Journal of Hazardous Materials 363 (2019) 277–285

Contents lists available at ScienceDirect

Journal of Hazardous Materials

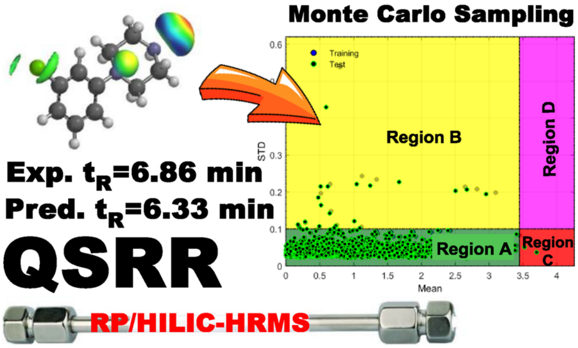
journal homepage: www.elsevier.com/locate/jhazmat

Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants

Reza Aalizadeh, Maria-Christina Nika, Nikolaos S. Thomaidis*

Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zographou, 15771, Athens, Greece

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
LC-HRMS
QSRR
Suspect and non-target screening
Transformation products
Biocides

ABSTRACT

Hydrophilic interaction liquid chromatography (HILIC) and reversed phase LC (RPLC) coupled to high resolution mass spectrometry (HRMS) are widely used for the identification of suspects and unknown compounds in the environment. For the identification of unknowns, apart from mass accuracy and isotopic fitting, retention time (t_R) and MS/MS spectra evaluation is required. In this context, a novel comprehensive workflow was developed to study the t_R behavior of large groups of emerging contaminants using Quantitative Structure-Retention Relationships (QSRR). 682 compounds were analyzed by HILIC-HRMS in positive Electrospray Ionization mode (ESI). Moreover, an extensive dataset was built for RPLC-HRMS including 1830 and 308 compounds for positive and negative ESI, respectively. Support Vector Machines (SVM) was used to model the t_R 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 t_R from various LC conditions. This paper provides validated models for predicting t_R in HILIC/RPLC-HRMS platforms to facilitate identification of new emerging contaminants by suspect and non-target HRMS screening, and were applied for the identification of transformation products (TPs) of emerging contaminants and biocides in wastewater and sludge.

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 [1,2]. 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 [3–5]. Schymanski et al. proposed a

* Corresponding author.
E-mail address: ntho@chem.uoa.gr (N.S. Thomaidis).

https://doi.org/10.1016/j.jhazmat.2018.09.047
Received 2 August 2018; Received in revised form 16 September 2018; Accepted 17 September 2018
Available online 18 September 2018
0304-3894/ © 2018 Elsevier B.V. All rights reserved.

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 [5]. Knowledge of t_R can also help reduce the number of plausible candidates and, subsequently, increase the chance of true identification [6,7]. Since the polar micropollutants and their transformation products (TPs) are the major focus in the aquatic environment [3], the complimentary use of hydrophilic interaction liquid chromatography (HILIC) with reversed phase liquid chromatography (RPLC) can provide additional experimental evidence and support to the identification of new compounds in the environment [6]. 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 [3,8]. In those cases, retention time prediction could support identification.

Several approaches have been presented to predict t_R in LC [9–20]. 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 [10–20]. 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 [19]. Apart from previous studies which have a limited applicability domain or showed high prediction errors, Falchi et al. [21] 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 [18,20–22]. Although the origin of error between the experimental and predicted t_R was investigated in a few studies [18,20,23], 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 [24]. 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 [24].

Although the use of HILIC is increasing as a complementary method to cover highly polar compounds and metabolites [6], few studies have reported modeling HILIC t_R [13,22]. 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, [13] 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.* Quantitative Structure-Retention Relationships (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.

Thus, the objectives of the current work 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 Monte Carlo Sampling (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.

2. Materials and methods

2.1. Sample preparation, instrumental analysis and dataset

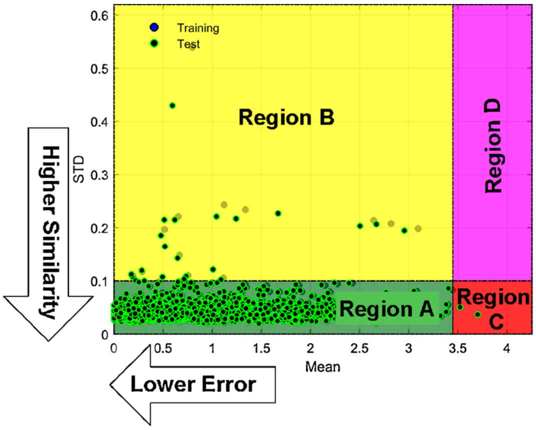
The reagents and solvents used to measure the t_R of reference standards in RPLC were explained previously [6,24]. All the details about the RPLC and HILIC separations and QToF-MS methods, the reagents and solvents can be found in supplementary material (SM, Appendix A), section SM 1.1 and SM 1.2. The lists of reference standards used for the modeling t_R are given in Tables B.1, B.2 and B.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 was also illustrated in the SM (Appendix A) Fig. A.1, section SM 1.3. There were insufficient compounds amenable to HILIC-(-)ESI to form a sufficiently large dataset for this work.

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/) [25]. The semi-empirical (AM1) [26,27] 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/) [28]. 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 [29]. 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 [30]. 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 (Fig. A.2–A.4). Details about affinity propagation and the heat maps for both RPLC and HILIC platforms are given in the SM (Appendix A), section SM 2.2. A genetic algorithm, written in MATLAB 8.5 [31], 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 Support Vector Machines (SVM), respectively. A brief introduction about the regression techniques used can be found in SM (Appendix A), SM 2.3 to SM 2.4. The accuracy of the models built to predict t_R was investigated using an external test set and cross-validation techniques. The full validation protocols and criteria are described in SM (sections SM 2.5) and in Table B.4.

2.3. Applicability domain studies and t_R 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.



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.

Scheme 1. The explanation of Monte-Carlo sampling (MCS) method used to define the applicability domain of the models developed to predict t_R.

Different methods are available to assess the presence of outliers [24,32]. A method called "OTrAMS" presented in [24] 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 [24]. 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 SM (Appendix A), section SM 2.6.

In addition to OTrAMS, another outlier detection procedure was developed using Monte-Carlo sampling method (MCS) [33] 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. Scheme 1 shows the interpretation of MCS results. MCS was set to 5000 iterations. As shown in Scheme 1, 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.

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 (Appendix B) Table B.5.

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 [34]. These compounds are often detected in effluents (incomplete removal) and the receiving environment [35–37] and data for their ozonation TPs is scarce. The identification workflow, along with RPLC/HILIC complementary analyses, is described in [34,38,39].

2.5. Screening of biocides in wastewater and sludge

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. [6], 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 [6,40]. The screening database, a complete list of biocides and pesticides (active ingredients), was compiled from regulatory databases [41–45]. Several other biocidal products such as Quaternary ammonium compounds (QACs) or disinfectants were collected from literature. [46–48]. 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 libraries [49–51]. This suspects list can be found in SM (Appendix B), Table B.6. The suspect screening workflow is described in detail in SM (Appendix A), section SM3.

3. Results and discussion

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:

$$\begin{aligned} t_R = {} & +2.3518(\pm 0.1335) + 0.7371(\pm 0.0204)\, logD_{(pH\ 3.60)} \\ & + 1.2389(\pm 0.0696)\, CIC1 + 0.5584(\pm 0.0396)\, SEigZ \\ & - 0.2198(\pm 0.0340)\, RDF020p + 0.4155(\pm 0.0306)\, AlogP \end{aligned} \tag{1}$$

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 [52]. More details about molecular descriptors selected can be found in supplementary materials (Appendix A), section SM 4.1.1. The elution of the compounds in RPLC-(+)ESI-HRMS is illustrated in Fig. A.5.

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 Appendix A (section SM 2.5) are listed in Table B.7 and the model meets all acceptance criteria. The OTrAMS results are shown in Fig. A.6 (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 [33] described above was applied, shown in Fig. A.6 (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), $\varepsilon = 0.2$ (insensitive loss function), $\gamma = 1.5$ (radial basis function (RBF)). These parameters are discussed in detail in SM (appendix A) section SM 2.4. The result of each optimization step is shown in Appendix A, section SM 4.1.3, Fig. A.7 (a–c). The predicted and experimental t_R values are listed in Table B.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.8). The molecular descriptors used here were investigated for the existence of inter-correlation cases and as listed in Table B.9, no cases with high inter-correlation were found.

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:

$$\begin{aligned} t_R = {} & +3.9078\,(\pm 0.2255) + 0.3016\,(\pm 0.0768) XlogP \\ & + 0.6128\,(\pm 0.0543) logD_{(pH\ 6.20)} + 0.1917\,(\pm 0.0543) RDF130m \\ & -1.377\,(\pm 0.3291) Mor16p - 0.3062\,(\pm 0.0695)\, nCconj \\ & + 0.1103\,(0.0178)\, MlogP^2 + 1.588\,(0.2461)\, B06[C{-}C] \\ & + 0.6183\,(0.1345)\, F04[Cl{-}Cl] \end{aligned} \tag{2}$$

XlogP is a measure of logP, logD is a measure of logP for the ionizable compounds at pH = 6.2, RDF130 m 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 Appendix A, section SM 4.2.1. The overall contribution of molecular descriptors to explain elution mechanism in (-)ESI-RP-LC-HRMS was investigated by Principal Component Analysis (PCA) (Fig. A.8).

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 introduced in Appendix A (section, SM 2.5) are listed in Table B.7. OTrAMS demonstrated that no outliers were detected for the training set (Fig. A.9 (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 3.1, to derive the distribution of compounds based on the origin of their errors. As shown in Fig. A.9(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 as described in Section 3.1. The internal parameters of SVM were optimized to C = 50 (a trade-off parameter), $\varepsilon = 0.08$ (insensitive loss function), $\gamma = 5$ (radial basis function (RBF)). The result of each optimization step is shown in Fig. A.10(a–c). The predicted and experimental t_R values are listed in Table B.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.8). The molecular descriptors used here were investigated for inter-correlation cases. As shown in Table B.10, no indication of inter-correlation is present.

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:

$$\begin{aligned} t_R = {} & +2.591(\pm 0.1323) - 1.233\,(\pm 0.0227) logD_{(pH\ 3.50)} \\ & - 0.1051\,(\pm 0.0204) GGI1 + 0.2293\,(\pm 0.0384) RDF020p \\ & + 0.2410\,(\pm 0.0322) H{-}050 + 1.332\,(\pm 0.1769)\, qnmax \\ & + 0.0807\,(0.0089)\, MlogP^2 + 0.8120\,(0.0370)\, AlogP \end{aligned} \tag{3}$$

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 [22,53]. More details about these molecular descriptors can be found in SM (Appendix A), section SM 4.3.1. The contribution of selected molecular descriptors to explain elution mechanism in HILIC-(+)ESI-HRMS was investigated by Principal Component Analysis (PCA) in Fig. A.11.

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.7. Three outliers (Prometryn, Irgarol-descyclopropyl and Arginine) were detected by OTrAMS for the test set (Fig. A.12 (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 Fig. A.12(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), $\varepsilon = 0.01$ (insensitive loss function) and $\gamma = 3$ (radial basis function (RBF)). The result of each optimization step is shown in Fig. A.13 (a–c). The predicted and experimental t_R values are listed in Table B.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.8). Inter-correlation results for the selected molecular descriptors are listed in Table B.11.

3.4. Acceptable error windows for predicted t_R

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.12 in SM (Appendix B). The strategy described above was used to calculate the acceptable error windows. Therefore, the 95th 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.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.

3.5. Comparison with literature models

Several approaches, previously developed and used to predict t_R [10–2124,54–56], were compared with the work presented here, and are shown in Table B.14. The studies [16,20] 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) [13,19]. The studies [15,20] 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. [12,16,17]. 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 [6,13,20,22]. 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 [15,20].

3.6. Application of t_R 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 [6]. 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.14(a) and A.14(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 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 Fig. A.14(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 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 [5]. 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 (Fig. A.15(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 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 (Fig. A.15(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 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 (Fig. A.16). 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 1).

3.7. Application of t_R prediction for the identification of biocides

The models developed here were applied to the suspect screening of over 273 biocides and active ingredients of pesticides in sewage sludge and wastewater samples. Nine target biocides, were treated as suspects and used for validation of the proposed screening workflow, including t_R prediction and MCS plot. The validation results are presented in SM 4.4.2. The identification methodology can be exemplified for the case of 5-Methylbenzotriazole (Fig. 1). Based on the mass accuracy, isotopic fitting and chromatographic peak score (Fig. 1(a)), two substances were met these conditions (5-Methylbenzotriazole and 2-Aminobenzimidazole), including the interpretation of MS/MS fragments using *in silico* fragmentations tools (Fig. 1(b)). The HILIC t_R prediction model could help to prioritize these suspects according to their degree of MEAN value in MCS plot (Fig. 1(c)). The spectra of reference standard was found in MassBank (SM880101) and the fragments (Fig. 1(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

Table 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

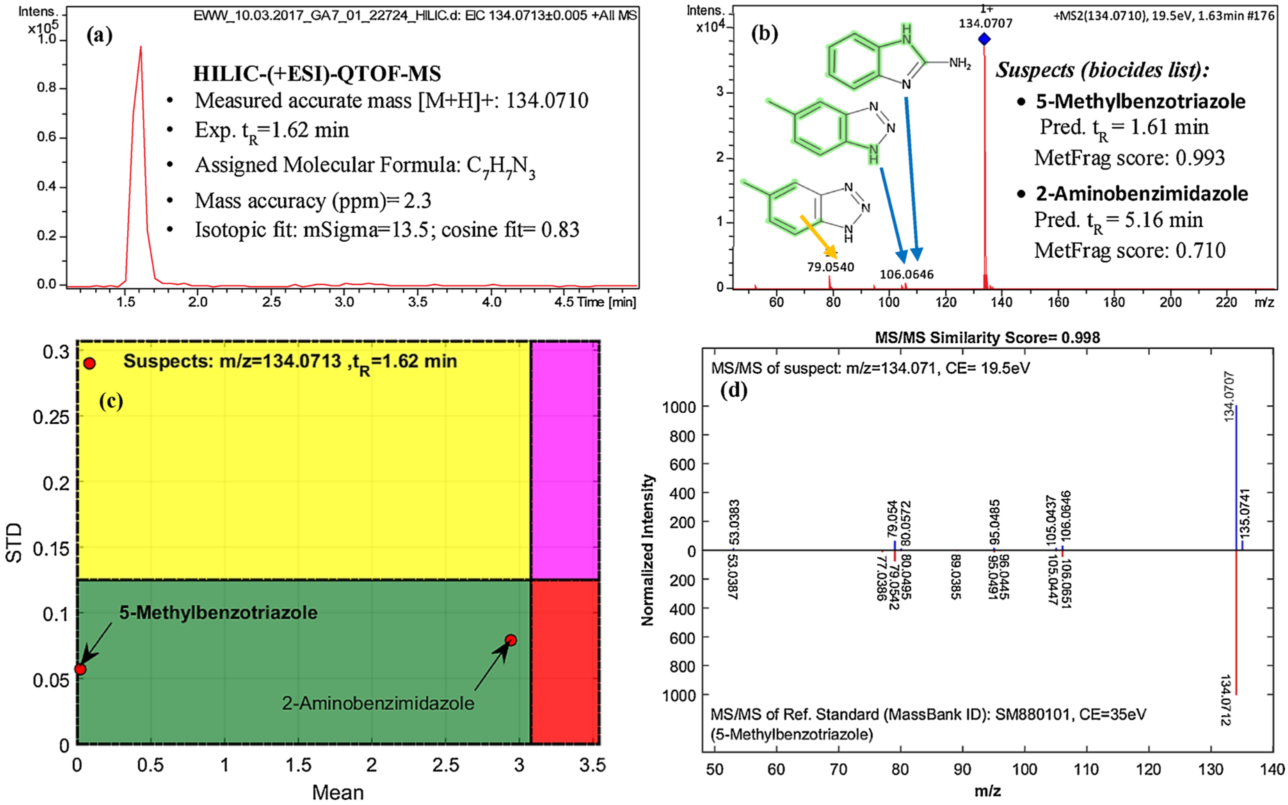


Fig. 1. Identification of 5-Methylbenzotriazole: (a) full MS chromatogram for the given mass (± 5 ppm); (b) MS/MS spectra and corresponding fragments; (c) MCS plot for evaluating the predicted t_R values; (d) confirmation step using spectra library. 5-Methylbenzotriazole was confirmed by reference standard later.

comparison (spectra similarity score of 0.998), and further by corresponding reference standard reaching to level 1.

26 biocides were identified through suspect screening, 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 Ethyldecyldi methylammonium (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 2 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(alkyl-chain)-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 [57]. 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 B.15. (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 [48]. 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 [48]. 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 [48]. 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 B.16). For the ion 242.2842, 74 candidates were retrieved from PubChem, and after applying all identification procedure said above,

Table 2
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
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
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.

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 B.17). 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.

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 Monte Carlo Sampling (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 transformation products (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. 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.

Conflict of interest

The authors declare no competing financialinterest.

Funding sources

R.A. acknowledges the scholarship and financial support from Hellenic Foundation for Research and Innovation (ELIDEK) with contract number 14484.

Acknowledgments

Authors gratefully acknowledge the contributions, proof reading and constructive discussions with Dr. Emma L. Schymanski (Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jhazmat.2018.09.047.

References

[1] S.D. Richardson, T.A. Ternes, Water analysis: emerging contaminants and current issues, Anal. Chem. 86 (2014) 2813–2848.

[2] A.A. Bletsou, J. Jeon, J. Hollender, E. Archontaki, N.S. Thomaidis, Targeted and non-targeted liquid chromatography-mass spectrometric workflows for identification of transformation products of emerging pollutants in the aquatic environment, TrAC Trend Anal. Chem. 66 (2015) 32–44.

[3] M. Krauss, H. Singer, J. Hollender, LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns, Anal. Bioanal. Chem. 397 (2010) 943.

[4] E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibáñez, T. Portolés, R. de Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipaničev, P. Rostkowski, J. Hollender, Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis, Anal. Bioanal. Chem. 407 (2015) 6237–6255.

[5] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying small molecules via high resolution mass spectrometry: communicating confidence, Environ. Sci. Technol. 48 (2014) 2097–2098.

[6] P. Gago-Ferrero, E.L. Schymanski, A.A. Bletsou, R. Aalizadeh, J. Hollender, N.S. Thomaidis, Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewater with LC-HRMS/MS, Environ. Sci. Technol. 49 (2015) 12333–12341.

[7] M. Hu, E. Müller, E.L. Schymanski, C. Ruttkies, T. Schulze, W. Brack, M. Krauss, Performance of combined fragmentation and retention prediction for the identification of organic micropollutants by LC-HRMS, Anal. Bioanal. Chem. 410 (2018) 1931–1941.

[8] C. Moschet, A. Piazzoli, H. Singer, J. Hollender, Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry, Anal. Chem. 85 (2013) 10312–10320.

[9] O.V. Krokhin, R. Craig, V. Spicer, W. Ens, K.G. Standing, R.C. Beavis, J.A. Wilkins, 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, Mol. Cell Proteomics 3 (2004) 908–919.

[10] F. Aicheler, J. Li, M. Hoene, R. Lehmann, G. Xu, O. Kohlbacher, Retention time prediction improves identification in nontargeted lipidomics approaches, Anal. Chem. 87 (2015) 7698–7704.

[11] V.I. Babushok, I.G. Zenkevich, Retention characteristics of peptides in RP-LC: peptide retention prediction, Chromatographia 72 (2010) 781–797.

[12] R. Bade, L. Bijlsma, T.H. Miller, L.P. Barron, J.V. Sancho, F. Hernandez, 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. 538 (2015) 934–941.

[13] D.J. Creek, A. Jankevics, R. Breitling, D.G. Watson, M.P. Barrett, K.E.V. Burgess, Toward global metabolomics analysis with hydrophilic interaction liquid chromatography–mass spectrometry: improved metabolite identification by retention time prediction, Anal. Chem. 83 (2011) 8703–8710.

[14] P.J. Eugster, J. Boccard, B. Debrus, L. Breant, J.L. Wolfender, S. Martel, P.A. Carrupt, Retention time prediction for dereplication of natural products (CxHyOz) in LC-MS metabolite profiling, Phytochemical 108 (2014) 196–207.

[15] J.B. Golubović, A.D. Protić, M.L. Zečević, B.M. Otašević, Quantitative structure retention relationship modeling in liquid chromatography method for separation of candesartan cilexetil and its degradation products, Chemometr. Intell. Lab. Syst. 140 (2015) 92–101.

[16] T.H. Miller, A. Musenga, D.A. Cowan, L.P. Barron, Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks, Anal. Chem. 85 (2013) 10330–10337.

[17] K. Munro, T.H. Miller, C.P. Martins, A.M. Edge, D.A. Cowan, L.P. Barron, Artificial neural network modelling of pharmaceutical residue retention times in wastewater extracts using gradient liquid chromatography-high resolution mass spectrometry data, J. Chromatogr. A 1396 (2015) 34–44.

[18] F. Ruggiu, P. Gizzi, J.L. Galzi, M. Hibert, J. Haiech, I. Baskin, D. Horvath, G. Marcou, A. Varnek, Quantitative structure-property relationship modeling: a valuable support in high-throughput screening quality control, Anal. Chem. 86 (2014) 2510–2520.

[19] E. Tyrkko, A. Pelander, I. Ojanpera, Prediction of liquid chromatographic retention for differentiation of structural isomers, Anal. Chim. Acta 720 (2012) 142–148.

[20] A.M. Wolfer, S. Lozano, T. Umbdenstock, V. Croixmarie, A. Arrault, P. Vayer, UPLC–MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling, Metabolomics 12 (2015) 8, https://doi.org/10.1007/s11306-015-0888-2.

[21] F. Falchi, S.M. Bertozzi, G. Ottonello, G.F. Ruda, G. Colombano, C. Fiorelli, C. Martucci, R. Bertorelli, R. Scarpelli, A. Cavalli, T. Bandiera, A. Armirotti, Kernel-based, partial least squares quantitative structure-retention relationship model for UPLC retention time prediction: a useful tool for metabolite identification, Anal. Chem. 88 (2016) 9510–9517.

[22] K. Goryński, B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, R. Kaliszan, 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 797 (2013) 13–19.

[23] A.D. McEachran, K. Mansouri, S.R. Newton, B.E.J. Beverly, J.R. Sobus, A.J. Williams, A comparison of three liquid chromatography (LC) retention time prediction models, Talanta 182 (2018) 371–379.

[24] R. Aalizadeh, N.S. Thomaidis, A.A. Bletsou, P. Gago-Ferrero, Quantitative structure–Retention relationship models to support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples, J. Chem. Inf. Model. 56 (2016) 1384–1398.

[25] J.J.P. Stewart, MOPAC2016™, (2016) (accessed 20 October 2016), http://www.openmopac.net/.

[26] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, J. Am. Chem. Soc. 107 (1985) 3902–3909.

[27] W. Thiel, Semiempirical quantum–chemical methods, Wiley Interdiscip. Rev. Comput. Mol. Sci. 4 (2014) 145–157.

[28] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON Software for Molecular Descriptors Calculation Milan, Italy, (2007) (accessed 20 November 2016), http://www.vcclab.org/lab/pclient/.

[29] Partitioning(logD) Marvin 6.3.1; ChemAxon, 2014; http://www.chemaxon.com/. (accessed 20 November 2016).

[30] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.

[31] Mathworks. The Mathworks Inc, 2005; https://www.mathworks.com/. (accessed 30 November 2016).

[32] T.I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M.T. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D.W. Stanton, J.J. van de Sandt, W. Tong, G. Veith, C. Yang, 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. 33 (2005) 155–173.

[33] D.-S. Cao, Y.-Z. Liang, Q.-S. Xu, H.-D. Li, X. Chen, A new strategy of outlier detection for QSAR/QSPR, J. Comput. Chem. 31 (2010) 592–602.

[34] C. Christophoridis, M.-C. Nika, R. Aalizadeh, N.S. Thomaidis, Ozonation of ranitidine: effect of experimental parameters and identification of transformation products, Sci. Total Environ. 557–558 (2016) 170–182.

[35] M. Ibáñez, V. Borova, C. Boix, R. Aalizadeh, R. Bade, N.S. Thomaidis, F. Hernández, UHPLC-QTOF MS screening of pharmaceuticals and their metabolites in treated wastewater samples from Athens, J. Hazard. Mater. 323 (2017) 26–35.

[36] M.E. Dasenaki, N.S. Thomaidis, Multianalyte method for the determination of

pharmaceuticals in wastewater samples using solid-phase extraction and liquid chromatography–tandem mass spectrometry, Anal. Bioanal. Chem. 407 (2015) 4229–4245.

[37] N.A. Alygizakis, P. Gago-Ferrero, V.L. Borova, A. Pavlidou, I. Hatzianestis, N.S. Thomaidis, Occurrence and spatial distribution of 158 pharmaceuticals, drugs of abuse and related metabolites in offshore seawater, Sci. Total Environ. 541 (2016) 1097–1105.

[38] V.G. Beretsou, A.K. Psoma, P. Gago-Ferrero, R. Aalizadeh, K. Fenner, N.S. Thomaidis, Identification of biotransformation products of citalopram formed in activated sludge, Water Res. 103 (2016) 205–214.

[39] D.E. Damalas, A.A. Bletsou, A. Agalou, D. Beis, N.S. Thomaidis, 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, Environ. Sci. Technol. 52 (2018) 6023–6031.

[40] P. Gago-Ferrero, V. Borova, M.E. Dasenaki, 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, Anal. Bioanal. Chem. 407 (2015) 4287–4297.

[41] P. Gago-Ferrero, A. Krettek, S. Fischer, K. Wiberg, L. Ahrens, Suspect screening and regulatory databases: a powerful combination to identify emerging micropollutants, Environ. Sci. Technol. 52 (2018) 6881–6894.

[42] European Chemical Agency (ECHA), Biocidal Active Substances, Regulation (EU) No 528/2012, https://echa.europa.eu/information-on-chemicals/biocidal-active-substances (accessed 01 June 2018).

[43] European Commission Health & Food Safety Directorate-General, Safety of the Food Chain, Pesticides and Biocides, Draft Working Document Air Iii Renewal Programme, SANCO/2012/11284 –rev. 21, (2018) (accessed 01 July 2018), https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_ppp_app-proc_air-3_sanco-2012-11284.pdf.

[44] AccuStandard, Biocide Standards Reference Guide, (2018) (accessed 01 July 2018), https://www.accustandard.com/assets/BIOCIDE_GUIDE.pdf.

[45] Pesticide Action Network, PAN Europe Study of Pesticide and Biocide Contamination of Fruit and Vegetables in Four EU Member States, (2009) (accessed 01 July 2018), https://www.pan-europe.info/old/Resources/Other/Pesticide_and_Biocide_Contamination_of_Fruit_and_Vegetables_results.pdf.

[46] W.-R. Liu, Y.-Y. Yang, Y.-S. Liu, L.-J. Zhang, J.-L. Zhao, Q.-Q. Zhang, M. Zhang, J.-N. Zhang, Y.-X. Jiang, G.-G. Ying, Biocides in wastewater treatment plants: mass balance analysis and pollution load estimation, J. Hazard. Mater. 329 (2017) 310–320.

[47] M. Ruff, M.S. Mueller, M. Loos, 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 Res. 87 (2015) 145–154.

[48] T. Ruan, S. Song, T. Wang, R. Liu, Y. Lin, G. Jiang, Identification and composition of emerging quaternary ammonium compounds in municipal sewage sludge in China, Environ. Sci. Technol. 48 (2014) 4289–4297.

[49] European MassBank (NORMAN MassBank), https://massbank.eu/MassBank/ (accessed 15 July 2018).

[50] mzCloud mass spectral database, https://www.mzcloud.org/ (accessed 15 July 2018).

[51] MoNA, MassBank of North America (MoNA), http://mona.fiehnlab.ucdavis.edu/ (accessed 15 July 2018).

[52] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, Prediction of hydrophobic (Lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods, J. Phys. Chem. A 102 (1998) 3762–3772.

[53] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, New York, 2008.

[54] M. Cao, K. Fraser, J. Huege, T. Featonby, S. Rasmussen, C. Jones, Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics, Metabolomics 11 (2015) 696–706.

[55] P.G. Boswell, J.R. Schellenberg, P.W. Carr, J.D. Cohen, A.D. Hegeman, 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, J. Chromatogr. A 1218 (2011) 6742–6749.

[56] D. Abate-Pella, D.M. Freund, Y. Ma, Y. Simón-Manso, J. Hollender, C.D. Broeckling, D.V. Huhman, O.V. Krokhin, D.R. Stoll, A.D. Hegeman, T. Kind, O. Fiehn, E.L. Schymanski, J.E. Prenni, L.W. Sumner, P.G. Boswell, Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods, J. Chromatogr. A 1412 (2015) 43–51.

[57] I. Ferrer, E.M. Thurman, Analysis of hydraulic fracturing additives by LC/Q-TOF-MS, Anal. Bioanal. Chem. 407 (2015) 6417–6428.