

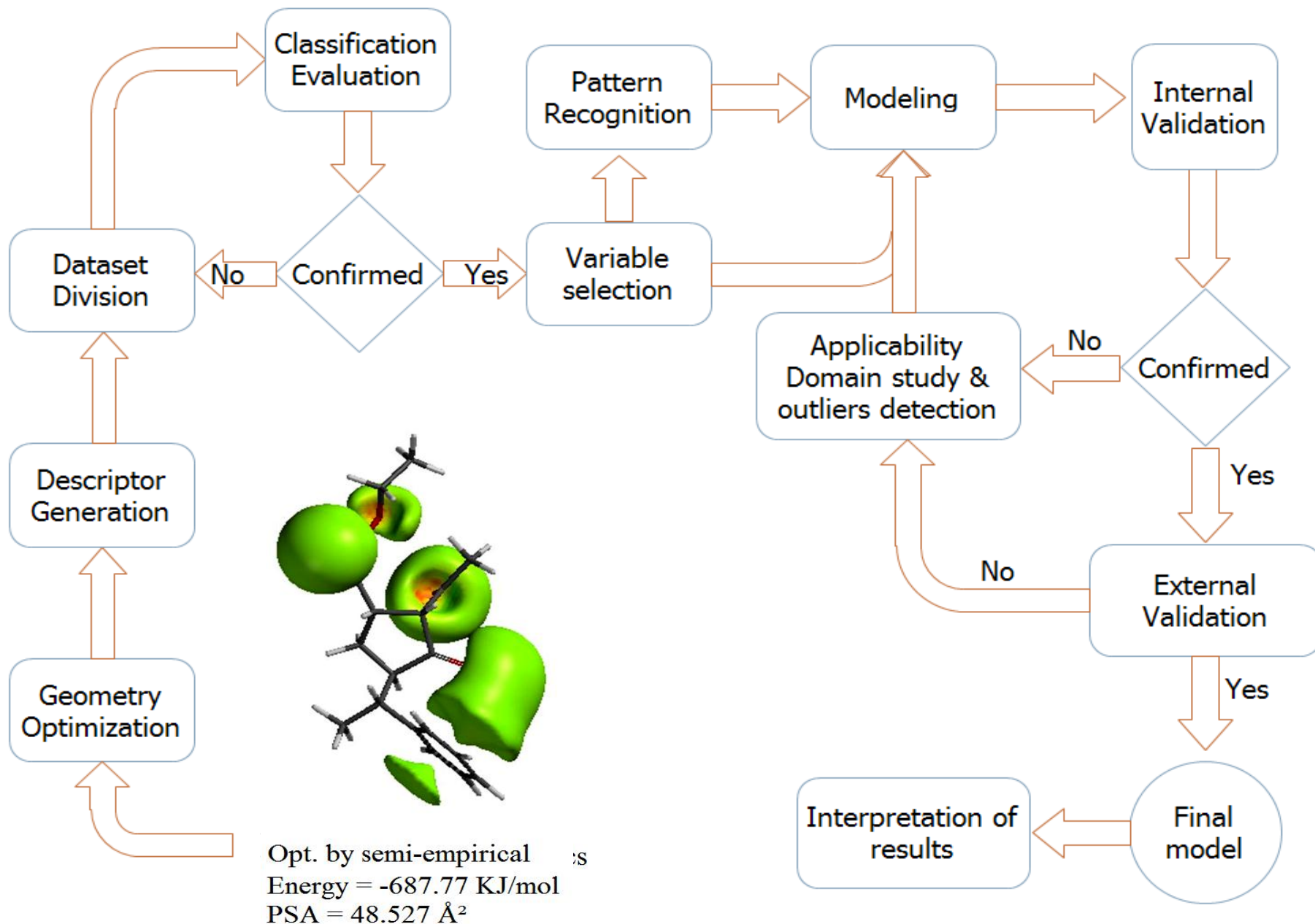
14th International Conference on
Environmental Science and Technology
CEST2015, 3-5 September 2015, Rhodes, Greece

Application of retention time prediction models for suspect and non-target HRMS screening of emerging contaminants in the aquatic environment

PRESENTER: REZA AALIZADEH



Laboratory of Analytical Chemistry,
Department of Chemistry
National and Kapodistrian
University of Athens



The screenshot shows the DRAGON software interface. The window title is "DRAGON". On the left, under "Running the program", there are five buttons: "Calculate descriptors", "Load descriptors", "Load responses", "View descriptors", and "Save descriptors". Below these is a small image of a dog with "EXIT" written on a brick wall. The main area is titled "Descriptor blocks" and has tabs for "0D", "1D", "2D", "3D", and "Others". It lists 22 descriptor categories, each with a question mark icon. At the bottom, there are icons for "Help", "Example Data", "Weightings", "Comments", "WHIM and GETAWAY", "Versions", and "Tips of the day". The text "Milano Chemometrics" is at the bottom center.

Running the program

- Calculate descriptors
- Load descriptors
- Load responses
- View descriptors
- Save descriptors

Descriptor blocks

0D 1D 2D 3D Others

- 1. constitutional descriptors
- 2. topological descriptors
- 3. walk and path counts
- 4. connectivity indices
- 5. information indices
- 6. 2D autocorrelations
- 7. edge adjacency indices
- 8. Burden eigenvalues
- 9. topological charge indices
- 10. eigenvalue-based indices
- 11. Randic molecular profiles
- 12. geometrical descriptors
- 13. RDF descriptors
- 14. 3D-MoRSE descriptors
- 15. WHIM descriptors
- 16. GETAWAY descriptors
- 17. functional group counts
- 18. atom-centred fragments
- 19. charge descriptors
- 20. molecular properties
- 21. 2D binary fingerprints
- 22. 2D frequency fingerprints

Descriptor list | Descriptor search

Help | Example Data | Weightings | Comments | WHIM and GETAWAY | Versions | Tips of the day

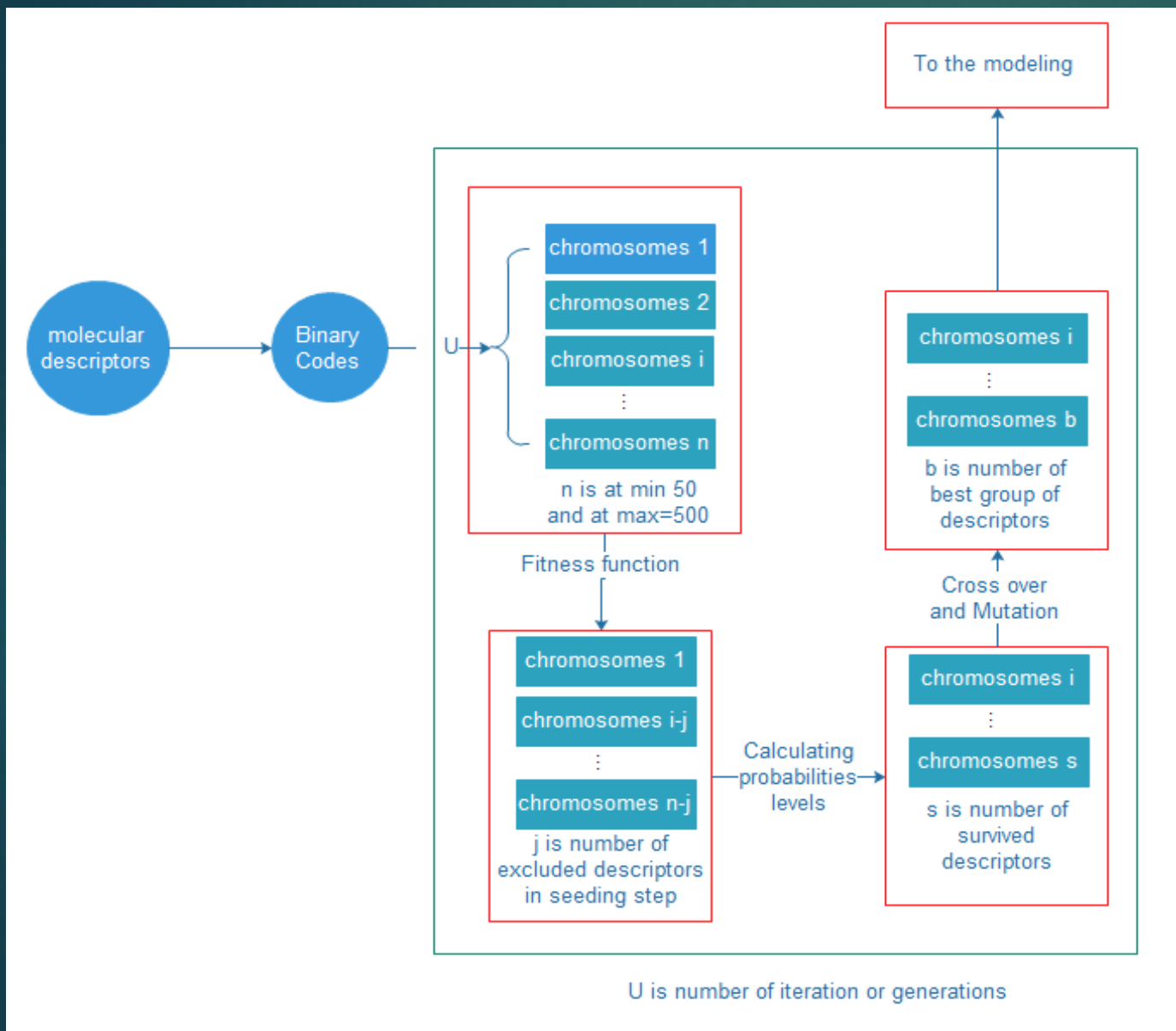
Milano Chemometrics

The screenshot displays the TrAMSclass software window, titled "Classification and Evaluation Toolbox". At the top, there are controls for "load and find optimum N cluster", including "load" and "optimize" buttons, a "Test set percent:" input field set to "20", and an "Export" button. The toolbox is organized into several sections:

- Classification Tools:** Modified k-Medoid-Clustering, Kennard Stone Method, kNN clustering, fuzzy c-means clustering, fuzzy K-means clustering, PCA, Dendrogram, Passing Msg between data, bagged_clustering_kmeans, Distance based optimum design, Bysian Classification, and Affinity Propagation clustering.
- supervised pattern recognition:** Multi method evaluation, Fuzzy kNN-clusering, and PLS-Discriminant Analysis.
- Unsupervised pattern recognition:** Kohenon self organizing map and CAIMAN (Matrix Analysis).
- Outliers detection:** Normalized Mean Distance, William plot, William plot+Cook's distance, Correlation Analysis, and A 4D-plot for origin of outliers.

At the bottom of the window, a note states: "This Toolbox was developed by Reza Aalizadeh at University of Athens."

Molecular Descriptor Selection



Modeling and whole procedure

6

QSAR

Descriptor Calculation
CDK tool Padel BlueDesc

Pre-Treatment
Correlation cut off: 0.9
Load Treatment

unsupervised variable reduction
V-WSP algorithm

Classification
Clustering K-means
Sorting Order for Cluster
Sorting Order for cluster+PCA
PCA
Kennard Stone Method
Modified k-Medoid-Clustering

supervised pattern recognition
Discriminant Analysis Classifying

Distance based optimal design
Precent of Train in Data: 80
Number of Solutions: 1
Run Export

Class & Influence Matrix Analysis
Classifying CAIMAN

Kohonen and CP-ANN
Classifying Kohonen

Variable Selection
GA (Q2LOO Fitness)
GA(LOF Fitness) Stepwise
Best Selection (validation)

Random Frog
Cross Validation Number: 1
The Number of Iterations: 10000
Initial Number of Variables: 2
Weight Samples: No
Random Frog center
GA(Q2LOO)-RandomFrog

MC-Uninformative Variable Elimination
Load Train: 80
The Number of Iterations: 10000
Number of Latent Variables: 20
Run center
GA(Q2LOO)-(MC-UVE)

Modelling
MLR PLS(TOMCAT)
MLR (LOF) ANN
SW-MLR SVM

Outlier detection and Validation
Y-randomization AD-MDI
William Plot AD (full)
Euclidean 3D-Plot (AD)
External Valid Validation

Modelling with simple PLS
Load Train: 80
Latent Variables: 20
Cross Validation: 10 Run
Methods: center
Number of Iterations: 10000
Print Process: Display
Order: Default
Sub-Calib.: 60 center
PLS-DA Predict Save model
Monte Carlo Sampling: 2500


PLS-Discriminative Analysis
Load Train: 80
Latent Variables: 20
Cross Validation: 10 Run
Methods: center
Number of Iterations: 10000
Print Process: Display
Order: Default
Weight: Default
Number of Components: 2
Sub-Calib.: 60 center
PLS-DA Predict Save model
Monte Carlo Sampling: 2500

Survival of the Fittest
Load Train: 80
Maximal Principle to Extract: 18
Fold Number CV Validation: 5
Number of Evolution: 50
Run center
Number of Descriptors: 6
GA(Q2LOO)-(CARS)

Subwindow Permutation Analysis
Load Train: 80
Monte Carlo Sampling: 1000
Fold N. Cross Validation: 3
Sample Population in Each MCS: 15
Number of PLS components: 12
Run center
Sub-Calib.: 75
Number of Descriptors: 6
GA(Q2LOO)-(SPA)

Cross Validation Analysis
Y-randomization PowerMV
Leave-G/O-out Bootstrap

Treat About Save Reset Exit



The screenshot displays the RefTrAMS software interface, which is divided into several functional panels on the left and two overlapping calculator windows on the right.

Main Interface Panels:

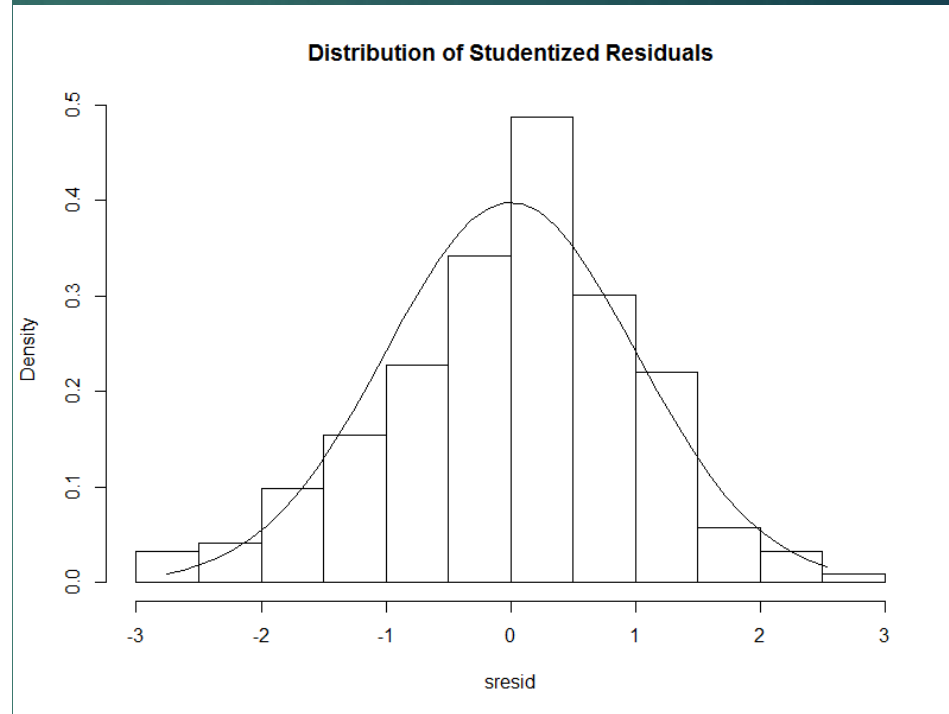
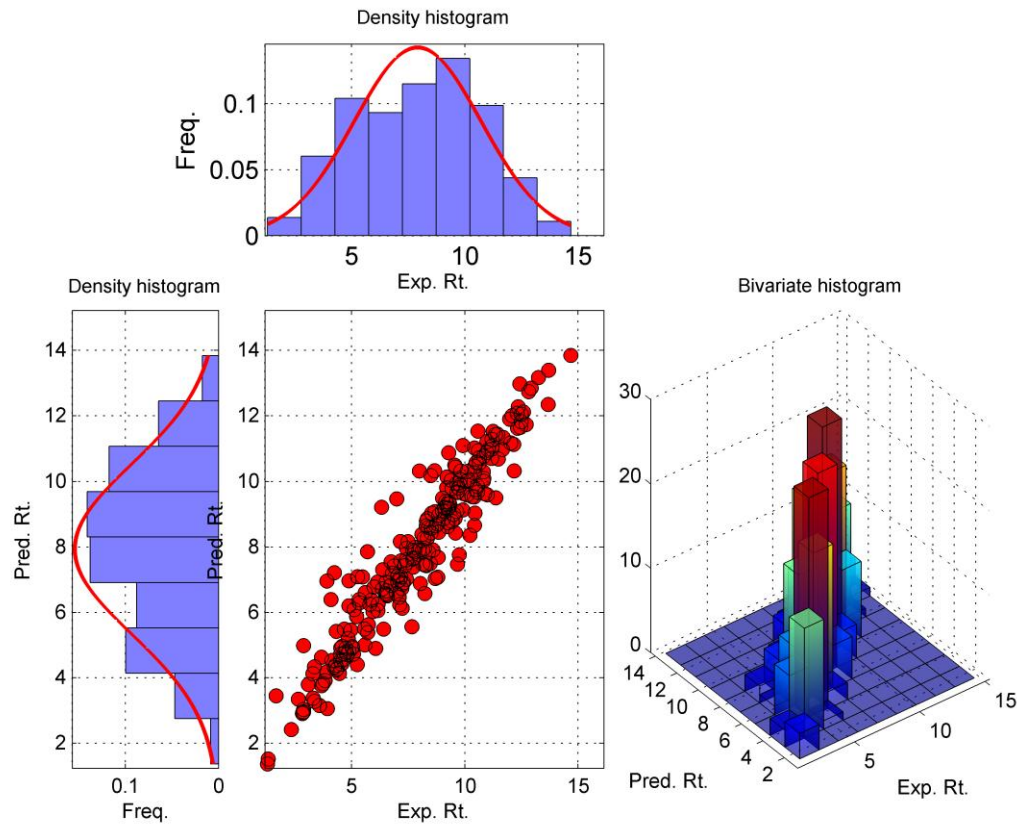
- Prediction of Rt:** Includes radio buttons for (Negative) ESI, (Positive) ESI, RP, and HILIC. Below are buttons for XlogP, LogD, Linear, and Non-Linear.
- Applicability domain study:** Includes buttons for Distribution, PCA, Dendrogram, and OTrAMS. There are checkboxes for Plot OTrAMS and Save OTrAMS.
- Advanced outlier study:** Includes buttons for Load and Mapping. A text input field for Molecule (Example: m25) and a Search button are present.
- Searching Database:** Includes a text input field for Name (Example: Amitrol) and a Search button.
- Footer:** Includes buttons for Manual, Reset, and Exit, followed by the text: "This package is part of RetTrAMS program."

RetLogD - LogD Calculator Window:

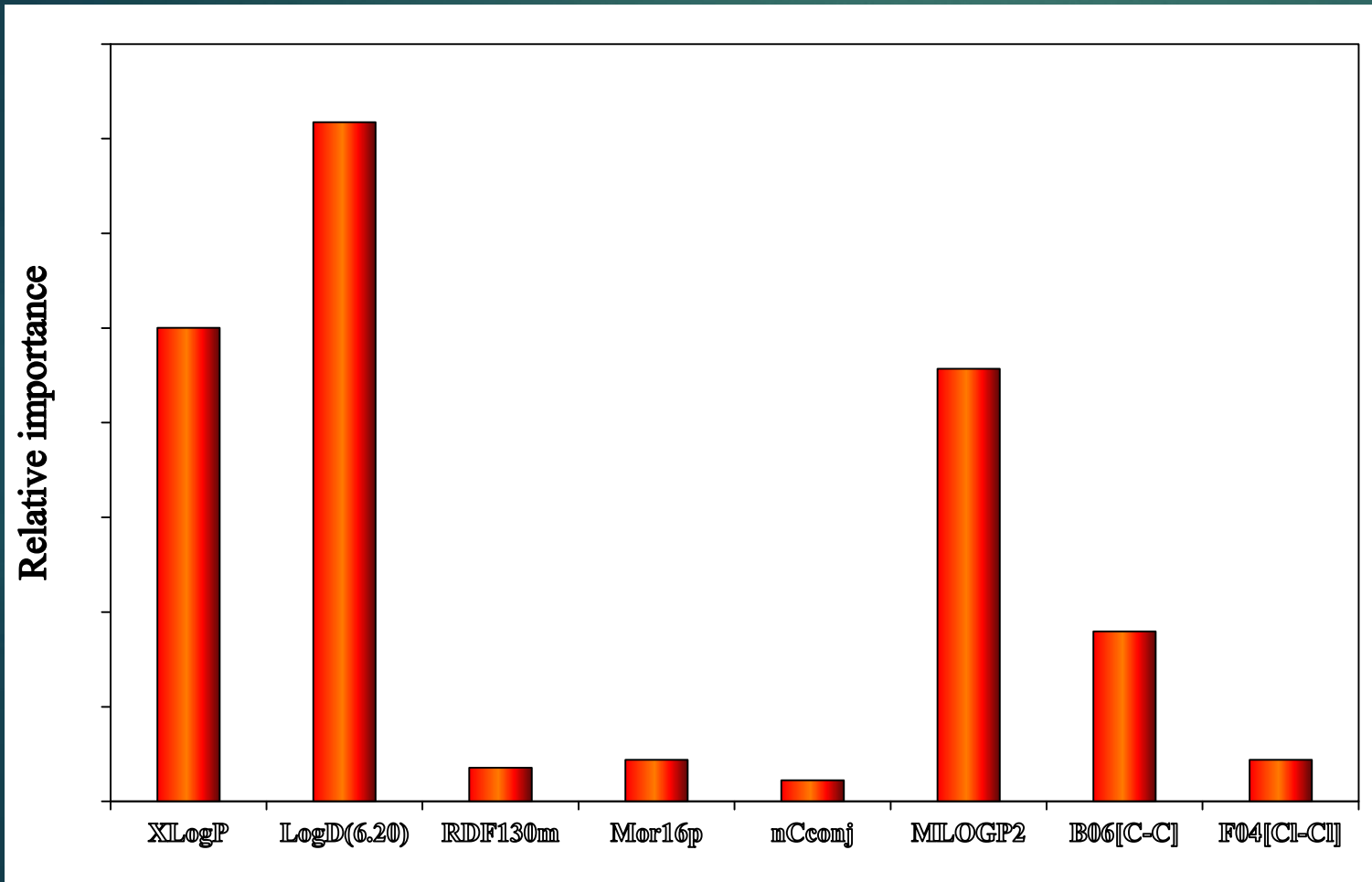
- Buttons: Download Online, RetLogD, X
- Options: Assign pH (Negative) (selected), Assign pH (Positive)
- Input: pH: Ex.: 3.6
- Options: LogD (selected), LogP
- Buttons: load, Calculate, Cancel
- Text: "LogD Calculator is part of RetTrams program. LogD calculation is performed by ChemAxon program."

XLOGPcal - XlogP Calculator Window:

- Buttons: XLOGPcal, X
- Input: SMILES: Ex.: NC1C=CC(Cl)=CC=1C(=O)C1C=CC=CC=1
- Buttons: Calculate, Padel, Cancel, Install Packages
- Text: "XlogP Calculator is part of RetTrams program."

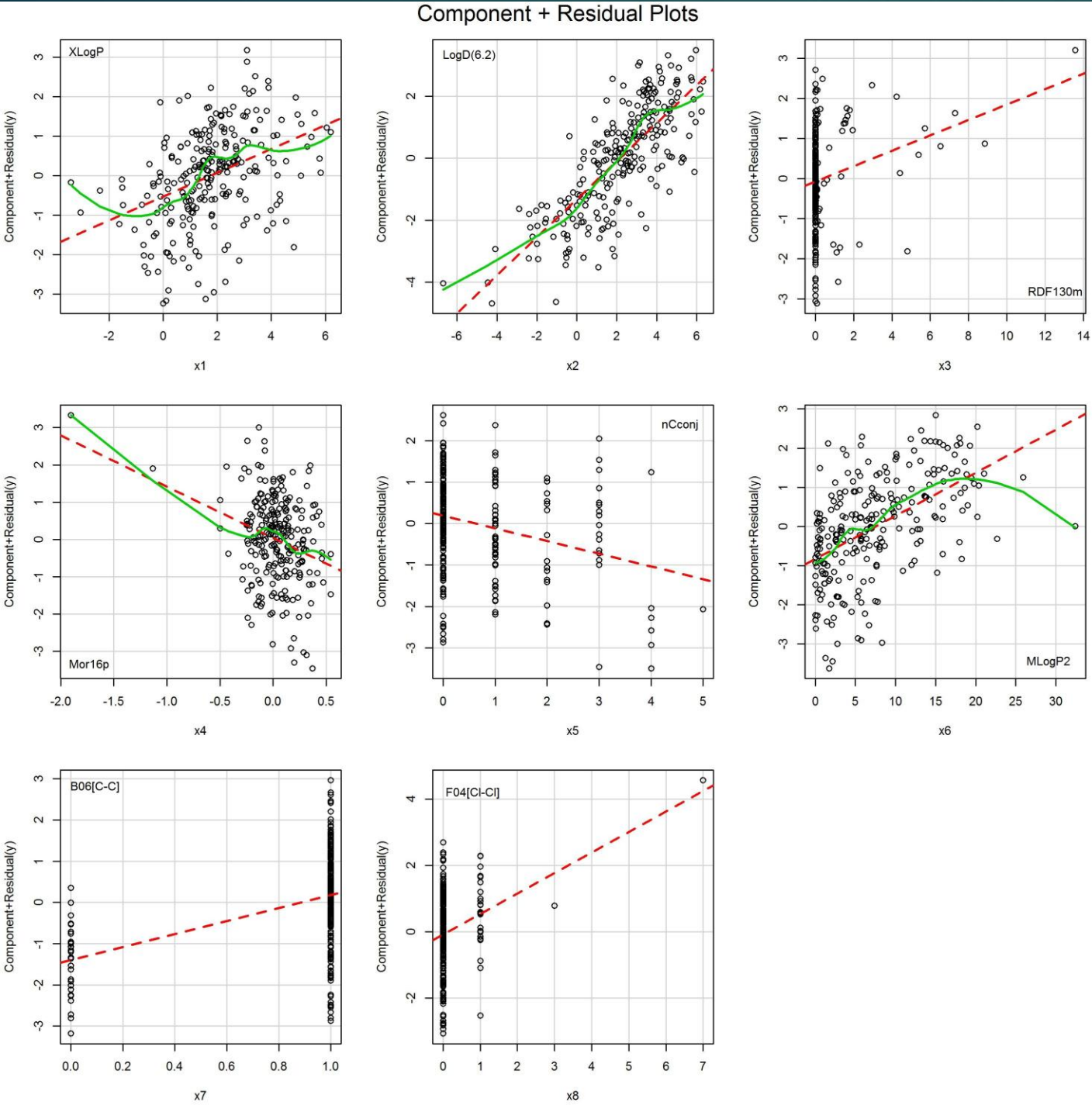


	Training			Test		
	R2	RMSE	F	R2	RMSE	F
MLR	0.841	1.116	156.651	0.844	1.103	36.012
SVM	0.911	0.842	278.615	0.857	1.052	35.009

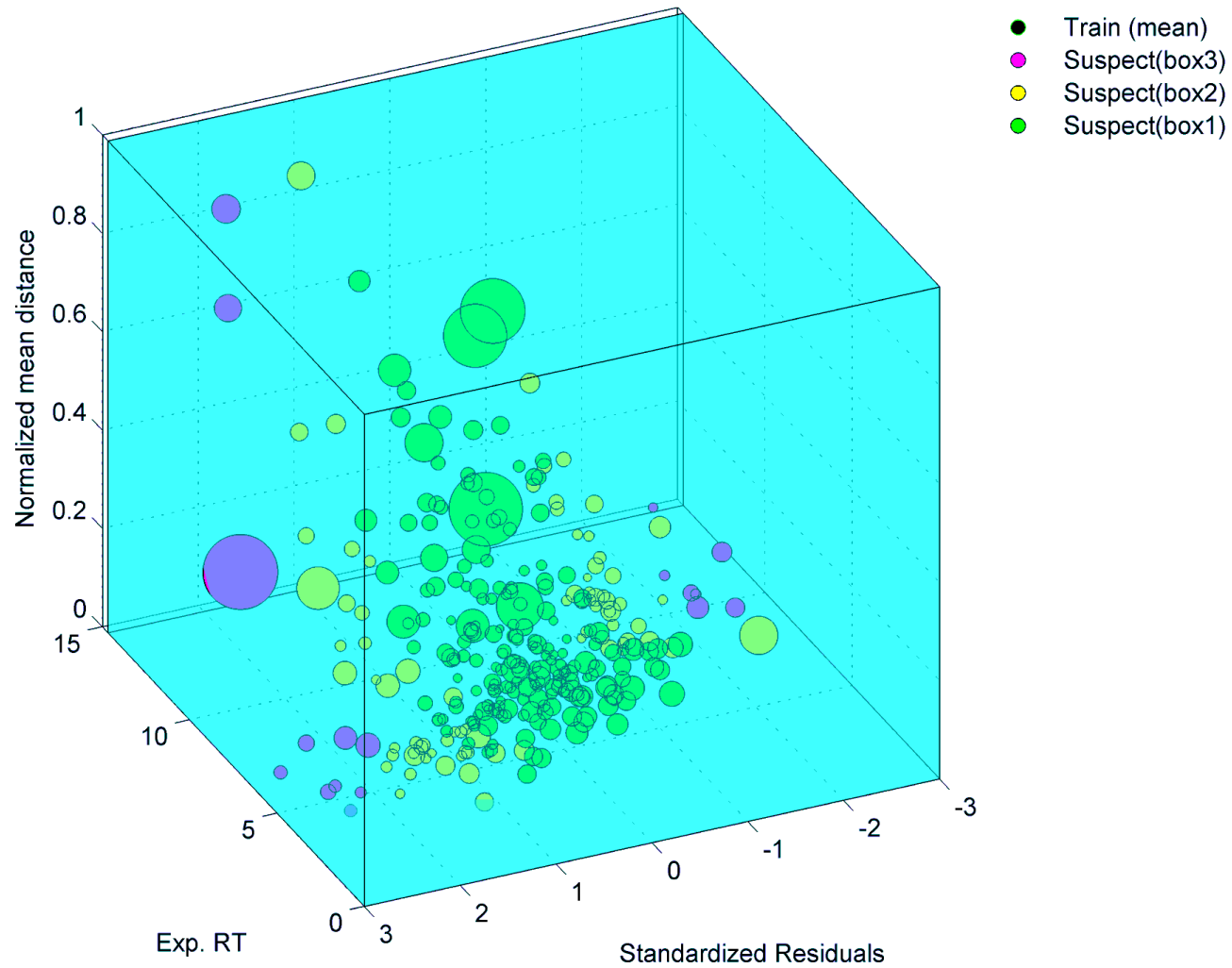


Negative Ionization

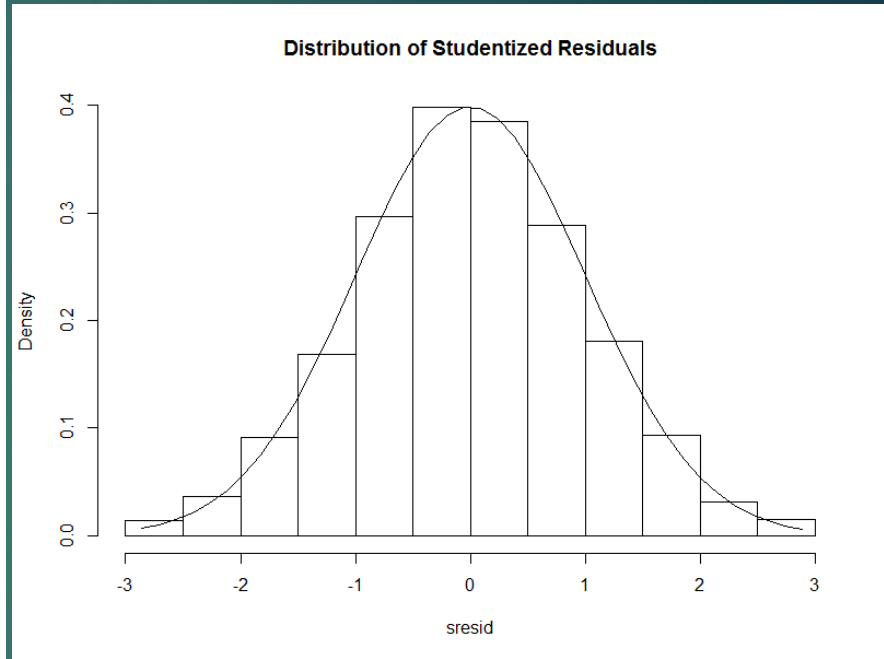
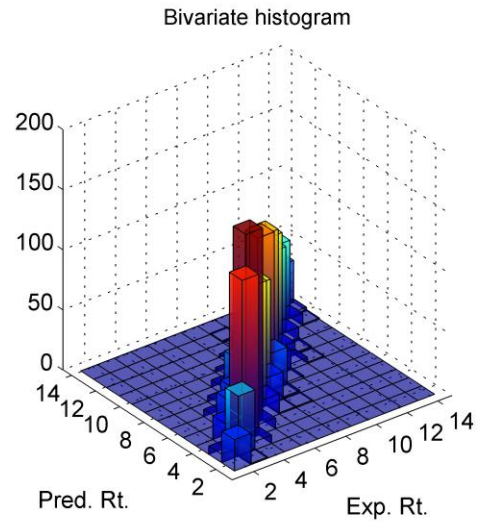
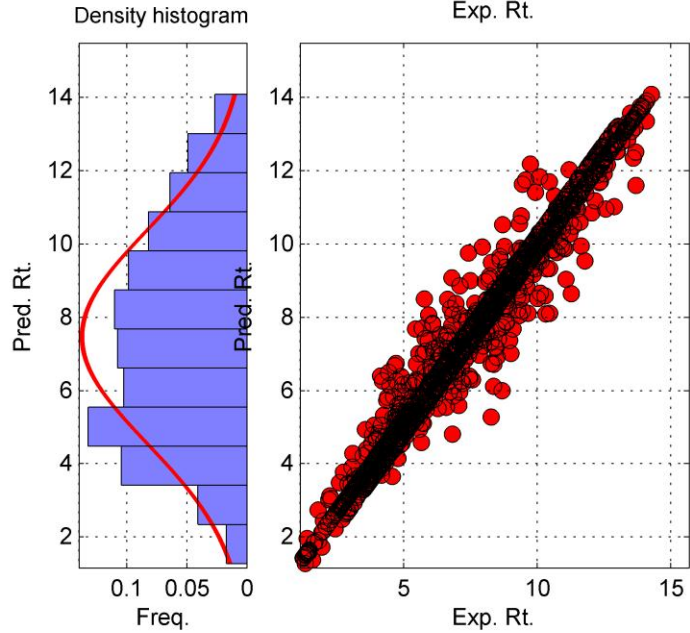
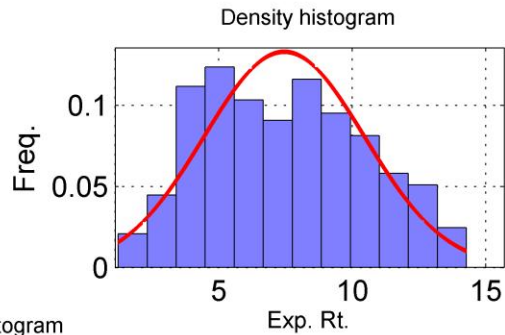
XLogP	+
LogD (pH=6.2)	+
RDF130m	+
Mor16p	-
nCconj	-
MLOGP2	+
B06(C-C)	+
F04(Cl-Cl)	+



RP_(-)ESI



$$\text{Standardized Residuals} = \frac{r(i)}{\left[\text{sqrt} \left(\frac{r' * r}{n - p} * (1 - h(i)) \right) \right]}$$



Training

R2 RMSE F

MLR 0.844 1.187 1569.088

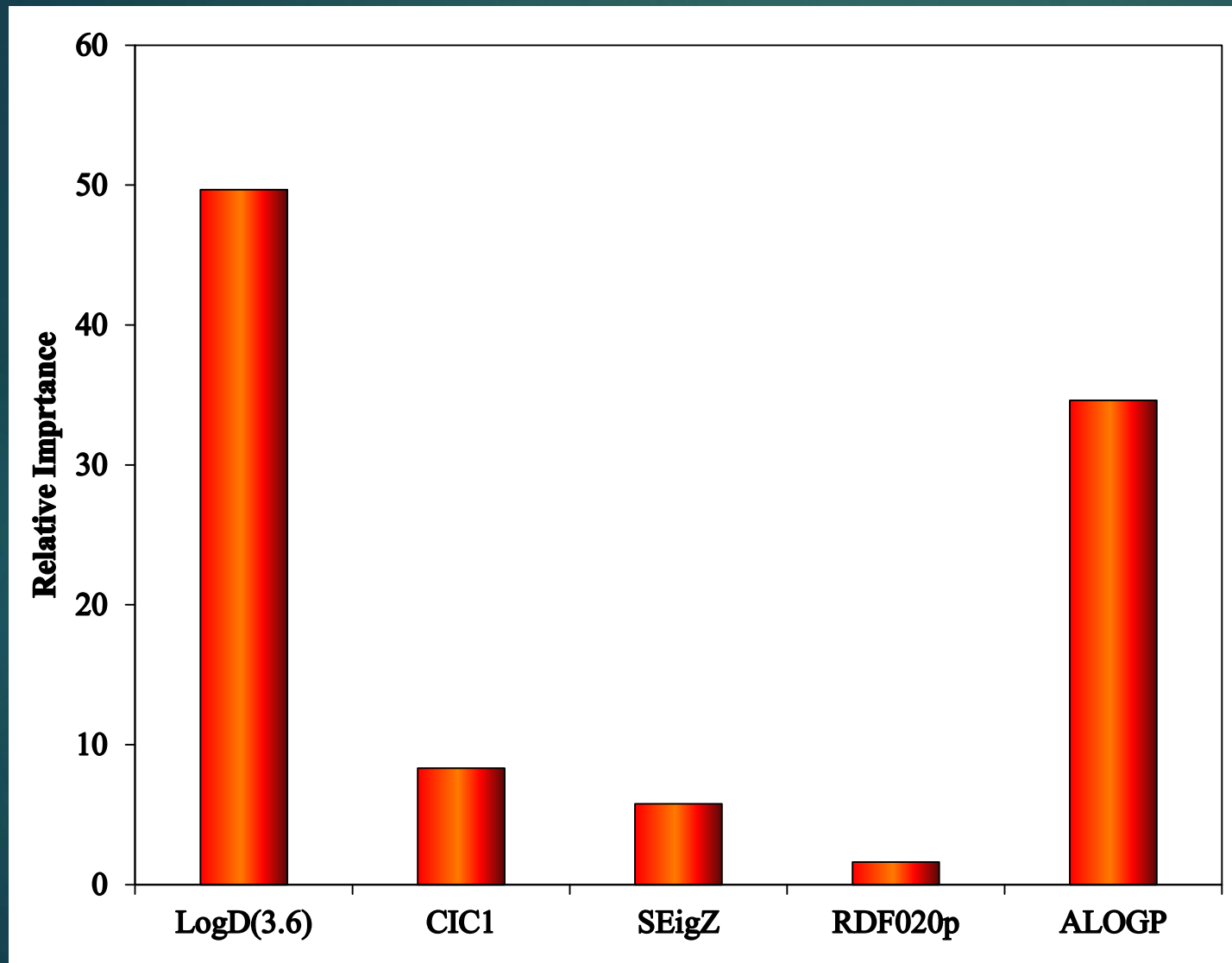
SVM 0.971 0.514 9281.657

Test

R2 RMSE F

0.848 1.179 433.091

0.881 1.042 574.810

**Positive Ionization**

LogD (pH=3.6) | +

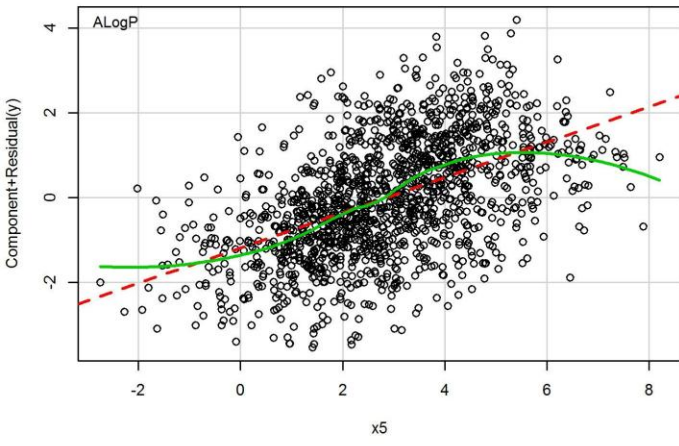
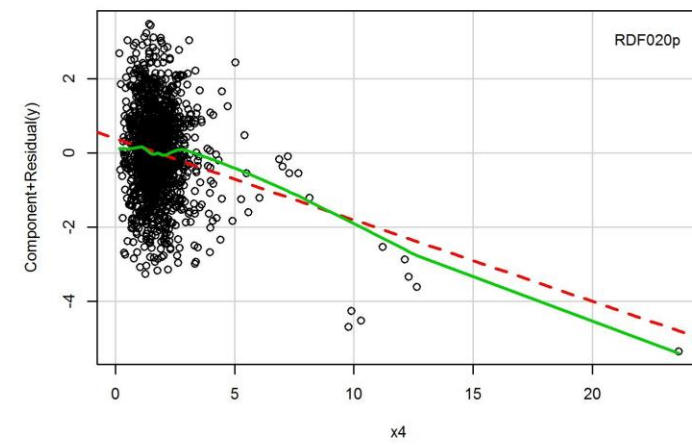
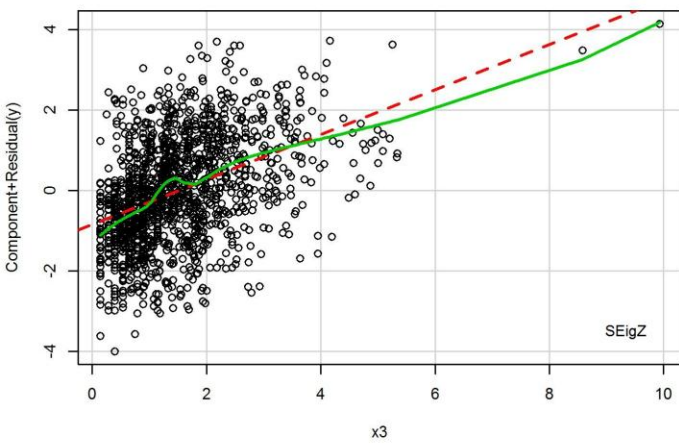
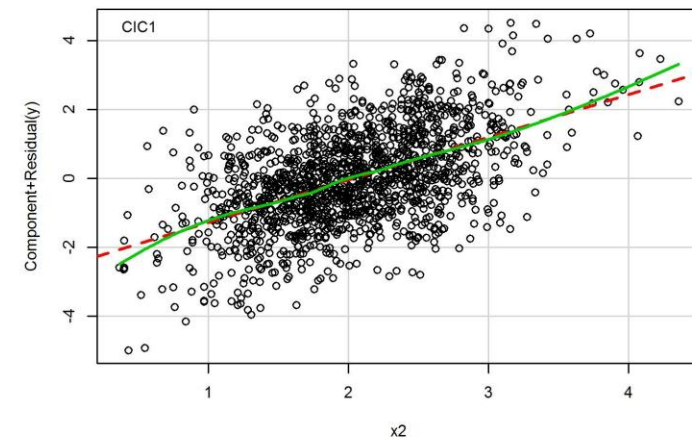
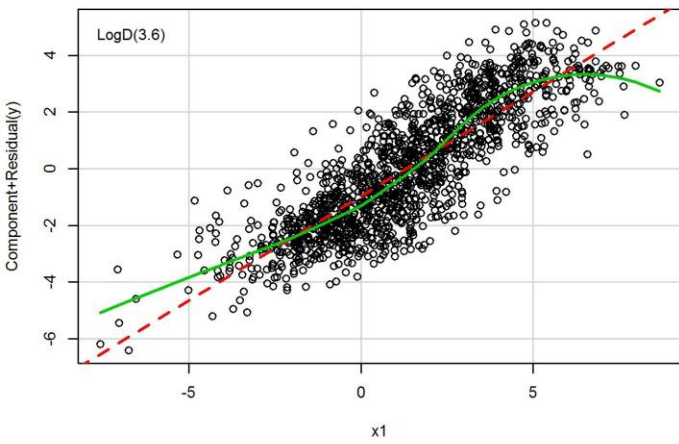
CIC1 | +

SEigZ | +

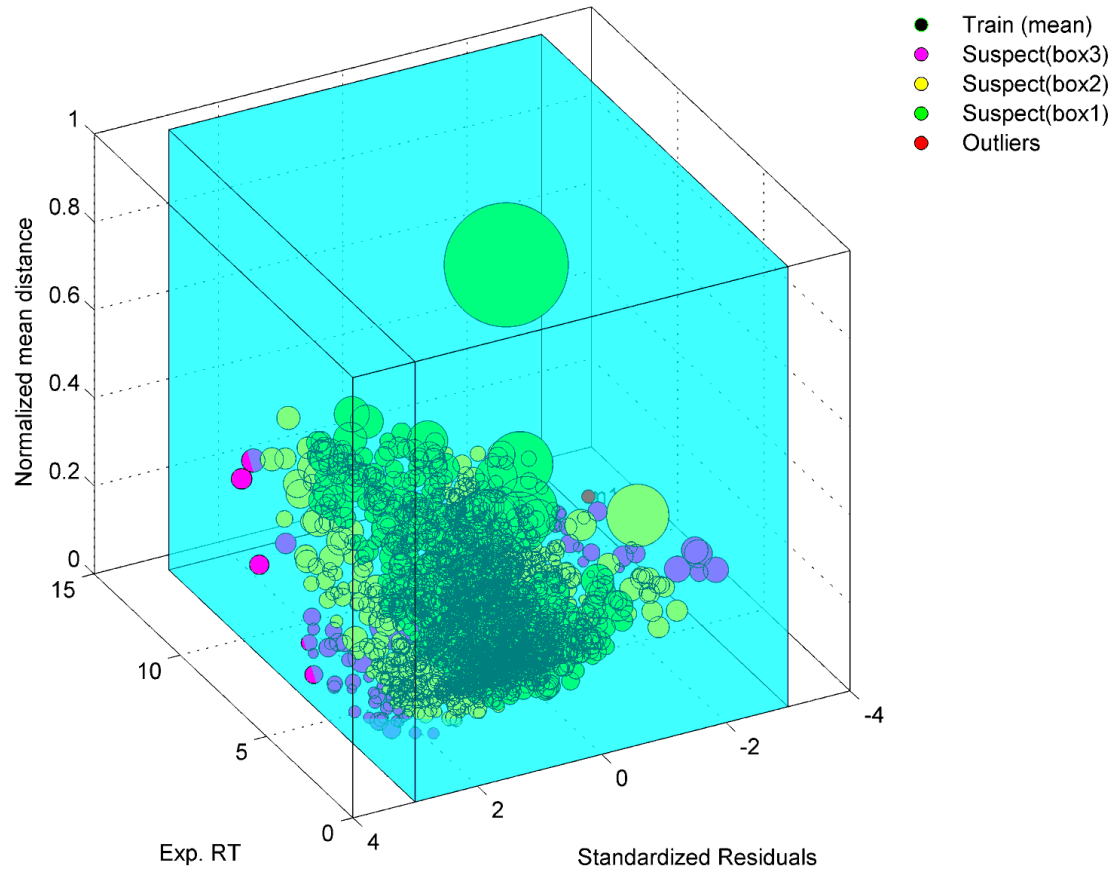
RDF020p | -

AlogP | +

Component + Residual Plots



RP_(+)ESI



Number of compounds inside each box Percent of compounds inside each box

box1	1259	69
box2	483	26
box3	87	5
box4	1	0

Protocols

16

- To accept or reject a suspect structure, perform RetTrAMS and OTrAMS → locate the points in boxes
- If the suspect compound locates in box 1 and box 2 → the suspect structure is accepted.
- If the suspect compound locates in box 3 → further validation should be done.
- If the suspect compound locates in box 4 → the suspect structure is rejected.

Acknowledgments

17

Nikolaos S. Thomaidis

Anna Bletsou

Pablo Gago Ferrero

Thanks for your attention!



This research has been co-financed by the European Union and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – ARISTEIA 624 (TREMPOOL project).