

**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**

Richard Bade<sup>1</sup>, Lubertus Bijlsma<sup>1</sup>, Thomas H. Miller<sup>2</sup>, Leon P. Barron<sup>2</sup>, Juan Vicente Sancho<sup>1</sup>, Felix Hernández<sup>1\*</sup>

<sup>1</sup>*Research Institute for Pesticides and Water, University Jaume I, Avda. Sos Baynat, E-12071 Castellón, Spain.*

<sup>2</sup>*Analytical & Environmental Sciences Division, Faculty of Life Sciences and Medicine, King's College London, 150 Stamford Street, London SE1 9NH, United Kingdom*

\*Corresponding author: [felix.hernandez@uji.es](mailto:felix.hernandez@uji.es) Tel.: +34 964 387366 Fax: +34 964 387368

Published in **Science of the Total Environment**

DOI: *10.1016/j.scitotenv.2015.08.078*

Issue: *538 (2015), pages 934-941*

**Abstract:**

The recent development of broad-scope high resolution mass spectrometry (HRMS) screening methods has resulted in a much improved capability for new compound identification in environmental samples. However, positive identifications at the ng/L concentration level rely on analytical reference standards for chromatographic retention time ( $t_R$ ) and mass spectral comparisons. Chromatographic  $t_R$  prediction can play a role in increasing confidence in suspect screening efforts for new compounds in the environment, especially when standards are not available, but reliable methods are lacking. The current work focuses on the development of artificial neural networks (ANNs) for  $t_R$  prediction in gradient reversed-phase liquid chromatography and applied along with HRMS data to suspect screening of wastewater and environmental surface water samples. Based on a compound  $t_R$  dataset of >500 compounds, an optimized 4-layer back-propagation multi-layer perceptron model enabled predictions for 85 % of all compounds to within 2 minutes of their measured  $t_R$  for training (n=344) and verification (n=100) datasets. To evaluate the ANN ability for generalization to new data, the model was further tested using 100 randomly selected compounds and revealed 95 % prediction accuracy within the 2-minute elution interval. Given the increasing concern on the presence of drug metabolites and other transformation products (TPs) in the aquatic environment, the model was applied along with HRMS data for preliminary identification of pharmaceutically-related compounds in real samples. Examples of compounds where reference standards were subsequently acquired and later confirmed are also presented. To our knowledge, this work presents for the first time, the successful application of an accurate retention time predictor and HRMS data-mining using the largest number of compounds to preliminarily identify new or emerging contaminants in wastewater and surface waters.

**Keywords:** Retention time prediction, artificial neural networks, time-of-flight high resolution mass spectrometry, screening of emerging contaminants

## 1. Introduction

The number of emerging contaminants in the aquatic environment is increasing, due to urbanization and subsequent societal and industrial needs (Pal et al., 2014). The development of liquid chromatography-high resolution mass spectrometry (LC-HRMS) technologies has revolutionized the analysis of emerging contaminants in environmental waters, and especially for screening of large numbers of compounds (Agüera et al., 2013; Gómez et al., 2010; Hernández et al., 2011; Hogenboom et al., 2009). HRMS instruments allow the recording of full-scan spectra with high mass accuracy and resolution, thus making it possible to search for any given compound based on its exact mass.

There has been much interest in improving the confidence in the identification of small molecules with HRMS; from potential positives through to detection and finally confirmation (Hernández et al., 2015a; Schymanski et al., 2014). The main distinguishing factor between these levels is the (non-) availability of reference standards. Suspect screening refers to compounds tentatively identified based solely on HRMS data and comparable spectral libraries. Confirmation requires reference standards. An additional tool to increase the confidence in the tentative identification of compounds for which standards are unavailable is reliable and accurate  $t_R$  prediction. This is of particular relevance in the case of degradation/transformation products (TPs), which can reach the aquatic environment in high concentrations, but commonly for which reference standards are less accessible. Chemical risk assessment is therefore significantly challenging for such compounds.

Prediction of  $t_R$  plays an important role in the qualitative identification of emerging contaminants. Many different approaches to  $t_R$  prediction exist and range from the simple (Kern et al., 2009; Nurmi et al., 2012) to the complex (Goryński et al., 2013; Ji et al., 2009; Kaliszan et al., 2003; Ukić et al., 2014a). For example,  $\log K_{ow}$  models can be derived using freely accessible data from chemical databases such as ChemSpider and PubChem, as well as freeware prediction sources such as VCCLABS. Its use in  $t_R$  prediction is extremely simple to implement. It is frequently used in environmental studies for the description of the fate of various pollutants and as a simple  $t_R$  predictor for TPs (Kern et al., 2009) and emerging contaminants (Bade et al., 2015; Nurmi et al., 2012). Alongside simple algorithms, other and more complex *in silico* approaches now exist which are based on quantitative structure-retention relationship (QSRR) modeling,

including artificial neural networks, support vector machines and random forests (Giaginis and Tsantili-Kakoulidou, 2012; Héberger, 2007). The principal aim of QSRR is to predict retention data from the molecular structure and its physicochemical properties, using a range of input descriptors and measured  $t_R$  data. One QSRR method gaining recent attention for broad screening using high resolution techniques is the use of artificial neural networks (ANNs), a predictive computing technique that has shown itself as a promising  $t_R$  predictor with potentially higher accuracy than classical models (Miller et al., 2013; Ukić et al., 2014b). The design of ANNs were inspired by the human brain and differ from classical computer programs in that they generally employ non-linear learning techniques using a set of case examples (i.e. a training dataset) (Kaliszan et al., 2003). In the training phase, the ANN requires a range of suitable molecular descriptors as well as the true output value (in this case, measured  $t_R$ ) to use for comparison with predicted values. At the same time, a second dataset of case examples is often used for verification and to assess overall ANN predictive error. The true output values in the verification set are generally not employed for learning, but the number of training cycles can be stopped by the user or the software when the overall measured error across all cases is at its minimum. Therefore, ANN learning is generally an iterative process and once an acceptable number of training cycles is reached, the optimized ANN can be applied to predict the output where experimentally derived data are unavailable (Miller et al., 2013). In some cases, a third dataset can be used after the model has been finalized to 'blind test' the predictive power of the network. Its use is even more pertinent for analyses where large number of new analytes are expected to occur and with potentially high variance from sample to sample, such as in environmental and municipal water samples. Therefore, since information from the sample includes chromatographic  $t_R$  as well as HRMS data, it makes this interpretation of suspect occurrence more accessible in the first instance.

The aim of this work was to develop and evaluate ANN for predictions of unknown chromatographic  $t_R$  in suspect screening of environmental waters. To the best of our knowledge, this method includes the largest range of physicochemically diverse compounds for this purpose ( $n=544$  in total) and includes both neutral and charged compounds eluted under gradient reversed-phase LC conditions. Lastly, this work aimed to improve upon a recent  $\log K_{ow}$ -based  $t_R$  prediction approach (Bade et al., 2015) using the ANNs as an alternative. This work, for the first time, presents the use of ANN for

identification of additional suspect compounds (including metabolites and TPs) in wastewater and surface water samples both with and without reference standards.

## **2. Experimental**

### **2.1 Reagents and Chemicals**

A total of 544 analytical grade reference materials were used for preparation of model solutions at 25 µg/L or 50µg/L (diluted from mixed standard solutions in methanol or acetonitrile with water) for ANN modeling of  $t_R$ . These included pesticides, drugs of abuse, human/veterinary pharmaceuticals and mycotoxins (See Supplementary Information (SI) Table S1 for all compounds used in this study). These covered a large range of molecular hydrophobicity ( $\log K_{ow}$  -3 to 9). Information relating to 595 standards was available (Bade et al., 2015), however after transforming the compounds using SMILES codes, some errors were observed, leading to incomplete data, and a further 42 were removed from the initial ANN method development (Section 3.1) to use in a subsequent blind test (Section 3.2). Further details relating to these compounds can be found elsewhere (Bade et al., 2015; Hernández et al., 2015b).

### **2.2 Water samples for suspect identification**

A total of 44 composite (24-h) influent and effluent wastewater (IWW and EWW) samples and grab surface water (SW) samples were used to demonstrate the application of the developed ANN model. All these samples were previously used in different studies performed at our lab using the same analytical instrumentation for analysis. (Hernández et al., 2015a) All measured  $t_R$  data herein were generated using ultra-high pressure liquid chromatography coupled to quadrupole-time of flight mass spectrometry (UHPLC-QTOF-MS).

### **2.3 UHPLC-QTOF MS**

A Waters Acquity UPLC system (Waters, Milford, MA, USA) was interfaced to a hybrid quadrupole-orthogonal acceleration-TOF mass spectrometer (XEVO G2 QTOF, Waters Micromass, Manchester, UK), using a electrospray ionization (ESI) Z-Spray interface operating in positive mode. The chromatographic separation was performed using an Acquity UPLC BEH C<sub>18</sub> 100 × 2.1 mm, 1.7µm particle size column (Waters) at a flow rate of 300 µl/min. Gradient elution was performed using mobile phases of A= H<sub>2</sub>O and

B= MeOH, both containing 0.01% HCOOH. The initial percentage of B was 10%, which was linearly increased to 90% in 14 min, followed by a 2 min isocratic period and, then, returned to initial conditions during 2 min. The total run time was 18 min. Nitrogen was used as the drying gas and nebulizing gas.

MS data were acquired over an  $m/z$  range of 50–1000. A capillary voltage of 0.7 kV and cone voltage of 20 V were used. Collision gas was argon 99.995 % (Praxair, Valencia, Spain). The desolvation temperature was set to 600°C, and the source temperature to 135°C. The column temperature was set to 40°C. MS data was acquired in MS<sup>E</sup> mode, selecting a collision energy of 4 eV for low energy (LE) and a ramp of 15–40 eV for high energy (HE). The LE and HE functions settings were for both a scan time of 0.4s. (Hernández et al., 2011; Ibáñez et al., 2013)

Processing of MS data was made using ChromaLynx XS application manager (within MassLynx v 4.1; Waters Corporation). The following parameters were used: mass window 0.020 Da (for positive ID  $\leq$  0.010 Da), peak width at 5% height: 6 seconds, peak-to-peak baseline noise: 1000 and threshold absolute area 500. When manually searching the data for all peaks in an eXtracted Ion Chromatogram (XIC), a chromatographic peak was thought viable when above an intensity threshold of 3000 counts.

#### 2.4 Molecular description and neural network optimization procedures

Compound  $\log D$  data (for a mobile phase of pH=3.2) were generated using Percepta PhysChem Profiler (ACD Laboratories, ON, Canada) and for all other descriptors, Parameter Client freeware was used (Virtual Computational Chemistry Laboratory, Munich, Germany). Canonical simplified molecular line entry system strings (SMILES) were created using ChemSpider freeware (Royal Society of Chemistry, UK) for 544 compounds and from these 16 molecular descriptors (as ANN inputs) were generated including the number of double and triple bonds (nDB or nTB), the number of carbon and oxygen atoms (nC or nO), the number of 4-9 membered rings (nR04-nR09), unsaturation index (UI), hydrophilic factor (Hy), Moriguchi and Ghose-Crippen  $\log P$  (MlogP and AlogP respectively) as well as with software predicted  $\log K_{ow}$  data (Tetko et al., 2005). Prediction of  $t_R$  (as the designated single output) via neural networks was performed using Trajan version 6.0 neural network simulator (Trajan Software Ltd., Lincolnshire, U.K.) and compared with experimentally determined  $t_R$  via correlation graphs as well as

assessment of residual errors. Table S1 has all compounds and their respective predicted and experimental  $t_R$  and ANN subset.

### **3. Results and Discussion**

In a previous study, we developed a simple  $t_R$  prediction model based on  $\log K_{ow}$  of nearly 600 compounds, predicted using freeware (Bade et al., 2015). This resulted in approximately 70 % of all compounds being predicted within 2 minutes of the measured  $t_R$ , and 95 % within 4 minutes. This technique was simple to implement and facilitated the removal of several false positives. However, when investigating unknowns and compounds for which reference standards were unavailable, it was concluded that a more robust, accurate and precise methodology was still needed. In this vein, ANNs were considered as an alternative. Recent work successfully used ANN for a similar purpose, albeit using a much smaller set of compounds of 86 and 166 compounds in either study and focused only on pharmaceuticals (Miller et al., 2013; Munro et al., 2015). It is unlikely that this fully represents the breadth of alternative compound classes and chemistries potentially occurring simultaneously in environmental waters. However, these models successfully predicted  $t_R$  for a range of blind test drug compounds in wastewater and urine to warrant further investigation here using a much larger case dataset.

#### **3.1 Prediction of $t_R$ using Artificial Neural Networks (ANNs)**

The molecular descriptors chosen were based on the previous work wherein more than 200 descriptors were evaluated (Miller et al., 2013). As the same type of reversed-phase column and LC system were used in both studies, the same descriptors were hoped to provide similar results. These descriptors were used again to also assess the possibility for transferring the model to another laboratory and to extend the prediction to a much larger set of chemically diverse compounds. Collinearity data for all molecular descriptors and retention time are given in the SI. Higher Pearson correlations were observed for hydrophobicity-based descriptors with retention time as was perhaps expected (maximum  $R = 0.823$  with  $\log K_{ow}$ ). Similarly, these descriptors also showed

some collinearity with each other ( $R \leq 0.889$ ) which prevented strong conclusions to be drawn regarding their relative importance to an ANN model. As molecular descriptors selected were from previous investigations, there is also the possibility that additional descriptors may have had more importance to retention prediction on this system. For example, retention on reversed-phase media is not only dependent on hydrophobic interactions, but also steric and shape effects. Large molecules may not interact with the stationary phase well and thus show reduced retention (Wilson et al., 2002). A simple Pearson correlation was examined for a selection of potentially relevant additional descriptors covering charge states, geometrical, topological and physicochemical properties. Overall, most of these descriptors showed weak relationships  $< 0.5$  except for the descriptor BTLA96 which showed a negative Pearson coefficient of  $-0.649$ .

Over 100,000 network architectures for each model type were initially investigated for their predictive ability across five different ANN model types including 3- and 4-layer multi-layer perceptrons (MLPs), generalized regression neural networks (GRNNs), radial basis functions (RBFs), linear neural networks and probabilistic neural networks (PNN). For training, 344 cases were used along with 100 cases for both verification and blind testing of network performance. Cases were randomly assigned at the beginning of each network test to prevent any bias from pre-selection. Upon selection of the 'best' model statistics (minimum/maximum values, interquartile ranges, standard deviations, medians and means) were generated to ensure that a fair representation of cases and descriptors were present in each dataset subset (see SI). The diversity of ANN types and architectures tested was balanced against the error generated. For network design and testing the omission of input descriptors was included as an option. One hundred of the best networks (software selected) were retained for further investigation which mainly comprised of MLPs, GRNNs and RBFs. Overall for this separation system, the best correlations of predicted versus experimentally measured  $t_R$  were observed using MLPs in comparison to all other types and these correlations were in agreement with previous works despite different compounds being used for training, verification and blind testing (Barron et al., 2009; Miller et al., 2013; Munro et al., 2015). The finalized network was found to be a 4-layer 16-19-9-1 MLP using all 16 molecular descriptors as inputs (**Figure 1**). The source code (in C) for this ANN has been attached in the SI. Reducing the number of descriptors further worsened predictive accuracy of the blind test set in general. This ANN type and architecture was chosen based on the lowest absolute



errors (i.e. predicted  $t_R$ -measured  $t_R$ ) in the training, verification and blind test sets across all networks. Therefore, ANN architecture was based on performance and the software designer tool was used to optimize the number and composition of hidden layers. The coefficients of determination ( $R^2$ ) were between 0.86 and 0.90 between the three sets, which was already a marked improvement on that obtained from our previous study at  $R^2= 0.67$ . Furthermore, the root-mean-squared error (RMSE) of the blind set of compounds (1.03 min) is less than half that of our previous work (2.19 min) (Bade et al., 2015).

The maximum measured  $t_R$  on this chromatographic system was 16.50 min (narasin; an antibiotic) and the lowest was 0.86 min (methamidophos; an organophosphate insecticide). For all compounds within this retention window of 15.64 min, the mean error in  $t_R$  prediction was <6 % using this ANN approach for all compounds. Overall, the mean absolute errors and standard deviations were recorded as  $0.97\pm 0.95$  min (training set);  $0.79\pm 0.85$  min (verification set);  $0.79\pm 0.69$  min (blind test set); and  $0.91\pm 0.89$  min (all sets combined). When focussing specifically on the blind test set which was used to simulate a true application of the approach, 95 % of compounds had predicted  $t_R$  values within 2.00 min of the measured value and the maximum error was 3.56 min for metosulam (**Figure 2**). However, across the other datasets some larger errors were recorded in isolated cases. **Table 1** shows that for all datasets, 90% of all 544 compounds could be predicted to within 2.00 min of the measured  $t_r$  value. Upon subdivision of the datasets, 85 % of the compounds in the training and verification sets and 95% of compounds in the blind test set were predicted to within two minutes of the measured value. The maximum error recorded within the training set was +6.25 min (for the beta blocker atenolol; measured  $t_r=2.49$  min); within the verification set was +5.19 min (for the anti-helminthic drug levamisole; measured  $t_r=3.02$  min) and within the blind test set was +3.56 min (for the pesticide, metosulam; measured  $t_r=7.54$  min).

Prediction errors for all cases were investigated again with respect to any apparent trends and it was found that a very slight over-estimation existed for poorly retained compounds, as well as the converse for strongly retained compounds. Recent work focussing on modelling a smaller number of compounds in wastewater also revealed a similarly slight bias, but used a different network type (a GRNN)(Munro et al., 2015). When examining those compounds with absolute errors >2.00 min (54 compounds in total across all sets), these were spread across the entire compound retention range (mean

measured  $t_R = 7.81 \pm 3.93$ ). However, a slight over-estimation was again apparent for 29 compounds with  $t_R$  from 1.34-5.38 min (mean measured  $t_R = 2.97$  min). Reduced under-estimation was observed for the remaining 25 compounds eluting between 5.38-16.30 min (mean measured  $t_R = 10.67$  min for these compounds). Seventeen compounds eluting  $< 5.38$  min were over-estimated and  $t_R$  for eleven were under-estimated when eluting  $> 10.00$  min.

The contribution of each descriptor towards the final prediction output was investigated using the ANN software sensitivity analysis tool. In this test, each molecular descriptor is removed and treated as missing by the ANN. A new predicted  $t_R$  is generated and a ratio calculated between the network error with a given input omitted to the error of the network with a complete input dataset. Ratios  $> 1$  indicated higher importance in the prediction. Perhaps not surprisingly for a reversed-phase chromatography system, the most important molecular descriptors and their measured error ratios were: logD (1.443), log $K_{ow}$  (1.182), AlogP (1.114), nO (1.096), UI (1.006) MlogP (1.023), Hy (1.017), nDB (1.012), nR04-nR09 (all 1.000-1.063), nTB (1.004) and nC (1.002) Hydrophobicity-based descriptors are likely to show importance as retention on reversed-phase media is primarily by van der Waals interactions. Again, while these descriptors together show their combined importance to the network, moderate collinearity between them means relative error ratios should be treated with caution for these descriptors (Table SI). However, such collinearity should not adversely affect predictive ability of the model. The lowest ratio was observed for nR04 with a ratio of 1.000 meaning no change in network performance was measured for its removal. Within the dataset, only 14 compounds had 4-membered rings (amoxicillin, ampicillin, cefaclor, cefadroxil, cefalexin, cefotaxime, cefquinome, cefuroxime, cloxacillin, dicloxacillin, heptenophos, oxacillin, oxasulfuron, and penicillin G). Inclusion of nR04 still resulted in better performance in comparison to any other type or architecture investigated during the network optimization stage and so was retained as a descriptor in the final ANN model. Error ratios discussed above represent that of the entire dataset (training, verification and blind test). Sensitivity analysis (see SI) of each set separately revealed excellent consistency across all sets, showing that predictive accuracy was likely to be the dominant contributors to error ratios rather than over-fitting of the training set alone.

### 3.2 Use of ANN as a $t_R$ predictor in environmental water samples

In wide-scope screening methods, where thousands of compounds are searched, it is of great importance to have secondary techniques for aid in the identification process. While reference standards can unequivocally confirm the identification of a compound, purchasing and maintaining standards for all compounds is prohibitively expensive. Data acquired from MS fragmentation of suspect compounds may also direct investigations. Furthermore, for many TPs, reference standards are not available and therefore alternate means for confidence in detection/identification are necessary. Although it is not comparatively informative as mass spectra,  $t_R$  prediction models can be very helpful in gaining more confidence in the obtained data and reducing time-consuming data processing. Most importantly, the application of  $t_R$  prediction is not to replace the use of reference standards, but to help (along with MS/MS data) to direct synthesis efforts for confirmation in the usual manner. Prediction of  $t_R$  is best used at the beginning of this process and is especially useful when at a certain exact mass (i.e. (de)protonated molecule of a suspect compound), more than one chromatographic peak appears in the corresponding XIC.

Along these lines, and to test the “blind” skills of the ANN in a real environmental application, the 100 blind compounds as well as an additional set of compounds from our previous study not initially used into the ANN method (a total of 142 compounds), including primarily metabolites and TPs were searched. None of these compounds were in the training or verification sets used for ANN development. From this list of 142 compounds, 46 were finally selected and searched in 44 water samples (EWW, IWW and SW) using ChromaLynx and the ANN predicted  $t_R$ , based on their possible occurrence in the environment (**Table S2**) (Gracia-Lor et al., 2011, 2010; Hernández et al., 2015a; Zuccato et al., 2006). For further confidence, and to see how many false positive chromatographic peaks (above the intensity threshold) could be disregarded, fragment ions were also included in the detection process (**Table S3**). When a compound was identified on the basis of the accurate mass of the (de)protonated molecule and at least one fragment ion in at least one sample it was included in this test, leaving 26 compounds of various chemical classes and including nine metabolites, for only some of which standards were available in our laboratory (**Table 2**). A  $t_R$  window of 2 minutes was used in this section, as 95% of all compounds in the ANN blind set were within this window, thereby giving high confidence that almost all compounds should be found. The

only compound found to have a  $t_R$  outside that of the ANN predicted 2 minute window was codeine. The incorporation of the ANN predicted  $t_R$  allowed almost half (49%) of all chromatographic peaks to be ignored, and even after removing codeine from the calculation, 48% could be disregarded. Furthermore, all but three compounds had a reduction in the median number of potential positive peaks, while 11 compounds had a median value of only one chromatographic peak remaining after the introduction of the ANN predicted window.

In this section, examples are shown in the identification of losartan (originally tentatively identified with  $t_R$  prediction before a reference standard was purchased) and the tentative identification of the metabolites 10,11-dihydroxy carbamazepine and O-desmethyl venlafaxine (no reference standard available).

### 3.2.1 Assignment with reference standards

Losartan is a pharmaceutical used to treat hypertension that has been both predicted (Howard and Muir, 2013; Oosterhuis et al., 2013) and detected in environmental waters (Hernández et al., 2015a; Matsuo et al., 2011). Its presence in a suspect list is thus warranted, and its exact mass was incorporated into our HRMS database. When screening WW and SW samples, XICs at the exact mass of losartan ( $m/z$  423.1700) resulted in two chromatographic peaks (4.63 and 10.13 minutes) (**Figure 2**, bottom left). The ANN  $t_R$  predictor was used and calculated a  $t_R$  of 9.95 minutes. This is almost exactly the  $t_R$  of one of the peaks, but it is worth noting that even after incorporating a  $\pm 2$  minute window only one peak warranted further investigation. Nevertheless, both peaks could have conceivably corresponded to losartan, therefore further research was conducted.

The LE and HE mass spectrum of the peak at 10.13 minutes was investigated for fragment ions:  $m/z$  207.0917 ( $C_{14}H_{11}N_2$ , -2.4ppm), 377.1552 ( $C_{22}H_{22}N_4Cl$ , +5ppm) and 405.1590 ( $C_{22}H_{22}N_4Cl$ , -2.3ppm) (**Figure 2**, right). Literature (Hernández et al., 2015a) and the mass spectral database MassBank (Horai et al., 2010) were also searched to aid in the confidence of these fragment ions. It was found that these three fragment ions did indeed correspond to losartan. As seen in the figure, all associated fragments corresponded to the peak at 10.13 minutes, following the assertion of the  $t_R$  predictor. Furthermore, it is clearly seen that none of the fragment ions correspond with that of the other peak in the LE (4.63 minutes). A standard was later purchased and injected, unequivocally confirming that the peak at 10.13 minutes did correspond to losartan.

### 3.2.2 Tentative identification of metabolites without standard reference materials

The elimination of false positives in suspect analysis is challenging, especially the environmental matrices investigated herein (SW and WW) as thousands of compounds may be present. Suspect screening, by definition, does not rely on reference standards (Krauss et al., 2010). While the exact mass capability of HRMS has gone some way to avoid false positives, even at narrow mass window chromatograms, matrix inferences can be present, thereby hindering confident identification (Bade et al., 2015; Croley et al., 2012). A precise  $t_R$  predictor can therefore be a great additional means of identification.

Most research of emerging contaminants in the environment has focused on parent compounds, however many compounds can be at least partially metabolized or degraded

in natural conditions (Jakimska et al., 2014). In this respect, the tentative identification of two major metabolites of carbamazepine and venlafaxine were explored: 10,11-dihydroxy carbamazepine and O-desmethyl venlafaxine (**Figure 3a**). More than one chromatographic peak was observed at the exact mass of each protonated molecule. The ANN  $t_R$  predictor was then used to try to help minimize the number of peaks to be analysed, with the predicted  $t_R$  calculated to be 6.29 and 5.76 minutes for 10,11-dihydroxy carbamazepine and O-desmethyl venlafaxine, respectively.

Two large (~4.00 and 6.66 minutes) and one small (~4.90 minutes) peaks are seen in the LE XIC of 10,11-dihydroxy carbamazepine ( $m/z = 271.1080$ ). The predicted  $t_R$  was calculated to be 6.29 minutes and by including a  $\pm 2$  minute window as in Section 3.3.1, the large peak at ~4.00 minutes could be disregarded, leaving the peaks at 6.66 and ~4.90 minutes. The ability to focus on fewer peaks is the primary aim and benefit of  $t_R$  prediction in environmental screening applications for unknowns. While having only one (correct) peak remaining is ideal, being able to disregard some peaks gives credence to the use of  $t_R$  prediction in the identification process. To aid in the differentiation of these peaks, fragment ions were sought in literature, whereby one group performed an MS/MS experiment with a QTOF instrument to find the fragment ions of 10,11-dihydroxy carbamazepine (271.1080, 236.0706, 210.0913 and 180.0808) (Ferrer and Thurman, 2012). As the XICs in the figure show, all fragment ions have the peak at 6.65 minutes in common. This provides great confidence that this peak is indeed from 10,11-dihydroxy carbamazepine, however for unequivocal identification, a standard would still have to be purchased.

The example of O-desmethyl venlafaxine represented the worst case scenario, where no peaks could be removed after application of the  $\pm 2$  minute limit window. However, it must be noted that this was a rare case and only occurred for three of 26 compounds. In the LE XIC, two large (4.69 and 5.00 minutes) and two small peaks (~4.20 and 6.50 minutes) are seen. Even after incorporating the ANN predicted  $t_R$  and associated window ( $5.76 \pm 2$  minutes), all four peaks are still of interest. With the peaks being so close together, even using a  $\pm 1.3$  minute window (corresponding to ~80% of compounds being successfully inside this window) would result in the correct elimination of only one small peak. In this situation, fragment ions have to be used to gain further information on the identity. Fragment ions were thus searched in literature (Herrera-Lopez et al., 2014) and investigated in the HE. **Figure 3b** shows all ions corresponding to the peak at 4.69

min. While the ANN predicted  $t_R$  and associated window did include O-desmethyl venlafaxine, all other peaks in the LE XIC were also inside, meaning that no further confidence could be gained by  $t_R$  prediction, rather through the investigation of fragment ions. Nevertheless, the combination of ANN predicted  $t_R$  and fragment ions led to the tentative identification of this compound.

While the examples explained here show the successful assignment of chromatographic peaks, it is impossible to have total confidence with  $t_R$  prediction. In cases where there is more than one peak in the XIC, and the predicted peak is found to be incorrect, the peaks slightly outside the prediction window will also have to be investigated. Nevertheless, with data processing nowadays being the most time consuming part of environmental screening methods, the time saved by incorporating  $t_R$  prediction outweighs the possibility of false negatives (and positives).

These examples clearly show the utility of ANN as a  $t_R$  predictor, not just for its ability to disregard some false positive peaks, but also for its accuracy and subsequent confidence for tentatively identified compounds. It is therefore recommended when performing large scope (e.g. >1000 compounds) screening of environmental samples to include accurate  $t_R$  prediction in the strategy used for identification. This is particularly useful in the investigation of metabolites and TPs, for which standards can be very costly or unavailable.

## **Conclusions**

This work showed the development and use of a  $t_R$  predictor based on artificial neural networks. In particular, a four layer multilayer perceptron successfully modeled retention of 544 compounds under these separation conditions. Overall, 90 % of all compounds eluted within 2.00 minutes of the predicted value and for 100 blind test compounds, 95% were predicted within this window. The network was applied to additional suspect compound occurrence in wide-scope screening based on the use of LC-HRMS, demonstrating that it can reduce the number of false negatives or positives. This saves time and effort in the tentative identification of the compounds detected, as only those chromatographic peaks that fit the predicted  $t_R$  need to be focused on. Several representative examples are given to illustrate the usefulness of the complementary use of precise  $t_R$  prediction in large suspect screening of emerging contaminants. It is recommended to include this prediction for the identification of suspect compounds, particularly in the investigation of metabolites and TPs of organic contaminants, for which reference standards are commonly less accessible.



## **Acknowledgments**

Richard Bade acknowledges the European Union for his Early Stage Researcher (ESR) contract as part of the EU-International Training Network SEWPROF (Marie Curie-PEOPLE Grant #317205).

The financial support from the Spanish Ministry of Economy and Competitiveness (Ref CTQ2012-36189), and Generalitat Valenciana, Spain (research group of excellence PROMETEO II/2014/023; ISIC 2012/016) is also acknowledged.

Thanks are also extended to the Biotechnology and Biological Sciences Research Council (BBSRC) and AstraZeneca (under the Global SHE research programme) for award of a CASE industrial scholarship for T. Miller who collaborated on this project (Reference: BB/K501177/1).

## References

- Agüera, A., Martínez Bueno, M.J., Fernández-Alba, A.R., 2013. New trends in the analytical determination of emerging contaminants and their transformation products in environmental waters. *Environ. Sci. Pollut. Res. Int.* 20, 3496–515. doi:10.1007/s11356-013-1586-0
- Bade, R., Bijlsma, L., Sancho, J. V., Hernández, F., 2015. Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. *Talanta* 139, 143–149. doi:10.1016/j.talanta.2015.02.055
- Barron, L., Havel, J., Purcell, M., Szpak, M., Kelleher, B., Paull, B., 2009. Predicting sorption of pharmaceuticals and personal care products onto soil and digested sludge using artificial neural networks. *Analyst* 134, 663–70. doi:10.1039/b817822d
- Croley, T.R., White, K.D., Callahan, J.H., Musser, S.M., 2012. The chromatographic role in high resolution mass spectrometry for non-targeted analysis. *J. Am. Soc. Mass Spectrom.* 23, 1569–78. doi:10.1007/s13361-012-0392-0
- Ferrer, I., Thurman, E.M., 2012. Analysis of 100 pharmaceuticals and their degradates in water samples by liquid chromatography/quadrupole time-of-flight mass spectrometry. *J. Chromatogr. A* 1259, 148–57. doi:10.1016/j.chroma.2012.03.059
- Giaginis, C., Tsantili-Kakoulidou, A., 2012. Quantitative Structure–Retention Relationships as Useful Tool to Characterize Chromatographic Systems and Their Potential to Simulate Biological Processes. *Chromatographia* 76, 211–226. doi:10.1007/s10337-012-2374-6
- Gómez, M.J., Gómez-Ramos, M.M., Malato, O., Mezcua, M., Fernández-Alba, A.R., 2010. Rapid automated screening, identification and quantification of organic micro-contaminants and their main transformation products in wastewater and river waters using liquid chromatography-quadrupole-time-of-flight mass spectrometry with an accurate-mass. *J. Chromatogr. A* 1217, 7038–54. doi:10.1016/j.chroma.2010.08.070
- Goryński, K., Bojko, B., Nowaczyk, A., Buciński, A., Pawliszyn, J., Kaliszan, R., 2013. 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, 13–19. doi:http://dx.doi.org/10.1016/j.aca.2013.08.025
- Gracia-Lor, E., Sancho, J. V, Hernandez, F., 2011. Multi-class determination of around 50 pharmaceuticals, including 26 antibiotics, in environmental and wastewater samples by ultra-high performance liquid chromatography-tandem mass spectrometry. *J. Chromatogr. a* 1218, 2264–2275. doi:10.1016/j.chroma.2011.02.026

- Gracia-Lor, E., Sancho, J. V, Hernandez, F., 2010. Simultaneous determination of acidic, neutral and basic pharmaceuticals in urban wastewater by ultra high-pressure liquid chromatography-tandem mass spectrometry. *J. Chromatogr. a* 1217, 622–632. doi:10.1016/j.chroma.2009.11.090
- Héberger, K., 2007. Quantitative structure-(chromatographic) retention relationships. *J. Chromatogr. A* 1158, 273–305. doi:10.1016/j.chroma.2007.03.108
- Hernández, F., Bijlsma, L., Sancho, J. V, Díaz, R., Ibáñez, M., 2011. Rapid wide-scope screening of drugs of abuse, prescription drugs with potential for abuse and their metabolites in influent and effluent urban wastewater by ultrahigh pressure liquid chromatography-quadrupole-time-of-flight-mass spectrometry. *Anal. Chim. Acta* 684, 87–97. doi:10.1016/j.aca.2010.10.043
- Hernández, F., Ibáñez, M., Botero-Coy, A.-M., Bade, R., Bustos-López, M.C., Rincón, J., Moncayo, A., Bijlsma, L., 2015a. LC-QTOF MS screening of more than 1,000 licit and illicit drugs and their metabolites in wastewater and surface waters from the area of Bogotá, Colombia. *Anal. Bioanal. Chem.* 407, 6405–6416. doi:10.1007/s00216-015-8796-x
- Hernández, F., Ibáñez, M., Portolés, T., Cervera, M.I., Sancho, J. V, López, F.J., 2015b. Advancing towards universal screening for organic pollutants in waters. *J. Hazard. Mater.* 282, 86–95. doi:10.1016/j.jhazmat.2014.08.006
- Herrera-Lopez, S., Hernando, M.D., García-Calvo, E., Fernández-Alba, a. R., Ulaszewska, M.M., 2014. Simultaneous screening of targeted and non-targeted contaminants using an LC-QTOF-MS system and automated MS/MS library searching. *J. Mass Spectrom.* 49, 878–893. doi:10.1002/jms.3428
- Hogenboom, A.C., van Leerdam, J.A., de Voogt, P., 2009. Accurate mass screening and identification of emerging contaminants in environmental samples by liquid chromatography-hybrid linear ion trap Orbitrap mass spectrometry. *J. Chromatogr. A* 1216, 510–9. doi:10.1016/j.chroma.2008.08.053
- Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., Nishioka, T., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45, 703–14. doi:10.1002/jms.1777
- Howard, P.H., Muir, D.C.G., 2013. Identifying new persistent and bioaccumulative organics among chemicals in commerce. III: Byproducts, impurities, and transformation products. *Environ. Sci. Technol.* 47, 5259–5266. doi:10.1021/es4004075
- Ibáñez, M., Bijlsma, L., van Nuijs, A.L.N., Sancho, J. V, Haro, G., Covaci, A., Hernández, F., 2013. Quadrupole-time-of-flight mass spectrometry screening for

- synthetic cannabinoids in herbal blends. *J. Mass Spectrom.* 48, 685–94.  
doi:10.1002/jms.3217
- Jakimska, A., Kot-Wasik, A., Namieśnik, J., 2014. The Current State-of-the-Art in the Determination of Pharmaceutical Residues in Environmental Matrices Using Hyphenated Techniques. *Crit. Rev. Anal. Chem.* 44, 277–298.  
doi:10.1080/10408347.2013.835244
- Ji, C., Li, Y., Su, L., Zhang, X., Chen, X., 2009. Quantitative structure-retention relationships for mycotoxins and fungal metabolites in LC-MS/MS. *J. Sep. Sci.* 32, 3967–3979. doi:10.1002/jssc.200900441
- Kaliszan, R., Baczek, T., Bucinkński, A., Buszewski, B., Sztupecka, M., 2003. Prediction of gradient retention from the linear solvent strength (LSS) model, quantitative structure-retention relationships (QSRR), and artificial neural networks (ANN). *J. Sep. Sci.* 26, 271–282. doi:10.1002/jssc.200390033
- Kern, S., Fenner, K., Singer, H.P., Schwarzenbach, R.P., Hollender, J., 2009. Identification of Transformation Products of Organic Contaminants in Natural Waters by Computer-Aided Prediction and High-Resolution Mass Spectrometry. *Environ. Sci. Technol.* 43, 7039–7046. doi:10.1021/es901979h
- Krauss, M., Singer, H., Hollender, J., 2010. LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal. Bioanal. Chem.* 397, 943–951. doi:10.1007/s00216-010-3608-9
- Matsuo, H., Sakamoto, H., Arizono, K., Shinohara, R., 2011. Behavior of pharmaceuticals in waste water treatment plant in Japan. *Bull. Environ. Contam. Toxicol.* 87, 31–35. doi:10.1007/s00128-011-0299-7
- Miller, T.H., Musenga, A., Cowan, D.A., Barron, L.P., 2013. Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks. *Anal. Chem.* 85, 10330–7. doi:10.1021/ac4024878
- Munro, K., Miller, T.H., Martins, C.P.B., Edge, A.M., Cowan, D. a., Barron, L.P., 2015. 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, 33–44.  
doi:10.1016/j.chroma.2015.03.063
- Nurmi, J., Pellinen, J., Rantalainen, A.-L., 2012. Critical evaluation of screening techniques for emerging environmental contaminants based on accurate mass measurements with time-of-flight mass spectrometry. *J. Mass Spectrom.* 47, 303–12. doi:10.1002/jms.2964
- Oosterhuis, M., Sacher, F., ter Laak, T.L., 2013. Prediction of concentration levels of metformin and other high consumption pharmaceuticals in wastewater and regional surface water based on sales data. *Sci. Total Environ.* 442, 380–388.  
doi:10.1016/j.scitotenv.2012.10.046

- Pal, A., He, Y., Jekel, M., Reinhard, M., Gin, K.Y.H., 2014. Emerging contaminants of public health significance as water quality indicator compounds in the urban water cycle. *Environ. Int.* 71, 46–62. doi:10.1016/j.envint.2014.05.025
- Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J., 2014. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ. Sci. Technol.* doi:10.1021/es5002105
- Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V., Radchenko, E., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y., Prokopenko, V. V., 2005. Virtual computational chemistry laboratory - design and description. *J. Comput. Aided. Mol. Des.* 19, 453–463. doi:10.1007/s10822-005-8694-y
- Ukić, Š., Novak, M., Žuvela, P., Avdalović, N., Liu, Y., Buszewski, B., Bolanča, T., 2014a. Development of gradient retention model in ion chromatography. Part I: Conventional QSRR approach. *Chromatographia* 77, 985–996. doi:10.1007/s10337-014-2653-5
- Ukić, Š., Novak, M., Žuvela, P., Avdalović, N., Liu, Y., Buszewski, B., Bolanča, T., 2014b. Development of gradient retention model in ion chromatography. Part II: Artificial Intelligence QSRR Approach. *Chromatographia* 77, 997–1007. doi:10.1007/s10337-014-2653-5
- Wilson, N.S., Dolan, J.W., Snyder, L.R., Carr, P.W., Sander, L.C., 2002. Column selectivity in reversed-phase liquid chromatography: III. The physico-chemical basis of selectivity. *J. Chromatogr. A* 961, 217–236. doi:10.1016/S0021-9673(02)00658-1
- Zuccato, E., Castiglioni, S., Fanelli, R., Reitano, G., Bagnati, R., Chiabrando, C., Pomati, F., Rossetti, C., Calamari, D., 2006. Pharmaceuticals in the environment in Italy: causes, occurrence, effects and control. *Environ. Sci. Pollut. Res. Int.* 13, 15–21. doi:10.1065/espr2006.01.004

**Table 1: Summary of predicted  $t_R$  errors for all ANN test sets. Numbers given in italics are those which fall below the proposed 2-min window limit**

Percentile of compounds	50	55	60	65	70	75	80	85	90	95
Predicted $t_R$ error/min										
All sets (n=544)	<i>0.66</i>	<i>0.75</i>	<i>0.85</i>	<i>0.97</i>	<i>1.05</i>	<i>1.19</i>	<i>1.39</i>	<i>1.56</i>	<i>1.96</i>	2.80
Training set (n=344)	<i>0.70</i>	<i>0.81</i>	<i>0.89</i>	<i>1.02</i>	<i>1.13</i>	<i>1.24</i>	<i>1.45</i>	<i>1.70</i>	2.21	2.87
Verification set (n=100)	<i>0.51</i>	<i>0.59</i>	<i>0.64</i>	<i>0.76</i>	<i>0.89</i>	<i>0.99</i>	<i>1.35</i>	<i>1.44</i>	<i>1.60</i>	2.43
Blind test set (n=100)	<i>0.70</i>	<i>0.78</i>	<i>0.84</i>	<i>0.89</i>	<i>1.01</i>	<i>1.08</i>	<i>1.16</i>	<i>1.34</i>	<i>1.63</i>	<i>1.99</i>

**Table 2: All compounds used for testing ANN predicted  $t_R$ , together with the number of samples each compound was detected, average number of peaks in the XIC and the  $\pm 2$  minute window and the predicted and sample  $t_R$**

Compound	Detection rate out of 44 samples	Median peaks per XIC (range)	Median peaks inside $\pm 2$ min $t_R$ window (range)	Predicted $t_R$ (min)	Sample $t_R$ (min)	Inaccuracy in predicted $t_R$ (min)
10,11-dihydroxy carbamazepine <sup>a</sup>	30	2 (1-4)	1 (1-3)	6.29	6.66	0.37
2-hydroxy-terbutylazine <sup>b</sup>	6	6 (4-8)	2 (1-3)	4.24	5.50	1.26
4-desmethoxy omeprazole <sup>a</sup>	33	1 (1-2)	1 (1-2)	8.15	6.29	-1.86
4-formylamino-antipyrine <sup>b,c</sup>	37	2 (1-8)	2 (1-4)	3.53	3.70	0.17
a-hydroxy metoprolol <sup>a</sup>	17	5 (2-10)	3 (2-4)	2.73	3.30	0.57
Benzoyllecgonine <sup>b</sup>	40	3 (1-6)	2 (1-5)	4.11	4.65	0.54
Bezafibrate <sup>b</sup>	2	6 (4-7)	1 (1-1)	10.80	10.78	-0.02
Caffeine <sup>a</sup>	40	3 (1-6)	1 (1-3)	3.17	3.83	0.66
Carbamazepine <sup>b</sup>	40	3 (1-10)	3 (1-5)	7.73	8.82	1.09
Carboxy losartan <sup>a</sup>	35	2 (1-3)	1 (1-1)	10.75	10.44	-0.31
Codeine <sup>b</sup>	25	5 (1-7)	1 (0-3)	4.85	2.46	-2.39
Cotinine <sup>a</sup>	32	10 (4-14)	5 (3-7)	2.00	1.81	-0.19
Diazinon <sup>b,c</sup>	4	3 (2-5)	2 (1-2)	11.64	12.50	0.86
Diclofenac <sup>b</sup>	13	1 (1-2)	1 (1-1)	11.71	12.17	0.46
Gemfibrozil <sup>b</sup>	13	4 (3-7)	2 (1-4)	12.29	13.34	1.05
Lidocaine <sup>a</sup>	40	5 (1-11)	2 (1-4)	5.21	4.24	-0.97
Lincomycin <sup>b,c</sup>	14	4 (2-8)	2 (1-4)	4.28	3.73	-0.55
Losartan <sup>b</sup>	42	2 (1-4)	2 (1-3)	9.95	10.13	-0.18
Metoprolol <sup>a</sup>	24	3 (1-9)	1 (1-3)	4.07	5.44	1.37
Naproxen <sup>b</sup>	32	7 (4-11)	2 (1-4)	10.06	10.54	0.48
O-desmethylvenlafaxine <sup>a</sup>	29	3 (2-5)	3 (1-5)	5.76	4.68	-1.08
Paraxanthine <sup>a</sup>	39	7 (2-12)	3 (2-4)	2.03	2.97	0.94
Ranitidine <sup>a</sup>	23	4 (2-6)	1 (1-1)	3.03	2.10	-0.93
Terbutylazine <sup>b,c</sup>	4	1 (1-3)	1 (1-2)	9.60	10.78	1.18
Trimethoprim <sup>b</sup>	32	7 (2-10)	4 (1-6)	3.11	3.52	0.41
Valsartan <sup>b</sup>	39	2 (1-7)	1 (1-3)	10.90	11.24	0.34

<sup>a</sup>: Standard not available in laboratory Sample  $t_R$  was based on HRMS data and the incorporation of fragment ions

<sup>b</sup>: Standard available in laboratory

<sup>c</sup>: Compound in blind set of ANN

## Figures for Captions

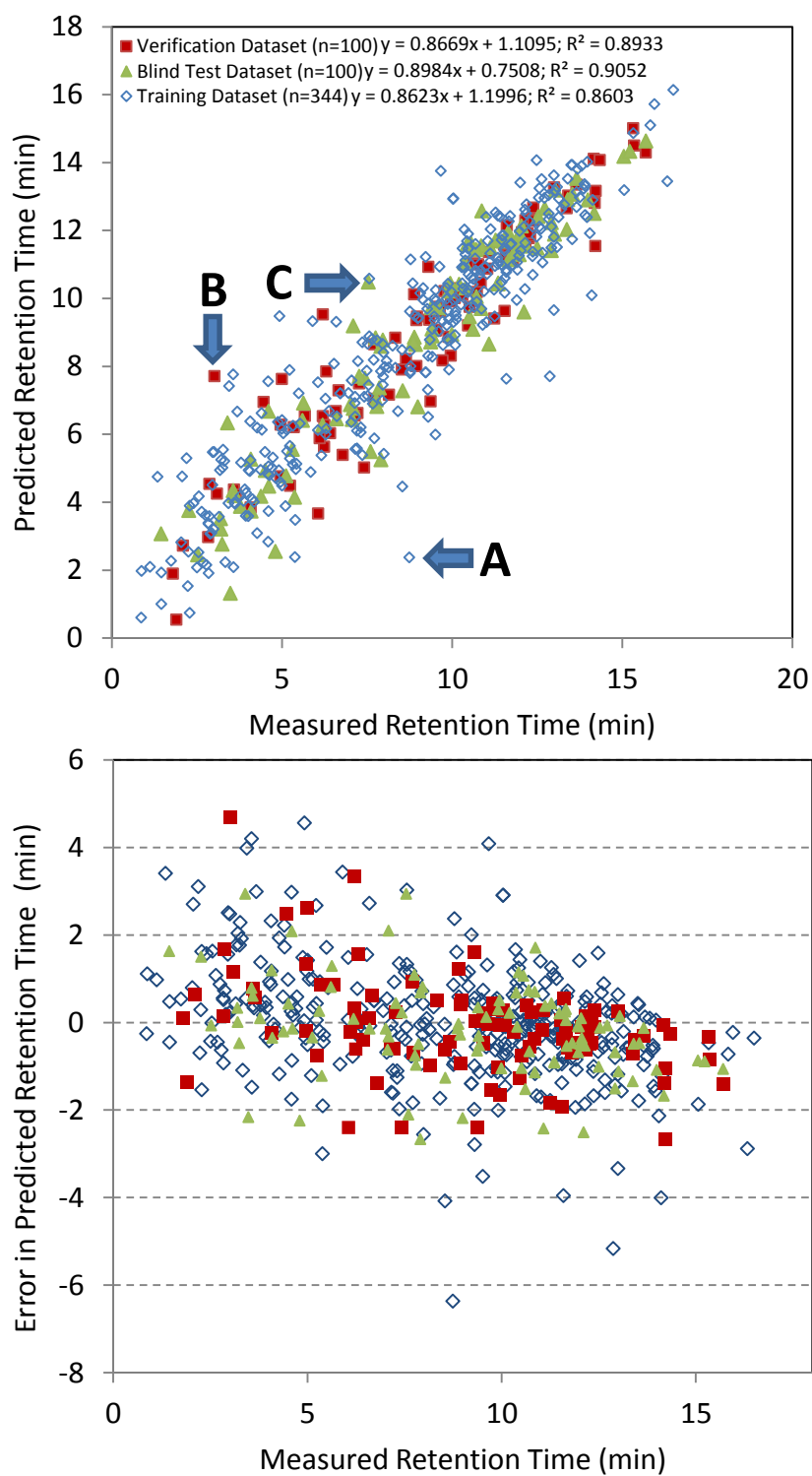
**Figure 1:** TOP: Correlation of measured and predicted  $t_R$  for all compounds. The worst outliers in each set are also shown: A=atenolol (training set, error of -6.21 minutes); B=levamisole (verification set, +5.19 minutes); C=metosulam (blind set, +3.56 minutes). BOTTOM: Residual errors for all compounds in each dataset. For training, verification and blind test sets, n=344, 100 and 100 compounds respectively.

**Figure 2:** LEFT: eXtracted Ion Chromatograms (XICs) of losartan in an EWW sample, together with ANN  $t_R$  prediction (9.95 min) and associated  $\pm 2$  minute window (dotted lines; centre dashed line is the ANN-predicted value). RIGHT: Mass spectra from LE (bottom) and HE (top), showing fragment ions (423.158, 377.156, 207.092).

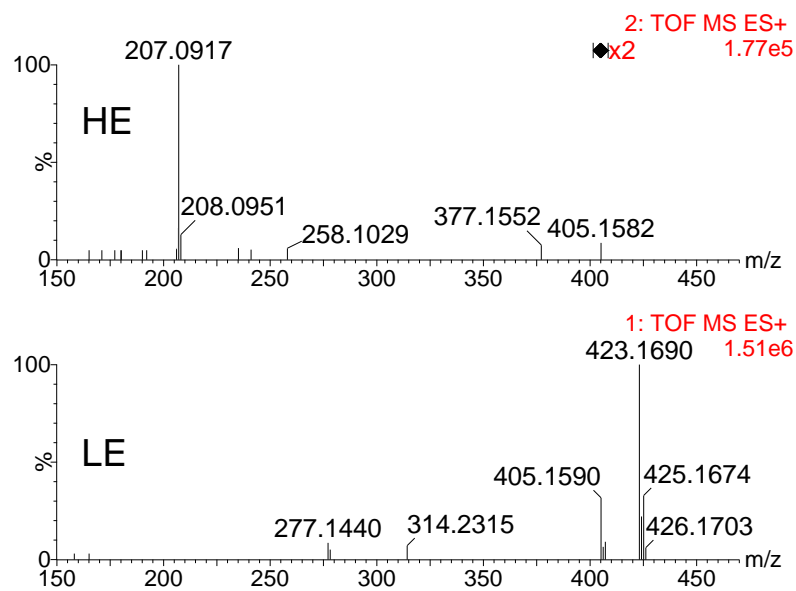
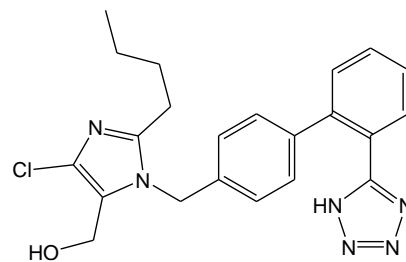
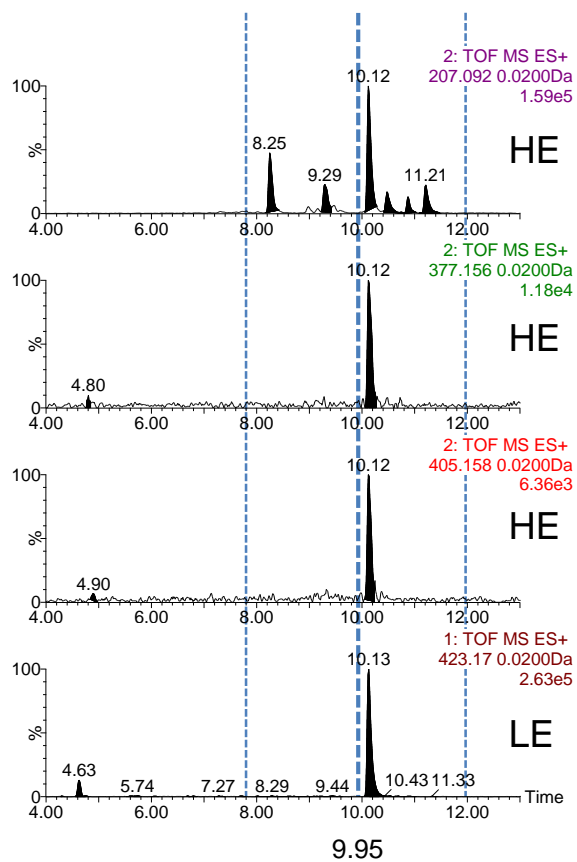
**Figure 3:** Tentative identification of 10,11-dihydroxy carbamazepine (a) and O-desmethyl venlafaxine (b) in an IWW and EWW sample respectively, together with ANN prediction and  $\pm 2$  minute window.



**Figure 4**



**Figure 5:**



**Figure 6:**

