Retention time prediction for 653 pesticides on a biphenyl liquid chromatography stationary phase using machine learning

Anthony Sullivan[1]; Leon P Barron[2]; Alan Barnes[3]; Neil Loftus[3]

[1]Shimadzu UK Limited, Milton Keynes, UK; [2]Department of Analytical, Environmental & Forensic Sciences, King's College London, UK; [3]Shimadzu Corporation, Manchester, UK

Overview

- A machine learning model for the accurate prediction of 653 pesticide retention times (t_R) on a biphenyl stationary phase was developed.
- Using a multi-layer perceptron neural network ensemble, prediction of 75% of all compounds lay within 39 seconds of measured t_R over a 12 min gradient elution method.
- 16 molecular descriptors were selected based on molecular structure and properties as input variables for the model. 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.

1. Introduction

Suspect screening large numbers of analytes by a single LC-MS/MS method has become more widespread in recent years with new advances in high speed data-dependent (DDA) or data-independent (DIA) acquisition methods. The process of molecular identification can however be challenging when it is not possible to measure an authentic standard. Retention time verification (or prediction) is a critical tool in suspect screening. The ability to predict retention times on C_{18} has recently been demonstrated using machine learning tools, but models have not been explored for other reversed-phase media which may offer alternative selectivity to enhance component identification. In this work, the prediction of retention times for a diverse chemical space is considered using artificial neural networks for a biphenyl stationary phase.

2. Materials and Methods

653 pesticides were measured by triple quadrupole LC-MS/MS (Shimadzu Corporation, Japan). Retention time prediction modelling initially generated over 5,000 molecular descriptors for each compound. Prioritization of descriptors was performed using collinearity assessment, t_R-correlation, genetic feature selection and user curation. A variety of machine learning model types were trained using these descriptors including linear, radial basis function, probabilistic neural networks, 3/4-layer multi-layer perceptrons (MLPs) and generalized regression neural networks.

Liquid chromatography		Mass spectrometry	
UHPLC	Nexera LC system	**LC-MS/MS**	LCMS-8060
Analytical column	Restek Raptor biphenyl (100 x 2.1 mm, 2.7 µm)	**Ionisation mode**	Heated electrospray
Column temperature	35ºC	**Polarity switching time**	5 msec
Flow rate	0.4mL/minute	**Pause time**	1 msec
Solvent A	Water + 2 mM ammonium formate, 0.002% formic acid	**Total MRM transitions**	1919 (1819 positive; 100 negative)
Solvent B	Methanol + 2 mM ammonium formate, 0.002% formic acid	**MRM Dwell**	4msec (target ion); 1msec (reference ion)
Gradient	Binary reversed phase (10.5min)	**Temp: Int. Block, DL**	350, 300, 150ºC
Injection volume	2µL sample (+ 40µL water)	**Gas: Heat, Dry, Neb,**	10, 10, 3, L/min

Table 1. LC and MS/MS acquisition parameters used to measure 653 pesticides.

3. Results

3.1 Applying a machine learning model for pesticide analysis on a biphenyl phase with MS/MS detection

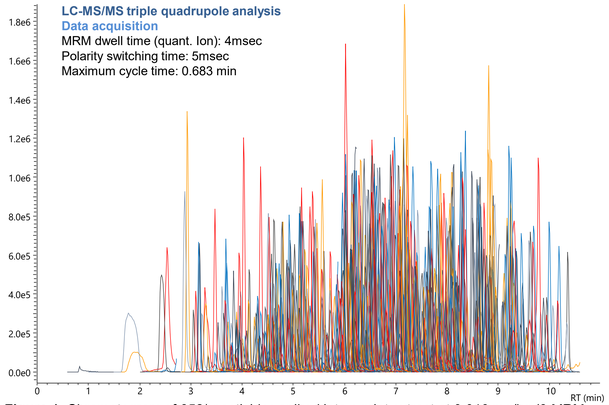


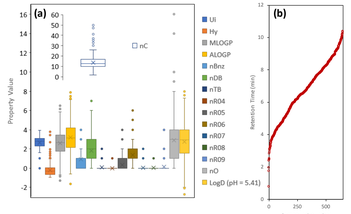
Figure 1. Chromatogram of 652* pesticides spiked into a mint extract at 0.010 mg/kg (3 MRMs per compound). *dimethirimol removed for clarity (measured t_R 5.00 min; predicted t_R 4.88 min)

The panel of compounds, broadly characterized as pesticides, contained a diverse chemical space including herbicides, fungicides, insecticides, avicides and antibacterials/antibiotics. The test model considered an LC-MS/MS MRM method developed to quantify and screen a panel of 653 pesticides from QuEChERS extracts of several food commodities.

Starting from 16 descriptors, four yielded no data in the main (nTB, nR04-05 and nR07-09), however, these were retained in the model as a small number of compounds did possess some of these features (Fig. 2). Twelve descriptors were successfully shortlisted for model training which included molecular properties (logD at pH 5.4, AlogP, Hy), constitutional indices constitution (nO, nN, nBO, nBT, nBM, SCBO) and topological indices (SNar, SCBO, Mi). By using an optimized ratio of 70:15:15 proportioned across the dataset for training, verification and blind testing, a 3-layer 16-5-1 MLP model offered the most consistent and accurate predictions across all three datasets and following ensembling with four replicated MLPs of the same architecture, even better consistency and performance was achieved (Fig. 3.). Excellent correlation between measured and predicted t_R was observed across all three datasets ($R^2 \geq 0.885$). The mean error and standard deviation for n=98 blind test compounds were 27 ±23 seconds which equated to <5 %. For the training (n=457) and verification sets (n=98), the mean errors and standard deviations were 28 ±25, and 28 ±20 seconds, respectively. Following a sensitivity analysis of the model, the most influential descriptor was logD followed by number of six membered rings (nR06), hydrophilic factor (Hy) and number of benzene rings (nBnz). This prediction is in line with predictions achieved previously using C_{18} media and therefore it was concluded that this model could be potentially used for higher assurance *in silico* tentative identification when analyzed using both C_{18} and biphenyl media.

3-2. Molecular Descriptors

For prediction of retention time, n=16 molecular descriptors were based on previous work. All artificial neural network modelling was performed using Trajan v6.0 software (Trajan Software Ltd., Lincolnshire, UK).



Abbreviation	Molecular Descriptor
Ui	Unsaturation index
Hy	Hydrophilic factor
MlogP	Moriguchi logP
AlogP	Ghose–Crippen logP
nBnz	Number of benzene-like rings
nDB	Number of double bonds
nTB	Number of triple bonds
nR04-nR09	Number of 4–9 membered rings
nC	Number of carbons
nO	Number of oxygens
logD	Calculated at pH 5.4

Figure 2. (a) Range of data for each descriptor used in the optimized t_R prediction model and (b) the coverage of measured t_R of all 653 compounds across the 12 min gradient runtime.

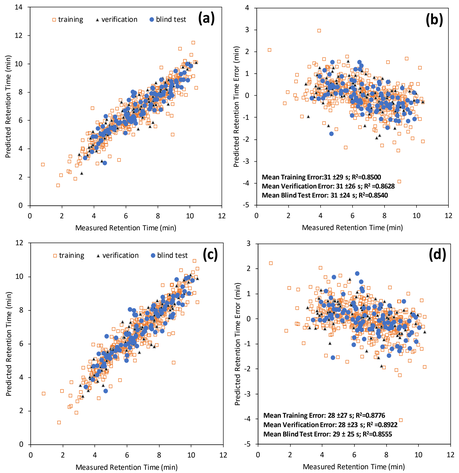


Figure 3. (a) Predicted versus measured t_R using a single 16-5-1 MLP model and its associated residual errors (b), (c) predicted versus measured t_R using an ensemble of four MLPs and its associated residual errors (d). Data split into training data (n=457); verification and blind test data (n=98 each).

3-3. Contribution of each descriptor to model predictions

The dependency of both the best single model and ensemble on each molecular descriptor was evaluated. Each molecular descriptor was systematically removed and the change in performance from the complete dataset calculated to produce an error ratio. The largest contribution to the prediction for both models was logD and in line with similar models on C_{18} media. 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.

Principal component analysis of the shortlisted descriptor data for all compounds revealed clear clustering for most molecules to define an applicability domain (Fig. 5.). A few outliers existed in principal component 2 which were identified as macromolecules. These could be explained due to the compounds containing a larger number of rings compared to other compounds.

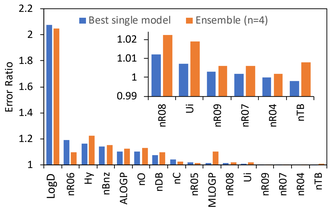


Figure 4. Sensitivity analysis of the single MLP model (blue) and ensemble model (orange). Error ratios >1 represent high model dependency on that descriptor.

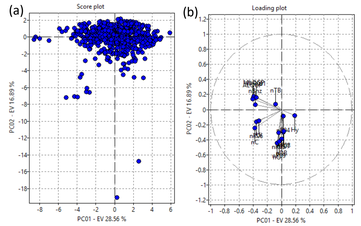


Figure 5. PCA of descriptor data for all 653 compounds showing the score plots for principal component 1 and 2 (a) and the associated loading plots (b).

4. Conclusions

- For the first time prediction of t_R for 653 pesticides was achieved on a biphenyl reversed-phase LC gradient.
- Four two-layer MLPs achieved the best results within an acceptance threshold set at ±39 seconds of the true value.
- This approach represents an efficient way to rapidly shortlist compounds before investing in expensive reference materials when performing suspect screening by LC-MS/MS.