Gradient Retention Time Prediction for 653 Pesticides on a Biphenyl LC Column Using Machine Learning

by Leon P. Barron[a] and Neil J. Loftus[b]*

[a]Dept. Analytical, Environmental & Forensic Sciences, King's College London, Franklin Wilkins Building, 150 Stamford St., London SE1 9NH, United Kingdom

[b]Shimadzu Corporation, Manchester, United Kingdom

**Email: leon.barron@kcl.ac.uk; Tel. +44 (0)20 7848 3842*

The development of a machine learning model for the accurate prediction of 653 pesticide retention times (t_R) on a biphenyl stationary phase is presented. Using an ensemble of four multi-layer perceptron neural networks, prediction of 75% of all compounds lay within 39 s of measured t_R over a 12 min gradient elution method. A total of 16 input variables were selected and included constitutional indices (number of double/triple bonds, carbons and oxygen atoms) ring descriptors (number of 4-9 membered and benzene rings) and molecular properties (unsaturation index, hydrophilic factor, unsaturation index, logP and logD). Correlation of blind test data was excellent (R^2=0.8555 for n=98 compounds). LogD contributed significantly more to predictions compared to other descriptors, but 6-membered/benzene ring and hydrophilicity descriptors contributed more on biphenyl than for separations on C_{18} media. Principal component analysis of descriptor data showed good clustering overall and a wide applicability domain. The ability to accurately predict t_R on biphenyl media represents an excellent opportunity for *in silico* suspect screening applications using an alternative selectivity to C_{18}, especially when coupled to high resolution mass spectrometry.

Introduction

Qualitative and quantitative analysis of environmental samples containing large numbers of analytes in single LC-MS/MS assays has become more widespread in recent years. This has been rapidly enabled by the development of fast scanning and trapping-type mass spectrometry instruments and most notably with high resolution accurate mass spectrometry (HRMS). Despite having accurate *m/z* measurements in both full-scan and tandem MS modes, isomers often exist that make identification challenging for some compounds, especially in complex matrices. Chromatographic retention time (t_R) is usually used to further distinguish compounds, where standards are available. Unfortunately, this is not always the case. For pharmaceuticals and illicit drugs, for example, the presence of Phase I and II metabolites still pose a challenge for confirmation *in silico* as reference materials of high purity either are not available or are prohibitively expensive to procure. Also, retention on C_{18} media is limited for many such polar compounds.

Where multiple unknowns exist in a sample, the prediction of t_R may rapidly enable shortlisting of candidates. Retention prediction has been the focus of significant research and has been particularly successful in gas chromatography [1-4]. However, t_R prediction in liquid chromatography (LC) has been more challenging. Mechanistic approaches, e.g., using linear solvation energy relationships, have been able to successfully predict retention of compounds using sets of measured t_R or retention factor *(k)* data gathered under an array of different experimental conditions, such as mobile phase composition, pH, temperature, flow rate, etc. However, the number of experiments generally required to build such models is often high and application to large numbers of (unknown) compounds, especially under gradient conditions has, on the whole, been very limited. Among other computational approaches, machine learning has been used for many years for t_R prediction of peptides [5]. Recently, machine learning has been used successfully for

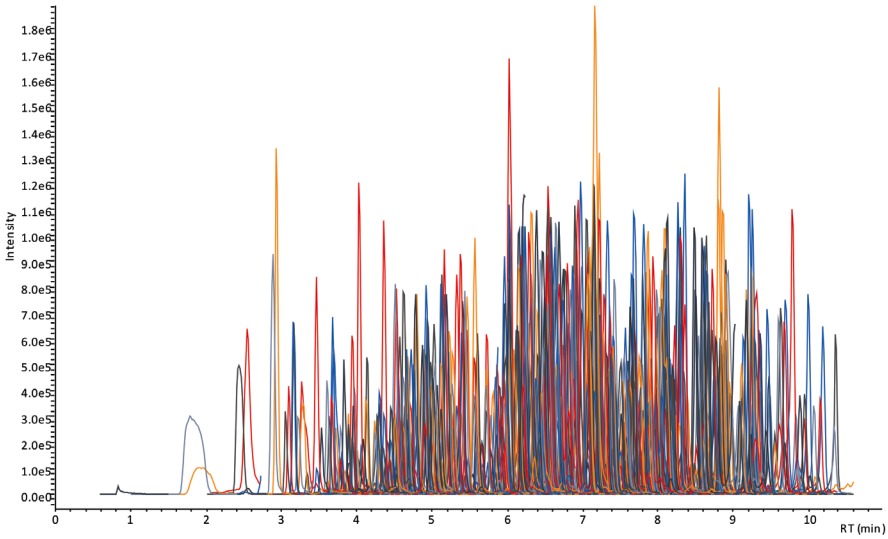

Figure 1. Multiple reaction monitoring chromatograms of 652 pesticides on the 100 x 2.1 mm, 2.7 µm Raptor biphenyl column. Note: dimethirimol data removed for clarity (measured t_R = 5.00 min; predicted t_R = 4.88 min).

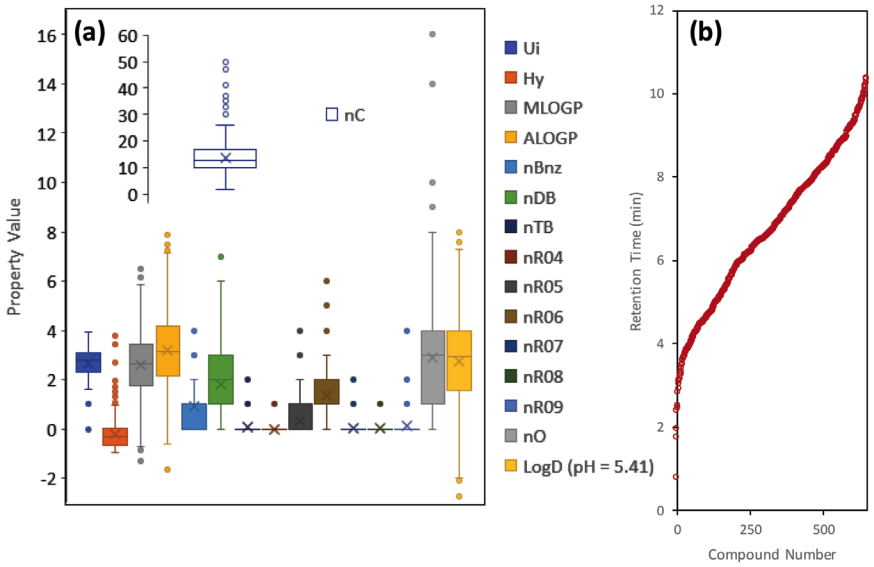

Figure 2. (a) Range of data for each descriptor used in the optimised t_R prediction model and (b) the coverage of measured t_R of all 653 compounds across the 12 min gradient runtime. For molecular descriptor abbreviations, see section 2.2.

small molecules including pharmaceuticals, metabolites, pesticides, herbicides and industrial chemicals [6-9]. This technique involves the use of computer algorithms that can learn to predict t_R by finding trends in compound structures, properties and functions. Recently, we published work using artificial neural networks to predict retention of 1,117 chemically diverse compounds across ten reversed-phase chromatography (RPLC) methods for a range of different applications and sample types [10]. Predictions were generally very good and with an average inaccuracy of 1.02 ±0.54 min across all methods. Longer runtimes generally yielded more inaccuracy, but it was found that inaccuracy was relative and within ~3-5 % of the retention ranges of all analytes measured. Therefore, faster separations in general enabled better accuracy. However, all of these methods involved separations performed on either C_{18} or C_8 media. Other RPLC phases exist, such as aromatic, polar embedded and hybrid phases, which can offer alternative selectivity to C_{18}, especially for isomers and more polar compounds. To our knowledge, no machine learning-based prediction models have yet been developed for other RPLC media.

Fast, selective and highly sensitive methods for an array of different compounds have recently been developed using biphenyl columns and have gained in popularity as an alternative to C_{18}. Shimadzu Corporation recently published a method for 646 pesticides using this phase [11]. The aim of this work was to train and evaluate a machine learning model to predict t_R for application to these compound types given the size of the dataset available for training models in this method. Particularly relevant to pesticides, the aromatic character in the stationary phase can improve separation of these compounds and offers an excellent alternative to C_{18}. Predictions of t_R on several different RPLC media in this way offers the possibility for rapid shortlisting of candidates when performing suspect screening on environmental samples.

2. Experimental

2.1 Retention time datasets

Retention data for 653 compounds were generated using a biphenyl column (Restek Raptor 100 x 2.1 mm, 2.7 µm) configured to a Shimadzu LCMS-8060 LC-MS/MS instrument in polarity switching mode and with MRM data for up to three transitions per compound. Multistep gradient elution was using mobile phase reservoirs containing water (mobile phase A) and methanol (mobile phase B) with both containing a buffer of 2 mM ammonium formate and 0.002% formic acid in each. A 0.002% formic acid concentration resulted in a higher signal intensity in MS/MS, particularly for negative ion mode, and this ion signal response was consistent within and between batch analyses. This approach has been reproduced within food safety applications but also within drugs of abuse testing [12]. The % relative standard deviation (%RSD) of n=100 replicate injections of a spiked apple matrix at 50 µg/L was previously measured at an average of 0.12 %.

The gradient was as follows: 3-10% B over 1 min; 10-55% B for 2 min; 55-100% B over 7.5 min; held at 100% B for 1.5 min followed by re-equilibration to 3 % B for 3 min. The column temperature was 35°C, the injection volume was 2 µL and the flow rate was 0.4 mL/min.

2.2 Descriptor generation, feature selection and retention time prediction

For all compounds, canonical simplified molecular input line entry system (SMILES) strings were generated using Chemspider freeware (Royal Society of Chemistry, UK). Molecular descriptors were generated using two licenced software packages ACD Labs Percepta for logD only (Advanced Chemistry Development Laboratories, ON, Canada) and Dragon version 7 for all other descriptors (Kode Chemoinformatics s.r.l., Pisa, Italy). For prediction of pesticide t_R, n=16 molecular descriptors were based on our previous work [10] including unsaturation index (Ui), hydrophilic factor (Hy), Ghose–Crippen logP (AlogP), Moriguchi logP (MlogP), number of benzene-like rings (nBnz), number of double and triple bonds (nDB/nTB), number of 4–9 membered rings (nR04-nR09), number of carbons (nC), number of oxygens (nO) and logD (calculated at pH 5.4). These descriptors were sub-selected from a larger set of >200 user-curated constitutional, topological and physicochemical descriptors deemed relevant to reversed-phase LC mechanisms in the Network Designer Tool in the neural network simulator package, Trajan v6.0.

2.3 Machine learning, optimisation and procedures

All artificial neural network modelling was performed using Trajan v6.0 software (Trajan Software Ltd, Lincolnshire, UK). The intelligent problem solver tool was used to optimise a suitable neural network and architecture in several steps, each comprising 15 min intervals for training. Briefly, the network type was first selected from a range of different types, including probabilistic neural networks (PNNs), generalised regression neural networks (GRNNs), radial basis function (RBFs), as well as three- and four-layer multilayer perceptrons (MLPs). Here, the MLPs were the best choice for t_R prediction, and are a type of feed-forward, non-linear model that comprises of: (a) an input layer (i.e., molecular descriptors); (b) one or two hidden layers which each contain an optimised number nodes which are

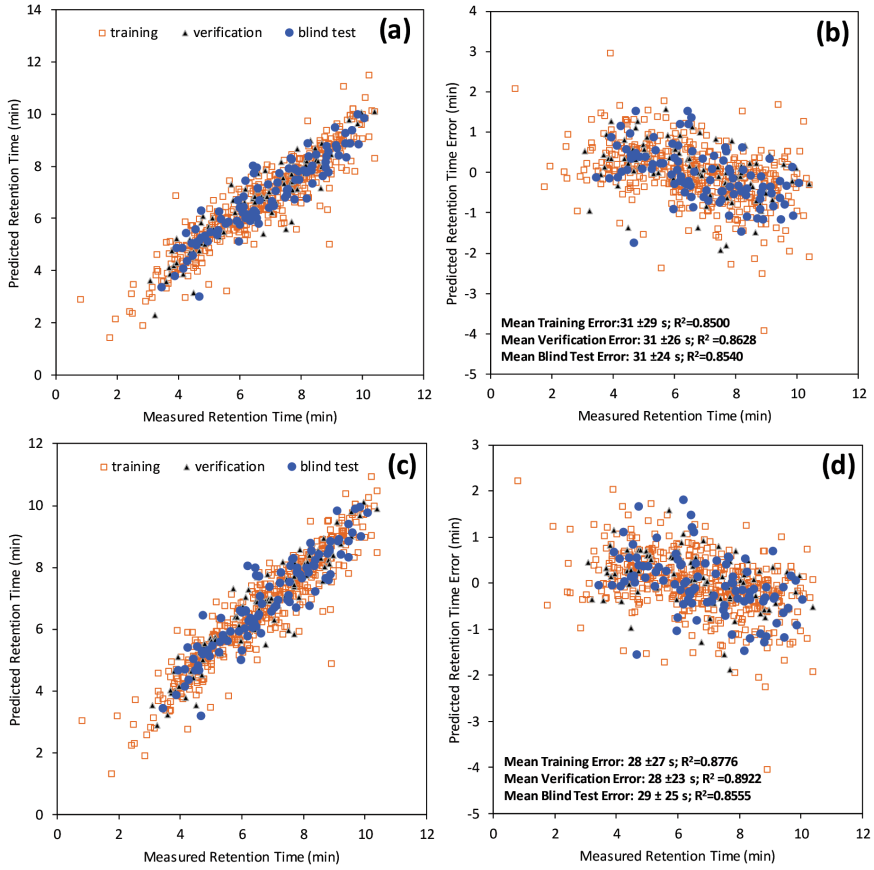

Figure 3. (a) Predicted versus measured t_R using a single 16-5-1 MLP model and its associated residual errors (b), (c) predicted versus measured t_R using an ensemble of four MLPs and its associated residual errors (d). Data split into training data (n=457); verification and blind test data (n=98 each).

interconnected and weighted; and (c) the output layer which collates the information from the hidden layer(s) and generates t_R via an activation function. The proportioning of all datasets was set at 70:15:15 across training, verification and blind test cases (optimised). Cases were randomly re-assigned for every neural network type investigated. The best model type was retained based on the lowest predicted errors obtained and with consistency across the three datasets. The best model overall was then replicated exactly via the custom network designer (n=6) and evaluated for correlation coefficient (R^2), slope, intercept, residual error, as well as overall accuracy, precision and sensitivity to molecular descriptor data.

3. Results and Discussion

3.1 Compound selection and retention behaviour on biphenyl media

The biphenyl column used here offers complimentary selectivity to C_{18} especially for aromatics and polar compounds and especially when used with methanolic mobile phases. According to the manufacturer, it is suitable for fast separations and increases the retention of early eluting species to minimise matrix suppression when used with mass spectrometry. The column is packed with superficially porous particles with a surface area of 130 m^2/g which offered high efficiency in this case.

Retention data for all 653 compounds covered most of the gradient separation space (Figure 1 and Figure 2(b)), which was considered desirable to allow models to learn more fully from quantitative structure-activity relationship (QSAR) data at each timepoint. Higher retention of polar compounds was observed meaning that model predictive accuracy for any new compounds eluting early could be less reliable. For example, and following elution of the first compound, aminopyralid at 0.806 min, the next compound to elute was methamidophos at 1.758 min. Only 9 compounds in total eluted within the first 3 min, after which the remaining compounds eluted in more rapid succession up to etofenprox at 10.367 min (Figure 2). Biphenyl media contain large electron clouds which promote enhanced π-π and dipole-induced dipole interactions in addition to van der Waals interactions. These early eluting compounds contained no phenyl rings, but all contained at least one double bond or displayed some level of aromaticity, and enough to result in significant retention from the void. The full selection of 653 compounds covered a wide range of polarities (e.g., -1.63≤AlogP≤7.88). A total of n=460 compounds had between one and four benzene-like rings and n=565 had between one and seven double bonds. However, correlation of AlogP with t_R was lower than expected at R^2=0.5496. Many compounds were partly or fully ionised under these slightly acidic mobile phase conditions (pH=5.4), with logD for all compounds between -2.75 (diuron) and 7.99 (acequinocyl). This descriptor was correlated to a larger extent with t_R (R^2 = 0.6279) than AlogP and better took into account the ionised portion of all compounds under mobile phase pH conditions. However, it alone could not be used to predict t_R reliably. In the main, nC was high across the board in comparison to nO, potentially resulting in preferential retention via van der Waals forces over dipole-induced dipole interactions. In previous works involving t_R prediction on C_{18} media, a prioritised list of 16 molecular descriptors enabled reliable models to be built for >1,117 drugs, pesticides and industrial chemicals. However, here some of these descriptors yielded no data in the main. These were nTB, nR04-05 and nR07-09. Nonetheless, these were retained in the model as a small number of compounds did possess some of these features and it was decided to test the generalisability of the C_{18} model to another reversed-phase medium as is.

3.2 Performance of the optimised model

During optimisation, it was quickly apparent that MLPs performed best. This was in line with previous models for C_8 or C_{18} media [7, 10]. The best neural network-type model had a 16-5-1 MLP architecture (Figure 3(a)). Fewer layers and nodes was desirable so that the model could be more easily interpreted and to enhance its stability for generalised application. An R^2 >0.85 was achieved for all three datasets, including the blind test data. Overall, excellent consistency was also observed between the training, verification and blind tests datasets showing that the model was not over-trained. All three yielded a mean

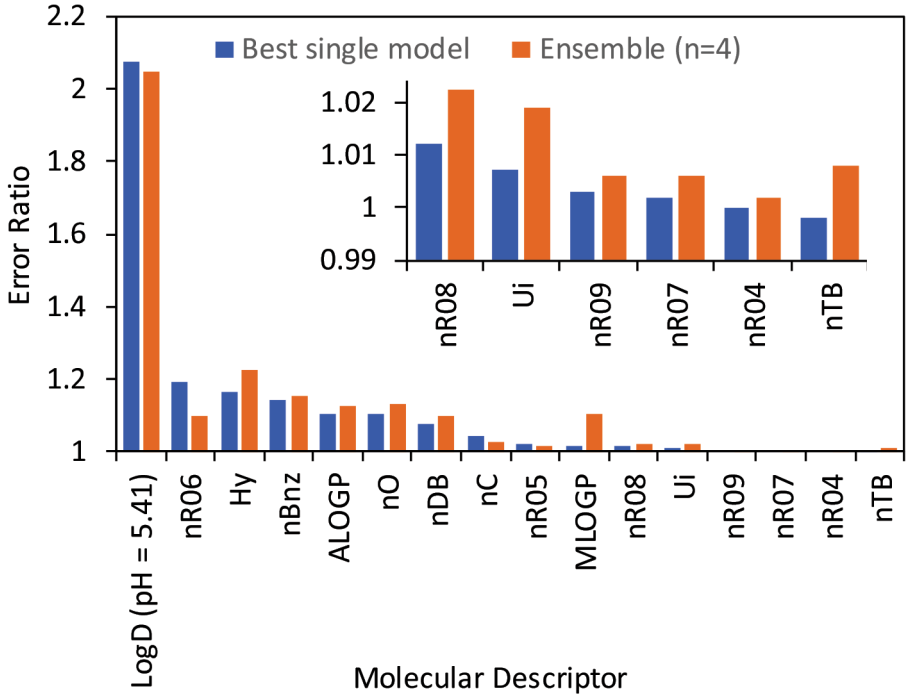

Figure 4. Sensitivity analysis of the single MLP model (blue) and ensemble model (orange). Error ratios >1 represent high model dependency on that descriptor.

error of 31 s from the measured t_R value which represented 4.3% of the analyte retention range, and again similar to previous performance on C_{18}. Over 86% of all compounds were predicted to within 60 s of the measured value (n=567) and the 75th percentile of all errors was 43 s, which was set as the threshold for matching. The worst performance was for triazoxide, mesotrione and cyclosulfamuron with errors of -3.91 min, 2.97 min and -2.5 min, respectively. Each of these have high sulphur or nitrogen content, which is not covered explicitly in the descriptor dataset. All three of these compounds were present in the training set. The worst predictions in the verification and blind test sets were for acibenzolar-S-methyl and allidochlor, respectively and errors for both were within 2 min. However, on the whole, this model generalised very well to this new stationary phase type.

Across all three datasets, there was a small negative bias to the prediction overall (-1.2 s). Closer inspection of Figure 3(b), however, revealed an underlying trend to the errors obtained. Early eluting compounds were very slightly over-predicted in comparison to later eluting compound (average error of the first and second half of all eluting compounds was +17.1 and -17.8 s respectively). The stability of artificial neural networks can be improved by 'ensembling' models. Four replicate MLPs were retrained and combined. Overall, the bias of the ensemble model reduced significantly to 0.22 s. The trend in bias reduced marginally to 15.2 and 13.3 s for the first and second half of eluting species, respectively, but was still evident. This ensemble model marginally reduced the average errors in comparison to the single model alone and the correlations improved generally (Figure 3(c) and (d)). Errors were 28 ±27 s, 28 ±23 s and 29 ±25 s for the training, verification and blind test sets respectively. Performance remained poorest for triaxoxide as before (error = -4.02 min), but slightly improved for both mesotrione (2.05 min) and cyclosulfuron (-2.234 min). Overall, there were fewer outliers than with a single MLP model. Therefore, it was decided to proceed with the ensemble as the preferred approach. The 75th percentile of all absolute errors was 39 s and this was set as the match threshold. In the blind test set for the ensemble model, a few compounds with structural commonality lay outside of this range that are worth noting. Specifically, these included several substituted nitroaniline species, such as butralin, isopropalin, pendimethalin and nitralin. Closer examination of these structures highlighted that no descriptor was included to represent nitro groups specifically, although it was hoped this would have been reflected in logD/logP data indirectly. This could also be partially explained by the existence of only two similar compounds (i.e., oryzalin and flumetralin) in the training set. However, overall, the performance of the ensemble model was considered excellent and represents the first successful prediction of gradient retention times of such a large number of compounds on a biphenyl stationary phase and to this accuracy level.

3.3 Collinearity, sensitivity analysis and applicability domain

Very low single descriptor correlations with t_R were observed and this further strengthened the need for the multi-input model approach (Figure 5). Indeed, initial investigations of simpler, multiple linear regression (MLR) models showed that, although a correlation existed between measured and predicted t_R (R^2=0.7896), it yielded inferior results to neural networks in general with mean t_R inaccuracy for all compounds of 2.35 ±0.83 min. The non-linear neural network approach was far superior, as shown above. Unlike in MLR, where coefficients can be interpreted to help understand contributions of each input to the output, interrogation of input dependency for neural networks is more complex. The dependency of both the best single model and ensemble on each molecular descriptor was then evaluated where each molecular descriptor was systematically removed and the change in performance from the complete dataset calculated to produce an error ratio. Values less than 1.0 indicated that the model was sub-optimal and that descriptor data was worsening predictions. As can be observed in Figure 4, by far the largest contribution to the prediction for both models was logD and in line with similar models on C_{18} media. The next most important descriptors were slightly different between the single and ensemble models and also to previous models on C_{18}. In particular, some descriptors were likely prioritised given the aromatic character on the stationary phase (e.g., nR06, Hy, nBnz). The high contribution of Hy in particular is likely to also reflect the observed effect of increased retention of polar, early eluting compounds as it is related to hydrophilicity [13].This molecular descriptor includes variables such as the number of hydrophilic groups (-OH, -SH, -NH), nC and nSK the number of atoms excluding hydrogen and was the second highest contributor to predictions using the ensemble model. As can also be seen, descriptors that were retained in the model design stage, but which had near zero values were deprioritised in both models (i.e., nTB, nR04, nR07-09). For the single model, error ratio values for nTB was <1.0 meaning there was a slight improvement in the model when it was removed. With the ensemble model however, and even though very few compounds possessed triple bonds, it still did contribute overall to the prediction. This is likely where stability of the ensemble approach was observed, leading to better generalisability.

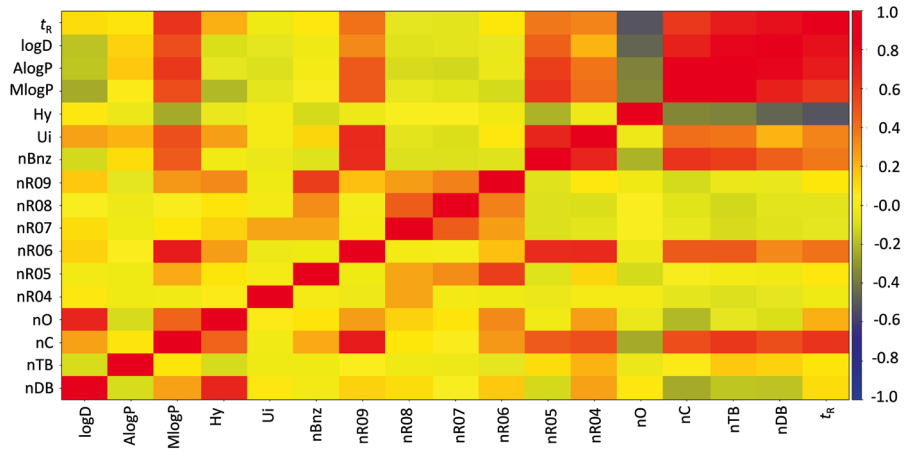

Figure 5. Collinearity analysis for all 16 molecular descriptors and t_R.

The descriptors used here were chosen from a larger list of >200 compound properties that were initially user-curated based on having known relevance to RPLC [7]. Following this, neural network-based descriptor selection was used to narrow the list down to a set of 16 molecular descriptors that yielded the best accuracy overall. An alternative approach to feature selection is to generate thousands of molecular descriptors at random and perform statistical feature selection using statistical algorithms [14, 15]. This may yield alternative descriptors giving similar (or better) performance to the user-curated approach, but may not necessarily have any relationship to mechanisms in RPLC. Both methods were performed here, and both gave similar results in initial experiments. Indeed, pursuit of alternative descriptors to generate a second parallel model may boost confidence in predictions on biphenyl phases, and similar to previous work in our laboratory on prediction of passive sampler uptake rates where both approaches were used. However, this was beyond the scope of this work as the generalisability of the previous C_{18} optimised model was considered a priority to enable simultaneous predictions across multiple types of RPLC methods using the same descriptor set. The limitation of a user-curated approach was that some moderate collinearity existed between some variables (Figure 5). Collinearity can add some unnecessary imprecision or inaccuracy to models if left undetected and can lead to overfitting (which was not observed here, due to good consistency between the training, verification and blind test set data). Particularly, collinearity can affect mechanistic interpretation of the model. The highest positive Pearson correlations of >0.8 were observed, unsurprisingly, for MlogP and AlogP and between AlogP and logD. Therefore, mechanistic interpretations between these specific descriptors in terms of relative weighting should be taken with caution. Both were among the most positively correlated descriptors with t_R along with logD and nC over all others. However, Pearson correlations for all descriptors with t_R was <0.8, showing that no one descriptor was likely useful to model retention accurately. As above, removal of the collinear descriptors worsened the predictions (likely as a result of learning from slight differences in calculation of logP, for example), so these were retained despite being collinear. All other correlations were below a threshold of 0.8 and were considered acceptable for use here. No excessive negative correlation was observed between any of the other descriptors, which might be expected from a user-curated approach. Principal component analysis of the shortlisted descriptor data for all compounds in Figure 6(a) revealed clear clustering for most molecules to define an applicability domain generally. A few outliers existed in principal component 2, which may highlight poor molecular description and a limited applicability domain for these molecules in particular. A closer examination revealed that most of these were macromolecules such as gibberellic acid, avermectins, doramectin, azadirachtin and spinosad which contained larger numbers of rings than the rest of the compounds (Figure 6(b)). However, the predicted t_R for these compounds were mostly within the 39 s threshold except for three compounds isonoruron (t_R absolute error =59 s), sulfosulfuron (43 s) and spinetoram (48 s). Therefore, the selected descriptors for these types of molecules were likely insufficient for accurate predictions, but this was considered a very minor limitation given that this represented <1% of the number of compounds in the dataset. Examination of the PCA data, however, was able to identify this to give added assurance to the user if needed.

Conclusion

Prediction of t_R for 653 pesticides on a biphenyl reversed-phase stationary phase under gradient elution conditions was possible using machine learning for the first time. In particular, an ensemble of four two-layer MLPs achieved the best results within an acceptance threshold set at ±39 s of the true value. Although the data was curated on an LC-MS/MS system in targeted mode, prediction of t_R becomes especially useful for unknown identification workflows using full-scan high resolution mass spectrometry. This approach represents an efficient way to rapidly shortlist suspect compounds before investing in expensive reference materials or synthesis.

References

1. J. Beens, R. Tijssen, J. Blomberg, Prediction of comprehensive two-dimensional gas chromatographic separations. A theoretical and practical exercise, Journal of Chromatography A 822(2) (1998) 233-251.
2. A. Burel, M. Vaccaro, Y. Cartigny, S. Tisse, G. Coquerel, P. Cardinael, Retention modeling and retention time prediction in gas chromatography and flow-modulation comprehensive two-dimensional gas chromatography: The contribution of pressure on solute partition, Journal of Chromatography A 1485 (2017) 101-119.
3. M. Harju, T. Hamers, J.H. Kamstra, E. Sonneveld, J.P. Boon, M. Tysklind, P.L. Andersson, Quantitative structure-activity relationship modeling on in vitro endocrine effects and metabolic stability involving 26 selected brominated flame retardants, Environmental Toxicology and Chemistry 26(4) (2007) 816-826.
4. N. Strehmel, J. Hummel, A. Erban, K. Strassburg, J. Kopka, Retention index thresholds for compound matching in GC-MS metabolite profiling, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences 871(2) (2008) 182-190.
5. K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Paša-Tolić, M.S. Lipton, K.J. Auberry, E.F. Strittmatter, Y. Shen, R.

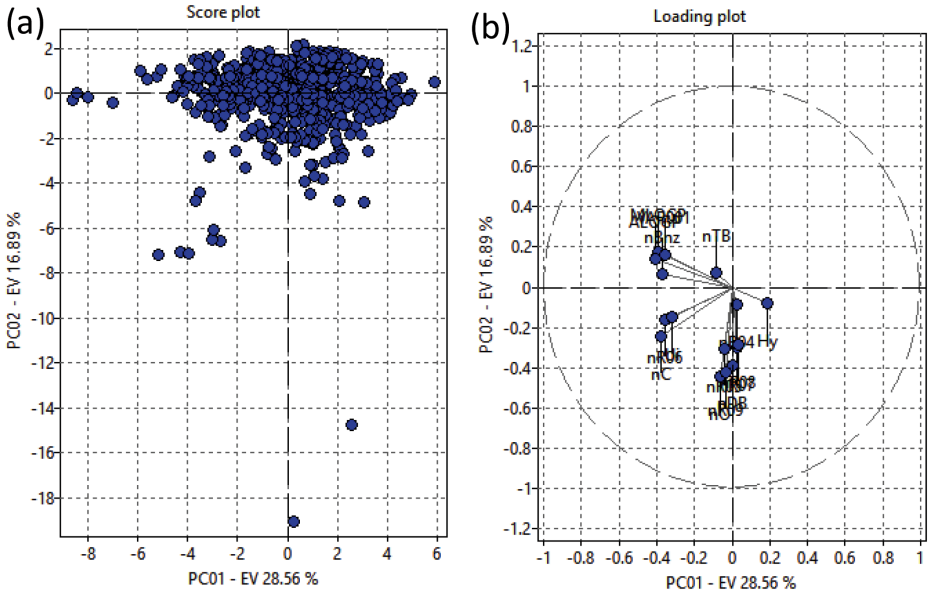

Figure 6. PCA of descriptor data for all 653 compounds showing the score plots for principal component 1 and 2 (a) and the associated loading plots (b).

Zhao, R.D. Smith, Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses, Analytical Chemistry 75(5) (2003) 1039-1048.

6. R. Bade, L. Bijlsma, T.H. Miller, L.P. Barron, J.V. Sancho, F. Hernández, 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, Science of the Total Environment 538 (2015) 934-941.
7. 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, Analytical Chemistry 85(21) (2013) 10330-10337.
8. C.B. Mollerup, M. Mardal, P.W. Dalsgaard, K. Linnet, L.P. Barron, Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry, Journal of Chromatography A 1542 (2018) 82-88.
9. R. Aalizadeh, M.C. Nika, N.S. Thomaidis, Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants, Journal of Hazardous Materials 363 (2019) 277-285.
10. L.P. Barron, G.L. McEneff, Gradient liquid chromatographic retention time prediction for suspect screening applications: A critical assessment of a generalised artificial neural network-based approach across 10 multi-residue reversed-phase analytical methods, Talanta 147 (2016) 261-270.
11. D.R. Baker, L. Fages, E. Capodanno, N. Loftus, Expanding capabilities in multi-residue pesticide analysis using the LCMS-8060, Shimadzu Corporation, Application News Document No. C136, 2016.
12. X. Li, Y. Wang, Q. Zhou, Y. Yu, L. Chen, J. Zheng, A sensitive method for digoxin determination using formate-adduct ion based on the effect of ionization enhancement in liquid chromatograph-mass spectrometer, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences 978-979 (2015) 138-144.
13. R. Todeschini, M. Vighi, A. Finizio, P. Gramatica, 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors, SAR and QSAR in environmental research 7(1-4) (1997) 173-193.
14. T.H. Miller, J.A. Baz-Lomba, C. Harman, M.J. Reid, S.F. Owen, N.R. Bury, K.V. Thomas, L.P. Barron, The First Attempt at Non-Linear *in Silico* Prediction of Sampling Rates for Polar Organic Chemical Integrative Samplers (POCIS), Environmental Science and Technology 50(15) (2016) 7973-7981.
15. T.H. Miller, M.D. Gallidabino, J.R. MacRae, S.F. Owen, N.R. Bury, L.P. Barron, Prediction of bioconcentration factors in fish and invertebrates using machine learning, Science of the Total Environment 648 (2019) 80-89.