



Predicting retention time for suspect and non-target HILIC-LC-HRMS screening of emerging contaminants in the aquatic environment

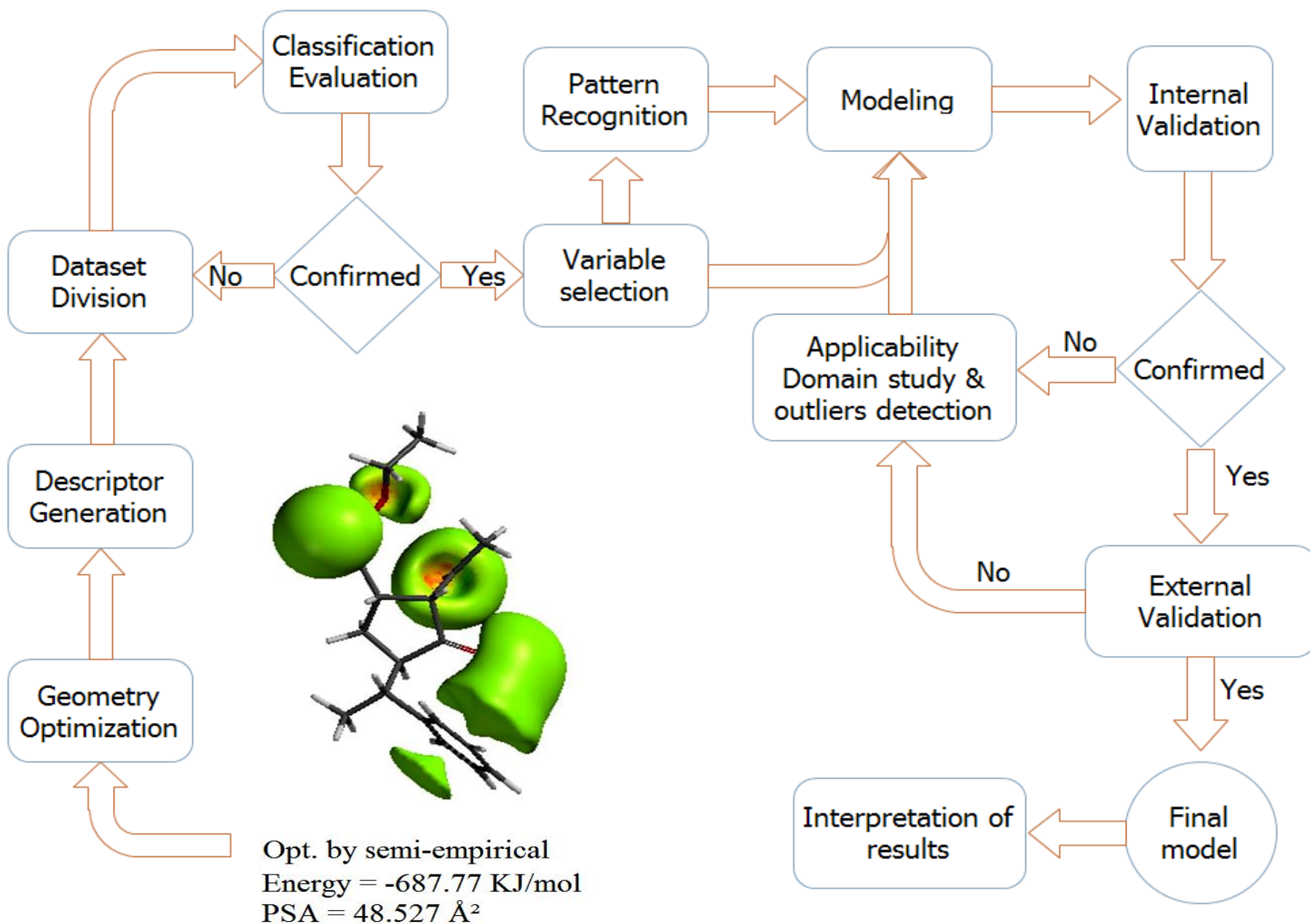
PRESENTER: REZA AALIZADEH



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

Workflow for performing a QSPR modeling

2



Descriptor calculation

3

The screenshot shows the DRAGON software interface. The window title is "DRAGON". On the left, under "Running the program", there are buttons for "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. On the right side, there are links for "About", "Address", "Handbook", "Tools", and "Thanks", each with an icon. At the bottom, there are links for "Help", "Example Data", "Weightings", "Comments", "WHIM and GETAWAY", "Versions", and "Tips of the day", each with an icon. 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

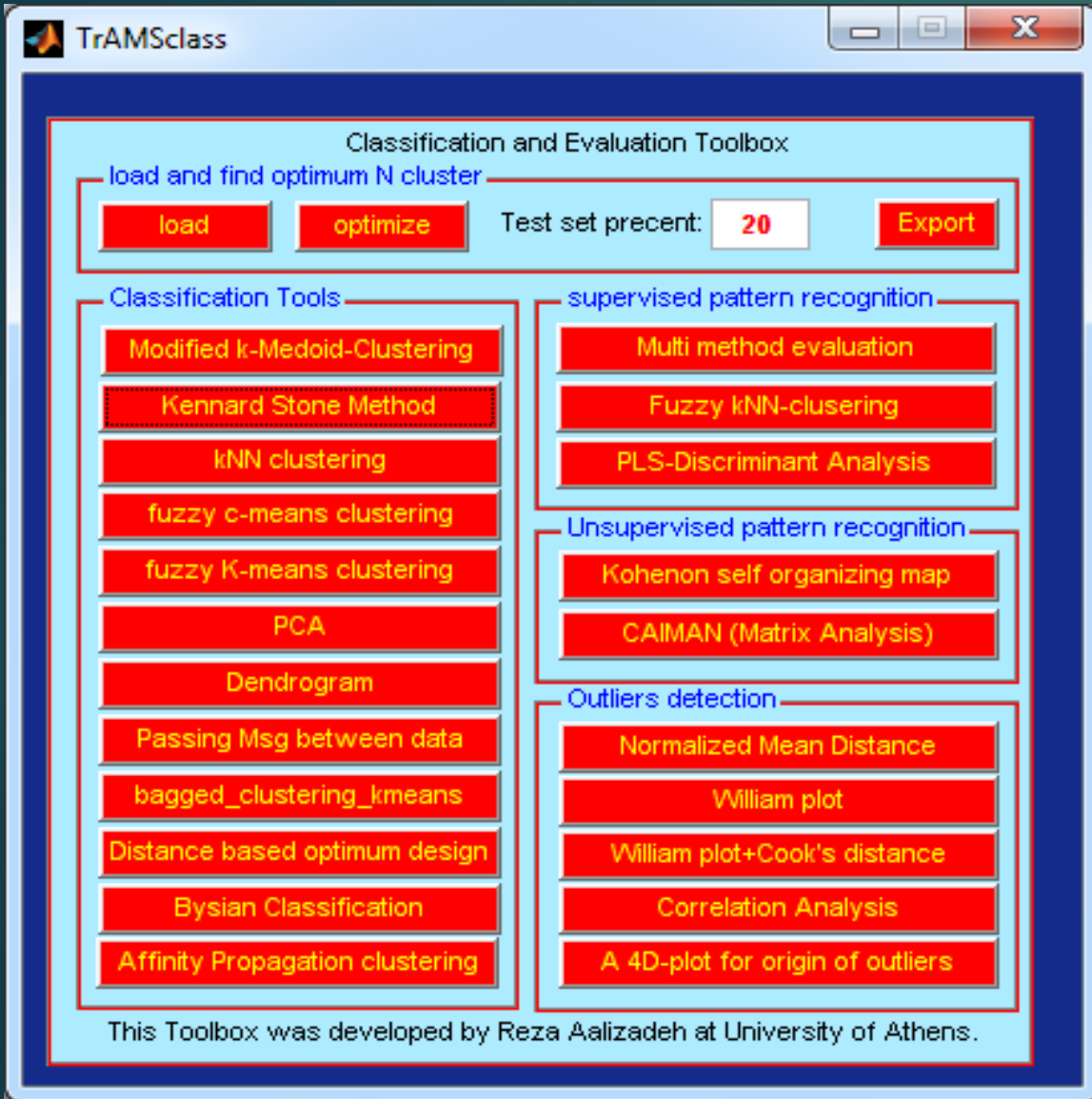
1. constitutional descriptors	2. topological descriptors
3. walk and path counts	4. connectivity indices
5. information indices	6. 2D autocorrelations
7. edge adjacency indices	8. Burden eigenvalues
9. topological charge indices	10. eigenvalue-based indices
11. Randic molecular profiles	12. geometrical 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

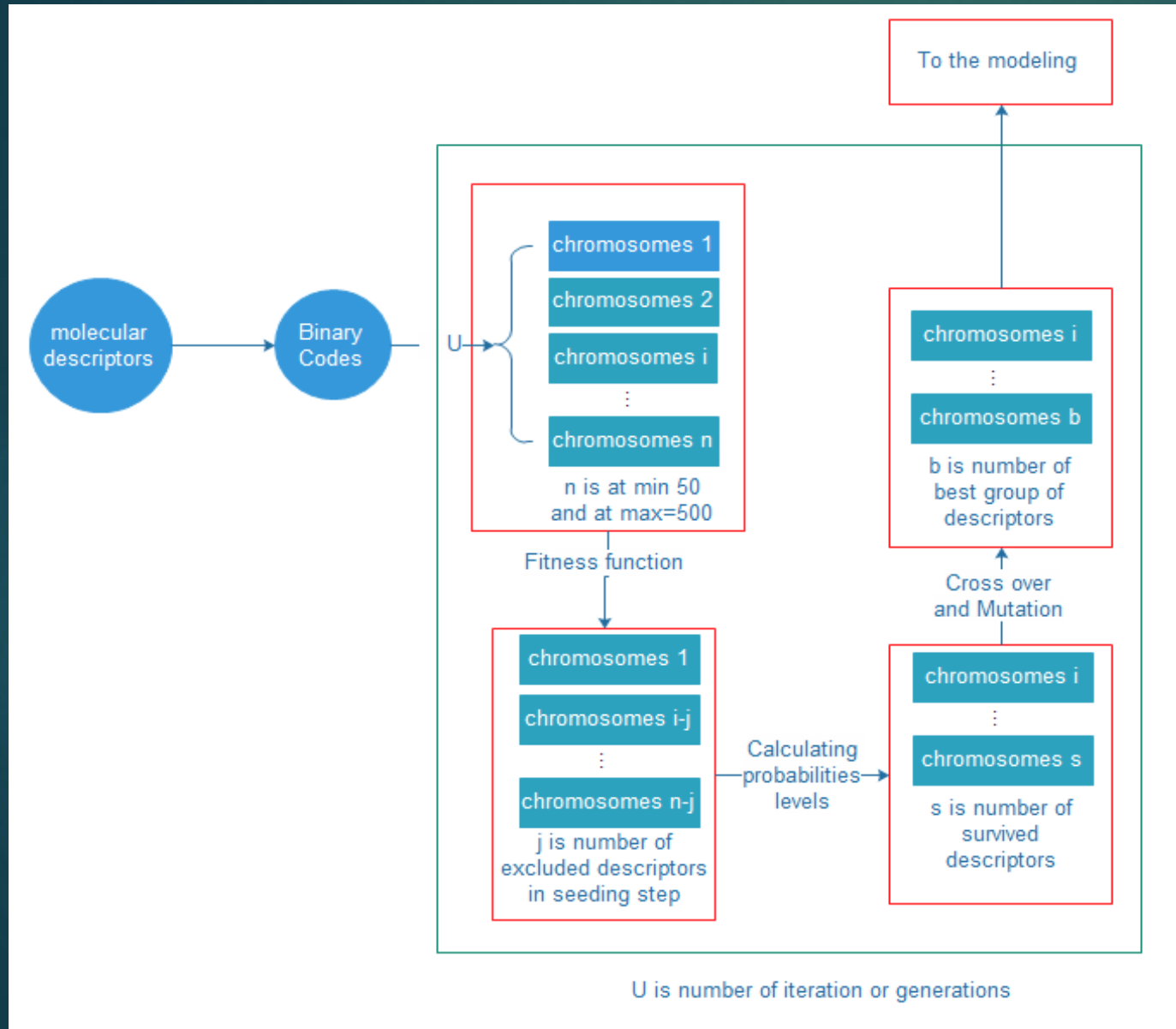
Milano Chemometrics

Dataset Division & Accuracy Assessment



Molecular Descriptor Selection

5



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 RetTrAMS software interface, which is organized into several functional panels on the left and two calculation windows on the right.

Main Interface Panels:

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

RetLogD - LogD Calculator Window:

- Buttons: load, Calculate, Cancel
- Options: Assign pH (Negative) (selected), Assign pH (Positive)
- LogD/LogP: LogD (selected), LogP
- Text: "LogD Calculator is part of RetTrams program. LogD calculation is performed by ChemAxon program."

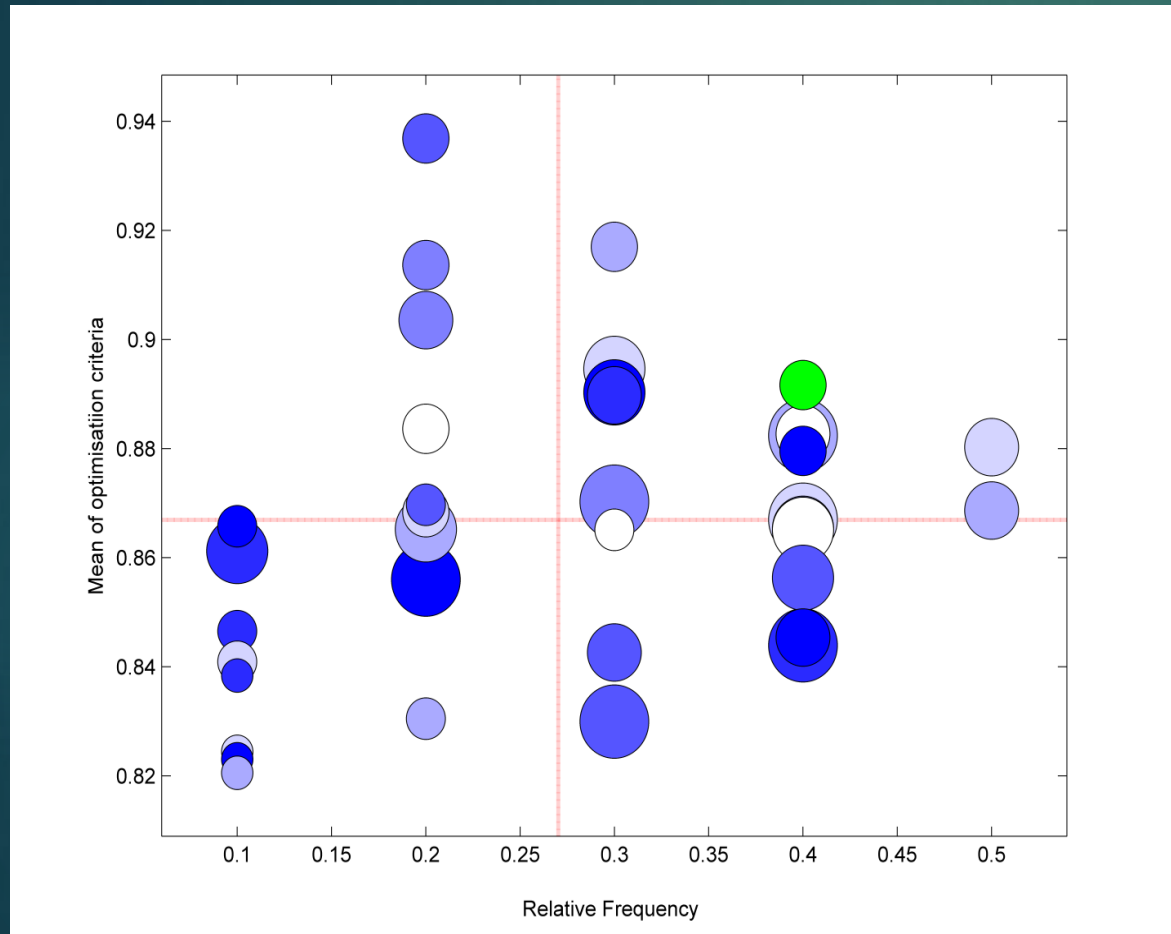
XLOGPcal - XlogP Calculator Window:

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

Results and discussion



'neurons'	'epochs'	'frequency'	'optimization criterion'
12x12	100	0.5	0.88028
12x12	150	0.5	0.86865
10x10	300	0.4	0.89164



fitting

error rate: 0.067
non-error rate: 0.933
accuracy: 0.956

plot class profiles
plot ROC cruves
view confusion matrix
view calculated class
view class weights

class	Spec	Sens	Prec
1	0.99	0.89	0.94
2	0.99	0.93	0.95
3	0.94	0.99	0.96

cross-validation

error rate: 0.110
non-error rate: 0.890
accuracy: 0.924

view confusion matrix
view predicted class
view class weights

class	Spec	Sens	Prec
1	0.97	0.87	0.87
2	0.97	0.84	0.87
3	0.93	0.97	0.96

prediction on external samples

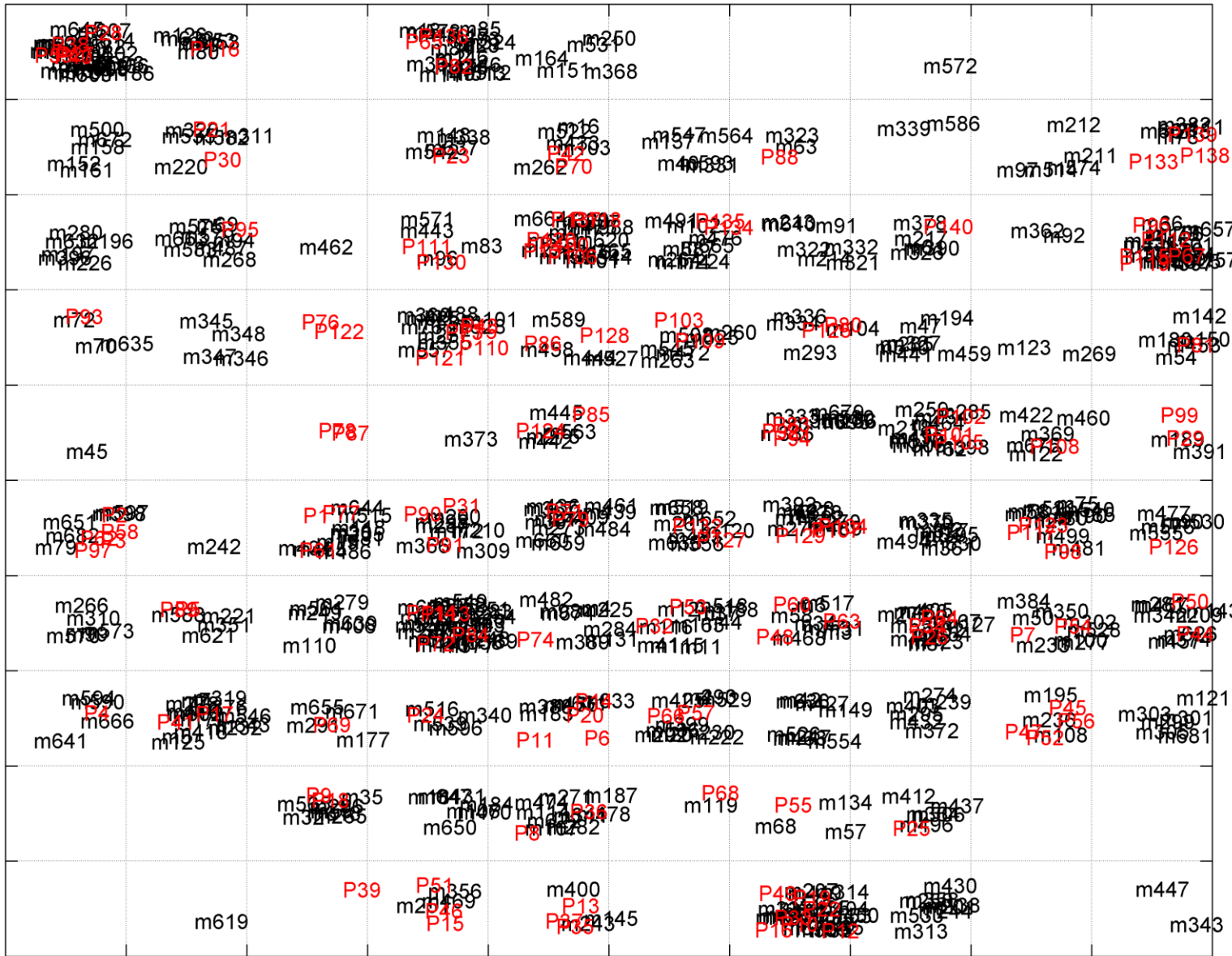
error rate: 0.177
non-error rate: 0.823
accuracy: 0.823

view samples in top map
view confusion matrix
view predicted class
view class weights

class	Spec	Sens	Prec
1	0.94	0.80	0.77
2	0.97	0.73	0.84
3	0.87	0.94	0.92

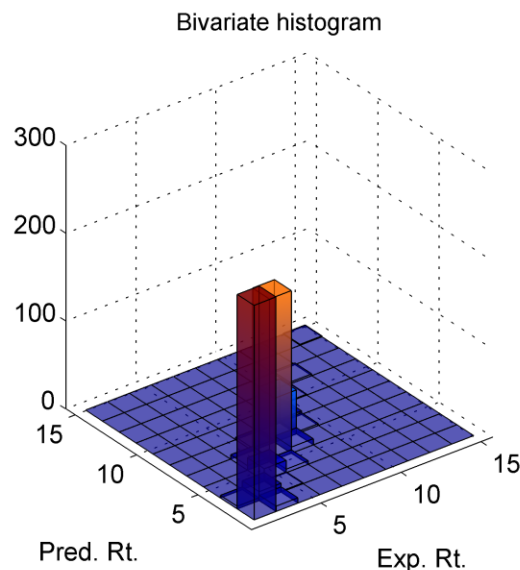
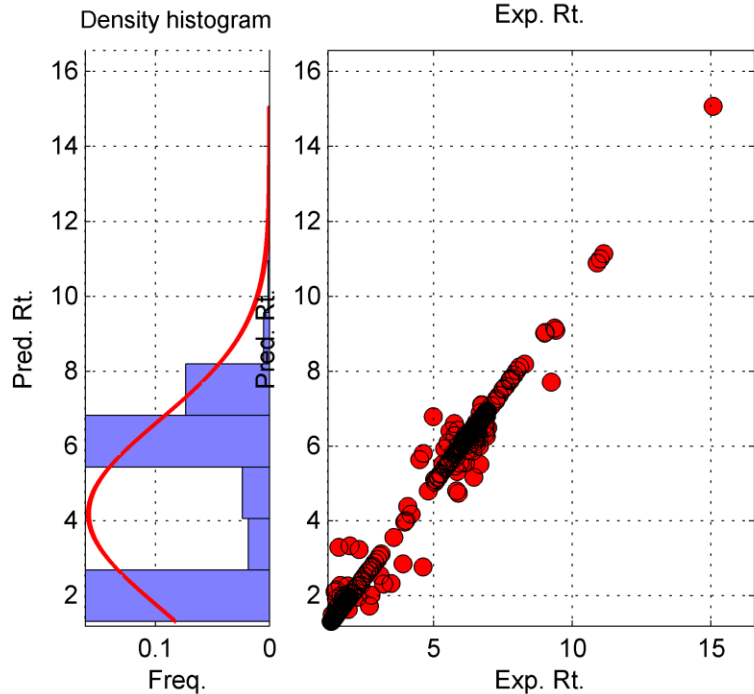
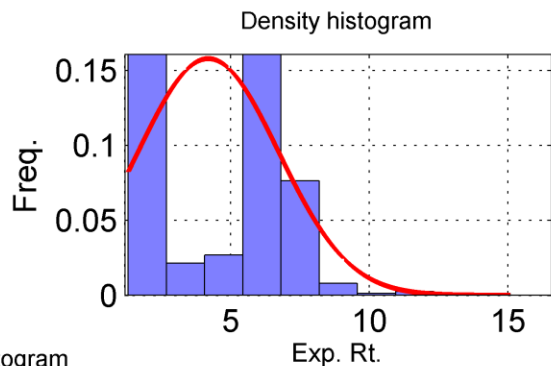
Optimization of Self-Organizing Maps (SOMs)

Map

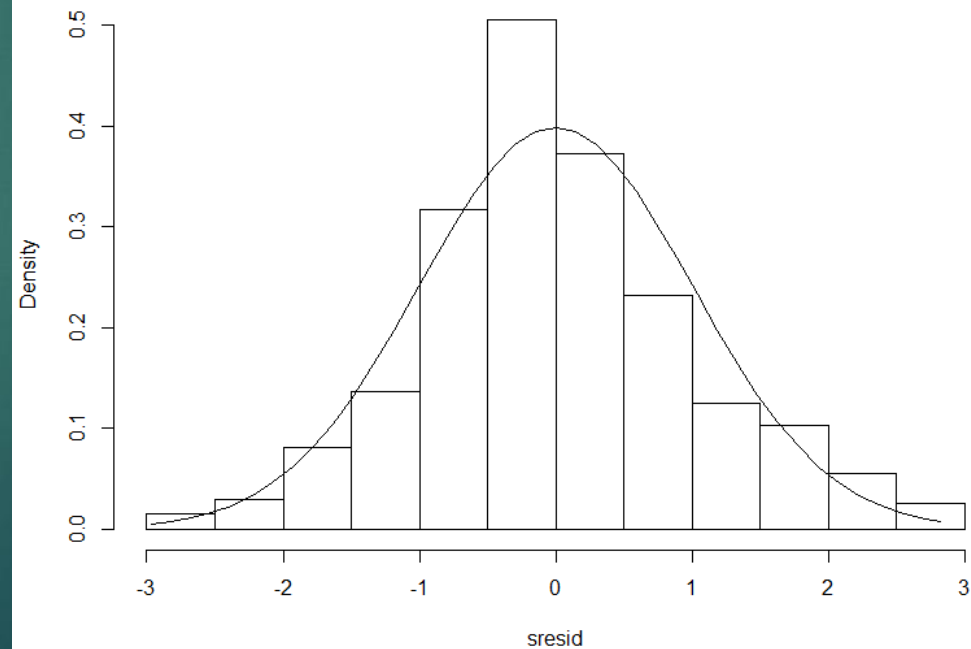


HILIC_(+)ESI

11



Distribution of Studentized Residuals



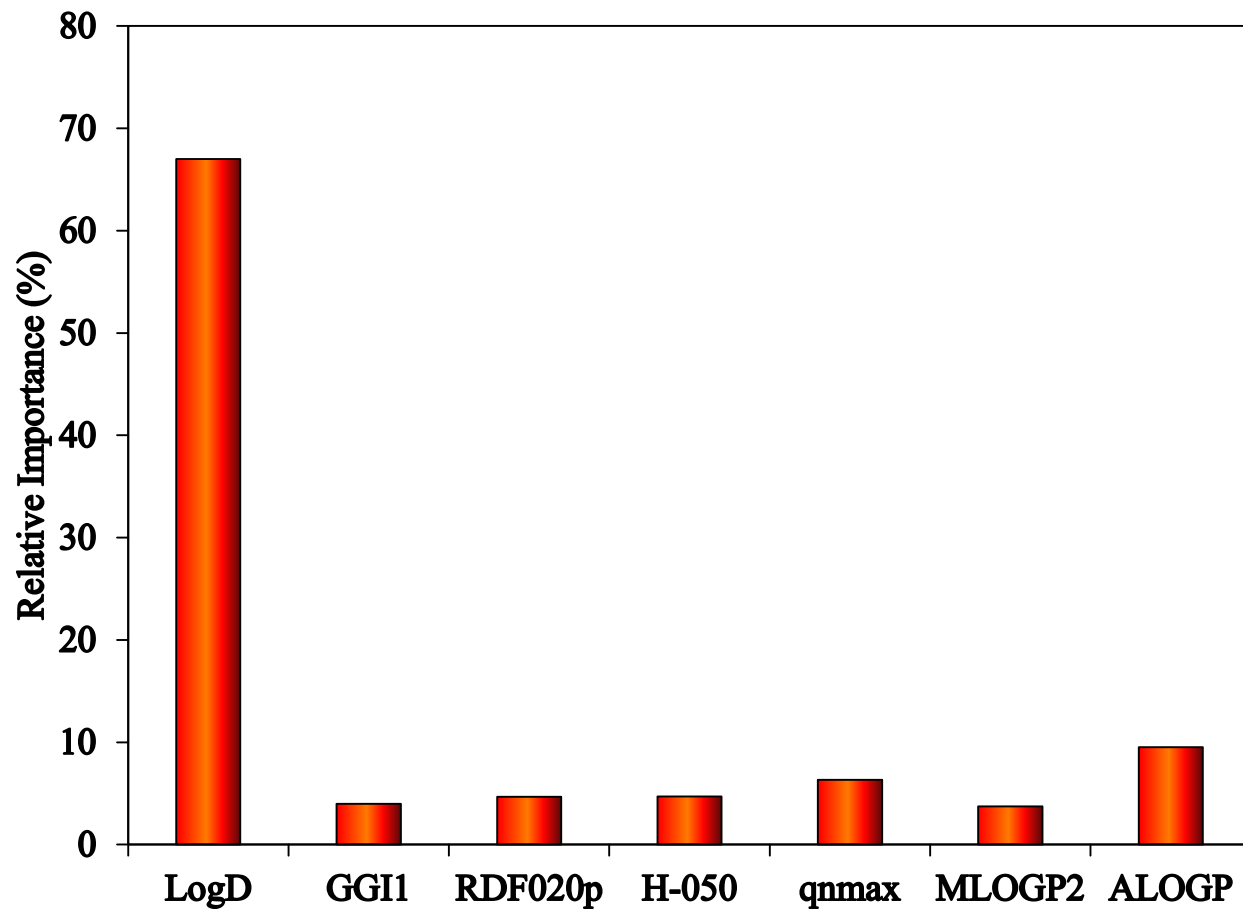
Training

Test

	R2	RMSE	F	R2	RMSE	F
MLR	0.896	0.814	658.174	0.865	0.912	132.049
SVM	0.989	0.260	7058.252	0.908	0.747	191.711

Variable Importance

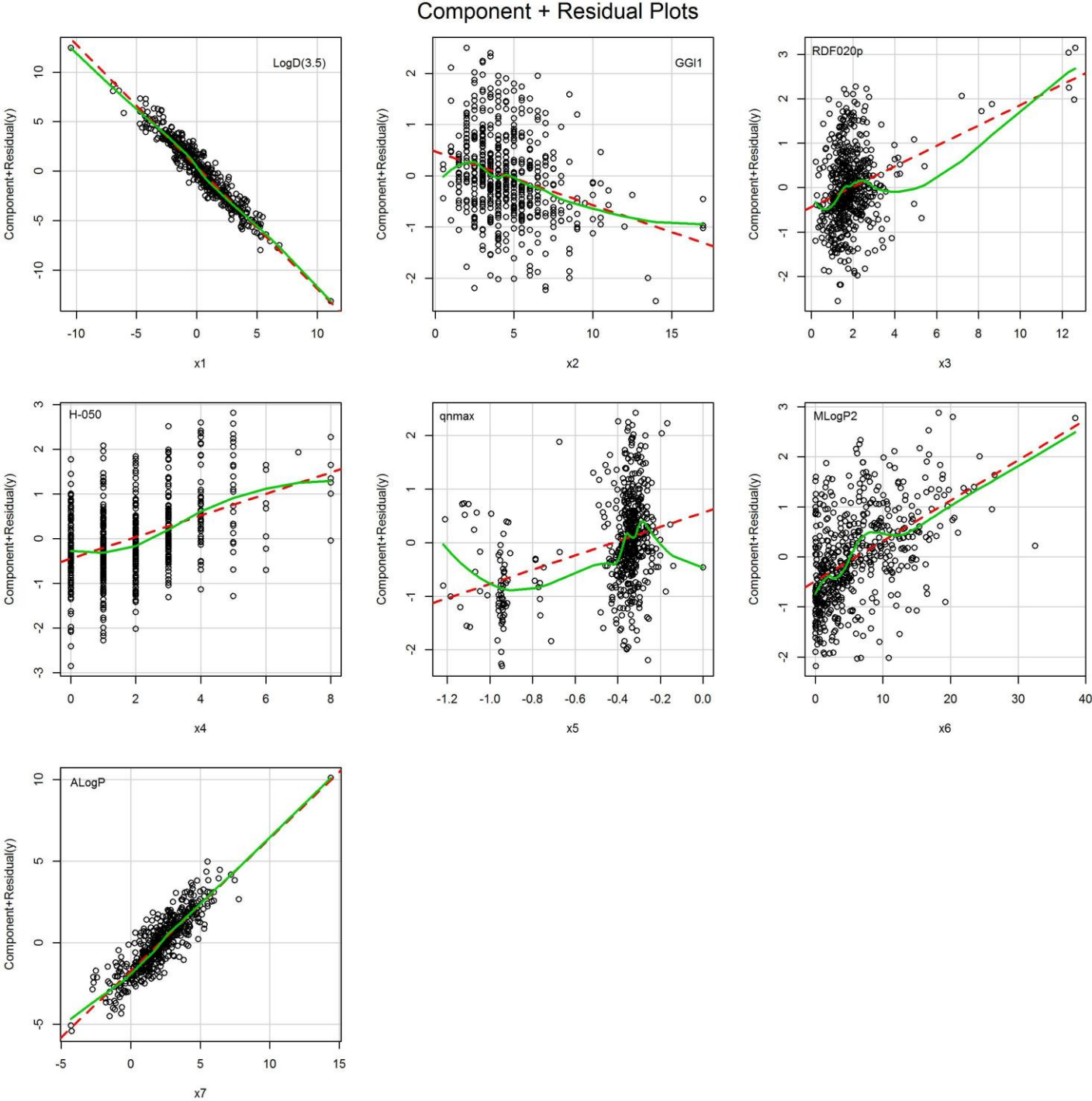
12



Positive Ionization	
LogD	-
GGI1	-
RDF020p	+
H-050	+
qnmax	+
MLOGP2	+
AlogP	+

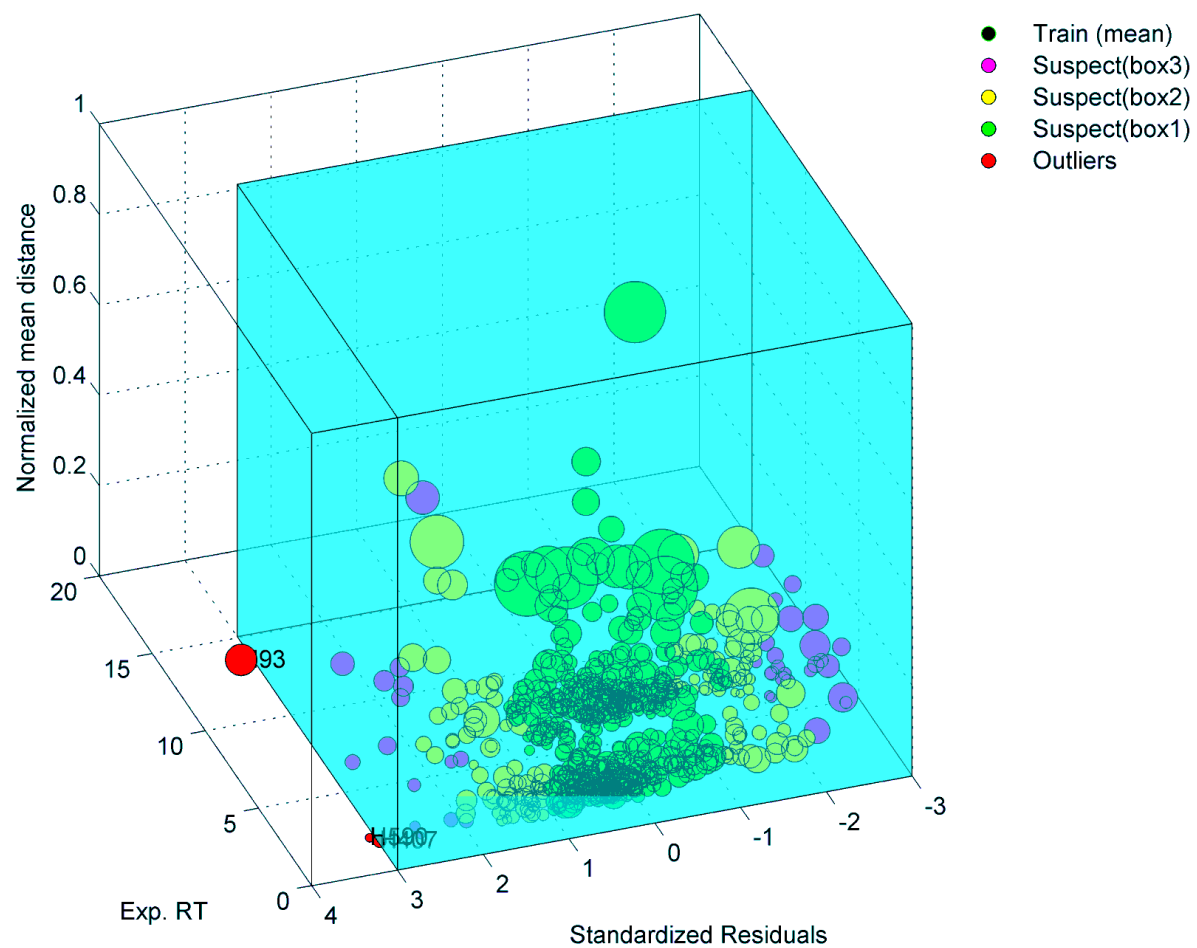
Variables effects over Rt

13



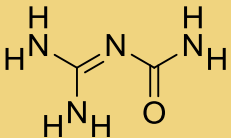
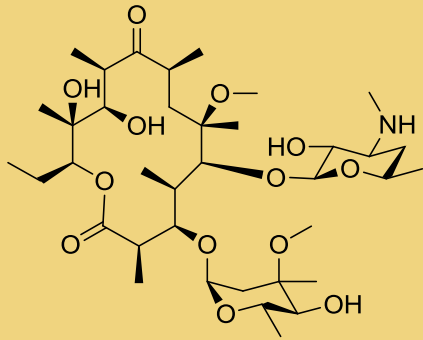
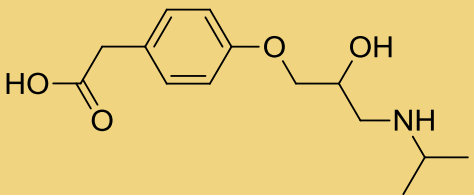
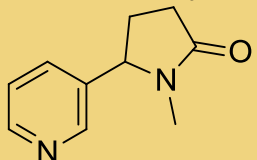
Applicability domain

14



	Number of compounds inside each box	Percent of compounds inside each box
box1	484	71
box2	153	22
box3	42	6
box4	3	0

Suspect screening

No.	Name, Structure and Formula	Parent	Elution Pattern recognition tool/ Exp. t_R	Exp. t_R	Pred. t_R
S13	Guanylurea, $C_2H_6N_4O$  N-desmethyl clarithromycin $C_{37}H_{67}NO_{13}$	Metformin	D(+) Mannose: 7.75 Melamine: 6.7	6.8	6.64
S3		Clarithromycin	Erythromycin: 6.51 Aliskiren: 6.19 Roxithromycin: 5.9 Tylosin: 6.55	6.06	5.95
S1	Atenolol acid, $C_{14}H_{21}NO_4$ 	Atenolol	Amlodipine: 6.49 Ranitidine: 6.5	6.73	6.69
S6	Cotinine, $C_{10}H_{12}N_2O$ 	Nicotine	4-Acetamidoantipyrine: 2.38 Paraxanthine: 2.26	2.42	2.45

Protocols

- 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

Nikolaos S. Thomaidis

Alexandros Markatis

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 (TREMOPOL project).