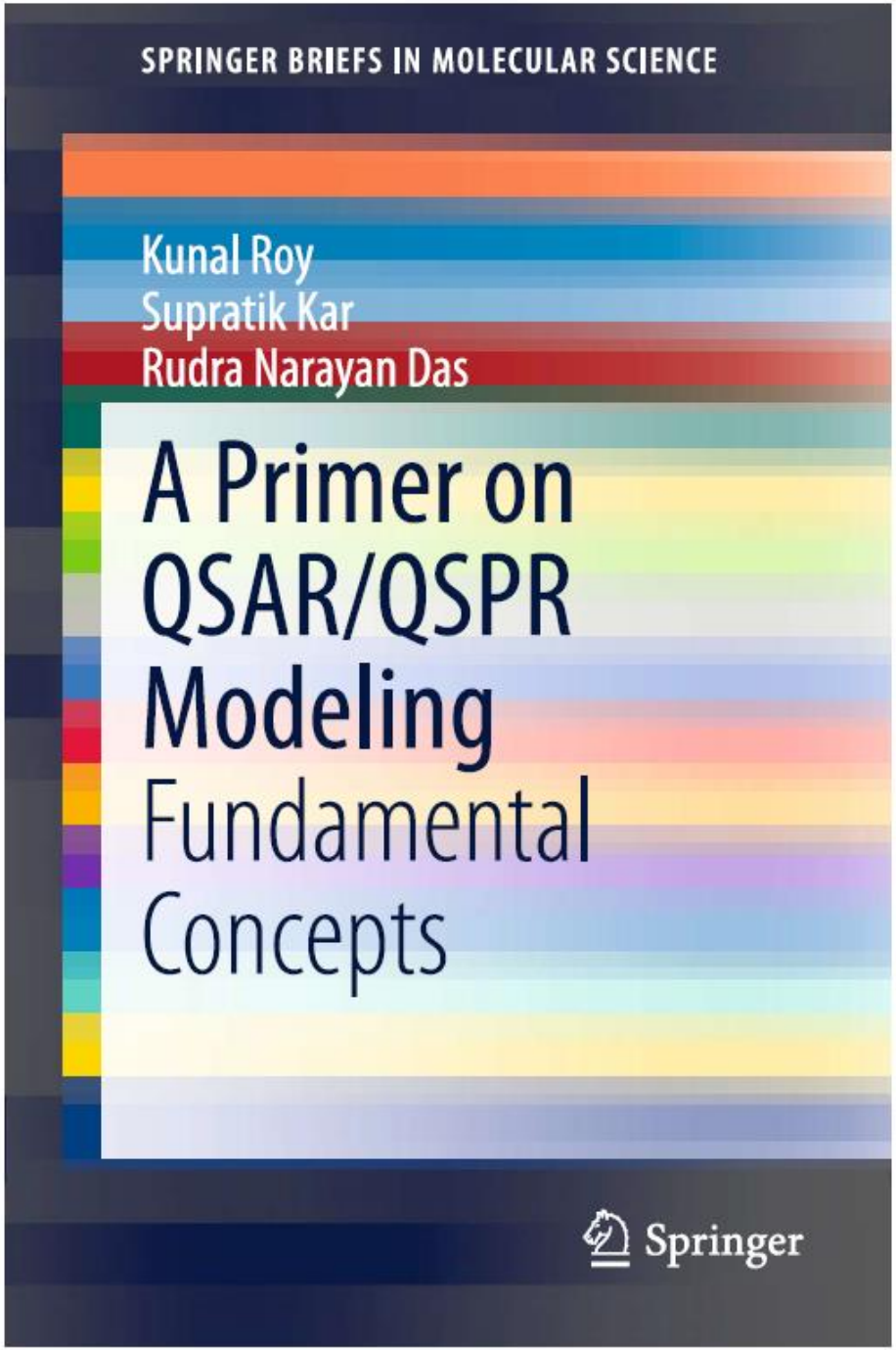
विज्ञान एवं प्रौद्योगिकी विभाग
DEPARTMENT OF
SCIENCE & TECHNOLOGY

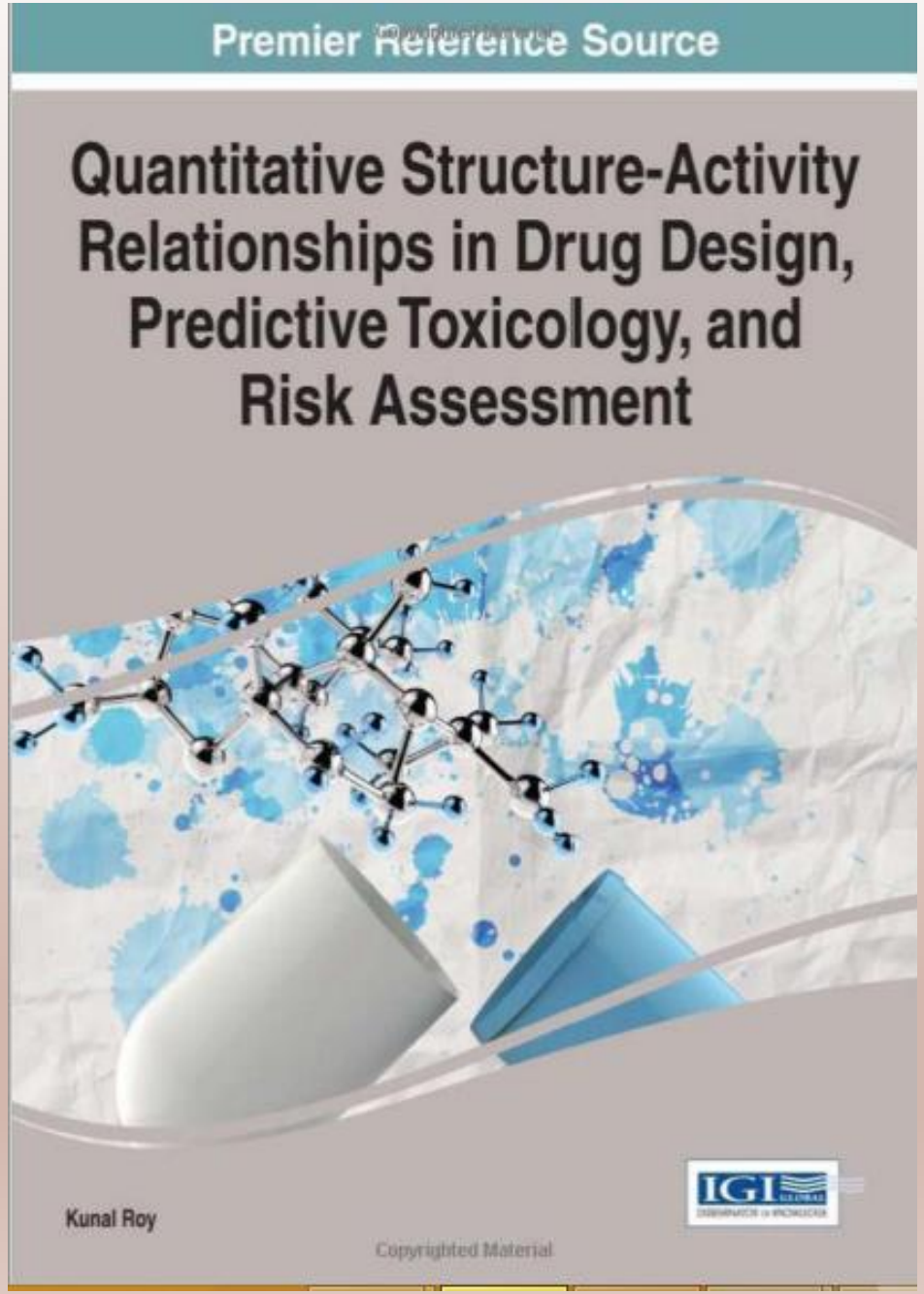
Software Programs developed in Java by
Pravin Ambure (ambure.pharmait@gmail.com)
Arkaprava Banerjee (arka.banerjee16@gmail.com)

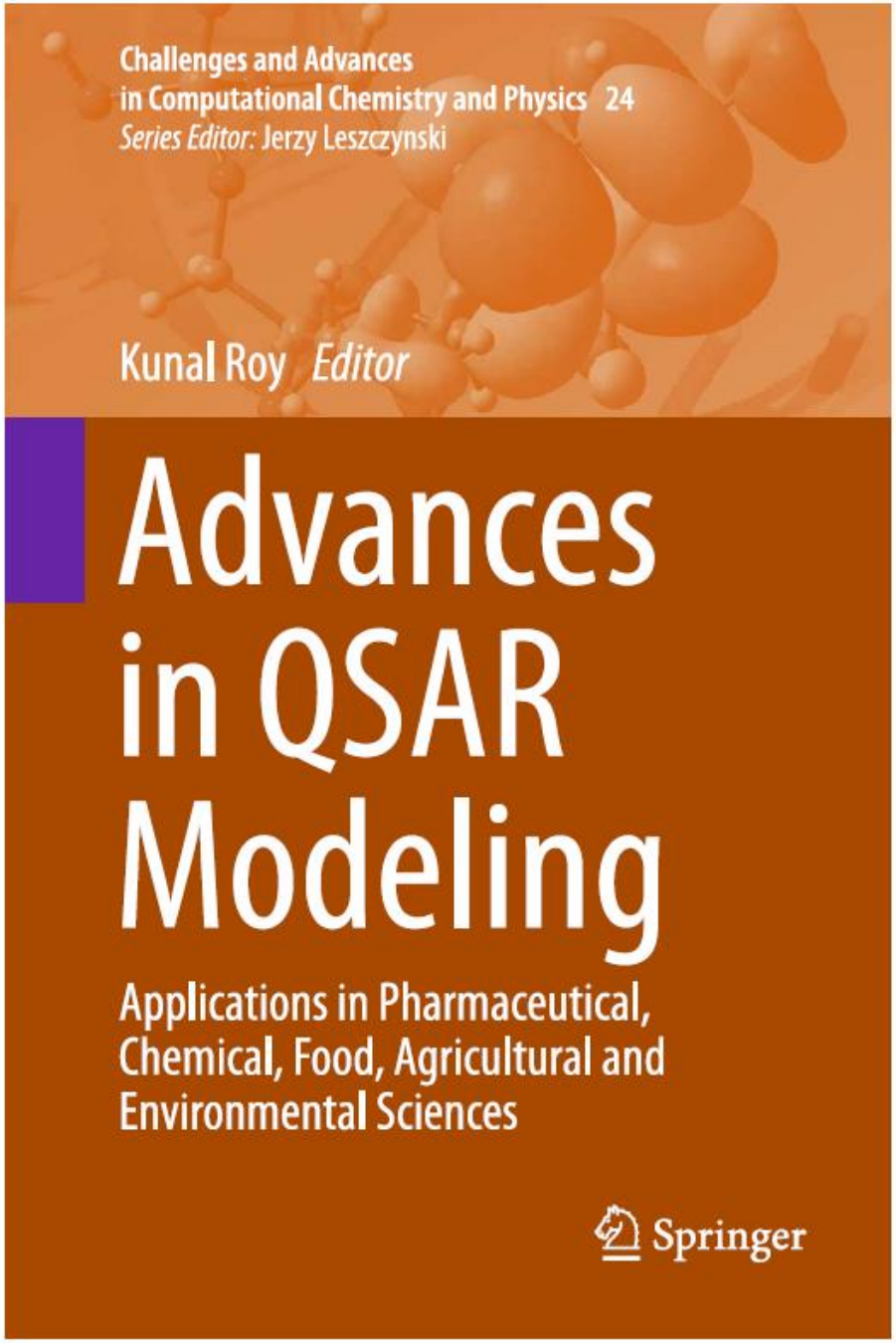
Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment

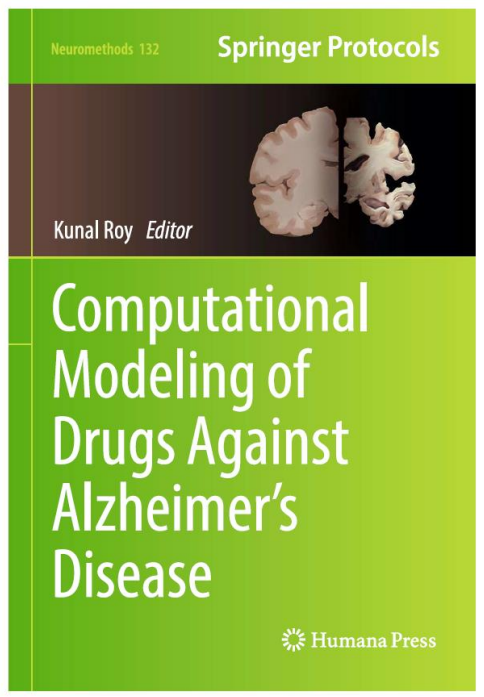


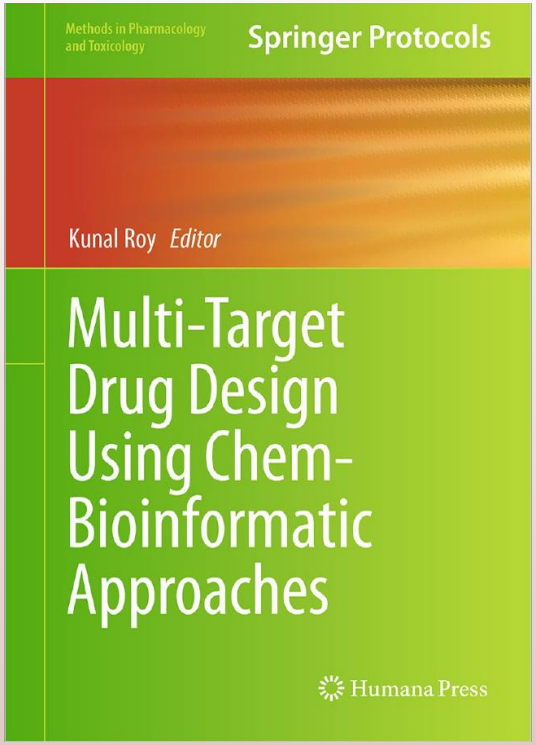
Kunal Roy, Supratik Kar
Rudra Narayan Das

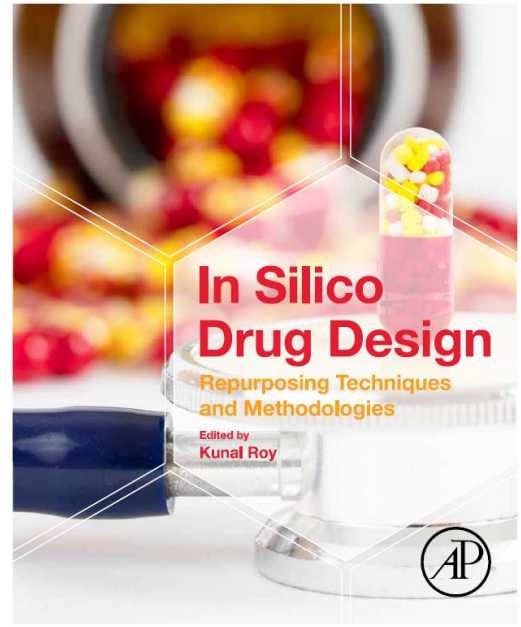

SPRINGER BRIEFS IN MOLECULAR SCIENCE
Kunal Roy
Supratik Kar
Rudra Narayan Das
A Primer on QSAR/QSPR Modeling
Fundamental Concepts
Springer

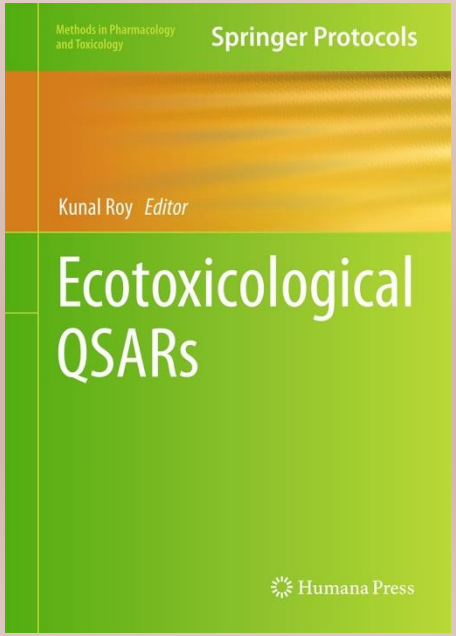

Premier Reference Source
Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment
Kunal Roy
IGI
Copyrighted Material


Challenges and Advances
in Computational Chemistry and Physics 24
Series Editor: Jerzy Leszczynski
Kunal Roy Editor
Advances in QSAR Modeling
Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences
Springer

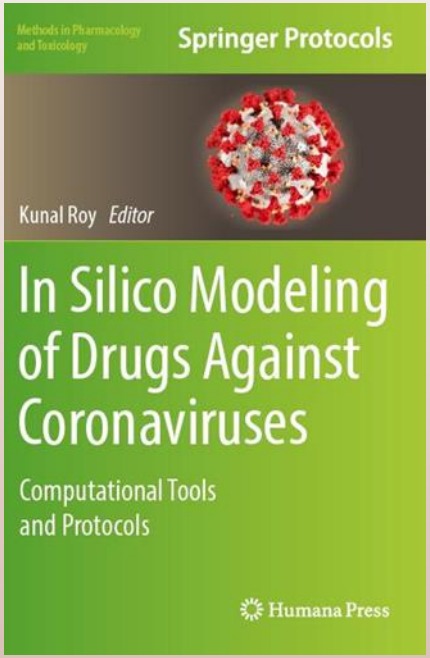

Neuromethods 132
Springer Protocols
Kunal Roy Editor
Computational Modeling of Drugs Against Alzheimer's Disease
Humana Press

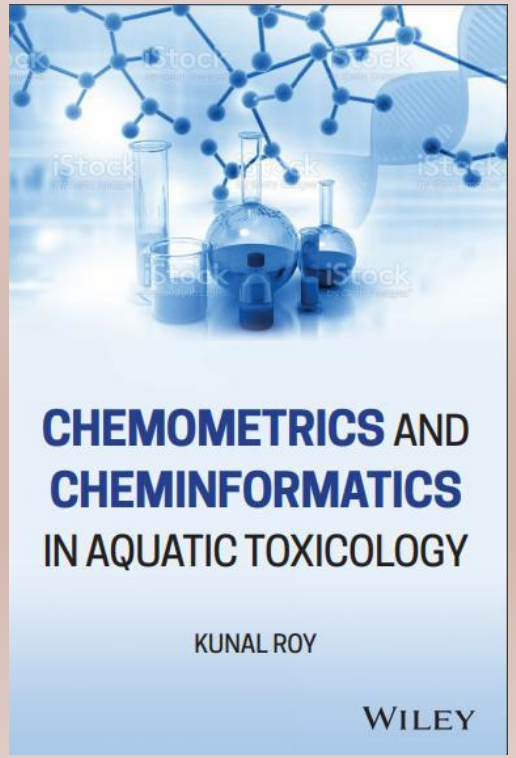

Methods in Pharmacology and Toxicology
Springer Protocols
Kunal Roy Editor
Multi-Target Drug Design Using Chem-Bioinformatic Approaches
Humana Press

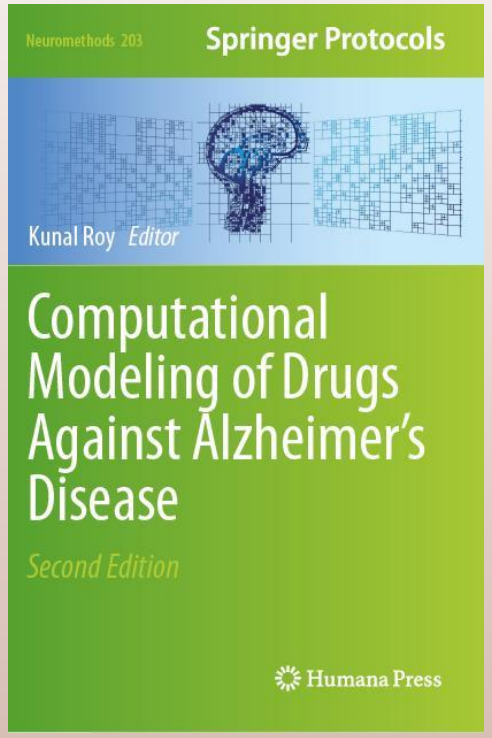

In Silico Drug Design
Repurposing Techniques and Methodologies
Edited by
Kunal Roy


Methods in Pharmacology and Toxicology
Springer Protocols
Kunal Roy Editor
Ecotoxicological QSARs
Humana Press

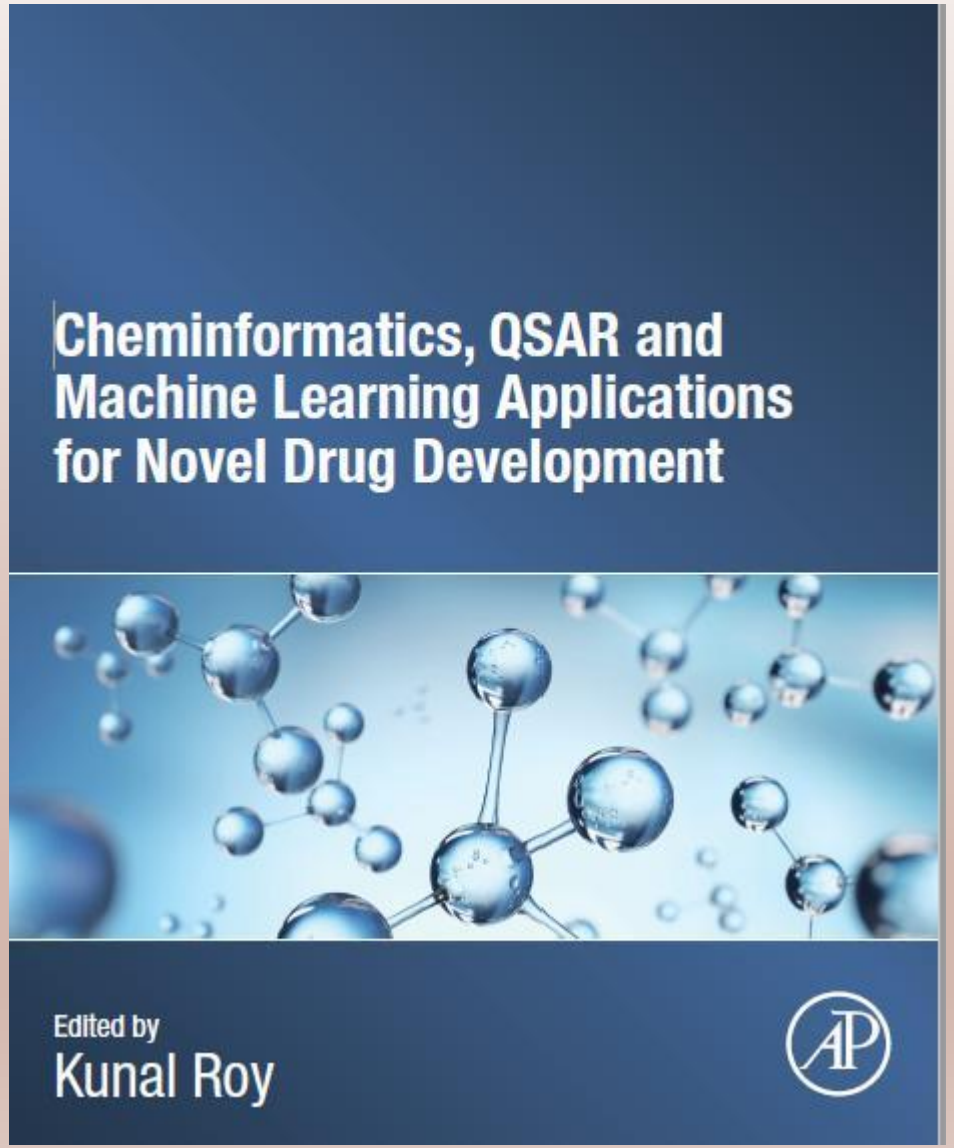
Recent titles (2021-2023)

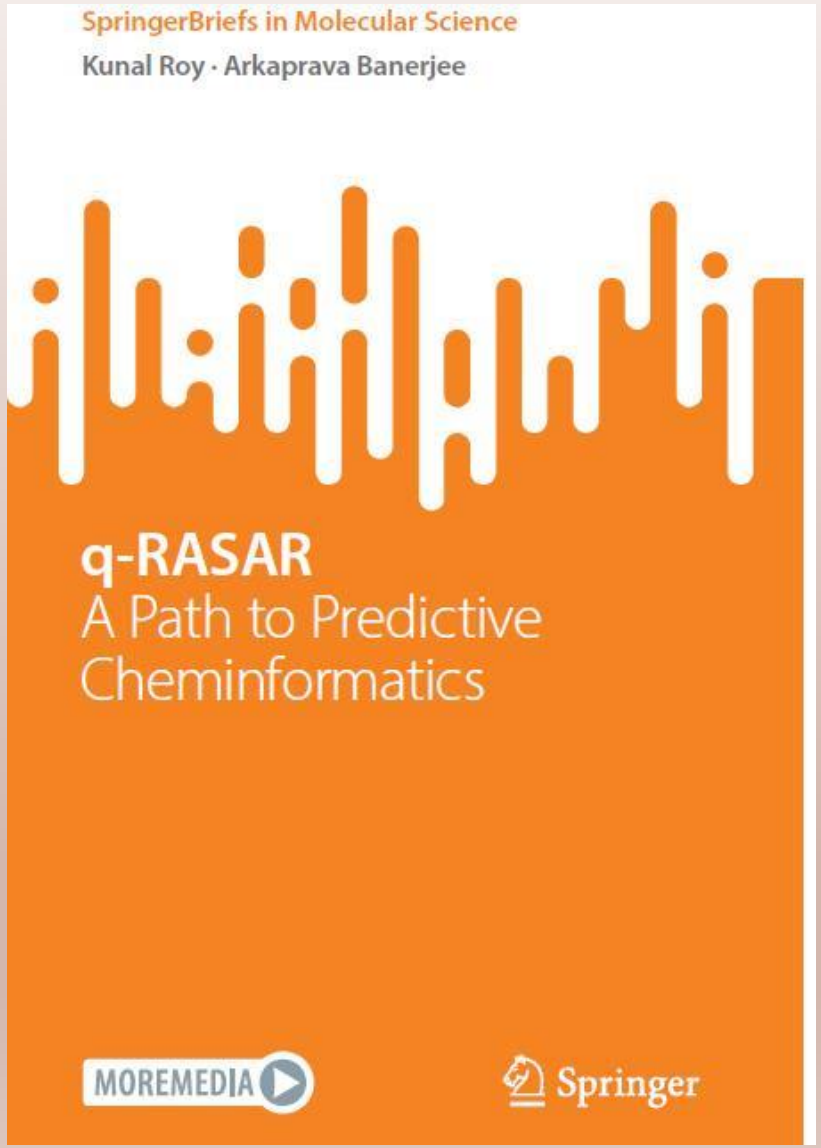






Recent titles (2023-2024)






Thank You!