NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES

DEPARTMENT OF CHEMISTRY

DOCTORAL THESIS

Development of quantitative structure property relationships to support non-target LC-HRMS screening

REZA AALIZADEH

MSc. CHEMIST

ATHENS 2019

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΧΗΜΕΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ανάπτυξη νέων μοντέελων συσχέτισης δομής – ιδιοτήτων για την υποστήριξη της μη στοχευμένης ανάλυσης περιβαλλοντικών δειγμάτων με φασματομετρία μάζας υψηλής διακριτικής ικανότητας

REZA AALIZADEH

MSc ΧΗΜΙΚΟΣ

ATHENS 2019

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ανάπτυξη νέων μοντέλων συσχέτισης δομής – ιδιοτήτων για την υποστήριξη της μη στοχευμένης ανάλυσης περιβαλλοντικών δειγμάτων με φασματομετρία μάζας υψηλής διακριτικής ικανότητας

REZA AALIZADEH

Αριθμός μητρώου: 001501

Επιβλέπων καθηγητής:

Νικόλαος Θωμαΐδης, Καθηγητής

Τριμελής συμβουλευτική επιτροπή:

Νικόλαος Θωμαΐδης, Καθηγητής

Αναστάσιος Οικονόμου, Καθηγητής

Ευάγγελος Γκίκας, Επίκουρος Καθηγητής

Επταμελής εξεταστική επιτροπή:

Νικόλαος Θωμαΐδης, Καθηγητής

Αναστάσιος Οικονόμου, Καθηγητής

Ευάγγελος Γκίκας, Επίκουρος Καθηγητής

Ιωάννης Ντότσικας, Καθηγητής

Θωμάς Μαυρομούστακος, Καθηγητής

Ειρήνη Παντερή, Καθηγητής

Χρήστος Κόκκινος, Επίκουρος Καθηγητής

Ημερομηνία εξέτασης: 15/03/2019

DOCTORAL THESIS

Development of quantitative structure property relationships to support non-target LC-HRMS screening

REZA AALIZADEH

Registration Number: 001501

Supervising Professor:

Dr. Nikolaos Thomaidis, Professor

Three-member consultative committee:

Dr. Nikolaos Thomaidis, Professor

Dr. Anastasios Economou, Professor

Dr. Evangelos Gikas, Assistant Professor

Seven-member examination committee:

Dr. Nikolaos Thomaidis, Professor

Dr. Anastasios Economou, Professor

Dr. Evangelos Gikas, Assistant Professor

Dr. Ionnis Dotsikas, Professor

Dr. Thomas Mavromoustakos, Professor

Dr. Irene Panderi, Professor

Dr. Christos Kokkinos, Assistant Professor

Defense Date: 15/03/2019

<This page intentionally left blank>

ΠΕΡΙΛΗΨΗ

Κατά την τελευταία δεκαετία, ένας μεγάλος αριθμός αναδυόμενων ρύπων έχουν ανιχνευθεί και ταυτοποιηθεί σε επιφανειακά ύδατα και λύματα, προκαλώντας ανησυχία για το υδάτινο οικοσύστημα, λόγω της πιθανής χημικής τους σταθερότητας. Η τεχνική της υγροχρωματογραφίας - φασματομετρίας μάζας υψηλής διακριτικής ικανότητας (LC-HRMS) αποτελεί μια αποτελεσματική τεχνική για την ανίχνευση αναδυόμενων ρύπων στο περιβάλλον. Η ταυτόχρονη δε ανάλυση των δειγμάτων με τις συμπληρωματικές τεχνικές της υγροχρωματογραφίας αντίστροφης φάσης (RPLC) και της υγροχρωματογραφίας υδρόφιλων αλληλεπιδράσεων (HILIC), συντελεί στην ταυτοποίηση «ύποπτων» ή και άγνωστων ρύπων με ποικίλες φυσικοχημικές ιδιότητες. Για την ταυτοποίηση τους, απαιτείται να πληρούνται συγκεκριμένα κριτήρια, τα οποία αξιολογούνται με βάση τη χρήση διαγνωστικών εργαλείων, όπως η ακριβής πρόβλεψη του χρόνου ανάσχεσης, η *in silico* θραυσματοποίηση και η πρόβλεψη της συμπεριφορά τους στον ιοντισμό.

Στο 3ο κεφάλαιο της παρούσας διδακτορικής διατριβής περιγράφεται η ανάπτυξη μιας ολοκληρωμένης πορείας εργασίας (workflow) για τη διερεύνηση των παραμέτρων που επηρεάζουν τον χρόνο έκλουσης μεγάλου αριθμού ενώσεων που συγκαταλέγονται στους αναδυόμενους ρύπους. Για τον σκοπό αυτό, πάνω από 2.500 αναδυόμενοι ρύποι χρησιμοποιήθηκαν για την ανάπτυξη του μοντέλου πρόβλεψης χρόνου ανάσχεσης για τις 2 υγροχρωματογραφικές τεχνικές (RP- και HILIC-LC-HRMS) και για ηλεκτροψεκασμό τόσο σε θετικό όσο και σε αρνητικό ιοντισμό (+/-ESI). Στη συνέχεια, πραγματοποιήθηκε εφαρμογή του μοντέλου για την υπολογιστική πρόβλεψη του χρόνου ανάσχεσης, για την ταυτοποίηση 10 νέων προϊόντων μετασχματισμού των φαρμακευτικών ενώσεων (tramadol, furosemide και niflumic acid) ύστερα από επεξεργασία με όζον.

Στο 4ο κεφάλαιο παρουσιάζεται η ανάπτυξη ενός καινοτόμου γενικευμένου χημειομετρικού μοντέλου το οποίο είναι ικανό να προβλέπει τον χρόνο έκλουσης κάθε πιθανού ρύπου, ανεξαρτήτου υγροχρωματογραφικής μεθόδου που χρησιμοποιείται, συμβάλλοντας σημαντικά στην σύγκριση αποτελεσμάτων από διαφορετικές LC-HRMS μεθόδους. Το συγκεκριμένο μοντέλο χρησιμοποιήθηκε για την ταυτοποίηση «ύποπτων» και άγνωστων ενώσεων σε διεργαστηριακές δοκιμές.

Το Κεφάλαιο 5, περιέχει την περιγραφή της ανάπτυξης ενός υπολογιστικού μοντέλου πρόβλεψης τοξικότητας αναδυόμενων ρύπων που ανιχνεύονται στο υδάτινο οικοσύστημα. Το συγκεκριμένο μοντέλο αποσκοπεί στην εκτίμηση του πιθανού περιβαλλοντικού κινδύνου για νέες ενώσεις που ταυτοποιήθηκαν μέσω σάρωσης «ύποπτων» ενώσεων και μη-στοχευμένης σάρωσης, για τις οποίες δεν είναι ακόμα διαθέσιμα πειραματικά δεδομένα τοξικότητας.

Τέλος, στο κεφάλαιο 6 παρουσιάζεται ένας αυτοματοποιημένος και συστηματικός τρόπος σάρωσης «ύποπτων» ενώσεων και μη-στοχευμένης σάρωσης σε δεδομένα από LC-HRMS. Η νέα αυτή αυτοματοποιημένη πορεία εργασίας, αποσκοπεί στην λιγότερο χρονοβόρα επεξεργασία των HRMS δεδομένων, και στην εφαρμογή της μη-στοχευμένης σάρωσης ώστε να είναι δυνατή η εφαρμογή τους σε καθημερινούς ελέγχους ρουτίνας ή/και για χρήση από τις κανονιστικές αρχές.

Περιοχή έρευνας: Αναλυτική Χημεία
Λέξεις κλειδιά: Χημειομετρία, σάρωση για ύποπτες ενώσεις, μη στοχευμένη ανάλυση, φασματομετρία μαζών υψηλής διακριτικής ικανότητας

ABSTRACT

Over the last decade, a high number of emerging contaminants were detected and identified in surface and waste waters that could threaten the aquatic environment due to their pseudo-persistence. As it is described in chapters 1 and 2, liquid chromatography high resolution mass spectroscopy (LC-HRMS) can be used as an efficient tool for their screening. Simultaneously screening of these samples by hydrophilic interaction liquid chromatography (HILIC) and reversed phase (RP) would help with full identification of suspects and unknown compounds. However, to confirm the identity of the most relevant suspect or unknown compounds, their chemical properties such as retention time behavior, MSn fragmentation and ionization modes should be investigated.

Chapter 3 of this thesis discusses the development of a comprehensive workflow to study the retention time behavior of large groups of compounds belonging to emerging contaminants. A dataset consisted of more than 2500 compounds was used for RP/HILIC-LC-HRMS, and their retention times were derived in both Electrospray Ionization mode (+/-ESI). These *in silico* approaches were then applied on the identification of 10 new transformation products of tramadol, furosemide and niflumic acid (under ozonation treatment).

Chapter 4 discusses about the development of a first retention time index system for LC-HRMS. Some practical applications of this RTI system in suspect and non-target screening in collaborative trials have been presented as well.

Chapter 5 describes the development of *in silico* based toxicity models to estimate the acute toxicity of emerging pollutants in the aquatic environment. This would help link the suspect/non-target screening results to the tentative environmental risk by predicting the toxicity of newly tentatively identified compounds.

Chapter 6 introduces an automatic and systematic way to perform suspect and non-target screening in LC-HRMS data. This would save time and the data analysis loads and enable the routine application of non-target screening for regulatory or monitoring purpose.

Subject Area: Analytical Chemistry

Keywords: Chemometrics, Suspect Screening, Non-target Screening, High Resolution Mass Spectrometry

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my supervisor Prof. Dr. Nikolaos S. Thomaidis in the Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens for the useful comments, remarks and engagement throughout the learning process of this thesis. He was always supportive whenever I ran into a trouble spot or had a question about my research or writing. I honestly thank him for being there and steering me in the right direction. I would like to acknowledge Dr. Peter C. von der Ohe (Environmental Protection Agency of Germany (UBA)), Dr. Emma Schymanski (Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg) and Dr. Maria-Christina Nika (National and Kapodistrian University of Athens) for their valuable comments and feedback during my research. I would like to thank the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) for providing the scholarship and financial support (under the HFRI Ph.D. Fellowship grant (GA. no. 14484)) during my study. Finally, I want to express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without them. Thank you.

Content

ΠΕΡΙΛΗΨΗ 6

ABSTRACT 8

CHAPTER 1: Challenges in identification of chemicals via liquid chromatography high resolution mass spectrometry (LC-HRMS) 18

1.1. Introduction 18

1.2. Screening Strategies 19

1.2.1. Target Screening 19

1.2.2. Suspect screening 20

1.2.3. Non-target screening 21

1.3. Importance of chemometrics methods in non-target screening 25

1.3.1. Current pre-processing data analysis approach used in HRMS 25

1.3.2. Current post-processing data analysis approach used in HRMS 26

1.3.2.1. Unsupervised chemometrics methods 27

1.3.2.2. Supervised chemometrics methods 27

1.4. Prioritization methods 28

1.5. Liquid chromatography and use of retention time in the non-target screening 29

1.6. Prediction of LC Retention Time 30

1.6.1. Use of the solvatochromic method 31

1.6.2. Quantitative Structure-Retention Relationship 31

1.6.3. Chemical Fingerprints 32

1.7. Retention Time Indices and prediction 33

1.7.1. Available RTI module for LC 33

1.7.1.1. Direct projection of various LC systems 33

1.7.1.2. Use of n-nitroalkane 34

1.8. Use of experimental and in silico MS/MS spectra 34

1.8.1. MS/MS spectral library 34

1.8.2. Use of in-silico fragmentation tool 35

1.8.3. Prediction of MS/MS fragments 35

1.9. Acute risk assessment in aquatic environment 36

1.9.1. Baseline toxicity and prediction of acute toxicity 36

1.9.2. In silico risk assessment 37

1.9.3. Read across approach 38

1.9.4. Derivation of predicted no-effect concentration 38

1.10. Automated suspect and non-target screening 38

CHAPTER 2: Scope and Objective 44
2.1. The analytical and computational chemistry problem 44
2.2. Research Objectives and Scope 44
CHAPTER 3: Development and Application of Retention Time Prediction Models (RPLC/HILIC-QToF-MS) in the Suspect and Non-target Screening of Emerging Contaminants 48
3.1. Introduction 48
3.2. Materials and methods 50
3.2.1. Sample preparation, instrumental analysis and dataset 50
3.2.2. QSRR workflows 53
3.2.3. Applicability domain studies and tR acceptable error windows 54
3.2.4. Experimental setup for the generation and identification of TPs of selected emerging contaminants 56
3.3. Results and discussion 56
3.3.1. RPLC-(+)ESI-HRMS 56
3.3.2. RPLC-(-)ESI-HRMS 57
3.3.3. HILIC-(+)ESI-HRMS 59
3.3.4. Acceptable error windows for predicted tR 60
3.3.5. Comparison with literature models 62
3.3.6. Application of tR prediction in the identification of transformation products 63
3.4. Conclusions 66
CHAPTER 4: Development and Prediction of Liquid Chromatographic Retention Time Indices 67
4.1. Introduction 67
4.2. Materials and Methods 69
4.2.1. Chemicals 69
4.2.2. Instrumentation and Procedure 70
4.2.3. Retention Time Indices 70
4.2.4. Intralaboratory and interlaboratory validation 72
4.2.5. External application and validation 73
4.2.6. QSRR workflows 73
4.2.7. Prediction Intervals 73
4.2.8. Stability test of RTI calibrants 74
4.3. Results and Discussion 75

4.3.1. Selection of RTI calibrants 75
4.3.2. Modeling Retention Time Indices 76
4.3.3. Intralaboratory Accuracy of RTI 78
4.3.4. Interlaboratories Accuracy of RTI 87
4.3.5. Evaluation of RTI proposed in –ESI 87
4.3.6. Evaluation of RTI proposed in +ESI 95
4.3.7. External evaluation 102
4.3.8. Stability test of RTI calibrants 102
4.3.9. Application of RTI in Suspect and Non-target Screening 107
4.4. Conclusions 108
CHAPTER 5: Prediction of Acute Toxicity of Emerging Contaminants on the Water Flea Daphnia magna by Ant Colony Optimization - Support Vector Machine QSTR models 110
5.1. Introduction 110
5.2. Materials and Method 112
5.2.1. Dataset preparation and chemical-toxicity curation 112
5.2.2. Molecular descriptors calculation 114
5.2.3. Molecular descriptor selection and modelling 114
5.2.4. Validation criteria 116
5.2.5. Identification of potential toxicity outliers using OTrAMS 116
5.3. Results and discussion 118
5.3.1. Data treatment 118
5.3.2. Derivation of the ACO-MLR and ACO-SVM models 119
5.3.3. Chemical toxicity mechanism in Daphnia magna 124
5.3.4. Prediction performance of the ACO-SVM and ACO-MLR models 127
5.3.5. Additional evaluation set and applicability domain 128
5.3.6. Identification of potentially erroneous toxicity values 131
5.4. Conclusions 132
CHAPTER 6: AutoNonTarget: an R package for automatic suspect and non-target screening 134
6.1. Introduction 134
6.2. Methodology 137
6.2.1. Samples collection and instrumentation 137
6.2.2. Overview of steps involved in the automatic suspect/non-target screening 137

6.2.3. Peak picking for HRMS data 139
6.2.4. Subtraction of analytical procedural blank from samples 142
6.2.5. Prioritization of MS peaks-list 145
6.2.5.1. Peaks list annotation 145
6.2.5.2. Time trend analysis 146
6.2.5.3. Chemometrics 147
6.2.5.3.1. Group comparison with statistical analysis 148
6.2.5.3.2. Volcano plot 148
6.2.5.3.3. Variable importance in projections in PLS-DA 149
6.2.5.3.4. S-plot for variable significant test in OPLS-DA 151
6.2.5.3.5. Variable prioritization in decision tree 151
6.2.6. Regulatory, suspect and public Databases 152
6.2.7. Assignment of molecular formula 154
6.2.8. MS/MS spectral interpretation and prediction 154
6.2.9. In-house QSRR models for retention time prediction 155
6.2.10. Diagnostic evidence 155
6.2.11. Experimental MS/MS spectral match 157
6.2.12. Confirmation by authentic reference standard 157
6.2.13. Aquatic risk assessment 158
6.3. Results and Discussion 158
6.3.1. Detected false positives by Deep Learner 158
6.3.2. Screening of biocides in wastewater and sludge 160
6.4. Conclusions 167
CHAPTER 7: Concluding Remarks 168
References 169
Supplementary Material 187

LIST OF FIGURES

Figure 1.1 General screening strategies used in HRMS 24

Figure 3.1 Distribution of cationic, anionic and neutral compounds in RPLC and HILIC 53

Figure 3.2 The explanation of Monte-Carlo sampling (MCS) method used to define the applicability domain of the models developed to predict tR 55

Figure 3.3 Identification of 5-Methylbenzotriazole: (a) full MS chromatogram for the given mass (±5ppm); (b) MS/MS spectra and corresponding fragments; (c) MCS plot for evaluating the predicted tR values; (d) confirmation step using spectra library. 5-Methylbenzotriazole was confirmed by reference standard later. 62

Figure 4.1 The cloud plot of experimental RTIs measured in various LC conditions with their acceptance CIs in (A) –ESI and (B) +ESI 82

Figure 4.2 The correlation between the experimental and predicted RTIs measured in various LC conditions for validation set used in (A) –ESI and (B) +ESI 87

Figure 4.3 Cloud map of RTI values in different LC conditions reported by (A) UFZ; (B) Eawag; (C) UJI in (–ESI) 89

Figure 4.4 QSRR based prediction results with its boundaries for the compounds reported by UFZ; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results 90

Figure 4.5 QSRR based prediction results with its boundaries for the compounds reported by Eawag; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI;(D) distribution of error associated with prediction results. 92

Figure 4.6 QSRR based prediction results with its boundaries for the compounds reported by UJI; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results. 94

Figure 4.7 Cloud map of RTI values in different LC conditions reported by (A) UFZ; (B) Eawag; (C) UJI in (+ESI) 96

Figure 4.8 QSRR based prediction results with its boundaries for the compounds reported by UFZ; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results. 97

Figure 4.9 QSRR based prediction results with its boundaries for the compounds reported by Eawag; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results. 99

Figure 4.10 QSRR based prediction results with its boundaries for the compounds reported by UJI; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results. 101

Figure 4.11 Stability test for the RTI calibrants 107

Figure 5.1 Overlap between the normally distributed toxicity levels and chemical space in training and test set. 119

Figure 5.2 Correlation between experimental and predicted Toxicity: A) ACO-MLR and B) ACO-SVM 122

Figure 5.3 A) The OTrAMS can be used to detect compounds with erroneous 124

Figure 5.4 Principal Component Analysis with molecular descriptors loadings and toxicity levels 126

Figure 5.5 Applicability domain study for the evaluation set using A) Normalized mean distance versus leverage values and B) OTrAMS 129

Figure 5.6 Correlation between experimental and predicted pLC_{50} for the evaluation set using ACO-SVM 131

Figure 6.1 AutoNonTarget workflow 138

Figure 6.2 The peak picking workflow used for LC-HRMS data processing 141

Figure 6.3 A schematic workflow of distinguishing between the false and true positive peaks using the deep learner algorithm 144

Figure 6.4 Gaussian curves for detection of m/zs (here it is TPs) 147

Figure 6.5 Volcano plot and prioritization of m/zs in the peaks list 149

Figure 6.6 Dot product basics for mass spectral comparison 156

Figure 6.7 False positive detected by “Deep Learner” as analytical procedural subtraction approach 159

LIST OF TABLES

Table 1.1 Publicly available data analysis tools for suspect/non-target screening in HRMS 40

Table 3.1 Retention time prediction results for the identification of ozonation transformation products of emerging contaminants 65

Table 4.1 Prediction performance of models developed for proposed RTI system. 76

Table 4.2 Intralaboratory results for the RTI proposed for (–ESI) mobile phase 80

Table 4.3 RTI and tR calibration curve for (–ESI) under different LC conditions 81

Table 4.4 Intralaboratory results for the RTI proposed for (+ESI) mobile phases 84

Table 4.5. RTI and tR linear equations for (+ESI) in different LC conditions 85

Table 4.6 RTI and tR calibration curve for (–ESI) reported by different laboratories 91

Table 4.7 RTI and tR linear equations for (+ESI) reported by different laboratories 98

Table 5.1 Intercorrelation between molecular descriptors and related VIF value. 120

Table 5.2 Internal and external evaluation and related accuracy measurements. 121

Table 5.3 Internal and external accuracy of current model in contrast to the previously proposed ones. 127

Table 6.1 List of identified biocides in influent, effluent wastewater (IWW & EWW) and sewage sludge of wastewater treatment plants (WWTP) of Athens (Greece) 164

< This page intentionally left blank>

CHAPTER 1

Challenges in identification of chemicals via liquid chromatography high resolution mass spectrometry (LC-HRMS)

1.1. Introduction

Over the last decades, thousands of substances with potential risks for human and aquatic life are disposed in the environment. Their rapid and accurate identification is emerged as an important field in both analytical and environmental science. The evolution of high resolution mass spectroscopy coupled with liquid chromatography (mainly, Hydrophilic interaction liquid chromatography (HILIC) and reversed phase LC (RPLC)) has opened up a new opportunities for the identification of polar and partially polar compounds in complex environmental samples. Identification procedures in LC-HRMS were detailed into three categories including target analysis (where reference standards are available), suspect screening (existence of suspected substances based on prior information) and finally non-target screening (no prior information is available nor the reference standards) [1]. In this context, especially in the suspect and non-target screening task, the use of computational methods such chemometrics (for the prioritization of m/z values in a typical HRMS peaks-list), cheminformatics (for elucidation of tentative candidates for a given mass of interest) and bioinformatics (for metabolic pathway analysis) are highly encouraged. Chemometrics is an inter-disciplinary field and it could be defined as the science of studying a chemical process/system by mathematics or multivariate statistics. In the case of HRMS data analysis, the main objective of chemometrics is to prioritize the peaks-list based on the variation of instrumental response factor (either maximum intensity or peak area) over a treatment chain or time trends. With this respect, the great deal of identification efforts can be focused only on those (m/z values) which their peak area or intensity vary significantly from one category to another. To achieve this, the unsupervised or supervised classification method should be followed depending on our prior knowledge of the categories of set of samples. For the identification of suspects or unknowns, apart from mass accuracy and isotopic fitting, retention time (tR) and MS/MS spectra evaluation is required. To this end, cheminformatics can be highly promising. It has major applications in development of *in silico* fragmentation MS/MS

methods to interpret an experimental MS/MS spectra resulting in the rational elucidation of target chemical structure or its use in development of Quantitative Structure-Property Relationships (QSRR) for retention time prediction. Retention time prediction is one of the key step to exclude false positives during elucidation of target chemical structure. Retention Time Indices (RTI) remains the best way to harmonize elution pattern of chemicals and further use them for promotion of identification confidence. This is very well established in Gas Chromatography (GC) and known as Kovats retention indices (use of n-alkanes for calculating the RI of chemicals) whereas such a valuable method remains unknown in LC. The risk assessment and its prediction are also an emerging technique to develop the chemical watch-list in the suspect and retrospective screening. This helps to prioritize the chemicals based on their imposed environmental risk and subsequently promote their regulations and terms of use.

1.2. Screening Strategies

1.2.1. Target Screening

Target screening is the most reliable screening approach that can be followed during low and high MS data treatment and it generally refers to conditions where the substance is already known and the reference standard is available to verify the detected compound. Therefore, the criteria that is used in the target screening is matching the mass accuracy, isotopic pattern, MS/MS fragments and retention time of the reference standard with the detected compound. To this end, low and high resolution mass spectrometry instruments can be used safely. LC coupled to triple quadrupole (LC-QqQ-MS/MS) is widely used analytical technique in target screening. LC-QqQ-MS/MS is a tandem mass spectrometer consisting of two quadrupole mass analyzers in series, with a (non-mass-resolving) quadrupole between them that acts as a cell for collision-induced dissociation. The QqQ analyzer promotes the application of various MS/MS modes such as product ion scan, precursor ion scan, neutral loss scan and selected reaction monitoring (SRM), which is the most predominant (due to increased selectivity, better accuracy and greater reproducibility).

In the product scan mode, the first quadrupole selects an ion of a known mass, which is to be fragmented in q2, and the third quadrupole scans the entire m/z range, providing the information on the fragments made. This method is generally performed

to identify transitions used for quantification. When operating the tandem MS with precursor scan mode, a certain product ion is selected in third quadrupole, and the precursor masses are scanned in the first quadrupole. In the neutral loss scan method, both first and third quadrupole are scanned together, but with a constant mass offset. This permits the selective recognition of all ions which, by fragmentation in q2, lead to the loss of a neutral fragment (such as H_2O, NH_3). In the Selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) modes, both first and third quadrupole are set at a specific mass, allowing only a distinct fragment ion from a certain precursor ion to be detected. With the use of LC-QqQ-MS/MS, efficient quantitative data can be obtained for the analysis of emerging contaminants and the identification and quantification of their Transformation products (TPs), especially in the field of pesticides and pharmaceutical compounds, where understanding metabolism of the substance is of great importance.

1.2.2. Suspect screening

Suspect screening is the technique of choice for the identification of compounds that is suspected to be present in the sample under analysis and their reference standard was not available at the time of analysis for confirmation. Full-scan MS with data-dependent acquisition mode (DDA) in the tandem mass spectrometry (full scan MS and its DDA MS^2 with inclusion list) is the best way to perform a successful suspect screening. Although DDA mode offers great sensitivity and certainty over deconvolution and interpretation of MS/MS fragments towards its parent ion, it can only acquire MSMS data for certain numbers of intense ions present in the full scan MS data. Therefore, having an inclusion list grantees that the MSMS fragments list will be derived for the precursor ion in the inclusion list. In suspect screening, an important step of the identification workflow is the use of computational (*in silico*) tools for prioritization of the candidates. Generally, the suspect screening starts with the exact mass accuracy match between compounds in the suspect list and the peaks in the samples. It is important to note that an intensity threshold value is applied to cutoff unclear spectra and an analytical procedural blank subtraction is required to avoid false positives identifications. The chromatographic retention time (tR) plausibility, isotopic pattern, and ionization efficiency are used as further filters to narrow down the number of candidates for a peak observed in the sample. Furthermore, using the

MS/MS or MSn operating mode, the chemical structure of the suspected compound can be elucidated via fragmentation pattern. At this stage use of retention time prediction, MSMS prediction tools as well as lookup methods for the experimental MSMS fragments list in the literature or mass spectrum library could help to further decrease the chance of false positives. Nevertheless, confirmation of the suspected and most probable candidate requires purchase of reference standard. Currently, automatic compilation of comprehensive suspect list and accurate prediction of retention time and MSMS fragmentation pattern are the main focus of literature in the suspect screening.

1.2.3. Non-target screening

Non-target screening is a screening strategy for identifying the compounds for which there is no previous knowledge available and it is usually carried out after target and suspect screening. Non-target screening is a challenging task and it generally starts after subtraction of analytical procedural blank and a prioritization task over a peaks-list -which could be contain over ten thousand masses for a sample- generated by HRMS instrument. The removal of noise peaks, mass recalibration and componentization of isotopes and adducts are usually carried out subsequently. The prioritization of the masses in the peaks-list could be simple peak score criteria [2], risk imposed counterparts (could be achieved through effect directed analysis (EDA) and HRMS) [3] or the masses that found to be important to characterize certain set of samples through chemometric methods [4]. Effect-directed analysis (EDA) is an approach to identify chemicals in complex mixtures which exert adverse effects. The combination of bioassays for effect detection, fractionation to reduce sample complexity, and chemical analysis mainly by HRMS found to be very effective tool for efficient and successful toxicant identification [5]. One of the areas where EDA is successfully applied is water resource monitoring. This however limits the application of this practice and it is mainly used over the certain samples (such as environmental samples) where there is potential toxicants present. The prioritized masses are called potential mass of interest and concerning the level of identification proposed by Schymanski et al., these masses are at the level of identification confidence 5. Non-target screening, is therefore, starts from this level reaching to probable chemical structure while the list of possible candidates are provided as a suspect list in the

suspect screening where the assignment of identification confidence starts from level 3. To proceed from level of identification 5 to 4 for a mass of interest, evaluation of molecular formula [6] and match between theoretical and experimental isotopic pattern or presence of supportive ions (such as adducts) are required. In case of known unknown and existence of possible candidates for the given molecular formula, the MS/MS spectrum should be interpretable for a retrieved candidate. At this stage, *in silico* fragmentation tools such as MetFrag [7] or CFM-ID [8] could be useful to rank the candidates based on their explained MS/MS fragments (match between *in silico* and experimental MS/MS fragments). In addition, the retention time prediction [9] and other physico-chemical (for instance use of complementary RP versus HILIC elution pattern [10]) data can contribute for further ranking of possible structures and facilitate the identification process [11] especially in the case of isobaric substances [10]. The impact of chromatographic resolution on mass measurement accuracy, mass measurement precision, and ion suppression (due to co-elution) have been studied at fundamental level and correct LC condition is a need for successful screening of complex mixes [12]. To reach to the level of identification confidence 2 (2a and 2b), more supportive data are required. Presence of diagnostic ions or fragmentation pattern between parent compounds and their transformation products are needed to increase the level of identification confidence to 2b. If the experimental MS/MS spectrum is available in the literature or spectrum library (MoNA (http://mona.fiehnlab.ucdavis.edu/), MassBank (https://massbank.eu/MassBank/), mzCloud (https://www.mzcloud.org/) and METLIN (https://metlin.scripps.edu/)) and it is matching to the observed one, then the level of identification confidence can be reached to 2a. Although, the comparison of experimental RTI values can be used to promote the identification confidence from level 3 to level 2, it is not included in the current identification scheme due to the lack of such a tool in liquid chromatography when these levels, proposed. A small modification to this scheme to include elution parameters would be needed and it should be evaluated rigorously. For non-target screening, high resolution mass spectrometry is needed in order to have mass accuracy for confirmation of molecular formula and a reliable interpretation of the MS/MS spectra, however there have been several efforts to perform non-target

screening with low resolution mass spectrometry instruments [13]. The general procedure followed in the screening strategies is shown in Figure 1.1.

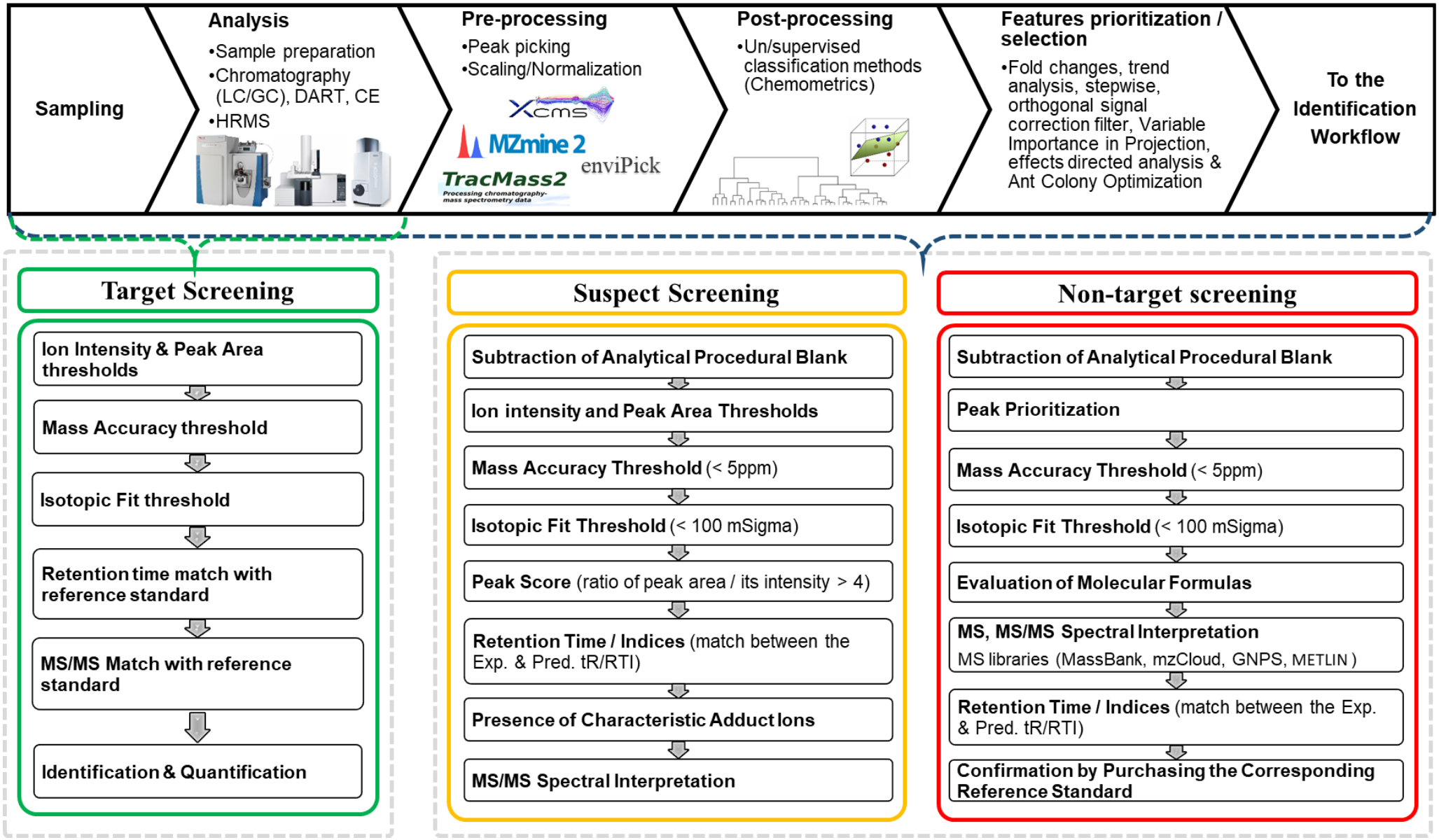


Figure 1.1 General screening strategies used in HRMS

1.3. Importance of chemometrics methods in non-target screening

Non-targeted analysis has gained great importance in the field of analytical chemistry especially in the metabolomics, environmental science, food science and especially the '-omics' related subjects. This is due to the fact that non-targeted analysis leads to derivation of qualitative and quantitative information of as many compounds as possible in the analyzed samples and provides a more holistic view of the composition of samples. As mentioned above, the bottleneck in the non-target screening is to first prioritize the peaks-list, and then elucidates the chemical structure, as the identification task becomes harder when several hundred candidates can be assigned to a mass within provided mass accuracy. Chemomterics plays crucial role in this aspect to reduce the efforts of identification task to only important component of the peaks-list. This way, the main focus and effort is to identify the masses that are important to explain characteristics of a sample.

1.3.1. Current pre-processing data analysis approach used in HRMS

Before performing the post-processing data analysis on the peak list, a normalization or transformation of the data might be necessary [14-16]. Normalization removes the effects of confounding variations related to experimental sources, such as experimental bias, analytical noise or instrumental sensitivity. If the signal of the detected biomarkers is stable, a simple normalization can be followed by estimating the relative ratio of the abundance of analytes to all other detected peaks [14]. However, the assumption of negligible overall concentration changes is naive when using MS instruments. The MS instruments can lose their sensitivity after passage of time, causing considerable changes in the total concentrations of analytes. In this case, scaling based on the total chromatogram can extremely distort the peak list. Nevertheless, spiking the samples with any internal standards, or adaptation of QC samples in each data acquisition procedure remains the best procedure to study variation in instrumental response factor [16, 17]. Compounds at lower concentrations can be manipulated easily by the analytical noise. To promote the comparison of different biomarkers, scaling the peak list is a necessity. Autoscaling is one of the most widely used scaling method in the chemometrics; in this method, each variable (subtracted from its mean) has equal (unit) variance by multiplying it with the inverse

of standard deviation (SD). Another well-known scaling method is “Pareto” which increases the importance of biomarkers with low signals without significantly amplifying the noise. Scaling should not be applied to LC-HRMS data blindly because regardless of the variation in the instrumental response factor, the intensity variation between (m/z)s provides important information [14]. Nevertheless, scaling the LC-MS data might help enhance the weights for low intense (bio)markers promoting them to be influential in the post-processing data analysis. Although, some of these scaling methods such as “Autoscaling” or “Pareto” may pave the road for discovery of biomarkers at lower concentration, most of these peaks (at lower intensity) might not be even detected in peak picking step. Therefore, the total effect of scaling/normalization/transformation of peak list needs to be explored carefully.

1.3.2. Current post-processing data analysis approach used in HRMS

HRMS datasets may contain irrelevant or redundant variables (originated from contamination in the ion source, carry over in LC column, analytical procedural blank or satellite peak) which can adversely affect the outcome of chemometircs method. Post-processing data analysis refers to set of actions that can be followed in order to decrease the complexity of HRMS results for an analyzed sample. Some types of sample could even add up to this complexity such as environmental samples where they are subjected to great exposure risk of pollutants from various sources. Tools are required to resolve these obstacles and measurements complexity (from automatic monitoring, applied treatment methods, etc), uncertainty, imprecision, multi-scalarity, heterogeneity, loss of instrumental sensitivity or matrix effect later. The fold change in the instrumental response factor (maximal peak intensity or peak area in the HRMS) between “before-and-after” a treatment process (such as influent and effluent wastewater) is the simplest way to find common, removed or generated features. In general, two-group tests (such as Welch’s t-test) allow researchers to determine the features whose levels are significantly different between two defined conditions. Trend analysis is another well-known approach to track the formation or degradation of a compound over time or use of ozonation/chloreation dose. This enables to quickly tag the peaks that appeared over time (as a result of formation (mainly the transformation products (TPs) or accumulation of pollutants in the downstream) or those that removed

(parent compounds or compounds that are easily biodegradable). There are other much advanced chemometric methods which can be used depending on the origin of research. These methods have gained growing applications in the field of foodomics (for food authenticity) and metabolomics (biological pathway network and analysis as well as discovery of biomarkers for various disease). These types of analysis can be done in a un/supervised manner in case the category of a sample is known.

1.3.2.1. Unsupervised chemometrics methods

Unsupervised classification methods are the most widely used ones for data exploratory analysis of HRMS raw data. Among which Principal Component Analysis (PCA), clustering analysis (such as hierarchical clustering) or unsupervised Kohenon self-organizing map (SOMs). PCA, the most widely applied technique, linearly decomposes the data array (which is the HRMS peaks-list) into lower dimensional space than original data array and provides score and loading plot to show the sample distribution and the significance of variables (in this case is the *m/z* values at recorded t_R).

1.3.2.2. Supervised chemometrics methods

Supervised techniques use the prior knowledge of categorical information of a sample alongside the peaks-list data to find a pattern for assigning the classes on a set of samples with a minimum misclassification error. The advantage of this kind of technique is the predictive capability of the models to be readily used over a new set of samples. Supervised classification techniques can be performed by linear methods, such as linear discriminant analysis (LDA) [18], soft independent modelling by class analogy (SIMCA) [19], partial least squares-discriminant analysis (PLS-DA) [20], orthogonal projections to latent structures discriminant analysis (OPLS-DA) [21] or non-linear methods, including support vector machine (SVM) and random forest (RF) [22], counterpropagation artificial neural networks (CP-ANNs) [23] and supervised SOMs (SKNs) [24]. These techniques sometimes can be easily over-fitted especially non-linear methods (support vector machine) and certain features selection methods should be used in parallel. It is worth to note the feature selection ability is inherited in RF when it searches for classification rules, but to avoid the convergence of RF to

suboptimal features space and subsequently avoid tuning the tree (describes the classification rules), it is recommended to use a feature selection tool beforehand [4].

1.4. Prioritization methods

Chemometric techniques used in the analysis of HRMS data often are accompanied by a feature selection methods (such stepwise variable selection or nature inspired techniques). This help solving the difficulties of searching whole variable space and performing non-target screening. Although this could be categorized as prioritization step in the workflow, it only focuses on the peaks-list.

EDA, as being said above, is an approach to identify chemicals in complex mixtures which exert adverse effects. The combination of bioassays for effect detection, fractionation to reduce sample complexity, and chemical analysis mainly by HRMS found to be very effective tool for efficient and successful toxicant identification [5]. Ecological risk assessment (ERA) has been extensively used to help understand the adverse effects caused by contaminants to the aquatic environment.

Detection of potentially persistent, bioaccumulating and toxic chemicals is the main purpose of prioritization step in the developed non-target workflow (Figure 1). To identify environmentally relevant contaminants, previous studies have focused on criteria such as ecotoxicity [25], exposure [26] or bioactivity [27]. An important step in ecotoxicity is effect assessment, i.e. the determination of the maximum concentration at which the aquatic life form is protected, known as the predicted no-effect concentration (PNECs). Comparison between the PNECs and quantification data of the detected compound is the mostly used approach in this realm [28]. The use of predicted based PNECs (P-PNECs) and development of semi-quantification methods remove the barrier of resources limit for implementation of bioassay tests or reference standards purchase when there are several candidates proposed for a peak (as the outcome of non-target screening). Other prioritization strategies have focused on pre-selected water contaminants, such as active pharmaceutical ingredients. These approaches prioritized their list of chemicals based on ecotoxicity data [29], biodegradation, bioaccumulation and ecotoxicity data, prescription dispensation [30], environmental concentrations, half-lives, octanol-water partition coefficients, and ecotoxicity data [31]. Attempts have also been made to start with large inventories of industrial chemicals or pharmaceuticals and use prioritization schemes to identify

potentially persistent and bioaccumulating substances [32]. This is the current focus of European network of reference laboratories, research centers and related organizations for monitoring of emerging environmental substances network of reference laboratories dealing with emerging environmental substances (NORMAN association) [33] to compile by far the largest chemical inventory of emerging contaminants (NORMAN SusDat) present in the environment.

These strategies are mainly successful in case of retrospective and suspect screening and they cannot be performed easily in combination with non-target screening. This is due to the fact that sometimes the level of identification cannot reach beyond level 3 which means there are few proposed chemical structures that are potential candidates for a peak. Therefore, most of the risk assessment based approach would not be helpful.

1.5. Liquid chromatography and use of retention time in the non-target screening

Retention information of an analyte in the sample could help prioritize the candidates list when they have reasonably different physico-chemical properties (such as polarity, ionization potency, van der Waals force (branched/unbranched-chain compounds) etc.). The chemical structure of the molecules in the sample will also give clues as to their elution order. This is particularly useful when two peaks are isobaric compounds (Isobaric compounds are compounds with the same nominal mass but with a different molecular formula), and the resolution of mass spectrometric instrument could not differentiate them, but they could be separated in the LC part. Reversed phase HPLC is a configuration in which the mobile phase used is more polar than the stationary phase. The name ‘Reversed Phase’ arises as this was the second, mode of chromatography after “Normal Phase” in which a polar stationary phase is used in conjunction with a less polar mobile phase. Generally, reversed phase stationary phases are hydrophobic and chemically bonded to the surface of silica particles.

The exact elution mechanism of compounds in the Hydrophilic Interaction Liquid Chromatography (HILIC) chromatography is still under investigation and modelling of retention time data is highly encouraged, however the general mechanism involves polar analyte partitioning into and out of a layer of water which is adsorbed onto the

surface of the polar stationary phase. If the eluent pH is adjusted, the stationary phase surface will be charged and therefore certain types of electrostatic interactions can be undertaken between analyte and the stationary phase. The pH of eluent will also affect the liquid/liquid partitioning behavior and subsequently, the partitioning of the analyte. Therefore, buffer concentration have a dominant effect on the retention in HILIC chromatography due to their effects in the degree of analyte and stationary phase ionization and the polarity of the eluent.

As being said, LC helps to resolve the issues observed with isobaric/isomeric compounds in mass spectrometry. Moreover, retention information of a peak can assist the identification task, especially when there are potential candidates for a peak. For the last 5 years, several studies have shown the added value of using retention time data during suspect/non-target screening [9, 34-40]. In one of the studies [36], the formation of TPs from citalopram during biological treatment process in the activated sludge was investigated. Retention time information is used alongside the mass spectrometric evidence during suspect and non-target strategies based on liquid chromatography quadrupole-time-of-flight mass spectrometry (LC-QTOF-MS). The complementary use of RPLC and HILIC for the identification of polar TPs, and the application of quantitative structure-retention relationship (QSRR) prediction models provided valuable information to facilitate the identification. In total, thirteen TPs were tentatively identified.

1.6. Prediction of LC Retention Time

Several retention time predictive models have been proposed for standard HPLC conditions [9, 34, 35, 39, 41-49] and commercially available (ACD/ChromGenius or Chromsword) to facilitate LC method development [50, 51]. These prediction models are mainly based on the chemical structure of the compound and the method development is supported by an internal database of similar compounds of known retention time. Some of the well-known methods used to predict retention time are based on solvatochromic parameters, quantitative structure-retention relationship (QSRR), which is mainly based on partitioning coefficient, and chemical fingerprints. Retention time prediction has been applied in various field such as several scientific fields, including proteomics, foodomics, metabolomic, forensic, environmental, pharmaceutical and medical sciences.

1.6.1. Use of the solvatochromic method

Unlike computationally derived molecular descriptors, the solvatochromic parameters should be determined experimentally. These parameters are combined with each other and are being correlated with capacity factor via a multiple linear regression technique resulting in so called solvation energy relationships (LSER) [52]. However, this is elaborative and the experimentation could be time consuming. The limitation of LSER approach somewhat hinders to use it to predict retention time or rapid method development. Solvatochromic descriptors are the excess molar refraction (R_2), the polarity/dipolarity (π_2^H), the ‘‘effective’’ hydrogen bond acidity and basicity ($\sum \alpha_2^H$) and ($\sum \beta_2^H$), respectively), and the McGowan’s characteristic volume (V_2/100) [52]. Depending on the mixed mode mobile phase and the stationary phase used, LSER can be much simplified [52].

1.6.2. Quantitative Structure-Retention Relationship

In 1977, the first three publications were published with the aim of finding correlation between chemical structures and their chromatographic behavior which is now called QSRR. Since then, a large number of efforts were made to derive robust mathematical models that not only predict the retention time of compounds, but also explain the chemical features affecting retention time values. Several good models have been reported for gas-chromatographic (GC) retention based on chemical features derived from molecular graphs and quantum chemical energy-related [45, 53, 54]. Generally, QSRR results for liquid chromatographic (LC) retention data present lower statistical quality than those reported for GC and this is due to the effect of chromatographic conditions such as stationary phase, column type, separation conditions and elution mechanism at different molecular level over retention behavior of compounds [55, 56]. Beside the lack of ability for inclusion of these effects to QSRR based models, little efforts were done to enrich the applicability domain of models for application of different type of compounds [9, 46]. Use of a data set consisted of large chemical diversity (i.e. increasing the chromatographic effects over retention time values) would also unable the models to find the rational chemical features and thus insufficient interpretations [46]. By growth of chemometics and introduction of new type of molecular features for 3D structure of molecules, capabilities of models were

increased to handle dataset with abnormal retention time [57]. Recent advances in both chromatographic science and chemometics caused a revolutionary enhancement of identification and interpretation of results however modeling of retention time in LC-HRMS is still a challenging work due to complexity of chromatographic and instrumental system [58, 59]. There is a need for computational tools such as QSRR to help the identification of unknown substances in the environment [1, 60]. The metabolomics field has greatly got benefit from reliable retention time prediction (as recently demonstrated by Creek et al. [61] using a QSRR method developed for hydrophilic interaction chromatography (HILIC). Since most of the compounds dealt in metabolomics are polar, HILIC platform is often chosen for analysis. In the environmental chemistry and specifically screening of emerging pollutants, the wide scope prediction is needed as the existed compounds in the environment can have wide range of polarity and diverse chemical properties. Most of the QSRR based models include a partitioning molecular descriptor such as Alog*P* or logD (using pk_a to correct log*P* for ionizable compounds) which they could be prone to errors as these molecular descriptors are basically derived from other predictive tools. Therefore, sophisticated methods should be used to trade-off between internal predictive power and external application of these models. In our previous study [9], we have used a large data set of emerging pollutants (approximately 800 compounds) described with 22 different types of molecular descriptors. With these descriptors we built linear (multiple linear regression) and nonlinear (artificial neural networks and support vector machine) robust models, paying attention to the selection of the final descriptors, the choice of the training and test set, the external validation, and the outlier identification.

1.6.3. Chemical Fingerprints

As said above, some of the classic descriptors, calculating the partitioning coefficient, are prone to errors and it is highly needed to have other molecular properties (which is comprised basically from chemical structure) to compensate the prediction error. In an interesting study published in 2016, Falchi et al. [43] have introduced new types of retention time models which were based on large collections of classical physicochemical and topological descriptors in combination with Canvas2D chemical

fingerprints. Canvas offers seven types of hashed fingerprints, MACCS keys, and customizable SMARTS-base structural keys [62]. All types of chemical fingerprints are represented, and a sparse storage scheme allowed each chemical feature to be mapped to a unique bit. They derived first kernel-based partial least squares (KPLS) model over a retention time data of a large chemical library of **1383** synthetic compounds.

1.7. Retention Time Indices and prediction

As said above, prior knowledge of the retention time of the plausible chemical structures would allow further reduction of candidates list that need to be investigated. However, sharing the liquid chromatographic retention time information is limited across laboratories due to uncertainty raised from LC conditions. Elution information in gas chromatography are routinely used and have a major role during identification procedure owing to Kovats retention indices (RI) in which enables cross checking RI of a suspect with library data and linking various GC conditions. However, there are little efforts coordinated for liquid chromatography to make its different conditions comparable. The reason is lack of experimental information about the elution of whole chemical space (for a large database) in different liquid chromatographic conditions. Moreover, currently, there is no sufficient information available about the calibrants as similar as Kovats RI system for LC.

1.7.1. Available RTI module for LC

1.7.1.1. Direct projection of various LC systems

According to Stanstrup and coworkers [63], most of compounds conserve their elution orders for similar chromatographic system (reversed phase and stationary phase (C-18)) and this could be extremely variable in different types of chromatographic columns (e.g., hydrophilic interaction liquid chromatography (HILIC)). By introduction of the strategy of directly mapping various LC conditions, this could help to share t_R information and include it during identification task [63]. This approach so called PredRet however requires a large number of compounds to project the t_R information with high confidence.

1.7.1.2. Use of n-nitroalkane

Hall and co-workers was also developed RTI system alongside a prediction model similar to Kovats RTI using n-nitroalkanes instead of alkanes [64]. Based on this approach, n-nitroalkanes elute before and after an unknown being measured and then, their retention times define 100 times the number of carbons. This however requires a detailed descriptions for analyzing the reproducibility of RTI system in different LC conditions (different gradient elution program, mobile phase compositions and stationary phase). The use of logarithmic scale for the retention times and also additive function for the number of carbon remain ambiguous since number of carbon is not solely, directly and equally correlated to the polarity measure in LC [65]. Moreover, the RTI system was proposed only for a structurally diverse group of 411 small molecules consisting of endogenous compounds, endogenous metabolites and drugs. Therefore, compounds chosen for testing this system were limited to the set of biological elements (C, H, N, O, S, and P) and contained at least one protonable atom to facilitate their detection in mass spectrometry. There is room however to extend this approach and to explore the effect of compounds containing halogens over reliability of this RTI system. Having used the n-nitroalkanes as calibrants will also limit application of these calibrants in negative ionization mode in MS, as they will not be ionized properly (i.e. the favorable ionization is +ESI owing to nitrogen in nitroalkanes).

1.8. Use of experimental and in silico MS/MS spectra

The identification of unknown compounds from mass spectral data is one of the most commonly-way to help elucidating the target chemical structure. With recent developments to high resolution, accurate mass spectrometry coupled with chromatographic separation has revolutionized the high-throughput analysis and opened up whole new ranges of substances that can be detected within the detection limit of the instruments.

1.8.1. MS/MS spectral library

Several online and publicly available databases have been developed to assist identification of specific types of compounds in suspect and non-target screening. These database provide only probable assignments that must be further evaluated by retention time matching and/or MS/MS analysis. In the absence of a pure reference

standard analyzed under identical analytical conditions, MS/MS data looked up against a reference MS/MS database are typically the most conclusive evidence for validating and putatively annotating a metabolite feature using MS information. Some of these databases include both chemical and MS/MS spectral data such as METLIN, HMDB (the human metabolome database) [66], FoodDB (food constituents), LMSD (biologically relevant lipids database), NIST (include mainly the MS information about authentic chemical standards of metabolites and compounds of industrial and environmental importance) whereas some other databases are chemical inventories such as KEGG, PubChem, ChemSpider, MetaCyc, ChEBI and Comptox EPA dashboard. MassBank, mzCloud and GNPS (contain MS/MS spectral data of natural products) are examples of the databases that exclusively include MS spectral data.

1.8.2. Use of in-silico fragmentation tool

The large majority of these substances or peaks detected in samples typically remain unidentified. As being said, when the reference standards are not available or not present in the spectral libraries or even sometimes numerous potential candidates are proposed, use of *in silico* computational tools are recommended. At this stage, several *in silico* fragmentation tools such as MetFrag [7] or CFM-ID [8] are proposed and used widely to interpret and predict MS/MS fragments, respectively. Some other advnaced tools, such as Mass Frontier, FingerID [67] or MetFusion [68], have been developed recently which are complementary to MetFrag/CFM-ID. For instance, FingerID uses a SVM model, which is trained from the mapping between the mass spectra and molecular fingerprints, to create *in silico* MS/MS fragments of the candidates.
In the other hand, MetFusion approach takes advantage of the spectral data for some compounds and performs a combined query of both MetFrag and MassBank, such that the scores of candidates with high chemical similarity to high-scoring reference spectra are increased. Herein, the MassBank scores are calculated on the basis of a modified cosine distance to compute the similarity between the query spectrum and the reference spectra and the results are ranked according to this spectral similarity.

1.8.3. Prediction of MS/MS fragments

Allen et al. [8] has introduced a stochastic, generative Markov model to investigate the fragmentation pattern in the small molecules. Implemented in a web service so called

CFM-ID (competitive fragment modelling), the MS/MS spectrum of a given compound can be predicted at low (10V), medium (20V) and high (40V) collision energies. The input chemical structure can be provided in SMILES or InChI format. A proton will then be added or removed ([M+H]+ or [M−H]− precursor ion) according to whether the user has specified positive or negative mode ionization. CFM-ID can also be used to assign fragments to spectra to rank the candidates. The CFM-ID web server provides a friendly and fast web interface to assist interpretation of tandem mass spectrometry data. Some other commercial programs (such as Mass Frontier (Thermo Scientific) and MS Fragmenter (ACD Labs)) have been developed which are rule-based, using thousands of manually curated patterns to predict fragmentations (for both EI and ESI).

1.9. Acute risk assessment in aquatic environment

Modern societies largely depend on a wide range of down-the-drain products, such as personal care products or household washing agents, containing multiple chemical compounds that finally end up in the aquatic environment, together with their environmental transformation products and manufacture by-products. Increasing contamination of freshwater resources with chemical pollutants has therefore become a major public concern in almost all parts of the world [69], resulting in the introduction of respective chemical regulations to assess associated risks and to ensure the restriction or ban of the most problematic compounds. In Europe, about 100,000 industrial chemicals are registered under the REACH Regulation, of which 30,000 to 70,000 are in daily use [70]. Any new compounds identified through non-target screening should be evaluated for any substantial toxicity in the aquatic environment. Some reports have also reported the excess toxicity of TPs or even degradation products of pharmaceuticals, biocides and pesticides in contrast to their parent compounds [71]. This raises concerns and expectations to have computational tools for quick estimation of toxicity of compounds when there is no reference standard available.

1.9.1. Baseline toxicity and prediction of acute toxicity

Chemicals are persistent and bioaccumulating, they may pose a hazard to the environment by acting as baseline toxicants (baseline toxicity or narcosis). Effect

concentrations of these neutral compounds are generally well predicted by any QSARs that include hydrophobicity in the model's structure. This is often illustrated by octanol-water partition coefficient (log Kow) whereas these baseline QSARs underestimate the acute toxicity for compounds with diverse polarity. Better surrogates are membrane vesicles created from phospholipid bilayers, so-called liposomes [72]. When the liposome-water partition coefficient (log Klipw) is used as descriptor in QSARs of toxicity, nonpolar and polar compounds fall on one regression line.

1.9.2. In silico risk assessment

The Ecological Structure Activity Relationships (ECOSAR available in EPI Suite EPA (US Environmental Protection Agency)) class program is a well-known predictive system that estimates aquatic toxicity. The tool estimates a chemical's acute (short-term) toxicity and chronic (long-term or delayed) toxicity to aquatic organisms, such as fish, aquatic invertebrates, and aquatic plants. Some other tools such as ToxTree follows classification based approach to find the toxicity and it is only limited to a few endpoints (this does not include the aquatic toxicity) and more generic applications (mutagenicity, degradation, DNA binding alerts etc). ToxTrAMS is a novel executable *in silico* toxicity program that has been developed in University of Athens in 2017 [73]. It includes a large amount of data in four spices (Pimephales promelas, daphnia magna, tetrahymena pyriformis and Pseudokirchneriella subcapitata) in the aquatic environment which is frequently used to assess acute toxicity. The main advantage of ToxTrAMS is the structure alert and estimation of toxicity in both classification and regression basis. The application domain of toxicity models poses significant challenges. The lack of available knowledge about the mechanisms of toxicity for many of the endpoints makes it impossible to apply deductive approaches and select appropriate compounds/training set to model. Therefore, "mechanistic interpretation of toxicity" comes after development of models and assessment of the molecular descriptors. VEGA is also another famous and freely available toolbox which offers tens of models (mainly based on CAESAR models) [74] for properties such as persistence, logP, bioconcentration factor (BCF), carcinogenicity, mutagenicity, skin sensitization.

1.9.3. Read across approach

Another method that is used to estimate the toxicity is case-specific read-across [75]. Read across toxicity approach is generally defined as a data gap-filling procedure in which the (aquatic) toxicity of a compound is considered to be equal to (the average toxicity of) similar and relevant chemical structure. Therefore, the known experimental data of similar compounds can be used for suspect compounds. The important steps in any read across approach are to assess (1) the similarity between target(s) and suspect compound(s) and (2) the uncertainties included in the read across workflow. A comprehensive basis is documented by REACH (Regulation (EC) No 1907/2006).

1.9.4. Derivation of predicted no-effect concentration

Predicted No-Effect Concentration (PNEC) is the concentration of a compound at which its adverse effects will most likely not emerge during long term or short term exposure. In the environmental risk assessment, PNECs are compared to the actual or semi-quantified concentration (AC) to determine whether the risk of a compound is acceptable or not. If AC/PNECs<1, the risk is not significant. The PNECs are usually calculated by dividing toxicological dose value (lethal concentration or median effect concentration) by an assessment factor. The median effective concentration (EC50) is the concentration of a compound in an environmental medium expected to produce a certain effect in 50% of test organisms (usually planktonic crustacean Daphnia) in a given population under a defined set of conditions. The Lethal Concentration 50 (LC50) is the concentration of a compound in water causing a death (50% of the tested population) to aquatic life. The EC50 and the LC50 are often used in ecotoxicology as an indicator of the toxicity of a compound to the environment. Assessment factors (AFs) are used to address the differences between laboratory data and real conditions, taking into account of interspecies and intraspecies differences. Assessment factors applied for long-term tests are smaller because the uncertainty of the extrapolation from labs to real environmental condition is reduced.

1.10. Automated suspect and non-target screening

All the information provided above are required for comprehensive identification of chemical profile of a sample and the effects that they could impose to the environment.

Yet another challenge with HRMS instruments is the derivation of massive amount of data which increase the workload of their subsequent evaluation. In addition, acquisition of full scan and MS/MS information simultaneously would provide even more data in a single run. To this end, semi-automated data-processing tools, that could incorporate all the methods which help increase the identification level (such as molecular formula matching, isotopic fitting, retention time prediction or MSMS match between reference standard and the observed spectra), are necessary [10, 76]. There are several computer assisted programs publicly available to overcome some steps (prioritization of peaks-list, detection of adducts or homologue series, chemometric analysis and *in silico* fragmentation tool (for ranking the structure based on the explained MSMS fragments) and their effects and fate in the environment [77]) in the data processing. Table 1.1 provides the overview of these tools with brief information about the extension of their application.

Table 1.1 Publicly available data analysis tools for suspect/non-target screening in HRMS

Tools	Functionality	Instrument Data Type	Year	Reference
METLIN	Metabolite Searching/ Level 3 - Tentative Candidates (*in-silico* MSMS spectra)/ Level 2a - Library Spectrum Match	GC/LC-HRMS	2018	https://metlin.scripps.edu/
MetaboAnalyst v4.0	Preprocessing/ Statistical Analysis/ Metabolomics Pathway analysis/ Level 3 - Tentative Candidates	GC/LC-HRMS	2018	https://www.metaboanalyst.ca/
metabomxtr	Statistical Analysis	GC/LC-HRMS	2018	https://www.bioconductor.org/packages/release/bioc/html/metabomxtr.html
metaMS	Preprocessing	GC/LC-HRMS	2018	https://www.bioconductor.org/packages/release/bioc/html/metaMS.html
MS-DIAL	data independent MS/MS deconvolution/ Annotation/ Statistical Analysis/ Level 2a - Library Spectrum Match	GC/LC-HRMS/ MS/MS	2018	http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/index.html
MS-FINDER	Annotation of peaks-list/ Statistical Analysis/ Level 3 - Tentative Candidates	LC-HRMS/ MS/MS	2018	http://prime.psc.riken.jp/Metabolomics_Software/MS-FINDER/index.html
MZmine	Pick peaking/ Annotation/ MS/ Level 3 - Tentative Candidates	CE/GC/LC-HRMS/ MS/MS	2018	http://mzmine.github.io/
XCMS	Pick peaking	GC/LC-HRMS/ MS/MS	2018	http://bioconductor.org/packages/release/bioc/html/xcms.html
KPIC2	Pick peaking/ Annotation/ Statistical Analysis	GC/LC-HRMS	2017	https://github.com/hcji/KPIC2
apLCMS	Preprocessing	LC-HRMS/ LC-FT-MS	2017	http://web1.sph.emory.edu/apLCMS/
cosmiq	Preprocessing	GC/LC-HRMS	2017	http://bioconductor.org/packages/release/bioc/html/cosmiq.html
nontarget	Annotation/ Homologue series detection	LC-HRMS	2017	https://cran.r-project.org/web/packages/nontarget/index.html
eMZed	Annotation/ MS/ Statistical Analysis/ Level 3 - Tentative Candidates	LC-HRMS	2017	http://emzed.ethz.ch/index.html

Tools	Functionality	Instrument Data Type	Year	Reference
AntDAS	Peak picking/ Annotation/ MS/ Level 3 - Tentative Candidates	LC-HRMS	2017	http://software.tobaccodb.org/software/antdas
IPO	Optimisation of peak picking parameters	LC-HRMS	2017	https://www.bioconductor.org/packages/release/bioc/html/IPO.html
MetaX	Workflow	LC-HRMS	2017	http://metax.genomics.cn/
MetMSLine	Workflow	LC-HRMS	2017	http://wmbedmands.github.io/MetMSLine/
NOREVA	Preprocessing/ Statistical Analysis	GC/LC-HRMS	2017	http://idrb.zju.edu.cn/noreva/
BatMass	Quality control and data exploration for Proteomics and Metabolomics	LC-MS	2016	https://pubs.acs.org/doi/10.1021/acs.jproteome.6b00021
MetFrag2.2	Prioritization of candidates/ MS/MS in-silico fragmentation tool	MS/MS	2016	https://msbi.ipb-halle.de/MetFragBeta/
PlantMAT	Phytochemical knowledge for the prediction of plant natural products such as saponins and glycosylated flavonoids	LC-HRMS	2016	https://sourceforge.net/projects/plantmat/
enviMass	Peak picking/ Annotation/ Statistical Analysis/ Level 4 - Unequivocal Molecular Formula	GC/LC-HRMS	2016	https://www.looscomputing.ch/eng/enviMass/ overview.htm
Ionwinze	Statistical Analysis	LC-HRMS	2016	https://sourceforge.net/projects/ionwinze/
MetFamily	Annotation/ MS/ Level 3 - Tentative Candidates	LC-HRMS/MS/MS	2016	http://msbi.ipb-halle.de/MetFamily/
mzOS	Annotation/ MS/ Level 3 - Tentative Candidates	LC-HRMS	2016	https://github.com/jerkos/mzOS
XCMS Online	Pick peaking/ Annotation/ MS/ Level 3 - Tentative Candidates	GC/LC-HRMS	2016	https://xcmsonline.scripps.edu/
enviPick	Pick peaking/ Annotation/ Level 4 - Unequivocal Molecular Formula	LC-HRMS	2016	https://cran.r-project.org/web/packages/enviPick/index.html
SMART	Statistical Analysis	GC/LC-MS	2016	http://www.stat.sinica.edu.tw/hsinchou/metabolomics/SMART.htm

Tools	Functionality	Instrument Data Type	Year	Reference
geoRge	Analyzing untargeted LC/MS data from stable isotope-labeling experiments	LC-HRMS	2016	https://github.com/jcapelladesto/geoRge
FlavonQ	Automated Data Processing Tool for Profiling Flavone	LC-HRAM-MS	2015	https://pubs.acs.org/doi/10.1021/acs.analchem.5b02624
Mass Frontier (commercial)	MS/MS *in-silico* fragmentation tool	LC-HRMS	2015	http://www.highchem.com/manual/
MET-XAlign	Preprocessing	LC-HRMS	2015	http://bioinfo.noble.org/manuscript-support/met-xalign/
Metabolome Searcher	Annotation/ MS/ Level 3 - Tentative Candidates	LC-HRMS	2015	http://procyc.westcent.usu.edu/cgi-bin/MetaboSearcher.cgi
mzMatch	Preprocessing/ Annotation/ MS/ Level 3 - Tentative Candidates	LC-HRMS	2015	http://mzmatch.sourceforge.net/index.php
xMSanalyzer	Preprocessing	LC-HRMS	2015	https://sourceforge.net/projects/xmsanalyzer/
MET-COFEA	Pick peaking	GC/LC-HRMS	2014	http://bioinfo.noble.org/manuscript-support/met-cofea/
CFM-ID 2.0	Prediction of MS/MS spectra (Electrosparay ionization and Electron Impact)	MS/MS	2014	http://cfmid.wishartlab.com/
AStream	Annotation/ MS/ Level 4 - Unequivocal Molecular Formula	LC-HRMS	2014	http://www.urr.cat/AStream/AStream.html
HayStack	Statistical Analysis	LC-HRMS	2014	http://binf-app.host.ualr.edu/haystack/
MarVis	Pathway Analysis/ MSEA		2014	http://marvis.gobics.de/
MetaboliteDetector	Preprocessing	GC-HRMS	2014	http://md.tu-bs.de/
MS2Analyzer	Annotation/ MS/ Level 2a - Library Spectrum Match	LC-HRMS/ MS/MS	2014	http://fiehnlab.ucdavis.edu/projects/MS2Analyzer/
TNO-DECO	Preprocessing	GC/LC-HRMS	2014	https://github.com/NetherlandsMetabolomicsCentre/TNO-DECO
TracMass2	Pick peaking	GC/LC-HRMS	2014	http://pubs.acs.org/doi/suppl/10.1021/ac403905h
decoMS2	Preprocessing	LC-HRMS	2013	http://pattilab.wustl.edu/software/decoms2/decoms2.php
MAVEN	Annotation/ MS/ Level 3 - Tentative Candidates	LC-HRMS	2013	http://genomics-pubs.princeton.edu/mzroll/index.php
MetaboQC	Optimisation	LC-HRMS	2013	http://evolution.haifa.ac.il/index.php/component/k2/item/146

Tools	Functionality	Instrument Data Type	Year	Reference
AMDORAP	Preprocessing	LC-HRMS	2012	http://amdorap.sourceforge.net/
MetaboSearch	Annotation/ MS/ Level 3 - Tentative Candidates	LC-HRMS	2012	http://omics.georgetown.edu/metabosearch.html
MetAlign	Preprocessing	GC/LC-HRMS	2012	http://www.wageningenur.nl/en/show/MetAlign-1.htm
MetExtract	Preprocessing	LC-HRMS	2012	https://code.google.com/archive/p/metextract/
Sequential design of experiments (DoE)	Optimisation of peak picking parameters	LC-HRMS	2012	https://pubs.acs.org/doi/10.1021/ac301482k
CAMERA	Annotation/ MS/ Level 4 - Unequivocal Molecular Formula	LC-HRMS	2012	http://bioconductor.org/packages/release/bioc/html/CAMERA.html
MeDDL	Visualization and analysis of small molecule metabolite GC−MS and LC−MS data for biomarker discovery	GC/LC-HRMS	2010	https://pubs.acs.org/doi/pdf/10.1021/ac100034u
MZedDB	Annotation/ MS/ Level 4 - Unequivocal Molecular Formula	LC-HRMS	2009	http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html
MathDAMP	Statistical Analysis	CE/GC-HRMS	2007	http://mathdamp.iab.keio.ac.jp/

CHAPTER 2

Scope and Objective

2.1. The analytical and computational chemistry problem

There has been growing interests and studies dealing with the occurrence and identification of organic emerging pollutants in various environmental media in the last ten years. The main focus of these publications are the identification and quantitation of new organic micropollutants, their fate in environmental counterparts or their ecotoxicological impact and degradation as well as TPs.

Currently, there are 100000 commercially registered compounds in Europe where the majority of them would end up in the water cycle. This even becomes a larger list considering the predicted metabolites or transformation products of all these pollutants. Specific strategies for target, suspect and non-target screening, together with the development of optimized analytical techniques are needed in the analysis of these chemicals in environmental samples.

The latest advances in HRMS have initiated a new trend in analytical data processing in recent years. This has led to opening Pandora's Box of chemical inventories needed to be screened in a routing basis for water monitoring quality. Targeted analytical methods are now often complemented with suspect and non-target data screening methods to trace presence of any new compound [78, 79].

As aforementioned, suspect and non-target screening is a more challenging and time-consuming task and often success is not granted. In this case, supportive experimental and *in silico* approaches are needed to be followed in order to find the proper way to identify compounds.

2.2. Research Objectives and Scope

The thesis is consisted of four studies. The first two studies are about application of retention time data in suspect and non-target screening whereas the other two chapters discuss about the aquatic risk assessment and development of automatic computational approach for suspect and non-target screening, respectively.

In the first study, a novel and comprehensive workflow was developed to study the retention time behavior of large groups of emerging contaminants using QSRR. 682 compounds were analyzed by HILIC-HRMS in positive ESI. Moreover, an extensive dataset was built for RPLC-HRMS including 1830 and 308 compounds for positive and negative ESI, respectively. SVM was used to model the retention time data. The applicability domains of the models were studied by Monte Carlo Sampling (MCS) methods. The MCS method was also used to calculate the acceptable error windows for the predicted retention time from various LC conditions. This study provided validated models for predicting retention time in HILIC/RPLC-HRMS platforms to facilitate identification of new emerging contaminants by suspect and non-target screening. Furthermore, these models were applied for the identification of transformation products (TPs) of emerging contaminants.

In the second study, a new method was developed to establish RTI for liquid chromatography. A true positive identification in case of unknowns requires supporting orthogonal information such as elution and mass spectrometric pattern matching analytical reference standards. In this regard, there is an emerging need to compare and harmonize liquid chromatographic retention time information between different labs for increasing the identification confidence in non-target or suspect screening workflows and provide evidence of the existence of a compound in any sample. This requires a unified index that is flexible and less system-dependent. Here we developed a RTI system for LC based on calibration of elution pattern using set of substances representative of the appropriate chemical space. These calibrants were selected chemometrically, using newly developed ant colony optimization similarity indices. This approach selects a set of compounds with maximum overlap with the retention times and chemical similarity indices with the rest of the compounds of the chemical space (here 2123 compounds) after appropriate training. A calibration set of 18 compounds with RTI set to range between 1 and 1000 was proposed as the most appropriate RTI system after rigorous intralaboratory evaluation. An interlaboratory comparison was coordinated within the NORMAN network to validate the proposed RTI system externally on completely different instrumentation and LC conditions. The proposed RTI system found to promote a higher confidence suspect screening via reduction of false positives and increase the comparability between

laboratories by allowing a comparison of retention times for standards measured in other laboratories.

The third part of the thesis was devoted to the development of a robust risk assessment approach for aquatic environment. A large dataset was compiled, with the experimental acute toxicity values (pLC_{50}) of 1353 compounds in Daphnia magna after 48-h of exposure. A novel quantitative structure–toxicity relationship (QSTR) model was developed, using Ant Colony Optimization to select the most relevant set of molecular descriptors, and SVM to correlate the selected descriptors with the toxicity data. A new method was also proposed to define the chemical space failure for a compound with unknown toxicity to avoid using the prediction results. The resulting ACO-SVM model was successfully applied on an additional evaluation set and the prediction results were found to be very accurate for those compounds that fall inside the defined applicability domain. In fact, compounds commonly found to be difficult to predict, such as quaternary ammonium compounds or organotin compounds were outside the applicability domain, while five representative homologues of LAS (non-ionic surfactants) were, on average, well predicted within one order of magnitude

Finally, the last study was to propose a workflow to screen a regulatory database in environmental related samples such as influent/effluent wastewater and sewage sludge samples (collected from the WWTP of Athens (Greece)). Both the so called “AutoSuspect” and “AutoNontarget” start with an optimized peak picking algorithm, using XCMS with IPO package behind the optimization task. This can be done either on the MS information recorded in data independent/dependent acquisition mode. Componentization and annotation of peaks list are then achieved using “CAMERA” and “nontarget” R package [80, 81]. This help assign chemical formula from regulatory database/online chemical databases (PubChem, ChemSpider etc) to every m/z in the peaks-list and focus on those m/z that are found to be potential precursor ion. The theoretical isotopic pattern was calculated for these chemical formula by “enviPat” [82] and then compared with extracted experimental isotopic pattern to exclude false chemical formula from hit list. All retention time of the detected m/z are converted to RTI through the calibration approach, described in second chapter of this thesis, to facilitate

identification procedure. The remaining candidates in the hit list are being evaluated by the available experimental MS/MS information MetFrag (*in silico* fragmentation approach) and library spectrum (MassBank, MoNA and Metlin). 40,053 compounds from Norman Susdat (list of compounds found mainly in aquatic environment, *http://www.norman-network.com/?q=node/236*), a list of 273 biocides and 70 million compounds (PubChem) have been screened by AutoSuspect and AutoNonTarget in the collected samples.

CHAPTER 3

Development and Application of Retention Time Prediction Models (RPLC/HILIC-QToF-MS) in the Suspect and Non-target Screening of Emerging Contaminants

3.1. Introduction

Nowadays, Liquid chromatography (LC) coupled to high resolution mass spectrometry (HRMS) plays a key role in the identification of new ("emerging") micropollutants in the aquatic environment [2, 83]. Two parallel approaches can be followed for the identification of emerging compounds that are not available as reference standards, namely suspect and non-target screening [37, 84, 85]. Schymanski et al. proposed a scheme for reporting the identification confidence, where the interpretation of fragmentation pattern in the deconvoluted MS/MS spectra, retention time (t_R) information (in addition to mass accuracy and the isotopic pattern of the precursor ion) are included as supporting experimental evidence for identification and chemical structural elucidation [85]. Knowledge of t_R can also help reduce the number of plausible candidates and, subsequently, increase the chance of true identification [37, 86]. Since the polar micropollutants and their TPs are the major focus in the aquatic environment [3], the complimentary use of HILIC with RPLC can provide additional experimental evidence and support to the identification of new compounds in the environment [37]. Nevertheless, the structure elucidation of isomeric compounds or TPs based only on their fragmentation pattern, may sometimes not be feasible, since they produce common fragments and the reference standards are not always available [84, 87]. In those cases, retention time prediction could support identification.

Several approaches have been presented to predict t_R in LC [35, 61, 88-97]. However, the accurate prediction of t_R for emerging contaminants has remained a challenge due to the lack of appropriate and wide dataset of t_R values, the non-representative selection of molecular descriptors with sophisticated methods to cover their diverse chemical structures and t_R elution behavior [35, 61, 89-97]. Tyrkko et al. used ACD/ChromGenius to predict t_R and applied it for the identification of unknowns, however the prediction error was large for most of the polar compounds and required the use of experimental confirmation to explain the origin of error [96]. Apart from previous studies which have a limited applicability domain or showed high prediction errors, Falchi et al. [98] followed a

robust workflow and proposed a model based on the combination of physicochemical properties and fingerprint information of more than 1383 synthetic compounds. While the effect of geometry optimization of chemical structures on prediction of t_R has remained vague, studies in which the optimization of the chemical structures was performed prior to modeling resulted in higher accuracy [95, 97-99]. Although the origin of error between the experimental and predicted t_R was investigated in a few studies [48, 95, 97], there is no clear agreement over acceptable error windows for predicted t_R. Relative acceptance windows was proposed as a way to include the effect of chemical structures in an earlier study [9]. With this approach, compounds that were similar to the compounds of the training set had a narrower acceptance windows compared to those that were less similar [9].

Although the use of HILIC is increasing as a complementary method to cover highly polar compounds and metabolites [37], few studies have reported modeling HILIC t_R [61, 99]. Therefore, there is a need for the prediction of t_R for tentatively identified polar micropollutants in HILIC. There is a few prior information of molecular descriptors available for HILIC. Creek et al. applied HILIC and t_R prediction in metabolite identification, [61] using logD and two charge-related molecular descriptors that comprised of pH, pK_a and formal charge state for 120 compounds. However, inter-correlation between logD and the charge related descriptors was observed. Structural based models, i.e. QSRR, capable of searching chemical space to define the correct polarity value for a compound may help to understand the elution mechanism in HILIC. A rigorously validated QSRR model with a wide applicability domain and no over-fitting can provide prediction results for any structure of interest (eluted in LC) with high accuracy.

The objectives of the current study were: (a) the development of validated QSRR models with a novel workflow and broad applicability domain for RPLC and HILIC HRMS platforms; (b) the development of a novel and easy-to-use visualization methods to provide information about the origin of error in predictions using MCS; (c) the development of a novel approach to define the acceptable error windows for predicted t_R; and (d) the demonstration of the applicability of QSRR models in the identification of new TPs of emerging contaminants and biocides in environmental samples.

3.2. Materials and methods

3.2.1. Sample preparation, instrumental analysis and dataset

The pesticide reference standards were donated by Bruker Daltonics (Bremen, Germany), at a concentration of 1 mgL^{-1} in methanol. The remaining standards were purchased from Sigma–Aldrich (Germany)**.** Individual stock solutions were prepared in methanol at 1 g L^{-1} and stored at -20 °C. Then, working solutions were prepared in methanol at a concentration of 1 mg L^{-1}. Acetonitrile (ACN) and methanol (MeOH), LC-MS grade, was purchased from Merck (Germany), whereas 2-propanol of LC-MS grade was from Fisher Scientific (Geel, Belgium). Sodium hydroxide monohydrate (NaOH) for trace analysis ≥99.9995%, ammonium acetate, ammonium formate and formic acid, all LC-MS grade, were purchased from Fluka, Sigma–Aldrich (Germany). Distilled water used for LC–MS analysis was provided by a Milli-Q purification apparatus (Millipore Direct-Q UV, Bedford, MA, USA). Regenerated cellulose (RC) syringe filters (15 mm diameter, 0.22 μm pore size) were provided from Phenomenex (Torrance, CA, USA).

An ultrahigh-performance liquid chromatography (UHPLC) system with a LPG-3400 pump (Dionex UltiMate 3000 RSLC, Thermo Fisher Scientific, Germany), interfaced to a QToF mass spectrometer (Maxis Impact, Bruker Daltonics, Bremen, Germany) was used for the screening analysis.

In RPLC, the chromatographic separation was performed on an Acclaim RSLC C18 column (2.1 × 100 mm, 2.2 μm) from Thermo Fisher Scientific (Driesch, Germany) preceded by a guard column, ACQUITY UPLC BEH C18 1.7 μm, VanGuard Pre-Column, Waters (Ireland), held at 30 ˚C. Mobile phase composition in positive ionization mode (PI) was (A) H_2O:MeOH (90:10) with 5 mM ammonium formate and 0.01% formic acid and (B) MeOH with 5 mM ammonium formate and 0.01% formic acid. For negative ionization mode (NI), the mobile phase was (A) H_2O:MeOH (90:10) with 5 mM ammonium acetate and (B) MeOH with 5 mM ammonium acetate. The gradient elution program was the same for the two ionization modes, lasting a total of 15.5 min, with 5 min of re-equilibration of the column for the next injection, as follows: 1% B (0.2 mL min^{-1}) for 1 min, increasing to 39 % in 2 min (flow rate 0.2 mL min^{-1}), and then to 99.9 % (flow rate 0.4 mLmin^{-1}) in the following 11 min. Then, it is held constant for 2 min (flow rate 0.48 mL min^{-1}) and then

initial conditions were restored within 0.1 min and the flow rate decreased to 0.2 mL min^{-1}. The injection volume was 5 µL.

In hydrophilic interaction liquid chromatography (HILIC), separation was performed on an ACQUITY UPLC BEH Amide column (2.1 × 100 mm, 1.7 µm) from Waters (Dublin, Ireland) preceded by a guard column of the same packaging material, kept at 40 ˚C. For PI, the aqueous phase consisted of H_2O with 1 mM ammonium formate and 0.01% formic acid and the organic phase was ACN/$H_2$0 95/5 with 1 mM ammonium formate and 0.01% formic acid. The adopted elution gradient started with 100% of organic phase, held for 2 minutes, decreasing to 5 % in 10 min, and held for the following 5 min. The initial conditions were restored within 0.1 min and held for 8 min. The flow rate was 0.2 mL/min and the injection volume was set to 5 µL.

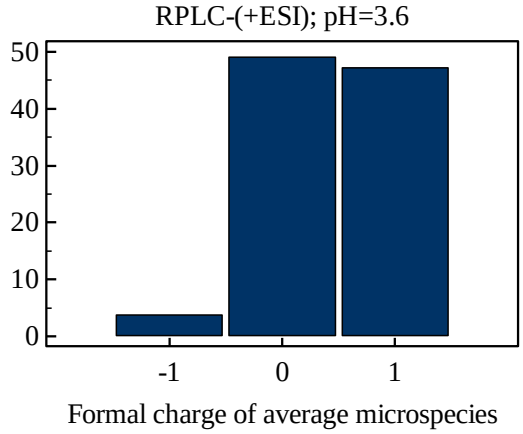
The operating parameters of the electrospray ionization interface (ESI) were for PI mode: capillary voltage, 2500 V; end plate offset, 500 V; nebulizer, 2 bar; drying gas, 8 L min^{-1}; dry temperature, 200 °C.

The QToF MS system was operated in broadband collision induced dissociation (bbCID) acquisition mode over the range *m/z* 50−1000 with a scan rate of 2 Hz. The Bruker bbCID mode provides MS and MS/MS spectra at the same time, at two different collision energies. At low collision energy (4 eV), MS spectra were acquired and at high collision energy (25 eV), fragmentation is taking place in the collision cell resulting in MS/MS spectra.

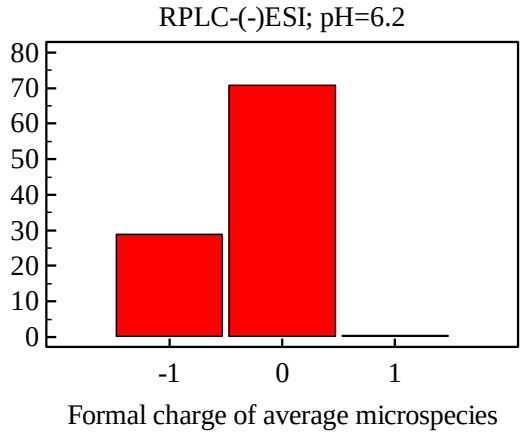
A QToF MS external calibration performed daily with a sodium formate solution, and a segment (0.1−0.25 min) in every chromatogram was used for internal calibration, using a calibrant injection at the beginning of each run. The sodium formate calibration mixture consisted of 10 mM sodium formate in a mixture of water/isopropanol (1:1). The theoretical exact masses of calibration ions in the range of 50−1000 Da were used for calibration. The instrument provided a typical resolving power of 36000−40000 during calibration (39274 at *m/z* 226.1593, 36923 at *m/z* 430.9137, and 36274 at *m/z* 702.8636). Mass spectra acquisition and data analysis was processed with Data Analysis 4.3 and Target Analysis 1.3 (Bruker Daltonics, Bremen, Germany).

The lists of reference standards used for modeling t_R are given in the supplementary material (SM, chapter 3, appendix B) Tables B.3.1, B.3.2 and B.3.3. The formal charge

of the compounds (their average microspecies) recorded in RPLC/HILIC with corresponding pH value was calculated using ChemAxon (“Partitioning(logD)” plugin, Marvin v6.3.1) to find the distribution of neutral/anionic/cationic compounds for each chromatographic system. The dataset compiled for RPLC-(+) ESI mode included 898 neutral, 69 anionic and 863 cationic compounds. The dataset for RPLC-(-)ESI had 218 neutral, 89 anionic and one cationic compound. Finally, the dataset compiled for HILIC-(+)ESI had 311 neutral, 25 anionic and 346 cationic compounds. The distribution of neutral/anionic/cationic compounds for each chromatographic system is also illustrated in the Figure 3.1. There were insufficient compounds amenable to HILIC-(-)ESI to form a sufficiently large dataset for this work.



Neutral compounds: 898
Anionic compounds: 69
Cationic compounds: 863



Neutral compounds: 218
Anionic compounds: 89
Cationic compounds: 1 (Triflumizole)

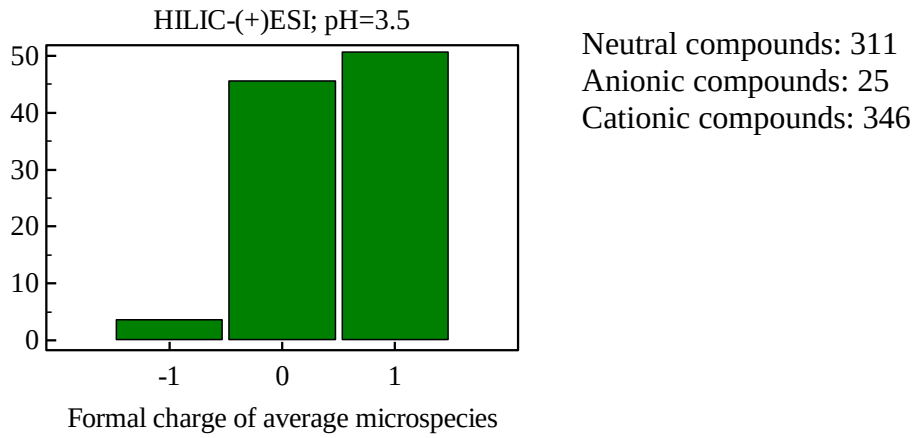


Figure 3.1 Distribution of cationic, anionic and neutral compounds in RPLC and HILIC

3.2.2. QSRR workflows

The t_R for each ESI mode (positive, negative) was modelled separately. The geometries of all chemical structures were optimized using MOPAC2016 (also available online at http://www.scbdd.com/mopac-optimization/optimize/) [100]. The semi-empirical (AM1) [101, 102] method was used to achieve the best geometrical conformer (lowest intermolecular energy). Molecular features of the optimized compounds were calculated using the E-dragon software (available online at http://www.vcclab.org/lab/pclient/) [103]. In addition, the lipophilicity of the optimized compounds in the aqueous phase at various pH (log D), were calculated at pH 3.6 (RPLC, ranging between -7.576 and 8.672) and pH 3.5 (HILIC, range -10.458 to 11.124) for positive ESI and at pH 6.2 for negative ESI (RPLC, range -6.700 to 6.325), using ChemAxon [104]. The dataset, including the molecular features with experimental t_R generated for each condition, was pre-treated by removing the constant and near constant molecular descriptors and further checked for the existence of co-linearity. The remaining molecular features were split into training and test sets using the affinity propagation method [105]. Here, the similarities between pairs of compounds were used as input to affinity propagation. Real-valued messages were exchanged between compounds until a high-quality set of exemplars and related clusters was derived gradually [105]. The genetic algorithm, written in MATLAB 8.5 [106], was used to select the most relevant molecular descriptors that correlated to the t_R. The selected descriptors were correlated linearly and non-linearly to t_R using Multiple Linear

Regressions (MLR) and SVM, respectively. The accuracy of the models built to predict t_R was investigated using an external test set and cross-validation techniques.

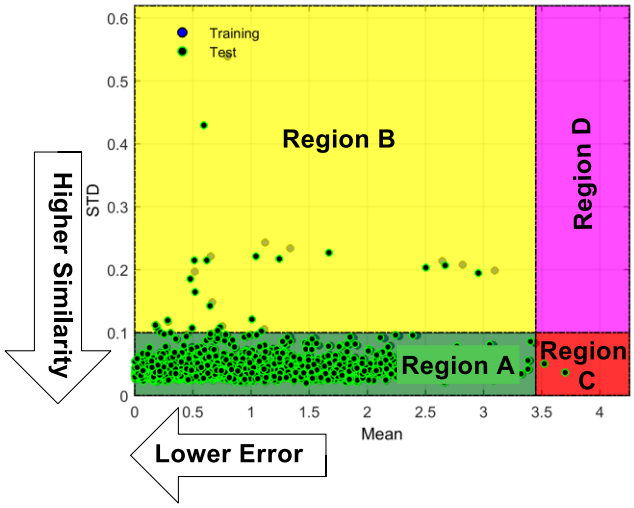
The internal accuracy for the proposed models was studied using concordance correlation coefficient values [107], correlation coefficient, F-value, Root Mean Square Error (RMSE), Y-randomization (shuffling the experimental t_R), R^2_p parameter (penalizes the correlation coefficient of a model) [108] and cross-validation techniques (leave one out and leave group out) [109]. For evaluation of the external predictive ability, the validation approaches introduced by Chirico and Gramatica [110, 111] were followed. High prediction accuracy is accomplished for external application if the models meet the criteria of acceptance for Concordance Correlation Coefficient (CCC) [110], OECD guidelines (Q^2_{F1}& Q^2_{F2}& Q^2_{F3}) [112], modified correlation coefficient (r^2_m) and Golbraikh-Tropsha method [113]. The details about the mathematical terms and formula of each validation procedure can be found in supplementary materials (Chapter 3, appendix B) Table B 3.4.

3.2.3. Applicability domain studies and tR acceptable error windows

The origin of residuals (error) between the experimental and predicted t_R occur mainly due to either errors in the reported t_R or chemical structural diversity from the training set used to build the model. Different methods are available to assess the presence of outliers [9, 114]. A method called “OTrAMS” presented in [9] was developed to not only define the applicability domain of the models, but also to decrease the chance of false positive structures in the case of suspect and non-target screening [9]. This method was established based on the effect of chemical structural diversity, standardized residuals (SDR) of predictions and the leverage value of each compound, which is proportional to the Hotelling T^2 and Mahalanobis distance. This approach provides a quick overview about the origin of errors in t_R information. More details about OTrAMS can be found in the supplementary material (SM, Chapter 3, appendix A), section SM 2.1.

In addition to OTrAMS, another outlier detection procedure was developed using MCS [115] to understand the origin of errors in t_R modelling. It is a robust technique to detect different kinds of outliers by developing many cross-predictive models. The results of this procedure are displayed by plotting the absolute values of mean of predictive residuals

(MEAN) versus standard deviations of predictive residuals (STD). The cut-off limit for MEAN and STD are defined based on the population density of compounds in the training set. Figure 3.2 shows the interpretation of MCS results. MCS was set to 5000 iterations. As shown in Figure 3.2, four regions are defined for interpreting outliers; the lower left area (region A) shows the data that are not outliers; the top left region (region B) of the plot shows the data points that are outliers due to structural diversity, but they can be still modelled; the bottom right area (region C) represents the samples that are outliers due to the observed t_R values (the area where the potential false positives exist); and the top right region (region D) of the plot displays the outliers due to large structural diversity and ambiguous observed t_R. Cut-off values for MEAN and STD were determined based on the distribution density of MEAN and STD of the training set.



Region A (color green): Compounds within the applicability domain; observed t_R is accepted.
Region B (color yellow): Structurally diverse compounds, t_R can still be accepted, but additional verification might be needed.
Region C (color red): Wrong t_R, the observed t_R cannot be accepted. These are potential false positives.
Region D (color magenta): Structurally diverse compounds that are outside the applicability domain. Other verification tools are needed.

Figure 3.2 The explanation of Monte-Carlo sampling (MCS) method used to define the applicability domain of the models developed to predict tR

MCS was also used to define acceptable error windows for predicted t_R. The strategy used to calculate the acceptable error windows was to find a threshold where 95% of the MEAN values (in MCS plot) locate. Therefore, the 95th quantile of MEAN was calculated, which is the mean of the prediction errors of each sample at 5000 times MCS, and used to derive the error windows. This approach was further tested on 13 datasets extracted from MassBank spectral records (http://massbank.eu/MassBank/, last visit July 2018). The details about the LC conditions of these datasets can be found in supplementary materials (Chapter 3, appendix B) Table B 3.5.

3.2.4. Experimental setup for the generation and identification of TPs of selected emerging contaminants

Ozonation batch experiments were conducted in sealed bottles by mixing a predefined amount of ozone saturated solution with an aqueous solution of selected emerging contaminants (tramadol, furosemide and niflumic acid), following the procedure already described in a previous study [116]. These compounds are often detected in effluents (incomplete removal) and the receiving environment [38, 117, 118] and data for their ozonation TPs is scarce. The identification workflow, along with RPLC/HILIC complementary analyses, is described in [36, 116, 119]. The chemical structure of these pharmaceuticals with their drugbank ID can be found in supplementary materials (Chapter 3, appendix B) Table B 3.6.

3.3. Results and discussion

3.3.1. RPLC-(+)ESI-HRMS

The best set of five molecular descriptors showing high correlation and prediction accuracy with t_R was selected by Genetic Algorithm (GA). A general linear model for RPLC-(+)ESI-HRMS based on affinity propagation-GA-MLR was obtained with the following equation:

$$t_R = +2.3518(\pm 0.1335) + 0.7371(\pm 0.0204)\,\mathrm{logD}_{(\mathrm{pH\ 3.60})} + 1.2389(\pm 0.0696)\,\mathrm{CIC1} + 0.5584(\pm 0.0396)\,\mathrm{SEigZ} - 0.2198(\pm 0.0340)\,\mathrm{RDF020p} + 0.4155(\pm 0.0306)\,\mathrm{AlogP} \qquad (3.1)$$

logD is the measure of hydrophobicity for the ionizable compounds, CIC1 is the Complementary Information Content index (neighborhood symmetry), SEigZ is the eigenvalue sum from Z weighted distance matrix of a Hydrogen-depleted Molecular Graph, RDF020p is Radial Distribution Function weighted by atomic polarizabilities and AlogP is logP estimated by Ghose–Crippen method [120]. More details about molecular descriptors selected can be found in supplementary materials (appendix A), section SM 2.1.1. The elution of the compounds in RPLC-(+)ESI-HRMS is illustrated in Figure A 3.1.

The proposed model was built based on 1461 compounds in the training set and validated using the techniques described above, including external evaluation on 369 compounds as test set. The statistical parameters, introduced in section 3.3.2, are listed in Table B 3.7 (SM, Chapter 3, Appendix B) and the model meets all acceptance criteria. The OTrAMS results are shown in Figure A 3.2 (a) and demonstrates that no outliers were present for this training set. Over 70% of the whole dataset was predicted with the error less than 1.0 min (6.67 % of LC run time). The Monte-Carlo cross-validation method [115] described above was applied, shown in Figure A 3.2 (b), indicates that the majority of compounds are located in Region A. These results suggest that the model is free from outliers in the training set and is thus well suited for prediction purposes.

The non-linear model was built using the same training set and molecular descriptors. The internal parameters of SVM were optimized based on the RMSE of leave one out cross-validated model as C=50 (a trade-off parameter), ε=0.2 (insensitive loss function), γ=1.5 (radial basis function (RBF)). The result of each optimization step is shown in supplementary material, Chapter 3, appendix A, section SM 2.1.3, Figure A 3.3 (a-c). The predicted and experimental t_R values are listed in Table B 3.1 for RPLC-(+)ESI-HRMS. The comparison results of two models (MLR and SVM) show that SVM has higher internal and external accuracy for prediction of t_R (Table B 3.8). The molecular descriptors used here were investigated for the existence of inter-correlation cases and as listed in Table B 3.9, no cases with high inter-correlation were found.

3.3.2. RPLC-(-)ESI-HRMS

The same workflow was applied to RPLC-(-)ESI-HRMS and the following equation was obtained with eight molecular descriptors selected by GA:

$$t_R = +3.9078\ (\pm 0.2255) + 0.3016\ (\pm 0.0768)\mathrm{XlogP} + 0.6128\ (\pm 0.0543)\mathrm{logD}_{(\mathrm{pH\ 6.20})} + 0.1917\ (\pm 0.0543)\mathrm{RDF130m} - 1.377\ (\pm 0.3291)\mathrm{Mor16p} - 0.3062\ (\pm 0.0695)\ \mathrm{nCconj} + 0.1103\ (0.0178)\ \mathrm{MlogP}^2 + 1.588\ (0.2461)\ \mathrm{B06[C-C]} + 0.6183\ (0.1345)\ \mathrm{F04[Cl-Cl]} \quad (3.2)$$

XlogP is a measure of logP, logD is a measure of logP for the ionizable compounds at pH=6.2, RDF130m is the Radial Distribution Function weighted by atomic mass, Mor16p is 3D-MoRSE weighted by atomic polarizabilities, nCconj is the number of non-aromatic conjugated C(sp2), $MlogP^2$ is the squared Moriguchi octanol-water partition coefficient, B06[C-C] is the presence/absence of C–C (carbon-carbon single bonds) and F04[Cl-Cl] is the frequency of Cl–Cl in a chemical graph. More details about these molecular descriptors can be found in supplementary material, Chapter 3, appendix A, section SM 2.2.1. The overall contribution of molecular descriptors to explain elution mechanism in (-)ESI-RP-LC-HRMS was investigated by Principal Component Analysis (PCA) (Figure A 3.4).

The model was developed based on 247 compounds (training set) and the validation protocols were applied to confirm the predictive power of the model, including external evaluation on 62 compounds. The evaluation of statistical parameters are listed in Table B 3.7. OTrAMS demonstrated that no outliers were detected for the training set (Figure A 3.5 (a)). Over 68% and 26% of the whole dataset (training and test set) were predicted with an error less than 1 min (6.67 % of LC run time) and 2 min (13.34 % of LC run time), respectively. The performance of the –ESI models was lower than those obtained in +ESI mode, due to the smaller dataset, which limits the ability to capture the variations in experimental t_R for these compounds. This is inherent to the ionization technique, as fewer compounds are ionizable in negative mode.

MCS was also used, as described in section 3.2.3, to derive the distribution of compounds based on the origin of their errors. As shown in Figure A 3.5 (b), the majority of compounds are in the area with low prediction errors (Region A and B). The results of

the outlier detection methods suggest that the training set is free of outliers and thus the model is acceptable for prediction purposes.

The non-linear model was also built and compared to affinity propagation-GA-MLR. The internal parameters of SVM were optimized to C=50 (a trade-off parameter), ε=0.08 (insensitive loss function), γ=5 (radial basis function (RBF)). The result of each optimization step is shown in Figure A 3.6 (a-c). The predicted and experimental t_R values are listed in Table B 3.2 for RPLC-(–)ESI-HRMS. Comparison of the two models (MLR and SVM) reveals that SVM has high internal and external accuracy for t_R prediction (Table B 3.8). The molecular descriptors used here were investigated for inter-correlation cases. As shown in Table B 3.10, no indication of inter-correlation is present.

3.3.3. HILIC-(+)ESI-HRMS

The best set of seven molecular descriptors showing high correlation and prediction accuracy with t_R was selected for HILIC by GA. A general linear model for HILIC-(+)ESI-HRMS based on affinity propagation-GA-MLR was obtained with the following equation:

$$\begin{aligned} t_R = {} & +2.591(\pm 0.1323) - 1.233\ (\pm 0.0227)\mathrm{logD}_{(\mathrm{pH\ 3.50})} - 0.1051\ (\pm 0.0204)\mathrm{GGI1} \\ & + 0.2293\ (\pm 0.0384)\mathrm{RDF020p} + 0.2410\ (\pm 0.0322)\mathrm{H{-}050} \\ & + 1.332\ (\pm 0.1769)\ \mathrm{qnmax} + 0.0807\ (0.0089)\ \mathrm{MlogP}^2 \\ & + 0.8120\ (0.0370)\ \mathrm{AlogP} \end{aligned} \qquad (3.3)$$

log D is a measure of log P for the ionizable compounds at pH=3.5, GGI1 is the Radial topological charge index of order 1, RDF020p is Radial Distribution Function weighted by atomic polarizabilities, H-050 is number of H attached to a heteroatom, qnmax is the maximum negative charge, while $MlogP^2$ and AlogP are the measures of logP for neutral compounds [99, 121]. More details about these molecular descriptors can be found in SM (Chapter 3, appendix A), section SM 2.3.1. The contribution of selected molecular descriptors to explain elution mechanism in HILIC-(+)ESI-HRMS was investigated by Principal Component Analysis (PCA) in Figure A 3.7.

The model was built on a training set of 542 compounds. The internal validation was followed as described in the section 3.1 and the external predictive ability of the model was evaluated using a test set of 140 compounds. The statistical parameters for the

developed model are listed in Table B 3.7. Three outliers (Prometryn, Irgarol-descyclopropyl and Arginine) were detected by OTrAMS for the test set (Figure A 3.8 (a)), while no outliers were observed for the training set. All in all, more than 93 % of the whole dataset was predicted with an error less than 1 min (71 %) and 2 min (22 %). MCS was also used. As shown in Figure A 3.8 (b), the majority of compounds are in Region A.

The non-linear model was also built and compared to affinity propagation-GA-MLR. The internal parameters of the SVM were optimized as C=100 (a trade-off parameter), ε=0.01 (insensitive loss function) and γ=3 (radial basis function (RBF)). The result of each optimization step is shown in Figure A 3.9 (a-c). The predicted and experimental t_R values are listed in Table B 3.3 for HILIC-(+)ESI-HRMS. Comparison of the two models (MLR and SVM) indicates that SVM has better internal and external accuracy for t_R prediction (Table B 3.8). Inter-correlation results for the selected molecular descriptors are listed in Table B 3.11.

3.3.4. Acceptable error windows for predicted tR

In order to define acceptable error windows for predicted t_R, experimental retention time data was retrieved from MassBank. Thirteen new QSRR models were developed from these data and evaluated by MCS. The accuracy of the models along with LC conditions and total number of compounds in each model can be found in Table B 3.12 in SM (appendix B). The strategy described above was used to calculate the acceptable error windows. Therefore, the 95^{th} quantile of MEAN from MCS was calculated for each dataset from MassBank. The acceptable error windows in predicted t_R is obtained by the individual MEAN cut-off value for each LC condition in 13 dataset. Table B 3.13 lists the results of various quantile values and acceptable error windows for each LC condition and dataset. This error windows is approximately 12% of the total chromatographic run time or maximum experimental t_R used in the training set during model development. In the view of these results, MCS is a useful technique to define the confidence intervals for t_R prediction and provides a reasonable confidence for the applicability domain of the models in case of suspect/non-target screening.

The identification methodology can be exemplified for the case of 5-Methylbenzotriazole (Figure 3.3) where the MCS plot could successfully distinguished 2-Aminobenzimidazole as false positive. Based on the mass accuracy, isotopic fitting and chromatographic peak

score (Figure 3.3 (a)), two substances were met these conditions (5-Methylbenzotriazole and 2-Aminobenzimidazole), including the interpretation of MS/MS fragments using in silico fragmentations tools (Figure 3.3 (b)). The HILIC t_R prediction model could help to prioritize these suspects according to their degree of MEAN value in MCS plot (Figure 3.3 (c)). The spectra of reference standard was found in MassBank (SM880101) and the fragments (Figure 3.3 (d)) at m/z 53.0383, 79.0540, 80.0572, 95.0485, 105.0437, 106.0646 and 134.0707 fit very well with the prioritized suspect (5-Methylbenzotriazole), corresponding to $[C_4H_5]^+$, $[C_6H_7]^+$, $[C_5H_6N]^+$, $[C_6H_7O]^+$, $[C_6H_5N_2]^+$, $[C_7H_8N]^+$, and $[C_7H_8N_3]^+$, respectively. Therefore, the identification was confirmed by t_R prediction, MCS plot, MS/MS comparison (spectra similarity score of 0.998), and further by corresponding reference standard reaching to level 1.

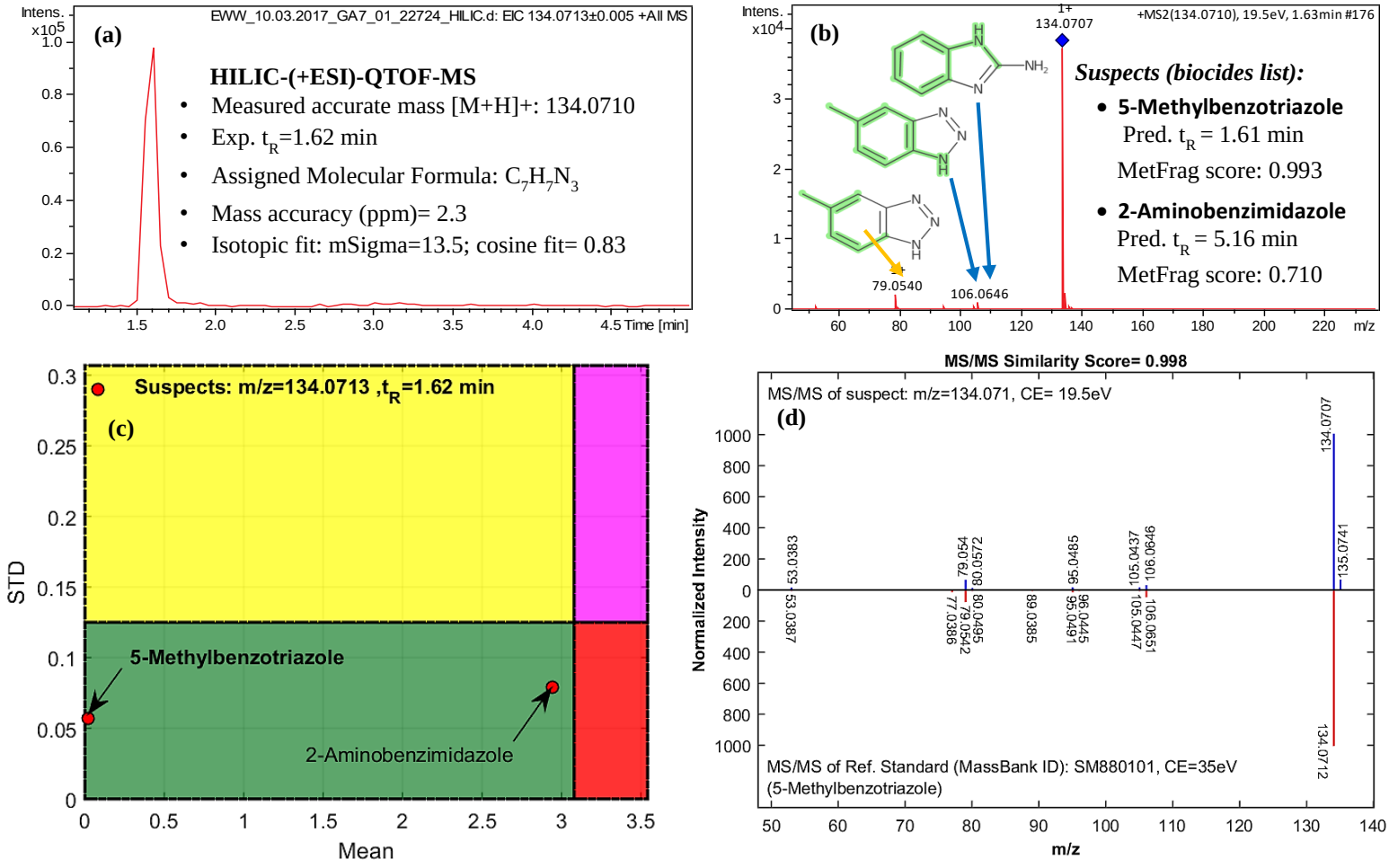


Figure 3.3 Identification of 5-Methylbenzotriazole: (a) full MS chromatogram for the given mass (±5ppm); (b) MS/MS spectra and corresponding fragments; (c) MCS plot for evaluating the predicted tR values; (d) confirmation step using spectra library. 5-Methylbenzotriazole was confirmed by reference standard later.

3.3.5. Comparison with literature models

Several approaches, previously developed and used to predict t_R [9, 35, 44, 61, 89-98, 122, 123], were compared with the work presented here, and are shown in Table B 3.14. The studies [93, 97] that applied non-linear regression methods (such as Artificial Neural Network (ANN) or SVM) modelled t_R with low prediction errors compared with those models that were proposed based on linear regression method (i.e. Partial Least Square (PLS) and MLR) [61, 96]. The studies [92, 97] that standardized the geometry of

compounds in their t_R model were found to be slightly better (in terms of internal fitting and prediction error) than those where no standardization steps were used. [35, 93, 94]. The models developed here for RPLC/HILIC platforms are based on a large number of emerging contaminants and offer high prediction accuracy in contrast to previous studies [37, 61, 97, 99]. Moreover, the applicability domain of the proposed models was carefully defined, which is very crucial for the removal of false positives, in contrast to two previously methods that were built based on large set of emerging contaminants but with no defined applicability domain [92, 97].

3.3.6. Application of tR prediction in the identification of transformation products

All developed models were used for the identification of some new ozonation TPs of emerging contaminants. t_R prediction was used either for enhancing the identification confidence of proposed TPs structures or finding the elution order of isomeric TPs structures.

Three series of ozonation experiments were conducted where the transformation of tramadol (TRA), furosemide (FUR) and niflumic acid (NA) was investigated, following suspect and non-target workflows [37]. Among the identified TPs of TRA after RPLC-(+)ESI-HRMS analysis, TRA-218 and TRA-282 were structurally elucidated based on the interpretation of their MS/MS spectra (Fig. A 3.10(a) and A 3.10(b), respectively). The proposed structures were highly supported by the t_R prediction results, since an error of 0.22 and -0.48 min, respectively, was derived (Table 3.1). Moreover, three isomeric TPs of TRA (with m/z 296) were detected at 3.5, 4.5 and 4.8 min. Based on common reactions between TRA and ozone, three possible structures could fit the proposed formula, following Criegee mechanism reaction. As displayed in Figure A 3.10(c), the MS/MS spectra of the three isomers were almost identical and no diagnostic fragments were detected. The t_R prediction contribution to the identification workflow was significant, since it indicated a distinct chromatographic separation of the three proposed structures, and the experimental t_R were in accordance with the predicted one, with errors ranging from -0.29 to 0.21 min (Table 3.1). Thus, based on the t_R prediction results, the identification of these three isomers (with estimated elution order), reached level 2b of identification confidence [85]. In the case of FUR ozonation TPs, several TPs were detected by RPLC-

(-)ESI-HRMS analysis. Among them, FUR-276, eluted at 3.0 min, was structurally elucidated based on the characterization of its fragments obtained though HRMS analysis (Figure A 3.11(a)). The proposed structure was further supported by the good fitting between the experimental and the predicted t_R (error of 0.21 min) and MCS plot reaching to level 2b (Table 3.1). Moreover, a TP of FUR with m/z 288, eluted at 3.80 min, was detected. Due to the low intensity of this TP, the acquisition of data dependent MS/MS spectra was not feasible, whereas the full MS/MS spectra was noisy and provided no information that could lead to structure elucidation (Figure A 3.11(b)). The proposed structure was included in the suspect FUR TPs (possible to be formed during the ozonation of FUR). Although the predicted t_R (-0.24 min error) was matching to the experimental one and it was in region A of MCS plot (Table 3.1), the level of identification was remained at 3, due to poor MS/MS spectra. Last but not least, t_R prediction was proven helpful in the identification of three isomeric hydroxylated TPs of NIF, eluted at 6.4, 8.1 and 8.9 min. Although an unequivocal formula could be proposed for the three isomers, their fragmentation pattern did not include any characteristic fragments to indicate the exact position where the hydroxylation took place (Figure A 3.12). The t_R prediction highly supported the identification of specific isomers, since the predictions were indicative for the proposed structures, and were identical to the experimental ones (errors from -0.23 to 1.02 min) (Table 3.1).

Table 3.1 Retention time prediction results for the identification of ozonation transformation products of emerging contaminants

Analysis	Parent compound	Transformation product	t_R experimental (min)	t_R predicted (min)	t_R error (min)	Applicability Domain
RPLC-(+)ESI-HRMS	Tramadol	TRA-218	3.31	3.53	0.22	Region A (MCS): accepted
		TRA-296 a	3.54	3.75	0.21	Region A (MCS): accepted
		TRA-296 b	4.50	4.29	-0.21	Region A (MCS): accepted
		TRA-296 c	4.81	4.52	-0.29	Region A (MCS): accepted
		TRA-282	3.72	3.24	-0.48	Region A (MCS): accepted
RPLC-(-)ESI-HRMS	Furosemide	FUR-276	3.04	3.25	0.21	Region A (MCS): accepted
		FUR-288	3.80	3.56	-0.24	Region A (MCS): accepted
RPLC-(+)ESI-HRMS	Niflumic acid	NIF-299 a	6.42	6.19	-0.23	Region A (MCS): accepted
		NIF-299 b	8.10	7.08	-1.02	Region A (MCS): accepted
		NIF-299 c	8.91	8.89	-0.02	Region A (MCS): accepted

3.4. Conclusions

Robust t_R prediction models have been developed based on a large number of emerging contaminants for two chromatographic systems (RPLC) and (HILIC) in two electrospray ionization modes. The non-linear models (SVM) showed high internal and external accuracy and accurate prediction results for suspect screening purposes. A new method, based on MCS, was developed to define the confidence intervals in t_R prediction. This technique incorporates the effect of chemical structures and their similarities compared with the training set to reduce the number of false positives or eliminate the wrong chemical structures assigned for the observed t_R. These models were applied in the suspect and non-target screening of TPs of three emerging pollutants (tramadol, furosemide and niflumic acid). Ten new TPs were tentatively identified using the t_R models and in silico fragmentation and the results proved the value of t_R prediction for newly identified TPs where the reference standards were difficult or impossible to obtain. The t_R models and MCS plot were also used to support the identification of 28 biocides in IWW, EWW and sewage sludge collected from WWTP of Athens which is discussed in details in the chapter 6 of the thesis.

CHAPTER 4

Development and Prediction of Liquid Chromatographic Retention Time Indices

4.1. Introduction

As aforementioned, high resolution mass spectrometry (HRMS) coupled with LC has been widely used to analyze the complex mixtures with wide polarities due to the high separation power and super sensitivity for the measurement of compounds at low concentration [124]. One of the major bottleneck using LC-HRMS is to correctly identify the true positive compounds among the pools of plausible candidates when MS-based information lead to multiple molecular structures [1]. This frequently happens in suspect and non-target screening workflows where the task is to identify the unknowns with certain level of confidence [125]. Therefore, confirming a structure becomes the task of eliminating false plausible candidates based on the available experimental information such as MS fragmentation pattern, retention time (t_R) and ionization behavior between the sample and authentic standards [11, 37, 126]. Prior knowledge about t_R information for the plausible candidates can significantly decrease the number of false positive candidates [37]. Although, t_R information can be useful during screening task, it is often neglected due to the difficulties of accurately predicting and mapping it between different LC conditions.

All studies published to date on the prediction of t_R in LC can be divided into two types; direct experimental t_R mapping [63, 127, 128], which can be used irrespective to specific LC condition, and t_R models established by quantitative structure−retention relationship (QSRR) approach which works locally for specific LC condition [9, 49, 89, 91, 98]. The common strategy to predict t_R of any pollutants is focused primarily on exploring the set of physicochemical descriptors (such as hydrophilicity, polarizability, electronegativity etc.) which yield insight into mechanisms of the elution (interpretability) in contrast to other approach (projection of t_R). The correlation between hydrophilicity and t_R is found to be the baseline and first indication over understanding the elution of a compound in RPLC [9, 49]. These quantitative structure−retention relationship (QSRR) models however rely on t_R information for a large number of compounds to offer the interpretable elution mechanism and correct prediction results [9, 98]. These models can carry a risk of

overfitting or applicable for specific groups of compounds that the QSRR models are trained for [49, 129]. This is generally due to inadequate number of compounds used during QSRR modelling procedure to define the chemical space boundaries or errors in the calculation of molecular descriptors. Development of large database and addressing the uncertainty in t_R prediction results would increase the reliability of these models [130]. Recently, we have developed a QSRR workflow supplemented with the application domain information in which outlying the strategies whether the predicted t_R should be accepted or rejected for a plausible candidate [9, 34]. Apart from the accuracy and interpretability offered by a QSRR model, this approach should be extended and coupled to t_R projection strategies to be applicable under various LC conditions. Nevertheless, development of RTI in LC is ultimate goal to facilitate screening task and control LC quality.

Elution information in gas chromatography are routinely used and have a major role during identification procedure owing to Kovats retention indices (RI) in which enables cross checking RI of a suspect with library data and linking various GC conditions. However, there are little efforts coordinated for LC to make different LC conditions comparable. The reason is lack of experimental information about the elution of whole chemical space (for a large database) in different LC conditions. Moreover, currently, there is not any sufficient information about the calibrants as similar as Kovats RI system for LC. According to Stanstrup and coworkers [63], most of compounds conserve their elution orders for similar chromatographic system (reversed phase and stationary phase (C-18)) and this could be extremely variable in different types of chromatographic columns (e.g., hydrophilic interaction liquid chromatography (HILIC)). By introduction of the strategy of directly mapping various LC conditions, this could help to share t_R information and include it during identification task [63]. This approach so called PredRet however requires a large number of compounds to project the t_R information with high confidence. Hall and co-workers was also developed RTI system similar to Kovats RTI using n-nitroalkanes instead of alkanes [64]. Based on this approach, n-nitroalkanes elute before and after an unknown being measured and then, their retention times define 100 times the number of carbons. This however requires a detailed descriptions for analyzing the reproducibility of RTI system in different LC conditions (different gradient elution

program, mobile phase compositions and stationary phase). The use of logarithmic scale for the retention times and also additive function for the number of carbon remain ambiguous since the number of carbon is not solely, directly and equally correlated to the polarity measure in LC [65]. Moreover, the RTI system was proposed only for a structurally diverse group of 411 small molecules consisting of endogenous compounds, endogenous metabolites and drugs. Therefore, compounds chosen for testing this system were limited to the set of biological elements (C, H, N, O, S, and P) and contained at least one protonable atom to facilitate their detection in mass spectrometry. There is room however to extend this approach and to explore the effect of compounds containing halogens over reliability of this RTI system. Having used the n-nitroalkanes as calibrants will also limit application of these calibrants in negative ionization mode in MS, as they may not ionized properly (i.e. the favorable ionization is +ESI owing to nitrogen in nitroalkanes).

We have therefore sought to build a RTI system in reversed-phase liquid chromatography (RPLC) that is based on calibration of elution pattern using set of substances representative of the appropriate chemical space (more than 1820 compounds from various groups of emerging contaminants). These calibrants were selected chemometrically, using ant colony optimization similarity indices (ACO-SI). This approach selects a set of compounds with maximum overlap with the retention times and chemical similarity indices with the rest of the compounds of the chemical space (here 2123 compounds) after appropriate training. Whole LC part was then linearly calibrated based on these 18 calibrants which have the continuous elution order with pre-normalized RTI values. A calibration set of 18 compounds with RTI set to range between 1 and 1000 was proposed as the most appropriate RTI system. This approach can help the community sharing the t_R information and predict it across different LC conditions and make t_R information more useful in suspect and non-target screening workflows.

4.2. Materials and Methods

4.2.1. Chemicals

The reference standards of the pesticides were donated to the laboratory by Bruker Daltonics (Bremen, Germany), at a concentration of 1 mgL^{-1} in methanol. The rest of the

compounds included in the study were all purchased from Sigma–Aldrich (Germany)**.** Individual stock solutions of these compounds were prepared in methanol at a concentration of 1 g L^{-1} and stored at -20 °C. Then, working solutions were prepared in methanol at a concentration of 1 mg L^{-1}. List of these chemicals can be found in [34]. Acetonitrile (ACN) and Methanol (MeOH) of LC-MS grade, was purchased from Merck (Germany), whereas 2-propanol of LC-MS grade was from Fisher Scientific (Geel, Belgium). Sodium hydroxide monohydrate (NaOH) for trace analysis ≥99.9995%, ammonium acetate, ammonium formate and formic acid, all LC-MS grade, were purchased from Fluka, Sigma–Aldrich (Germany). Distilled water used for LC–MS analysis was provided by a Milli-Q purification apparatus (Millipore Direct-Q UV, Bedford, MA, USA). Regenerated cellulose (RC) syringe filters (15 mm diameter, 0.22 µm pore size) were provided from Phenomenex (Torrance, CA, USA).

4.2.2. Instrumentation and Procedure

An ultrahigh-performance liquid chromatography (UHPLC) system with a LPG-3400 pump (Dionex UltiMate 3000 RSLC, Thermo Fisher Scientific, Germany), interfaced to a Quadrupole Time of Flight (Q-ToF) mass spectrometer (Maxis Impact, Bruker Daltonics, Bremen, Germany) was used for the screening analyses. The same chromatographic and instrumental conditions described previously [9, 37] were used to record t_R of large number of compounds belonging to emerging contaminants in RPLC platform. More details about the instrumental and chromatographic conditions are described in the supplementary material (SM, Chapter 4, appendix A), section SM 4.1.1

4.2.3. Retention Time Indices

The idea was to select the most appropriate set of calibrants out of a dataset of RTs of 2123 compounds taking into account the RT distribution and the chemical similarity of the selected calibrants with the rest of the compounds in the chemical space. Selection of potential compounds as RTI calibrants were achieved chemometrically. There were seven steps behind the selection of RTI calibrants. Firstly, two datasets developed for negative and positive Electrospray Ionization (ESI) modes were processed separately to calculate the molecular descriptors [34]. These molecular descriptors were calculated

using Padel [131] and E-DRAGON [132] based on the final 3D standardized structures. In addition, ChemAxon [133] was used to calculate logD [134] at pH of mobile phases used in LC part which was 3.6 (for positive ESI) and 6.2 (for negative ESI)). Each ESI mode (positive, negative) was processed separately [34]. These 3D structures were obtained out of various conformers (the conformer with the lowest energy was retained as final 3D structure out of several conformers) using Balloon [135]. Totally, 3200 molecular descriptors were calculated for each compound, representing the constitutional, topological, geometric, electrostatic, hydrophobicity, steric effect, quantum related chemical descriptors, chemical fingerprints (Pubchem chemical fingerprint) [136] and various 3D molecular descriptors [137-139]. Constant and near constant molecular descriptors were removed from each dataset. The main datasets, consisting of 303 (-ESI) and 1820 compounds (+ESI), were further processed with collinearity removal. A threshold for the collinearity removal was set to 0.9 in which the molecular descriptors showing a less correlation with the retention time was removed while its collinear pair was retained (the final numbers of 545 and 607 molecular descriptors were obtained for positive and negative ESI, respectively). Secondly, Principal Component Analysis (PCA) was performed on the retained molecular descriptors (excluding retention time) to project the chemical properties of the compounds based on their covariance in three principal components (PCs). Thirdly, a matrix contains retention time information and these three PCs was prepared separately for each ESI as input for identifying the potential RTI calibrants. t_R information along with PCs are the best representative data that can address the similarity between compounds as well as the elution in LC. In forth step, the algorithm starts creating several subset of compounds; then, it calculates the overlap of normal distribution (objective function) between these subset of compounds (RTI calibrants) and the rest of the compounds at predefined desired number of calibrants for each subset. In fifth step, the algorithm seeds to detect the potential subsets of compounds that their selection increases the overlap of normal distribution of RTI calibrants and rest of chemical space. In sixth step, it combines all the compounds between subsets to achieve the highest objective function. Creating the large number of population of compounds, seeding the subsets of compounds and their combinations required a dynamic algorithm to train itself for selecting the best couple of compounds as calibrants. For such as case,

a nature inspired optimization algorithms found to be helpful [140]. Here, we used Ant Colony Optimization (ACO) [141, 142] as a technique for compounds selection. More details about ACO can be found in the supplementary material (SM, Chapter 4, appendix A), section SM 4.1.2. The optimum number of 18 compounds out of 5, 8, 10, 12, 15, 18, 20, 22 and 25 compounds as LC calibrants was achieved for both ESI modes comparing their overlap values and prediction accuracy of RTI system for 30 (+ESI) and 15 (-ESI) external compounds. The RTI system was proposed considering the minimum and maximum elution that was observed for a generic RPLC method and scaled up to 1000. This scale system between 1 and 1000 was set to have large RTI unit between compounds that have different elution and compare the error more realistically. The RTI system proposed here was formulated as below:

$$RTI = \frac{(t_{R\,x} - t_{R\,min})}{(t_{R\,max} - t_{R\,min})} \times 1000 \qquad \rightarrow \qquad RTI = \alpha(t_{R\,c}) + C \qquad 4.1$$

where $t_{R\,x}$ and $t_{R\,c}$ are the t_R observed for the calibrants and a compound, $t_{R\,min}$ and $t_{R\,max}$ are the minimum and maximum t_R observed for the calibrants, respectively. α and C are the slope and the intercept at 99% confidence interval. The linear correlation is the RTI calibration equation and it is used to transform any experimental t_R in one system into RTI. The RTI values can be used afterwards to harmonize the elution of compounds in various LCs. This could also facilitate the evaluation of LC quality according to the degree of deviation from linearity (or use of lack of fit to examine the residuals after calibration curve development) for RTI calibrants.

4.2.4. Intralaboratory and interlaboratory validation

An intra-laboratory evaluation was followed considering various C18 columns (Acclaim RSLC C18, Atlantis T3 C18, XBridge C18 Waters, and Acquity UPLC BEH C18 column), mobile phase compositions (MeOH, H_2O and ACN with and without buffer system) and gradient elution program (in addition to our pervious method [9], two other gradient elution program were adopted from literature [124, 143]) to evaluate the reproducibility of the proposed RTI system. More details about the instrumental and chromatographic conditions are described in the supporting material (Chapter 4; Appendix B) Table B 4.1. In addition, the interlaboratory comparisons were organized with the aim to evaluate

externally the accuracy of RTI system from one lab to another with completely different instrumentation. The proposed RTI system was evaluated by three laboratories (Swiss Federal Institute for Aquatic Science and Technology (Eawag), Helmholtz-Centre for Environmental Research (UFZ) and University Jaume I (UJI)). Eawag used two different LC conditions [124], UFZ conducted the evaluation under the LC condition reported in and UJI applied the LC condition reported in to evaluate the RTI system. More details about the instrumental and chromatographic conditions for all the participants are described in the supplementary material (SM, Chapter 4, appendix A), section SM 4.1.3. The t_R information for the calibrants as well as set of compounds as a blind set (for evaluating the external prediction capability of proposed RTI system) were reported by each laboratory.

4.2.5. External application and validation

Fifteen laboratories evaluated the proposed RTI system externally apart from the core team. The details about their LC conditions can be found in Table B 4.2 (supplementary material (SM, Chapter 4, appendix B)).

4.2.6. QSRR workflows

The modelling workflow introduced in our previous study [34] was followed to predict RTI. Some details about each step of this workflow can be found in the supplementary material (SM, Chapter 4, appendix A), section SM 4.1.4.

4.2.7. Prediction Intervals

There were two layers of uncertainties comprised from to the QSRR models and RTI versus t_R calibration curve. In case of comparing the experimental RTI and its accuracy pair wisely among different LC conditions, compounds that are falling inside the confidence interval of 99% (CI) of the calibration curve were considered to be identical throughout various LC conditions. In other words, to accept that the difference between two RTIs values are not significantly different, the means of RTIs were compared statistically. This was rigorously done using student t-test, ANOVA, least significant difference (LSD) [144] and multiple comparison procedures [145, 146] which is much accurate in terms of derivation of correct significance level for multiple pairwise

comparison. In case of single RTI value between two labs, lower and upper CIs were used from calibration curve to perform these tests. Therefore, the difference between two experimental RTIs measured for a compound by two labs is significant when

$$\left|\overline{RTI}_{lab1} - \overline{RTI}_{lab2}\right| \geq t_{N-k,1-\frac{\alpha}{2}} \times \sqrt{S_I^2\left(\frac{1}{n_{lab1}} + \frac{1}{n_{lab2}}\right)} \qquad 4.2$$

where $t_{N-k,1-\frac{\alpha}{2}}$ is critical student t value at N−k degrees of freedom for a significance level set to α. N is the total number of observations, k is the number of labs and S_I^2 is the estimation of the variance within the labs.

In our previous work concerning the uncertainty in prediction results via QSRR method, we coupled leverage, standardized residuals (relative t_R error window) and normalized mean distance in a single 3D bubble plot so called OTrAMS [9]. It was found to be very robust method to decrease the chance of false positives for a single retention time value. This is a very useful tool to handle false positives in cases where several plausible candidates were present for a single t_R/RTI value. For a compound without an experimental t_R/RTI measured, the only way to address the applicability domain of the QSRR based models is to use the chemical space boundaries. Therefore, if the error is the function of chemical space failure, warning leverage values versus normalized mean distance can be used to define the applicability domain [73]. These two methods (OTrAMS and chemical space boundaries) were used here during RTI modelling and applied over prediction results of RTI in interlaboratory study. The details about OTrAMS and chemical space boundaries can be found in the supplementary material (SM, Chapter 4, appendix A), section SM 4.1.5.

4.2.8. Stability test of RTI calibrants

Four mixtures (two mixtures of 18 compounds for each RTI calibrants set) were prepared at the concentration of 2 mgL^{-1} in pure methanol (at the final volume of 250 µL). Stability test was performed by analyzing each mixture at 0, 6, 18, 24, 36 and 48 hours of storage time at two different temperature (-18 and +2 °C). The mixtures were returned to refrigerator and freezer after each analysis time point and stored for the next injection.

4.3. Results and Discussion

Correct selection of calibrants covering various elution characteristics (being sensitive to the pH and compositions of mobile phase) can lead to a more accurate harmonisation of t_R values from one LC configuration to another. In addition, a RTI system and its prediction are of need to avoid to fit/map various t_Rs every time and use only RTI prediction versus the experimental one (calibrated by 18 RTI calibrants). Here, the selected calibrants represent range of polarity, molecular properties and they have consistent elution order in various LC conditions. Moreover, QSRR based models can help decrease chance of plausible candidates in case of suspect and non-target screening.

4.3.1. Selection of RTI calibrants

In the RTI system proposed here, a nature inspired algorithm was used to dynamically calculate the overlap between normal distribution (of experimental t_R and chemical space) of any selected calibrants with rest of compounds. This enabled to test various combinations of compounds and select the best set of calibrants that has highest normal distribution overlap on their chemical space (calculated from molecular descriptors and chemical figerprints) and experimental t_R values. Figure A 4.1 (supplementary material (SM, Chapter 4, appendix A)) shows the final overlap that is achieved based on 18 set of compounds as RTI calibrants. The overlap of normal distribution between calibrants and rest of dataset are shown separately for experimental t_R and chemical space in Figure S1. The optimal number of calibrants was achieved using the lowest residuals derived between the predicted and experimental RTI and t_R for 30 and 15 compounds as validation set in positive and negative ESI, respectively. These compounds were selected with the same algorithm as the calibrants, using ACO-SI. The errors observed for these compounds, after using various number of calibrants, are shown in Figure A 4.2 (supplementary material (SM, Chapter 4, appendix A)). The lowest distribution of error (between ±1) is derived using 18 calibrants. However, the less error is seen for +ESI owing to the large database compiled.

4.3.2. Modeling Retention Time Indices

Two models were built, after calibrating the t_R values of large number of emerging contaminants to RTI (±ESI) values, using ACO to select the most respective molecular descriptors and Support Vector Machine (ACO-SVM) to non-linearly correlate these molecular descriptors with RTI. The performance of the models for the proposed RTI system can be found in Table 4.1. Both models show high correlation coefficient and leave one out cross-validation as well as low Root Mean Square Error (RMSE). The linear models (ACO-Multiple Linear Regression) however showed less prediction accuracy than ACO-SVM.

Table 4.1 Prediction performance of models developed for proposed RTI system.

	Training				Test		
	R^2	RMSE	F	Q^2_{LOO}	R^2	RMSE	F
RTI for (–ESI):							
ACO-MLR	0.835	89.221	301.849	0.827	0.801	84.606	57.456
ACO-SVM	0.984	27.709	3596.46	0.813	0.833	75.703	71.898
RTI for (+ESI):							
ACO-MLR	0.847	89.187	2011.48	0.846	0.835	92.630	446.150
ACO-SVM	0.945	53.605	6232.62	0.864	0.868	82.642	610.690

Regarding the molecular descriptors selected and used behind RTI model (–ESI), ACO identified logD (pH=6.2) as the most important descriptor (with importance of 67.7%). LogD (distribution coefficient at certain pH) is the most expected molecular descriptor to describe the general mechanism of hydrophobicity in the chromatography. Largest absolute eigenvalue of Burden modified matrix - n4 /weighted by relative mass (SpMax4_Bhm) (with importance of 19.6%), electronic features of the molecule relative to molecular size (ETA_BetaP) (with importance of 4.5%) and the chemical fingerprint of atom paired (C-S) (with importance of 8.2%) follow as potential molecular descriptors. Other three descriptors reflect the molecular size, electronic profile, which can reveal the

ionic interaction of compounds with stationary phase, and chemical fingerprint of C-S in molecular structure. The linear equation to derive experimental RTIs from t_R values of calibrants as well as QSRR based predicted RTIs in –ESI are formulated below:

$$Exp.\ RTI_{(-ESI)} = 76.8986\ (\pm 0.1200) t_{R_{(-ESI)}} - 128.2275\ (\pm 0.01473) \qquad 4.3$$

$$\begin{aligned} Pred.RTI_{(-ESI)} &= +90.2684\ (\pm 78.7783) + 68.6920\ (\pm 2.9550)\ logD_{pH=6.2} \\ &+ 173.2919\ (\pm 20.5676)\ SpMax4_{Bhm} - 264.7767\ (\pm 46.8641)\ ETA_{BetaP} \\ &- 97.9374\ (\pm 14.5943)\ PubchemFP293[C-S] \end{aligned} \qquad 4.4$$

LogD(pH=3.6), Hybridization ratio (HybRatio), 3D topological distance based autocorrelation lag 5 / weighted by covalent radius (TDB5r) and charged partial surface area (THSA) [147] were selected and used to model RTI in + ESI. LogD was already discussed in –ESI and showed high importance (68.3%) among the selected descriptors in RTI proposed for +ESI as well. Hybridization ratio is calculated by dividing the numbers of carbon with sp^3 hybridization to total numbers of carbon (sp^3 and sp^2 hybridization) in a molecule. It shows the geometry and bonding properties for a molecule. High HybRatio reflects the less conjugated carbon (sp^2) and high sp^3 carbon which causes an increase in hydrophobicity of a molecule and to some extent increase of RTI. HybRatio however showed lowest importance (1.8%) among three other selected molecular descriptors. TDB5r (with importance of 7.9%) shows the size of an atom that forms part of one covalent bond in topological distance of 5 in a molecule. Covalent bond occurs between carbons and carbon-hydrogen and thus the mechanism of act is based on the effect of hydrophobicity on RTI. THSA (with importance of 22.0%) was also selected as the second most important molecular descriptor after logD. It reveals the electronic profile of a compound, indicating the ionic interaction of the compounds with stationary phase. The linear equation to derive experimental RTIs from t_R values of calibrants as well as QSRR based predicted RTIs in +ESI are formulated below:

$$Exp.\ RTI_{(+ESI)} = 76.3787\ (\pm 0.01194) t_{R_{(+ESI)}} - 99.9112\ (\pm 0.1006) \qquad 4.5$$

$$Pred.RTI_{(+ESI)} = -57.0034\ (\pm 19.6192) + 70.9028\ (\pm 1.1186)\ logD_{pH=3.6} + 159.8775\ (\pm 11.5547)\ HybRatio + 62.2194\ (\pm 10.2845)\ TDB5r + 0.5516\ (\pm 0.0226)\ THSA \qquad 4.6$$

The dataset used to build the models as well as experimental and predicted RTIs for ±ESI modes can be found in Table B 4.3 and Table B 4.4. Figure A 4.3 A&B (supplementary material (SM, Chapter 4, appendix A), section SM 4.2.2) show the results of OTrAMS for the predicted RTI using equation 4 and 6. As it can be seen, all the compounds are within the acceptance threshold of ±3SR (box3). In addition to applicability domain approach applied for compounds with known t_R and RTI, defining the chemical space boundaries with unknown t_R and RTI was done by projecting the leverage values of chemicals against their distance from training set used to build RTI models. This extremely helps to identify the potential source of inaccuracies in prediction of RTI when the error is the subject of chemical space failure. Figure A 4.3 C&D (supplementary material (SM, Chapter 4, appendix A), section SM 4.2.2) show the chemical space boundaries for the compounds used in each ESI platform. It is found that all chemicals are inside the chemical space boundaries and there are not any substantial outliers as a result of chemical space failure.

4.3.3. Intralaboratory Accuracy of RTI

The proposed RTI system was evaluated internally in four different LC conditions to examine whether the RTI system works properly by changing the mobile phase composition, gradient elution program and column type. The accuracy of proposed RTI system was studied in terms of RMSE, square correlation coefficient (R^2), distribution of residuals derived from experimental and predicted RTI/t_R and the true positive harmonisation rate of the elution of compounds in different LC conditions. The true positive harmonisation rate (TPHR) is derived by the following equation:

$$TPHR = \frac{No.True\ Harmonised\ RTIs\ within\ 99\%\ CIs}{No.detected\ compounds} * 100 \qquad 4.7$$

TPHR indicates the percentage of RTI values for the compounds that have the overlap of experimental RTIs/its 99% CIs in various LC conditions in contrast to the RTI values derived from the main LC condition. To accept whether the difference between RTI values from different LC conditions is significant or not, the multiple comparison procedure was applied using LSD as main decision criteria. The retention time values observed for RTI calibrants in each LC condition described in SM 4.1.3.1 can be found in Table 4.2. Dinoterb and Valproic acid are found to be outside of prediction limit while building the RTI calibration equation in LC condition 4. All in all, the internal and external accuracy of proposed RTI system in –ESI found to be reliable and work well in mobile phase composition of MeOH:H_2O with the TPHR of above 100%. The outcomes of evaluation of internal and external accuracy for the proposed RTI system in –ESI can be found in Table 4.3. TPHR is also calculated by multiple comparison procedure and the results are listed in Table 4.3 and also visualized as a cloud plot. In this cloud plot, the bubble size is proportional to the CIs of experimental RTI (at 99% CIs from calibration curve) in which the overlap between these bubbles correspond to the successful harmonisation of elution of compounds from one LC condition to another. The cloud plot of experimental RTIs for –ESI platform is shown in Figure 4.1A. The result of multiple comparison procedure for evaluating the internal accuracy of RTI in –ESI can be found in Table B 4.5.

Table 4.2 Intralaboratory results for the RTI proposed for (–ESI) mobile phase

Compound	Mol. Formula	$[M-H]^-$	Main LC (t_R, min)	LC1 (t_R min)	LC2 (t_R min)	LC3 (t_R min)	LC4 (t_R min)	RTI
Amitrole	C2H4N4	83.0363	1.67	1.56	0.86	0.86	0.81	1
Benzoic acid	C7H6O2	121.0295	2.88	3.12	1.72	1.37	2.56	93.08
Acephate	C4H10NO3PS	182.0046	3.09	3.17	1.87	1.31	1.11	109.23
Salicylic acid	C7H6O3	137.0244	3.58	3.77	2.37	1.82	1.31	146.92
Simazine 2-Hydroxy	C7H13N5O	182.1047	4.96	5.18	3.67	3.27	1.36	253.08
Tepraloxydim Peak1	C17H24ClNO4	340.1321	5.26	5.53	3.97	3.67	8.64	276.15
Bromoxynil	C7H3Br2NO	273.8509	5.35	5.68	3.92	3.67	8.14	283.08
MCPA	C9H9ClO3	199.0167	6.49	6.58	4.82	4.82	5.77	370.77
Valproic acid	C8H16O2	143.1078	7.04	7.33	5.28	5.23	14.23	413.08
Phenytoin	C15H12N2O2	251.0826	7.16	7.38	5.28	5.23	11.4	422.31
Flamprop	C16H13ClFNO3	320.0495	7.49	7.63	6.03	5.88	13.6	447.69
Benodanil	C13H10INO	321.9734	7.99	8.34	6.23	6.18	12.5	486.15
Dinoterb	C10H12N2O5	239.0673	8.13	8.44	6.48	6.43	19.82	496.92
Inabenfide	C19H15ClN2O2	337.0749	9.23	9.49	7.53	7.79	16.33	581.54
Coumaphos	C14H16ClO5PS	361.0072	10.98	11.39	9.29	10.49	22.04	716.15
Triclosan	C12H7Cl3O2	286.9439	12.02	12.25	10.39	12.4	23.28	796.15
AvermectinB1a	C48H72O14	871.4849	13.64	14.3	12.5	15.96	24.98	920.77
Salinomycin	C42H70O11	749.4845	14.67	15.66	13.1	17.06	26.09	1000.00

LC1: Atlantis T3 C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate; Flow rate: multi-flow-rate gradient; Run time: 15 min
LC2: Acclaim RSLC C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate; Flow rate: 0.200 mL/min; Run time: 25 min
LC3: XBridge C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate; Flow rate: 0.200 mL/min; Run time: 25 min
LC4: XBridge C18; Mobile Phase: H2O (with 5 mM ammonium acetate)/ACN; Flow rate: 0.300 mL/min; Run time: 25 min

Table 4.3 RTI and tR calibration curve for (–ESI) under different LC conditions

LC conditions	RTI versus t_R equation	Standard Error	Internal accuracy	External accuracy (n=15)
LC1	RTI = 73.12 (t_R) –121.6	Intercept : ±5.800 Slope: ±0.6835	R^2=0.9986	R^2=0.9703, RMSE= 36.26, TPHR= 100%
LC2	RTI = 79.50 (t_R) –30.98	Intercept: ±9.424 Slope: ±1.386	R^2=0.9952	R^2=0.953, RMSE=50.14, TPHR= 100 %
LC3	RTI = 58.15 (t_R) + 67.66	Intercept: ±21.16 Slope: ±2.698	R^2=0.9667	R^2=0.932, RMSE=54.74, TPHR= 100 %
LC4	RTI = 32.42(t_R) + 67.00	Intercept: ±29.33 Slope: ±2.056	R^2=0.9470 Dinoterb and Valproic acid are outside of the prediction limit	R^2=0.872, RMSE=110.608, TPHR= 66.66 %

LC1: Atlantis T3 C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate; Flow rate: multi-flow-rate gradient; Run time: 15 min
LC2: Acclaim RSLC C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate; Flow rate: 0.200 mL/min; Run time: 25 min
LC3: XBridge C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate; Flow rate: 0.200 mL/min; Run time: 25 min
LC4: XBridge C18; Mobile Phase: H2O (with 5 mM ammonium acetate)/ACN; Flow rate: 0.300 mL/min; Run time: 25 min

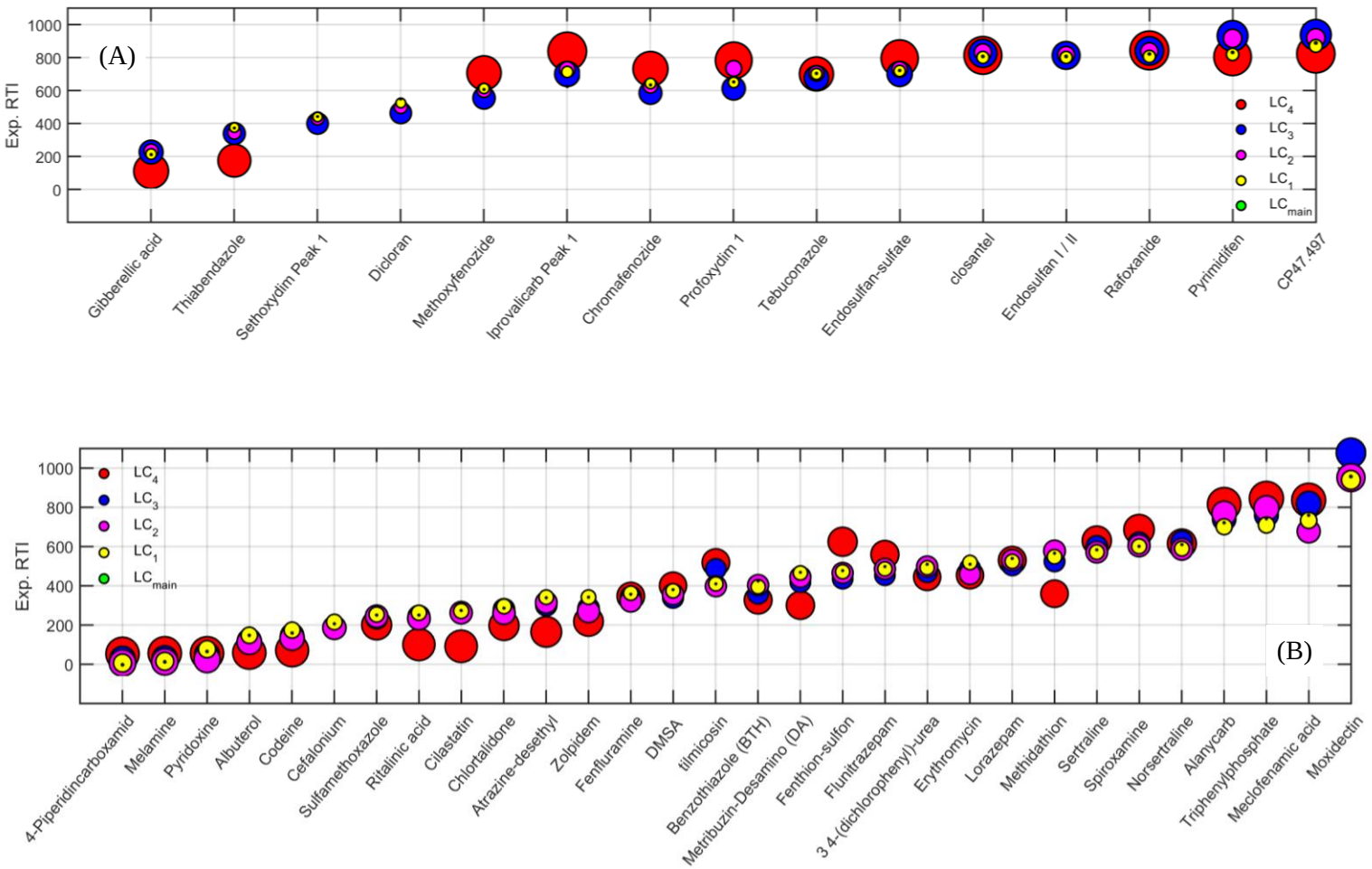


Figure 4.1 The cloud plot of experimental RTIs measured in various LC conditions with their acceptance CIs in (A) –ESI and (B) +ESI

The t_R values observed in +ESI for RTI calibrants in each LC condition described in SM 4.1.3.1 can be found in Table 4.4. The internal and external performance of the proposed RTI system in +ESI can be found in Table 4.5. Concerning the calibration quality (linearity and squared correlation coefficient), it is found that RTI proposed for +ESI generally has better outcomes in contrast to RTI for –ESI. Moreover, LC condition 4 remained less accurate than other LC conditions. Similarly, the internal and external accuracy of proposed RTI system in +ESI found to be reliable and work well in mobile phase composition of MeOH:H_2O, performing at any column type and gradient elution program, with TPHR of 100%. The TPHR are calculated from multiple comparison procedure algorithm and the cloud plot is presented in Figure 4.1B. The result of multiple comparison procedure for evaluating the internal accuracy of RTI in +ESI can be found in Table B 4.6. Student t-test was also used to compare the most divers internal LC conditions (main LC (the nominal one) and LC 4) based on experimental t_R and RTI values for 12 randomly selected substances from each LC condition. Table B 4.7 shows the results of student t-test. As it can be seen, most t_R values vary significantly from the nominal LC conditions to LC 4 condition (for instance CP47.497 with t_R value of 13.24 (LC main) and 24.29 (LC 4)) while the RTI values are steady (RTI value in LC main=869.18 and LC 4= 868.04). Student t-test explains the same effect which two LC conditions are identical at 99% CIs based on RTI values while differ significantly based on t_R values.

Table 4.4 Intralaboratory results for the RTI proposed for (+ESI) mobile phases

Calibrants	Mol. Formula	$[M+H]^+$	Main LC (t_R min)	LC1 (t_R min)	LC2 (t_R min)	LC3 (t_R min)	LC4 (t_R min)	RTI
Guanylurea	C2H6N4O	103.0614	1.31	1.49	1.25	1.30	1.23	1
Amitrol	C2H4N4	85.05087	1.39	1.55	1.25	1.30	1.23	6.11
Histamine	C5H9N3	112.0869	1.58	1.38	1.32	1.71	1.42	20.63
Chlormequate	C5H13ClN	123.0809	1.67	1.80	1.66	1.30	1.27	27.50
Methamidophos	C2H8NO2PS	142.0086	2.76	2.98	2.60	2.29	1.43	110.77
Vancomycin	C66H75Cl2N9O24	1448.437	3.26	3.46	5.34	4.03	2.06	148.97
Cefoperazone	C25H27N9O8S2	646.1497	4.36	4.57	5.34	4.75	4.15	233.00
Trichlorfon	C4H8Cl3O4P	256.9299	5.23	5.57	5.83	5.32	3.89	299.47
Butocarboxim	C7H14N2O2S	191.0849	6.07	6.43	6.88	6.22	7.35	363.64
Dichlorvos	C4H7Cl2O4P	220.9532	7.00	7.28	8.05	7.26	11.26	434.68
Tylosin	C46H77NO17	916.5264	7.88	8.64	8.66	8.56	13.78	501.91
TCMTB	C9H6N2S3	238.9766	9.25	9.62	11.7	10.39	20.97	606.57
Rifaximin	C43H51N3O11	786.3596	10.06	10.41	12.65	11.54	20.67	668.45
Spinosad A	C41H65NO10	732.4681	11.34	11.72	14.16	14.35	23.03	766.23
Emamectin B1a	C49H75NO13	886.5311	12.40	12.77	15.82	16.61	24.59	847.21
Abamectin	C48H72O14	873.4995	13.64	14.10	15.22	16.06	23.99	941.94
Nigericin	C40H68O11	725.4834	13.94	15.70	18.84	17.55	24.54	964.86
Ivermectin B1a	C48H74O14	875.5151	14.40	14.78	16.92	16.01	25.01	1000

LC1: Atlantis T3 C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate and 0.01% formic acid; Flow rate: multi-flow-rate gradient; Run time: 15 min

LC2: Acclaim RSLC C18; Mobile Phase: H2O/MeOH with 0.1% formic acid; Flow rate: 0.200 mL/min; Run time: 25 min

LC3: Acquity UPLC BEH C18; Mobile Phase: H2O/MeOH with 0.1% formic acid; Flow rate: 0.200 mL/min; Run time: 25 min

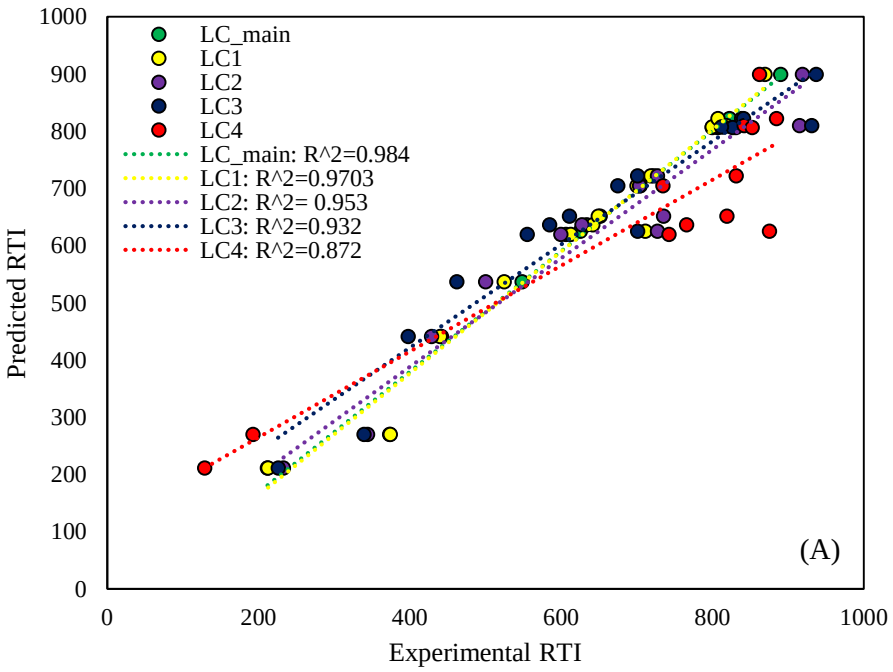
LC4: Acquity UPLC BEH C18; Mobile Phase: H2O/ACN with 0.1% formic acid; Flow rate: 0.300 mL/min; Run time: 25 min

Table 4.5. RTI and tR linear equations for (+ESI) in different LC conditions

LC conditions	RTI versus t_R equation	Standard Error	Internal accuracy	External accuracy (n=30)
LC1	RTI = 72.40(t_R) – 98.67	Intercept : ±10.21 Slope: ±1.151	R^2=0.9960	R^2=0.9589, RMSE=46.61, TPHR = 100%
LC2	RTI = 59.60(t_R) – 66.91	Intercept: ±20.82 Slope: ±2.020	R^2=0.9820	R^2=0.9286, RMSE=63.78, TPHR = 100%
LC3	RTI = 60.02(t_R) – 47.39	Intercept: ±19.04 Slope: ±1.909	R^2=0.9841	R^2=0.9298, RMSE=64.78, TPHR = 100%
LC4	RTI = 35.62(t_R) + 12.29	Intercept: ±29.74 Slope: ±1.942	R^2=0.9523	R^2=0.7789, RMSE=119.59 TPHR = 84.61%

LC1: Atlantis T3 C18; Mobile Phase: H2O/MeOH with 5 mM ammonium acetate and 0.01% formic acid; Flow rate: multi-flow-rate gradient; Run time: 15 min
LC2: Acclaim RSLC C18; Mobile Phase: H2O/MeOH with 0.1% formic acid; Flow rate: 0.200 mL/min; Run time: 25 min
LC3: Acquity UPLC BEH C18; Mobile Phase: H2O/MeOH with 0.1% formic acid; Flow rate: 0.200 mL/min; Run time: 25 min
LC4: Acquity UPLC BEH C18; Mobile Phase: H2O/ACN with 0.1% formic acid; Flow rate: 0.300 mL/min; Run time: 25 min

All in all, the proposed RTI system can harmonize the t_R information throughout different LC conditions. In addition, two QSRR based models were built to facilitate the identification of compound during suspect and non-target screening. The predicted RTI can be compared to the experimental one that was generated from calibration curve through the 18 calibrants introduced above. In this case, uncertainty and chemical space boundaries (using OTrAMS and leverage values versus normalized mean distance) can be accurately defined enabling to use RTI models for various LC conditions with high confidence. The plot of experimental versus predicted RTI based on QSRR models for these four LC conditions are presented in Figure 4.2. High correlation is observed between the predicted and experimental RTI for the validation set used in these four LC conditions.



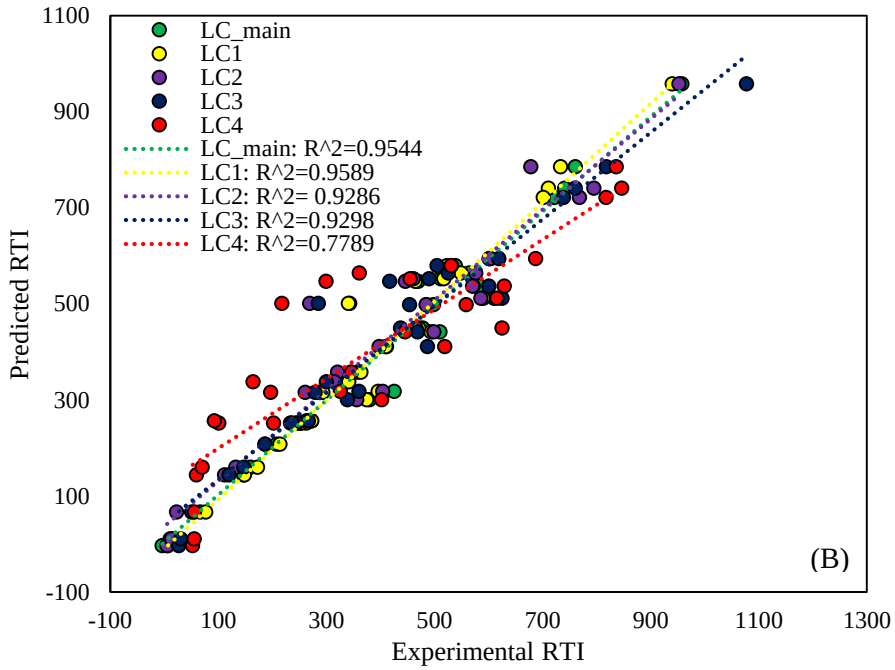


Figure 4.2 The correlation between the experimental and predicted RTIs measured in various LC conditions for validation set used in (A) –ESI and (B) +ESI

4.3.4. Interlaboratories Accuracy of RTI

Following the successful application of proposed RTI system in various LC conditions in both ±ESI, the inter-laboratory trials were performed independently by Helmholtz Centre for Environmental Research (UFZ), Swiss Federal Institute of Aquatic Science and Technology (Eawag) and University Jaume I (UJI). The results suggested that the proposed RTI system can be applied with high confidence to share the LC information throughout various LC conditions.

4.3.5. Evaluation of RTI proposed in –ESI

UFZ has successfully performed the proposed RTI system in –ESI and could observe high internal (linear equation derived from RTI versus t_R) and external accuracy (a large set of 208 compounds as evaluation set). Moreover, the 15 compounds existed in our dataset and the one provided by UFZ were compared by multiple comparison procedure

(Table B 4.8) and projected by their experimental RTI values on cloud plot (Figure 4.3A). All the compounds were overlapped and found to fall inside the RTI CIs. The width of RTI CIs mainly depends on the accuracy of calibration curve built between t_R and RTI. Compounds measured in –ESI are mainly acidic and partially polar compounds, thus the elution of these compounds could be highly sensitive to pH value of the mobile phase in LC condition. This could be source of large RTI CIs if the pH value was not adjusted. The prediction results were also found highly correlated with experimental values. All in all, 109 (52.40%) compounds were predicted within box 1 (within ±90 RTI unit), 67 compounds (32.21%) within box 2 (within $+90 \leq |RTI| < +175$ in RTI unit) and 32 compounds (15.38%) within box 3 (within $+175 \leq |RTI| < +275$ in RTI unit) (Figure 4.4A). Therefore, approximately 84.62 % of whole compounds successfully could be predicted with high confidence through the RTI system proposed. Moreover, no outliers, as the function of chemical space failure, are found (Figure 4.4B). The correlation between the experimental and predicted RTI as well as distribution of the error can be found in Figure 4.4C and D, respectively. These prediction intervals derived from OTrAMS method would ensure that the QSRR models are applicable and can be used to elucidate or prioritize the target structure among plausible candidates. TPHR as well as internal and external accuracy for the results reported by UFZ can be found in Table 4.6.

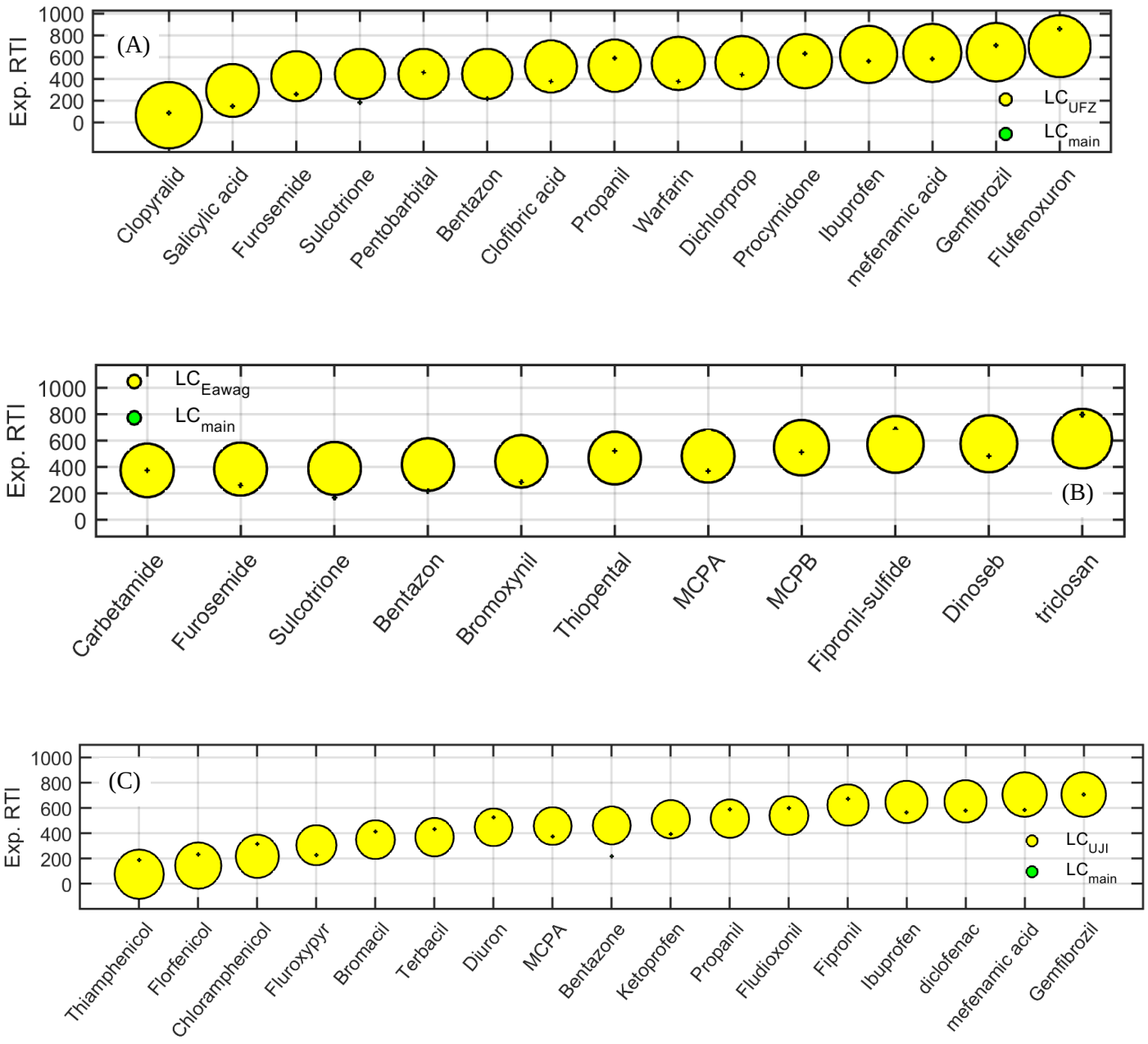


Figure 4.3 Cloud map of RTI values in different LC conditions reported by (A) UFZ; (B) Eawag; (C) UJI in (–ESI)

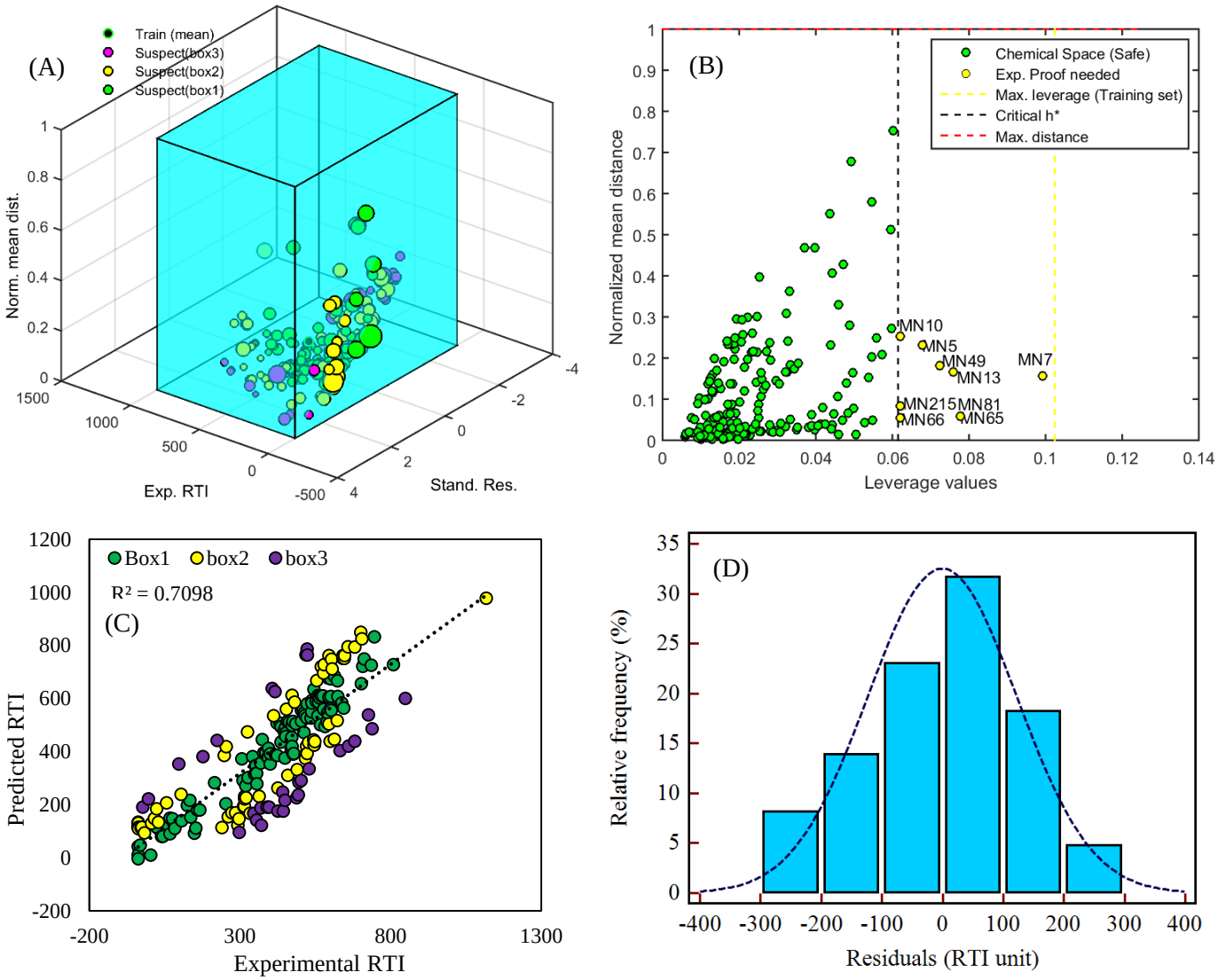


Figure 4.4 QSRR based prediction results with its boundaries for the compounds reported by UFZ; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results

Table 4.6 RTI and tR calibration curve for (–ESI) reported by different laboratories

Labs	RTI versus t_R equation	Standard Error	Internal accuracy	External accuracy
UFZ	RTI = 54.28 (t_R) – 60.61	Intercept : ±85.95 Slope: ±8.691	R^2=0.7360	R^2=0.7098, RMSE=122.4049, TPHR= 100 % (n=15)
Eawag	RTI = 36.95 (t_R) – 65.01	Intercept: ±86.26 Slope: ±5.761	R^2=0.7600	R^2=0.6805, RMSE=146.3575, TPHR=100 % (n=11)
UJI	RTI = 62.98 (t_R) –124.4	Intercept: ±63.13 Slope: ±6.310	R^2=0.8770	R^2=0.6873, RMSE=123.6278, TPHR=94.12 % (n=17)

Eawag was also evaluated the RTI system and reported 27 compounds in –ESI. Most of these compounds reported were surfactants which found to be difficult to be predicted, as their elution do not follow proportionally by their logD values. Moreover, 11 compounds existed in our dataset and at the same time provided by Eawag were projected by their experimental RTI values on cloud plot (Figure 4.3B). These RTI values were also statistically compared by multiple comparison procedure and the results can be found in Table B 4.9. Although, all the compounds are overlapped and found to fall inside the RTI CIs, the width of RTI CIs are as large as the ones derived for UFZ. It was expected as the mobile phase composition and its pH are the same in these two labs. This reflects the fact that adjustment of pH has major influence on the internal accuracy of LC condition when operating at negative ESI mode. The prediction results were also found highly correlated with experimental values. Totally, 10 (37%) compounds were predicted within box 1 (within ±80 RTI unit), 10 compounds (37%) within box 2 (within $+80 \leq |RTI| < +170$ in RTI unit) and 7 compounds (26%) within box 3 (within $+170 \leq |RTI| < +250$ in RTI unit) (Figure 4.5A). Therefore, 74 % of whole compounds successfully could be predicted with high confidence through the RTI system proposed. Moreover, no outliers, as the function of chemical space failure, are found (Figure 4.5B). The correlation between the experimental and predicted RTI as well as distribution of the error can be found in Figure 4.5C and D, respectively. TPHR along with internal and external accuracy for the results reported by Eawag are tabulated in Table 4.6.

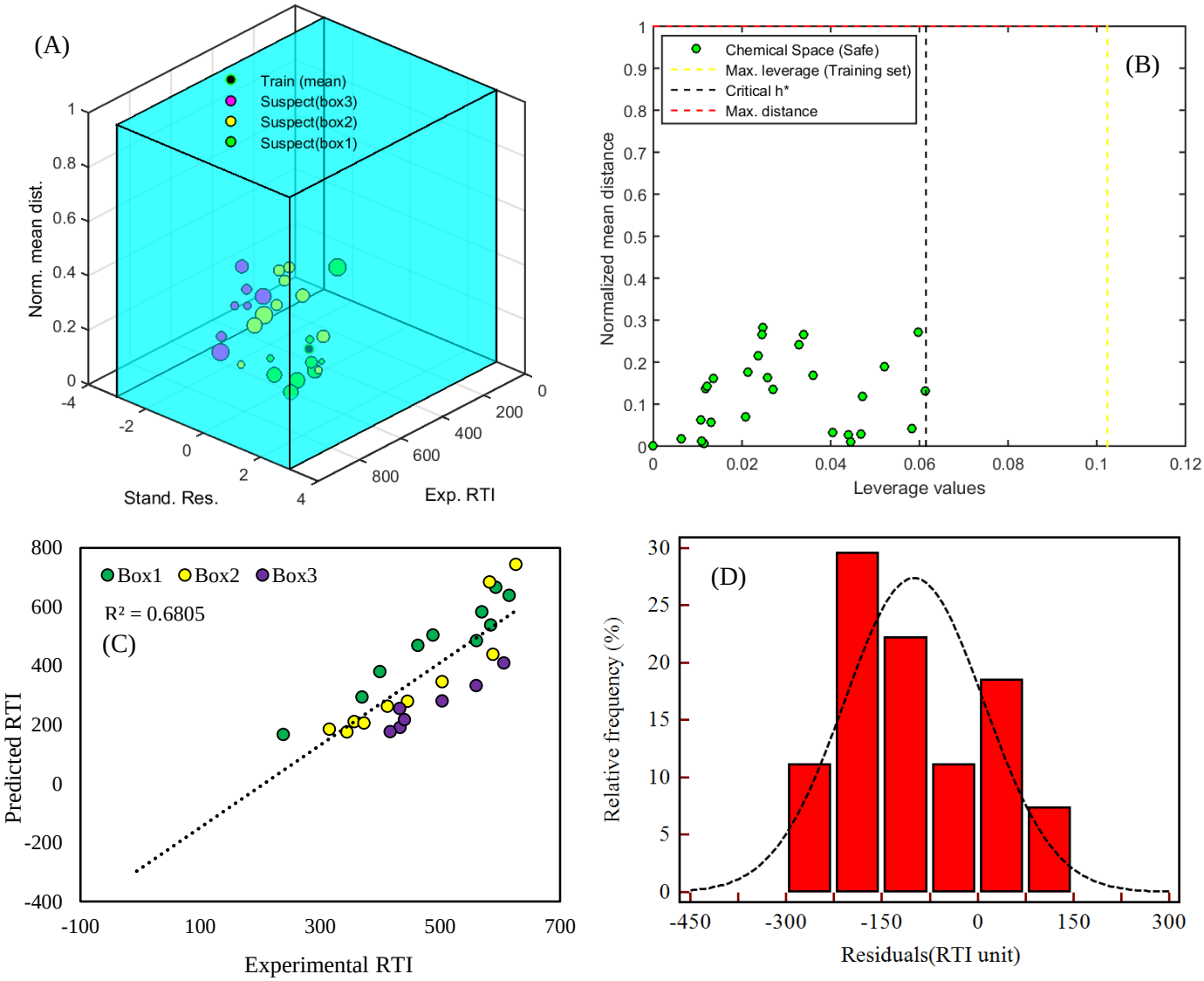


Figure 4.5 QSRR based prediction results with its boundaries for the compounds reported by Eawag; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI;(D) distribution of error associated with prediction results.

UJI was evaluated the RTI system with the LC condition described in SM 4.1.3.2 and reported 27 compounds in –ESI. In addition, 17 common compounds existed in our dataset and provided by UJI were projected by their experimental RTI values on the cloud plot (Figure 4.3C). The result of multiple comparison procedure can be found in Table B 4.10. Although, all the compounds are overlapped and found to fall inside the RTI CIs,

the width of RTI CIs are relatively smaller than those obtained for UFZ and Eawag. UJI did not use a buffer system and the pH is alike the one used in UFZ and Eawag which could be the source of major effect on ionic interaction between compounds and the stationary phase. This could be of major concern, when the analyses are operating in –ESI mode, as most of compounds detectable in –ESI contain acidic functional group. The prediction results were also found highly correlated with experimental values. Totally, 12 (44.44%) compounds were predicted within box 1 (within ±90 RTI unit), 12 compounds (44.44%) within box 2 (within $+90 \leq |RTI| < +170$ in RTI unit) and 3 compounds (11.12%) within box 3 (within $+170 \leq |RTI| < +250$ in RTI unit) (Figure 4.6 A). Therefore, 88.88 % of whole compounds successfully could be predicted with high confidence through the RTI system proposed. Moreover, no outliers are found to be due to the chemical space failure (Figure 4.6 B). The correlation between the experimental and predicted RTI as well as distribution of the error can be found in Figure 4.6C and D, respectively. TPHR together with internal and external accuracy for the results reported by UJI are listed in Table 4.6.

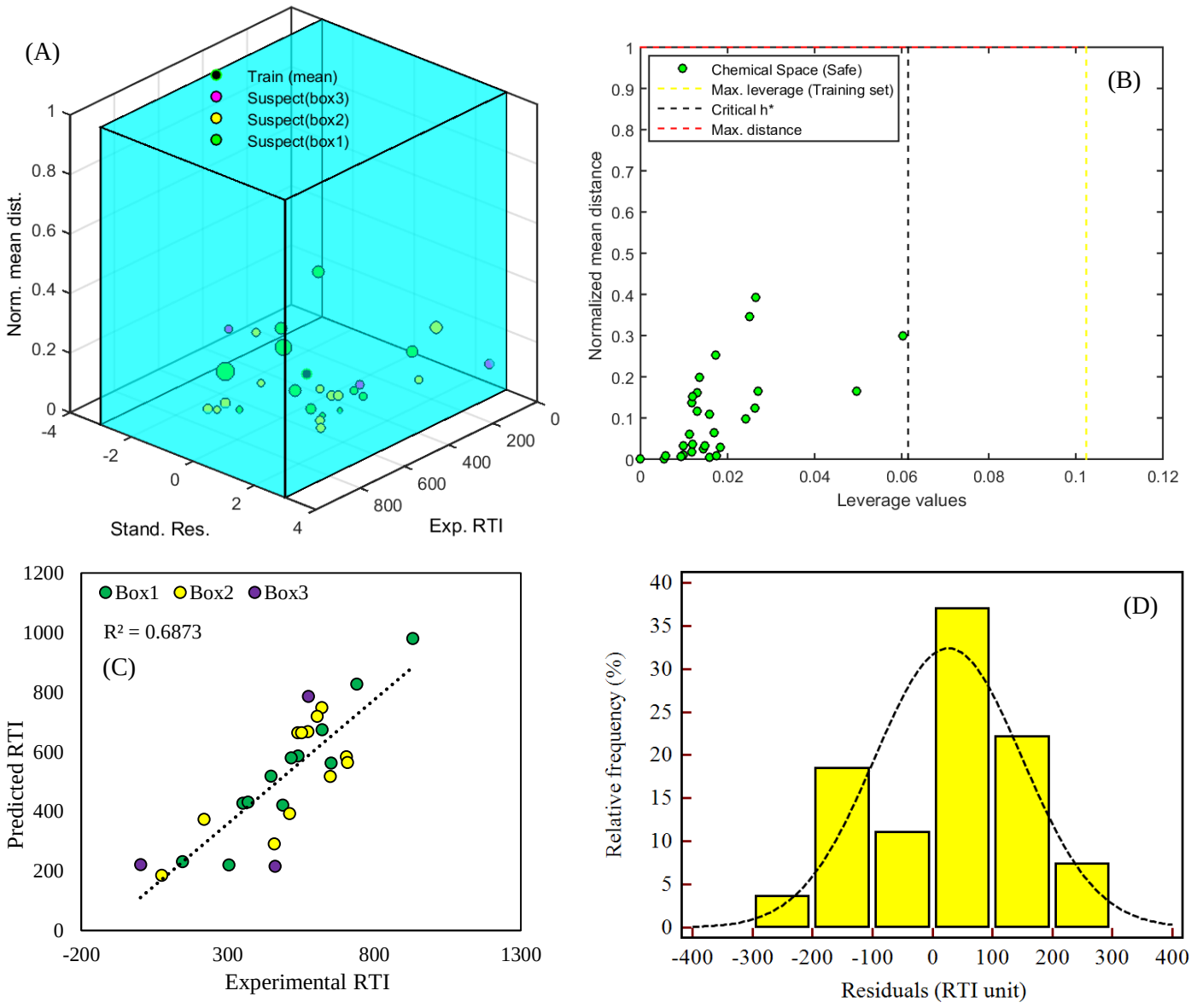


Figure 4.6 QSRR based prediction results with its boundaries for the compounds reported by UJI; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results.

4.3.6. Evaluation of RTI proposed in +ESI

Similarly done in –ESI, the same inter-laboratory study was carried out for the 18 compounds proposed as RTI calibrants in +ESI mode.

UFZ has successfully performed the proposed RTI system in +ESI and could observe high internal (quality of calibration curve between RTI and t_R) and external accuracy (a large set of 607 compounds as evaluation set). 22 common compounds existed in our dataset and the one provided by UFZ were projected by their experimental RTI values on the cloud plot (Figure 4.7A). 18 (81.82%) out of 22 compounds are overlapped and found to fall inside the RTI CIs. The result of multiple comparison procedure can be found in Table B 4.11. Since the internal accuracy (R^2= 0.9597) of calibration curve is high, the width of RTI CIs observed is significantly lower than those derived in –ESI mode. The prediction results were also highly correlated with experimental values. All in all, 340 (56.01%) compounds were predicted within box 1 (within ±90 RTI unit), 170 compounds (28.01%) within box 2 (within $+90 \leq |RTI| < +180$ in RTI unit) and 62 compounds (10.21%) within box 3 (within $+180 \leq |RTI| < +270$ in RTI unit) (Figure 4.8A). Additionally, 32 compounds were detected as outliers due to chemical dissimilarity and ambiguous retention time (such as 3-Nitro benzanthrone (Experimental RTI= 36.77; predicted RTI= 539.6361) which has constant logD value of 2.302 in pH=2-9, and it is not a polar compound to be expected to elute very early). Three compounds (Dimethyl dioctadecyl ammonium (which is extremely long chain (alkyl) quaternary ammonium compounds (QAC)), diatrizoate and N,N-Dimethyl sulfamide) are also found to be outlier due to the chemical dissimilarity (Figure 4.8B). Therefore, approximately 84.02 % of whole compounds could be predicted successfully with high confidence. The correlation between the experimental and predicted RTI as well as distribution of the error can be found in Figure 4.8C and D, respectively. TPHR as well as internal and external accuracy for the results reported by UFZ can be found in Table 4.7.

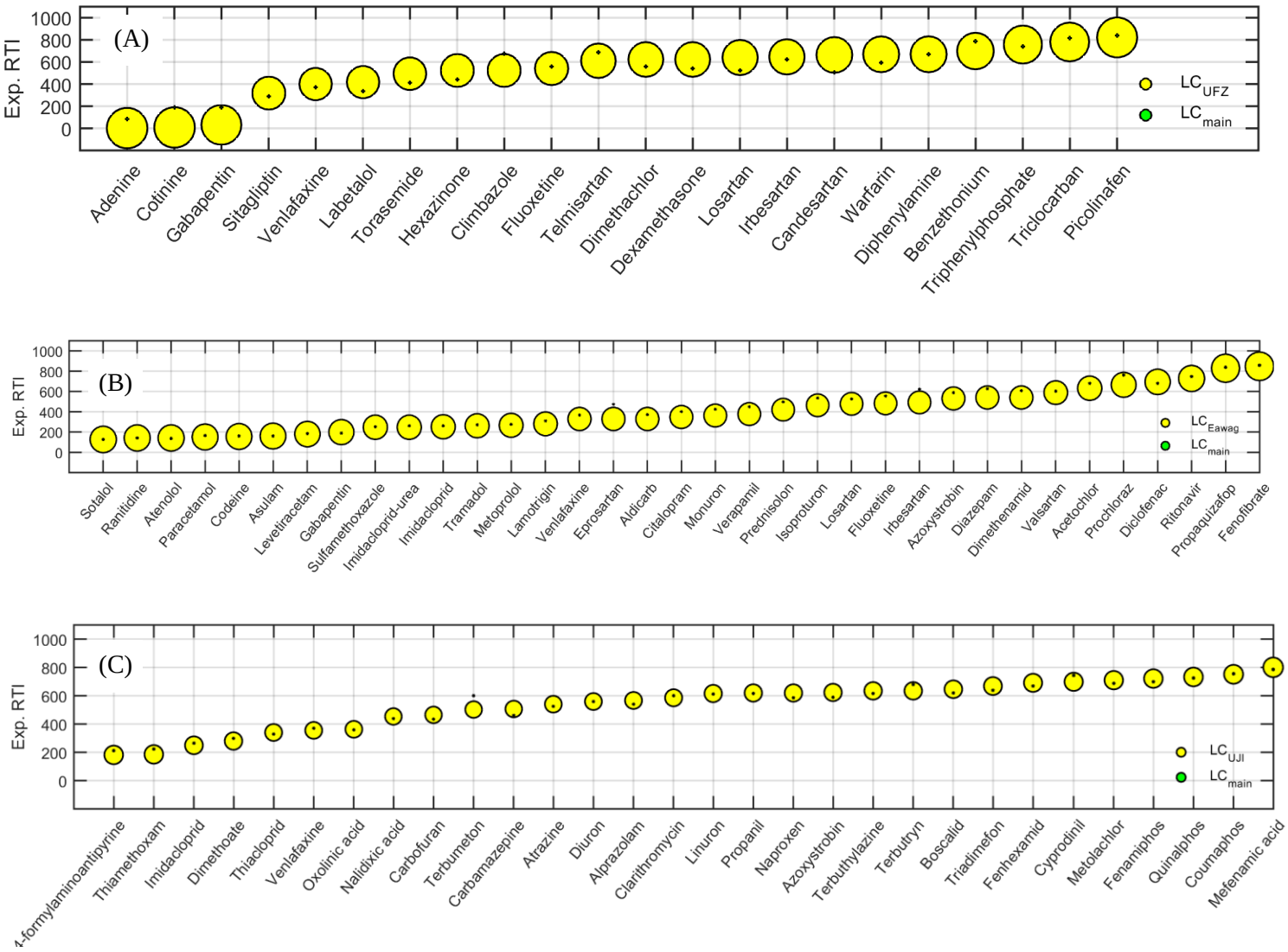


Figure 4.7 Cloud map of RTI values in different LC conditions reported by (A) UFZ; (B) Eawag; (C) UJI in (+ESI)

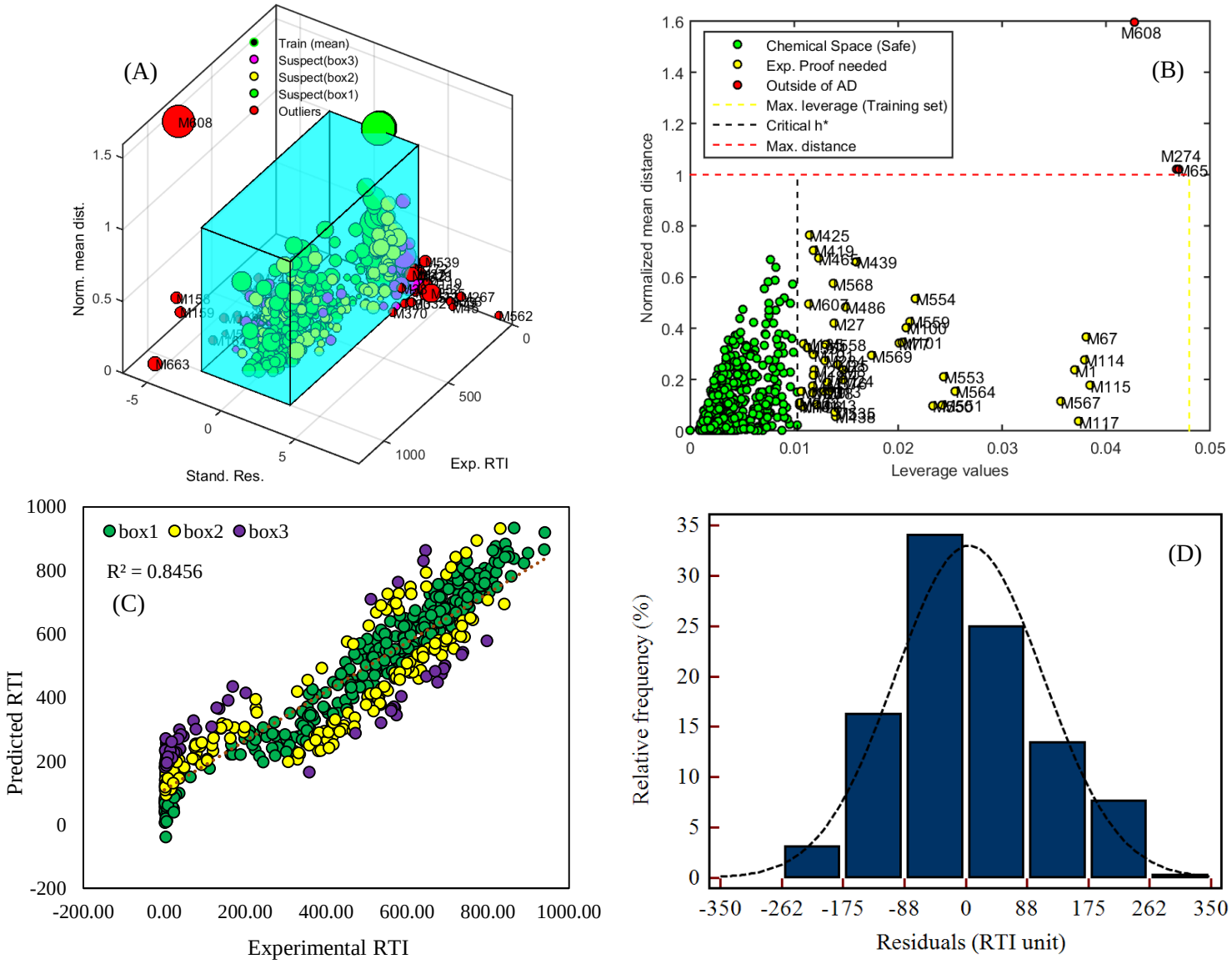


Figure 4.8 QSRR based prediction results with its boundaries for the compounds reported by UFZ; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results.

Table 4.7 RTI and tR linear equations for (+ESI) reported by different laboratories

Labs	RTI versus t_R equation	Standard Error	Internal accuracy	External accuracy
UFZ	RTI = 62.17 (t_R) – 23.53	Intercept : ±31.30 Slope: ±3.535	R^2=0.9597	R^2=0.8456, RMSE=105.8193, TPHR=100% (n=22)
Eawag	RTI = 60.03 (t_R) + 18.00	Intercept: ±27.27 Slope: ±3.068	R^2=0.9672	R^2=0.8835, RMSE=90.4343, TPHR=97.14% (n=35)
UJI	RTI = 63.34 (t_R) –18.72	Intercept: ±13.12 Slope: ±1.408	R^2=0.9926	R^2=0.8757, RMSE=81.8010, TPHR=96.67% (n=30)

Similarly, Eawag evaluated the RTI system and reported 55 compounds in +ESI. 35 compounds existed in our dataset and the one provided by Eawag were projected by their experimental RTI values on the cloud plot (Figure 4.7B). The result of multiple comparison procedure for these 35 compounds can be found in Table B 4.12. 33 (TPHR=94.29%) of 35 compounds are overlapped and found to fall inside the RTI CIs. The width of RTI CIs are also found to be smaller for polar compounds (eluting below 200 RTI) than non-polar ones. This is because the t_R versus RTI shows a polynomial (second-order) relationship as the t_R values increase and it deviates from linearity. Therefore, the fraction of error for partially and non-polar compounds can associate with the error caused by the linear calibration curve of chromatogram. The prediction results were highly correlated with experimental values. Totally, 40 (72.73%) compounds were predicted within box 1 (within ±75 RTI unit), 13 compounds (23.64%) within box 2 (within $+75 \leq |RTI| < +175$ in RTI unit) and 2 compounds (3.63%) within box 3 (within $+175 \leq |RTI| < +235$ in RTI unit) (Figure 4.9A). Therefore, 96.36 % of whole compounds successfully could be predicted with high confidence through the RTI system proposed. Moreover, no outliers are found due to the chemical dissimilarity (Figure 4.9B). The correlation between the experimental and predicted RTI as well as distribution of the error can be found in Figure 4.9C and D, respectively. Internal and external accuracy for the results reported by Eawag are tabulated in Table 4.7.

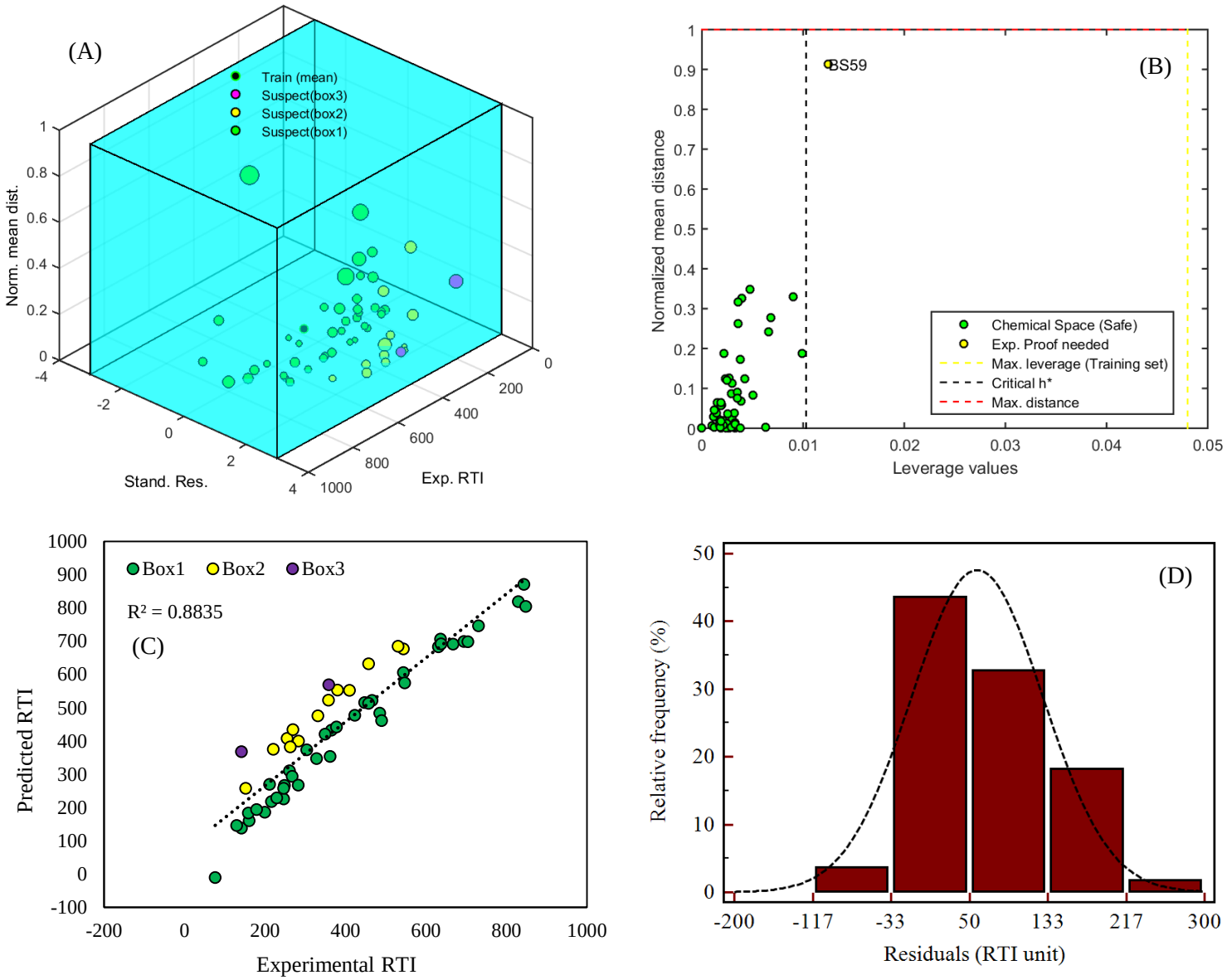


Figure 4.9 QSRR based prediction results with its boundaries for the compounds reported by Eawag; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results.

UJI was evaluated the RTI system and reported 44 compounds in +ESI. In addition, 30 compounds existed in our dataset and the one provided by UJI were projected by their experimental RTI values on cloud plot (Figure 4.7C). The result of multiple comparison procedure for these 30 compounds can be found in Table B 4.13. 29 (TPHR=96.67%) of

30 compounds are overlapped and found to fall inside the RTI CIs. The width of RTI CIs are also lower than that obtained by UFZ and Eawag. These RTI CIs were also remained constant throughout the chromatogram which indicates that the calibration curve is very well established (R^2=0.9926). The prediction results were highly correlated with experimental values (R^2=0.8757). Totally, 28 (63.64%) compounds were predicted within box 1 (within ±80 RTI unit), 14 compounds (31.82%) within box 2 (within $+80 \leq |RTI| < +160$ in RTI unit) and 1 compound (2.27%) within box 3 (within $+160 \leq |RTI| < +250$ in RTI unit) (Figure 4.10A). Therefore, 95.46 % of whole compounds could be predicted successfully with high confidence through the RTI system proposed. Moreover, none of the compounds are outlier due to the chemical dissimilarity (Figure 4.10B) and only one compound (Methamidophos) was detected as outliers due to its ambiguous elution (Experimental RTI= 69.9536 and predicted RTI= 351.5146). The correlation between the experimental and predicted RTI as well as distribution of the error can be found in Figure 4.10C and D, respectively. TPHR together with internal and external accuracy for the results reported by UJI are listed in Table 4.7.

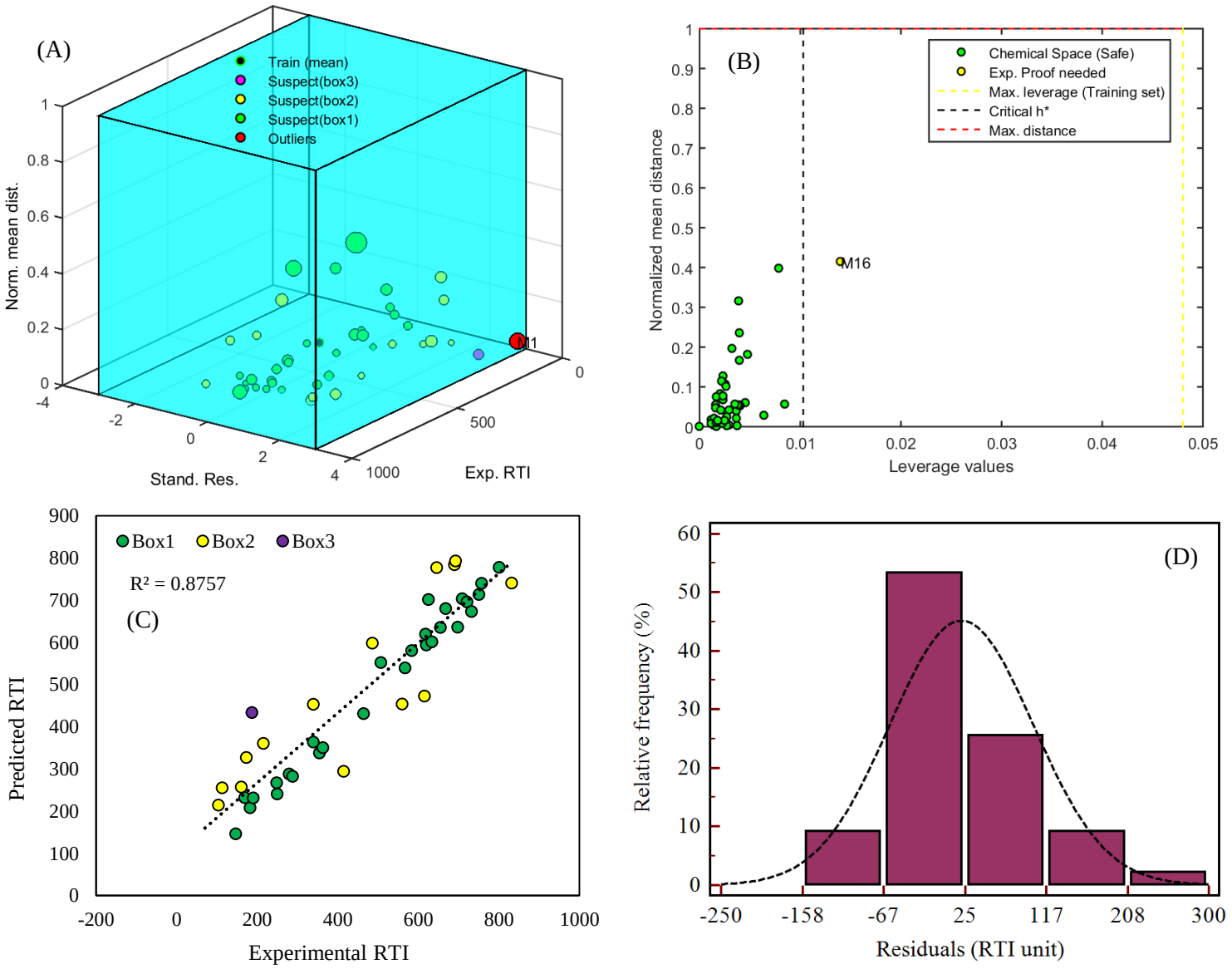


Figure 4.10 QSRR based prediction results with its boundaries for the compounds reported by UJI; (A) OTrAMS method; (B) leverage versus normalized mean distance to define the chemical space failure; (C) correlation between experimental and predicted RTI; (D) distribution of error associated with prediction results.

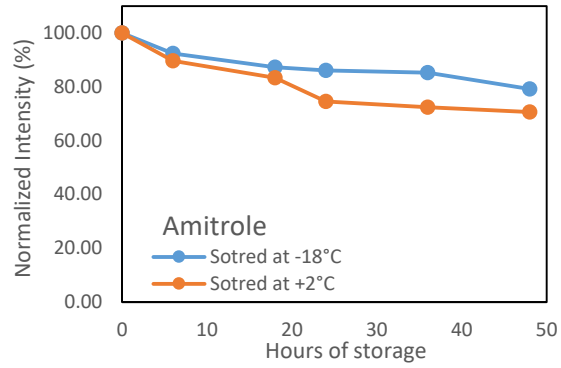
4.3.7. External evaluation

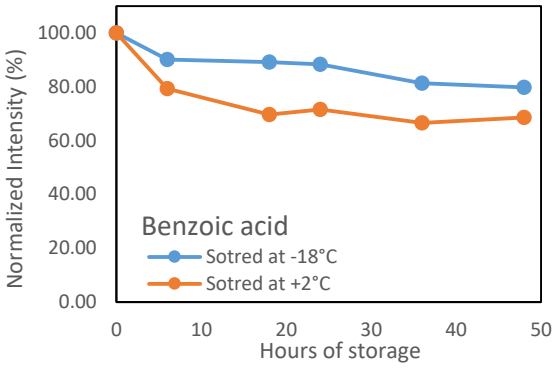
Fifteen laboratories were evaluated the RTI system proposed here with their own LC conditions listed in Table B 4.2. The evaluation result of RTI system in each LC condition is briefly explained in Table B 4.14. The internal accuracy for the calibration curve of 18 calibrants are higher than 0.94 square correlation coefficient in both ±ESI modes. High accuracy was also achieved for set of compounds as evaluation set after calibrating t_R in each laboratory. For instance, the first laboratory tested RTI system uses completely different mobile phase composition and stationary phase (Kinetex Biphenyl core-shell particle) comparing to the one used here as the nominal case, and high correlation was observed between experimental and predicted RTI for their external evaluation set. Table B 4.15 lists several examples where the ability of harmonization of tR values via RTI system is demonstrated. As it can be seen, regardless of the LC conditions used by each laboratory, the RTI values for each compound are statistically identical while the t_R values are significantly different.

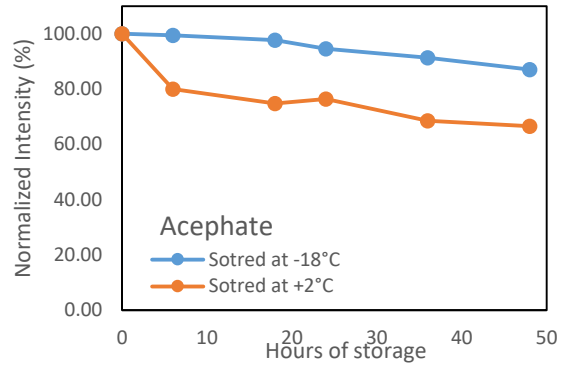
4.3.8. Stability test of RTI calibrants

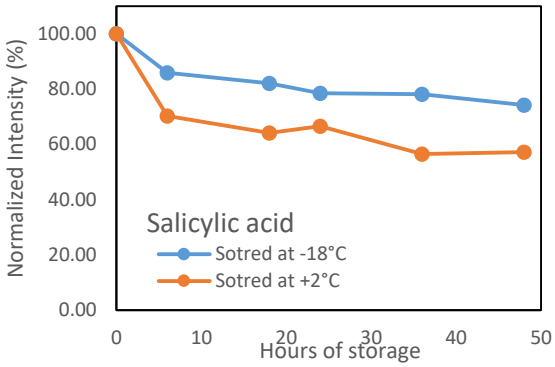
The stability test of each individual calibrant in the prepared mixture is provided in Figure 4.11. As it can be seen, the RTI calibrants are stable within 48 hours of consecutive analysis.

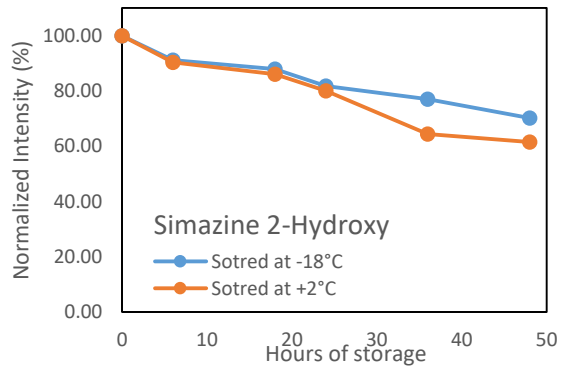
RTI calibrants for -ESI:

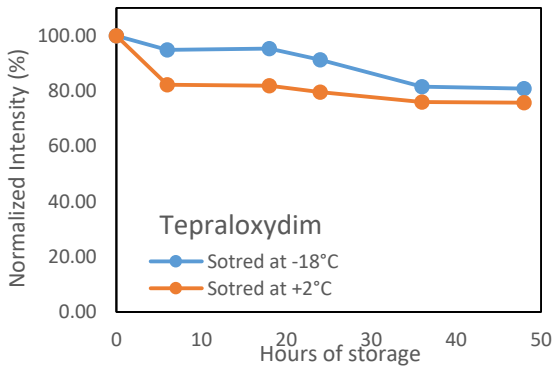


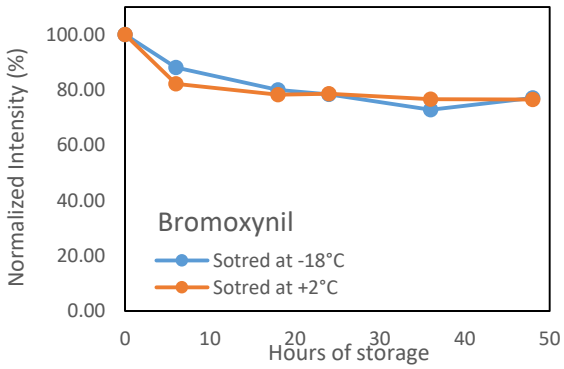


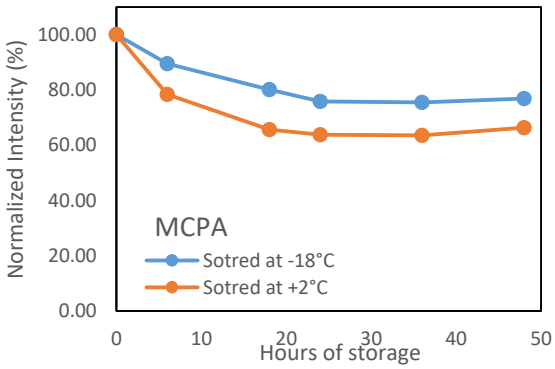


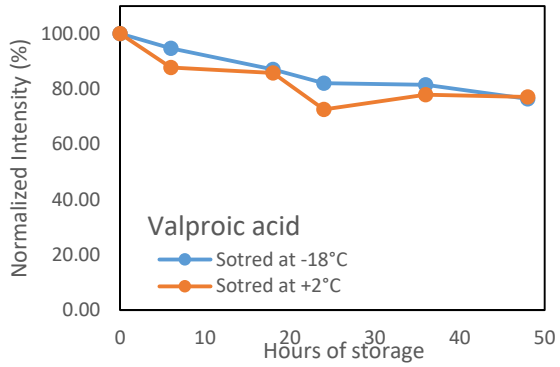


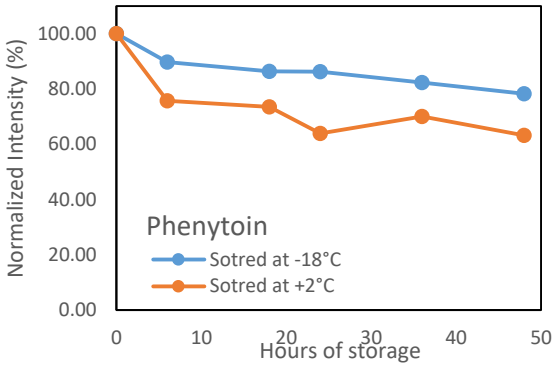


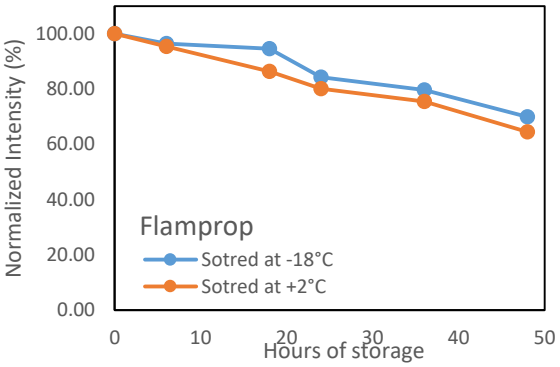


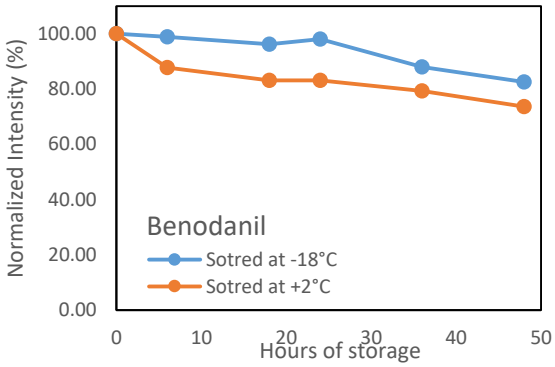


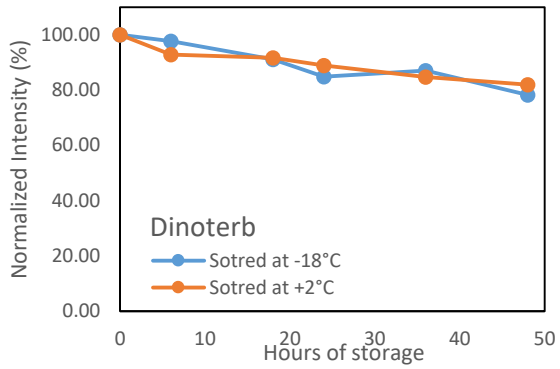


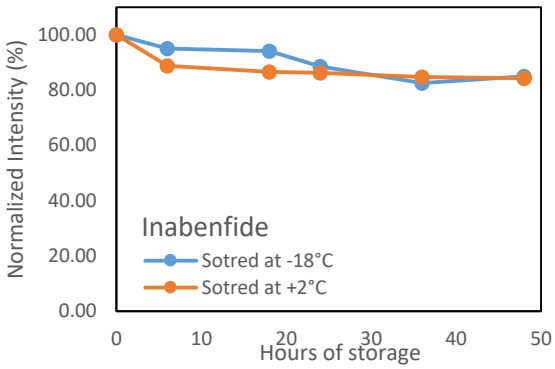

Valproic acid
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

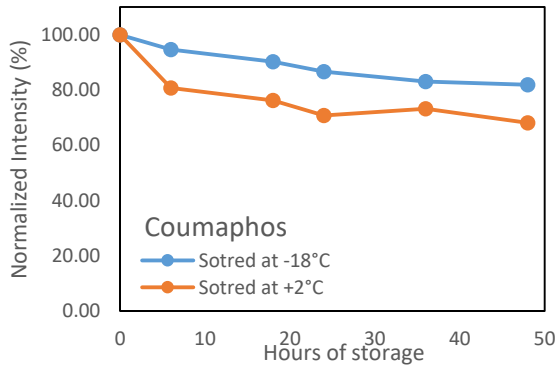

Phenytoin
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

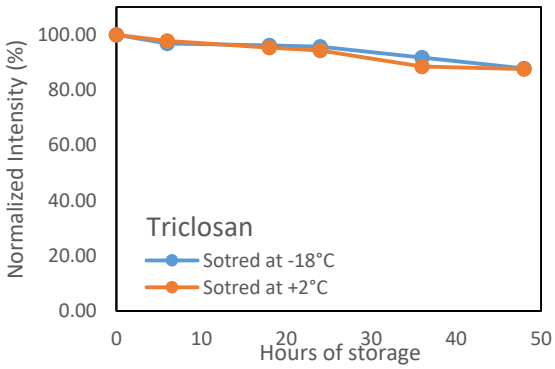

Flamprop
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

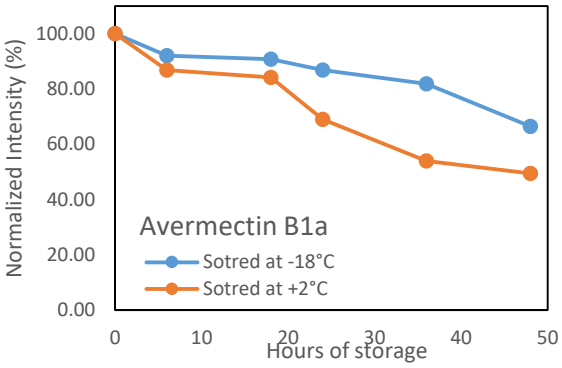

Benodanil
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

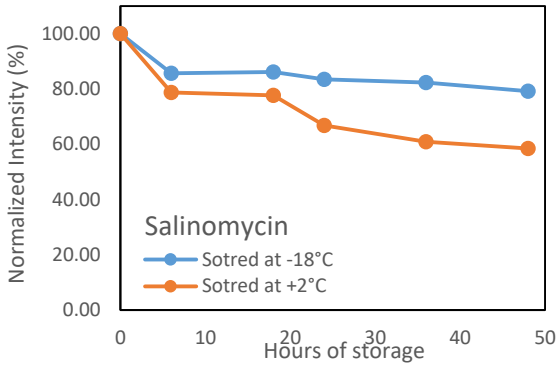

Dinoterb
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage


Inabenfide
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

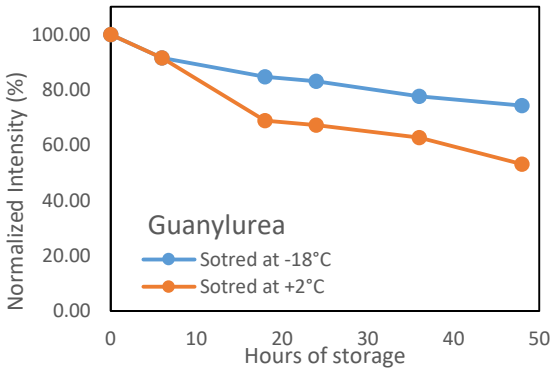

Coumaphos
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

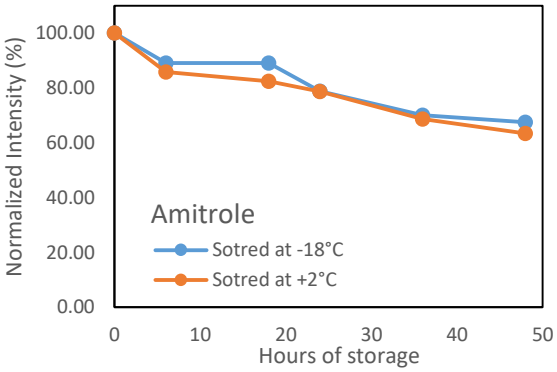

Triclosan
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

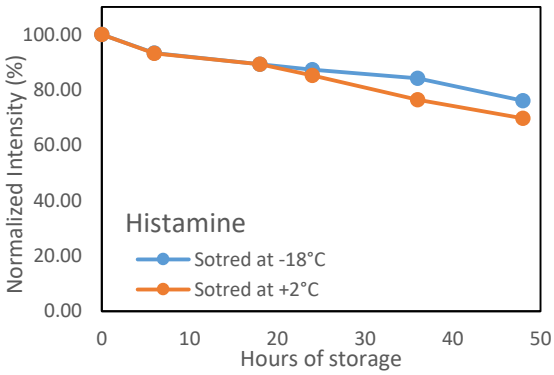


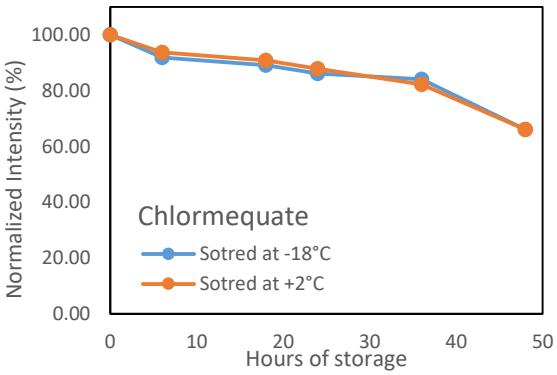


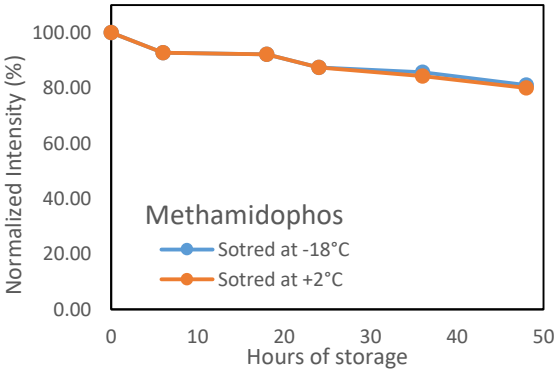
RTI calibrants for +ESI:

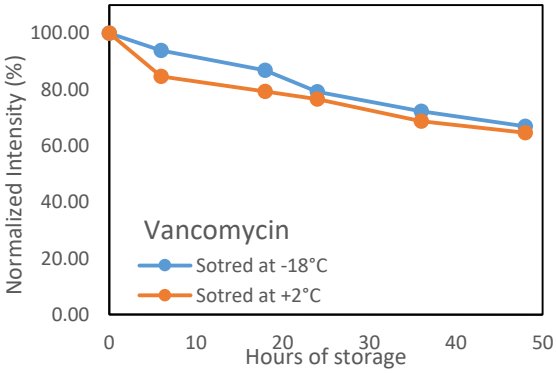


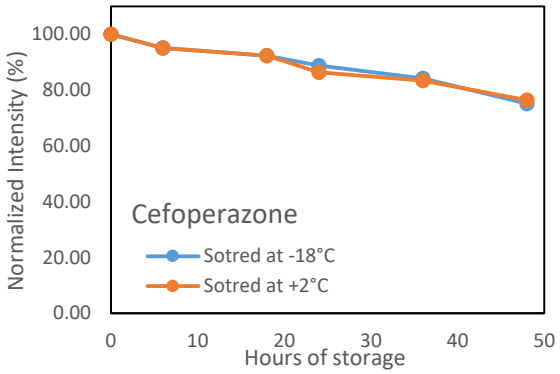


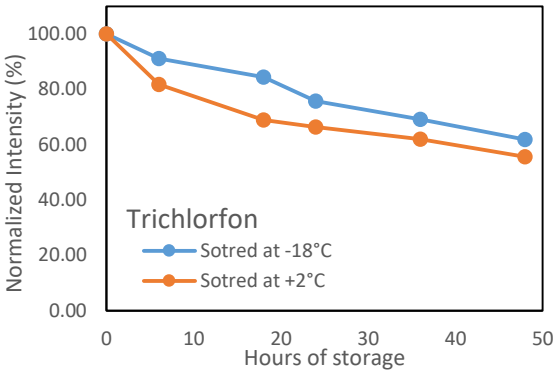


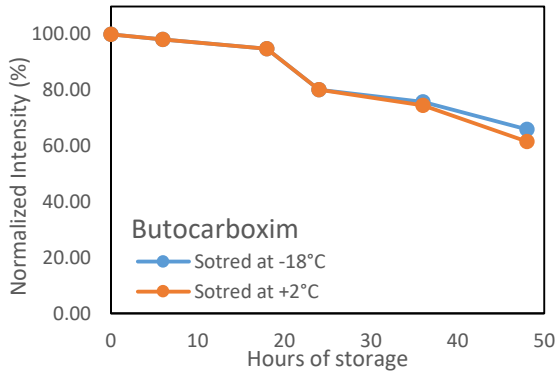


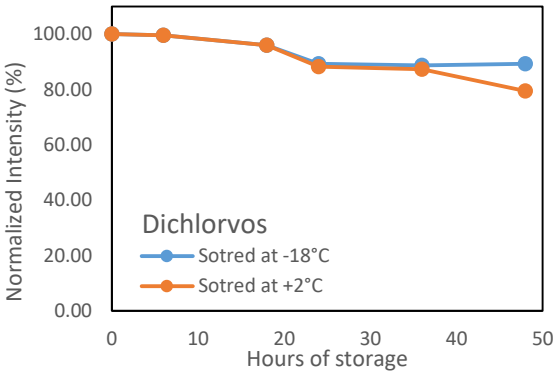


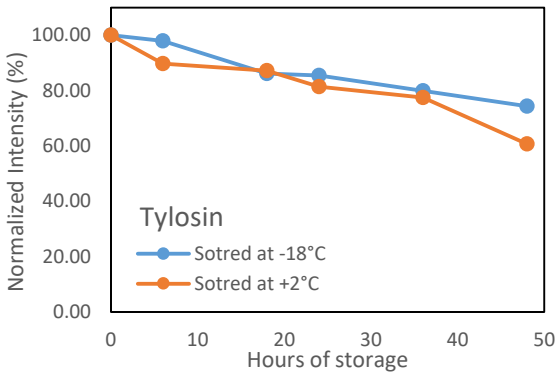


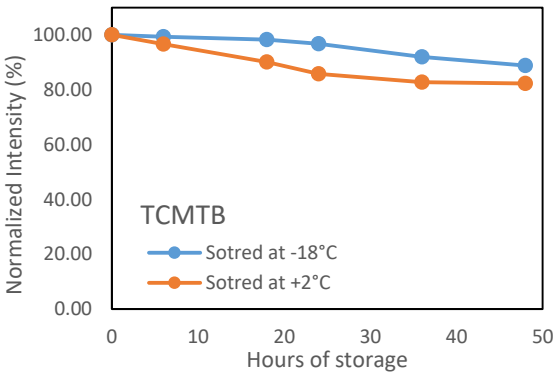

Cefoperazone
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

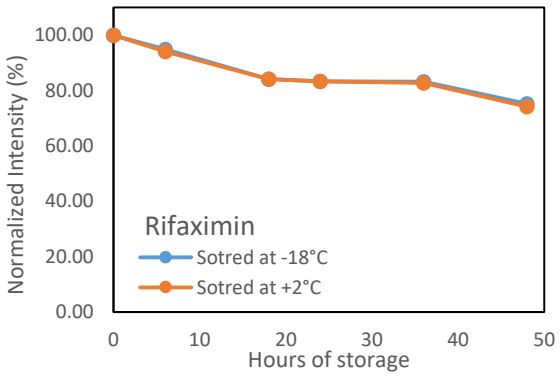

Trichlorfon
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

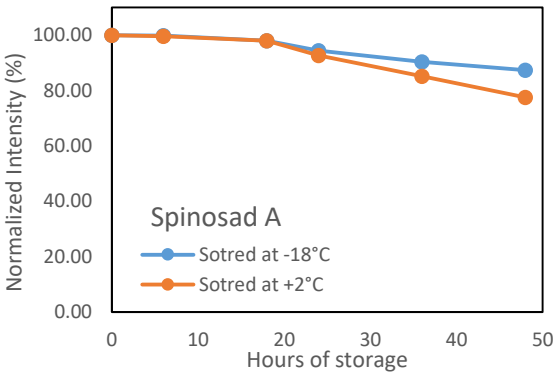

Butocarboxim
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage


Dichlorvos
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage


Tylosin
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage


TCMTB
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage


Rifaximin
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage


Spinosad A
Sotred at -18°C
Sotred at +2°C
Normalized Intensity (%)
Hours of storage

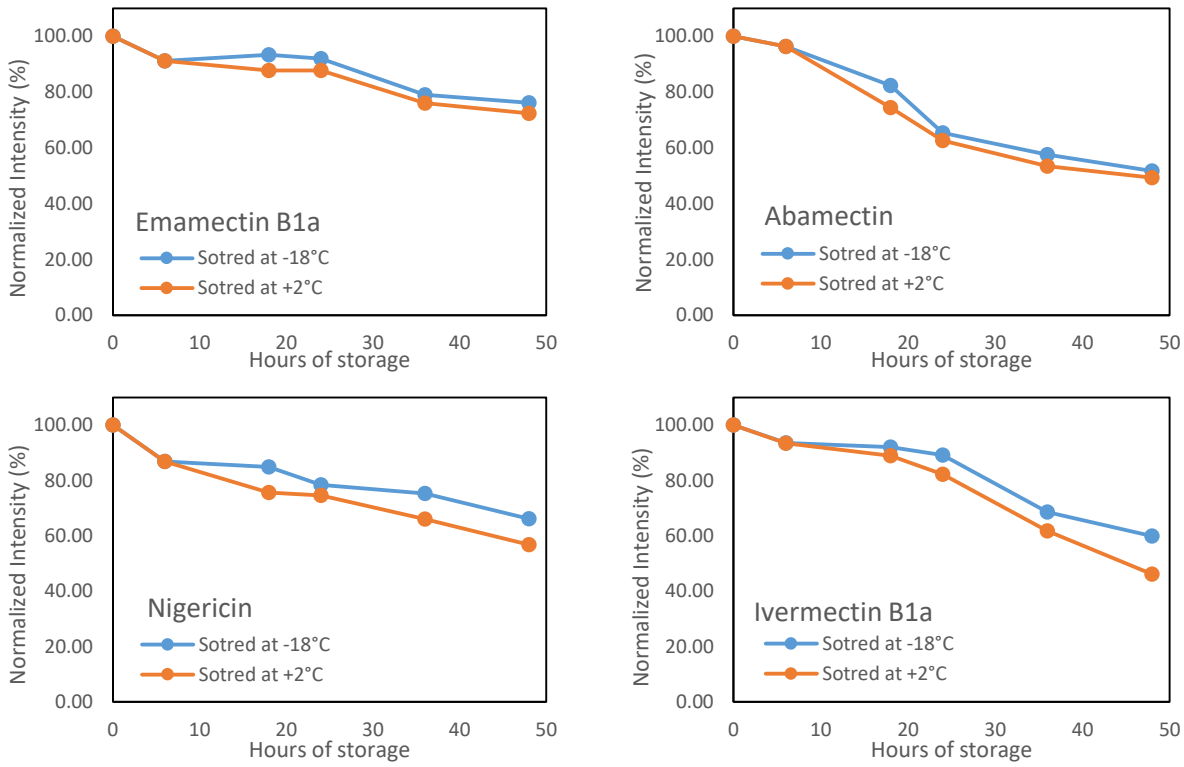


Figure 4.11 Stability test for the RTI calibrants

4.3.9. Application of RTI in Suspect and Non-target Screening

The proposed RTI approach has been applied successfully in a collaborative trial coordinated to perform untargeted analysis of indoor dust samples [148]. Twenty-one participants reported results among which 12 participants (analyzing the dust samples with LC-HRMS) utilized the RTI system successfully. Through the RTI system, some compounds such as diethylene glycol and morphine are found to be false positive in the report of some laboratories (although the reported level of identifications were level 3 and 4) [125]. Network of reference laboratories, research centers and related organizations for monitoring of emerging environmental substances (NORMAN network), in connection with the Digital Sample Freezing Platform (DSFP) (online LC-HRMS data archive and the retrospective screening platform (https://norman-data.eu/)) have already utilized the proposed RTI system in routine analysis of emerging contaminants and incorporated it in DSFP platform.

4.4. Conclusions

A simple approach was proposed to harmonize retention time (t_R) in various liquid chromatographic (LC) conditions. The harmonization of t_R is based on the calibration of LC via 18 calibrants into predefined retention time indices (RTI: 1-1000). This indices (RTI) was built based on establishing a linear calibration curve between elution of the 18 compounds and their RTI values in various LC conditions. These calibrants were selected chemometrically, using ant colony optimization similarity indices which trains to identify the potential subsets of compounds that their selection increases the overlap of normal distribution of calibrants and rest of chemical space. This approach was dynamically assessing the overlap of normal distribution of various combinations of compounds (calibrants) with the rest of unselected compounds (>2080). Intra-laboratory and interlaboratory evaluations showed a successful application rate of above 90% in ±ESI. The proposed RTI system could be subject of LC quality assurance by checking the cloud plot, multiple comparison procedure and confidence intervals for the calibrants. Moreover, it can be used to detect the LC variability over time or column ageing and recalibrate the LC condition. Quantitative structure−retention relationship (QSRR) models were applied to predict the RTI. These models are significantly helpful to achieve a certain identification confidence level in case of several plausible candidates. Thus, these models are not LC system dependent and can be applied under any LC condition which shows high internal accuracy and acceptable degree of linearity for calibration curve. These models were simple, accurate and interpretable which were mainly based on logD and ionic interaction between compounds and stationary phase. Applicability domain and chemical space boundaries of the models were also addressed to facilitate the validity of prediction results.

The successful rate of correctly harmonizing the t_R values was found to be lower for ACN (90.05% after comparison of additional 422 compounds between 9 LC condition) than methanol (100%) as mobile phase composition. The pH of mobile phase was also found to affect the elution pattern of calibrants in –ESI. Thus, these calibrants can be used to adjust the best LC condition as the accuracy is increasing at correct pH range in –ESI. We believe that this RTI system, not only will help with the identification process throughout various LC conditions, but also can be used to check the reproducibility and

quality of LC conditions used to perform the identification tasks. The QSRR models with defined applicability domain behind this system will also allow researchers to complete the feature annotation and compound identification process with high confidence.

.

CHAPTER 5

Prediction of Acute Toxicity of Emerging Contaminants on the Water Flea Daphnia magna by Ant Colony Optimization - Support Vector Machine QSTR models

5.1. Introduction

Modern societies largely depend on a wide range of down-the-drain products, such as personal care products or household washing agents, containing multiple chemical compounds that finally end up in the aquatic environment, together with their environmental transformation products and manufacture by-products. Increasing contamination of freshwater resources with chemical pollutants has therefore become a major public concern in almost all parts of the world,[69] resulting in the introduction of respective chemical regulations to assess associated risks and to ensure the restriction or ban of the most problematic compounds. In Europe, about 100,000 industrial chemicals are registered under the REACH Regulation, of which 30,000 to 70,000 are in daily use.[70] Depending on the amount produced or imported (in quantities above 1.0 tonne/year), companies are required to submit at least a base set of ecotoxicity data during registration, consisting of acute toxicity toward algae, fish and invertebrates to evaluate the associated risks.

Following the public request to avoid animal testing, companies are now allowed to submit waiver or estimates of alternative prediction methods instead of experimental data for these compounds, resulting in associated uncertainties as regards the potential risk of these compounds. This is especially problematic for compounds that have already been found in the environment and hence may be a potential harm to it. The use of these compounds, could be restricted or banned in a post authorization process, given a proper evaluation of their risks.

Given the large number of compounds that need to be evaluated in the REACH context, a promising way forward would be to include low cost screening methods that allow the identification of chemicals with substantial toxicity for the endpoints of interest as a first tier to justify the request for an experimental study. Moreover, such models could also be used to derive preliminary PNECs (Predicted No Effect Concentration) for

compounds already found in the environment to prioritize them based on potential risks. [149, 150]

Several QSTR (Quantitative Structure-Toxicity Relationship) models have been developed to explore the acute toxicity of various substances belonging to pharmaceuticals, pesticides or personal care products in aquatic environment and particularly toward the water flea Daphnia magna.[151-158] In most toxicity prediction surveys, the prediction models suffers from small datasets[152, 154] which can limit the subsequent application of these models. Another aspect is the quality of fitting results and the methods used to model toxicity[158] which can lead to either under-fitting or over-fitting issues. Considering various types of molecular descriptors that can explain the toxicity of a compound is highly encouraged as some studies proved that logkow (logarithmic octanol/water partition coefficient) is not the only molecular descriptor to be used to accurately model toxicity.[151, 153, 155, 159] Predicting toxicity becomes an easier and more accurate task if a dataset is a result of a curation step to filter out structurally diverse chemicals.[156] However, having larger datasets is beneficial due to future applications by covering a larger chemical space for so far untested compounds. Therefore, accurate aquatic toxicity predictions require a large set of compounds, additional molecular descriptors to compensate for example the logP/ $\log K_{ow}$ failure in ionizable compounds for robust modeling techniques with defined applicability domain. It is still a bottleneck to define the applicability domain of a model for a new compound with unknown toxicity as the only information available is the chemical structure and its similarity towards an existing dataset.[156, 160]

In toxicity modeling, logP/ $\log K_{ow}$ is the only molecular descriptor that we have a clear a priori knowledge about its mechanism of toxic action.[156] This is mainly due to the interaction between the cell membrane and the compound, which can correlate indirectly to its water solubility .[161] It is expected to have higher toxicity for compounds showing higher logP/ $\log K_{ow}$ values. It is still ambiguous or not well established how different molecular properties, such as polarizability, electronegativity, solute interaction or hydrogen bonding[162] and molecular descriptors that are a function of pH (like pk_a and number of donor/acceptor functional group) can affect the toxicity values of a compound. Besides QSTR models, [163, 164] other computational prediction methods, such as k-

Nearest Neighborhood (kNN) being a sort of automatic read-across from existing data, have been used to estimate the acute toxicity to standard test species, such as Pimephales promelas[165] or Daphnia magna.[156] It has been proposed to use multiple computational methods, to verify experimental or predicted toxicity data in a consensus approach[156], assuming that the weaknesses of one model will be counterbalanced by the strengths of the others and vice versa, provided that all models have been validated appropriately.

The aim of the current paper was to derive a SVM (Support Vector Machine) model for the acute toxicity to Daphnia magna, using the one of the most comprehensive datasets so far, and to verify its applicability domain with a diverse evaluation set of compounds, covering various classes of chemicals. Moreover, the tool OTrAMS [9] is introduced, allowing to identify the most likely correct toxicity of two or more deviant experimental test values or to verify the prediction result of another QSAR model (in terms of consensus), combining a test of the chemical applicability domain and the toxicity likelihood.

5.2. Materials and Method

5.2.1. Dataset preparation and chemical-toxicity curation

Experimental acute toxicity data (48-h LC_{50}, lethal concentration 50%) for a total of 2174 tests toward the water flea Daphnia magna was compiled from Kühne et al.[156], Cassotti et al. [155], Sangion and Gramatica [151, 152], T.E.S.T [157], the OCHEM platform[166] and five values for C_{10}-C_{14} LAS from the updated HERA report (HERA 2004, http://www.heraproject.com). This dataset was then split into a training and a test dataset for model development, based on Kühne et al.[156], as well as a third evaluation dataset consisting of the remaining data sources (including redundant toxicity entries from other laboratories) as described below. These original data entries were then chemically curated, following eight main steps[167]: (1) the initially available chemical identifiers (CAS number or SMILES) were unified into InChI; (2) 2D structures of the InChI were created and the dative bonds (e.g. nitro group) were standardized using Open Babel (http://openbabel.org/docs/current/)[168]; (3) Salts, metals and solvents were removed from the chemical structure; (4) the octect number was fixed and hydrogens were added;

(5) 2D structures were created using Open Babel and 3D structures were obtained out of various tautomer forms (the tautomer with the lowest energy was retained to get one structure out of different forms of a duplicate entries) using Balloon[135]; (6) Create a master SDF file with optimized 3D structures for all entries; (7) optimized InChI chemical identifier were created from the SDF file; (8) Duplicates to the main dataset of Kühne et al.[156] were moved to the evaluation set by comparing their optimized InChI, and a second SDF file with the retained structures of the evaluation set was created.

After chemical curation, 22 very large molecules (i.e. > 600 Daltons) were removed from the datasets (mainly due to less uptake/permeability by cellular membrane) [161] and 39 further compounds were removed because the reported experimental toxicity exceeded the water solubility [156, 169] of the compound by more than ten times. Moreover, structures that are generally known to be difficult to predict in the current QSTR models, such as surfactants, ionizable compounds (and their salts) or permanent charged substances (such as quaternary ammonium compounds) were removed from the main set and put in the evaluation set. All in all, the 2174 toxicity values refer to 1353 unique compounds that were split into toxicologically consistent training (1026) and test (327) sets to derive a robust model, as well as 660 compounds (including 220 new compounds) of the external validation set that were used to test the accuracy and to verify the applicability domain under realistic conditions. The final list of compounds used for the test and training dataset and the evaluation set can be found in Table A 5.1 and Table A 5.2 (supplementary material (SM, Chapter 5, appendix A)), respectively. The division into training and test set was achieved by Kennard-Stone algorithm. [170] Kennard-Stone algorithm starts by selecting the pair of points (in our cases compounds and created molecular descriptors) that are the furthest apart. The selected compounds were assigned to the training sets and removed from the list of compounds. Then, the next pair of compounds which are furthest apart are assigned to the test set and removed from the compound list. In a third step, the procedure assigns each remaining compound alternatively to the training and test sets based on the distance to the compounds already selected. The distance function used here is Euclidean distance.

5.2.2. Molecular descriptors calculation

After performing the chemical curation workflow, the final 3D optimized structures were used as input to calculate molecular descriptors using Padel.[131] 2756 molecular descriptors were calculated for each compound, representing the constitutional, topological, geometric, electrostatic, hydrophobicity, steric effect, quantum related chemical descriptors, chemical fingerprints (Pubchem fingerprint)[171] and various 3D molecular descriptors[137-139]. The calculated molecular descriptors were then screened for existence of constant and near constant cases for their removal. The combined test and training dataset consisting of 1353 compounds and 1225 molecular descriptors for each compound was further processed with collinearity removal. A threshold of above 0.9 for the inter-correlation removal was set. In case of detection of high inter-correlation for couple of descriptors, a molecular descriptor showing a lower correlation with the toxicity was removed while its pair was retained. To this end, the toxicity values were converted to logarithmic scale of pLC_{50} (mol/L) as recommended by Artem et al.[130] to allow for molecule-to-molecule activity comparisons. Finally, 826 molecular descriptors were retained for the molecular descriptors selection step.

5.2.3. Molecular descriptor selection and modelling

ACO is applied for selecting the molecular descriptors, and MLR and SVM was used for building the regression methods to model toxicity data. ACO is a swarm intelligence algorithm that is inspired from behavior of ants searching for food resources nearby their nest using pheromone deposition without any visual information [141, 142]. ACO is presumed to be a good method for molecular descriptor selection because it can solve complex optimization or feature selection problems [172]. For an ACO based feature selection case, the algorithm starts with the generation of a certain number of ants (here we set this number to 200 ants) placed randomly on the graph that represents a random molecular descriptor to start. From every starting feature, ant constructs a path through the search space (which is corresponding to a feature subset). Thus, each node (in a graph) relates to a molecular descriptor and each edge shows the traversal of an ant to travel from one feature to another. The number of artificial pheromones [0, 1] for an edge is associated with the popularity of the particular traversal by previous ants. Therefore,

ants can make a probabilistic decisions at each node which traverse to use next, based on the amount of artificial pheromone and related traversal degree. This continues until the minimum degree for the objective function (here, the fitness function was root mean square error (RMSE) of a leave-one-out cross-validation (Q^2_{LOO})) has reached a minimum, otherwise, the information in each edge were being updated and a new set of ants were created and the whole process was started all over again [141, 142]. We set the maximum number of iteration to 100, while the desired number of molecular descriptors was followed up to 6 features by checking the Variance Inflation Factor (VIF) threshold. VIF addresses the multi-collinearity And we used the acceptance cut-off value of below 5.0 to add another descriptor [173]. A descriptor exceeding this threshold can adversely affect the model and it can be an indication to rather exclude it from the selection of molecular descriptors. The Evaporation Rate (ER) is a method that causes all pheromone values to decrease uniformly. ER was set to 0.05 (this value is being kept constant during performing ACO and is generally a small value (0.01-0.05)) [142]. From a practical point of view, pheromone evaporation is required to prevent a too rapid convergence of the algorithm toward a sub-optimal chemical space. It presents a useful form of forgetting and cause exploration of new areas in the chemical space. The ACO algorithm was written and performed in MATLAB.

Multiple Linear Regression (MLR) and SVM are two modeling techniques used to correlate the molecular descriptors selected by ACO with the respective toxicity end-point. In both techniques, the RMSE of $Q^2_{LOO}$was considered to train the models [9]. The MLR is simple and requires no need for optimization of any internal parameter. SVM is based on linear or nonlinear RBF kernels, and thus can be applied to improve correlation of data with nonlinear nature. In SVM, the basic idea is to map the data X into a higher-dimensional feature space via a nonlinear mapping function (Φ) and then to do linear regression in this space [174]. However, SVM is more complex, using three internal parameters (C (which is a regularization constant), ε (ε-insensitive loss function), the kernel type (γ), and corresponding kernel parameters (here was radial basis function (RBF))) and should be optimized before proceeding to the final training/prediction step. More information about these parameters can be found in Vapnik [175].

5.2.4. Validation criteria

More details about the validation methods used here can be found in chapter 3 (section 3.2.2, validation criteria for QSAR methods) of the thesis.

5.2.5. Identification of potential toxicity outliers using OTrAMS

Here a brief description of all applicability domain methods is provided, as it is very important step in *in silico* toxicity approaches. This is generally based on the idea that similar compounds could show similar effects in the environment (similarity done in case of read across approach where sufficient amount of experimental data is available and certain percentage of chemical similarity exists between compounds in the dataset). The applicability domain defining the chemical space that a model is capable of covering should be part of any QSAR/QSTR workflow [176-178]. The Williams plot is considered to be a robust method to measure the applicability domain of any proposed model.[179] It is based on the leverage versus standardized residual values that the leverages are being estimated from molecular descriptors selected to build the QSPR model and is calculated as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i \qquad and \qquad h^* = 3(p+1)/n \tag{5.1}$$

where X is the molecular descriptors matrix, T is the indicator of the training set, x_i is the descriptor vector of each compound, n is the number of training set compounds, p is the number of molecular descriptors used as modeling variables, and h* is the warning leverage value and is a cut-off value to show that the chemical structures exceeding this threshold are outliers due to their high chemical structures dissimilarity [179]. The commonly used cut-off value for Standardized Residuals (SR) is ±3δ that covers 99% of the normally distributed data. Compounds which locate outside of this cut-off value would be considered outliers due to the abnormal response observed, however, compounds outside of the leverage cut-off value, but inside the standardized residual limits are considered as good leverages.

In addition to this method, Euclidean distance can be measured for training and test set, and then the mean distance for test set compounds can be normalized based on mean distance of training set versus observed property. This shows how the diversity of

chemical structures behaves toward the target property [180]. A test set compound outside the cut-off value of 1.0 (derived from normalization of mean distance of the training set), are considered to be outside of applicability domain of the model, and the training set is not representative for this test set compound. Last but not least, recently, Roy et. al [178]. developed a simple approach so called “Standardization approach (SD)” that found to be useful to identify chemically diverse compounds with unknown experimental endpoints. This method stands on calculation of a standard score and its normal distribution.

In our previous work, we coupled leverage vs standardized residuals and normalized mean distance in a single 3D bubble plot, so called OTrAMS, in which the z-axis shows the standardized residuals, the y-axis shows the normalized mean distance and the x-axis relates to the standardized observed property (i.e. acute toxicity) [9]. The bubbles size is proportional to the leverage values and are coded with color representing SR values (green (less than $-1.0 \leq SR \leq 1.0$), yellow ($1.0 < SR \leq 2.0$ or $-2.0 \leq SR < -1.0$), purple ($2.0 < SR \leq 3.0$ or $-3.0 \leq SR < -2.0$) and red ($SR > 3.0$ or $SR < -3.0$)). SR include the effect of chemical structure dissimilarity in the error calculation and, thus are considered more accurate in terms of understanding the origin of errors between experimental and predicted toxicity. Therefore, the window of acceptance for the error is relative to their similarity to the training set.

This helps to distinguish between compounds with good and bad leverage and to understand if the observed error is due to the deviating chemical structure or experimental measurement (which still could be correct, but would not be represented by the model). While the Williams plot is incapable of distinguishing between compounds found to be of good or bad leverage, this method could address whether the compound falls outside of the normalized mean distance of the training set. This is potentially a very useful tool to handle false positives in cases where several plausible candidates (e.g. isomers) were present for a single end-point or to decide which toxicity values to trust more when there is no numerical agreement between different data sources (e.g. due to wrong unit).

For a compound without an experimental measurement, the only way to address the applicability domain is to study the chemical space. Therefore, if the error is the function of chemical space failure, hat values, warning leverage values and also normalized mean

distance can be used to define the applicability domain. To this end, we used OTrAMS while training the model to understand the origin of outliers. In addition, we used a new method to show if the error is a function of chemical space failure, using hat values (a threshold applied by warning leverage value) versus normalized mean distance. This method offers several conditions where chemical space failure of a compound with unknown toxicity could be addressed: (a) The chemical space failure zone is the area above the normalized mean distance of 1 and leverage values higher than the warning leverage cut off. Any compounds found to be there are outside of the applicability domain of the proposed QSTR model and should not be predicted with this model; (b) The safe zone is the area where compounds are within the warning leverage and normalized mean distance limit. These predictions are accepted because these compounds are highly similar to the compounds in the training set used to build the model; (c) Last but not least, for compounds in the area that is exceeding the warning leverage cut-off limit, but they are within the limit of normalized mean distance and the maximum leverage value (in the training set), the prediction results are less reliable and in case of a resulting concern, values should be verified experimentally.

5.3. Results and discussion

5.3.1. Data treatment

The total dataset was divided into a consistent dataset of 1353 organic compounds together with 826 not inter-correlated molecular descriptors, split into a training set (including 1026 compounds used to build the QSTR model) and a test set (including 327 compounds used to evaluate the external prediction accuracy) and an external validation set of 660 compounds. The division of the consistent dataset was done carefully considering the fair distribution for both experimental toxicity and chemical space. This was done using the Kennard-Stone algorithm [170]. The ratio of 75 to 25 percent was used to divide the whole dataset into a training and a test subset. Figure 5.1 shows the overlap of the selected compounds in both datasets in terms of normal distribution of their toxicity level and chemical space. This proves that there is no bias toward the selection of training and test set compounds, i.e. their toxicity and chemical structure are normally distributed in both sets.

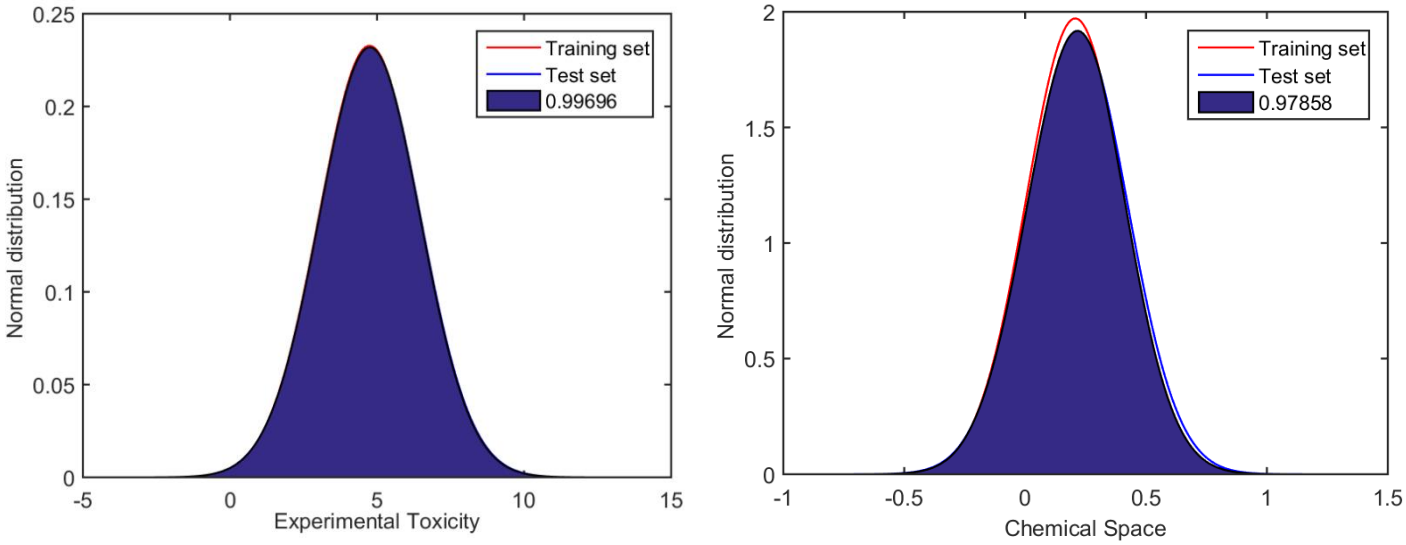


Figure 5.1 Overlap between the normally distributed toxicity levels and chemical space in training and test set.

5.3.2. Derivation of the ACO-MLR and ACO-SVM models

After division of the consistent dataset, ACO was used on the training set to select the best set of molecular descriptors. The final model was selected based on the lowest RMSE in the cross validation and fitting results. A MLR model with 6 molecular descriptors was developed according to validation criteria and defined rules for applicability domain studies without outliers as follows:

$$pLC_{50} = 2.2246\ (\pm 0.1034) + 0.2083\ (\pm 0.0293)\ AlogP + 1.7458\ (\pm 0.2587)\ AATSC0p + 0.2825\ (\pm 0.0336)\ CrippenlogP - 0.0527\ (\pm 0.0090)\ minsOH + 0.5630\ (\pm 0.0517)\ MLFER_{BH} + 0.3009\ (\pm 0.0332)\ XlogP \quad (5.2)$$

Correlation between molecular descriptors and variance inflation factor were calculated to show that the descriptors selected by ACO are not confounded, especially considering three different measures of logP that are believed to increase the accuracy of representing the correct logP for a compound in a consensus manner. For comparison we build a model with a consensus (geomen) logP, which had a somewhat lower predictability. This is making the proposed ACO-MLR model more accurate in terms of logP measurements, as all of them (AlogP, CrippenlogP and XlogP) have a positive mean

effect toward the prediction of toxicity in Daphnia magna. Table 5.1 shows the inter-correlation matrix and VIF for the selected descriptors.

Table 5.1 Intercorrelation between molecular descriptors and related VIF value.

	ALogP [a]	AATSC0p [b]	CrippenLogP [c]	minsOH [d]	MLFER_BH [e]	XLogP [f]	VIF [g]
ALogP	1.000						1.715
AATSC0p	0.224	1.000					1.246
CrippenLogP	0.600	0.058	1.000				3.279
minsOH	-0.197	-0.080	-0.195	1.000			1.063
MLFER_BH	-0.169	0.288	-0.083	0.000	1.000		1.240
XLogP	0.547	0.082	0.795	-0.191	-0.213	1.000	2.986

[a] Ghose-Crippen measure of logP (hydrophobicity)

[b] Average centered Broto-Moreau autocorrelation - lag 0 / weighted by polarizabilities

[c] Crippen's logP (hydrophobicity)

[d] Minimum atom-type E-State: -OH

[e] Overall or summation solute hydrogen bond basicity

[f] Measure of logP (hydrophobicity)

[g] Variance inflation factor

SVM was performed to seek any improvement on the prediction results by optimizing the internal parameters. The lowest RMSE of Q^2_{LOO} was observed at C=3, ε=0.1 and γ=0.5. Validation parameters were also calculated for the ACO-SVM model (the same molecular descriptors selected in ACO-MLR were used). The results showed higher internal and external accuracy compared to those of the linear ACO-MLR (Table 5.2). More details about the validation criteria can be found in section 5.1.4.

Table 5.2 Internal and external evaluation and related accuracy measurements.

Statistical parameters	Training set (n=1026)		Test set (n=327)		Evaluation set (n=660) (605: after removal of outliers)	
	ACO-MLR	ACO-SVM	ACO-MLR	ACO-SVM	ACO-MLR	ACO-SVM
Q^2_{LOO}	0.600	0.695	--	--	--	--
R^2	0.607	0.920	0.733	0.831	0.447	0.692
RMSE	1.074	0.498	0.900	0.707	1.004	0.716
CCC	0.756	0.953	0.830	0.908	0.655	0.823
r^2m	--	--	0.671	0.823	0.378	0.668
MAE	--	--	0.708	0.546	0.792	0.519
Q^2_{F1}	--	--	0.726	0.831	0.461	0.726
Q^2_{F2}	--	--	0.726	0.831	0.392	0.691
Q^2_{F3}	--	--	0.724	0.830	0.657	0.825
$(R^2 - R_0^2)/R^2 < 0.1$	--	--	0.0097	0.0001	0.0540	0.0018
$(R^2 - R_0'^2)/R^2 < 0.1$	--	--	0.2396	0.0319	0.5926	0.1352
$0.85 \leq k \leq 1.15$	--	--	1.0011	1.0037	0.9502	1.0038
$0.85 \leq k' \leq 1.15$	--	--	0.9673	0.9768	1.0021	0.9706
MAE-based criteria 95% data	--	--	Good	Good	Moderate	Good

The correlation plot between the experimental and the predicted toxicity using ACO-MLR and ACO-SVM are shown in Figure 5.2. Y-randomization was also applied and showed that there is no correlation as to the result of chance and that there is a meaningful correlation between the selected molecular descriptors and their toxicity. The results of Y-randomization after shuffling 20 times are listed in Table A 5.3.

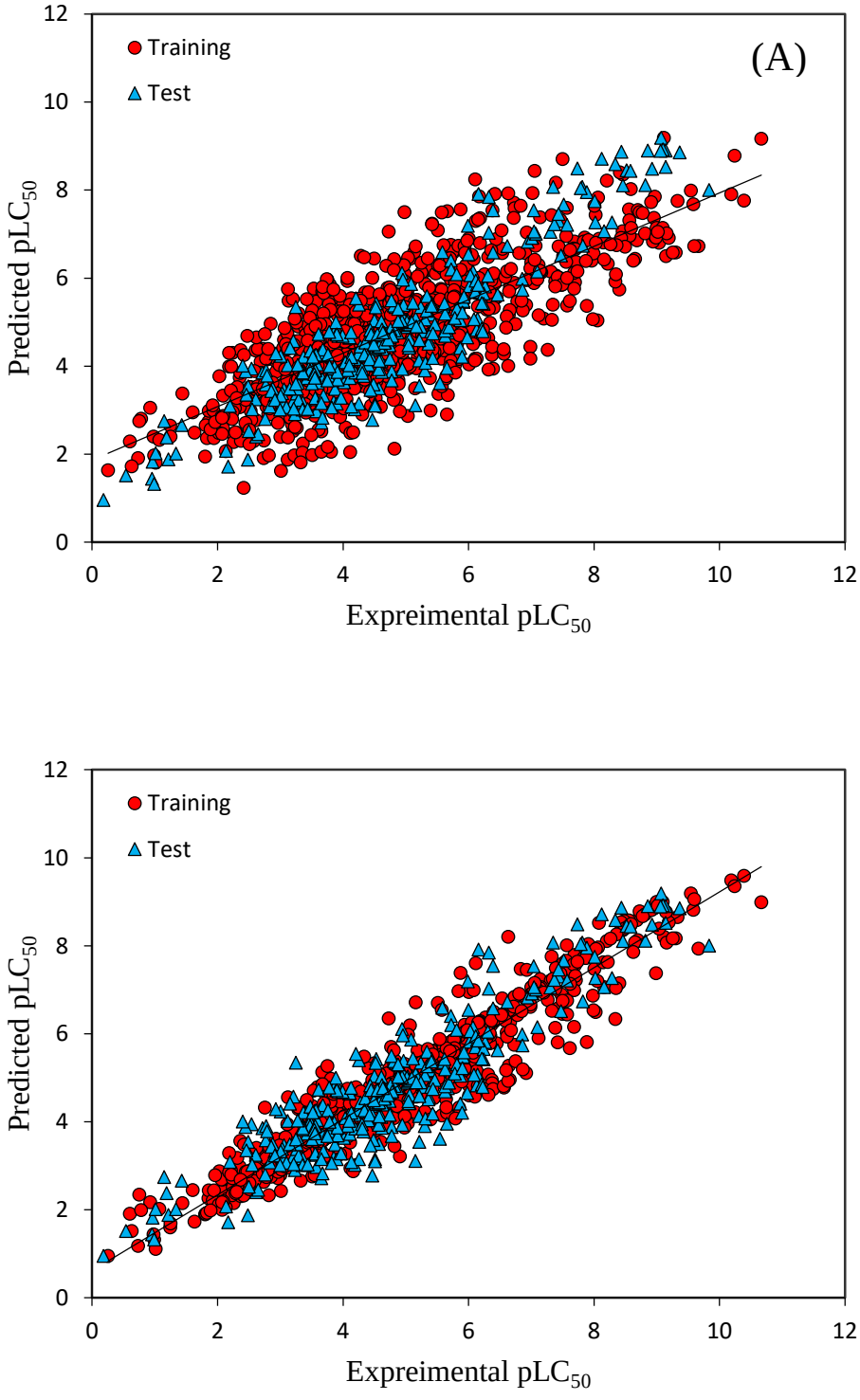
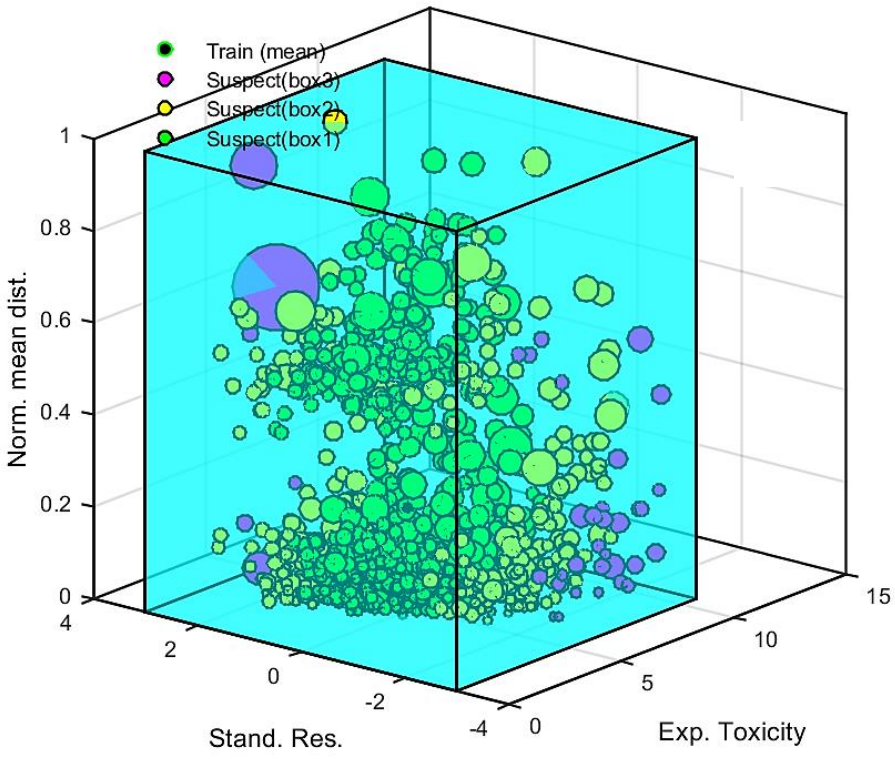


Figure 5.2 Correlation between experimental and predicted Toxicity: A) ACO-MLR and B) ACO-SVM

The proposed models were finally checked for existence of any outliers using the bubble plot of OTrAMS approach which is based on combination of Williams plot and normalized mean distance (Figure 5.3A). As discussed above, the cut-off limit for normalized mean distance is 1.0 as well as the warning leverage value can define whether a new compound is covered by the chemical space of the training set or not. If a compound shows a large bubble, but still falls inside the normalized mean distance, it could be treated as good leverage. Therefore, such a compound with an accepted SR value (in case of known toxicity values) can increase the chemical space edge.

All in all, 993 compounds (73.4%, from both datasets) are predicted with an error less than 1.0 SR (approximately 0.5 fold deviation in logarithmic unit) and another 220 compounds (16.3%) are predicted less than 2.0 SR (approximately up to 1.0 fold deviation in logarithmic unit). Therefore, the model covers 89.7% of compounds with prediction errors of less than 1.0 logarithmic unit (Figure 5.3B).



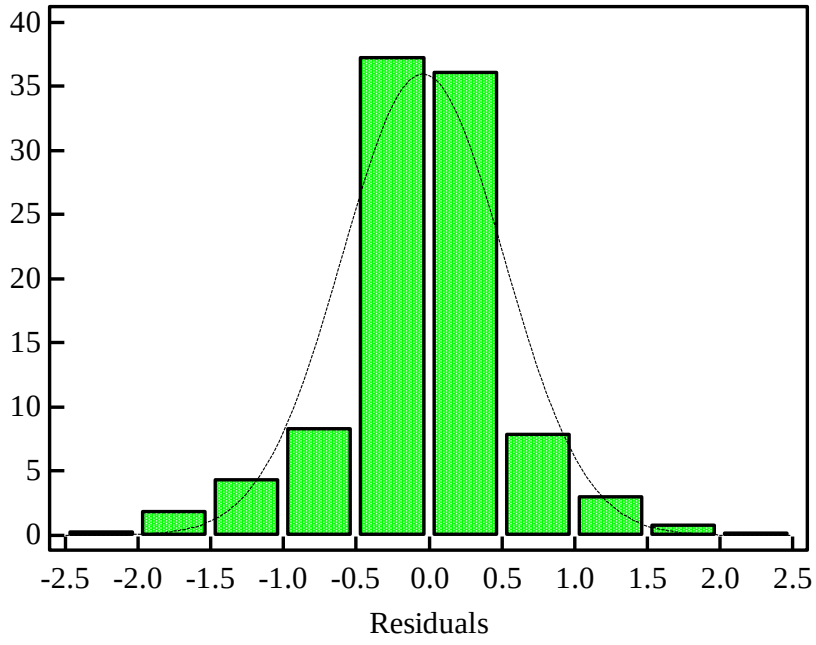


Figure 5.3 A) The OTrAMS can be used to detect compounds with erroneous predictions and to understand the origin of the error as result of either chemical space failure or experimental toxicity error. B) The distribution of errors (by ACO-SVM) in terms of log unit. The plots show all substances of the training and test set together.

5.3.3. Chemical toxicity mechanism in Daphnia magna

The molecular descriptors selected by ACO were further studied to understand the mechanism of toxic action (MOA) toward Daphnia magna. In this regard, Principal Component Analysis (PCA) was used to derive the loading of molecular descriptors on the basis of their contribution in two PCs (PC1 and PC2) with regard to their toxicity level. As it can be seen from Figure 5.4, ALogP, CrippenLogP and XLogP with relative importance of 18.92, 31.53 and 28.36, respectively, were found to be high in compounds that are very toxic (pLC50 in log mol/L >7) or toxic (7 ≥ pLC50 in log mol/L ≥ 5). AlogP represents the octanol-water partitioning coefficient logP, estimated by Ghose–Crippen method.[120] It describes the affinity of a molecule or a moiety toward a lipophilic environment. logP, though, has the meaning of the association of non-polar compounds in an aqueous environment comprising from the tendency of water to exclude non-polar molecules. XlogP is another atom-additive method for calculating logP.[181] It calculates

the logP value for a given compound by summing the contributions from component atoms and correction factors. CrippenLogP is also a measure of hydrophobicity with some new atom type classification system for use in atom-based calculation of logP.[182] These three measures of logP are methodologically independent and together in the model structure not confounded, but making the logP measurement more accurate through a quasi-consensus model (mix of AlogP, CrippenLogP and XLogP), as their VIF and inter-correlation values are less than the 5.0 and 0.9, respectively.

From Figure 5.4, it can be concluded that hydrophobicity is positively correlated with toxicity. This is in accordance with the experimental observation that a compound with high logP has a higher uptake in the cellular membrane.[161] It is believed that this mechanism, known as narcosis or baseline toxicity, will ultimately lead to the death of the test organism, simply by disturbing the function of the cell membrane.

Another descriptor selected by ACO is AATSC0p (relative importance of 8.00) which is average centered Broto-Moreau autocorrelation - lag 0 / weighted by atomic polarizabilities [183]. In this descriptor, the molecule atoms represent a set of discrete points in space, and the atomic property as a function is evaluated at those points. AATSC0p belongs to the 2D autocorrelation molecular descriptors and has previously shown to be important in toxicity prediction.[151] As it can be seen from Figure 5.4, the AATSC0p vector points towards the highly toxic compounds, suggesting that there is positive correlation between the experimental toxicity and polarizability. In particular, AATSC0p accounts for increasing tendency of the charge distribution of a molecule to be distorted from its normal shape and increases its interaction with cellular membranes.

The next descriptor is minsOH (relative importance of 5.53) which stands for Minimum atom-type E-State of –OH in a molecular graph. It belongs to the atom type electrotopological state and gives a more accurate and chemically meaningful expression to the role of functional groups, such –OH, in molecules.[184] As it can be seen from Figure 5.4, increasing the minsOH inversely affect the toxicity values and causes a compound to be less toxic/not-harmful. It can be explained easily as the addition of –OH to the molecular graph increases the polarity of compounds (i.e. make it more hydrophilic) hindering its uptake by cellular membranes.

The last molecular descriptor selected was MLFERBH (relative importance of 7.66) which is the overall or summation of solute - hydrogen bond basicity [162], revealing the solvent-solute interactions in liquid phases. This descriptor can explain the hydrogen bond donor counts or atomic charges and makes the charge density distorted from its normal shape. The loading plot shows that MLFERBH points to the same direction as AATSC0p (Figure 5.4). It might reflect that, as the hydrogen becomes more basic (increasing the hydrogen bond acceptor atom in highly polar atoms), the toxicity increases. It can also indirectly reflect the effect of pH on the experimental toxicity. From Figure 5.4, it is clear that this descriptor explains the toxicity of harmful compounds more systematically than those that are highly toxic and can be well understood via their intermediate logP values.

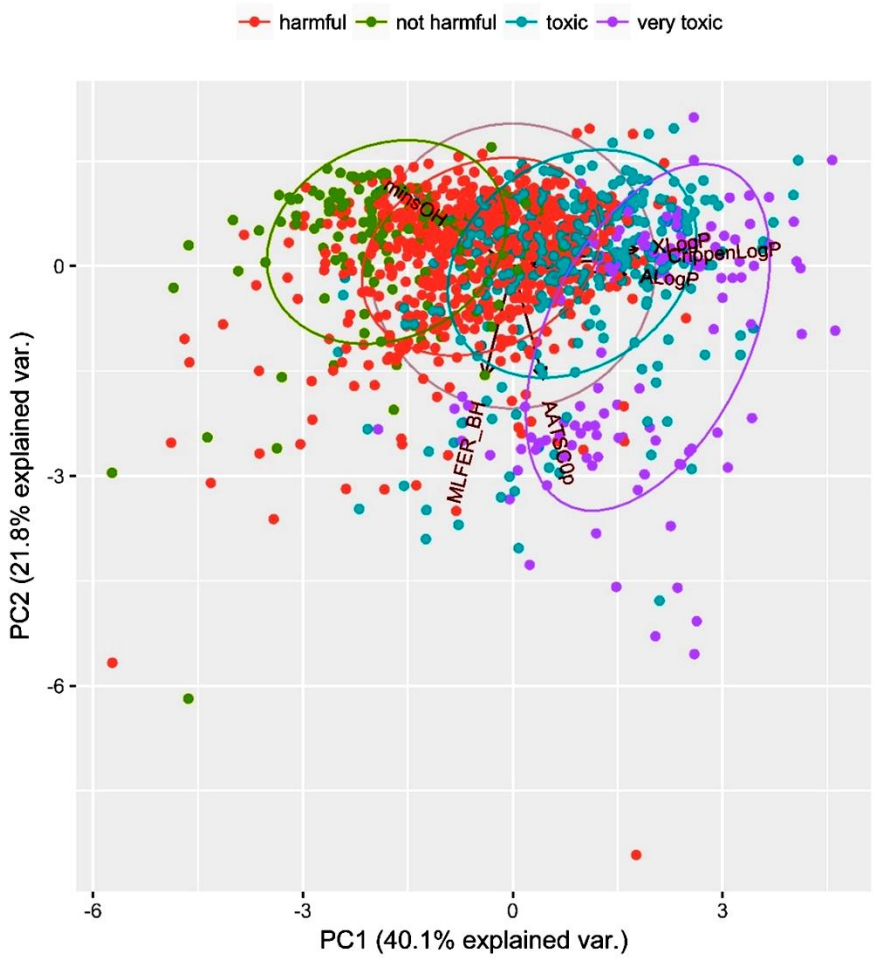


Figure 5.4 Principal Component Analysis with molecular descriptors loadings and toxicity levels

5.3.4. Prediction performance of the ACO-SVM and ACO-MLR models

The ACO-SVM model proposed here shows higher accuracy, as compared to previous prediction models (Table 3). With a R^2 of 0.92, more than 90% of the variance in the training set compounds was explained. Moreover, the application domain covered about 95% of the 1026 training set compounds. Similarly, the prediction power for the test set was still high, with R^2 of 0.83 and 90% of compounds being within the applicability domain. The absolute prediction accuracy was also high compared to other models [156],[155], with RMSE of 0.50 and 0.70 for the training and test set, respectively.

Table 5.3 Internal and external accuracy of current model in contrast to the previously proposed ones.

	No. Compounds	R^2_{Train}	RMSE	Q^2_{LOO}	R^2_{Test}	modelling
Current work	1353	0.920	0.498	0.695	0.831	ACO-SVM
R. Kühne et al. [156]	1365	0.750	0.860	0.750	--	MLR
A. P. Toropova et al. [185]	758	0.739	0.802	0.736	0.838	Monte Carlo-based
M. Cassotti et al. [186]	546	0.780	--	0.780	0.720	GA-kNN
M. Cassotti et al. [160]	436	0.780		0.780	0.730	GA-kNN
M. Moosus and U. Maran [187]	253	0.740	--	0.714	0.634	MLR
R.Vikas [188]	252	0.677	0.886	0.647	--	MLR
A. P. Toropova et al. [189]	297	0.708	1.040	0.697	0.790	Monte Carlo-based
A. P. Toropova et al. [190]	297	0.725	0.889	0.715	0.768	Monte Carlo-based
S. Kar and K. Roy [153]	222	0.695	--	0.678	0.741	PLS
A. R. Katritzky et al. [47]	130	0.759	--	0.728	0.737	MLR

On the contrary, the ACO-MLR was much less accurate, with R^2 of 0.60 and 0.73 for the two main datasets. Similarly, the absolute prediction accuracy in terms of RMSE was also lower, with 1.07 and 0.90, respectively. Overall, this model is still fair as compared to previous models [156],[155], also given the same high applicability domain as for the

training set, probably due to the similar distribution of toxicity and chemical properties among both datasets (**Figure 5.1**).

5.3.5. Additional evaluation set and applicability domain

A set of 660 additional toxicity values (Table A 5.2), i.e. including 220 new compounds, was compiled from various sources to externally evaluate the accuracy of the proposed models and to define their applicability domains via normalized mean distance versus leverage values analysis. This method shows whether the chemical space covered by a model is representative for a compound with an unknown toxicity value, adding certainty in the terms of chemical structure failure.

Out of 660 compounds, 515 compounds were predicted between ±1.0 SR, 90 compounds were predicted within $1.0 < SR \leq 2.0$ or $-2.0 \leq SR < -1.0$ (referring to 1 log unit), while 28 compounds were predicted with SR between $2.0 < SR \leq 3.0$ or $-3.0 \leq SR < -2.0$. 27 compounds were predicted above 3.0 SR. As it can be seen from Figure 5.5, 13 compounds are found to be outside of the chemical space, or application domain, of the proposed ACO-SVM model.

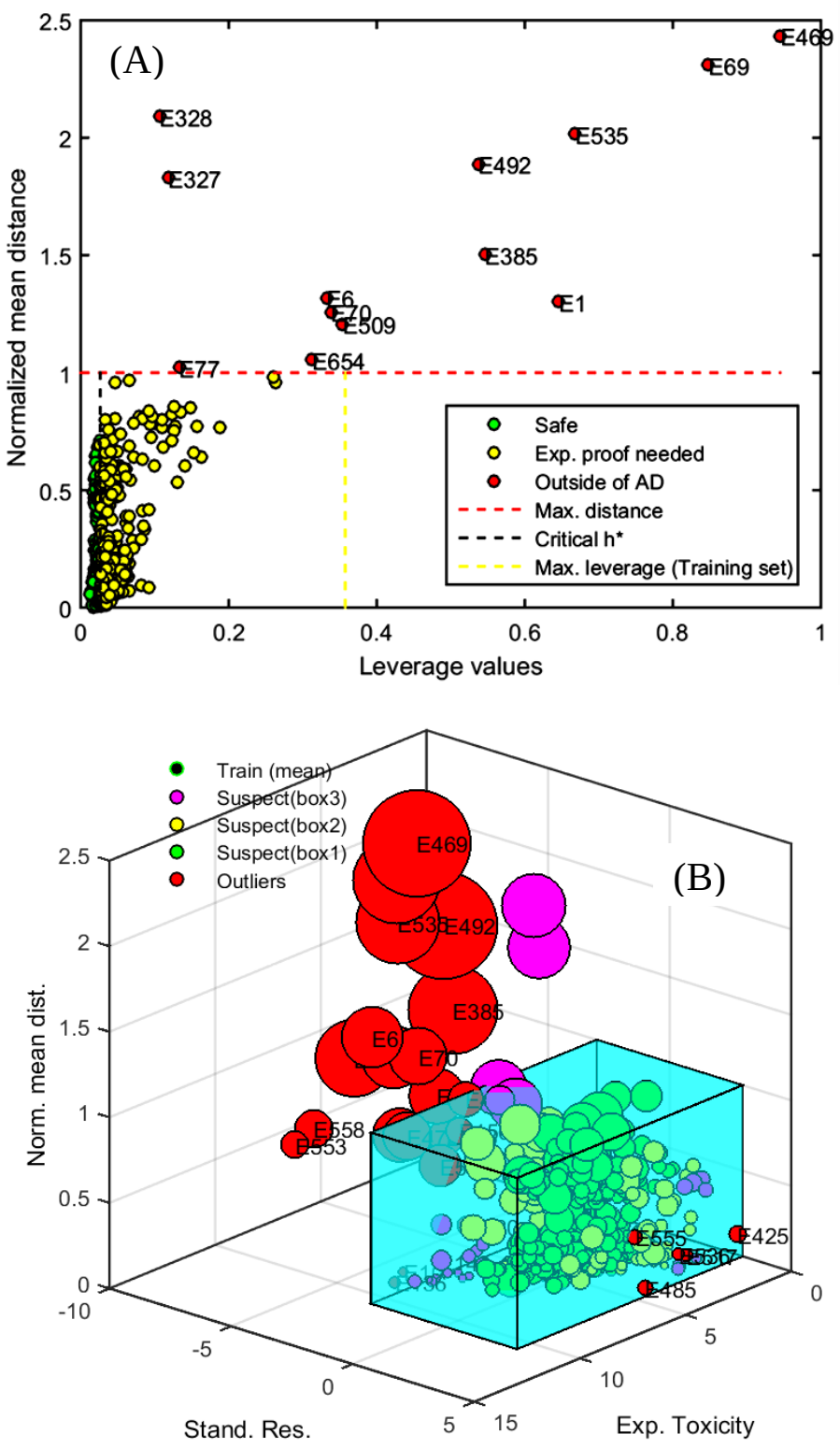


Figure 5.5 Applicability domain study for the evaluation set using A) Normalized mean distance versus leverage values and B) OTrAMS

Moreover, SD approach was used as simple approach to identify any compounds that are chemical structurally speaking outliers. Interestingly, SD identified mostly all the compounds that were found to be outlier via OTrAMS and normalized mean distance versus leverage values introduced here. These results can be found in Table A 5.4. SD based outlier detection however suggested some compounds such as E25, E51, E110 or E514 as structurally outlier while the prediction results were highly accurate. This could be due to the fact that this method is neglecting the relative importance of predictors and their intercorrelations.

The correlation between experimental and predicted toxicity values for the evaluation set is shown in Figure 5.6. As it can be seen, most of compounds are well predicted, suggesting that the proposed ACO-SVM model can cover a large chemical domain.

Therefore, the major source for the observed prediction errors is structural dissimilarity. Besides chemical descriptors as used to build the model, additional features were used to define the chemical domain of the model: For instance, compounds E1 and E2 have tin in their moiety and the structure is symmetric with smaller organic moiety which makes the logP measurement complex. Permanently charged molecules, such as quaternary ammonium compounds, as well as organometallic substances (e.g. organotins) are therefore considered to be outside the applicability domain of the proposed model, The remaining 14 compounds with residuals > 3 SR are prediction errors of the model, which however, can be explained by structural alerts for excess toxicity in Daphnia magna.[169] For these compounds, the model underestimated the toxicity of the compound due to a specific chemical group that made the compound much more toxic than generally assumed by logP.

On the other hand, there were also positive examples of an extended chemical domain for the ACO-SVM model. For example, five homologue structures of anionic surfactants, belonging to the class of LAS were well predicted, with a mean RMSE of about 0.6 SR. This example highlights the potential of the newly proposed model for future applications for compounds with scarce availability of toxicity data.

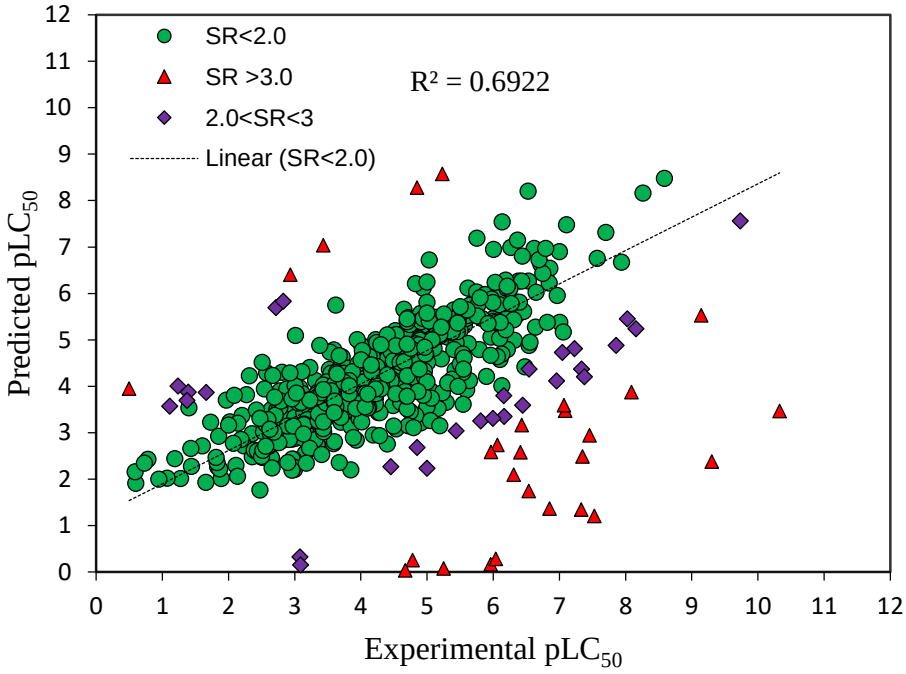


Figure 5.6 Correlation between experimental and predicted pLC_{50} for the evaluation set using ACO-SVM

5.3.6. Identification of potentially erroneous toxicity values

We used a well-defined training and test set to allow for the derivation of a robust model in terms of reliable toxicity values[155]. OTrAMS was used to verify experimental toxicity data while studying the origin of errors. In fact, the OTrAMS tool found no substantial outliers in the training set. Similarly, all the test set compounds fell into the chemical space of the proposed model, somewhat confirming our assumption of a consistent dataset.

In addition to the 13 compounds of the additional evaluation set that were clearly out of the chemical domain, OTrAMS could provide experimental proof to distinguish if a compound has good leverage or is an outlier (Figure 5.5). Compounds further identified with abnormal toxicity are highlighted in purple and can be found in Table A 5.2. For instance, E3 and E4 are marked in purple coded color, showing SR less than 3.0. E3 is very similar to the training set and it should have a higher toxicity due to the phosphorous group present. Therefore, in this case, the reported experimental toxicity might be too low. On the contrary, E4 has cyanide and hydroxide group which may inversely affect

lipophilicity and thus decrease the toxicity. Another example is the carbamate insecticide Pirimicarb (E149) that was recently identified as one of the most harmful compounds for European River basins, whose toxicity was underestimated by the model by more than 3 orders of magnitude. The reason is most likely the very specific binding of the carbamate group to the AChE receptor, which is not reflected in the descriptors used here. Nevertheless, these cases can be identified a priory when applying respective structural alerts to identify compounds with expected high excess toxicity. This example highlights again the need to combine all kinds of information to assess the risk of a compound.

5.4. Conclusions

A large dataset was used to explore the pLC_{50} estimation towards Daphnia magna. Ant colony optimization found to be an effective tool to search and select the most representative features among a pool of 867 molecular descriptors. It was found that the prediction of toxicity improves when using non-linear regression methods such as support vector machine. The accuracy of internal and external datasets suggested that the proposed ACO-SVM model is well established.

The application domain of the ACO-SVM model was also confirmed using the OTrAMS method. 1213 compounds (89.7%) were predicted with less than a one-fold logarithmic error in the main dataset used to build and validate the ACO-SVM toxicity model. Using leverage values with two cut off limits (warning leverage value and maximum leverage that has been observed in the training set) versus normalized mean distance showed to be also helpful to define the chemical space failure or to verify prediction results. In addition to a measure of logP as common molecular descriptor of acute aquatic toxicity, it was found that charge density (as results of polarizability and hydrogen bond acceptor) can affect the toxicity values positively. On the other hand, Minimum atom-type E-State of –OH (minsOH) showed to affect toxicity values negatively, with highly toxic compounds having lower minsOH values.

The proposed ACO-SVM model was also applied to an additional evaluation dataset. The results indicated that more than 515 out of 660 compounds are well predicted and fall inside the one-fold logarithmic error unit. The combined chemical and toxicological

applicability domain of the proposed OTrAMS is considered to be a useful tool to identify the most probable toxicity entry in case several reported experimental toxicity values somewhat disagree by far or to verify the plausibility of an experimental result of a new study. Similarly, OTrAMS might be used to verify the prediction results of other QSAR models. In case OTrAMS confirms that the prediction falls into the expected space of the ACO-SVM model, the prediction would confirm the consensus of the two models.

CHAPTER 6

AutoNonTarget: an R package for automatic suspect and non-target screening

6.1. Introduction

Over the last decades, thousands of substances with potential risks for human and aquatic life are disposed in the environment. Their rapid and accurate identification is emerged as an important “omics” research field. The evolution of high resolution mass spectroscopy coupled with liquid chromatography (LC-HRMS) has become the dominant technique for the large-scale detection, qualitative and quantitative profiling of polar and partially polar compounds in complex environmental samples. Identification procedures in LC-HRMS were detailed into three categories including target analysis (where reference standards are available), suspect screening (existence of suspected substances based on prior information) and finally non-target screening (no prior information is available nor the reference standards) [1]. While LC–HRMS can generate extensive amount of data, manual interpretation of large-scale and complex chemical profile is labor-intensive, time-consuming, and prone to human error.

Chemical databases [191] and mass spectral libraries [192] have been increasingly used for compound identification including various quality control protocols [192, 193] established to increase the success rate of identification procedure especially in case of non-target screening. Generally, the chemical libraries are being searched over an observed accurate mass or molecular formula in suspect and non-target screening as a first step in compound identification. Each chemical data source often include diverse chemical identifiers such as systematic IUPAC name, product names, molecular formula, CAS numbers, structural identifiers (e.g., SMILES, InChI Strings, InChIKeys). Although a molecular formula is sufficient for suspect/non-target screening, correct chemical structure is vital to proceed with *in silico* approaches. A single accurate mass or molecular formula might yield multiple matched near isobaric or isomeric compounds (within the mass accuracy of the instrument), and hence, ambiguous results need further assessments using other orthogonal information such as chromatographic retention time and fragments from tandem mass spectrometry. On the other hand, MS/MS spectral libraries of authentic compounds have also been dramatically increased and the quality

of the mass spectrum data [194] have been improved causing it to compare between various types of HRMS instruments [195]. Currently, the experimental MS/MS spectra of a compound of interest can be searched within available scientific resources in the literature or spectrum library (MoNA (http://mona.fiehnlab.ucdavis.edu/), MassBank (https://massbank.eu/MassBank/), mzCloud (https://www.mzcloud.org/) and METLIN (https://metlin.scripps.edu/)). However, reference MS/MS data for the natural products are often limited in these libraries (GNPS) and manual interpretation of the MS/MS spectra requires great deal of time and knowledge.

Numerous data analysis approaches have been developed to facilitate UHPLC-HRMS based data analysis [7, 80, 81, 196-217] among which XCMS and Mzmine2 are the most widely employed ones. However, several limitations do exist in practical applications. These tools are specific to certain area of data analysis strategy and does not necessary comply with the identification confidence [125]. For instance, XCMS is only applied to derive peaks list, CAMERA can only be used to annotate the peaks list created by XCMS; and next effort is to assign chemical formula considering the isotopic pattern and supportive adducts where no suitable/accurate methods have been developed yet. HRMS data analysis is very complicated and some of the challenges still remained intact. For example, the Gaussian-smoothing-based peak detection (especially for grouping the peaks) strategy faces the difficulty of identifying closely overlapped peaks. In other case, the binning mass values with large steps, such as 0.1 Da or 1 Da, could cause to miss several numbers of co-eluted components while the division of the entire scans with equal steps (such as 0.001 Da or lower) could split ions of a single pseudo spectra into different extract ion chromatogram (EICs) [205].

After peak picking step, masses should be prioritized in the created peaks list. The prioritized masses are called potential mass of interest and concerning the level of identification proposed by Schymanski et al. [125], these masses are at the level of identification confidence 5. Non-target screening starts from this level reaching to probable chemical structure while the list of possible candidates are provided as a suspect list in the suspect screening where the assignment of identification confidence starts from level 3. To proceed from level of identification 5 to 4 for a mass of interest,

evaluation of molecular formula [6] and match between theoretical and experimental isotopic pattern or presence of supportive ions (such as adducts) are required. In case of known unknown and existence of possible candidates for the given molecular formula, the MS/MS spectrum should be interpretable for a retrieved candidate. At this stage, *in silico* fragmentation tools such as MetFrag [7] or CFM-ID [8] could be useful to rank the candidates based on their explained MS/MS fragments (match between *in silico* and experimental MS/MS fragments). In addition, the retention time prediction [9] and other physico-chemical (for instance use of complementary RP versus HILIC elution pattern [10]) data can contribute for further ranking of possible structures and facilitate the identification process [11] especially in the case of isobaric substances [10]. To reach the level of identification confidence 2 (2a and 2b), more supportive data are required. Presence of diagnostic ions or fragmentation pattern between parent compounds and their transformation products are needed to increase the level of identification confidence to 2b. If the experimental MS/MS spectrum is available in the literature or spectrum library and it is matching to the observed one, then the level of identification confidence can be reached to 2a. The comparison of experimental RTI values can be used to promote the identification confidence from level 3 to level 2 [148].

Take environmental analysis, for example, the number of chemicals detected in the surface water is overwhelmingly increasing [78]. All the steps noted in the previous paragraph need to be automatized to help the water quality monitoring agencies for comprehensive chemical analysis in sewage/wastewater-treatment systems. In this section, a complete suspect/non-targeted data analysis strategy (AutoNonTarget) is presented for UHPLC–QTOF-MS-based chemical analysis of influent wastewater and sewage sludge samples. In this strategy, the peak detection, annotation, candidate's retrieval, molecular formula assignment, *in silico* prediction tools (retention time and MS/MS interpretation) as well as look up method for mass spectral library search are performed. Special emphasis is given for the identification of biocides, to illustrate the application, in the receiving environmental of Athens, Greece [34].

6.2. Methodology

6.2.1. Samples collection and instrumentation

Eight influent (IWW) and 8 effluent (EWW) wastewater samples (8 consecutive days in March 2017 from the wastewater treatment plant (WWTP) of Athens, Greece) were analyzed, according to ref. [37], to study the possible detection of biocides. In addition, 64 sewage sludge and IWW samples from the same WWTP (sampled again on 8 consecutive days in March, period 2010-2017) were also screened. The sample preparation method used for preparing the sewage sludge and influent/effluent wastewater samples were given in our previous studies [37, 218]. The instrumentation and LC conditions are as same as section 3.2.1, Chapter 3, of this thesis.

6.2.2. Overview of steps involved in the automatic suspect/non-target screening

As said in section 6.1, there are several steps needed to be performed to fully analyze the chemical profile of a sample. Figure 6.1 provides the overview of these methods together with corresponding level of identification for each step. Let's discuss these steps briefly. The raw data of the samples were exported to open source data format and a clean peaks-list was extracted using the XCMS R package (here CentWave algorithm was used) [3]. Following the componentization and annotation of the peak lists by CAMERA and nontarget R packages [4, 5], a list of masses of interest was created (deconvolution of the most probable precursor ions). Afterwards, the full scan MS chromatogram of the procedural blank samples were subtracted from the full scan MS chromatograms of the treated samples. This was done in R environment using a sophisticated machine learning approach. The suspect and non-target screening strategy was then used to identify these masses of interest. This was done automatically using our in-house R package [6]. In case of suspect screening, candidates match to the given mass from a prepared suspects list after applying the mass accuracy and isotopic pattern filter, thus there is a prior knowledge about the probable candidate and chemical structure. However, in the non-target screening, candidates are yet to be proposed for a given mass, and the identification procedure starts after finding set of mass of interest.

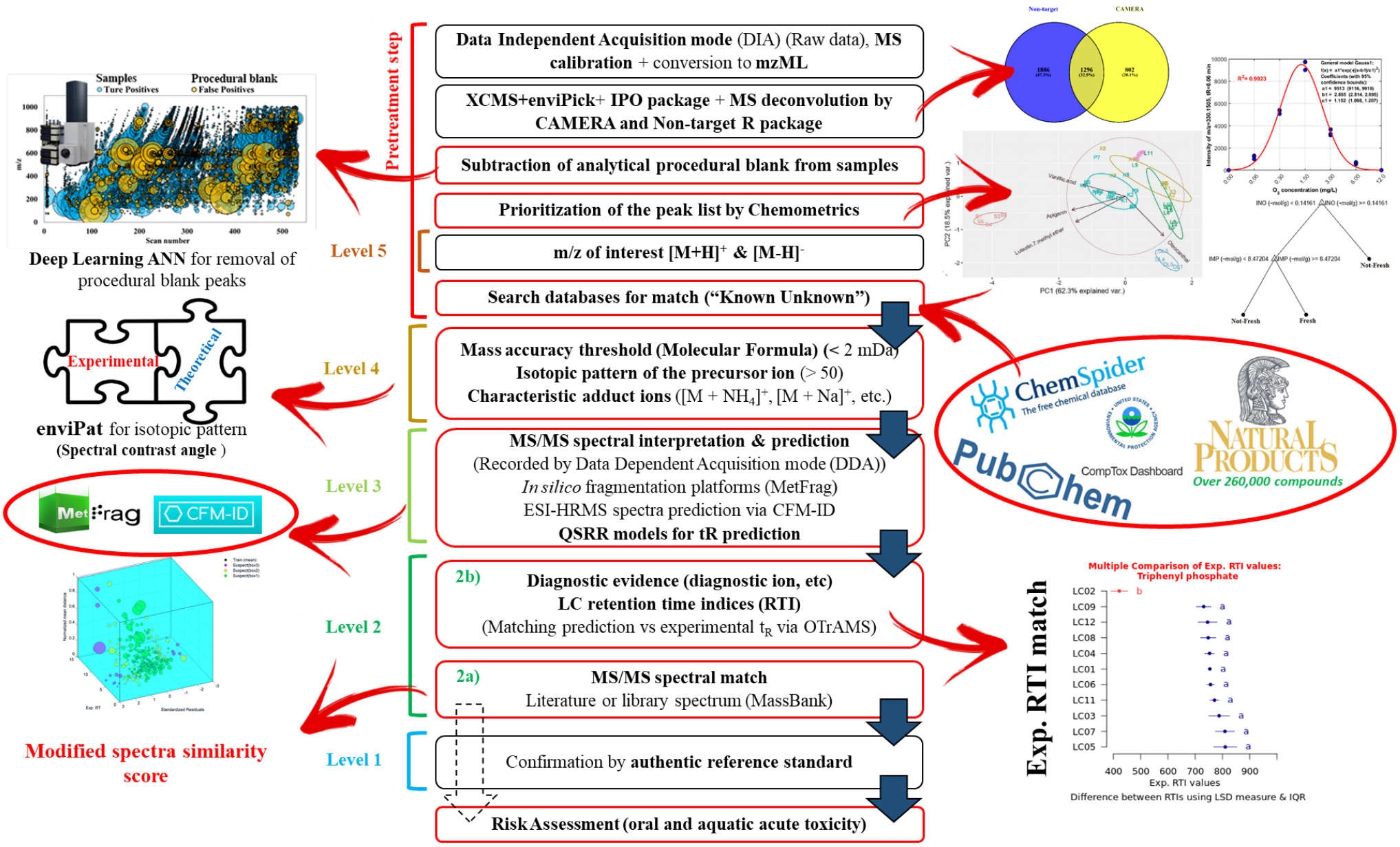


Figure 6.1 AutoNonTarget workflow

6.2.3. Peak picking for HRMS data

Through non-targeted LC-HRMS analysis, a peak list of thousand masses (centroid data) can be extracted for the analyzed samples which includes the intensity of each m/z in certain scan number (t_R). This peak list is the result of any peak picking algorithm that developed specifically for HRMS raw data [197, 199, 205, 219]. Here, XCMS has been used to extract the peaks list from the influent/effluent wastewater and sewage sludge samples analyzed by UHPLC-QToF-MS in both ESI mode. XCMS is proved to have high performance due to its robust and sensitive detection of potential region-of-interesting mass traces (ROIs) and high efficiency of centWave algorithm [220]. Below a general workflow to generate final peaks list from LC-HRMS raw data is discussed.

Prior to peak picking procedure, the raw data of the samples were exported to open source file format such as mzXML. Although, most of the commercial software developed for HRMS instruments supports data export, ProteoWizard, an open source software, used to export the raw data [221]. XCMS accepts any of file format of mzXML/mzML and as a first and most used peak picking algorithm requires an optimization step to adjust peak picking internal parameters. IPO R package [222] optimizes these parameters by using natural, stable ^{13}C isotopic peaks to calculate a peak picking score. Generally, the optimization task is better to be done with pooled samples instead of all analyzed ones. Retention time correction (the common “obiwarp/loess” method) [223] is optimized by minimizing relative retention time differences within each peak group. Grouping parameters (the XCMS method “density”) are optimized by increasing the number of peak groups that show one peak from each injection of a pooled sample. In IPO R package, the optimization task is achieved by Box-Behnken designs of experiment [224]. XCMS has several important parameters such as ppm (the tolerated mass deviation), minimum and maximum chromatographic peak width, and “snthresh” ratio (the chromatographic signal-to-noise threshold). Preferably, prefilter (a threshold of which an m/z to be considered as a true peak if it appears in k consecutive scan at J intensity threshold (k,J)) can be applied to exclude any false peaks in selected ROIs. Several additional steps such as retention time correction and alignment as well as peaks grouping across samples would be needed to merge peaks list from each sample. Filling any missing peaks and

also annotation of the detected m/z features are highly encouraged to avoid any adducts/isotopic peaks to be cofounded with their molecular ions in post-processing step. Here, the annotations of peaks list was done by complementary use of “CAMERA” [80] and “Non-target” R package [81]. The only disadvantage of the general workflow described above is that it is time consuming owning to optimization step. Figure 6.2 illustrates the peaks picking procedure applied to HRMS raw data.

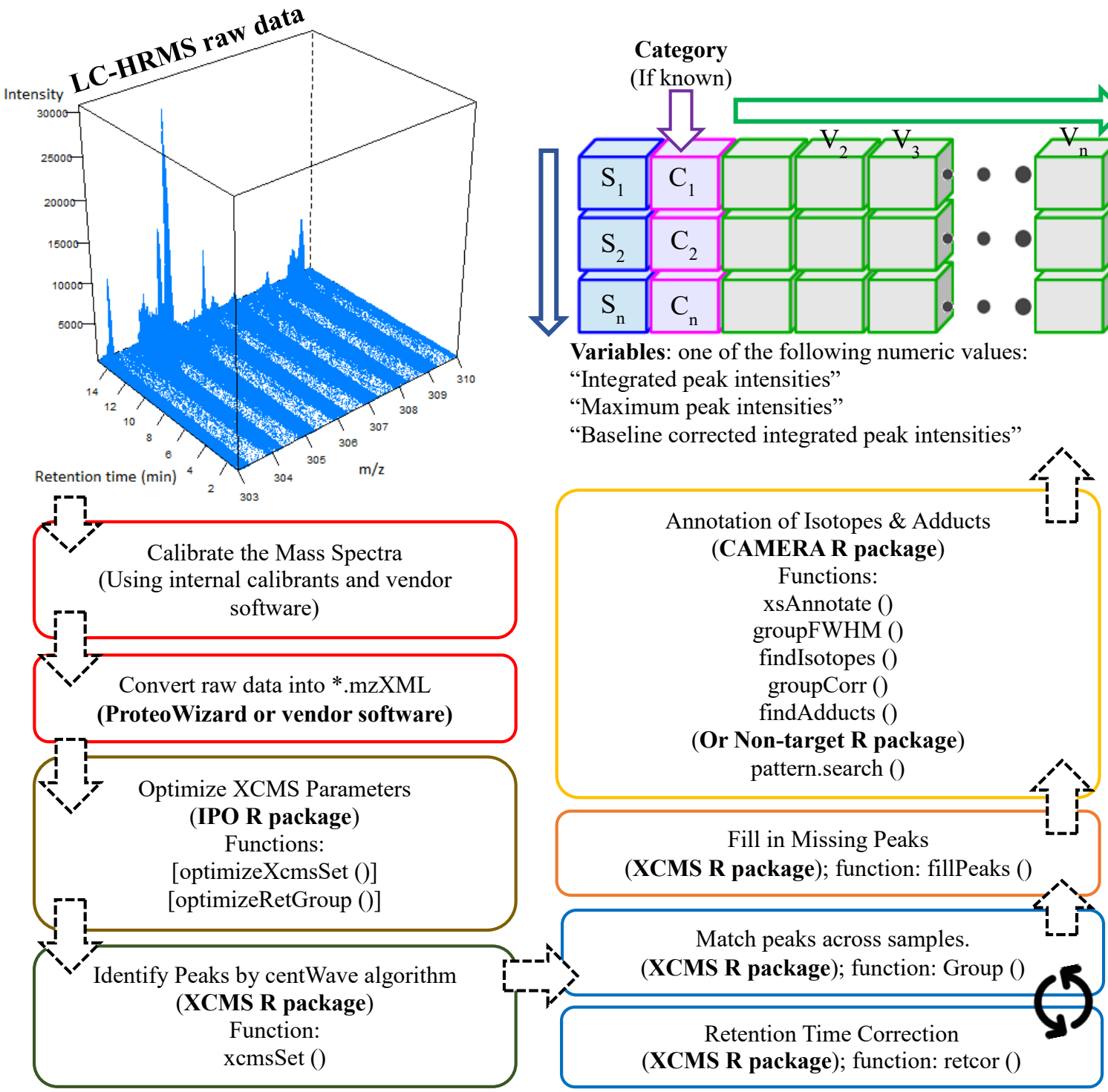


Figure 6.2 The peak picking workflow used for LC-HRMS data processing

6.2.4. Subtraction of analytical procedural blank from samples

Most of untargeted data processing tools produce peaks list from complex samples of which may contain erroneous features, such as duplicate or isotopic peaks and peaks originated from analytical procedural blank or contamination in the ion source of MS instrument. Manually curating an untargeted dataset involves removing duplicate features and analytical procedural blank peaks, isotopic features or combining multiple ion adducts belonging to a same molecule is a time-consuming and error-prone task [225]. There is a need for an automated method of identifying unsolicited features in the final peaks list before proceeding to advanced statistical analysis and identification of unknowns. Here we have used a novel machine learning approach and included in the "AutoNonTarget" workflow in order to derive probability values of a peak to be false or true positive considering the mass accuracy, retention time tolerance, intensity variation (fold changes), dilution factor (between samples and blank) and similarity across chromatographic peak shapes [226]. The deep learning artificial neural network (Deep Learner ANN) was behind the machine learning approach and showed the most accurate detection of false positives in contrast to simple flagging approach [227]. The setup to perform a deep leaner ANN for subtraction of analytical procedural blank is briefly discussed below. A modified version of an R package "mxnet" was used to build the convolutional neural network capable of deconvoluting the information an input extracted ion chromatograms (EIC). The input signal passes at first from two serially connected sets of layers. Each layer set contains one convolutional layer, followed by a layer that locates non linearities by the "tanh" activation function and a final layer that pools the maximum areas of the previous layer. The pooling layer output of the first set of layers described previously is the input to the second set of layers, which again consists of a convolutional layer, a nonlinearity detector layer and a max pooling layer.

The first convolutional layer consists of 20 filters of size 8x8 and stride 1x1 and the second one of 50 filters of size 8x8 and again stride 1x1. Both pooling layers use a filter of size 3x3 and a stride of 3x3. The padding parameter is by default computed automatically inside the "mxnet" functions if the filter size and the stride are defined by the user. After the above sets of input layers, two fully connected sets of layers follow. The first set of

fully connected layers at first contains a layer that flattens the output of the previous layers, meaning that it transforms the input array into a 2D array by collapsing higher dimensions. A deep neural network with 600 hidden nodes follows. The deep network tries to locate areas of the 2D array that behave similarly under affine transformations. Affine transformations preserve point's straight lines and planes. Objects in the 2D input array that remain similar under a combination of reflection, rotation, scaling and translation transformations, are detected by this deep network. The first set of connected layers ends with a final layer that has no hidden layers and simply applies the "tanh" activation function for further detection of non-linearity. The second set of fully connected layers is a deep network with 3 hidden nodes and again uses the "tanh" activation function in every node. Finally, the output of the whole system is a network that applies the "softmax" activation function. The "softmax" function or normalized exponential function is the most suitable activation function for classification problems, because its output can be interpreted as a probability distribution over all possible classes. After the above steps, the network tries to find the optimum values for all the network parameters, intensities of all the filters and weights of the fully connected layers, in order to minimize the classification error by using "backpropagation" for calculating the effect of each parameter on the output error and stochastic gradient descent for finding the global minimum of the classification error. A training set of 1200 highly diversified EICs was used which consists of all the scenario where a peak could face (noisy background, clear true/false positive, co-eluted peaks, equal/higher instrumental response in analytical procedural blank than samples and vice versa) while the test set consisted of 7221 EICs was used to evaluate the performance. For the total accuracy comparison, the area under the ROC curve (AUC) was used as estimator of the false and true positive peaks discrimination [228]. Figure 6.3 illustrates the overall procedure followed to remove the peaks originated from analytical procedural blank.

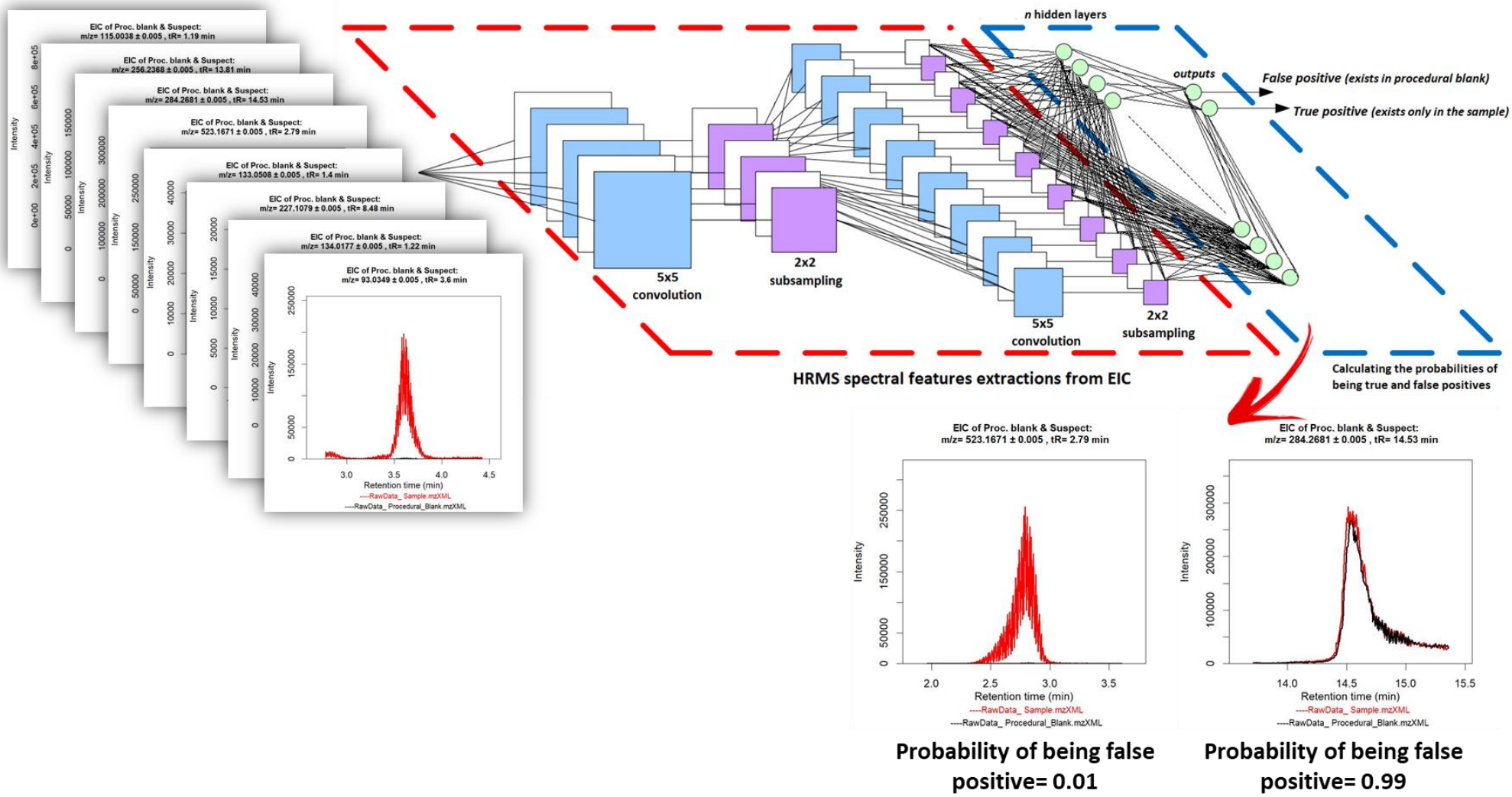


Figure 6.3 A schematic workflow of distinguishing between the false and true positive peaks using the deep learner algorithm

6.2.5. Prioritization of MS peaks-list

6.2.5.1. Peaks list annotation

Peaks in the same scan/EIC could present different high-resolution m/z values. Therefore, their annotation should be performed to cluster peaks that come from the same compound. An annotation of these ion species reduces the number of features yet to screened or used in the subsequent analysis. From two annotated ions, the molecular mass can be calculated or searched in the chemical databases and afterwards, it can be used to calculate elemental composition of the neutral compound. Here, “CAMERA” [80] and “nontarget” [81] R packages used to create annotation table of the peaks list. These packages use a dynamic rule set created from the combination of lists of observable ions. Each rule shows a specific ion species with the mass difference to the related molecular mass and ion charge. All m/z-differences within a compound spectrum are matched against these dynamic rules set. Matches with the same molecular mass are combined into hypothesis groups. In ESI, uncharged compounds are ionized via adduct formation with cations or anions or abstraction of protons. Moreover, neutral losses may happen to form the fragment ions. The main Isotopic ions, adduct ions, and isotopic adduct ions used in these packages are $[M+H]^+$, $[M+1+H]^+$, $[M+2+H]^+$, $[M+3+H]^+$, [M+Na]+, $[M+K]^+$, $[M+2Na-H]^+$, $[M+2K-H]^+$, $[M+NH3]^+$, $[M-H2O+H]^+$, and $[M-2H2O+H]^+$ (for peak annotation in +ESI). The ions used in rule set for–ESI are $[M-H]^-$, $[M-H+NaCOOH]^-$, $[M-2H+Na]^-$, $[2M-2H+Na]^-$, $[M -H+HCOOH]^-$, $[2M-H]^-$, $[2M-2H+K]^-$, $[M-2H+K]^-$, $[M-2H]^{2-}$. This helps to prioritize the peaks list to focus only the molecular ions of components, i.e., $[M+H]^+$ or $[M-H]^-$ for subsequent data analysis. CAMERA also get use of the chromatographic peak shape similarity. It utilizes the HRMS raw data to obtain the extracted ion chromatograms (EIC) for each feature and calculates a pointwise pearson correlation of the intensities between the chromatographic peak boundaries for all pairs of features in a compound spectrum. Second, we include the pearson correlation of intensities across all samples for each pair of features in a compound spectrum.

6.2.5.2. Time trend analysis

In HRMS, a mass could be interesting by observing its intensity trend across samples in a particular sampling period or even its high abundance [229]. However, this does not mean that all these m/zs can explain the classes of the samples. Trend analysis is helpful to trace the compounds that are accumulated in the receiving environment throughout the sampling period and have low biodegradation. Increasing trend can also happened in case of formation of TPs from a parent compound which reveal the quick detection of these TPs in the sample [227]. An automatic approach based on Gaussian curve fitting was developed to explore potential masses with interesting trends among thousands of peaks, extracted from XCMS. The purpose of this method was to fit Gaussian curve to peaks list from XCMS according to the samples label (it could be ozonation dose, sampling period, applied treatment process or sampling points (upstream/downstream) with a dynamic algorithm to evaluate the quality of the fitting based on the squared correlation coefficient for all the m/z in the peaks list. The Gaussian curve fits to peak shape and therefore, can be used to study m/zs that have been formed and then removed from one level (like ozonation dose or different days of sampling) to another. Furthermore, half Gaussian curve would also describe those m/zs that have increasing trend in their intensities as for instance ozonation dose increases. Generally, a Gaussian curve can be fitted to the set of data points as follows:

$$f(x) = \sum_{i=1}^{n} a_i e^{\left[-\left(\frac{x-b_i}{c_i}\right)^2\right]} \qquad (6.1)$$

where *a* is the amplitude, *b* is the centroid (location), *c* is related to the peak width, x is the samples label. We have validated this newly developed tool for the detection of TPs formed after ozonation of citalopram [230] and included in the “AutoNonTarget” workflow. *n* is the number of Gaussian peaks in the intensity vs level curve (it could be 1 and 2). When *n* is 1, the m/zs that formed and removed from one level to another, increased or decreased over the levels that can be detected. When *n* is 2, m/zs that are being formed and removed repetitively (formed-removed-formed or vice versa) can be detected. Figure 6.4 illustrates the Gaussian based trend analysis for detecting a TP after various

ozonation dose (level of ozonation dose (i.e. 0 (0 mg/L), 1 (0.06 mg/L), 2 (0.3 mg/L), 3 (1.50 mg/L), 4 (3.00 mg/L), 5 (6.00 mg/L), 6 (12.0 mg/L)).

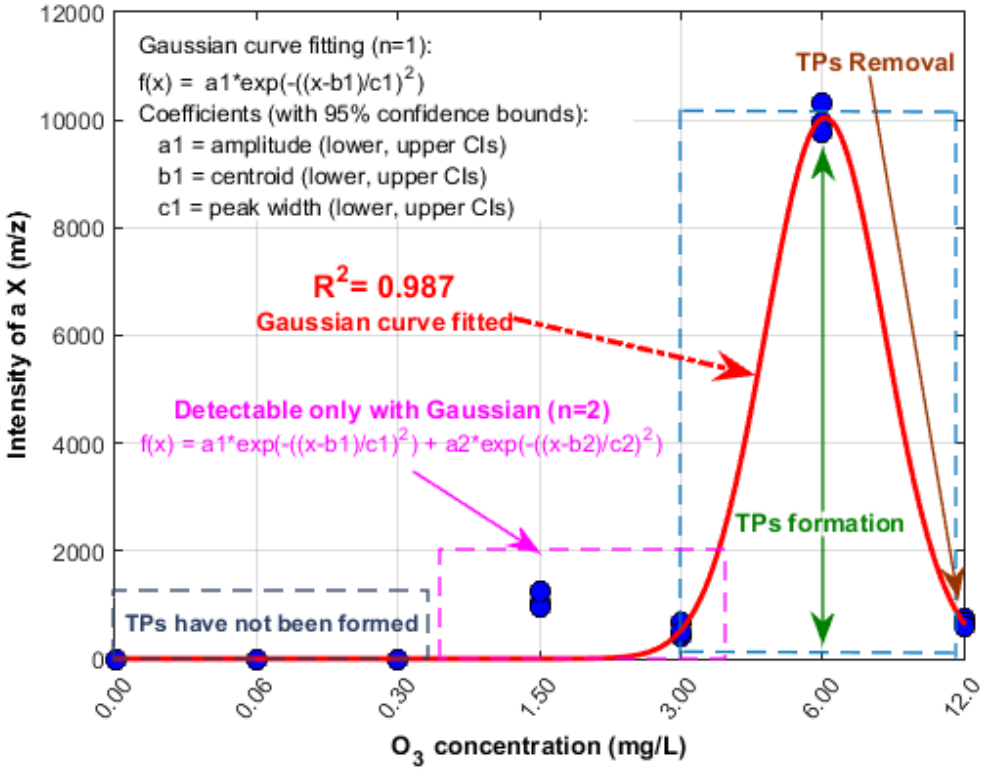


Figure 6.4 Gaussian curves for detection of m/zs (here it is TPs)

6.2.5.3. Chemometrics

Apart from focusing on the potential molecular ions, one can focus on the subset of these m/zs that can explain characteristic of the samples. This is important step when a non-target screening approach is used [37] due to great deal of efforts needed for their identification. This is also common when the question is to classify set of samples into their related groups according to their chemical profile. In most cases, features selection remains as the best solution to avoid overfitting in development of any classification models. It simplifies the model structures and can limit the identification efforts on the peaks that are meaningful to the classification problem. Therefore, sophisticated methods are needed to select relevant m/zs contributing to the classification problems such as simple analysis of variance (ANOVA) [204], volcano plot [231, 232], variables importance

in loading information from a PLS-DA [233], orthogonal-PLS (S-plot) [234] or decision tree [18] and nature inspired features selection algorithms [4]. Currently, “AutoNonTarget” includes PCA (unsupervised techniques for data exploratory analysis), ANOVA, volcano plot, PLS-DA, O-PLSDA, Random Forest and LDA (linear discriminant analysis) [18].

6.2.5.3.1. Group comparison with statistical analysis

A simple one-way analysis of variance (ANOVA) or posthoc multiple comparison test can be used in order to determine those (m/z)s that are significant in differentiating samples from different classes. In general, in a binary classification problem or two-group experimental setup, the fold changes (variation in the intensity of (m/z)s) can be evaluated statistically using the Welch-t test (to obtain the class-regulated data for each m/z) or *p*-value (to filter in the (m/z)s that their intensity changes are significant between two groups of samples) [202, 204]. ANOVA is extensively used to evaluate the significance of (m/z)s in HRMS data [235-239].

6.2.5.3.2. Volcano plot

An alternate visualization is a volcano plot [231]. Volcano plot is sometimes used for visualization of statistical results of omics data such as differential expression of genes measured through microarrays. The volcano plot has the power to show which m/z shows a stronger combination of fold change and statistical significance. They represent significance from a statistical test (such as a p-value) on the y-axis (all p values are transformed with log10 and a limit of +2 (-log10(p-value)) and fold-change on the x-axis (mainly with log2 transformation to apply the limits of ±1 log2(fold change)). As a consequence, m/zs that have a relatively low fold-change between the two samples appear near the center in the volcano plot whereas the m/zs that have significant p-values and fold changes are found in the upper-right or upper-left. Common m/zs between two groups are also located between fold and above p-value threshold. Figure 6.5 exemplifies the different area of volcano plot that can be used to focus certain m/zs in the peaks list.

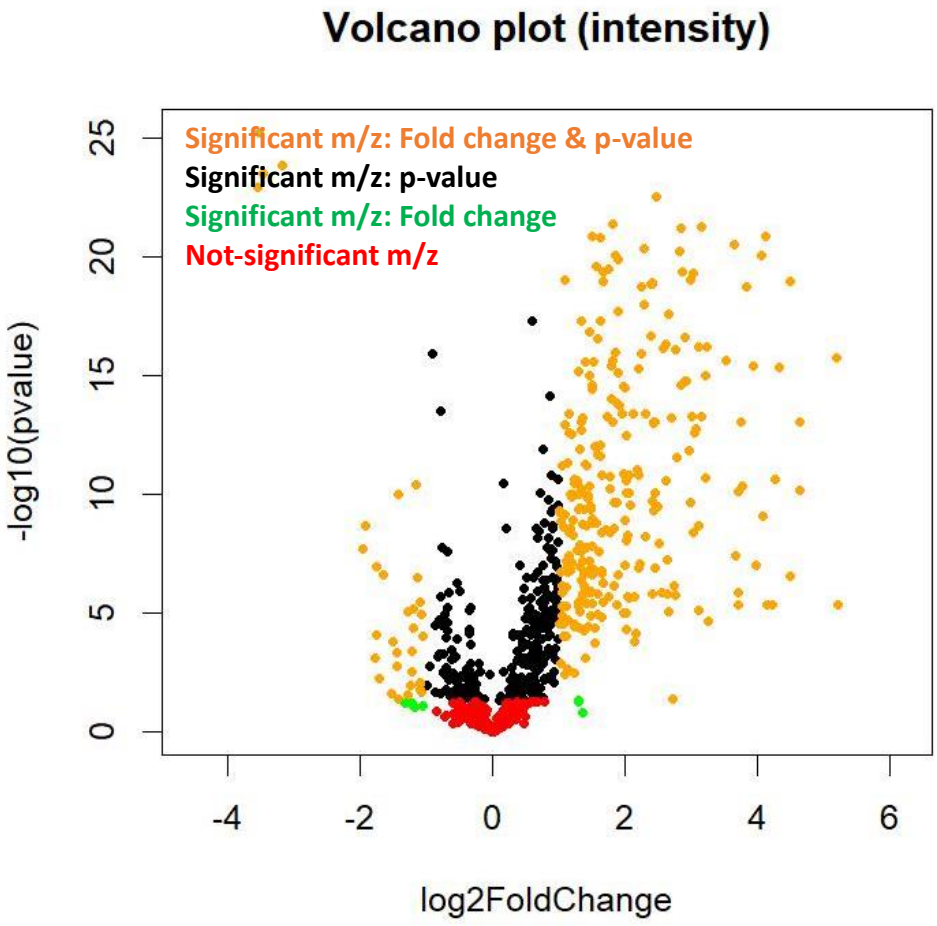


Figure 6.5 Volcano plot and prioritization of m/zs in the peaks list

6.2.5.3.3. Variable importance in projections in PLS-DA

The use of Principal Component Analysis (PCA) as a first-pass method to identify chemical profile differences derived from mass spectral data between samples is remarkably common practice in chemistry. Generally, data showing a good distribution in PCA score plot towards their classes, subsequently, it results in robust supervised methods development. PLS-DA is a linear classification method that combines the properties of partial least squares regression with the discrimination power of a

classification method [20, 240]. In fact, PLS has been modified for classification application and usually provides similar results with LDA. However, PLS-DA offers variable selection advantage inherited in PLS method. In PLS-DA, the classification model is generated by searching for Latent Variables (LVs) with a maximum covariance with the given classes. LVs are the relevant sources of data variability which are linear combinations of the original m/zs (GC/LC-HRMS peaks list). Therefore, PLS-DA provides a graphical visualization and classification model of the data and explains the patterns and relations between classes and samples by LV score and loading plot [20]. Loadings are the coefficients of m/zs in the linear combinations which can be interpreted as the influence of each m/z on each LV, while scores show the coordinates of the samples in the LV projection hyperspace [240]. The optimal number of LVs is usually selected by means of cross validation procedure and the attributed misclassification error. After building a PLS-DA model, the predicted classes will be returned according the total number of classes (i.e. for ***N*** number of classes, ***N*** number of vectors will be created including the prediction results varying between 0 to 1). Therefore, for each sample, with the prediction values in-between 0 and 1: a n-th value closer to zero denotes that the sample does not belong to the i-th class, while a value closer to one the opposite. Bayes theorem can be used to create a classification rule to correctly derive a threshold (the class threshold is fixed at value which the number of false positives and false negatives is minimized) and to assign a class for given sample [240].

Having known the variance explained in the projected dimension (especially in the latent variables from a PLS-DA model) [241], hundred over thousands of (m/z)s in a single peaks list created from LC-HRMS can be prioritized. This can be done using a well-known method so called Variable Importance in PLS-DA Projections (VIP) [241, 242]. The idea behind this measure is to accumulate the importance of each m/z being reflected by loading weights (***w***) from each component in PLS-DA structure. Generally, an m/z is significantly important if its VIP is above 1 and this threshold can be used to select most relevant (m/z)s for classification and subsequent identification task. VIP measure is already applied in the HRMS data analysis [23, 233].

6.2.5.3.4. S-plot for variable significant test in OPLS-DA

The recent modification of PLS-DA is the OPLS-DA [21] which is inspired from OPLS algorithm [243]. OPLS-DA method also known as orthogonal projection to latent variable is another useful tool for deriving the variables importance towards the classes of which the samples belong. OPLS-DA is a supervised method that pairs a peaks list with a corresponding matrix Y (contains the class information). OPLS-DA theory is as same as PLS-DA, but the only difference is that it integrates an orthogonal signal correction filter [244] to minimize the effect of variations that are orthogonal to the prediction results (uncorrelated variables (m/z)s) before constructing the final LVs. Therefore, only the Y-predictive variation is used to model the data. The main advantages of OPLS-DA over PLS-DA are better class discrimination and more robust identification of important features. These significance values (or m/zs weights in the model coefficient) can be extracted from the loading matrix of the model. Additionally, the loadings from an OPLS-DA model can be shown by means of an “S-plot” in which the modeled covariance p[1] is plotted on the x-axis and the correlation profile p(corr)[1] is plotted on the y-axis. m/zs values with higher p[1] values in both positive and negative directions have a larger impact on the variance between the groups, while peaks with higher p(corr)[1] values have more reliability. Therefore, data points that fall in the upper right and lower left quadrants have a high impact on the model and represent possible class-specific biomarkers. Welch-t test can also be applied to the s-plot to obtain the class-regulated data for each m/z quickly. The OPLS-DA method normally is applied when there are only two classes comprising Y.

6.2.5.3.5. Variable prioritization in decision tree

Tree-based approaches [22] consist of algorithms based on rule induction that is partitioning the dataset space into several class subspaces. Basically, the data set is recursively split into smaller subsets where each subset contains samples belonging to as few classes as possible. In each split (node), the partitioning is done in a way to reduce the entropy of new subsets. This continues until a final classification model is built which consists of a collection of nodes (tree) that describes the classification rule for given dataset. The best partitioning solution can be obtained by univariate strategies in which

the algorithm searches the single variable (m/z in HRMS) that gives the purest subsets (lower classification error) at each binary split; all the samples that satisfy the rule are grouped in one subset, otherwise into another.

In decision tree, variables can be selected based on the error (misclassification rate in out of bag samples) attributed to them. In other words, a variable that introduces lower error in each node of the tree will be selected [245]. This variable importance measure can be formulated as:

$$VI(X^f) = \frac{1}{ntree}\sum_{t}\left(\widetilde{OOB}_t^f error - OOB_t error\right) \quad 6.2$$

where for each tree (*t*) of a forest, OOB_t is the samples which are not included in the bootstrap samples to construct *t*. OOB_t*error* is the mean square error (MSE) of a single tree on OOB_t. $\widetilde{OOB}_t^f error$ is the error of perturbed sample created by randomly permuting the values of X^f (variable) in OOB_t. Therefore, RF internally acts as feature selection and selects features that explain low OOB error. The importance score for the *j*-th variable (m/z) is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences. Variables then can be ranked based on their importance score (important variables are those with large value for this score). Sometimes, to promote a decision tree approach to explore other subspace of variables, a nature-inspired metaheuristic algorithm might be needed. The measure of importance score has been used in the field of foodomics by HRMS [18].

6.2.6. Regulatory, suspect and public Databases

When the peaks list is prioritized, the next step is to find candidates for these peaks from online publicly available chemical databases such as ChemSpider (http://www.chemspider.com/), FooDB (http://www.foodb.ca/), comptox EPA dashboard (https://comptox.epa.gov/dashboard), REACH-ECHA (https://echa.europa.eu/), Norman SusDat (https://www.norman-network.com/?q=node/236), NIST (https://www.nist.gov/), HMDB (human metabolome database (http://www.hmdb.ca/)), PubChem (https://pubchem.ncbi.nlm.nih.gov/) or pre-compiled suspects list. Thereby, information

on the parent compound (e.g. molecular formula, substructures) can help to restrict the databases search and possible structures are likely to be proposed for the peak. However, databases contain mostly only Eps, but many TPs are not included yet. Even after filtering and strict criteria and thresholds in the above parameters, the number of peaks, which correspond to non-targets can exceed the number of 1,000. It is clear that elucidation of all those peaks would demand a great amount of time and effort; prioritization of the most intense peaks is a common strategy. Here to evaluate the capability of "AutoNonTarget", a complete list of biocides and pesticides (active ingredients) was compiled from regulatory databases [79, 246-249]. Biocides are class of active substances used for destroying any harmful organisms, but are meant to be harmless to human beings. The high concentration levels of biocides in the environment can cause some severe adverse effects to many life forms, and to some extent even human beings [250]. Therefore, they should be monitored carefully and it is vital to understand their removal efficiency during WWTP. For the suspect screening purpose, a complete list of biocides and pesticides (active ingredients) was compiled from regulatory databases [79] such as European Chemical Agency (ECHA); European commission health & food safety directorate-general, SANCO/2012/11284 –rev. 20, (EU) Pesticides and Biocides; and Pesticide Action Network (PAN), Europe study of pesticide and biocide; and biocide reference guide. Several other biocidal products such as Quaternary ammonium compounds (QACs) or disinfectants were collected from literature. [250-252]. The final suspects list includes 273 biocides and active ingredients of pesticides alongside their chemical identifiers, predicted t_R and three most common and abundant MS/MS fragments from spectra library (MoNA (http://mona.fiehnlab.ucdavis.edu/), MassBank (https://massbank.eu/MassBank/) and mzCloud (https://www.mzcloud.org/)) [105, 114, 253]. This suspects list can be found in SM (appendix A), Table A 6.1.

"AutoNonTarget" can also search online databases (Pubchem or any of those mentioned in the previous paragraph) with a mass accuracy threshold defined by the user. This could be accompanied by user based inputs such as molecular formula, inclusion/exclusion list, number of references/patents or usage/production data, certain atom types, substructure or neutral compounds or compounds with positive or negative charges. Another capability

is the batch mode candidates retrieval for whole peaks list with specific rules for various mass ranges.

6.2.7. Assignment of molecular formula

Generally, two ways can be used to obtain a chemical formula for an observed m/z value. First and most accurate solution is to utilize 7 golden rules to create and evaluate molecular formula [6] and subsequently compare isotopic abundance patterns of mass spectra and the created molecular formula. Isotope ratios are measured from the very beginning of mass spectrometry. Natural occurring elements can be monoisotopic (F, Na, P, I) or polyisotopic (H, C, N, O, S, Cl, Br). Using isotopic pattern generators one can calculate the contribution to the abundances of the M+1, M+2, M+3 isotope ions in mass spectra, where M+• or M-• reflect the molecular ion [254, 255]. However, the number of elements increase in complex and large molecules, the computation of correct isotope ratios becomes more complicated. A molecular formula can also be assigned after searching the online chemical abstracts service or public chemical database within mass accuracy of the instrument. After retrieval of candidates, the molecular formula of the candidates can be evaluated basically from isotopic pattern. “AutoNonTarget” utilizes the “enviPat” R package [82] to calculate the isotopic pattern of the chemical formula of each candidate. Then, it uses dot product method to compare the theoretical and experimental isotopic peaks.

6.2.8. MS/MS spectral interpretation and prediction

The large majority of these substances or peaks detected in samples typically remain unidentified. As being said, when the reference standards are not available or not present in the spectral libraries or even sometimes numerous potential candidates are proposed, use of *in silico* computational tools are recommended. At this stage, several *in silico* fragmentation tools such as MetFrag [7] or CFM-ID [8] are proposed and used widely to interpret and predict MS/MS fragments, respectively. After retrieval of candidates, the MS/MS fragments of the m/zs of interest are being extracted from the raw data (recorded either by DIA or DDA mode). Afterwards, MetFrag is used to interpret the fragments and rank the candidates according to the errors of explained fragments. Similarly, CFM-ID is

done as complementary approach not to only rank the candidates, but also to predict the MS/MS fragments for to compare with it with experimental ones in case of not existing in the mass spectrum libraries.

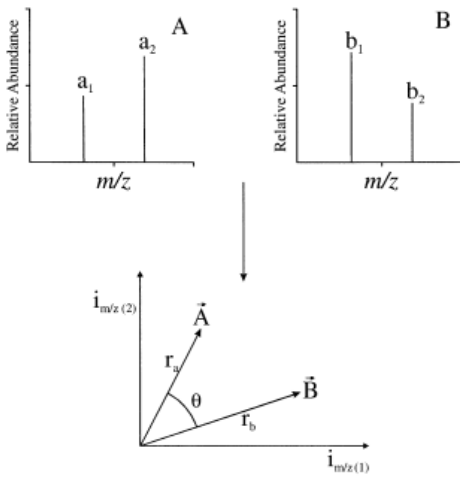
6.2.9. In-house QSRR models for retention time prediction

The QSRR models introduced in chapter 3 of this thesis have been used for prioritization of the candidates alongside the MCS. This is done considering the MEAN error of each predicted error from Monte Carlo sampling plot.

6.2.10. Diagnostic evidence

Diagnostic evidence, as discussed in chapter 1, refers to sort of information that increase the identification confidence when there is no MSMS fragments of reference standard is available. This may include the ink on the cover of a package material and later on its migration to food content, or comparison of fragmentation pathway from a parent compound and its transformation product. In some cases, specific ions prove the substructure of a compound and can be used to interpret/elucidate the chemical structure easily. User can define a diagnostic ion list (requires recursive SMARTS (a language for describing molecular patterns) substructure information [256] and m/z values) to search them within MS/MS fragments list and produce output that contains the potential candidates including the ions. As an extra evaluation step, the predicted MS/MS fragments of the retrieved candidates are being compared to the experimental ones to reach probable candidates.

There are four main approaches to compare mass spectral data from library with a newly observed experimental one. These methods are 1) probability-based matching [257, 258], 2) dot product (cosine similarity approach) [259], 3) weighted dot product [259, 260] and 4) mass spectral tree search (mzCloud). PBM computes the similarity between two spectra from the statistical probabilities of observing identical peaks between them and it is relatively complex. Dot product approach is the simplest approach and accurate enough to compare two mass spectra. Figure 6.6 shows the concept behind the dot product mass spectral similarity.



$$cos\theta = Spec.Similarity = \frac{\sum_{i=1}^{i=n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Figure 6.6 Dot product basics for mass spectral comparison

X and Y are two vectors obtained from mass spectra (encoded intensity data for the same m/zs). A similarity of 1 means the two vectors are identical, and a similarity of 0 means they are orthogonal and independent of each other. The only issues with this approach is that mass spectra recorded at different collision energy would give low similarity score owning to vector length difference ($r_{\boldsymbol{a}}$ vs $\boldsymbol{b}$). Last but not least, MS^n trees approach links ion-fragmentation pathways with substructure relationships in a hierarchical order. The significant aspect of MSn trees is that they can reveal both the dependency of precursor/product ion and product ion/product ion within the same MS^n stage or between different MS^n stages and finally help link all product ions to specific precursor ions. Two spectra are similar when all products ions are observed in fragmentation tree regardless of other ions that are generated due to different collision energy or different HRMS instrument. There a few tools available (MassFrontier or mzCloud) to use this interesting approach and computational tools are needed to achieve interpretation of fragmentation trees.

For “AutoNonTarget”, we have used modified version of dot product approach to compare MS/MS data from mass spectrum library (MS/MS data are extracted mainly from MassBank, GNPS, MoNA, Metlin, HMDB and FooDB) with experimental data. The steps used in modified dot product approach for mass spectra similarity purpose are noted below.

- Calculate error (ppm) between m/z of reference spectra and suspect
- Find those that are below 5 ppm
- Count m/z matched and divide it by total number of m/z in ref spectra (intensity threshold included)
- Calculate conventional dot product score
- Calculate mass spectra similarity as follows:
- $Spec.Sim. = \alpha * dot\ product\ score + \beta * score\ from\ counting\ matched\ m/zs$

α and β are weights set by two criteria: user defined/ default (α =0.81 and β=0.19 (when the collision energy is the same for the MS/MS of reference and suspected spectra); otherwise, α =0.5 and β=0.5). This compensates the low score data when the collision engery is different.

Comparison of experimental RTI value of reference standard and a retrieved candidate is also another approach to increase identification confidence and reach a probable structure. This requires a construction of a RTI bank, a database like NIST, for LC. A huge list of experimentally derived RTI data (~4000 compounds) was compiled and used to compare it with a matched candidates by means of multiple comparison procedure.

6.2.11. Experimental MS/MS spectral match

Herein, the MassBank scores are calculated on the basis of a modified cosine distance to compute the similarity between the query spectrum and the reference spectra and the results are ranked according to this spectral similarity.

6.2.12. Confirmation by authentic reference standard

To reach level one, researcher should provide the raw data of the reference standards mixtures prepared and ran alongside the other samples. After reach to level of identification above 2, these candidates (their m/zs, retention time and retention time indices as well as MS/MS) are being searched within these raw data to confirm them at the level of identification 1. Alternatively, researchers can use the reference standards mixture raw data to look up these compounds in the samples (much like target screening) with a csv file including the information of targeted compounds (exist in the mixture).

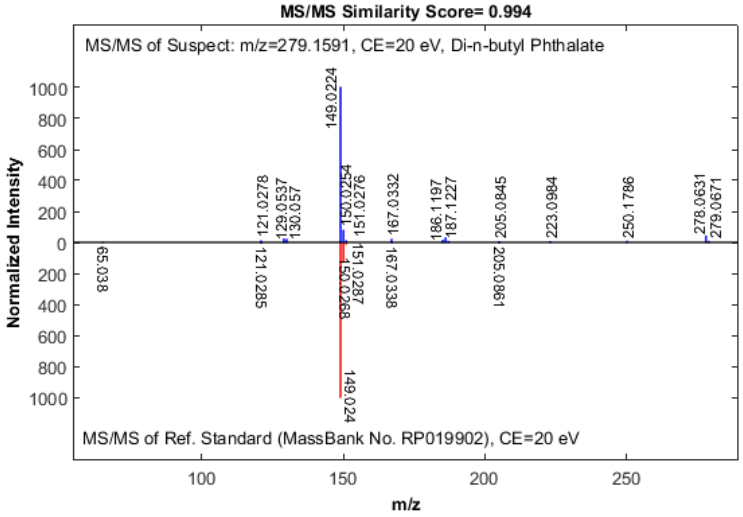
6.2.13. Aquatic risk assessment

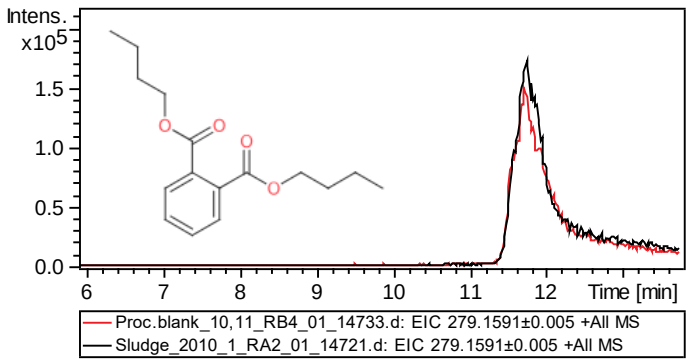
The QSTR models introduced in chapter 5 of this thesis have been used for estimating the aquatic acute toxicity of any compounds that is reached to level of identification confidence above 2.

6.3. Results and Discussion

6.3.1. Detected false positives by Deep Learner

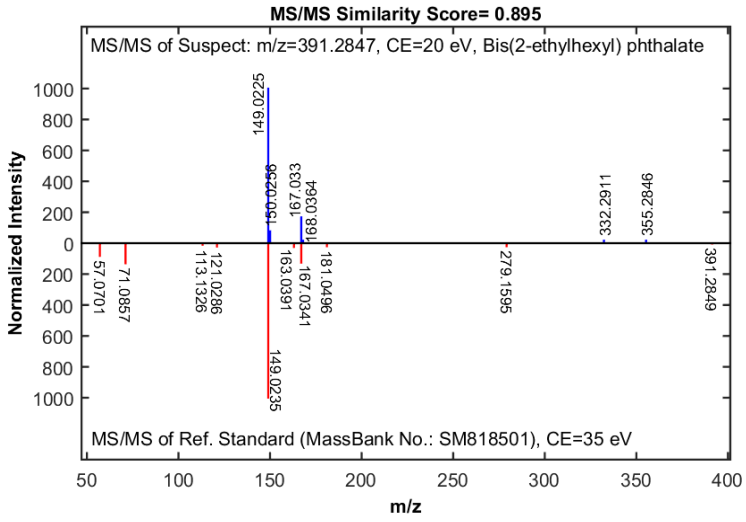
In this step, comparison of the sample with control or analytical procedural blank samples is important to exclude irrelevant peaks. The removal of noise peaks, mass recalibration and componentization of isotopes and adducts is usually carried out automatically as the next step. During the analysis of IWW/EWW and sewage sludge samples, some false positives (Di-n-butyl Phthalate, Triphenyl Phosphate and Bis(2-ethylhexyl) Phthalate) were detected which are mainly plasticizers with the probability above 0.9 (derived from Deep Learner ANN approach for analytical procedural blank removal). Figure 6.7 shows the plasticizers identified as false positive.

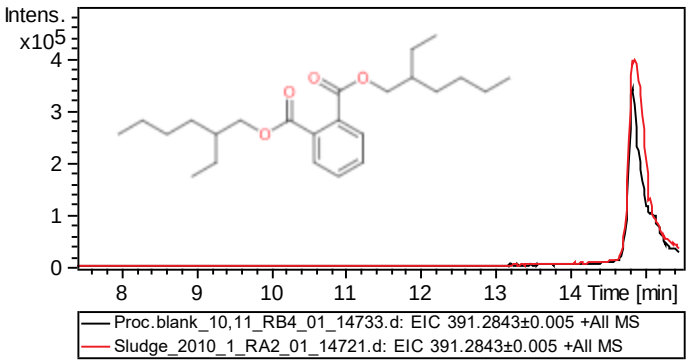




Probability Values derived by Deep Learner		
False Positive	True Positive	Noise Peak
0.999	0.001	0.000

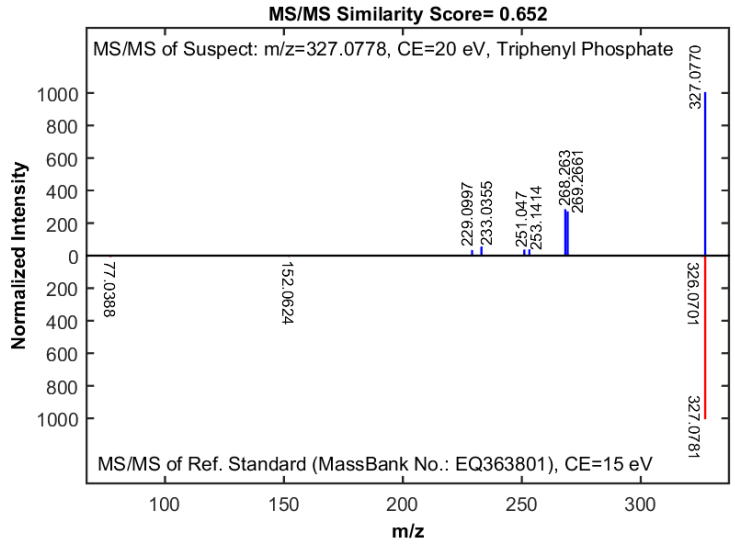
Identified compound:
Di-n-butyl Phthalate
Identification level: **2a**
Exp. t_R= 11.78 min, Pred. t_R= 11.52 min

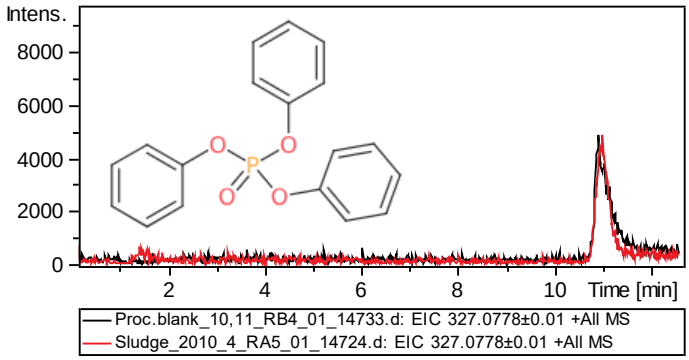




Probability Values derived by Deep Learner		
False Positive	True Positive	Noise Peak
0.989	0.011	0.000

Identified compound:
Bis(2-ethylhexyl) Phthalate
Identification level: **2a**
Exp. t_R= 14.99 min, Pred. t_R= 15.10 min





Probability Values derived by Deep Learner		
False Positive	True Positive	Noise Peak
0.999	0.000	0.001

Identified compound:
Triphenyl Phosphate
Identification level: **2b** (predicted and experimental tR matches). Most of the fragments are at low intensity, but explained by *in silico* fragmentation tool. A clear MSMS spectra is required.
Exp. t_R= 11.10 min, Pred. t_R= 11.80 min

Figure 6.7 False positive detected by "Deep Learner" as analytical procedural subtraction approach

6.3.2. Screening of biocides in wastewater and sludge

"AutoNonTarget" was applied to the suspect screening of over 273 biocides and active ingredients of pesticides in sewage sludge and wastewater samples. Nine target biocides (Azoxystrobin, DEET, 5-Methylbenzotriazole, Fluometuron, Fludioxonil, Triclocarban, Benzoic acid, Terbutylazine, and Climbazole) were treated as suspects and used for validation of the proposed automatic screening workflow.

Two very intense peaks corresponding to *m/z* 404.1250 and 192.1392 were detected at 8.89 min and 8.02 min and matched to azoxystrobin and DEET, from the biocide suspects list, after applying the mass accuracy and isotopic fit filtering (cosine fit (a simple dot product result between theoretical (calculated by enviPat [82]) and extracted isotopic pattern for given molecular formula (threshold >0.35)), respectively. Both these compounds had the error of predicted t_R below ±0.14 min and most of the fragments were explained by *in silico* fragmentation tools (MetFrag [7] and CFM-ID [8]). MCS results (region A: not false positive) were also in favor of accepting the predicted t_R values for these two compounds. These facts made these suspects suitable candidates for validation of the proposed identification workflow and they have been further confirmed by reference standards at the level of identification confidence of 1, according to their t_R values and MS/MS fragmentations match. Azoxystrobin and DEET were detected and identified in all IWW, EWW and sewage sludge. The biodegradability half-life of azoxystrobin and DEET were predicted to be approximately 4 days in the receiving environment [261, 262]. Table A 6.2 (SM, Chapter 6, Appendix A) provides the full identification procedure for all detected compounds in this study (n=28). The data from the spectral libraries (MoNA (http://mona.fiehnlab.ucdavis.edu/), MassBank (https://massbank.eu/MassBank/) and mzCloud (https://www.mzcloud.org/)) were also used to increase the level of identification confidence in some of the detected compounds such as 5-Methylbenzotriazole (m/z=134.0710, t_R=1.62 min), Fluometuron (m/z=231.0756, t_R=7.92 min), Fludioxonil (m/z=247.0324, t_R=9.71 min), Triclocarban (m/z=312.9711, t_R=12.06 min), Benzoic acid (m/z=121.0291, t_R=4.70 min), Decanoic acid (m/z=171.1391, t_R=9.69 min), Terbuthylazine (m/z=230.1161, t_R=9.32 min), Ketoconazole (m/z=531.1560, t_R=9.69 min), Climbazole (m/z=293.1055, t_R=9.84 min). For some of these compounds (5-Methylbenzotriazole, Fluometuron, Fludioxonil, Benzoic

acid, Terbutylazine and Climbazole), the reference standards were already available and therefore, the level of identification confidence was reached to 1 after evaluating t_R and MS/MS info. This proves the reliability of the applied screening workflow and use of t_R prediction models as well as MCS plot during the identification procedure. Two compounds (Decanoic acid and Ketoconazole) were identified at the level of identification confidence 2a, because the standards were not available but the predicted *versus* experimental t_R were acceptable (Region A and B in the MCS plot), and the MS/MS similarity score (modified dot product between library spectra and observed one) [263, 264] were at least above 0.7. All of these identified compounds were detected in the EWW, and have a biodegradability half-life between 3-7 days.

26 biocides were identified through suspect screening via “AutoNonTarget”, including different classes such as preservatives, disinfectants, repellents, veterinary hygiene and quaternary ammonium compounds (QACs). Four candidates for two other ions (m/z 214.2539 and 242.2842) (QACs: Undecyltrimethylammonium (ATMAC-11) and Ethyldecyldimethylammonium (DADMAC-2:10); Tridecyltrimethylaminium (ATMAC-13) and Butyl-decyl-dimethyl-ammonium (DADMAC-4:10)) were identified and reported for the first time via non-target screening strategy. Table 6.1 provides the list of 28 identified biocides in the influent/effluent wastewater and sewage sludge samples from WWTP of Athens.

Among the suspect screening of biocides and the identification results, several homologous series (QACs) have been detected (n=13). The fragmentation of these homologous was straightforward where, for instance, the benzylic amine bond breaks in Benzyl-dimethyl-n(alkylchain)-ammonium chloride (BAC-n(the alkyl chain number) and leads to the diagnostic ion at m/z 91.0542, known as the tropylium ion, and the related fragments corresponding to the unique alkyl chain substructure for each one of the homologous [265]. BAC-10, BAC-12, BAC-14 and BAC-16 were identified successfully considering the t_R prediction models, MCS plot, observing the diagnostic ion at m/z 91.0542 as well as matching the list of observed fragments to those previously reported in the literature. Therefore, the identification reached to level 2a. The full identification procedure, including the extracted ion chromatogram, MCS plot as well as MS/MS fragmentation can be found in Table A 6.2. (n-Alkyl)-trimethyl-ammonium (ATMACs)

homologous series were also detected and identified through t_R prediction model and MS/MS fragmentation pattern. Breaking the bonds in ATMACs homologous leads to the diagnostic ion at m/z 60.0807 which is trimethyl-ammonium ion [252]. ATMAC-12, ATMAC-14, ATMAC-16 and ATMAC-18 were identified at level 2a, as the predicted t_R was matching to the experimental one (MCS plot) and the MS/MS fragmentation pattern was similar to those that reported in the literature [252]. For these 4 ATMACs, the diagnostic ion was observed at high intensity and the MS/MS spectra was easily interpretable. However two other ATMACs (ATMAC-10 and ATMAC-20) did not present this diagnostic ion at high intensity and the MS/MS spectra was not clear. Therefore, these two QACs were tentatively identified at a level of identification 3. Another set of abundant homologous (paired and mixed di(n-alkyl)dimethylammonium (DADMAC)) were detected in the sewage sludge samples. Two paired DADMACs (Dioctyldimethylammonium bromide (DADMAC-8:8) and Didecyldimethylammonium bromide (DADMAC-10:10)) as well as a mixed DADMAC (Dimethyloctyldecylammonium bromide (DADMAC-8:10)) were tentatively identified at level of 2a, 3 and 2a, respectively. The predicted t_R and MCS plot were acceptable for the DADMAC-8:8 and DADMAC-8:10 and their MS/MS fragments were explicable among which two fragments (m/z 158.1896 and m/z 186.2201) were matching to the reported ions in the literature [252]. DADMAC-10:10 was also tentatively identified at level of identification 3 after observing only a single diagnostic fragment (m/z 186.2209) and predicted t_R match.

Through non-target screening two new QACs have been found at m/z 214.2539 and 242.2842. For the ion 214.2539, 60 candidates were retrieved from PubChem after applying mass accuracy and isotopic fit filter. MetFrag was used to prioritize these 60 candidates based on their explained MS/MS fragments. Having used t_R models and MCS plot, two most probable candidates were ATMAC-11 or DADMAC-2:10. ATMAC-11 was then assigned to this ion due to the lower t_R prediction error than DADMAC-2:10, however the diagnostic ion for ATMAC homologous (m/z 60.0807 which is trimethyl-ammonium ion) was not observed in the MS/MS spectra. Therefore, it is tentatively identified at the level of identification 3 (list of explained MS/MS fragments for m/z 214.2539, based on in silico fragmentation tool (MetFrag), can be found in Table A 6.3). For the ion 242.2842, 74 candidates were retrieved from PubChem, and after applying all identification

procedure said above, two most probable candidates (ATMAC-13 and DADMAC-4:10) were assigned to this m/z. ATMAC-13 was assigned to this ion due to the lower t_R prediction error than DADMAC-4:10, however some more evidence are required to confirm this structure. Therefore, ATMAC-13 is tentatively identified at the level of identification 3 (list of explained MS/MS fragments for m/z 242.2842, based on in silico fragmentation tool (MetFrag), can be found in Table A 6.4). These new detected QAC homologous were also found at high abundance in IWW and EWW. Further investigations on the occurrence and fate of these newly identified water soluble ATMACs and mixed DADMACs, as well as the potential ecological effects of QACs are still warranted and it will be the subject of further studies in order to better evaluate their behavior in the environment. Most of the identified biocides were found to be present in EWW, with a predicted biodegradation half-time of 3-17 days (pseudo-persistent compounds). Two new quaternary ammonium compounds (QACs) were also tentatively identified via non-target screening strategy.

Table 6.1 List of identified biocides in influent, effluent wastewater (IWW & EWW) and sewage sludge of wastewater treatment plants (WWTP) of Athens (Greece)

Compound Name	CAS No.	Class of Biocide	Measured m/z	Exp. t_R (Pred. t_R) (min)	LC-HRMS platform	Identified in	Level of identification confidence
Azoxystrobin	131860-33-8	Preservatives	404.1250	8.89 (9.02)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	1
5-Methylbenzotriazole	29878-31-7	Benzotriazoles	134.0710	1.62 (1.61)	HILIC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	1
DEET	134-62-3	Repellents & attractants	192.1392	8.02 (7.99)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	1
Fluometuron	2164-17-2	Herbicide	231.0756	7.92 (8.07)	RPLC-(-ESI)QTOF-MS	Sewage Sludge, IWW& EWW	1
Fludioxonil	131341-86-1	Preservatives	247.0324	9.71 (8.16)	RPLC-(-ESI)QTOF-MS	Sewage Sludge	1
Triclocarban	101-20-2	Cleaning products	312.9711	12.06 (11.17)	RPLC-(-ESI)QTOF-MS	Sewage Sludge	1
Benzoic acid	65-85-0	Veterinary hygiene	121.0291	4.70 (3.59)	RPLC-(-ESI)QTOF-MS	Sewage Sludge, IWW & EWW	1
Lauric acid	143-07-7	Repellents & attractants	199.1706	11.64 (10.28)	RPLC-(-ESI)QTOF-MS	Sewage Sludge, IWW & EWW	3
Decanoic acid	334-48-5	Repellents & attractants	171.1391	9.69 (8.84)	RPLC-(-ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a

Compound Name	CAS No.	Class of Biocide	Measured m/z	Exp. t_R (Pred. t_R) (min)	LC-HRMS platform	Identified in	Level of identification confidence
Pelargonic acid	112-05-0	Disinfectants & algaecides	157.1234	8.76 (8.27)	RPLC-(-ESI)QTOF-MS	Sewage Sludge, IWW & EWW	3
Terbutylazine	5915-41-3	Herbicides	230.1161	9.32 (9.26)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	1
Ketoconazole	65277-42-1	Fungicides	531.1560	9.69 (10.32)	RPLC-(+ESI)QTOF-MS	Sewage Sludge	2a
Climbazole	38083-17-9	Fungicides	293.1055	9.84 (9.98)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	1
Benzyldimethyldecyl ammonium chloride (BAC-10)	965-32-2	QACs [a]	276.2695	10.10 (10.59)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a
Benzyldimethyldodecyl ammonium chloride (BAC-12)	139-07-1	QACs [a]	304.3004	11.49 (11.11)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a
Benzyldimethyltetradecyl ammonium chloride (BAC-14)	139-08-2	QACs [a]	332.3311	12.58 (11.82)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a
Benzyldimethylhexadecyl ammonium chloride (BAC-16)	122-18-9	QACs [a]	360.3625	13.46 (12.30)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a
Decyltrimethyl ammonium bromide (ATMAC-10)	2082-84-0	QACs [a]	200.2370	11.24 (8.45)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	3

Compound Name	CAS No.	Class of Biocide	Measured m/z	Exp. t_R (Pred. t_R) (min)	LC-HRMS platform	Identified in	Level of identification confidence
Dodecyltrimethyl ammonium bromide (ATMAC-12)	1119-94-4	QACs [a]	228.2682	10.96 (8.83)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a
Tetradecyltrimethyl ammonium bromide (ATMAC-14)	1119-97-7	QACs [a]	256.2998	12.21 (10.13)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a
Hexadecyltrimethyl ammonium bromide (ATMAC-16)	57-09-0	QACs [a]	284.3314	13.46 (12.22)	RPLC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	2a
Trimethyloctadecyl ammonium bromide (ATMAC-18)	1120-02-1	QACs [a]	312.3631	14.24 (12.16)	RPLC-(+ESI)QTOF-MS	Sewage Sludge	2a
Eicosyltrimethyl ammonium bromide (ATMAC-20)	7342-61-2	QACs [a]	340.3934	14.94 (12.20)	RPLC-(+ESI)QTOF-MS	Sewage Sludge	3
Dioctyldimethyl ammonium bromide (DADMAC-8:8)	3026-69-5	QACs [a]	270.3159	11.09 (10.54)	RPLC-(+ESI)QTOF-MS	Sewage Sludge	2a
Didecyldimethyl ammonium bromide (DADMAC-10:10)	2390-68-3	QACs [a]	326.3788	13.12 (12.54)	RPLC-(+ESI)QTOF-MS	Sewage Sludge	3
Dimethyloctyldecyl ammonium bromide (DADMAC-8:10)	N.A.	QACs [a]	298.3471	12.28 (11.58)	RPLC-(+ESI)QTOF-MS	Sewage Sludge	2a
ATMAC-11 / DADMAC-2:10 [b]	N.A.	QACs [a]	214.2530	5.94 (5.79 & 5.44)	HILIC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	3
ATMAC-13 / DADMAC-4:10 [b]	N.A.	QACs [a]	242.2845	5.88 (5.92 & 5.94)	HILIC-(+ESI)QTOF-MS	Sewage Sludge, IWW & EWW	3

[a] Quaternary Ammonium Compounds (QACs)

[b] Identified through non-target screening workflow

6.4. Conclusions

This study presents an automatic data analysis workflow (“AutoNonTarget”) for UHPLC-HRMS-based target, suspect and non-target screening. Novel methods for EIC extraction, analytical procedural blank removal, chemometrics and prioritization, retrieval of candidates and detection of false positives are developed. Using a suspect list of 273 biocides as well as ability for retrieval of candidates from Pubchem, and screening them in the environmental samples (influent/effluent wastewater and sewage sludge) collected from WWTP of Athens (Greece), the advantages of our new approach to data evaluation during suspect/non-target screening analysis could be demonstrated. 28 biocides are identified in the collected samples. This was the first report of identification of biocides in the receiving environment of Athens (Greece). Most of the identified biocides were found to be present in effluent wastewater having the predicted biodegradation time of 3-17 days. Two new quaternary ammonium compounds (QACs) were also tentatively identified *via* non-target screening strategy. Use of this new approach to data evaluation especially in non-target screening analyses opens the possibilities of various other applications for regulatory body to get advantage of comprehensive monitoring of chemicals in environmental counterparts.

CHAPTER 7

Concluding Remarks

There has been a great development during the last decade in the field of environmental analysis especially with the advancement of LC-HRMS and related computational tools. These methods based on HRMS can provide valuable information about the occurrence, fate and distribution of analytes in the receiving environment. However, supportive tools would be needed to facilitate screening approaches (suspect/non-target screening).

Suspect screening is a technique for the identification of compounds based on previously available knowledge such as the origin of samples, exposed chemicals or contaminants, or list of expected chemical components. This approach is for a specific purpose and can provide fast, valuable and reliable information. However, there is still a need for more complete compound databases and suspect list compilation approach, mass spectral libraries and computational tools for prioritization of candidates (*in silico* fragmentation or retention time prediction models). In this thesis, new retention time models are presented for RPLC and HILIC with an advanced uncertainty measure tool for ranking the candidates based on their mean predictive error from Monte Caro sampling result.

As for Non-target Screening, it is vital for a comprehensive environmental analysis, because the majority of the compounds remain unknown in the samples. We have established a novel retention time indices for LC-HRMS and used it in several collaborative trials within Norman network. This is a major accomplishment, as RTI values of experimental data and reference standards can be compared to prioritize the candidates or even detect false positive. Chapter 6 of this thesis has also introduced a new automatic tool to facilitate the identification procedure in LC-HRMS.

Although we have developed some advanced *in silico* approaches to facilitate the suspect and non-target screening, the major focus was given to resolve retention time elution pattern of emerging pollutants and their applications in false positive removal. To enable the use of toxicity models introduced in chapter 5, there is a need for quantitative/semi-quantitative approach to compare the PNEC values. Moreover, use of instrumental response factor and ionization potential as a key feature to rank candidates in a typical suspect/non-target screening task remained a challenging task.

References

1. M. Krauss, H. Singer, and J. Hollender, LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns, *Analytical and Bioanalytical Chemistry*, vol. 397, no. 3, 2010, pp. 943-951.
2. E. L. S. Pablo Gago-Ferrero, Anna A. Bletsou, Reza Aalizadeh, Juliane Hollender, Nikolaos S. Thomaidis, Extended Suspect and Non-Target Strategies to Characterize Emerging Polar Organic Contaminants in Raw Wastewater with LC-HRMS/MS, *Environmental Science &*, vol. 49, no., 2015, pp. 12333-12341.
3. W. Brack et al., Effect-directed analysis supporting monitoring of aquatic environments — An in-depth overview, *Science of The Total Environment*, vol. 544, no., 2016, pp. 1073-1118.
4. N. P. Kalogiouri, R. Aalizadeh, and N. S. Thomaidis, Application of an advanced and wide scope non-target screening workflow with LC-ESI-QTOF-MS and chemometrics for the classification of the Greek olive oil varieties, *Food Chemistry*, vol. 256, no., 2018, pp. 53-61.
5. M. Krauss, Chapter 15 - High-Resolution Mass Spectrometry in the Effect-Directed Analysis of Water Resources, *Comprehensive Analytical Chemistry*, S. Pérez, P. Eichhorn and D. Barceló, eds, Elsevier, 2016, pp. 433-457.
6. T. Kind and O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, *BMC Bioinformatics*, vol. 8, no. 1, 2007, pp. 105.
7. C. Ruttkies et al., MetFrag relaunched: incorporating strategies beyond in silico fragmentation, *Journal of Cheminformatics*, vol. 8, no. 1, 2016, pp. 3.
8. F. Allen et al., CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra, *Nucleic Acids Research*, vol. 42, no. W1, 2014, pp. W94-W99.
9. R. Aalizadeh, N. S. Thomaidis, A. A. Bletsou, and P. Gago-Ferrero, Quantitative Structure–Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples, *Journal of Chemical Information and Modeling*, vol. 56, no. 7, 2016, pp. 1384-1398.
10. E. L. Schymanski et al., Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis, *Analytical and Bioanalytical Chemistry*, vol. 407, no. 21, 2015, pp. 6237-6255.
11. J. Hollender, E. L. Schymanski, H. P. Singer, and P. L. Ferguson, Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go?, *Environmental Science & Technology*, vol. 51, no. 20, 2017, pp. 11505-11512.
12. T. R. Croley, K. D. White, J. H. Callahan, and S. M. Musser, The Chromatographic Role in High Resolution Mass Spectrometry for Non-Targeted Analysis, *Journal of The American Society for Mass Spectrometry*, vol. 23, no. 9, 2012, pp. 1569-1578.
13. P. Rostkowski, P. Haglund, C. Dye, and M. Schlabach, *Non-target screening of environmental samples by low and high resolution time of flight mass spectrometry (TOF-MS)*, in *13th International Conference on Environmental Science and Technology (CEST), SEP 05-07, 2013, Athens, GREECE*, T. D. Lekkas, Editor. 2013, Global Nest, Secretariat.

14. R. A. van den Berg et al., Centering, scaling, and transformations: improving the biological information content of metabolomics data, *BMC Genomics*, vol. 7, no. 1, 2006, pp. 142.
15. S. Castillo, P. Gopalacharyulu, L. Yetukuri, and M. Orešič, Algorithms and tools for the preprocessing of LC–MS metabolomics data, *Chemometrics and Intelligent Laboratory Systems*, vol. 108, no. 1, 2011, pp. 23-32.
16. R. Arneberg et al., Pretreatment of Mass Spectral Profiles: Application to Proteomic Data, *Analytical Chemistry*, vol. 79, no. 18, 2007, pp. 7014-7026.
17. H. G. Gika et al., High temperature-ultra performance liquid chromatography–mass spectrometry for the metabonomic analysis of Zucker rat urine, *Journal of Chromatography B*, vol. 871, no. 2, 2008, pp. 279-287.
18. N. P. Kalogiouri, R. Aalizadeh, and N. S. Thomaidis, Investigating the organic and conventional production type of olive oil with target and suspect screening by LC-QTOF-MS, a novel semi-quantification method using chemical similarity and advanced chemometrics, *Analytical and Bioanalytical Chemistry*, vol. 409, no. 23, 2017, pp. 5413-5426.
19. L. A. Berrueta, R. M. Alonso-Salces, and K. Héberger, Supervised pattern recognition in food analysis, *Journal of Chromatography A*, vol. 1158, no. 1, 2007, pp. 196-214.
20. M. Barker and W. Rayens, Partial least squares for discrimination, *Journal of Chemometrics*, vol. 17, no. 3, 2003, pp. 166-173.
21. M. Bylesjö et al., OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, *Journal of Chemometrics*, vol. 20, no. 8-10, 2006, pp. 341-351.
22. L. Breiman, Random forests, *Machine Learning*, vol. 45, no. 1, 2001, pp. 5-32.
23. N. P. Kalogiouri, N. A. Alygizakis, R. Aalizadeh, and N. S. Thomaidis, Olive oil authenticity studies by target and nontarget LC–QTOF-MS combined with advanced chemometric techniques, *Analytical and Bioanalytical Chemistry*, vol. 408, no. 28, 2016, pp. 7955-7970.
24. F. Marini, J. Zupan, and A. L. Magrì, Class-modeling using Kohonen artificial neural networks, *Analytica Chimica Acta*, vol. 544, no. 1, 2005, pp. 306-314.
25. P. M. Bastos and P. Haglund, The use of comprehensive two-dimensional gas chromatography and structure–activity modeling for screening and preliminary risk assessment of organic contaminants in soil, sediment, and surface water, *Journal of Soils and Sediments*, vol. 12, no. 7, 2012, pp. 1079-1088.
26. H. P. Singer, A. E. Wössner, C. S. McArdell, and K. Fenner, Rapid Screening for Exposure to “Non-Target” Pharmaceuticals from Wastewater Effluents by Combining HRMS-Based Suspect Screening and Exposure Modeling, *Environmental Science & Technology*, vol. 50, no. 13, 2016, pp. 6698-6707.
27. J. E. Rager et al., Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring, *Environment International*, vol. 88, no., 2016, pp. 269-280.
28. Q. Huang et al., Derivation of aquatic predicted no-effect concentration (PNEC) for ibuprofen and sulfamethoxazole based on various toxicity endpoints and the associated risks, *Chemosphere*, vol. 193, no., 2018, pp. 223-229.
29. H. Sanderson et al., Ranking and prioritization of environmental risks of pharmaceuticals in surface waters, *Regulatory Toxicology and Pharmacology*, vol. 39, no. 2, 2004, pp. 158-183.

30. Å. Wennmalm and B. Gunnarsson, Public Health Care Management of Water Pollution with Pharmaceuticals: Environmental Classification and Analysis of Pharmaceutical Residues in Sewage Water, *Drug Information Journal*, vol. 39, no. 3, 2005, pp. 291-297.
31. E. R. Cooper, T. C. Siewicki, and K. Phillips, Preliminary risk assessment database and risk ranking of pharmaceuticals in the environment, *Science of The Total Environment*, vol. 398, no. 1, 2008, pp. 26-33.
32. P. H. Howard and D. C. G. Muir, Identifying New Persistent and Bioaccumulative Organics Among Chemicals in Commerce II: Pharmaceuticals, *Environmental Science & Technology*, vol. 45, no. 16, 2011, pp. 6938-6946.
33. V. Dulio et al., Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations, *Environmental Sciences Europe*, vol. 30, no. 1, 2018, pp. 5.
34. R. Aalizadeh, M.-C. Nika, and N. S. Thomaidis, Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants, *Journal of Hazardous Materials*, vol. 363, no., 2019, pp. 277-285.
35. R. Bade et al., Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis, *Sci Total Environ*, vol. 538, no., 2015, pp. 934-941.
36. V. G. Beretsou et al., Identification of biotransformation products of citalopram formed in activated sludge, *Water Research*, vol. 103, no., 2016, pp. 205-214.
37. P. Gago-Ferrero et al., Extended Suspect and Non-Target Strategies to Characterize Emerging Polar Organic Contaminants in Raw Wastewater with LC-HRMS/MS, *Environmental Science & Technology*, vol. 49, no. 20, 2015, pp. 12333-12341.
38. M. Ibáñez et al., UHPLC-QTOF MS screening of pharmaceuticals and their metabolites in treated wastewater samples from Athens, *Journal of Hazardous Materials*, vol. 323, no., 2017, pp. 26-35.
39. M. Hu et al., Performance of combined fragmentation and retention prediction for the identification of organic micropollutants by LC-HRMS, *Analytical and Bioanalytical Chemistry*, vol. 410, no. 7, 2018, pp. 1931-1941.
40. Q. Zhang et al., A strategy to improve the identification reliability of the chemical constituents by high-resolution mass spectrometry-based isomer structure prediction combined with a quantitative structure retention relationship analysis: Phthalide compounds in Chuanxiong as a test case, *Journal of Chromatography A*, vol. 1552, no., 2018, pp. 17-28.
41. T. Bączek and R. Kaliszan, Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics, *PROTEOMICS*, vol. 9, no. 4, 2009, pp. 835-847.
42. A. A. Klammer, X. Yi, M. J. MacCoss, and W. S. Noble. *Peptide Retention Time Prediction Yields Improved Tandem Mass Spectrum Identification for Diverse Chromatography Conditions*. 2007. Berlin, Heidelberg: Springer Berlin Heidelberg.
43. F. Falchi et al., Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification, *Analytical Chemistry*, vol. 88, no. 19, 2016, pp. 9510-9517.

44. M. Cao et al., Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics, *Metabolomics*, vol. 11, no. 3, 2015, pp. 696-706.
45. Z. Dashtbozorgi, H. Golmohammadi, and E. Konoz, Support vector regression based QSPR for the prediction of retention time of pesticide residues in gas chromatography–mass spectroscopy, *Microchemical Journal*, vol. 106, no. 0, 2013, pp. 51-60.
46. R. Kaliszan, QSRR: Quantitative Structure-(Chromatographic) Retention Relationships, *Chemical Reviews*, vol. 107, no. 7, 2007, pp. 3212-3246.
47. A. R. Katritzky et al., Quantitative Structure–Activity Relationship (QSAR) Modeling of EC50 of Aquatic Toxicities for Daphnia magna, *Journal of Toxicology and Environmental Health, Part A*, vol. 72, no. 19, 2009, pp. 1181-1190.
48. A. D. McEachran et al., A Comparison of Three Liquid Chromatography (LC) Retention Time Prediction Models, *Talanta*, no., 2018.
49. K. Gorynski et al., Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds, *Anal Chim Acta*, vol. 797, no., 2013, pp. 13-19.
50. E. F. Hewitt, P. Lukulay, and S. Galushko, Implementation of a rapid and automated high performance liquid chromatography method development strategy for pharmaceutical drug candidates, *Journal of Chromatography A*, vol. 1107, no. 1, 2006, pp. 79-87.
51. E. Tyrkkö, A. Pelander, and I. Ojanperä, Prediction of liquid chromatographic retention for differentiation of structural isomers, *Analytica Chimica Acta*, vol. 720, no., 2012, pp. 142-148.
52. M. Rosés and E. Bosch, Linear solvation energy relationships in reversed-phase liquid chromatography. Prediction of retention from a single solvent and a single solute parameter, *Analytica Chimica Acta*, vol. 274, no. 1, 1993, pp. 147-162.
53. R.-J. Hu et al., QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors, *Talanta*, vol. 68, no. 1, 2005, pp. 31-39.
54. Y. Du and Y. Liang, Data mining for seeking accurate quantitative relationship between molecular structure and GC retention indices of alkanes by projection pursuit, *Computational Biology and Chemistry*, vol. 27, no. 3, 2003, pp. 339-353.
55. M. Turowski et al., Selectivity of stationary phases in reversed-phase liquid chromatography based on the dispersion interactions, *Journal of Chromatography A*, vol. 911, no. 2, 2001, pp. 177-190.
56. R. I. J. Amos et al., Molecular modeling and prediction accuracy in Quantitative Structure-Retention Relationship calculations for chromatography, *TrAC Trends in Analytical Chemistry*, vol. 105, no., 2018, pp. 352-359.
57. R. Bouwmeester, L. Martens, and S. Degroeve, Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction, *Analytical Chemistry*, no., 2019.
58. L. Wu et al., Quantitative structure-ion intensity relationship strategy to the prediction of absolute levels without authentic standards, *Analytica Chimica Acta*, vol. 794, no., 2013, pp. 67-75.

59. B. Zonja, A. Delgado, S. Pérez, and D. Barceló, LC-HRMS Suspect Screening for Detection-Based Prioritization of Iodinated Contrast Media Photodegradates in Surface Waters, *Environmental Science & Technology*, vol. 49, no. 6, 2015, pp. 3464-3472.
60. M. Molíková, M. J. Markuszewski, R. Kaliszan, and P. Jandera, Chromatographic behaviour of ionic liquid cations in view of quantitative structure-retention relationship, *Journal of Chromatography A*, vol. 1217, no. 8, 2010, pp. 1305-1312.
61. D. J. Creek et al., Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction, *Analytical Chemistry*, vol. 83, no. 22, 2011, pp. 8703-8710.
62. J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman, Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods, *Journal of Molecular Graphics and Modelling*, vol. 29, no. 2, 2010, pp. 157-170.
63. J. Stanstrup, S. Neumann, and U. Vrhovsek, PredRet: prediction of retention time by direct mapping between multiple chromatographic systems, *Anal Chem*, vol. 87, no. 18, 2015, pp. 9421-9428.
64. L. M. Hall et al., Development of Ecom(5)(0) and retention index models for nontargeted metabolomics: identification of 1,3-dicyclohexylurea in human serum by HPLC/mass spectrometry, *J Chem Inf Model*, vol. 52, no. 5, 2012, pp. 1222-1237.
65. Adolfo Te´llez, Martı´ Rose´s, and E. Bosch, Modeling the Retention of Neutral Compounds in Gradient Elution RP-HPLC by Means of Polarity Parameter Models, *Analytical Chemistry*, vol. 81, no., 2009, pp. 9135-9145.
66. A. C. Guo et al., HMDB 3.0—The Human Metabolome Database in 2013, *Nucleic Acids Research*, vol. 41, no. D1, 2012, pp. D801-D807.
67. M. Heinonen, H. Shen, N. Zamboni, and J. Rousu, Metabolite identification and molecular fingerprint prediction through machine learning, *Bioinformatics*, vol. 28, no. 18, 2012, pp. 2333-2341.
68. M. Gerlich and S. Neumann, MetFusion: integration of compound identification strategies, *Journal of Mass Spectrometry*, vol. 48, no. 3, 2013, pp. 291-298.
69. R. P. Schwarzenbach et al., The Challenge of Micropollutants in Aquatic Systems, *Science*, vol. 313, no. 5790, 2006, pp. 1072-1077.
70. R. Loos et al., EU-wide survey of polar organic persistent pollutants in European river waters, *Environmental Pollution*, vol. 157, no. 2, 2009, pp. 561-568.
71. D. Hernández-Moreno et al., Acute hazard of biocides for the aquatic environmental compartment from a life-cycle perspective, *Science of The Total Environment*, vol. 658, no., 2019, pp. 416-423.
72. A. Baumer, K. Bittermann, N. Klüver, and B. I. Escher, Baseline toxicity and ion-trapping models to describe the pH-dependence of bacterial toxicity of pharmaceuticals, *Environmental Science: Processes & Impacts*, vol. 19, no. 7, 2017, pp. 901-916.
73. R. Aalizadeh, P. C. von der Ohe, and N. S. Thomaidis, Prediction of acute toxicity of emerging contaminants on the water flea Daphnia magna by Ant Colony Optimization-Support Vector Machine QSTR models, *Environmental Science: Processes & Impacts*, vol. 19, no. 3, 2017, pp. 438-448.
74. A. Cassano et al., CAESAR models for developmental toxicity, *Chemistry Central Journal*, vol. 4, no. 1, 2010, pp. S4.

75. T. W. Schultz et al., A strategy for structuring and reporting a read-across prediction of toxicity, *Regulatory Toxicology and Pharmacology*, vol. 72, no. 3, 2015, pp. 586-601.
76. E. L. Schymanski et al., Critical Assessment of Small Molecule Identification 2016: automated methods, *Journal of Cheminformatics*, vol. 9, no. 1, 2017, pp. 22.
77. B. I. Escher and K. Fenner, Recent Advances in Environmental Risk Assessment of Transformation Products, *Environmental Science & Technology*, vol. 45, no. 9, 2011, pp. 3835-3847.
78. N. A. Alygizakis et al., Exploring the Potential of a Global Emerging Contaminant Early Warning Network through the Use of Retrospective Suspect Screening with High-Resolution Mass Spectrometry, *Environmental Science & Technology*, vol. 52, no. 9, 2018, pp. 5135-5144.
79. P. Gago-Ferrero et al., Suspect Screening and Regulatory Databases: A Powerful Combination To Identify Emerging Micropollutants, *Environmental Science & Technology*, vol. 52, no. 12, 2018, pp. 6881-6894.
80. C. Kuhl et al., CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets, *Analytical Chemistry*, vol. 84, no. 1, 2012, pp. 283-289.
81. M. Loos and H. Singer, Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data, *Journal of Cheminformatics*, vol. 9, no. 1, 2017, pp. 12.
82. M. Loos et al., Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees, *Analytical Chemistry*, vol. 87, no. 11, 2015, pp. 5738-5744.
83. S. D. Richardson and T. A. Ternes, Water Analysis: Emerging Contaminants and Current Issues, *Analytical Chemistry*, vol. 86, no. 6, 2014, pp. 2813-2848.
84. M. Krauss, H. Singer, and J. Hollender, *Anal. Bioanal. Chem.*, vol. 397, no. 3, 2010, pp. 943.
85. J. J. E. Schymanski, R. Gulde, K. Fenner, M. Ruff, H. Singer, J. and Hollender, Identifying small molecules via high resolution mass spectrometry: communicating confidence, *Environmental Science & Technology*, vol. 48, no., 2014, pp. 2097-2098.
86. M. Hu et al., Performance of combined fragmentation and retention prediction for the identification of organic micropollutants by LC-HRMS, *Analytical and Bioanalytical Chemistry*, no., 2018.
87. C. Moschet, A. Piazzoli, H. Singer, and J. Hollender, Alleviating the Reference Standard Dilemma Using a Systematic Exact Mass Suspect Screening Approach with Liquid Chromatography-High Resolution Mass Spectrometry, *Analytical Chemistry*, vol. 85, no. 21, 2013, pp. 10312-10320.
88. O. V. Krokhin et al., An Improved Model for Prediction of Retention Times of Tryptic Peptides in Ion Pair Reversed-phase HPLC: Its Application to Protein Peptide Mapping by Off-Line HPLC-MALDI MS, *Molecular & Cellular Proteomics*, vol. 3, no. 9, 2004, pp. 908-919.
89. F. Aicheler et al., Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches, *Anal Chem*, vol. 87, no. 15, 2015, pp. 7698-7704.
90. V. I. Babushok and I. G. Zenkevich, Retention Characteristics of Peptides in RP-LC: Peptide Retention Prediction, *Chromatographia*, vol. 72, no. 9-10, 2010, pp. 781-797.

91. P. J. Eugster et al., Retention time prediction for dereplication of natural products (CxHyOz) in LC-MS metabolite profiling, *Phytochemistry*, vol. 108, no., 2014, pp. 196-207.
92. J. B. Golubović, A. D. Protić, M. L. Zečević, and B. M. Otašević, Quantitative structure retention relationship modeling in liquid chromatography method for separation of candesartan cilexetil and its degradation products, *Chemometrics and Intelligent Laboratory Systems*, vol. 140, no., 2015, pp. 92-101.
93. T. H. Miller, A. Musenga, D. A. Cowan, and L. P. Barron, Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks, *Anal Chem*, vol. 85, no. 21, 2013, pp. 10330-10337.
94. K. Munro et al., Artificial neural network modelling of pharmaceutical residue retention times in wastewater extracts using gradient liquid chromatography-high resolution mass spectrometry data, *J Chromatogr A*, vol. 1396, no., 2015, pp. 34-44.
95. F. Ruggiu et al., Quantitative structure-property relationship modeling: a valuable support in high-throughput screening quality control, *Anal Chem*, vol. 86, no. 5, 2014, pp. 2510-2520.
96. E. Tyrkko, A. Pelander, and I. Ojanpera, Prediction of liquid chromatographic retention for differentiation of structural isomers, *Anal Chim Acta*, vol. 720, no., 2012, pp. 142-148.
97. A. M. Wolfer et al., UPLC–MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling, *Metabolomics*, vol. 12, no. 1, 2015.
98. F. Falchi et al., Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification, *Anal Chem*, vol. 88, no. 19, 2016, pp. 9510-9517.
99. K. Goryński et al., Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds, *Analytica Chimica Acta*, vol. 797, no., 2013, pp. 13-19.
100. J. J. P. Stewart, *MOPAC2016*™. 2016.
101. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, *Journal of the American Chemical Society*, vol. 107, no. 13, 1985, pp. 3902-3909.
102. W. Thiel, Semiempirical quantum–chemical methods, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 2, 2014, pp. 145-157.
103. R. Todeschini, V. Consonni, A. Mauri, and M. Pavan, *Talete srl, DRAGON*, in *software for molecular descriptors calculation* 2007: Milan, Italy.
104. Partitioning(logD), *Marvin 6.3.1*. 2014, ChemAxon (http://www.chemaxon.com)"
105. B. J. Frey and D. Dueck, Clustering by Passing Messages Between Data Points, *Science*, vol. 315, no. 5814, 2007, pp. 972-976.
106. Mathworks. *Genetic algorithm and direct search toolbox users guide*. 2005.
107. J.-H. Lii et al., Molecular mechanics (MM2) calculations on peptides and on the protein Crambin using the CYBER 205, *Journal of Computational Chemistry*, vol. 10, no. 4, 1989, pp. 503-513.

108. I. Mitra, A. Saha, and K. Roy, Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants, *Molecular Simulation*, vol. 36, no. 13, 2010, pp. 1067-1079.
109. A. Tropsha, P. Gramatica, and V. K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR & Combinatorial Science*, vol. 22, no. 1, 2003, pp. 69-77.
110. N. Chirico and P. Gramatica, Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, *J Chem Inf Model*, vol. 51, no. 9, 2011, pp. 2320-2335.
111. N. Chirico and P. Gramatica, Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection, *J Chem Inf Model*, vol. 52, no. 8, 2012, pp. 2044-2058.
112. O. f. E. C.-o. a. Development and (OECD). *Guidance document on the validation of (quantitative)structure-activity relationship [(Q)SAR] models, 2007. OECD Web Site. http://www.oecd.org/officialdocuments/displaydocumentpdf/ ?cote=env/jm/mono%282007%292&doclanguage=en (accessed June 14, 2018)*.
113. A. Golbraikh and A. Tropsha, Beware of q2!, *Journal of Molecular Graphics and Modelling*, vol. 20, no. 4, 2002, pp. 269-276.
114. T. I. Netzeva et al., Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52, *Altern Lab Anim*, vol. 33, no. 2, 2005, pp. 155-173.
115. D.-S. Cao et al., A new strategy of outlier detection for QSAR/QSPR, *Journal of Computational Chemistry*, vol. 31, no. 3, 2010, pp. 592-602.
116. C. Christophoridis, M.-C. Nika, R. Aalizadeh, and N. S. Thomaidis, Ozonation of ranitidine: Effect of experimental parameters and identification of transformation products, *Science of The Total Environment*, vol. 557–558, no., 2016, pp. 170-182.
117. M. E. Dasenaki and N. S. Thomaidis, Multianalyte method for the determination of pharmaceuticals in wastewater samples using solid-phase extraction and liquid chromatography–tandem mass spectrometry, *Analytical and Bioanalytical Chemistry*, vol. 407, no. 15, 2015, pp. 4229-4245.
118. N. A. Alygizakis et al., Occurrence and spatial distribution of 158 pharmaceuticals, drugs of abuse and related metabolites in offshore seawater, *Science of The Total Environment*, vol. 541, no., 2016, pp. 1097-1105.
119. D. E. Damalas et al., Assessment of the Acute Toxicity, Uptake and Biotransformation Potential of Benzotriazoles in Zebrafish (Danio rerio) Larvae Combining HILIC- with RPLC-HRMS for High-Throughput Identification, *Environmental Science & Technology*, vol. 52, no. 10, 2018, pp. 6023-6031.
120. A. K. Ghose, V. N. Viswanadhan, and J. J. Wendoloski, Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods, *The Journal of Physical Chemistry A*, vol. 102, no. 21, 1998, pp. 3762-3772.
121. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, New York, Wiley-VCH, 2008.

122. P. G. Boswell et al., Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles, *Journal of Chromatography A*, vol. 1218, no. 38, 2011, pp. 6742-6749.
123. D. Abate-Pella et al., Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods, *Journal of Chromatography A*, vol. 1412, no., 2015, pp. 43-51.
124. E. L. Schymanski et al., Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry, *Environmental Science & Technology*, vol. 48, no. 3, 2014, pp. 1811-1818.
125. E. L. Schymanski et al., Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence, *Environmental Science & Technology*, vol. 48, no. 4, 2014, pp. 2097-2098.
126. J. Guo et al., Extended Virtual Screening Strategies To Link Antiandrogenic Activities and Detected Organic Contaminants in Soils, *Environmental Science & Technology*, vol. 51, no. 21, 2017, pp. 12528-12536.
127. D. Abate-Pella et al., Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods, *Journal of Chromatography A*, vol. 1412, no. Supplement C, 2015, pp. 43-51.
128. P. G. Boswell et al., A study on retention “projection” as a supplementary means for compound identification by liquid chromatography–mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments, *Journal of Chromatography A*, vol. 1218, no. 38, 2011, pp. 6732-6741.
129. R. Bade, L. Bijlsma, J. V. Sancho, and F. Hernandez, Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water, *Talanta*, vol. 139, no., 2015, pp. 143-149.
130. A. Cherkasov et al., QSAR Modeling: Where Have You Been? Where Are You Going To?, *Journal of Medicinal Chemistry*, vol. 57, no. 12, 2014, pp. 4977-5010.
131. C. W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *Journal of Computational Chemistry*, vol. 32, no. 7, 2011, pp. 1466-1474.
132. I. V. Tetko et al., Virtual Computational Chemistry Laboratory – Design and Description, *Journal of Computer-Aided Molecular Design*, vol. 19, no. 6, 2005, pp. 453-463.
133. . 2014, Partitioning(logD) Marvin 6.3.1, ChemAxon, http://www.chemaxon.com.
134. L. Xing and R. C. Glen, Novel Methods for the Prediction of logP, pKa, and logD, *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 4, 2002, pp. 796-805.
135. M. J. Vainio and M. S. Johnson, Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm, *Journal of Chemical Information and Modeling*, vol. 47, no. 6, 2007, pp. 2462-2474.
136. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt.
137. V. Consonni, R. Todeschini, M. Pavan, and P. Gramatica, Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies, *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 3, 2002, pp. 693-705.

138. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, *Handbook of Molecular Descriptors*, eds, Wiley-VCH Verlag GmbH, 2008, pp. 1-523.
139. R. Todeschini and P. Gramatica, New 3D Molecular Descriptors: The WHIM theory and QSAR Applications, *3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity*, Dordrecht, H. Kubinyi, G. Folkers and Y. C. Martin, eds, Springer Netherlands, 1998, pp. 355-380.
140. P. Žuvela, J. J. Liu, K. Macur, and T. Bączek, Molecular Descriptor Subset Selection in Theoretical Peptide Quantitative Structure–Retention Relationship Model Development Using Nature-Inspired Optimization Algorithms, *Analytical Chemistry*, vol. 87, no. 19, 2015, pp. 9876-9883.
141. M. Dorigo, M. Birattari, and T. Stützle, Ant colony optimization artificial ants as a computational intelligence technique, *IEEE Computational Intelligence Magazine*, vol. 1, no. 4, 2006, pp. 28-39.
142. M. Dorigo and C. Blum, Ant colony optimization theory: A survey, *Theoretical Computer Science*, vol. 344, no. 2-3, 2005, pp. 243-278.
143. Y. Moulard et al., Use of benchtop exactive high resolution and high mass accuracy orbitrap mass spectrometer for screening in horse doping control, *Analytica Chimica Acta*, vol. 700, no. 1, 2011, pp. 126-136.
144. Y. Dodge, Least Significant Difference Test, *The Concise Encyclopedia of Statistics*, New York, NYeds, Springer New York, 2008, pp. 302-304.
145. J. n. Pizarro, E. Guerrero, and P. L. Galindo, Multiple comparison procedures applied to model selection, *Neurocomputing*, vol. 48, no. 1, 2002, pp. 155-173.
146. C. Hartmann et al., Reappraisal of Hypothesis Testing for Method Validation: Detection of Systematic Error by Comparing the Means of Two Methods or of Two Laboratories, *Analytical Chemistry*, vol. 67, no. 24, 1995, pp. 4491-4499.
147. D. T. Stanton and P. C. Jurs, Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies, *Analytical Chemistry*, vol. 62, no. 21, 1990, pp. 2323-2329.
148. P. Rostkowski et al., The Strength in Numbers: Comprehensive Characterization of House Dust using Complementary Mass Spectrometric Techniques, *Analytical and Bioanalytical Chemistry*, no., 2018, pp. Under Review.
149. P. C. von der Ohe et al., A new risk assessment approach for the prioritization of 500 classical and emerging organic microcontaminants as potential river basin specific pollutants under the European Water Framework Directive, *Science of the Total Environment*, vol. 409, no. 11, 2011, pp. 2064-2077.
150. J. Slobodnik et al., Identification of river basin specific pollutants and derivation of environmental quality standards: A case study in the Slovak Republic, *TrAC Trends in Analytical Chemistry*, vol. 41, no., 2012, pp. 133-145.
151. A. Sangion and P. Gramatica, Ecotoxicity interspecies QAAR models from Daphnia toxicity of pharmaceuticals and personal care products, *SAR and QSAR in Environmental Research*, vol. 27, no. 10, 2016, pp. 781-798.
152. A. Sangion and P. Gramatica, Hazard of pharmaceuticals for aquatic environment: Prioritization by structural approaches and prediction of ecotoxicity, *Environment International*, vol. 95, no., 2016, pp. 131-143.

153. S. Kar and K. Roy, First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals, *Chemosphere*, vol. 81, no. 6, 2010, pp. 738-747.
154. S. Cassani et al., Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling, *Journal of Hazardous Materials*, vol. 258–259, no., 2013, pp. 50-60.
155. M. Cassotti et al., Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN method, *ATLA Alternatives to Laboratory Animals*, vol. 42, no. 1, 2014, pp. 31-41.
156. R. Kühne et al., Read-Across Prediction of the Acute Toxicity of Organic Compounds toward the Water Flea Daphnia magna, *Molecular Informatics*, vol. 32, no. 1, 2013, pp. 108-120.
157. T. Martin. *Toxicity Estimation Software Tool (TEST)*. 2016 [cited 2016 24/09]; version 4.2:[
158. H. Sanderson and M. Thomsen, Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action, *Toxicology Letters*, vol. 187, no. 2, 2009, pp. 84-93.
159. K. Roy and G. Ghosh, Exploring QSARs with Extended Topochemical Atom (ETA) Indices for Modeling Chemical and Drug Toxicity, *Current Pharmaceutical Design*, vol. 16, no. 24, 2010, pp. 2625-2639.
160. M. Cassotti, V. Consonni, A. Mauri, and D. Ballabio, Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards Daphnia magna, *SAR and QSAR in Environmental Research*, vol. 25, no. 12, 2014, pp. 1013-1036.
161. A. Böhme, A. Laqua, and G. Schüürmann, Chemoavailability of Organic Electrophiles: Impact of Hydrophobicity and Reactivity on Their Aquatic Excess Toxicity, *Chemical Research in Toxicology*, vol. 29, no. 6, 2016, pp. 952-962.
162. J. A. Platts, D. Butina, M. H. Abraham, and A. Hersey, Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach, *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 5, 1999, pp. 835-845.
163. C. D. John, The History and Development of Quantitative Structure-Activity Relationships (QSARs), *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, vol. 1, no. 1, 2016, pp. 1-44.
164. K. Roy, S. Kar, and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, 2015.
165. G. Schüürmann, R.-U. Ebert, and R. Kühne, Quantitative Read-Across for Predicting the Acute Fish Toxicity of Organic Compounds, *Environmental Science & Technology*, vol. 45, no. 10, 2011, pp. 4616-4622.
166. I. Sushko et al., Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, *Journal of Computer-Aided Molecular Design*, vol. 25, no. 6, 2011, pp. 533-554.
167. E. L. Schymanski et al., The NORMAN Suspect List Exchange: Facilitating European Collaboration on Suspect Screening, *Environmental Health Perspectives*, no., 2016.

168. N. M. O'Boyle et al., Open Babel: An open chemical toolbox, *Journal of Cheminformatics*, vol. 3, no. 1, 2011, pp. 33.
169. P. C. Von Der Ohe et al., Structural alerts - A new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay, *Chemical Research in Toxicology*, vol. 18, no. 3, 2005, pp. 536-555.
170. R. W. Kennard and L. A. Stone, Computer Aided Design of Experiments, *Technometrics*, vol. 11, no. 1, 1969, pp. 137-148.
171. H. C. Maureen B. Tracy et al., Precision enhancement of MALDI-TOF MS using high resolution peak detection and label-free alignment, *Proteomics*, vol. 8, no., 2008, pp. 1530–1538.
172. M. Shahlaei, Descriptor selection methods in quantitative structure-activity relationship studies: a review study, *Chemical reviews*, vol. 113, no. 10, 2013, pp. 8093-8103.
173. E. Pourbasheer et al., Prediction of PCE of fullerene (C60) derivatives as polymer solar cell acceptors by genetic algorithm–multiple linear regression, *Journal of Industrial and Engineering Chemistry*, vol. 21, no., 2015, pp. 1058-1067.
174. M. Goodarzi, M. P. Freitas, and R. Jensen, Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)uracil derivatives using MLR, PLS and SVM regressions, *Chemometrics and Intelligent Laboratory Systems*, vol. 98, no. 2, 2009, pp. 123-129.
175. V. N. Vapnik, Methods of Function Esthnation, *The Nature of Statistical Learning Theory*, New York, NYeds, Springer New York, 2000, pp. 181-216.
176. S. Weaver and M. P. Gleeson, The importance of the domain of applicability in QSAR modeling, *Journal of Molecular Graphics and Modelling*, vol. 26, no. 8, 2008, pp. 1315-1326.
177. G. Domenico et al., Applicability Domain for QSAR Models: Where Theory Meets Reality, *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, vol. 1, no. 1, 2016, pp. 45-63.
178. K. Roy, S. Kar, and P. Ambure, On a simple approach for determining applicability domain of QSAR models, *Chemometrics and Intelligent Laboratory Systems*, vol. 145, no., 2015, pp. 22-29.
179. T. I. Netzeva et al., Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships, *Alternatives to Laboratory Animals*, vol. 33, no. 2, 2005, pp. 1-19.
180. H. Golmohammadi, Z. Dashtbozorgi, and W. E. Acree Jr, Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine, *European Journal of Pharmaceutical Sciences*, vol. 47, no. 2, 2012, pp. 421-429.
181. R. Wang, Y. Fu, and L. Lai, A New Atom-Additive Method for Calculating Partition Coefficients, *Journal of Chemical Information and Computer Sciences*, vol. 37, no. 3, 1997, pp. 615-621.
182. S. A. Wildman and G. M. Crippen, Prediction of Physicochemical Parameters by Atomic Contributions, *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 5, 1999, pp. 868-873.

183. R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, ed. R. Mannhold, H. Kubinyi and G. Folkers, New York, Wiley-VCH, 2009, p. 1257.
184. L. H. Hall and L. B. Kier, Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information, *Journal of Chemical Information and Computer Sciences*, vol. 35, no. 6, 1995, pp. 1039-1045.
185. A. P. Toropova et al., Monte Carlo–based quantitative structure–activity relationship models for toxicity of organic chemicals to Daphnia magna, *Environmental Toxicology and Chemistry*, vol. 35, no. 11, 2016, pp. 2691-2697.
186. M. Cassotti et al., Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN method, *Altern Lab Anim*, vol. 42, no. 1, 2014, pp. 31-41.
187. M. Moosus and U. Maran, Quantitative structure–activity relationship analysis of acute toxicity of diverse chemicals to Daphnia magna with whole molecule descriptors, *SAR and QSAR in Environmental Research*, vol. 22, no. 7-8, 2011, pp. 757-774.
188. Reenu and Vikas, Exploring the role of quantum chemical descriptors in modeling acute toxicity of diverse chemicals to Daphnia magna, *Journal of Molecular Graphics and Modelling*, vol. 61, no., 2015, pp. 89-101.
189. A. P. Toropova, A. A. Toropov, E. Benfenati, and G. Gini, QSAR Models for Toxicity of Organic Substances to Daphnia magna Built up by Using the CORAL Freeware, *Chemical Biology & Drug Design*, vol. 79, no. 3, 2012, pp. 332-338.
190. A. P. Toropova et al., CORAL: QSAR modeling of toxicity of organic chemicals towards Daphnia magna, *Chemometrics and Intelligent Laboratory Systems*, vol. 110, no. 1, 2012, pp. 177-181.
191. A. D. McEachran et al., “MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies, *Journal of Cheminformatics*, vol. 10, no. 1, 2018, pp. 45.
192. M. Vinaixa et al., Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects, *TrAC Trends in Analytical Chemistry*, vol. 78, no., 2016, pp. 23-35.
193. E. L. Schymanski and A. J. Williams, Open Science for Identifying “Known Unknown” Chemicals, *Environmental Science & Technology*, vol. 51, no. 10, 2017, pp. 5357-5359.
194. M. A. Stravs, E. L. Schymanski, H. P. Singer, and J. Hollender, Automatic recalibration and processing of tandem mass spectra using formula annotation, *Journal of Mass Spectrometry*, vol. 48, no. 1, 2013, pp. 89-99.
195. H. Oberacher et al., Annotating Nontargeted LC-HRMS/MS Data with Two Complementary Tandem Mass Spectral Libraries, *Metabolites*, vol. 9, no. 1, 2018, pp. 3.
196. F. Qiu et al., PlantMAT: A Metabolomics Tool for Predicting the Specialized Metabolic Potential of a System and for Large-Scale Metabolite Identifications, *Analytical Chemistry*, vol. 88, no. 23, 2016, pp. 11373-11383.
197. C. A. Smith et al., XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification, *Analytical Chemistry*, vol. 78, no. 3, 2006, pp. 779-787.
198. C. C. Grigsby et al., Metabolite Differentiation and Discovery Lab (MeDDL): A New Tool for Biomarker Discovery and Mass Spectral Visualization, *Analytical Chemistry*, vol. 82, no. 11, 2010, pp. 4386-4395.

199. T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, *BMC Bioinformatics*, vol. 11, no. 1, 2010, pp. 395.
200. R. Tautenhahn et al., metaXCMS: Second-Order Analysis of Untargeted Metabolomics Data, *Analytical Chemistry*, vol. 83, no. 3, 2011, pp. 696-700.
201. M. Eliasson et al., Strategy for Optimizing LC-MS Data Processing in Metabolomics: A Design of Experiments Approach, *Analytical Chemistry*, vol. 84, no. 15, 2012, pp. 6869-6876.
202. R. Tautenhahn, G. J. Patti, D. Rinehart, and G. Siuzdak, XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data, *Analytical Chemistry*, vol. 84, no. 11, 2012, pp. 5035-5039.
203. S.-Y. Wang, C.-H. Kuo, and Y. J. Tseng, Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Methods, *Analytical Chemistry*, vol. 85, no. 2, 2013, pp. 1037-1046.
204. H. Gowda et al., Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses, *Analytical Chemistry*, vol. 86, no. 14, 2014, pp. 6931-6939.
205. E. Tengstrand, J. Lindberg, and K. M. Åberg, TracMass 2—A Modular Suite of Tools for Processing Chromatography-Full Scan Mass Spectrometry Data, *Analytical Chemistry*, vol. 86, no. 7, 2014, pp. 3435-3442.
206. H. Tsugawa et al., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, *Nature Methods*, vol. 12, no., 2015, pp. 523.
207. W. Zhang et al., MET-COFEA: A Liquid Chromatography/Mass Spectrometry Data Processing Platform for Metabolite Compound Feature Extraction and Annotation, *Analytical Chemistry*, vol. 86, no. 13, 2014, pp. 6245-6253.
208. M. Woldegebriel and G. Vivó-Truyols, Probabilistic Model for Untargeted Peak Detection in LC–MS Using Bayesian Statistics, *Analytical Chemistry*, vol. 87, no. 14, 2015, pp. 7345-7355.
209. M. Zhang, J. Sun, and P. Chen, FlavonQ: An Automated Data Processing Tool for Profiling Flavone and Flavonol Glycosides with Ultra-High-Performance Liquid Chromatography–Diode Array Detection–High Resolution Accurate Mass–Mass Spectrometry, *Analytical Chemistry*, vol. 87, no. 19, 2015, pp. 9974-9981.
210. W. Zhang et al., MET-XAlign: A Metabolite Cross-Alignment Tool for LC/MS-Based Comparative Metabolomics, *Analytical Chemistry*, vol. 87, no. 18, 2015, pp. 9114-9119.
211. D. M. Avtonomov, A. Raskind, and A. I. Nesvizhskii, BatMass: a Java Software Platform for LC–MS Data Visualization in Proteomics and Metabolomics, *Journal of Proteome Research*, vol. 15, no. 8, 2016, pp. 2500-2509.
212. J. Capellades et al., geoRge: A Computational Tool To Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics, *Analytical Chemistry*, vol. 88, no. 1, 2016, pp. 621-628.

213. E. Gorrochategui, J. Jaumot, S. Lacorte, and R. Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC Trends in Analytical Chemistry*, vol. 82, no., 2016, pp. 425-442.
214. Y.-J. Liang et al., SMART: Statistical Metabolomics Analysis—An R Tool, *Analytical Chemistry*, vol. 88, no. 12, 2016, pp. 6334-6341.
215. H.-Y. Fu et al., AntDAS: Automatic Data Analysis Strategy for UPLC–QTOF-Based Nontargeted Metabolic Profiling Analysis, *Analytical Chemistry*, vol. 89, no. 20, 2017, pp. 11083-11090.
216. O. D. Myers et al., One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks, *Analytical Chemistry*, vol. 89, no. 17, 2017, pp. 8696-8703.
217. H. Ji et al., KPIC2: An Effective Framework for Mass Spectrometry-Based Metabolomics Using Pure Ion Chromatograms, *Analytical Chemistry*, vol. 89, no. 14, 2017, pp. 7631-7640.
218. P. Gago-Ferrero, V. Borova, M. E. Dasenaki, and N. S. Thomaidis, Simultaneous determination of 148 pharmaceuticals and illicit drugs in sewage sludge based on ultrasound-assisted extraction and liquid chromatography–tandem mass spectrometry, *Analytical and Bioanalytical Chemistry*, vol. 407, no. 15, 2015, pp. 4287-4297.
219. M. Loos, M. Ruff, H. Singer, and J. Hollender, *Clustering-based ion chromatogram extraction and peak-picking for high-resolution LC-MS data*, in *International Mass Spectrometry Conference*. 2014: Geneva, Switzerland.
220. R. Tautenhahn, C. Böttcher, and S. Neumann, Highly sensitive feature detection for high resolution LC/MS, *BMC Bioinformatics*, vol. 9, no. 1, 2008, pp. 504.
221. M. C. Chambers et al., A cross-platform toolkit for mass spectrometry and proteomics, *Nature Biotechnology*, vol. 30, no., 2012, pp. 918.
222. G. Libiseller et al., IPO: a tool for automated optimization of XCMS parameters, *BMC Bioinformatics*, vol. 16, no. 1, 2015, pp. 118.
223. J. T. Prince and E. M. Marcotte, Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping, *Analytical Chemistry*, vol. 78, no. 17, 2006, pp. 6140-6152.
224. S. L. C. Ferreira et al., Box-Behnken design: An alternative for the optimization of analytical methods, *Analytica Chimica Acta*, vol. 597, no. 2, 2007, pp. 179-186.
225. B. C. DeFelice et al., Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography–Mass Spectroscopy (LC-MS) Data Processing, *Analytical Chemistry*, vol. 89, no. 6, 2017, pp. 3250-3255.
226. R. A. a. N. S. T. Polykarpos Beikos, *Minimizing Analytical Procedural Mass Spectral Features as False Positive Peaks in Untargeted Liquid Chromatography – High Resolution Mass Spectrometry Data Processing*, in *11th Aegean Analytical Chemistry Days (AACD2018)*. 2018: Chania, Crete, Greece.
227. N. A. Alygizakis, P. Gago-Ferrero, J. Hollender, and N. S. Thomaidis, Untargeted time-pattern analysis of LC-HRMS data to detect spills and compounds with high fluctuation in influent wastewater, *Journal of Hazardous Materials*, vol. 361, no., 2019, pp. 19-29.

228. Davide Ballabio and Viviana Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, *Analytical Methods*, vol. 5, no., 2013, pp. 3790.
229. M. M. Plassmann, E. Tengstrand, K. M. Åberg, and J. P. Benskin, Non-target time trend screening: a data reduction strategy for detecting emerging contaminants in biological samples, *Analytical and Bioanalytical Chemistry*, vol. 408, no. 16, 2016, pp. 4203-4208.
230. R. A. a. N. S. T. Maria-Christina Nika, Removal and transformation of citalopram and four of its biotransformation products during ozonation experiments, *Water Research*, no., 2019, pp. Under Review.
231. M. Hur et al., A global approach to analysis and interpretation of metabolic data for plant natural product discovery, *Natural product reports*, vol. 30, no. 4, 2013, pp. 565-583.
232. N. Kumar, M. A. Hoque, and M. Sugimoto, Robust volcano plot: identification of differential metabolites in the presence of outliers, *BMC bioinformatics*, vol. 19, no. 1, 2018, pp. 128-128.
233. A. Bajoub et al., Comparing two metabolic profiling approaches (liquid chromatography and gas chromatography coupled to mass spectrometry) for extra-virgin olive oil phenolic compounds analysis: A botanical classification perspective, *Journal of Chromatography A*, vol. 1428, no., 2016, pp. 267-279.
234. J. Trygg and T. Lundstedt, Chapter 6 - Chemometrics Techniques for Metabonomics, *The Handbook of Metabonomics and Metabolomics*, Amsterdam, J. C. Lindon, J. K. Nicholson and E. Holmes, eds, Elsevier Science B.V., 2007, pp. 171-199.
235. A. Bajoub et al., Potential of LC–MS phenolic profiling combined with multivariate analysis as an approach for the determination of the geographical origin of north Moroccan virgin olive oils, *Food Chemistry*, vol. 166, no., 2015, pp. 292-300.
236. R. García-Villalba et al., Characterization and quantification of phenolic compounds of extra-virgin olive oils with anticancer properties by a rapid and resolutive LC-ESI-TOF MS method, *Journal of Pharmaceutical and Biomedical Analysis*, vol. 51, no. 2, 2010, pp. 416-429.
237. S. Ben Brahim et al., LC–MS phenolic profiling combined with multivariate analysis as an approach for the characterization of extra virgin olive oils of four rare Tunisian cultivars during ripening, *Food Chemistry*, vol. 229, no., 2017, pp. 9-19.
238. S. Ammar et al., LC-DAD/ESI-MS/MS characterization of phenolic constituents in Tunisian extra-virgin olive oils: Effect of olive leaves addition on chemical composition, *Food Research International*, vol. 100, no., 2017, pp. 477-485.
239. L. Olmo-García et al., Development and validation of LC-MS-based alternative methodologies to GC–MS for the simultaneous determination of triterpenic acids and dialcohols in virgin olive oil, *Food Chemistry*, vol. 239, no., 2018, pp. 631-639.
240. D. Ballabio and V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA., *Anal. Methods*, vol. 5, no. 16, 2013, pp. 2.
241. T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, *Chemometrics and Intelligent Laboratory Systems*, vol. 118, no., 2012, pp. 62-69.
242. I.-G. Chong and C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, vol. 78, no. 1-2, 2005, pp. 103-112.

243. J. Trygg and S. Wold, Orthogonal projections to latent structures (O-PLS), *Journal of Chemometrics*, vol. 16, no. 3, 2002, pp. 119-128.
244. S. Wold, H. Antti, F. Lindgren, and J. Öhman, Orthogonal signal correction of near-infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, vol. 44, no. 1, 1998, pp. 175-185.
245. A. L. a. M. Wiener, Classification and Regression by randomForest, *R News*, vol. 2, no., 2002, pp. 18-22.
246. E. C. A. (ECHA). *Biocidal Active Substances, Regulation (EU) No 528/2012*. [cited 2018 01/June/2018]; Available from: https://echa.europa.eu/information-on-chemicals/biocidal-active-substances.
247. E. C. H. F. S. Directorate-General. *Safety of the food chain, Pesticides and Biocides, Draft Working Document Air Iii Renewal Programme, SANCO/2012/11284 –rev. 21* 2018 [cited 2018 01/July/2018]; Available from: https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_air-3_sanco-2012-11284.pdf.
248. AccuStandard. *Biocide Standards Reference Guide*. 2018 01/July/2018]; Available from: https://www.accustandard.com/assets/BIOCIDE_GUIDE.pdf.
249. P. A. Network. *PAN Europe Study of Pesticide and Biocide Contamination of Fruit and Vegetables in Four EU Member States*. 2009 [cited 2018 01/July/2018]; Available from: https://www.pan-europe.info/old/Resources/Other/Pesticide_and_Biocide_Contamination_of_Fruit_and_Vegetables_results.pdf.
250. W.-R. Liu et al., Biocides in wastewater treatment plants: Mass balance analysis and pollution load estimation, *Journal of Hazardous Materials*, vol. 329, no., 2017, pp. 310-320.
251. M. Ruff, M. S. Mueller, M. Loos, and H. P. Singer, Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry – Identification of unknown sources and compounds, *Water Research*, vol. 87, no., 2015, pp. 145-154.
252. T. Ruan et al., Identification and Composition of Emerging Quaternary Ammonium Compounds in Municipal Sewage Sludge in China, *Environmental Science & Technology*, vol. 48, no. 8, 2014, pp. 4289-4297.
253. MoNA. *MassBank of North America (MoNA)*. 15/July/2018]; Available from: http://mona.fiehnlab.ucdavis.edu/.
254. L. Patiny and A. Borel, ChemCalc: A Building Block for Tomorrow's Chemical Infrastructure, *Journal of Chemical Information and Modeling*, vol. 53, no. 5, 2013, pp. 1223-1228.
255. T. Pluskal, T. Uehara, and M. Yanagida, Highly Accurate Chemical Formula Prediction Tool Utilizing High-Resolution Mass Spectra, MS/MS Fragmentation, Heuristic Rules, and Isotope Pattern Matching, *Analytical Chemistry*, vol. 84, no. 10, 2012, pp. 4396-4403.
256. A. Drefahl, CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures, *Journal of cheminformatics*, vol. 3, no. 1, 2011, pp. 1-1.

257. G. M. Pesyna, R. Venkataraghavan, H. E. Dayringer, and F. W. McLafferty, Probability based matching system using a large collection of reference mass spectra, *Analytical Chemistry*, vol. 48, no. 9, 1976, pp. 1362-1368.
258. S. E. Stein and D. R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 9, 1994, pp. 859-866.
259. K. X. Wan, I. Vidavsky, and M. L. Gross, Comparing similar spectra: from similarity index to spectral contrast angle, *Journal of the American Society for Mass Spectrometry*, vol. 13, no. 1, 2002, pp. 85-88.
260. Z. B. Alfassi, On the normalization of a mass spectrum for comparison of two spectra, *Journal of the American Society for Mass Spectrometry*, vol. 15, no. 3, 2004, pp. 385-387.
261. K. Mansouri, C. M. Grulke, R. S. Judson, and A. J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, *Journal of Cheminformatics*, vol. 10, no. 1, 2018, pp. 10.
262. A. J. Williams et al., The CompTox Chemistry Dashboard: a community data resource for environmental chemistry, *Journal of Cheminformatics*, vol. 9, no. 1, 2017, pp. 61.
263. Reza Aalizadeh, Emma L. Schymanski, and N. S.Thomaidis, *AutoSuspect: an R package to Perform Automatic Suspect Screening based on Regulatory Databases*, in *15th International Conference on Environmental Science and Technology*. 2017: Rhodes, Greece.
264. N. S.Thomaidis, *Target, suspect and non-target HRMS screening approaches for food authenticity and quality: from research to industrial applications in: 4th European AMS Workshop: Ambient Mass Spectrometry on Food And Natural Products, Session 16, Prague, Czech Republic, 7-10 November 2017*. 2017.
265. I. Ferrer and E. M. Thurman, Analysis of hydraulic fracturing additives by LC/Q-TOF-MS, *Analytical and Bioanalytical Chemistry*, vol. 407, no. 21, 2015, pp. 6417-6428.

Supplementary Material

Chapter 3

Appendix A. Supplementary data

Appendix A provides a document file which describes the development and validation of all the QSRR models and additional data and graphics (**Figure A 3.1-A 3.14**).

Appendix B. Supplementary data

Appendix B provides an Excel file complement to this section which consists all additional Tables (**Tables B 3.1-B 3.14**).

Chapter 4

Appendix A. Supplementary data

Appendix A provides a document file which describes the development and validation of retention time indices as well as LC conditions and additional data and graphics (**Figure A 4.1-A 4.3**).

Appendix B. Supplementary data

Appendix B provides an Excel file complement to this section which consists all additional Tables (**Tables B 4.1-B 4.15**).

Chapter 5

Appendix A. Supplementary data

Appendix B provides an Excel file complement to this section which consists all additional Tables (**Tables A 5.1-A 5.4**).

Chapter 6

Appendix A provides an Excel file complement to this section which consists all additional Tables (**Tables A 6.1-A 6.4**).