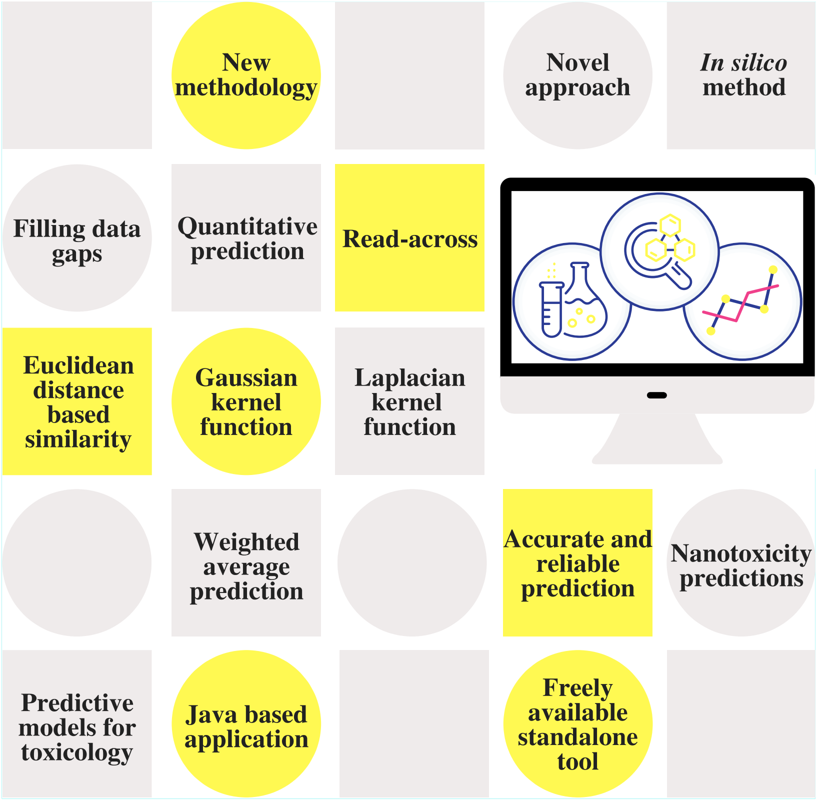
CHEMICAL READ-ACROSS PREDICTIONS OF ECOTOXICITY DATA



New methodology

Novel approach

In silico method

Filling data gaps

Quantitative prediction

Read-across

Euclidean distance based similarity

Gaussian kernel function

Laplacian kernel function

Weighted average prediction

Accurate and reliable prediction

Nanotoxicity predictions

Predictive models for toxicology

Java based application

Freely available standalone tool

Kunal Roy

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India.

In silico methods of toxicity assessment

- ❑**The effect of hazardous chemicals and pollutants on the ecosystem is a matter of great concern.**

- ❑**Since there is large number of chemicals currently in common use (approx. 100,000) and new chemicals are registered at a very high rate (1000 per year), it is obvious that our human and material resources are insufficient to obtain experimentally even basic information on environmental fate and effects for all these chemicals.**

- ❑**Thus, it is necessary to develop quantitative models that will accurately and readily predict environmental behaviour of large sets of chemicals.**

In silico methods of toxicity assessment

- Time and cost effective
- Avoids animal experimentation
- Supports “3R” Principles
- Can be applied for virtual compounds
- Supported by various organizations like

  European Centre for the Validation of Alternative Methods (ECVAM)

  International Organizations of Medical Sciences

  REACH (Registration, Evaluation and Authorization of Chemicals) regulations

  US EPA

  Organization for Economic Cooperation and Development (OECD)

Roy K, *Expert Opin Drug Discov,* **2007**, *2*, 1567-1577

What is 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**.

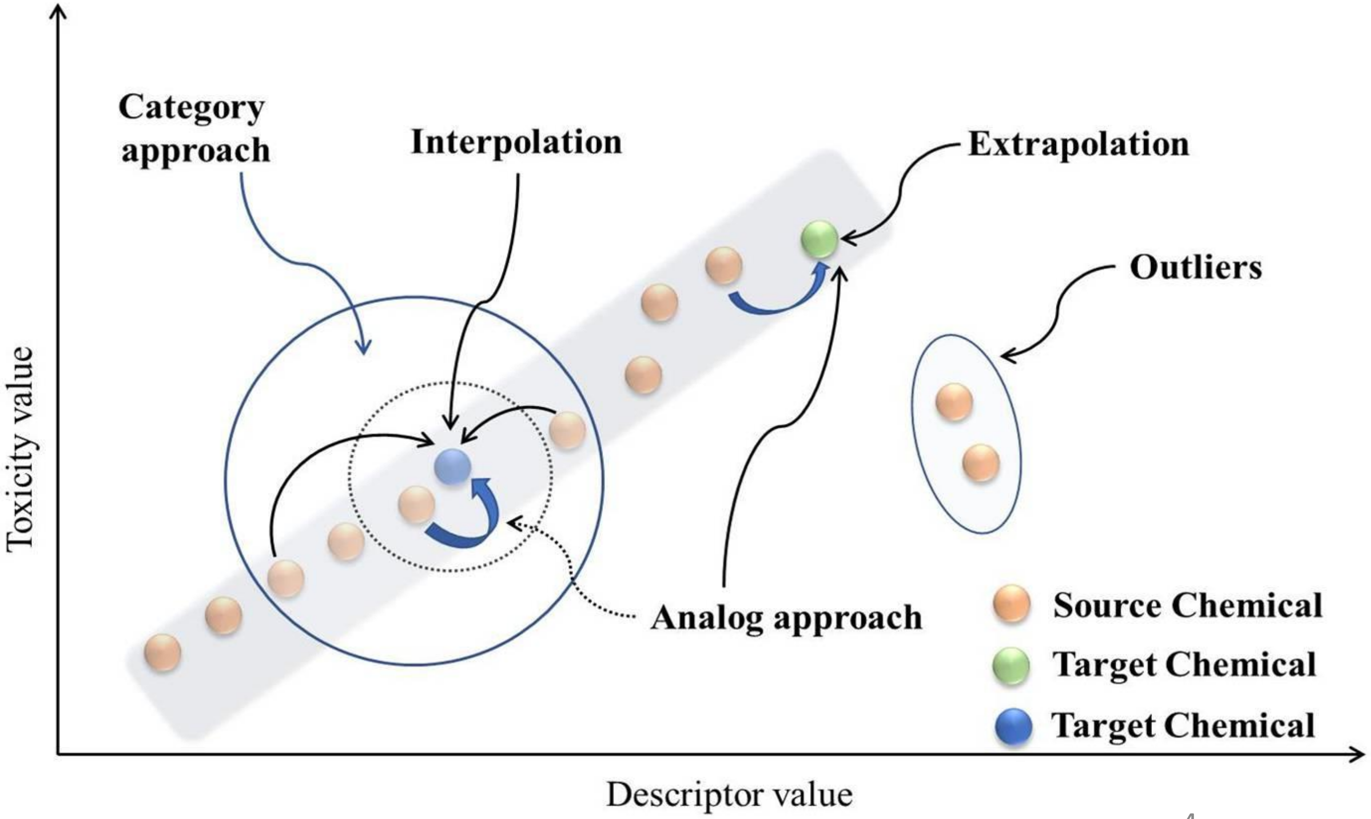
•Used for **data gap filling**.

•Defined chemical category is necessary.

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

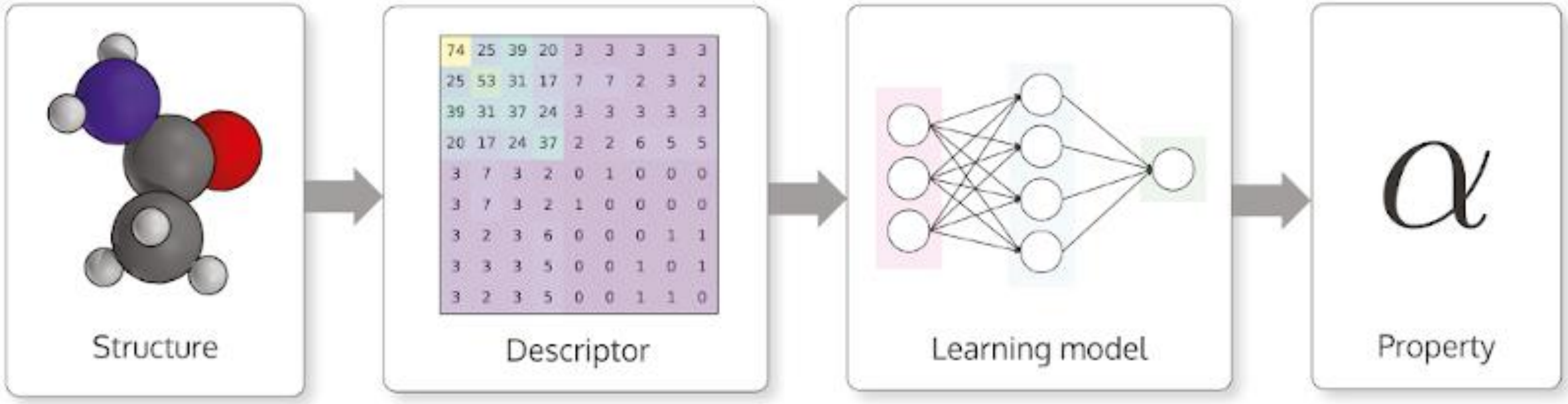
•**Analog** approach

•**Category** approach

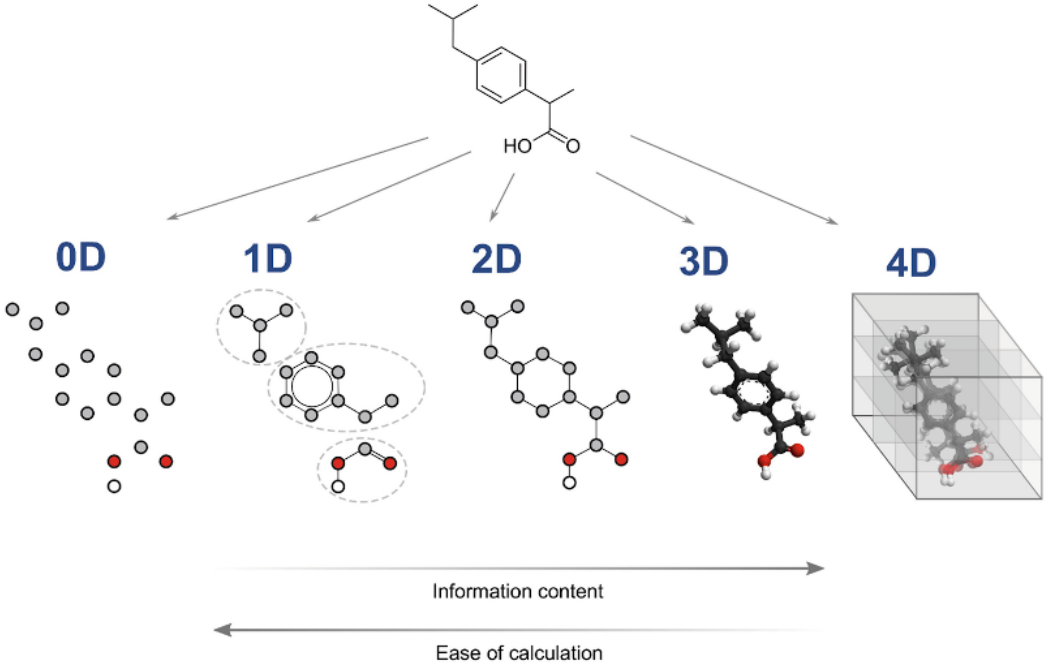


Vink, S.R. *et al.*, *Regul Toxicol Pharmacol*. **2010**, *58*, 64–71

What is Similaritv?

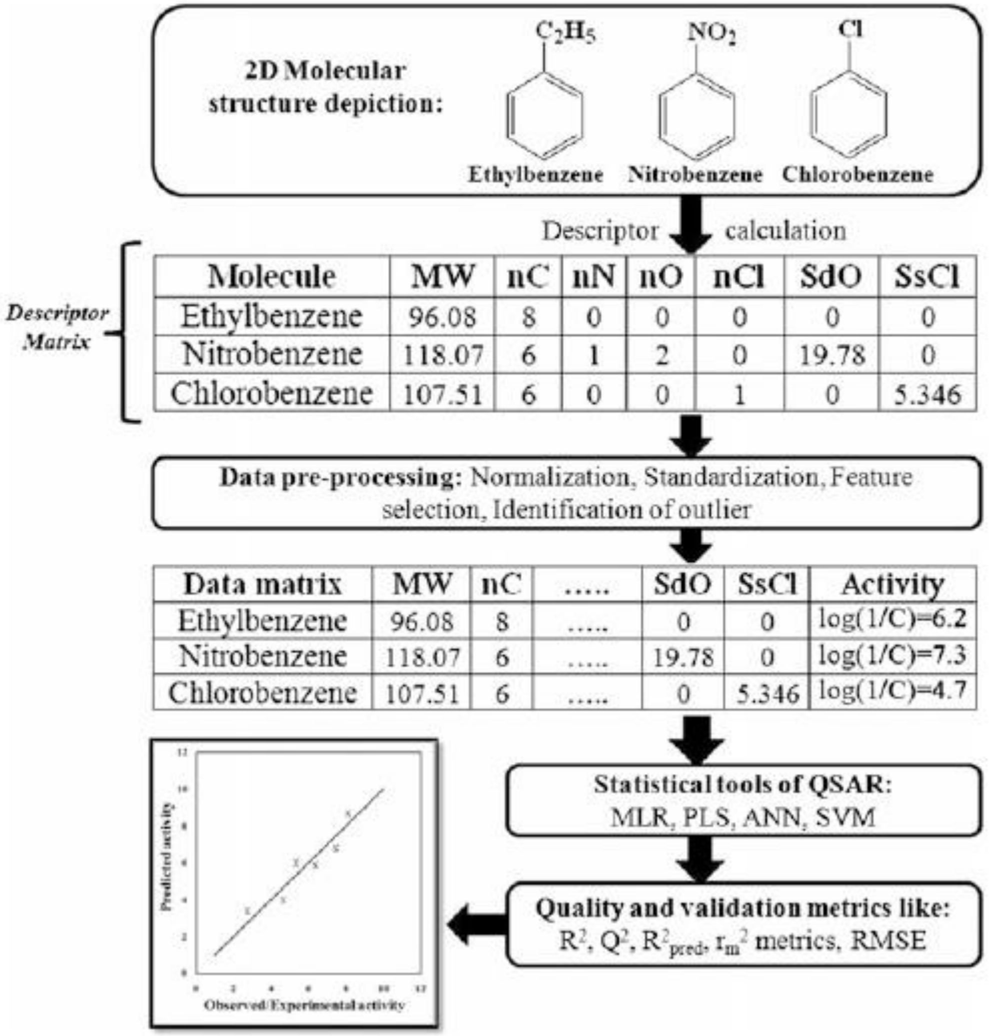


74	25	39	20	3	3	3	3	3
25	53	31	17	7	7	2	3	2
39	31	37	24	3	3	3	3	3
20	17	24	37	2	2	6	5	5
3	7	3	2	0	1	0	0	0
3	7	3	2	1	0	0	0	0
3	2	3	6	0	0	0	1	1
3	3	3	5	0	0	1	0	1
3	2	3	5	0	0	1	1	0



https://chemintelligence.com/blog/machine-learning-descriptors-molecules

What is Similarity?



2D Molecular structure depiction:

Ethylbenzene Nitrobenzene Chlorobenzene

Descriptor calculation

Descriptor Matrix

Molecule	MW	nC	nN	nO	nCl	SdO	SsCl
Ethylbenzene	96.08	8	0	0	0	0	0
Nitrobenzene	118.07	6	1	2	0	19.78	0
Chlorobenzene	107.51	6	0	0	1	0	5.346

Data pre-processing: Normalization, Standardization, Feature selection, Identification of outlier

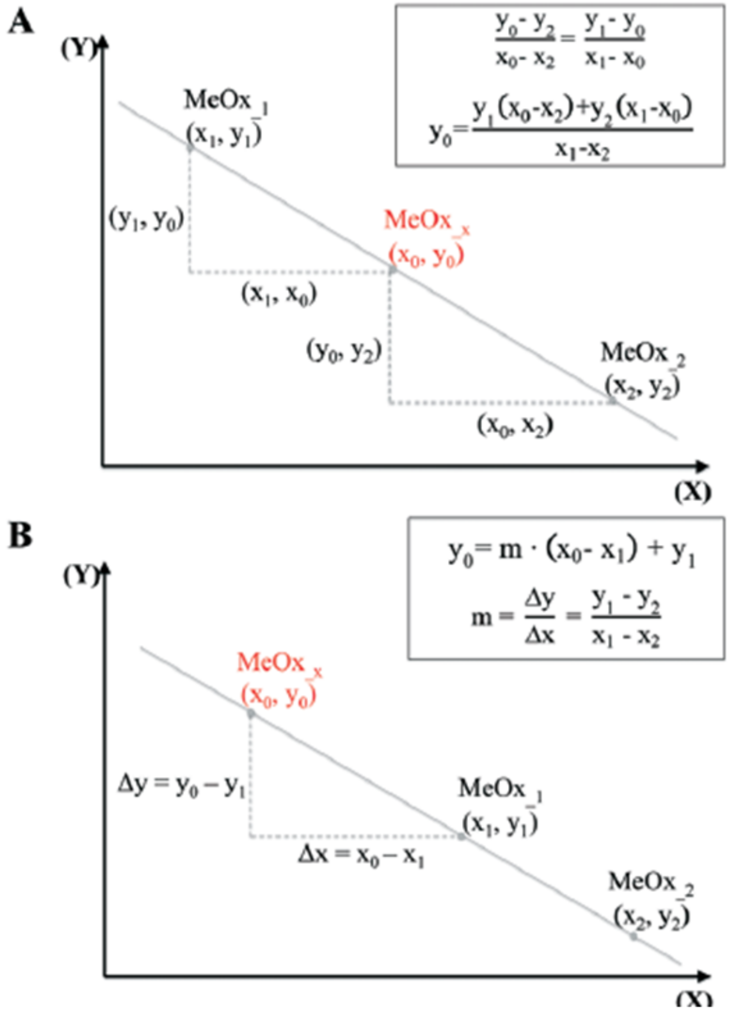
Data matrix	MW	nC	.....	SdO	SsCl	Activity
Ethylbenzene	96.08	8	.....	0	0	log(1/C)=6.2
Nitrobenzene	118.07	6	.....	19.78	0	log(1/C)=7.3
Chlorobenzene	107.51	6	.....	0	5.346	log(1/C)=4.7

Statistical tools of QSAR:
MLR, PLS, ANN, SVM

Quality and validation metrics like:
R^2, Q^2, R^2_{pred}, r_m^2 metrics, RMSE

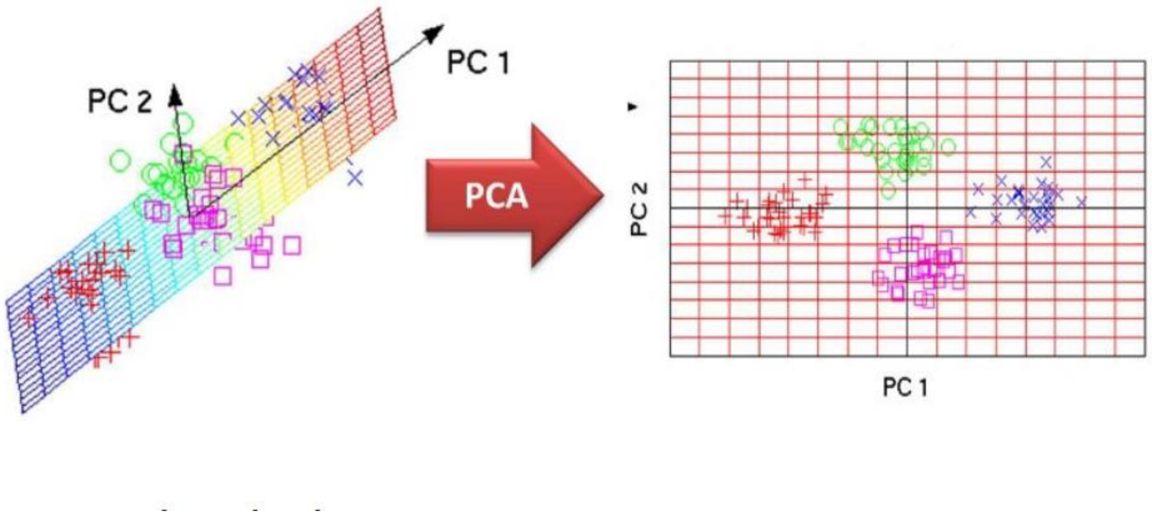
Roy, Kar and Das, A Primer on QSAR/QSPR Modeling (SpringerBrief), Springer, 2015

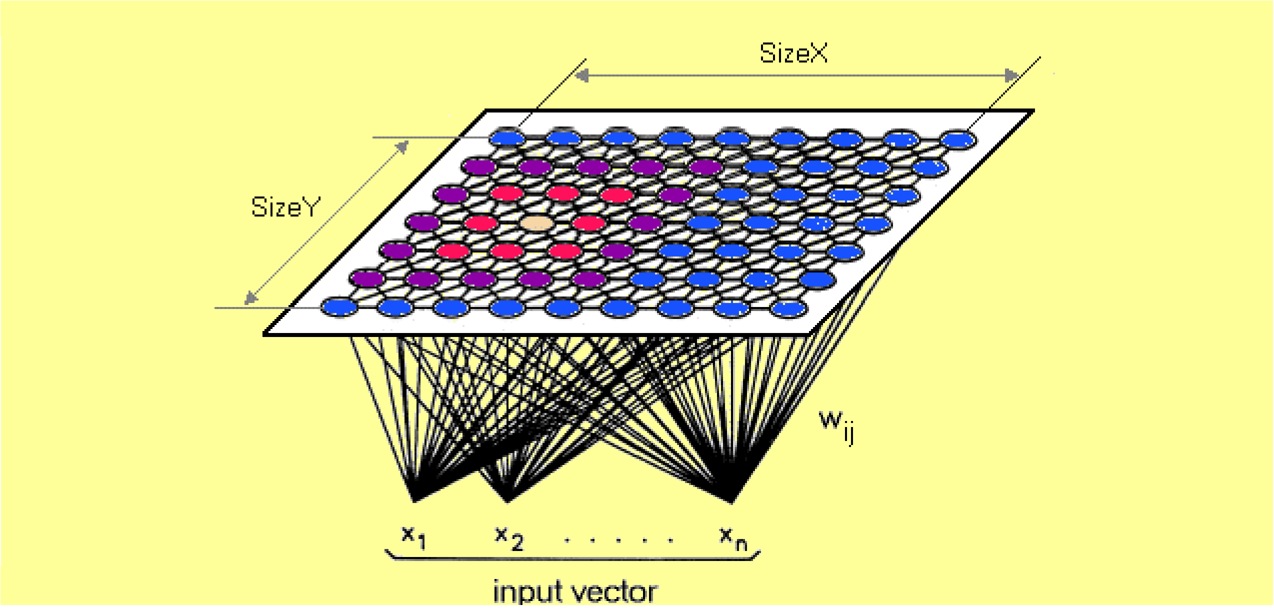
What is Similarity?



Gajewicz A, Environ. Sci.: Nano, 2017, 4, 346

What is Similarity?





https://towardsdatascience.com/feature-extraction-using-principal-component-analysis-a-simplified-visual-demo-e5592ced100a

https://towardsdatascience.com/self-organizing-maps-1b7d2a84e065

PCA

SOM

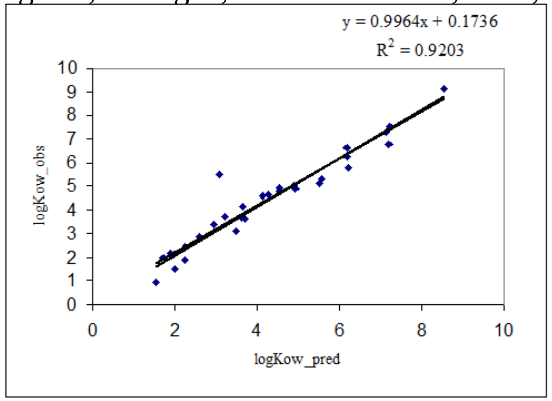
QSAR (Quantitative Structure-Activity Relationship)

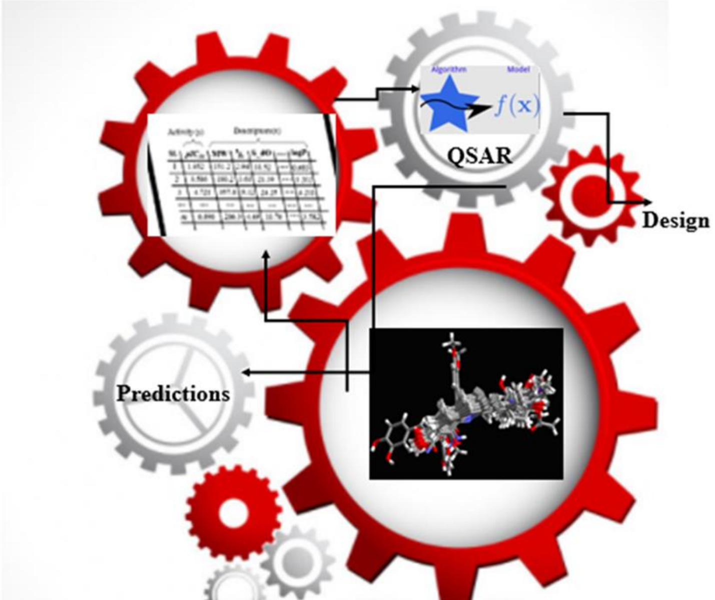
❑**QSAR deals with development of predictive models correlating <u>biological activity</u> (including therapeutic and toxic) of chemicals (drugs/toxicants/environmental pollutants) with <u>descriptors</u> representative of molecular structure and/or property by application of <u>statistical tools</u>.**

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

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \dots\dots\dots$$

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





Metrics for judging quality of QSAR models

❑ **Determination coefficient**

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{calc})^2}{\sum (Y_{obs} - \overline{Y})^2}$$

❑ **Explained variance**

$$R_a^2 = \frac{(n-1)R^2 - p - 1}{n - p - 1}$$

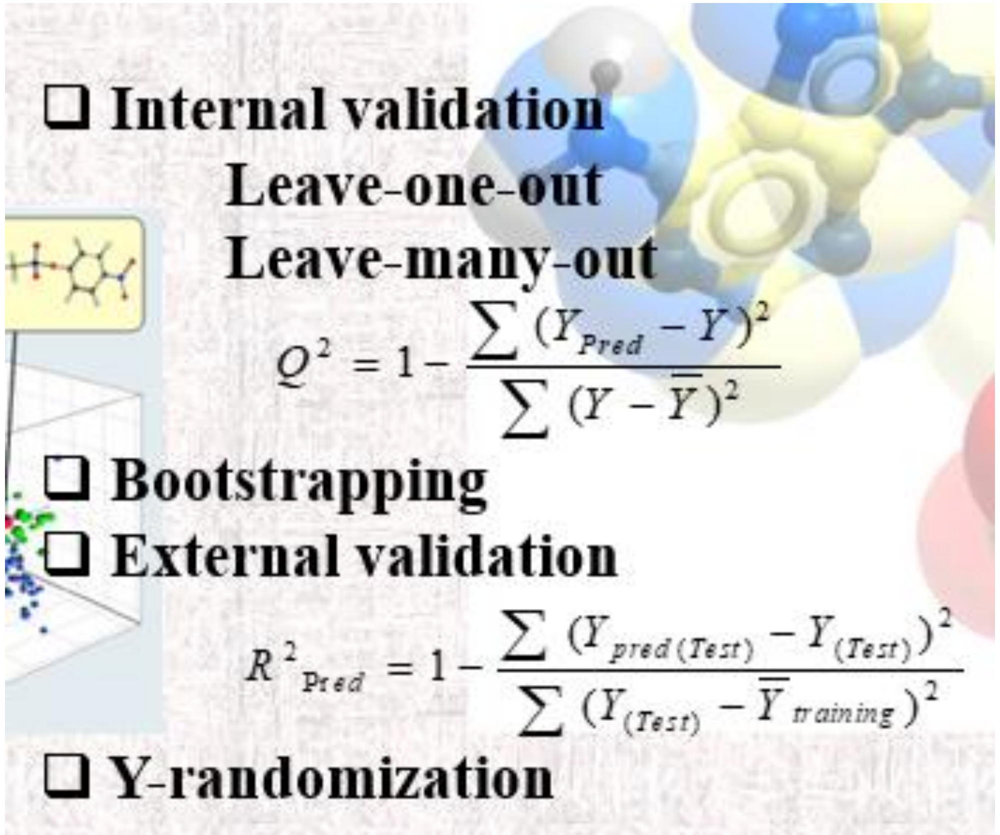
❑ **Variance ratio**

$$F = \frac{\dfrac{\sum (Y_{cal} - \overline{Y})^2}{p}}{\dfrac{\sum (Y_{obs} - Y_{cal})^2}{n - p - 1}}$$

❑ **Standard error of estimate**

$$s = \sqrt{\frac{(Y_{obs} - Y_{calc})^2}{n - p - 1}}$$

Validation of QSAR models



- ❑ Internal validation
  - Leave-one-out
  - Leave-many-out

$$Q^2 = 1 - \frac{\sum (Y_{Pred} - Y)^2}{\sum (Y - \overline{Y})^2}$$

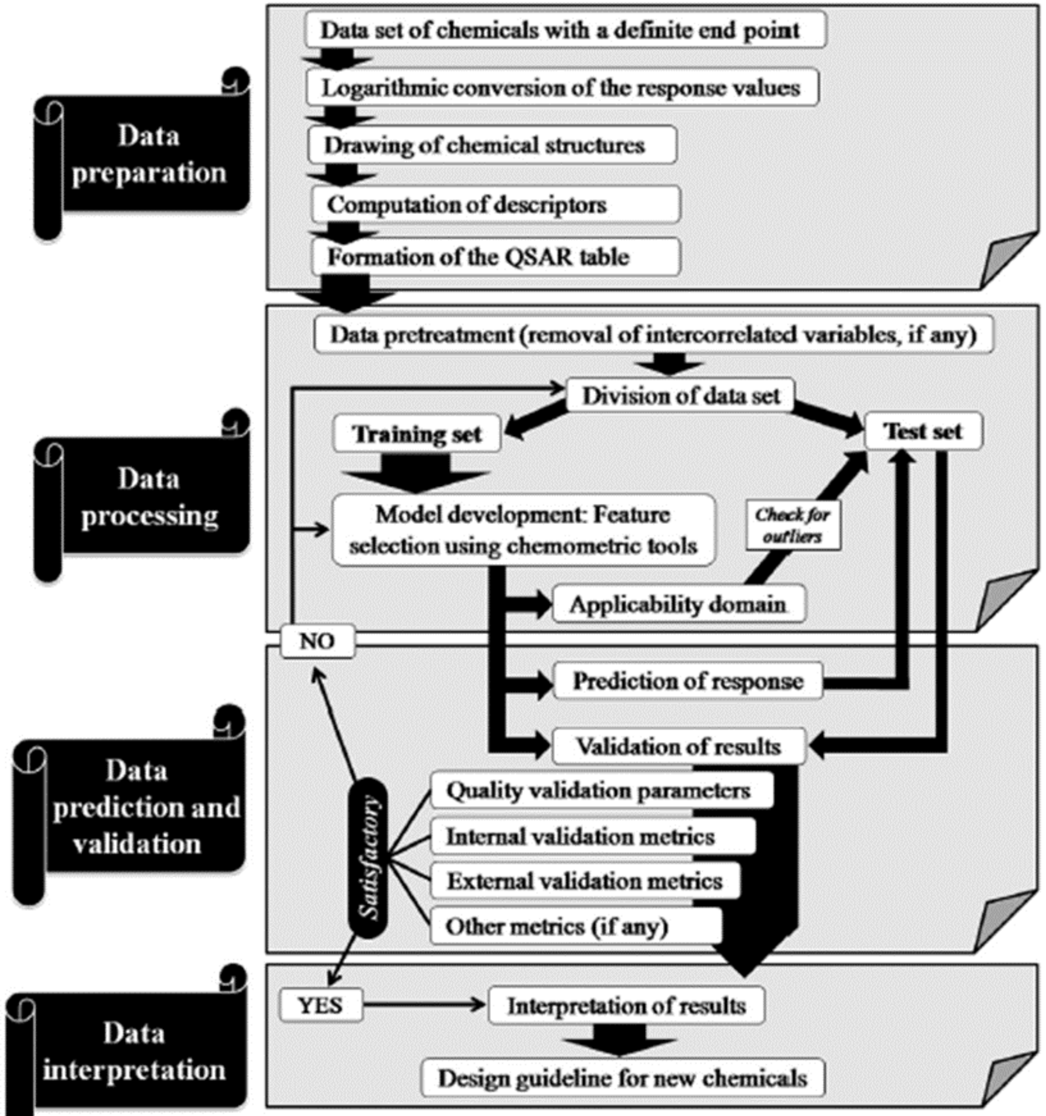
- ❑ Bootstrapping
- ❑ External validation

$$R^2_{Pred} = 1 - \frac{\sum (Y_{pred(Test)} - Y_{(Test)})^2}{\sum (Y_{(Test)} - \overline{Y}_{training})^2}$$

- ❑ Y-randomization

Algorithm
Model
f(x)
QSAR
Design
Predictions

Steps in QSAR model development



Data preparation

Data set of chemicals with a definite end point

Logarithmic conversion of the response values

Drawing of chemical structures

Computation of descriptors

Formation of the QSAR table

Data processing

Data pretreatment (removal of intercorrelated variables, if any)

Division of data set

Training set

Test set

Model development: Feature selection using chemometric tools

Check for outliers

Applicability domain

NO

Data prediction and validation

Prediction of response

Validation of results

Satisfactory

Quality validation parameters

Internal validation metrics

External validation metrics

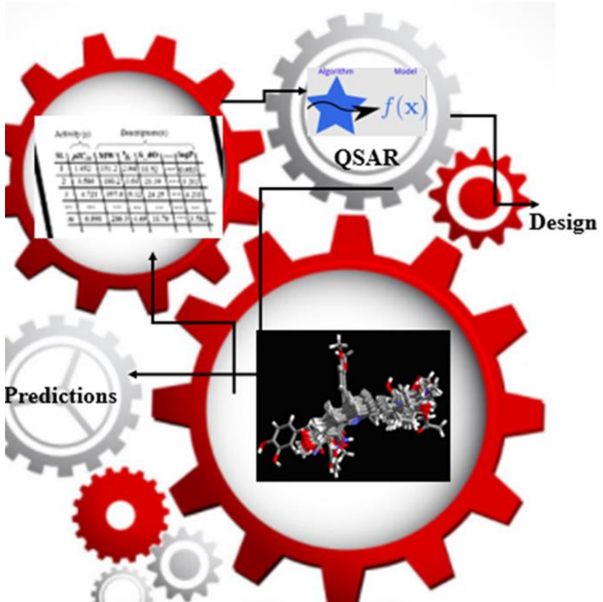
Other metrics (if any)

YES

Data interpretation

Interpretation of results

Design guideline for new chemicals



Why Read-across instead of QSAR?

- QSAR is not suitable for small datasets
- Read-across is not a statistical fitting process
- Calculation is comparatively easier than QSAR
- Alternative tool for hazard assessment, aimed at filling data gaps
- For nano-toxicity, the data sets are usually small; thus, application of quantitative read-across is more suitable than statistical fitting approaches

Chemical read-across predictions of Nanotoxicity data

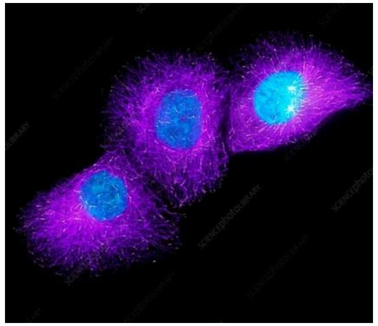
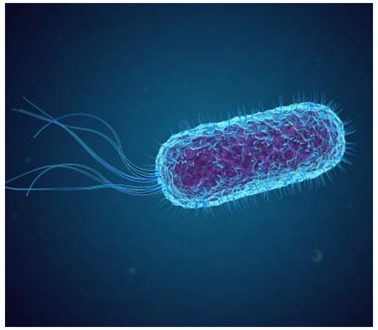
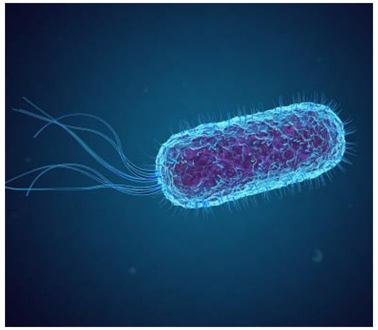
- To develop an easier and efficient method for quantitative read-across predictions

- Quantitative toxicity prediction of various small datasets (specifically, toxicity of metal oxide nano-particles) using a new method

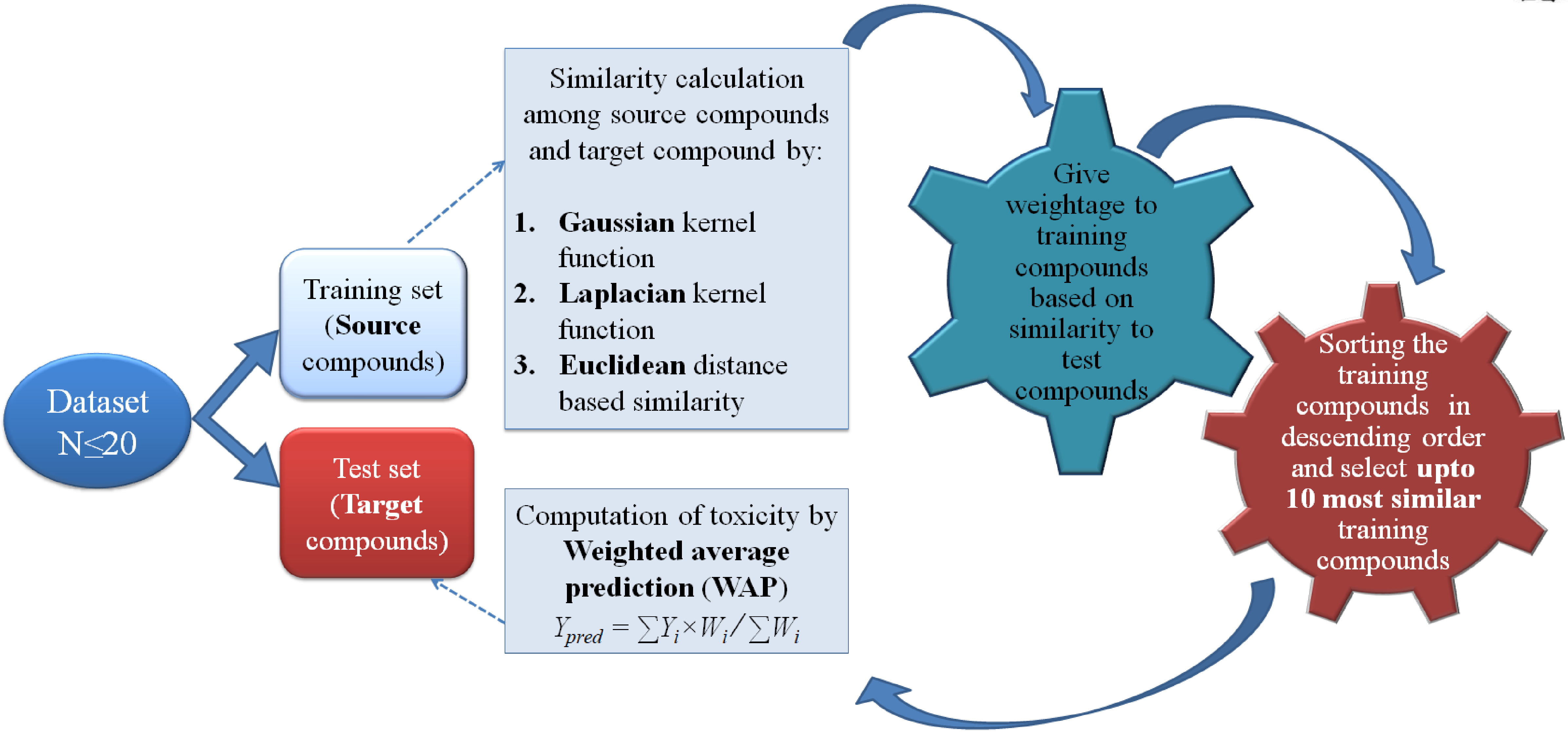
- Comparison of the results with the previous methods

- Development of an application for Read-across predictions.

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

Dataset	Endpoint	Descriptors	Data points
Dataset 1 *Environ. Sci.: Nano*, **2017**, *4*, 1389	**pLC**$_{50}$ of metal NPs against a **human ketatinocyte (HaCaT)** cell line. 	Mulliken Electro negativity of the cluster (χ^{c}), and the enthalpy of formation of a metal oxide nano-cluster representing a fragment of the surface (ΔH^{c}_{f}).	**18**
Dataset 2 *Environ. Sci.: Nano*, **2017**, *4*, 1389	**pEC**$_{50}$ of metal NPs against bacteria ***Escherichia coli***. 	The enthalpy of formation of gaseous cations having the same oxidation state as those in the metal oxide structure (ΔH_{Me+}), and the charge of the metal cation corresponding to a given oxide (**Me+**).	**17**
Dataset 3 *Environ. Sci.: Nano*, **2017**, *4*, 1389	**pLC**$_{50}$ of metal NPs against bacteria ***Escherichia coli*** under **dark condition**. 	Enthalpy of formation of gaseous cations having the same oxidation state as those in the metal oxide structure (ΔH_{Me+}), and the absolute electro negativity of the metal oxide (**LZELEHHO**).	**16**

Schematic representation of the proposed methodology



Dataset N≤20

Training set (**Source** compounds)

Test set (**Target** compounds)

Similarity calculation among source compounds and target compound by:

1. **Gaussian** kernel function
2. **Laplacian** kernel function
3. **Euclidean** distance based similarity

Give weightage to training compounds based on similarity to test compounds

Sorting the training compounds in descending order and select **upto 10 most similar** training compounds

Computation of toxicity by **Weighted average prediction (WAP)**

$$Y_{pred} = \sum Y_i \times W_i / \sum W_i$$

Gaussian Kernel function similarity estimation

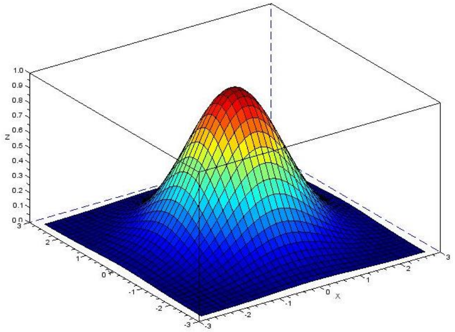
The function Gaussian kernel is a variant on the radial basis function kernel defined as:

$$\boldsymbol{f = \exp((-\|X-Y\|\hat{}2)/2\sigma^2)}$$

Where X, Y are the input vectors and $\|X-Y\|$ is the Euclidean distance between two vectors.

Say X and Y are two vectors each of length n

$X=\|X_1, X_2, X_3 ,\ldots, X_n\|;\ Y=\|Y_1, Y_2, Y_3, \ldots, Y_n\|$



d (X, Y) = $\|X-Y\| = \text{sqrt}((X_1-Y_1)^2+(X_2-Y_2)^2+ \ldots+(X_n\text{-}Y_n)^2)$

σ is a variable number. We have predicted the toxicity using different values of σ (**0.25**, **0.5**, **0.75**, **1.0**, **1.5**, **2.0**)

Laplacian Kernel function similarity estimation

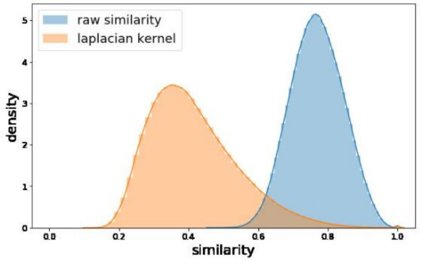
The function Laplacian kernel is a variant on the radial basis function kernel defined as:

$$\boldsymbol{\kappa\ (X, Y) = \mathrm{exp}((-\Upsilon \| X - Y \|_1)}$$

Where X, Y are the input vectors and $\|X-Y\|_1$ is the Manhattan distance between two vectors.

Say X and Y are two vectors each of length n

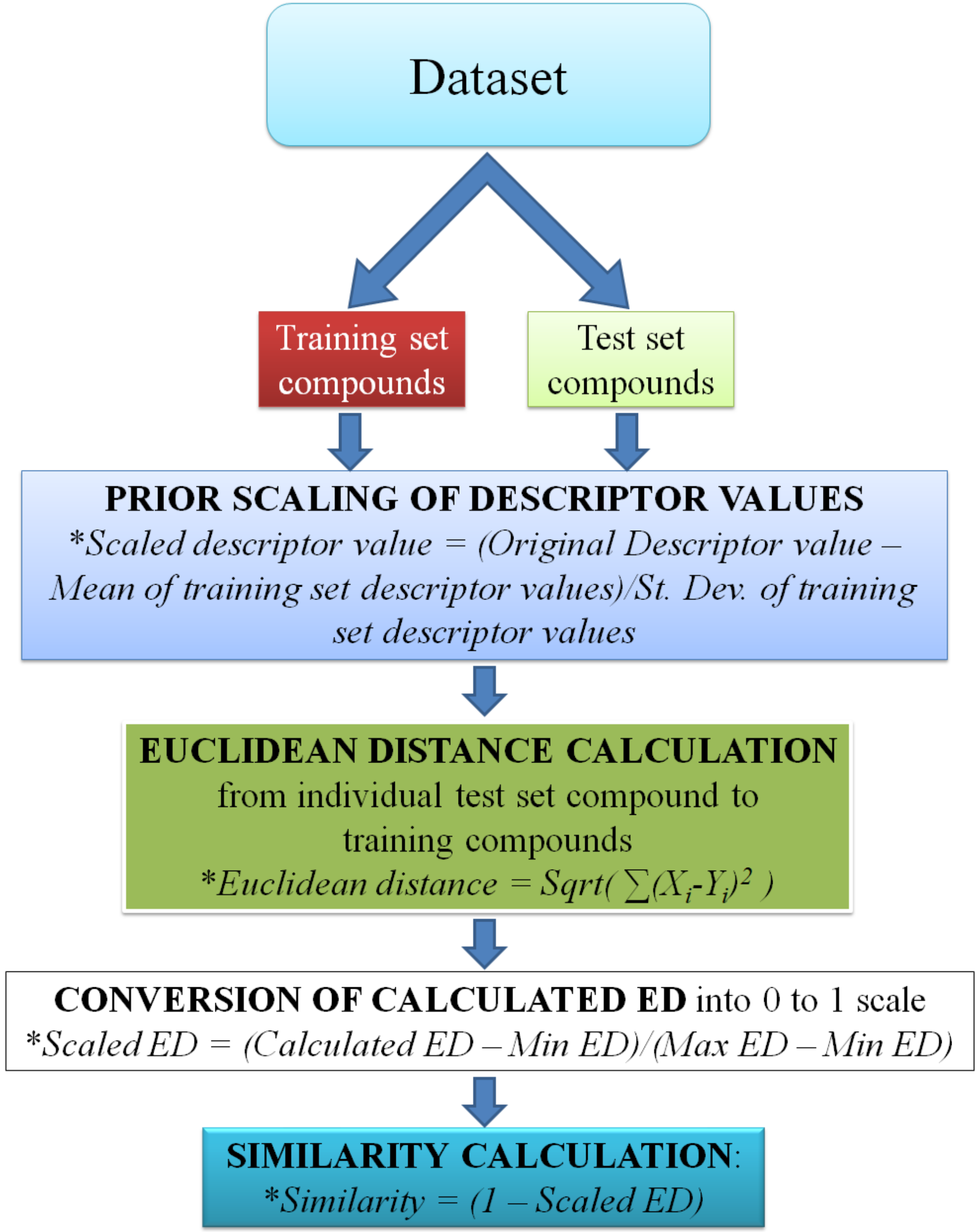
$X=\|X_1, X_2, X_3, \ldots, X_n\|$ $Y=\|Y_1, Y_2, Y_3, \ldots, Y_n\|$



$$\mathrm{d1}\ (X, Y) = \|X-Y\|_1 = (X_1 - Y_1) + (X_2 - Y_2) + (X_3 - Y_3) + \ldots. + (X_n - Y_n)$$

ϒ is a variable number. We have predicted the toxicity using different values of ϒ (**0.25**, **0.5**, **0.75**, **1.0**, **1.5**, **2.0**)

Euclidean distance based similarity estimation



Dataset

Training set compounds

Test set compounds

PRIOR SCALING OF DESCRIPTOR VALUES
**Scaled descriptor value = (Original Descriptor value – Mean of training set descriptor values)/St. Dev. of training set descriptor values*

EUCLIDEAN DISTANCE CALCULATION
from individual test set compound to training compounds
**Euclidean distance =* $Sqrt(\sum(X_i\text{-}Y_i)^2)$

CONVERSION OF CALCULATED ED into 0 to 1 scale
**Scaled ED = (Calculated ED – Min ED)/(Max ED – Min ED)*

SIMILARITY CALCULATION:
**Similarity = (1 – Scaled ED)*

Validation metrics

Quantitative validation metrics	
$\mathbf{Q^2_{F1}}$	$Q^2_{F1} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \overline{Y_{training}})^2}$
$\mathbf{Q^2_{F2}}$	$Q^2_{F2} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \overline{Y_{test}})^2}$
Root mean square error of prediction ($\mathbf{RMSE_p}$)	$RMSE_p = \sqrt{\frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{n_{test}}}$

Quantitative terms- $\boldsymbol{Y_{obs(test)}}$: Observed activity of test set compounds; $\boldsymbol{Y_{pred(test)}}$: Predicted activity of test set compounds; $\boldsymbol{\overline{Y_{training}}}$: Average observed activity of training set compounds; $\boldsymbol{\overline{Y_{test}}}$: Average observed activity of test set compounds; $\boldsymbol{n_{test}}$= number of compounds in the test set.

Classification-based metrics	
Sensitivity (%)	$Sensitivity = \frac{TP}{TP + FN}$
Specificity (%)	$Specificity = \frac{TN}{TN + FP}$
Precision (%)	$Precision = \frac{TP}{TP + FP}$
Accuracy (%)	$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$
F-measure (%) (harmonic mean of recall)	$F - measure(\%) = \frac{2}{\frac{1}{Precision} + \frac{1}{Sensitivity}}$
G-means (geometric mean)	$G - means = \sqrt{SpecificityXSensitivity}$
Cohen's kappa (K)	$P_r(a) = \frac{(TP + TN)}{(TP + FP + TN + FN)}$ $P_r(e) = \frac{\{(TP + FP)X(TP + FN)\} + \{(TN + FP)X(TN + FN)\}}{(TP + FN + FP + TN)^2}$ $Cohen's\ K = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$
Matthews correlation coefficient (MCC)	$MCC = \frac{(TPXTN) - (FPXFN)}{\sqrt{(TP + FP)X(TP + FN)X(TN + FP)X\ (TN + FN)}}$

Classification-based terms- $\boldsymbol{TP}$**:** True positive; $\boldsymbol{TN}$: True negative; $\boldsymbol{FP}$: False positive; $\boldsymbol{FN}$: False negative; $\boldsymbol{P_r(a)}$: relative observed agreement between the predicted classification of the model and the known classification; $\boldsymbol{P_r(e)}$:hypothetical probability ofchance agreement.

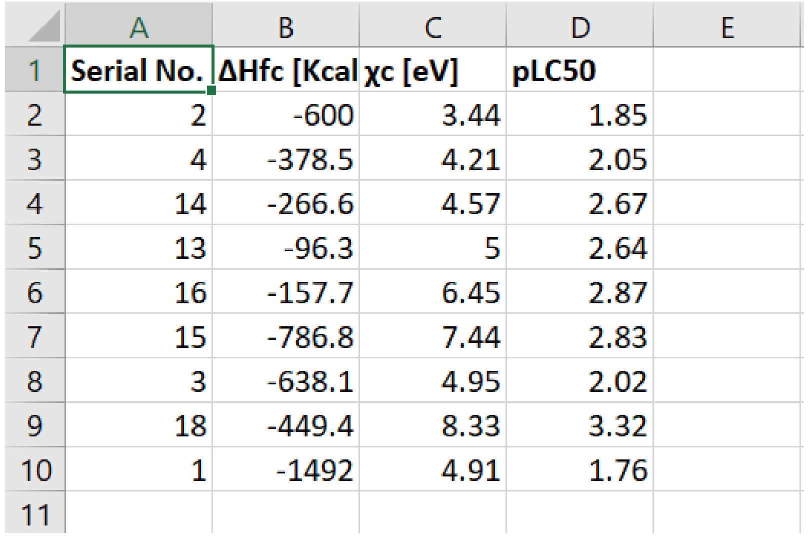
Software Development – A Java based application for quantitative read across

Software: Quantitative Read Across for Nanotoxicity Prediction available at https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home

- A java based application has been developed.
- It needs training set and test set data as input in ***.xlsx** format.
- User has to provide **σ value**, **ϒ value**, **number of similar training compounds**, **distance threshold**, and **similarity threshold** as input information.
- The program generates two output files namely Biological activity (predicted response), and sorted experimental response with respect to distance and similarity.

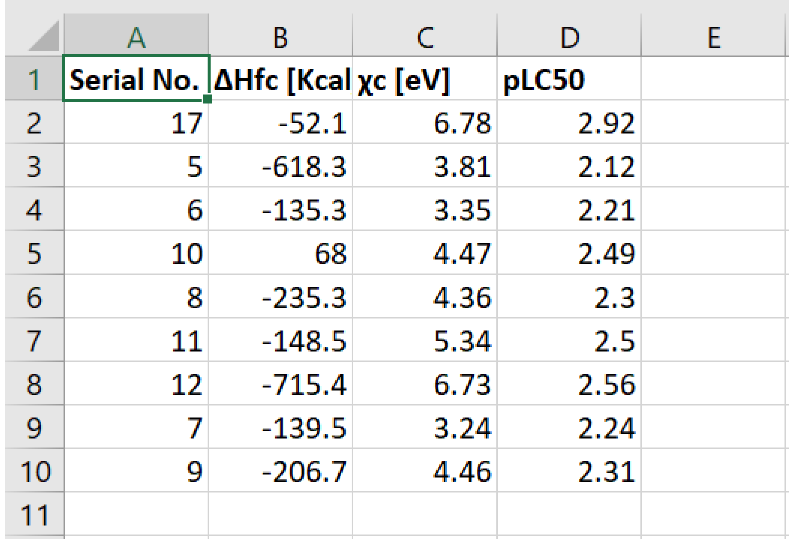
Input files:

Train.xlsx

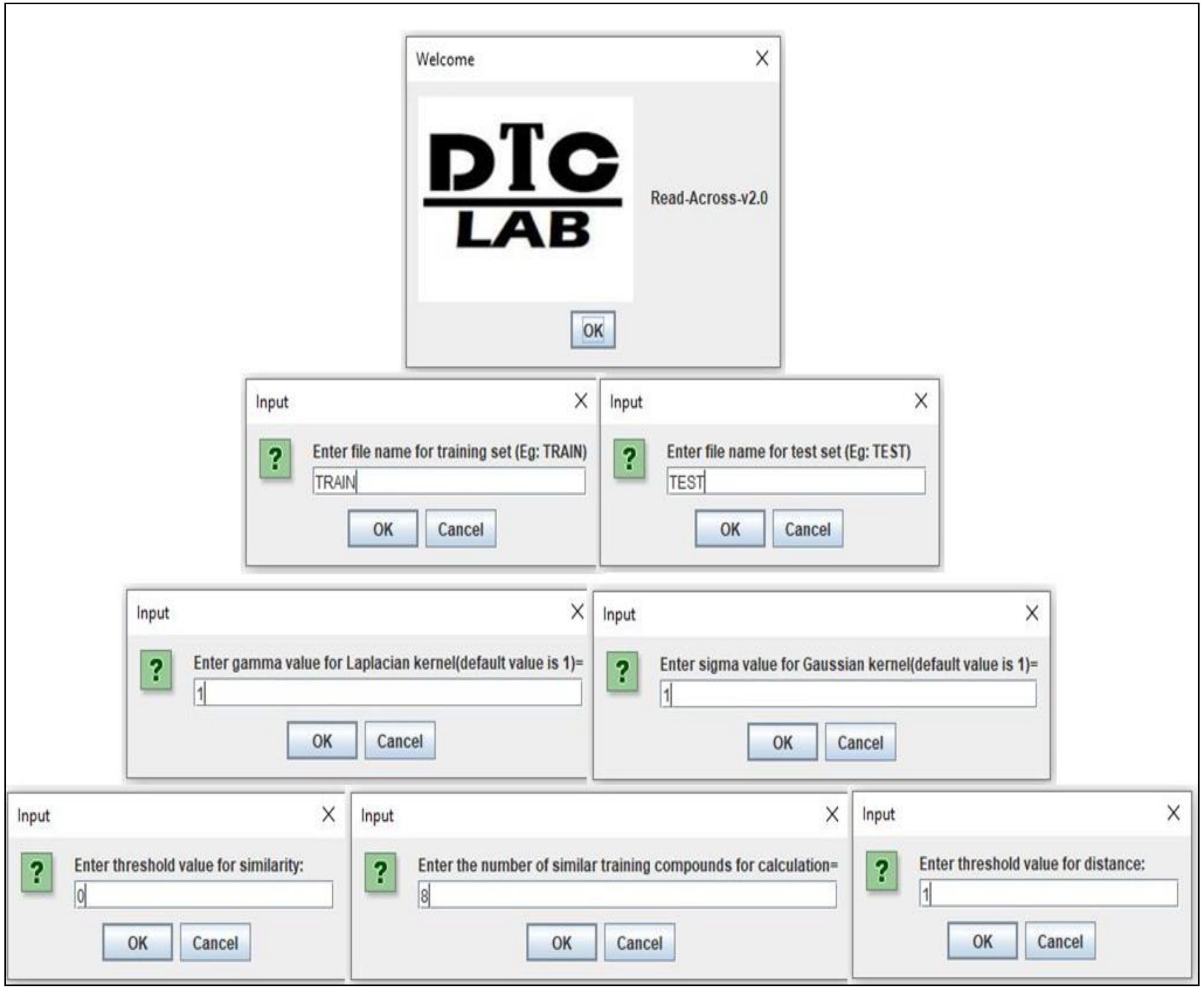


	A	B	C	D	E
1	**Serial No.**	**ΔHfc [Kcal**	**χc [eV]**	**pLC50**	
2	2	-600	3.44	1.85	
3	4	-378.5	4.21	2.05	
4	14	-266.6	4.57	2.67	
5	13	-96.3	5	2.64	
6	16	-157.7	6.45	2.87	
7	15	-786.8	7.44	2.83	
8	3	-638.1	4.95	2.02	
9	18	-449.4	8.33	3.32	
10	1	-1492	4.91	1.76	
11					

Test.xlsx



	A	B	C	D	E
1	**Serial No.**	**ΔHfc [Kcal**	**χc [eV]**	**pLC50**	
2	17	-52.1	6.78	2.92	
3	5	-618.3	3.81	2.12	
4	6	-135.3	3.35	2.21	
5	10	68	4.47	2.49	
6	8	-235.3	4.36	2.3	
7	11	-148.5	5.34	2.5	
8	12	-715.4	6.73	2.56	
9	7	-139.5	3.24	2.24	
10	9	-206.7	4.46	2.31	
11					



Read-Across-v2.0

Enter file name for training set (Eg: TRAIN)

TRAIN

Enter file name for test set (Eg: TEST)

TEST

Enter gamma value for Laplacian kernel(default value is 1)=

1

Enter sigma value for Gaussian kernel(default value is 1)=

1

Enter threshold value for similarity:

0

Enter the number of similar training compounds for calculation=

8

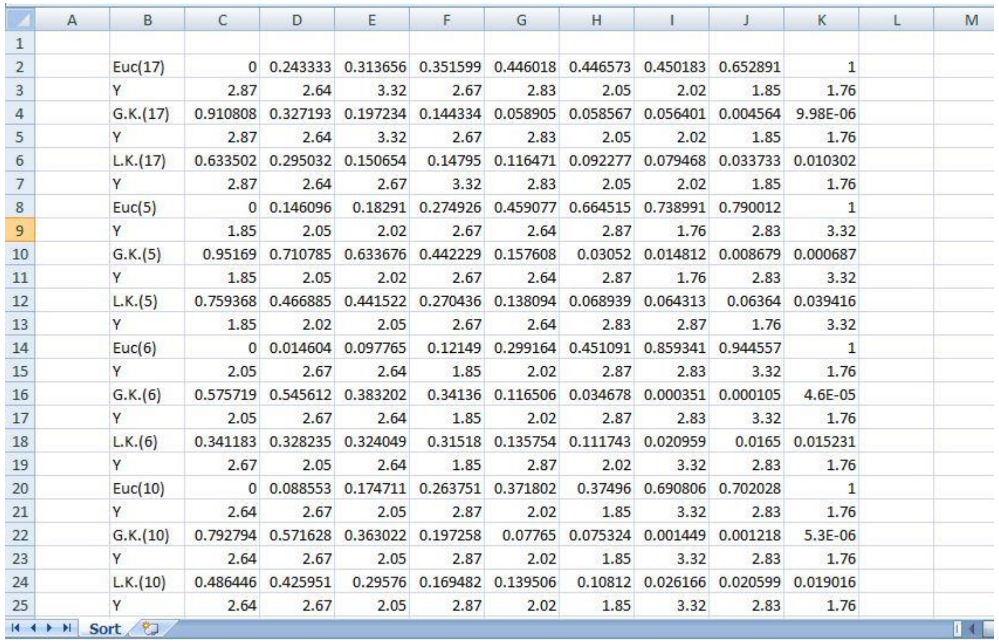
Enter threshold value for distance:

1

Snapshot of the developed program “Read-Across-v2.0”.

Program Output

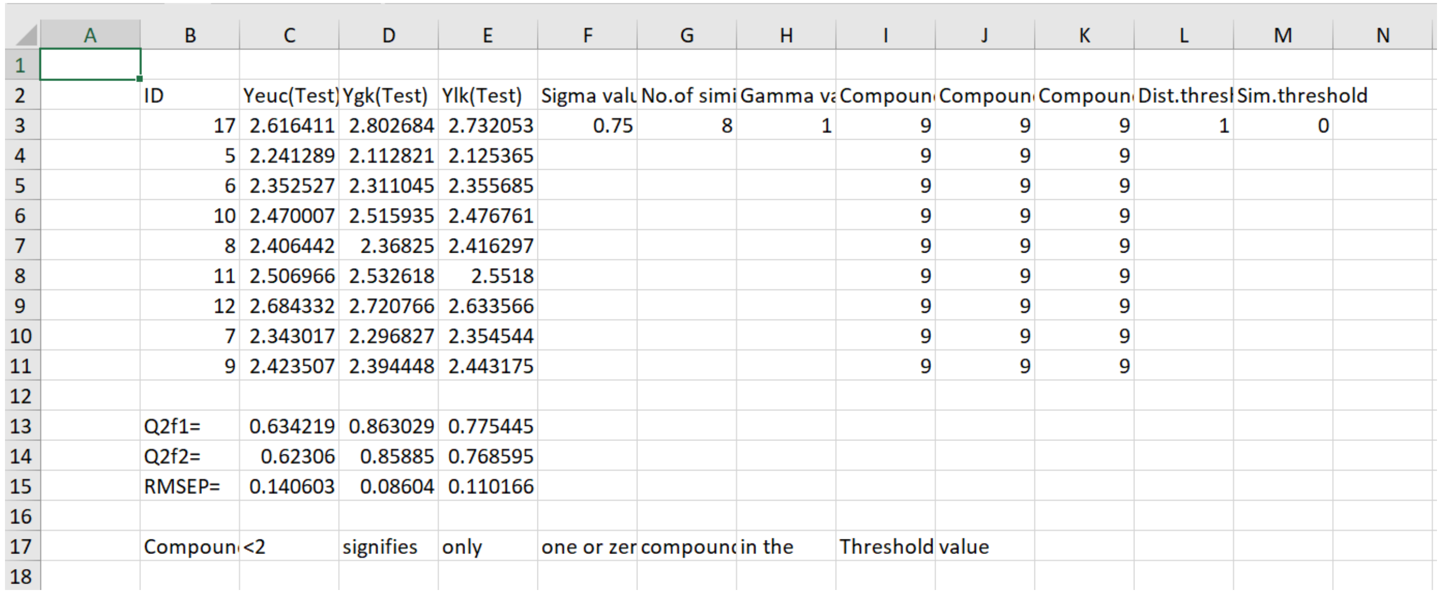
Sort.xlsx



	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2		Euc(17)	0	0.243333	0.313656	0.351599	0.446018	0.446573	0.450183	0.652891	1		
3		Y	2.87	2.64	3.32	2.67	2.83	2.05	2.02	1.85	1.76		
4		G.K.(17)	0.910808	0.327193	0.197234	0.144334	0.058905	0.058567	0.056401	0.004564	9.98E-06		
5		Y	2.87	2.64	3.32	2.67	2.83	2.05	2.02	1.85	1.76		
6		L.K.(17)	0.633502	0.295032	0.150654	0.14795	0.116471	0.092277	0.079468	0.033733	0.010302		
7		Y	2.87	2.64	2.67	3.32	2.83	2.05	2.02	1.85	1.76		
8		Euc(5)	0	0.146096	0.18291	0.274926	0.459077	0.664515	0.738991	0.790012	1		
9		Y	1.85	2.05	2.02	2.67	2.64	2.87	1.76	2.83	3.32		
10		G.K.(5)	0.95169	0.710785	0.633676	0.442229	0.157608	0.03052	0.014812	0.008679	0.000687		
11		Y	1.85	2.05	2.02	2.67	2.64	2.87	1.76	2.83	3.32		
12		L.K.(5)	0.759368	0.466885	0.441522	0.270436	0.138094	0.068939	0.064313	0.06364	0.039416		
13		Y	1.85	2.02	2.05	2.67	2.64	2.83	2.87	1.76	3.32		
14		Euc(6)	0	0.014604	0.097765	0.12149	0.299164	0.451091	0.859341	0.944557	1		
15		Y	2.05	2.67	2.64	1.85	2.02	2.87	2.83	3.32	1.76		
16		G.K.(6)	0.575719	0.545612	0.383202	0.34136	0.116506	0.034678	0.000351	0.000105	4.6E-05		
17		Y	2.05	2.67	2.64	1.85	2.02	2.87	2.83	3.32	1.76		
18		L.K.(6)	0.341183	0.328235	0.324049	0.31518	0.135754	0.111743	0.020959	0.0165	0.015231		
19		Y	2.67	2.05	2.64	1.85	2.87	2.02	3.32	2.83	1.76		
20		Euc(10)	0	0.088553	0.174711	0.263751	0.371802	0.37496	0.690806	0.702028	1		
21		Y	2.64	2.67	2.05	2.87	2.02	1.85	3.32	2.83	1.76		
22		G.K.(10)	0.792794	0.571628	0.363022	0.197258	0.07765	0.075324	0.001449	0.001218	5.3E-06		
23		Y	2.64	2.67	2.05	2.87	2.02	1.85	3.32	2.83	1.76		
24		L.K.(10)	0.486446	0.425951	0.29576	0.169482	0.139506	0.10812	0.026166	0.020599	0.019016		
25		Y	2.64	2.67	2.05	2.87	2.02	1.85	3.32	2.83	1.76		

Sort

Biological Activity.xlsx



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2		ID	Yeuc(Test)	Ygk(Test)	Ylk(Test)	Sigma valu	No.of simi	Gamma va	Compoun	Compoun	Compoun	Dist.thresl	Sim.threshold	
3		17	2.616411	2.802684	2.732053	0.75	8	1	9	9	9	1	0	
4		5	2.241289	2.112821	2.125365				9	9	9			
5		6	2.352527	2.311045	2.355685				9	9	9			
6		10	2.470007	2.515935	2.476761				9	9	9			
7		8	2.406442	2.36825	2.416297				9	9	9			
8		11	2.506966	2.532618	2.5518				9	9	9			
9		12	2.684332	2.720766	2.633566				9	9	9			
10		7	2.343017	2.296827	2.354544				9	9	9			
11		9	2.423507	2.394448	2.443175				9	9	9			
12														
13		Q2f1=	0.634219	0.863029	0.775445									
14		Q2f2=	0.62306	0.85885	0.768595									
15		RMSEP=	0.140603	0.08604	0.110166									
16														
17		Compoun	<2	signifies	only	one or zer	compound	in the	Threshold value					
18														

Results and Discussion

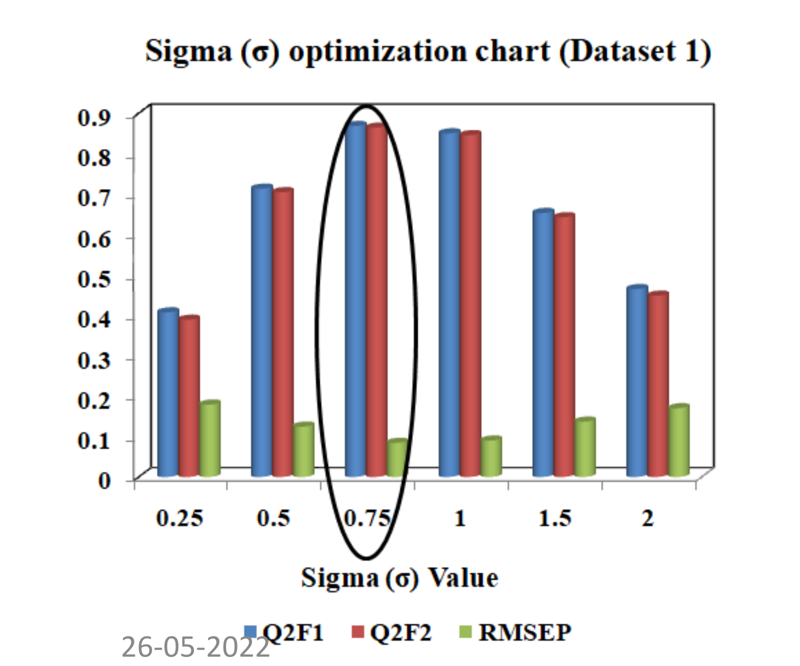
Toxicity prediction by Euclidean distance-based similarity estimation

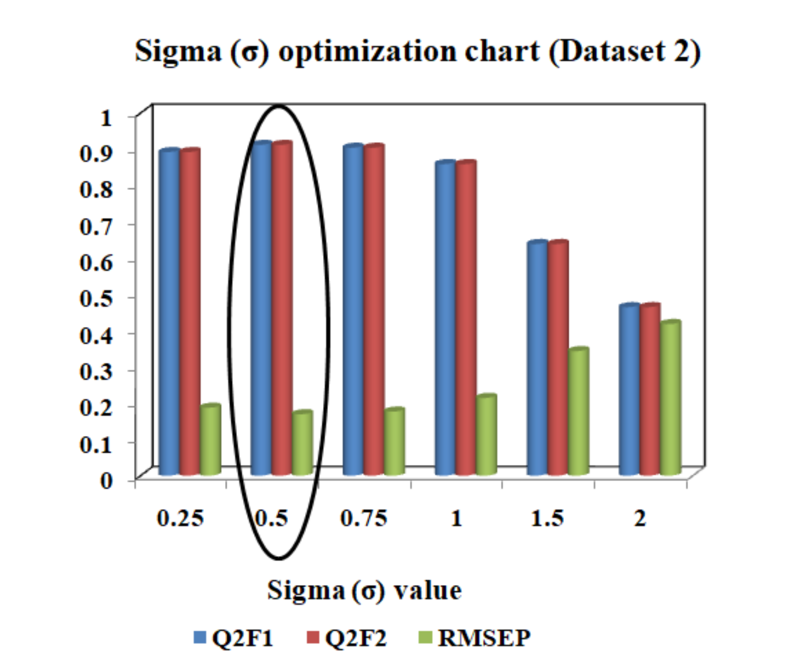
Dataset	No. of compounds in training set	Q^2_{F1}	Q^2_{F2}	$RMSE_p$
Dataset 1	9	0.63	0.62	0.14
Dataset 2	8	0.45	0.45	0.42
Dataset 3	8	0.77	0.69	0.60

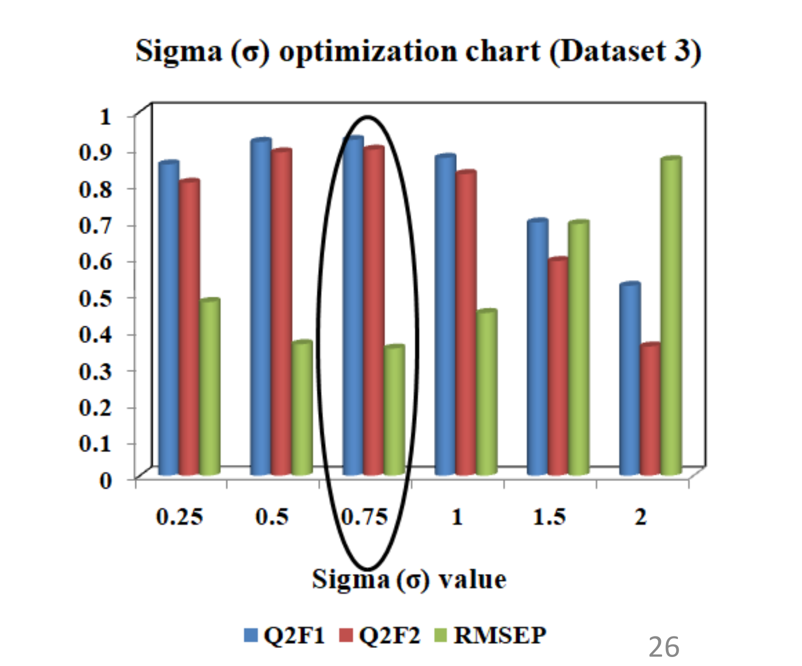
Toxicity prediction by Gaussian kernel function similarity estimation

Sigma (σ) optimisation

GAUSSIAN KERNEL	Dataset 1			Dataset 2			Dataset 3		
Sigma value	Q2F1	Q2F2	RMSEP	Q2F1	Q2F2	RMSEP	Q2F1	Q2F2	RMSEP
σ = 0.25	0.41	0.39	0.18	0.89	0.89	0.19	0.85	0.80	0.48
σ = 0.50	0.71	0.70	0.12	**0.91**	**0.91**	**0.17**	0.92	0.89	0.36
σ = 0.75	**0.87**	**0.86**	**0.08**	0.90	0.90	0.18	**0.92**	**0.90**	**0.35**
σ = 1.00	0.85	0.85	0.09	0.86	0.86	0.21	0.87	0.83	0.45
σ = 1.50	0.65	0.64	0.14	0.64	0.64	0.34	0.70	0.59	0.69
σ = 2.00	0.46	0.45	0.17	0.46	0.46	0.42	0.52	0.35	0.87



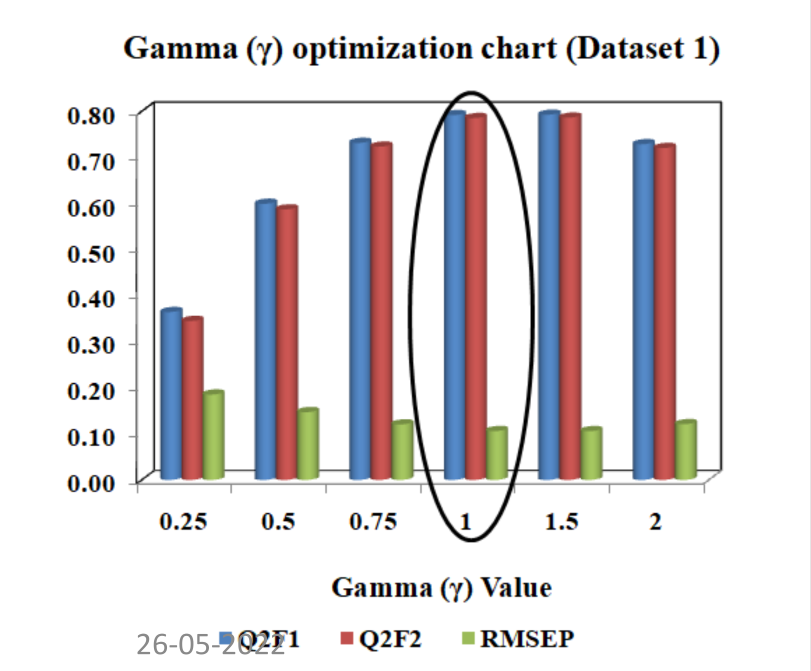


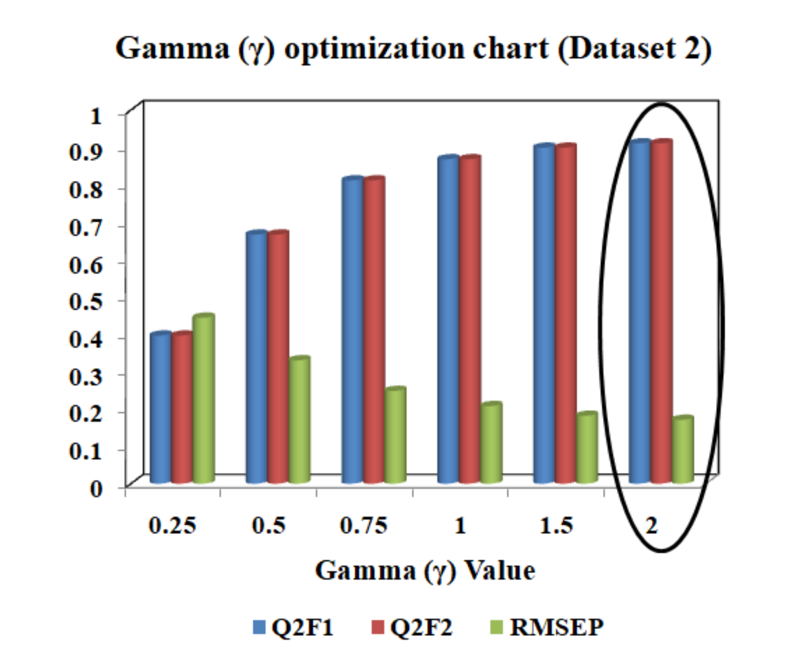


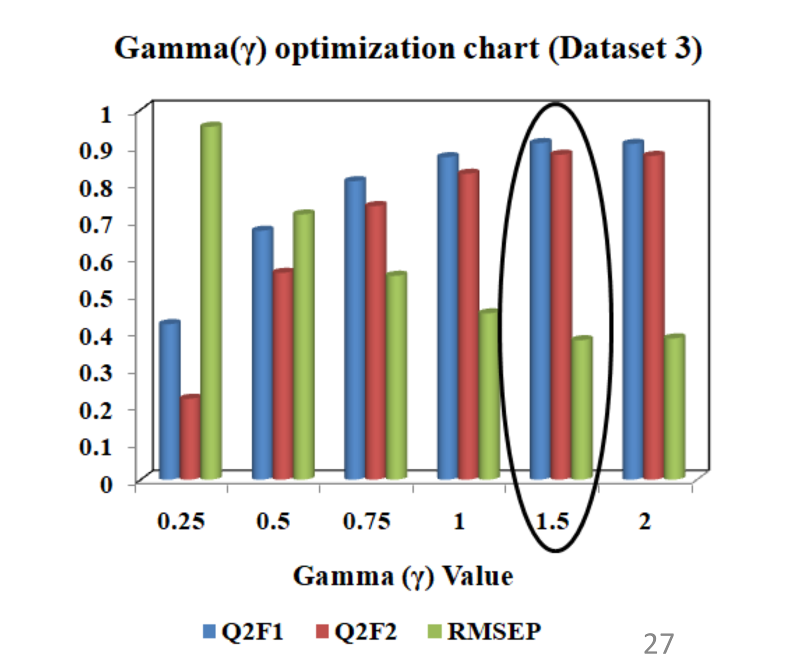
Toxicity prediction by Laplacian kernel similarity estimation

Gamma (γ) optimisation

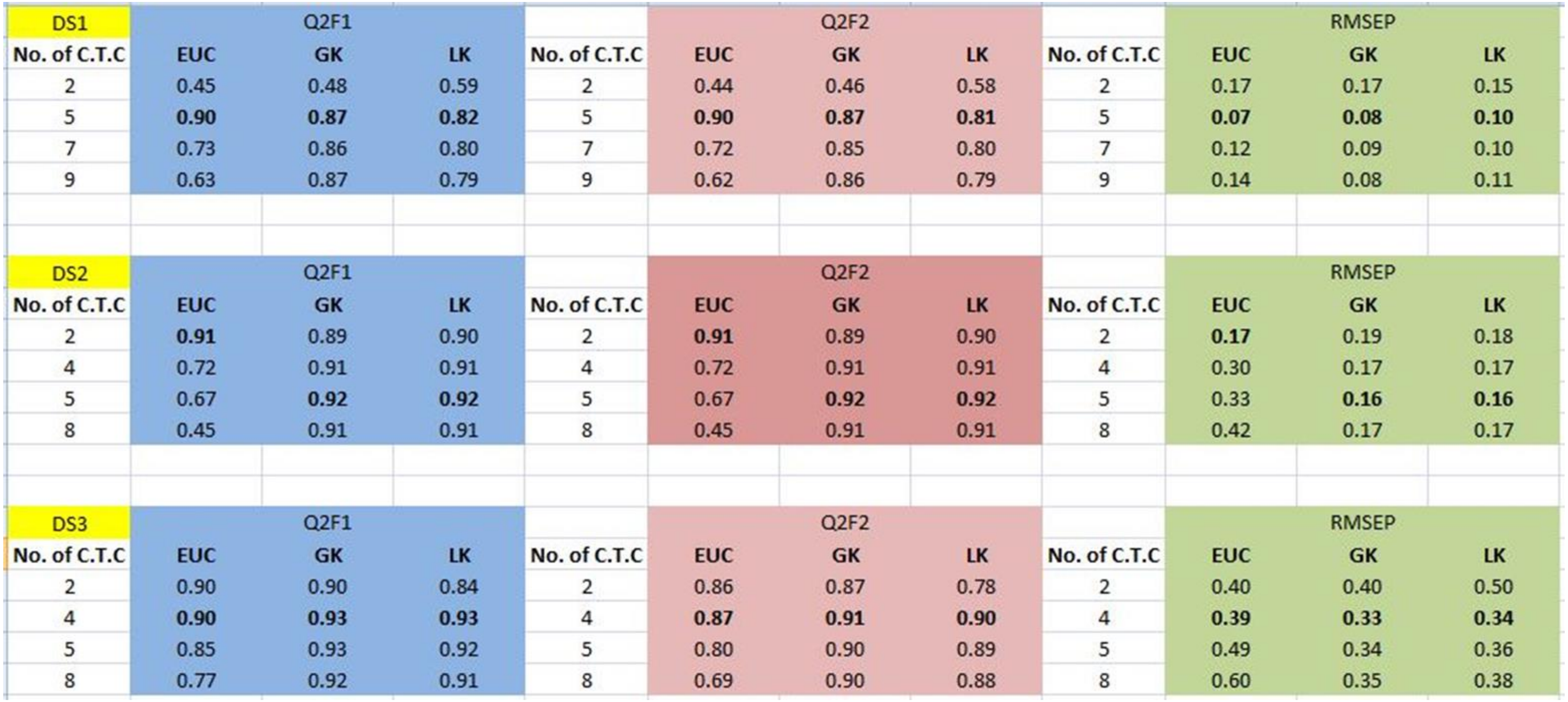
LAPLACIAN KERNEL	Dataset 1				Dataset 2				Dataset 3		
Gamma value	Q2F1	Q2F2	RMSEP		Q2F1	Q2F2	RMSEP		Q2F1	Q2F2	RMSEP
γ = 0.25	0.36	0.34	0.19		0.40	0.40	0.44		0.42	0.22	0.95
γ = 0.50	0.60	0.59	0.15		0.67	0.67	0.33		0.67	0.56	0.72
γ = 0.75	0.73	0.72	0.12		0.81	0.81	0.25		0.81	0.74	0.55
γ = 1.00	**0.79**	**0.79**	**0.11**		0.87	0.87	0.21		0.87	0.83	0.45
γ = 1.50	0.79	0.79	0.11		0.90	0.90	0.18		**0.91**	**0.88**	**0.38**
γ = 2.00	0.73	0.72	0.12		**0.91**	**0.91**	**0.17**		0.91	0.87	0.38







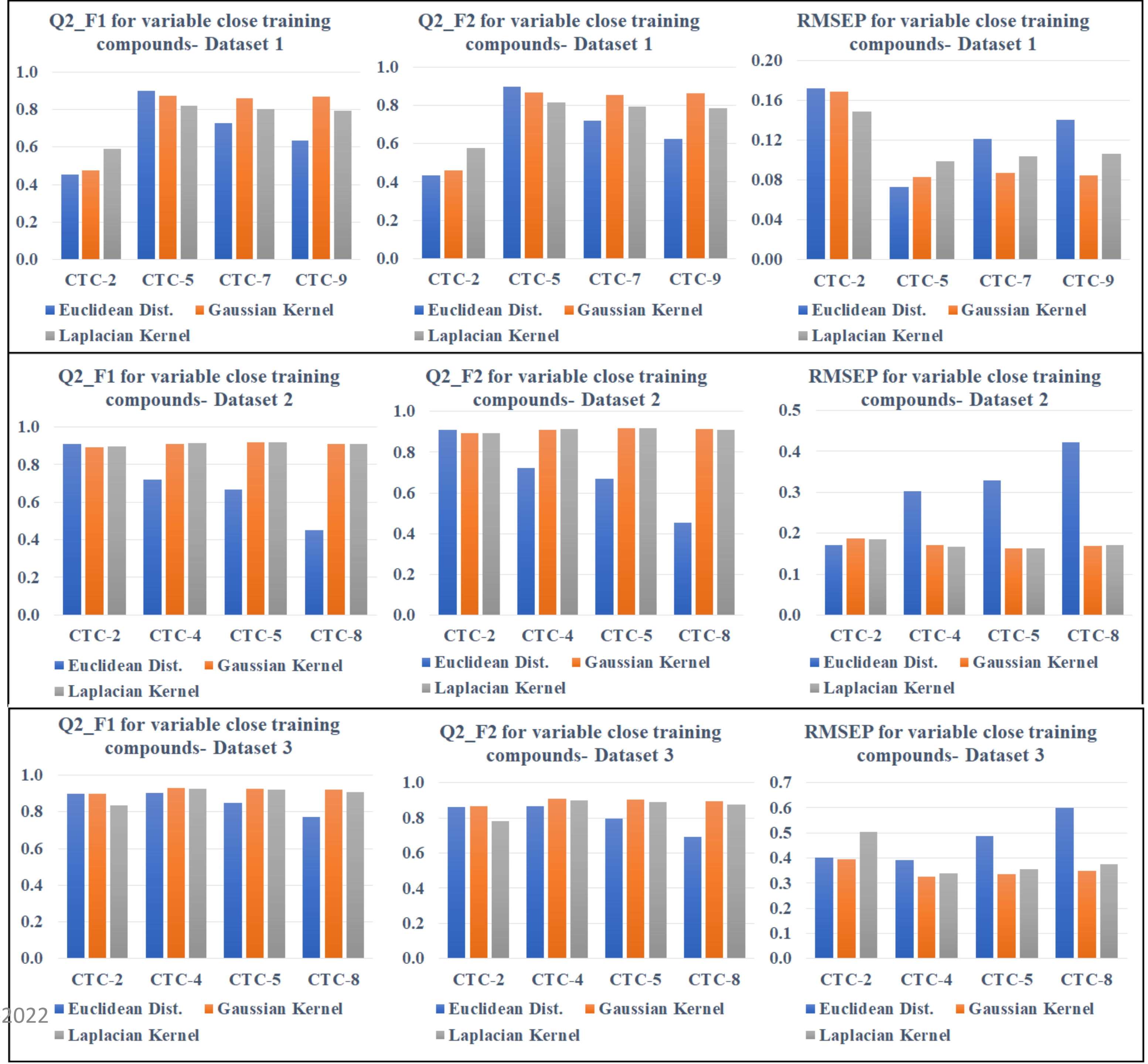
Effects of number of close training compounds on the toxicity prediction in new algorithm



DS1	Q2F1				Q2F2				RMSEP		
No. of C.T.C	EUC	GK	LK	No. of C.T.C	EUC	GK	LK	No. of C.T.C	EUC	GK	LK
2	0.45	0.48	0.59	2	0.44	0.46	0.58	2	0.17	0.17	0.15
5	**0.90**	**0.87**	**0.82**	5	**0.90**	**0.87**	**0.81**	5	**0.07**	**0.08**	**0.10**
7	0.73	0.86	0.80	7	0.72	0.85	0.80	7	0.12	0.09	0.10
9	0.63	0.87	0.79	9	0.62	0.86	0.79	9	0.14	0.08	0.11

DS2	Q2F1				Q2F2				RMSEP		
No. of C.T.C	EUC	GK	LK	No. of C.T.C	EUC	GK	LK	No. of C.T.C	EUC	GK	LK
2	**0.91**	0.89	0.90	2	**0.91**	0.89	0.90	2	**0.17**	0.19	0.18
4	0.72	0.91	0.91	4	0.72	0.91	0.91	4	0.30	0.17	0.17
5	0.67	**0.92**	**0.92**	5	0.67	**0.92**	**0.92**	5	0.33	**0.16**	**0.16**
8	0.45	0.91	0.91	8	0.45	0.91	0.91	8	0.42	0.17	0.17

DS3	Q2F1				Q2F2				RMSEP		
No. of C.T.C	EUC	GK	LK	No. of C.T.C	EUC	GK	LK	No. of C.T.C	EUC	GK	LK
2	0.90	0.90	0.84	2	0.86	0.87	0.78	2	0.40	0.40	0.50
4	**0.90**	**0.93**	**0.93**	4	**0.87**	**0.91**	**0.90**	4	**0.39**	**0.33**	**0.34**
5	0.85	0.93	0.92	5	0.80	0.90	0.89	5	0.49	0.34	0.36
8	0.77	0.92	0.91	8	0.69	0.90	0.88	8	0.60	0.35	0.38



Q2_F1 for variable close training compounds- Dataset 1

Q2_F2 for variable close training compounds- Dataset 1

RMSEP for variable close training compounds- Dataset 1

Q2_F1 for variable close training compounds- Dataset 2

Q2_F2 for variable close training compounds- Dataset 2

RMSEP for variable close training compounds- Dataset 2

Q2_F1 for variable close training compounds- Dataset 3

Q2_F2 for variable close training compounds- Dataset 3

RMSEP for variable close training compounds- Dataset 3

Figure. a) Bar diagram representing the effect of number of close training compounds on the metric values of Dataset 1; b) Bar diagram representing the effect of number of close training compounds on the metric values of Dataset 2; c) Bar diagram representing the effect of number of close training compounds on the metric values of Dataset 3.

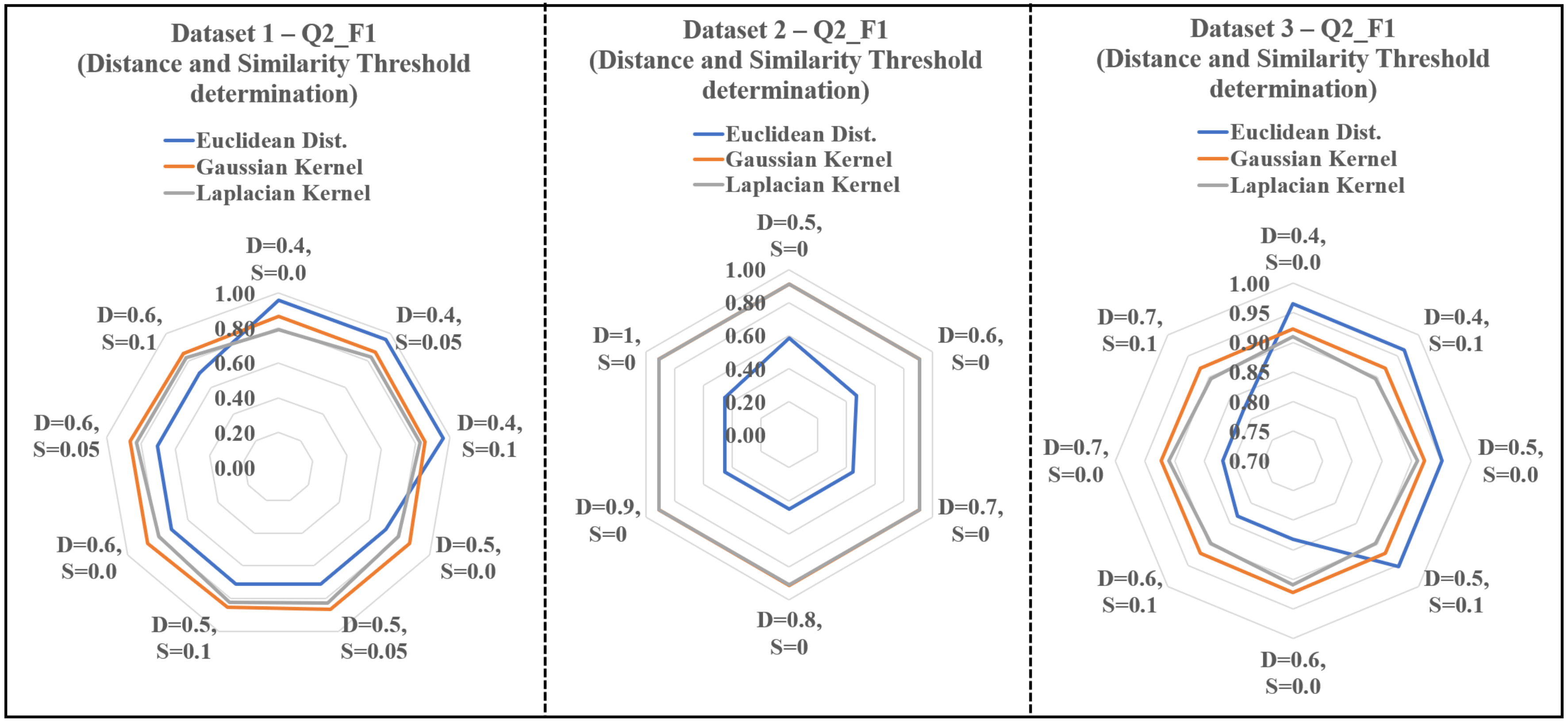
Distance and similarity threshold optimization for the new similarity based read-across algorithm

Dataset 1	Q2F1				Q2F2				RMSEP		
Threshold	EUC	GK	LK		EUC	GK	LK		EUC	GK	LK
D=0.4, S=0.0	0.96	0.87	0.79		0.96	0.86	0.79		0.05	0.08	0.11
D=0.4, S=0.05	**0.96**	**0.87**	**0.83**		**0.96**	**0.86**	**0.82**		**0.05**	**0.09**	**0.10**
D=0.4, S=0.1	0.96	0.85	0.82		0.96	0.85	0.82		0.05	0.09	0.10
D=0.5, S=0.0	0.71	0.87	0.79		0.70	0.86	0.79		0.13	0.08	0.11
D=0.5, S=0.05	0.71	0.87	0.83		0.70	0.86	0.82		0.13	0.09	0.10
D=0.5, S=0.1	0.71	0.85	0.82		0.70	0.85	0.82		0.13	0.09	0.10
D=0.6, S=0.0	0.71	0.87	0.79		0.70	0.86	0.79		0.13	0.08	0.11
D=0.6, S=0.05	0.71	0.87	0.83		0.70	0.86	0.82		0.13	0.09	0.10
D=0.6, S=0.1	0.71	0.85	0.82		0.70	0.85	0.82		0.13	0.09	0.10

Dataset 2	Q2F1				Q2F2				RMSEP		
Threshold	EUC	GK	LK		EUC	GK	LK		EUC	GK	LK
D=0.5, S=0	**0.59**	**0.91**	**0.91**		**0.59**	**0.91**	**0.91**		**0.37**	**0.17**	**0.17**
D=0.6, S=0	0.47	0.91	0.91		0.47	0.91	0.91		0.42	0.17	0.17
D=0.7, S=0	0.45	0.91	0.91		0.45	0.91	0.91		0.43	0.17	0.17
D=0.8, S=0	0.45	0.91	0.91		0.45	0.91	0.91		0.42	0.17	0.17
D=0.9, S=0	0.45	0.91	0.91		0.45	0.91	0.91		0.42	0.17	0.17
D=1, S=0	0.45	0.91	0.91		0.45	0.91	0.91		0.42	0.17	0.17

Dataset 3	Q2F1				Q2F2				RMSEP		
Threshold	EUC	GK	LK		EUC	GK	LK		EUC	GK	LK
D=0.4, S=0.0	**0.96**	**0.92**	**0.91**		**0.95**	**0.90**	**0.88**		**0.23**	**0.35**	**0.38**
D=0.4, S=0.1	0.96	0.92	0.90		0.95	0.89	0.86		0.23	0.35	0.40
D=0.5, S=0.0	0.95	0.92	0.91		0.93	0.90	0.88		0.28	0.35	0.38
D=0.5, S=0.1	0.95	0.92	0.90		0.93	0.89	0.86		0.28	0.35	0.40
D=0.6, S=0.0	0.83	0.92	0.91		0.77	0.90	0.88		0.51	0.35	0.38
D=0.6, S=0.1	0.83	0.92	0.90		0.77	0.89	0.86		0.51	0.35	0.40
D=0.7, S=0.0	0.82	0.92	0.91		0.76	0.90	0.88		0.53	0.35	0.38
D=0.7, S=0.1	0.82	0.92	0.90		0.76	0.89	0.86		0.53	0.35	0.40

Q^2_{F1}

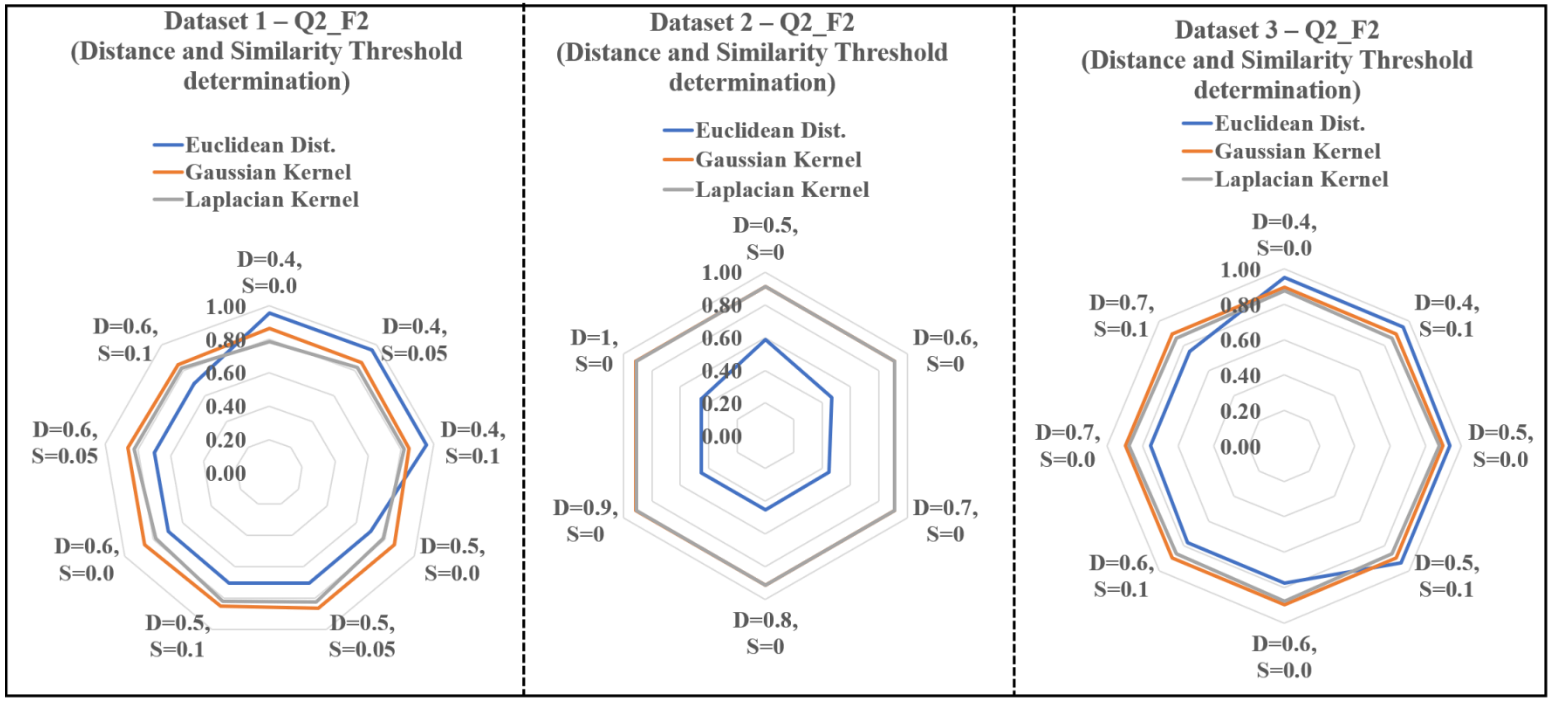


Dataset 1 – Q2_F1 (Distance and Similarity Threshold determination)

Dataset 2 – Q2_F1 (Distance and Similarity Threshold determination)

Dataset 3 – Q2_F1 (Distance and Similarity Threshold determination)

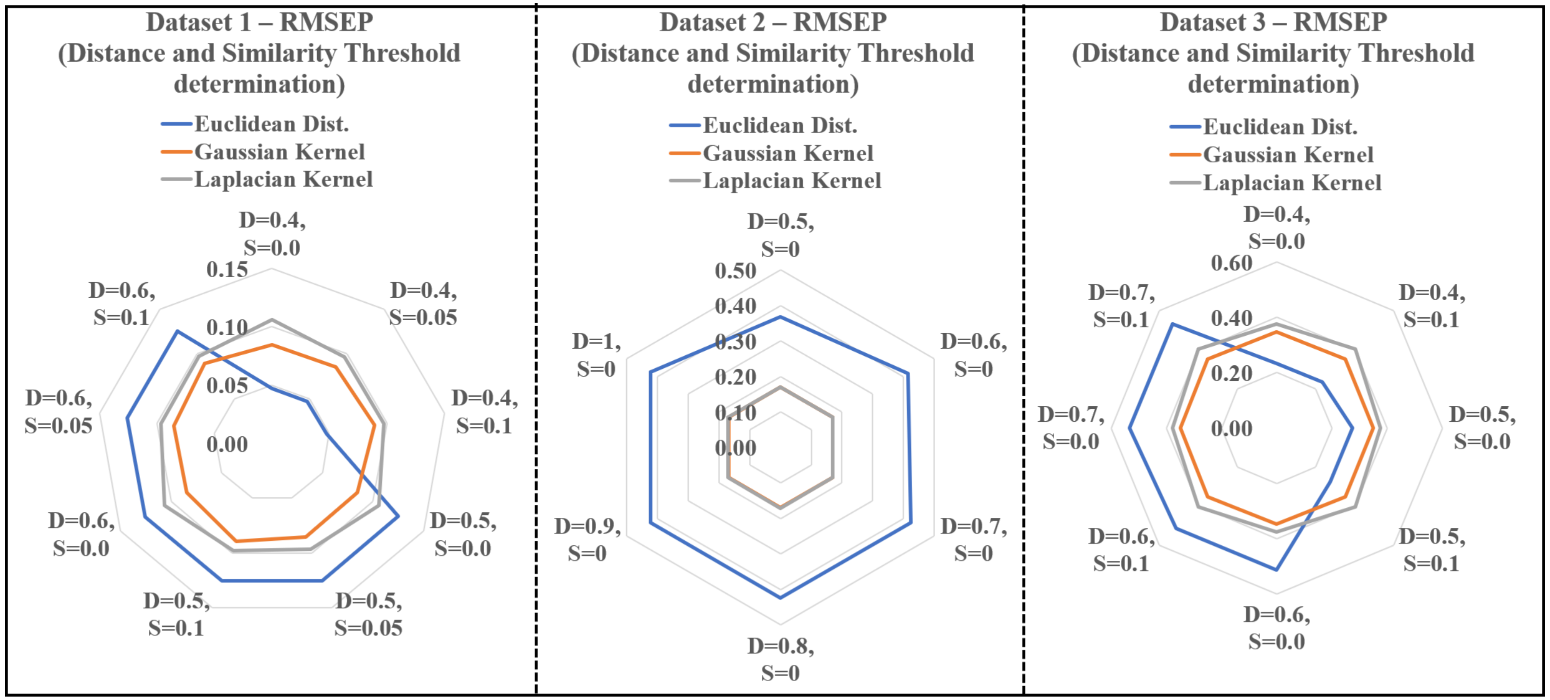
Q^2_{F2}


Dataset 1 – Q2_F2 (Distance and Similarity Threshold determination)

Dataset 2 – Q2_F2 (Distance and Similarity Threshold determination)

Dataset 3 – Q2_F2 (Distance and Similarity Threshold determination)

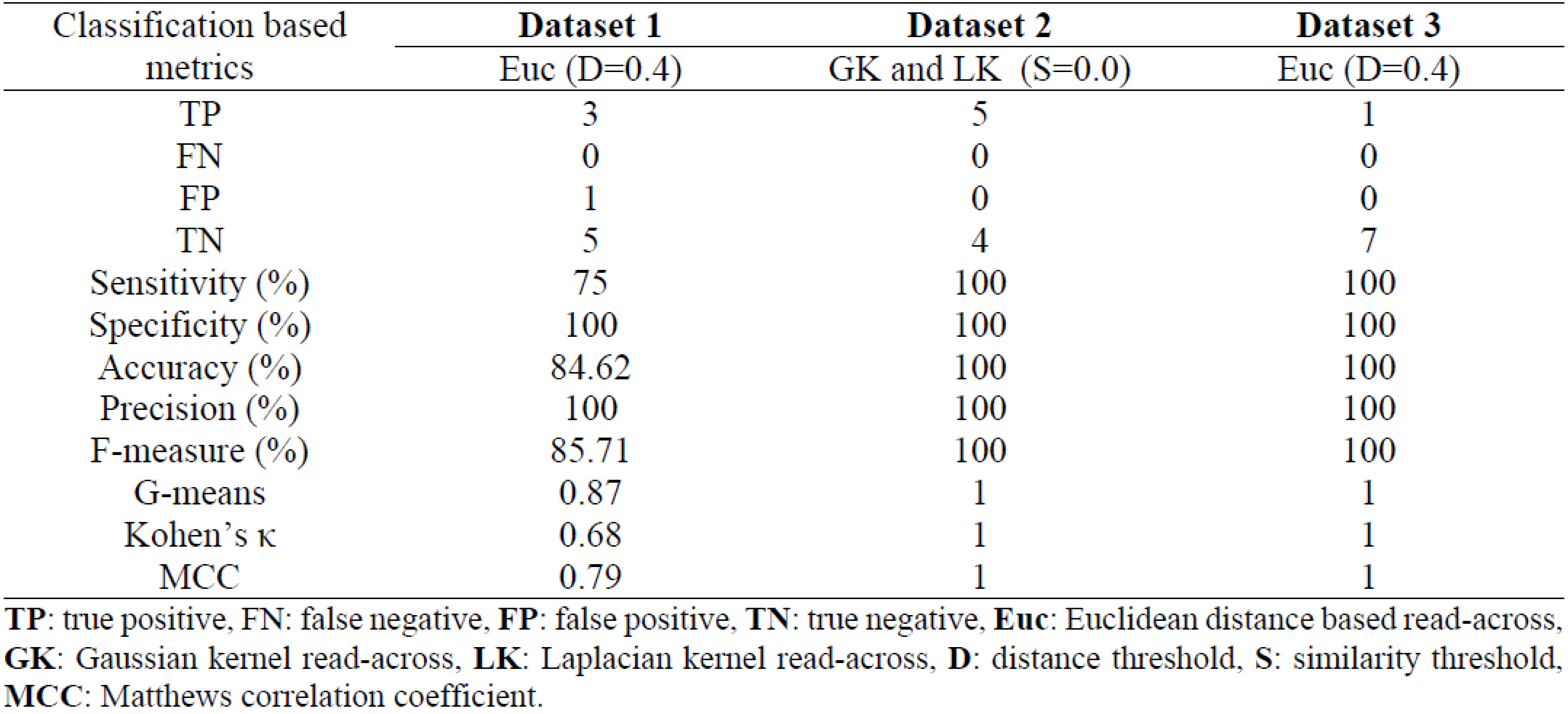
$RMSE_p$


Dataset 1 – RMSEP (Distance and Similarity Threshold determination)

Dataset 2 – RMSEP (Distance and Similarity Threshold determination)

Dataset 3 – RMSEP (Distance and Similarity Threshold determination)

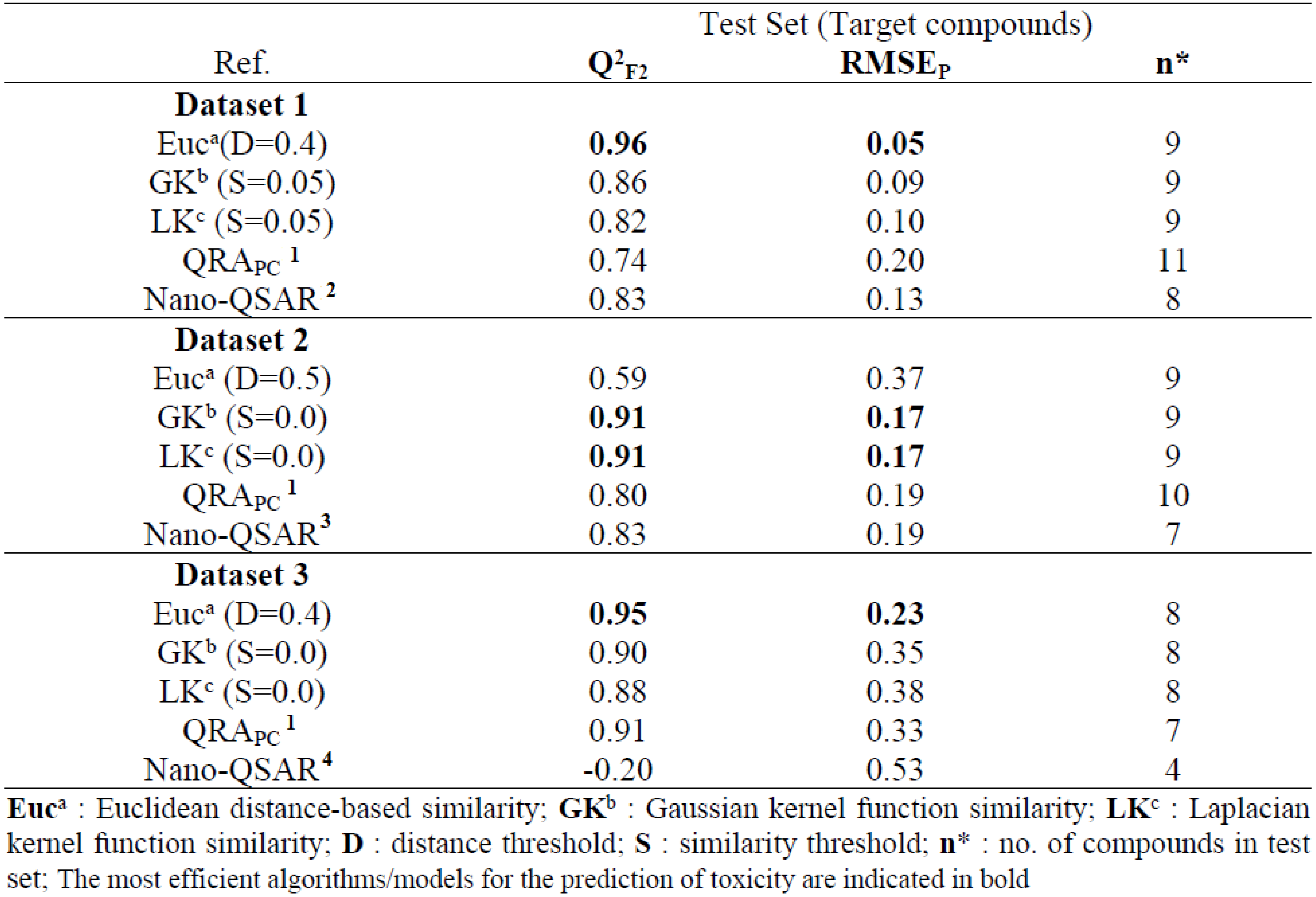
Evaluation of similarity-based read-across algorithm by classification-based metrics



Classification based metrics	**Dataset 1**	**Dataset 2**	**Dataset 3**
	Euc (D=0.4)	GK and LK (S=0.0)	Euc (D=0.4)
TP	3	5	1
FN	0	0	0
FP	1	0	0
TN	5	4	7
Sensitivity (%)	75	100	100
Specificity (%)	100	100	100
Accuracy (%)	84.62	100	100
Precision (%)	100	100	100
F-measure (%)	85.71	100	100
G-means	0.87	1	1
Kohen’s κ	0.68	1	1
MCC	0.79	1	1

TP: true positive, FN: false negative, **FP**: false positive, **TN**: true negative, **Euc**: Euclidean distance based read-across, **GK**: Gaussian kernel read-across, **LK**: Laplacian kernel read-across, **D**: distance threshold, **S**: similarity threshold, **MCC**: Matthews correlation coefficient.

Comparison of performance of new similarity-based algorithm with previously published *in silico* models

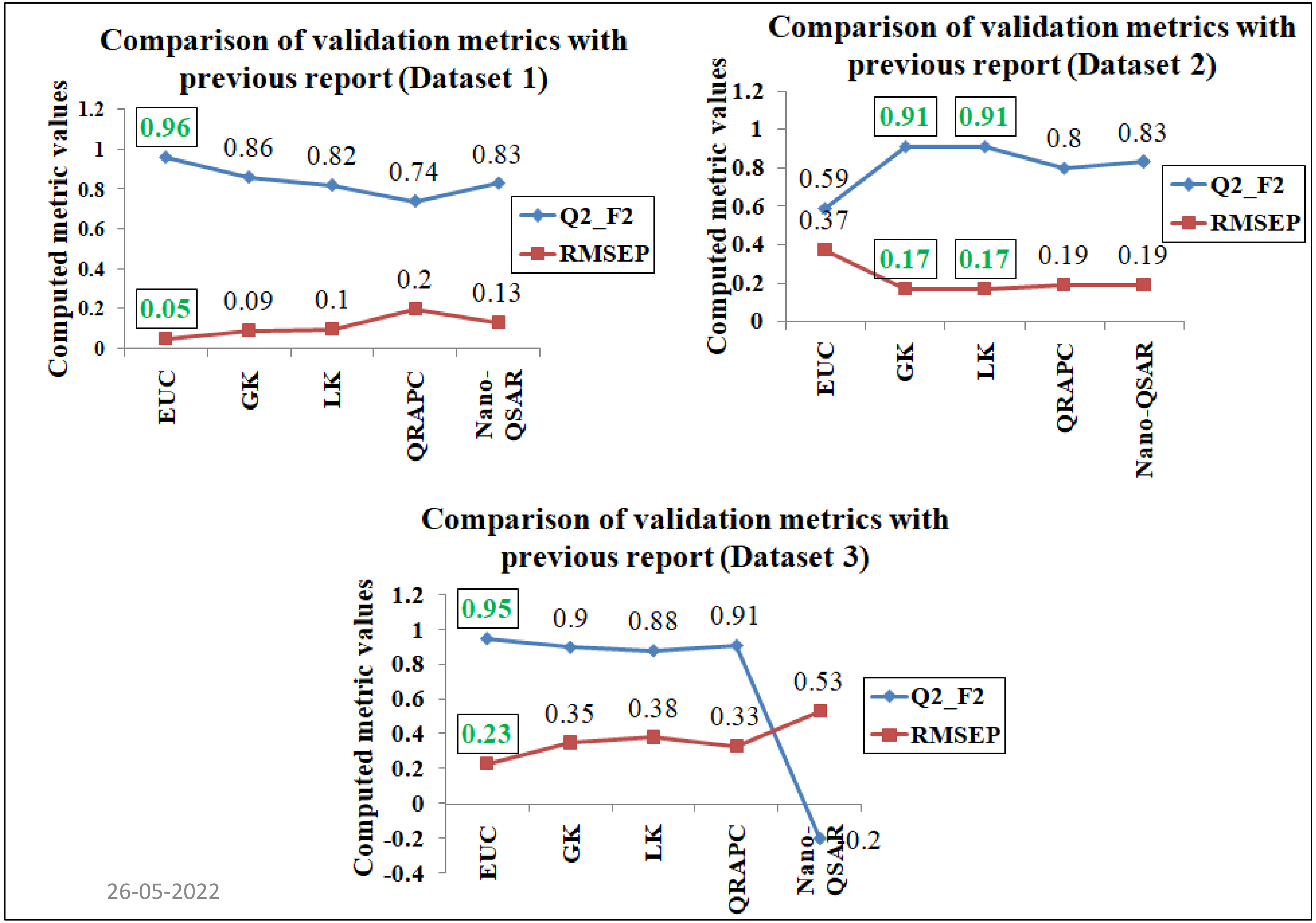


	Test Set (Target compounds)		
Ref.	**Q^2_{F2}**	**RMSE$_P$**	**n***
Dataset 1			
Euc[a](D=0.4)	**0.96**	**0.05**	9
GK[b] (S=0.05)	0.86	0.09	9
LK[c] (S=0.05)	0.82	0.10	9
QRA$_{PC}$ [1]	0.74	0.20	11
Nano-QSAR [2]	0.83	0.13	8
Dataset 2			
Euc[a] (D=0.5)	0.59	0.37	9
GK[b] (S=0.0)	**0.91**	**0.17**	9
LK[c] (S=0.0)	**0.91**	**0.17**	9
QRA$_{PC}$ [1]	0.80	0.19	10
Nano-QSAR[3]	0.83	0.19	7
Dataset 3			
Euc[a] (D=0.4)	**0.95**	**0.23**	8
GK[b] (S=0.0)	0.90	0.35	8
LK[c] (S=0.0)	0.88	0.38	8
QRA$_{PC}$ [1]	0.91	0.33	7
Nano-QSAR[4]	-0.20	0.53	4

Euc[a] : Euclidean distance-based similarity; **GK**[b] : Gaussian kernel function similarity; **LK**[c] : Laplacian kernel function similarity; **D** : distance threshold; **S** : similarity threshold; **n*** : no. of compounds in test set; The most efficient algorithms/models for the prediction of toxicity are indicated in bold

1. A. Gajewicz, *Environ. Sci. Nano*, 2017, **4**, 1389–1403.
2. A. Gajewicz, N. Schaeublin, B. Rasulev, S. Hussain, D. Leszczynska, T. Puzyn and J. Leszczynski, Nanotoxicology, 2015, 9, 313–325.
3. T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T. P. Dasari, A. Michalkova, H.-M. Hwang, A. Toropov, D. Leszczynska and J. Leszczynski, Nat. Nanotechnol., 2011, 6, 175–178.
4. K. Pathakoti, M. J. Huang, J. D. Watts, X. He and H. M. Hwang, J. Photochem. Photobiol. B Biol., 2014, 130, 234–240.

Comparison of performance of new similarity-based algorithm with previously published *in silico* models



Comparison of validation metrics with previous report (Dataset 1)

Comparison of validation metrics with previous report (Dataset 2)

Comparison of validation metrics with previous report (Dataset 3)

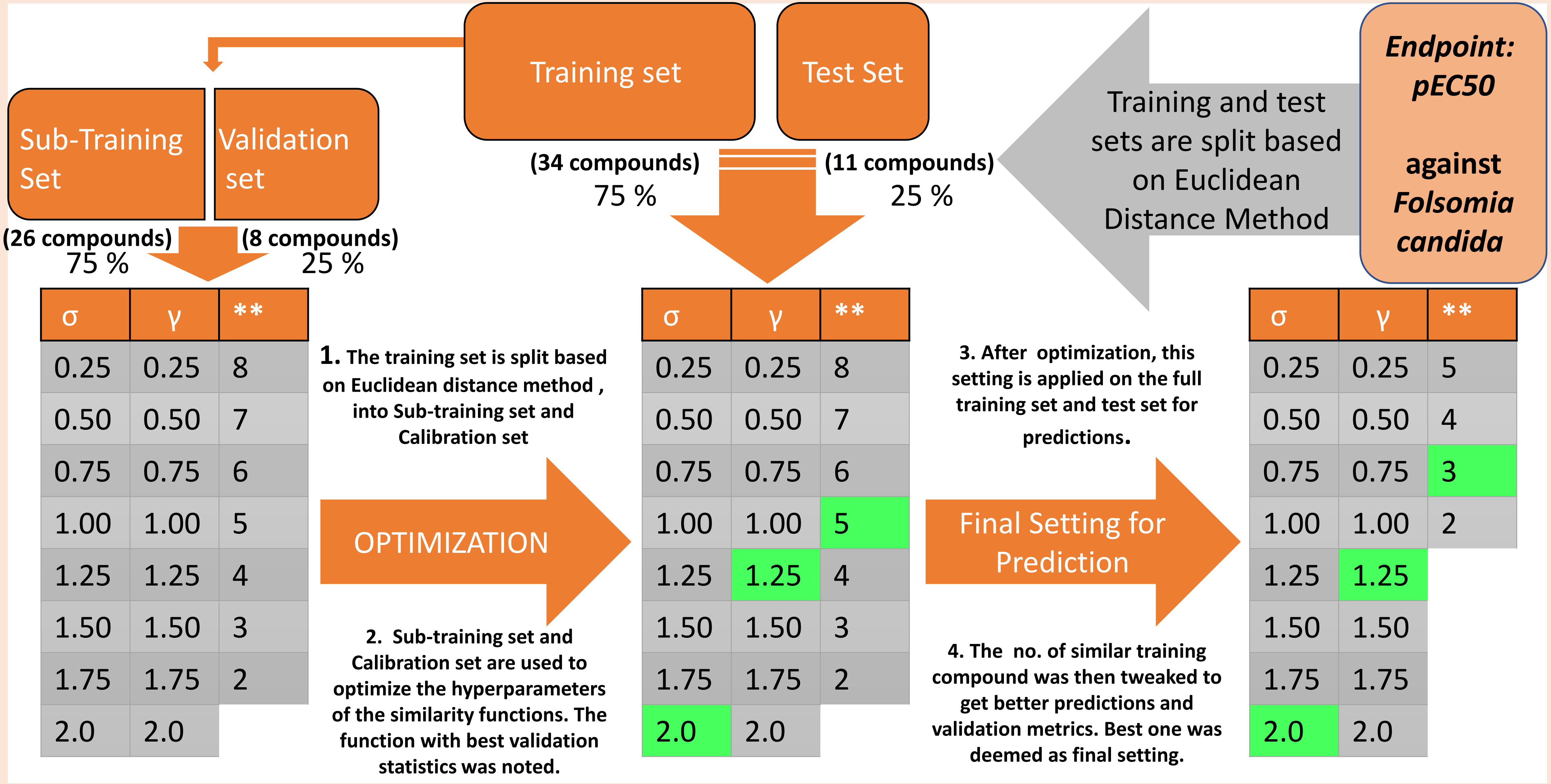
Figure:
Graphical representation of external validation metrics (**Q^2_{F2}, $RMSE_P$**) obtained from the new similarity based methods and previously published methods (**QRA_{PC}** and **Nano-QSAR**)

Summary of Nano-read-across studies

- A new quantitative read-across algorithm based on various similarity estimation techniques was introduced.
- Euclidean distance, Gaussian kernel function, and Laplacian kernel function – used for similarity estimation.
- Optimization of sigma and gamma values of Gaussian and Laplacian kernel function, respectively.
- Assessment of effect of number of close training compounds to the prediction quality was performed→ 2-5 close training compounds can efficiently predict the toxicity of query compounds.
- A distance threshold for the Euclidean distance similarity estimation and a similarity threshold for the Gaussian and Laplacian kernel function similarity estimations– better results. Suitable distance threshold = 0.4 to 0.5; suitable similarity threshold = 0.00 to 0.05.
- A simple java based computer program has also been developed (available at: https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home).
- The new similarity-based read-across algorithm and the designed software are easy to use, efficient, and an expert independent alternative method for the toxicity prediction of MeOx nanoparticles.

WORKFLOW

Read across prediction of soil ecotoxicity against *Folsomia candida*



Endpoint: pEC50

against Folsomia candida

Training and test sets are split based on Euclidean Distance Method

Training set

(34 compounds)
75 %

Test Set

(11 compounds)
25 %

Sub-Training Set

(26 compounds)
75 %

Validation set

(8 compounds)
25 %

σ	γ	**
0.25	0.25	8
0.50	0.50	7
0.75	0.75	6
1.00	1.00	5
1.25	1.25	4
1.50	1.50	3
1.75	1.75	2
2.0	2.0	

1. The training set is split based on Euclidean distance method , into Sub-training set and Calibration set

OPTIMIZATION

2. Sub-training set and Calibration set are used to optimize the hyperparameters of the similarity functions. The function with best validation statistics was noted.

σ	γ	**
0.25	0.25	8
0.50	0.50	7
0.75	0.75	6
1.00	1.00	**5**
1.25	**1.25**	4
1.50	1.50	3
1.75	1.75	2
2.0	2.0	

3. After optimization, this setting is applied on the full training set and test set for predictions.

Final Setting for Prediction

4. The no. of similar training compound was then tweaked to get better predictions and validation metrics. Best one was deemed as final setting.

σ	γ	**
0.25	0.25	5
0.50	0.50	4
0.75	0.75	**3**
1.00	1.00	2
1.25	**1.25**	
1.50	1.50	
1.75	1.75	
2.0	2.0	

🟩 - **Values selected**

Pal et al, unpublished work

**** - Number of similar training compounds**

Results

- *At the final setting*

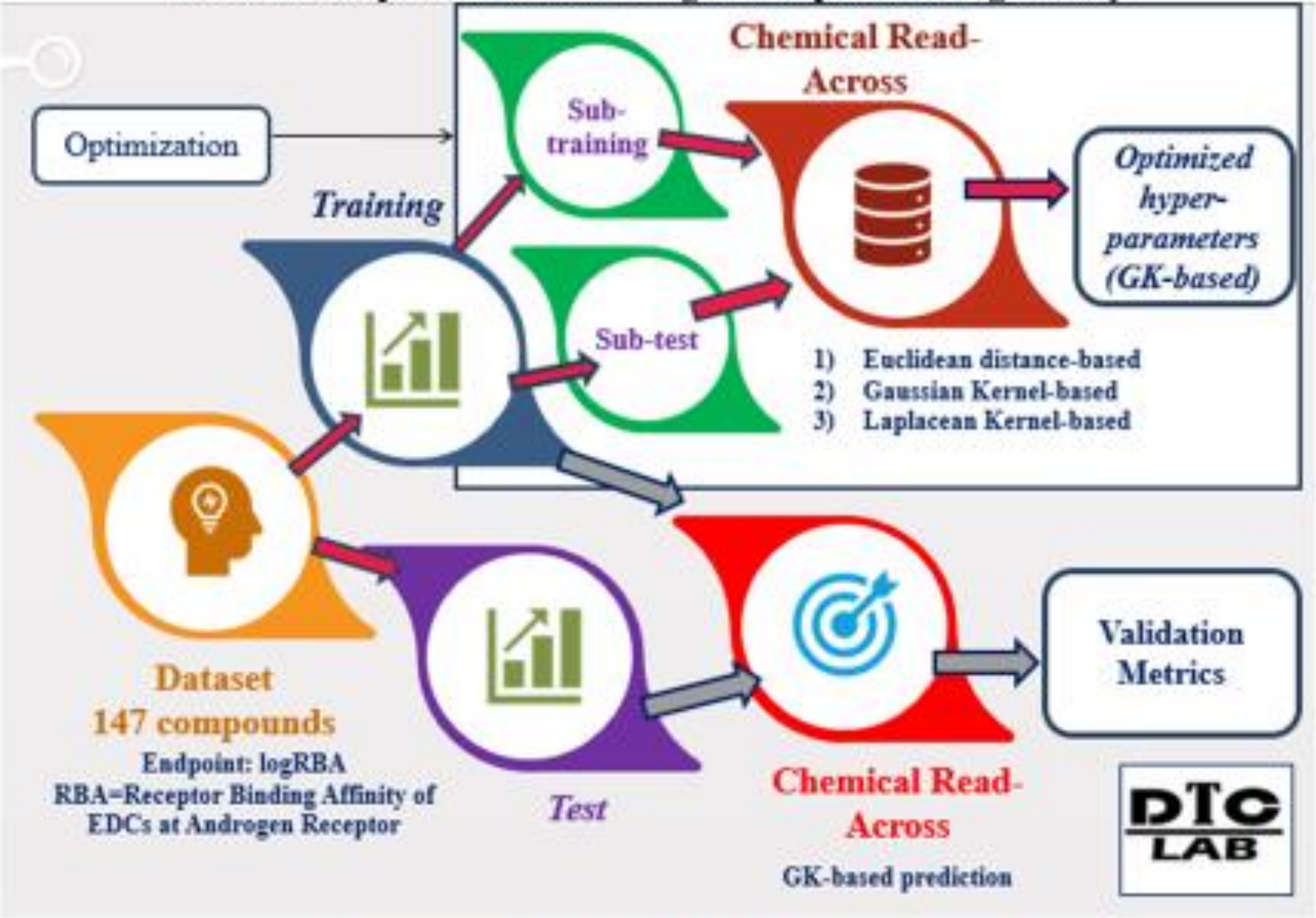
σ = 2.00
γ = 1.25

No. of similar Training Compounds = 3

	Yeuc(Test)	Ygk(Test)	Ylk(Test)
Q^2_{F1}	0.7613	0.7747	0.7393
Q^2_{F2}	0.7007	0.7174	0.6731
$RMSE_P$	0.7668	0.7449	0.8012

Gaussian kernel based function was found to be best here

Read across prediction of androgen receptor binding affinity



Banerjee et al, unpublished work

Workflow

Results for Chemical Read-Across

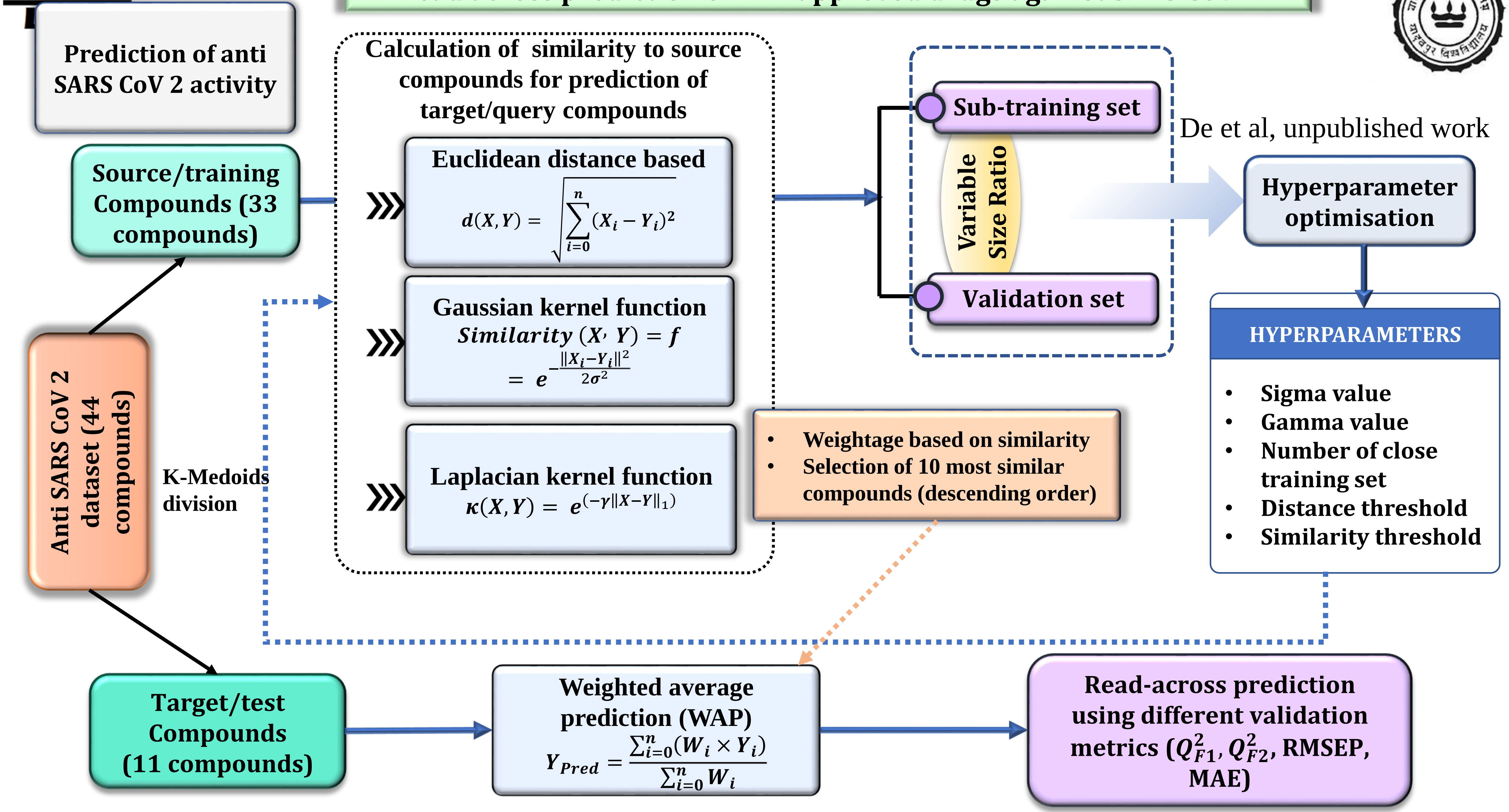
Validation Metrics: **Q^2_F1 :0.635** **Q^2_F2 :0.635**

(for Gaussian Kernel-based similarity consideration)

Sigma value: 1
Gamma value: 1
No. of similar training compounds: 10
Distance Threshold: 1
Similarity Threshold: 0

Optimized Hyper-parameters

Read across prediction of FDA approved drugs against SARS CoV 2



Prediction of anti SARS CoV 2 activity

Anti SARS CoV 2 dataset (44 compounds)

K-Medoids division

Source/training Compounds (33 compounds)

Target/test Compounds (11 compounds)

Calculation of similarity to source compounds for prediction of target/query compounds

Euclidean distance based

$$d(X,Y) = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$

Gaussian kernel function

$$Similarity\ (X, Y) = f = e^{-\frac{\|X_i - Y_i\|^2}{2\sigma^2}}$$

Laplacian kernel function

$$\kappa(X,Y) = e^{(-\gamma\|X-Y\|_1)}$$

- Weightage based on similarity
- Selection of 10 most similar compounds (descending order)

Sub-training set

Variable Size Ratio

Validation set

De et al, unpublished work

Hyperparameter optimisation

HYPERPARAMETERS

- Sigma value
- Gamma value
- Number of close training set
- Distance threshold
- Similarity threshold

Weighted average prediction (WAP)

$$Y_{Pred} = \frac{\sum_{i=0}^{n}(W_i \times Y_i)}{\sum_{i=0}^{n} W_i}$$

Read-across prediction using different validation metrics (Q^2_{F1}, Q^2_{F2}, RMSEP, MAE)

- The **antiviral** dataset consists of **44 compounds**
- **Training set** is composed of **33 compounds**, **test set** is composed of **11 compounds**
- **Four combination of features** (as described by M1, M2, M3 and M4) were used for read across prediction

MODEL FEATURES

Combination No.	FEATURES
M1	nROR, F06[C-Cl], NsNH2, VE1sign_Dz(p)
M2	nROR, F06[C-Cl], NsNH2, nRCOOR
M3	nROR, F06[C-Cl], NsNH2, VE1_B(e)
M4	nROR, F06[C-Cl], NsNH2, VE1_H2

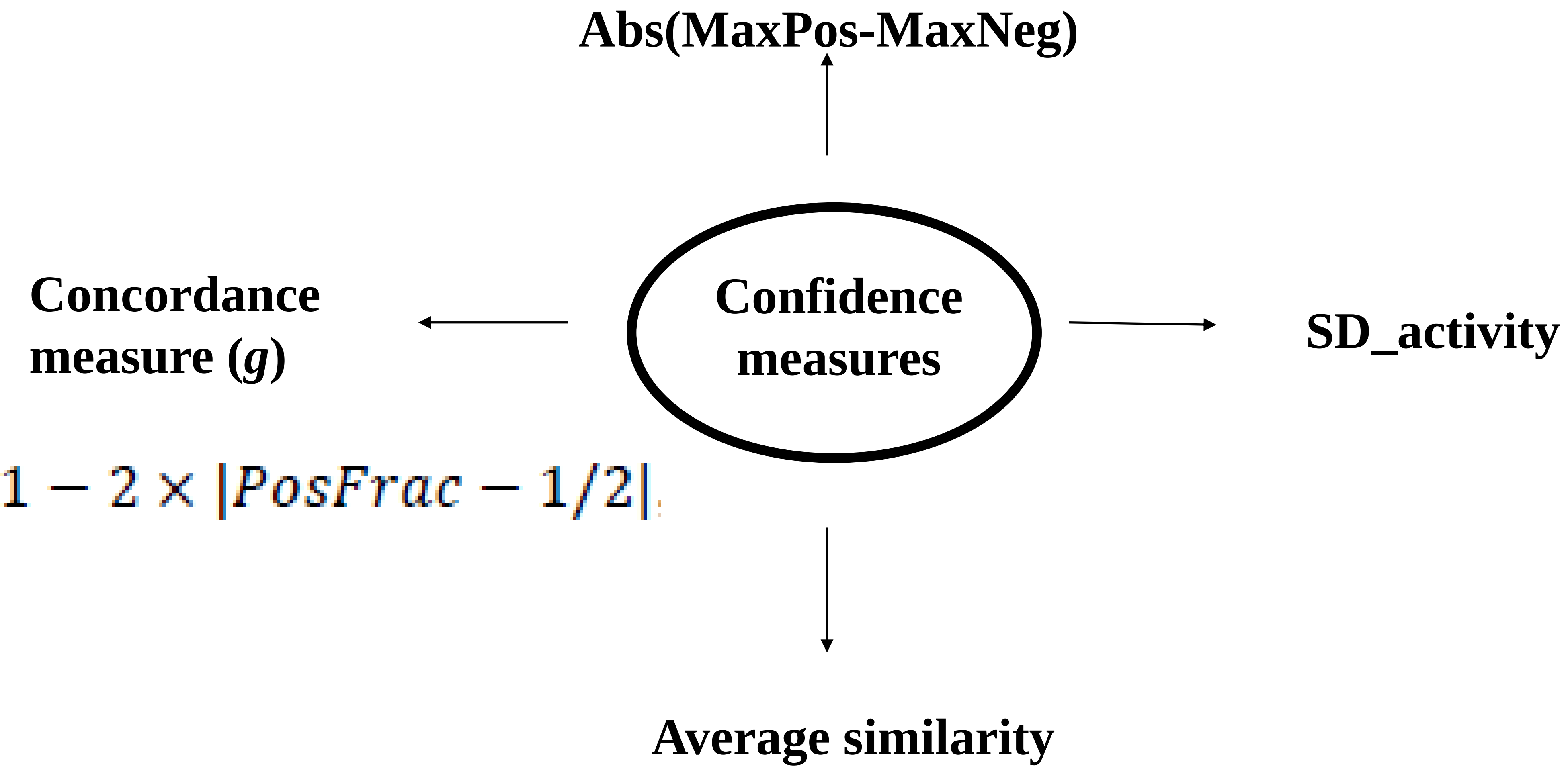
HYPERPARAMETER OPTIMISATION

Combination No.	Sigma value	Gamma value	No. of close training compounds	Distance threshold	Similarity threshold
M1	1.5	1.5	10	0.5	0
M2	1	1	10	0.6	0
M3	0.75	1.5	10	0.5	0
M4	0.75	1.75	10	0.6	0

- **READ ACROSS PREDICTION RESULTS**

Validation metrics	M1			M2			M3			M4		
	Pred Y_{euc}	Pred Y_{gk}	Pred Y_{lk}	Pred Y_{euc}	Pred Y_{gk}	Pred Y_{lk}	Pred Y_{euc}	Pred Y_{gk}	Pred Y_{lk}	Pred Y_{euc}	Pred Y_{gk}	Pred Y_{lk}
Q^2_{F1}	0.879	0.893	0.909	0.870	0.912	0.911	0.862	0.912	0.892	0.722	0.931	**0.932**
Q^2_{F2}	0.878	0.893	0.909	0.870	0.912	0.911	0.862	0.912	0.892	0.722	0.931	**0.932**
RMSEP	0.152	0.143	0.132	0.157	0.129	0.131	0.162	0.130	0.144	0.230	0.115	**0.114**
MAE	0.127	0.121	0.118	0.135	0.124	0.119	0.142	0.114	0.132	0.163	0.100	**0.104**

Reliability of Quantitative Read-Across Predictions



Abs(MaxPos-MaxNeg)

Concordance measure (*g*)

Confidence measures

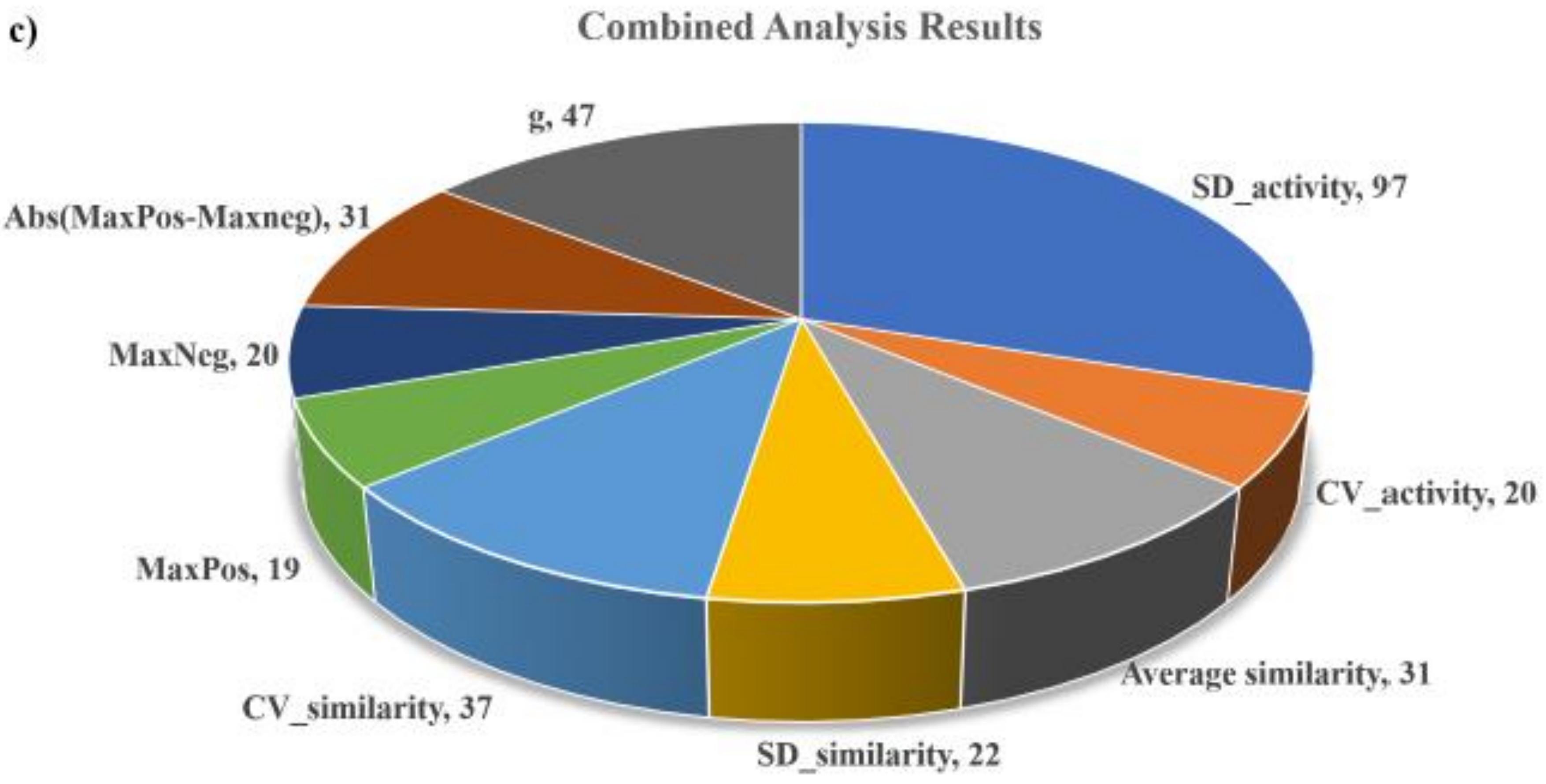
SD_activity

Average similarity

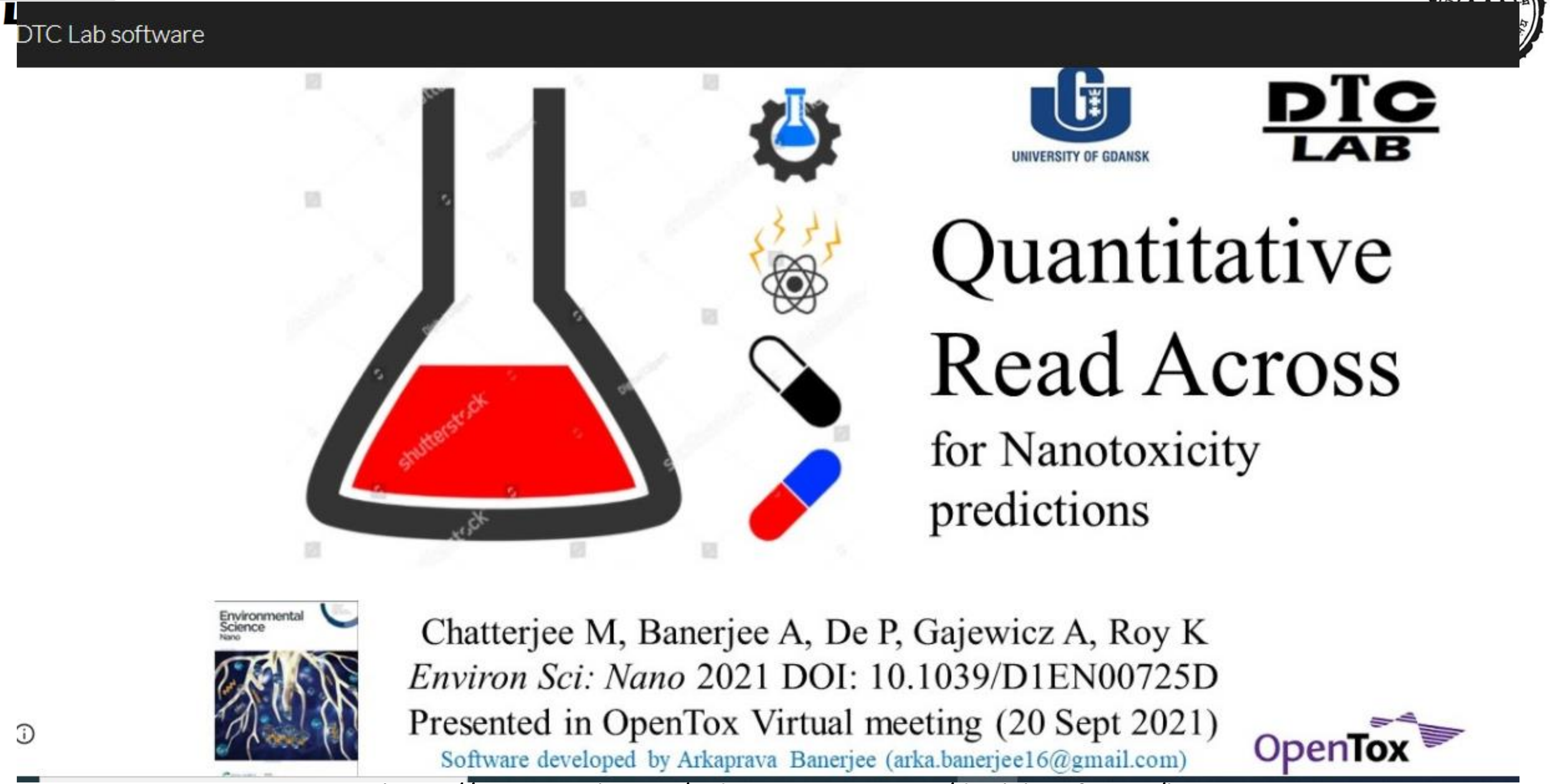
$$g = 1 - 2 \times |PosFrac - 1/2|$$

$$95\%\ confidence\ interval\ of\ read - across\ predictions\ =\ weighted\ average + t_{95\%} \times \frac{s_{weighted}}{\sqrt{n}}$$

Reliability of Quantitative Read-Across Predictions



Banerjee A, Chatterjee M, De P, Roy K, 2022 (Submitted)



DTC Lab software

Quantitative Read Across

for Nanotoxicity predictions

Chatterjee M, Banerjee A, De P, Gajewicz A, Roy K
Environ Sci: Nano 2021 DOI: 10.1039/D1EN00725D
Presented in OpenTox Virtual meeting (20 Sept 2021)
Software developed by Arkaprava Banerjee (arka.banerjee16@gmail.com)

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

Environmental
Science
Nano

ROYAL SOCIETY
OF CHEMISTRY

PAPER

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d1en00725d

A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data†

Mainak Chatterjee,[a] Arkaprava Banerjee,[a] Priyanka De,[a]
Agnieszka Gajewicz-Skretna[b] and Kunal Roy*[a]

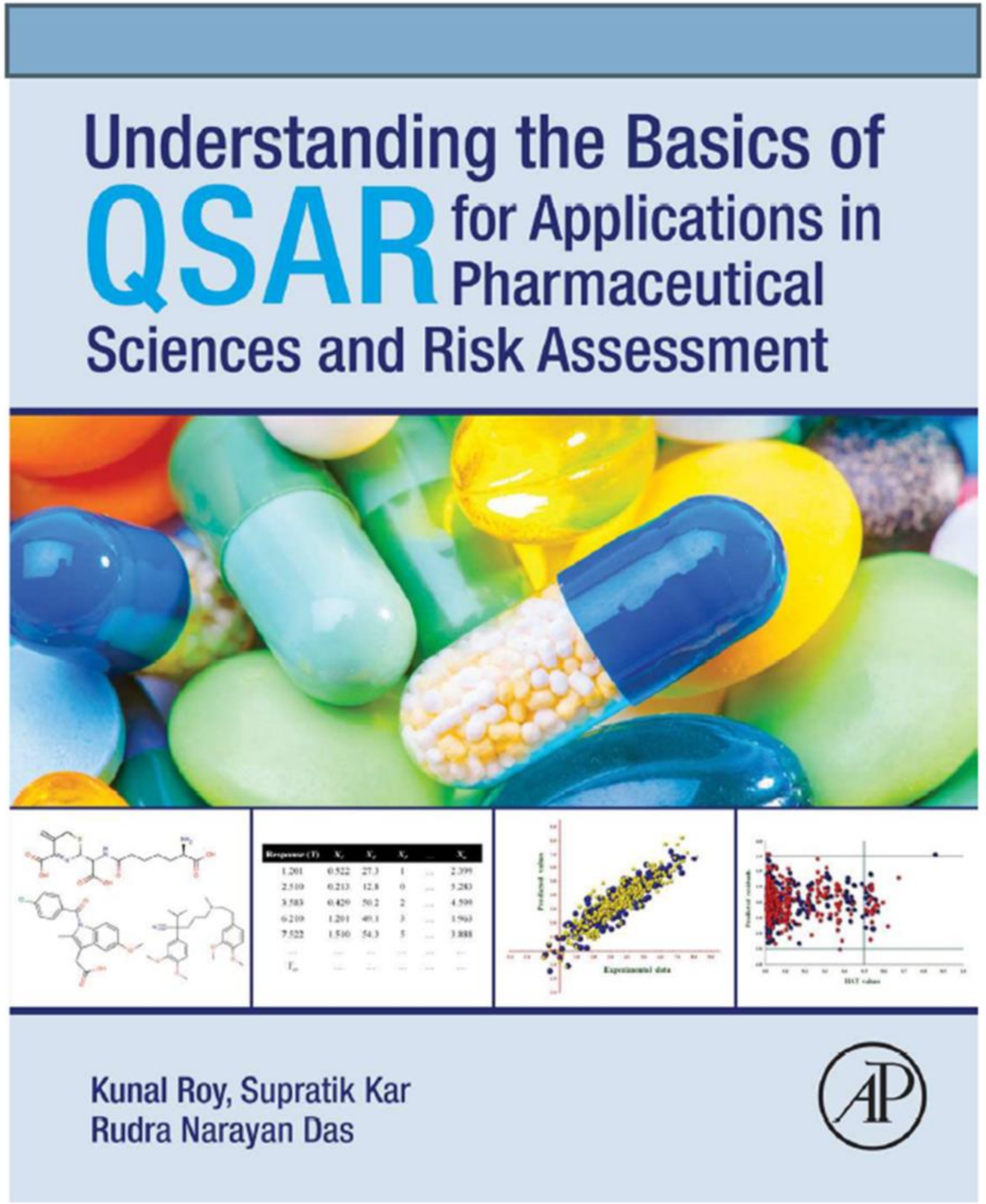
https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home

Acknowledgements

SERB INDIA

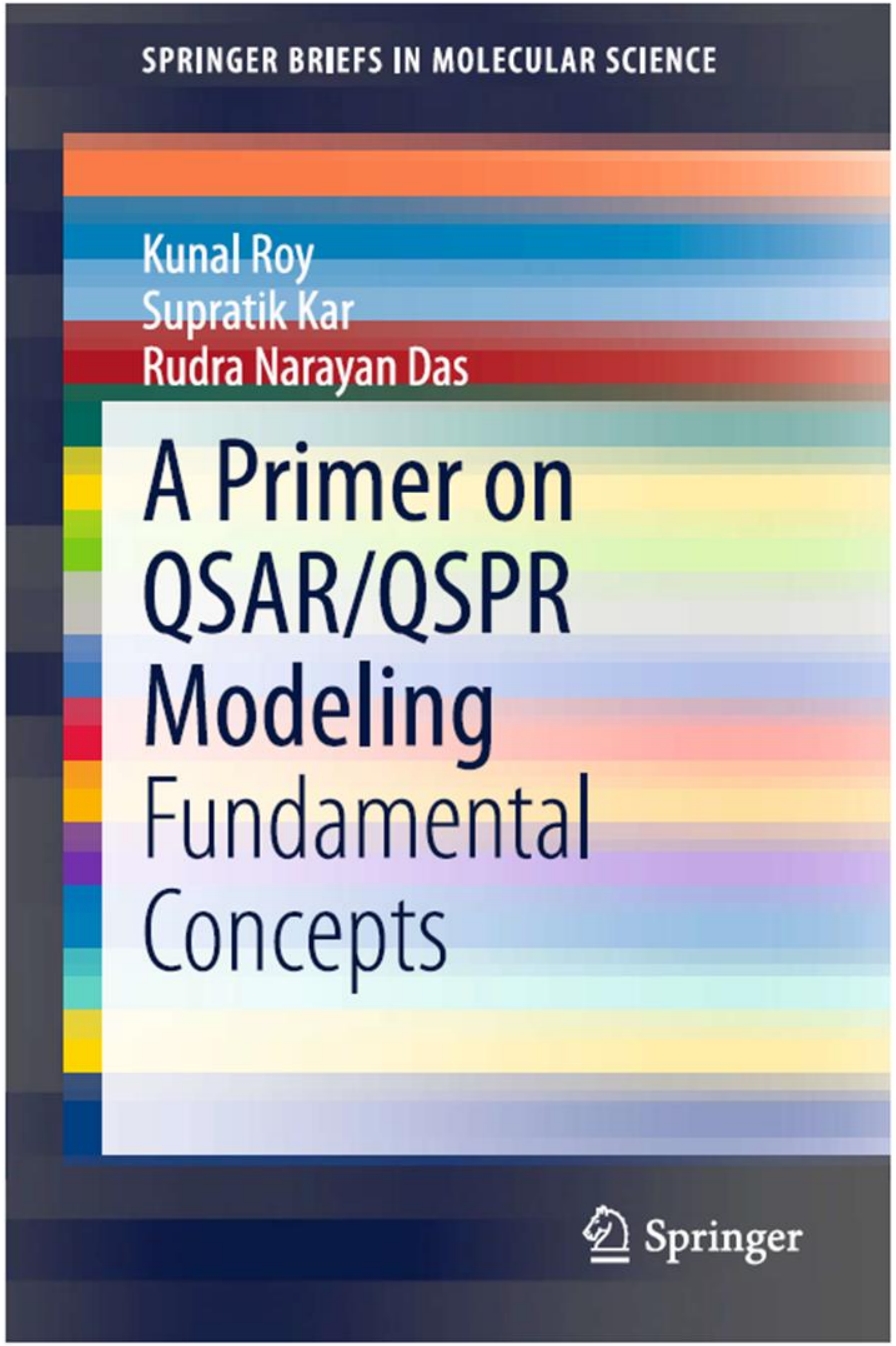
विज्ञान एवं प्रौद्योगिकी विभाग
DEPARTMENT OF
SCIENCE & TECHNOLOGY

सत्यमेव जयते



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

Kunal Roy, Supratik Kar
Rudra Narayan Das



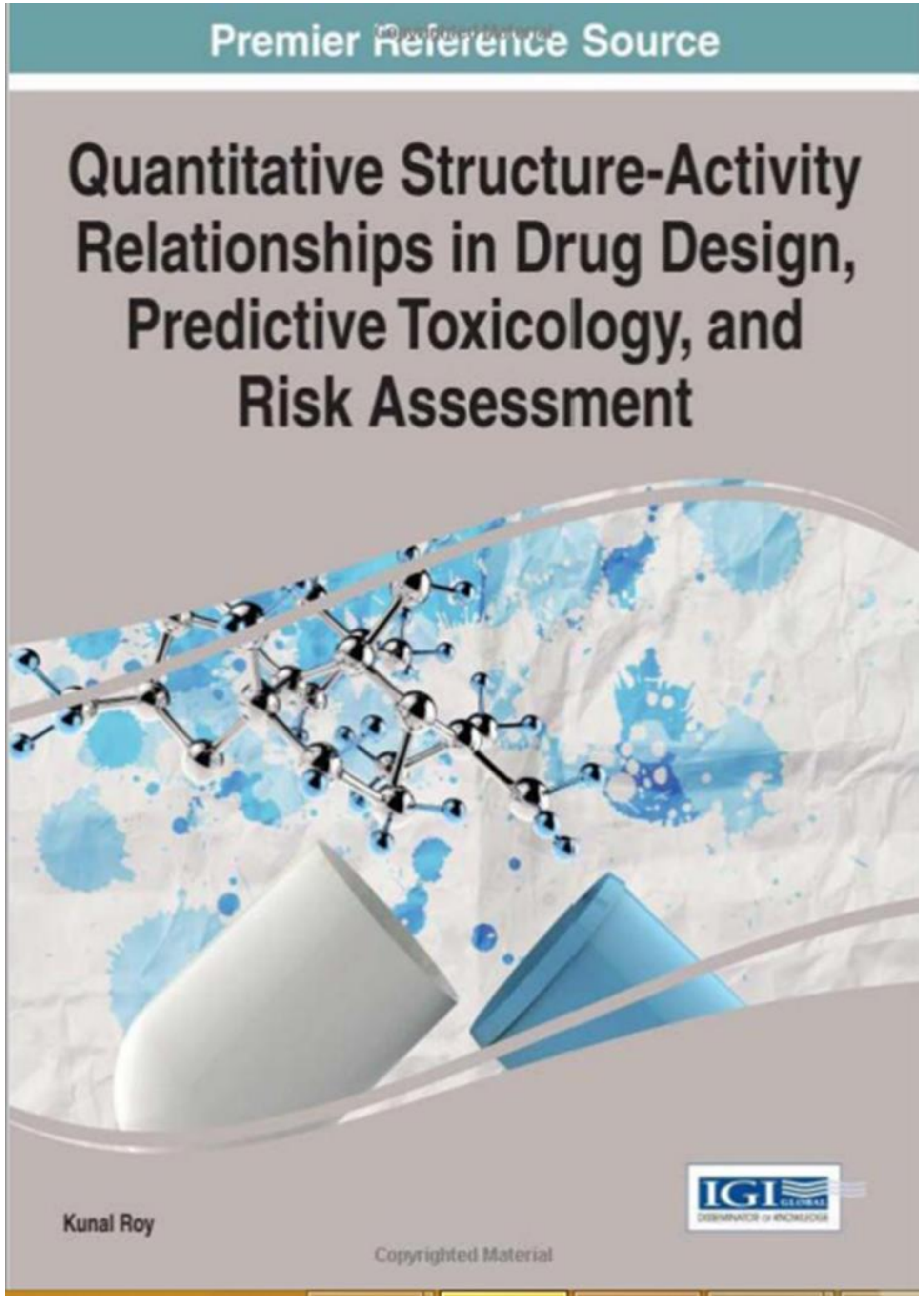
SPRINGER BRIEFS IN MOLECULAR SCIENCE

Kunal Roy
Supratik Kar
Rudra Narayan Das

A Primer on QSAR/QSPR Modeling

Fundamental Concepts

Springer

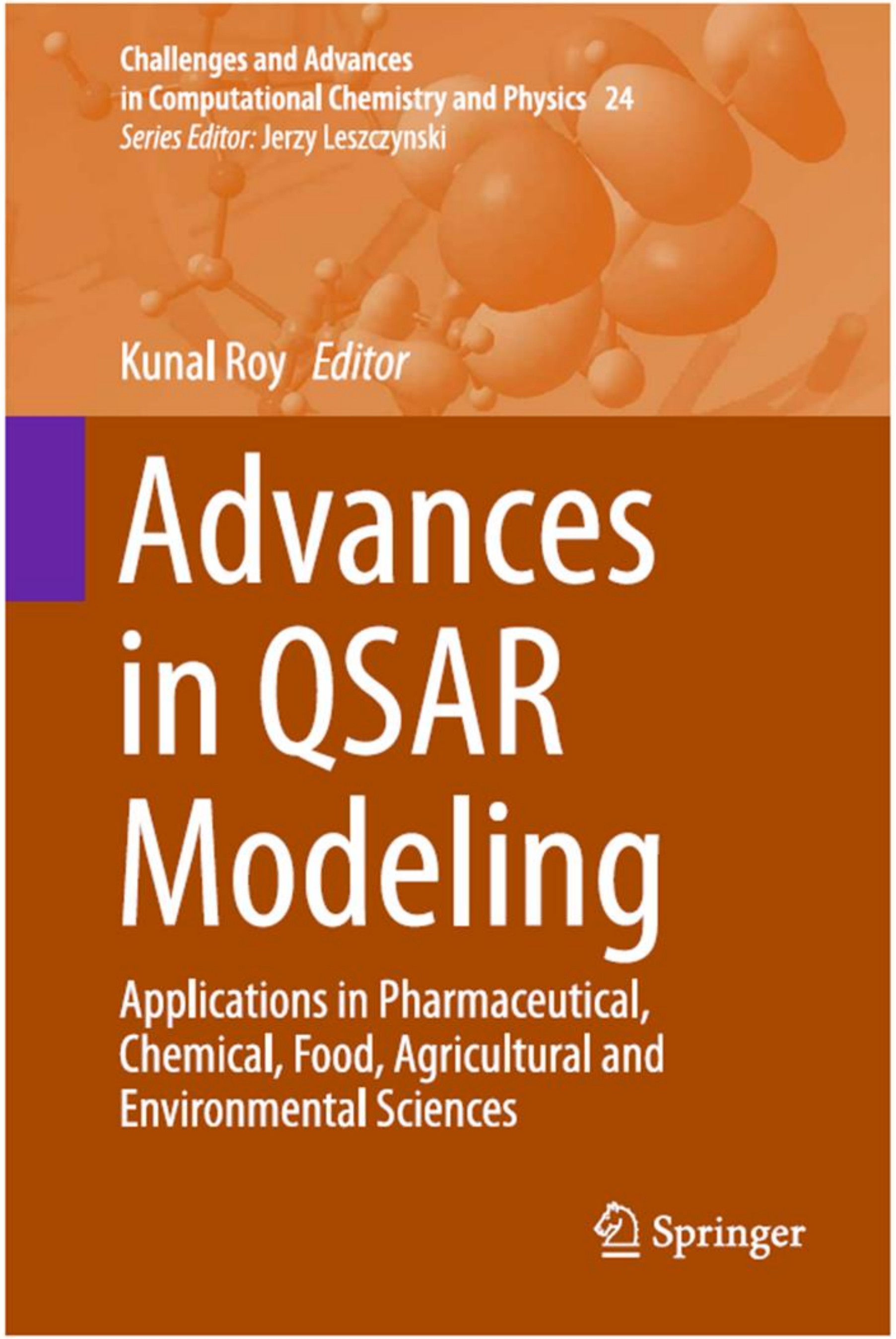


Premier Reference Source

Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment

Kunal Roy

IGI Global



Challenges and Advances in Computational Chemistry and Physics 24

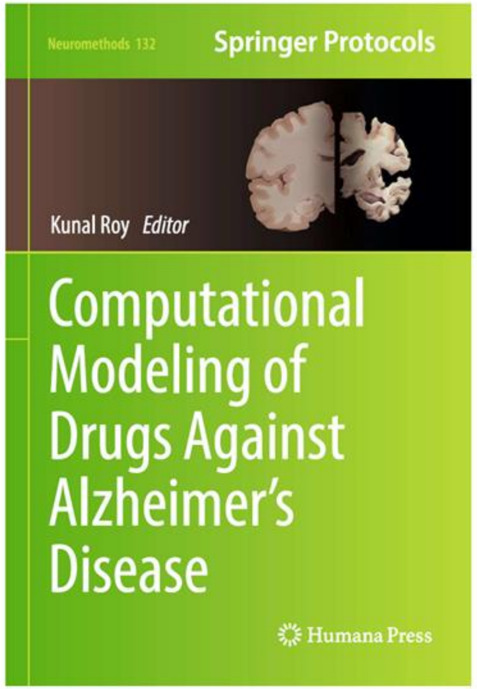
Series Editor: Jerzy Leszczynski

Kunal Roy *Editor*

Advances in QSAR Modeling

Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences

Springer



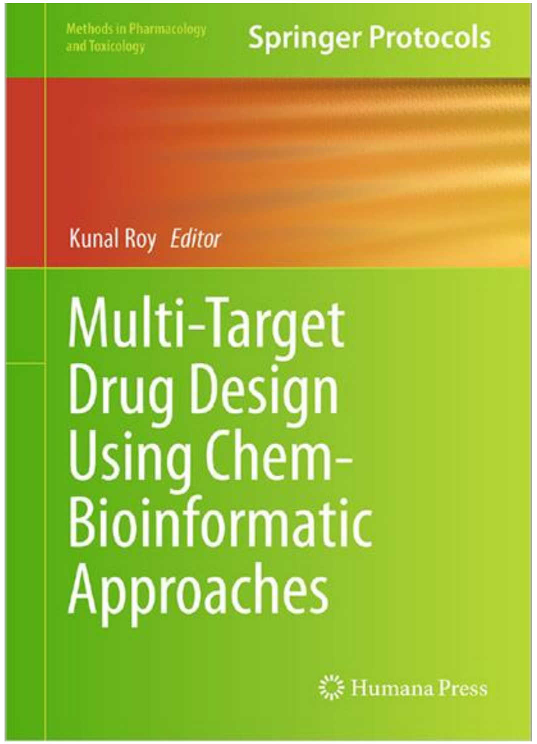
Neuromethods 132

Springer Protocols

Kunal Roy *Editor*

Computational Modeling of Drugs Against Alzheimer's Disease

Humana Press



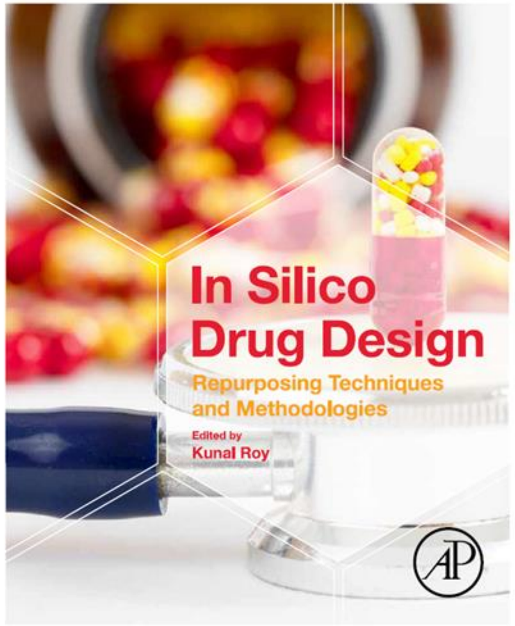
Methods in Pharmacology and Toxicology

Springer Protocols

Kunal Roy *Editor*

Multi-Target Drug Design Using Chem-Bioinformatic Approaches

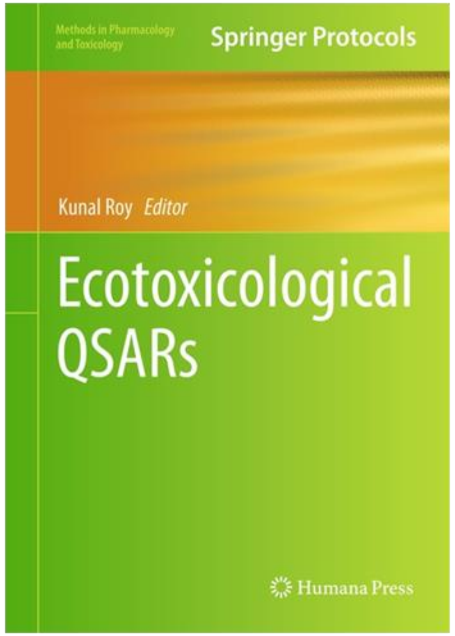
Humana Press



In Silico Drug Design

Repurposing Techniques and Methodologies

Edited by
Kunal Roy



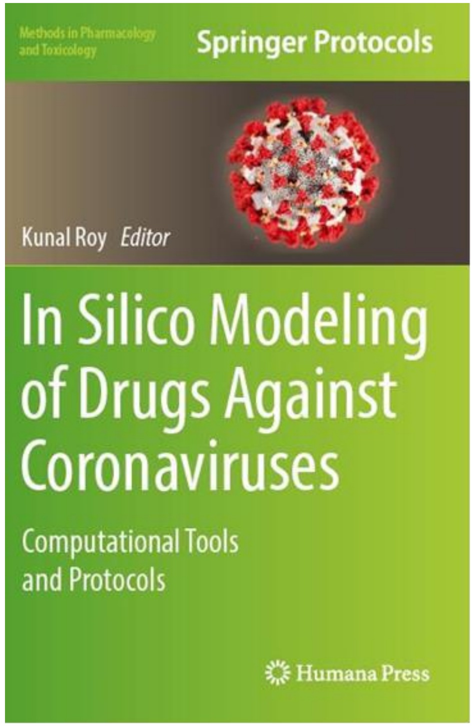
Methods in Pharmacology and Toxicology

Springer Protocols

Kunal Roy *Editor*

Ecotoxicological QSARs

Humana Press



Methods in Pharmacology and Toxicology

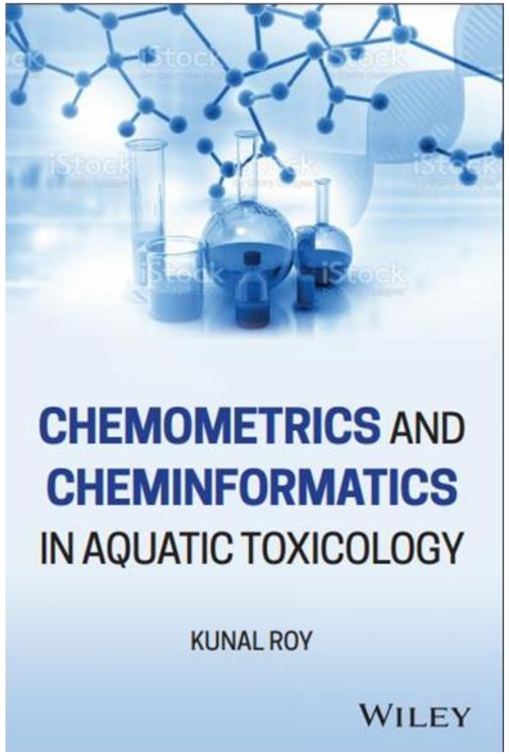
Springer Protocols

Kunal Roy *Editor*

In Silico Modeling of Drugs Against Coronaviruses

Computational Tools and Protocols

Humana Press



CHEMOMETRICS AND CHEMINFORMATICS IN AQUATIC TOXICOLOGY

KUNAL ROY

WILEY

Thank you!