UHPLC Retention Time Prediction Using ACD/ChromGenius Increases Efficiency in Investigative Monitoring of Emerging Contaminants

Anthony Gravell,[1] Graham A Mills,[2] and Matthew J Binnington[3]

[1]Natural Resources Wales. Swansea, UK,
anthony.gravell@naturalresourceswales.gov.uk
[2]School of Pharmacy and Biomedical Sciences, University of Portsmouth, Portsmouth, UK,
graham.mills@port.ac.uk
[3]Advanced Chemistry Development, Inc. Toronto, ON, Canada,
matthew.binnington@acdlabs.com

Introduction

The regulatory protection of European Union (EU) surface waters from chemical pollution is provided by the EU Water Framework Directive.[1] Though routine monitoring for key contaminants represents the primary measure for preserving surface water quality in most of Europe, investigative monitoring for novel pollutants is a critical means to characterize chemical unknowns. However, these emerging contaminants often lack information on chromatographic retention times, impeding the process of identification in analytical samples. This classification challenge is particularly important when considering the plethora of emerging pharmaceuticals, pesticides, metabolites, and industrial chemicals for which well-characterized retention time data are not available.

In this Application Note the retention time prediction capabilities of ACD/ChromGenius are outlined, and a suggestion of how these can be incorporated into a typical investigative water monitoring workflow is also included. Specifically, ultra-high performance liquid chromatography (UHPLC) retention time estimates for over 100 pharmaceuticals are evaluated by comparing predictions to measured values from a curated online database. Though the current example focuses on environmental monitoring of pharmaceutical compounds, the software may be applied to chemicals from any number of industries including food and beverages, oil and gas, and agrochemicals.

Methodology

An innovative approach for investigative monitoring of emerging aquatic contaminants is illustrated in Figure 1. A passive sampling device for polar contaminants (Chemcatcher®)[2] is deployed and the resultant total extract is non-selectively analyzed via UHPLC. High-resolution, accurate-mass mass spectrometry is then used to determine chemical formulae from the mass spectra of chromatographic peaks of interest. Available databases are then searched for target formulae to screen proposed structures, leading to multiple structural candidates. ChromGenius may then be introduced to facilitate peak identification for poorly characterized chemicals by estimating retention times based on structural similarity to known molecules. Inclusion of ChromGenius in such a protocol can significantly reduce the time requirement for identification of individual compounds, and collectively shorten the investigation of a catalogue of unknowns by weeks if not months.
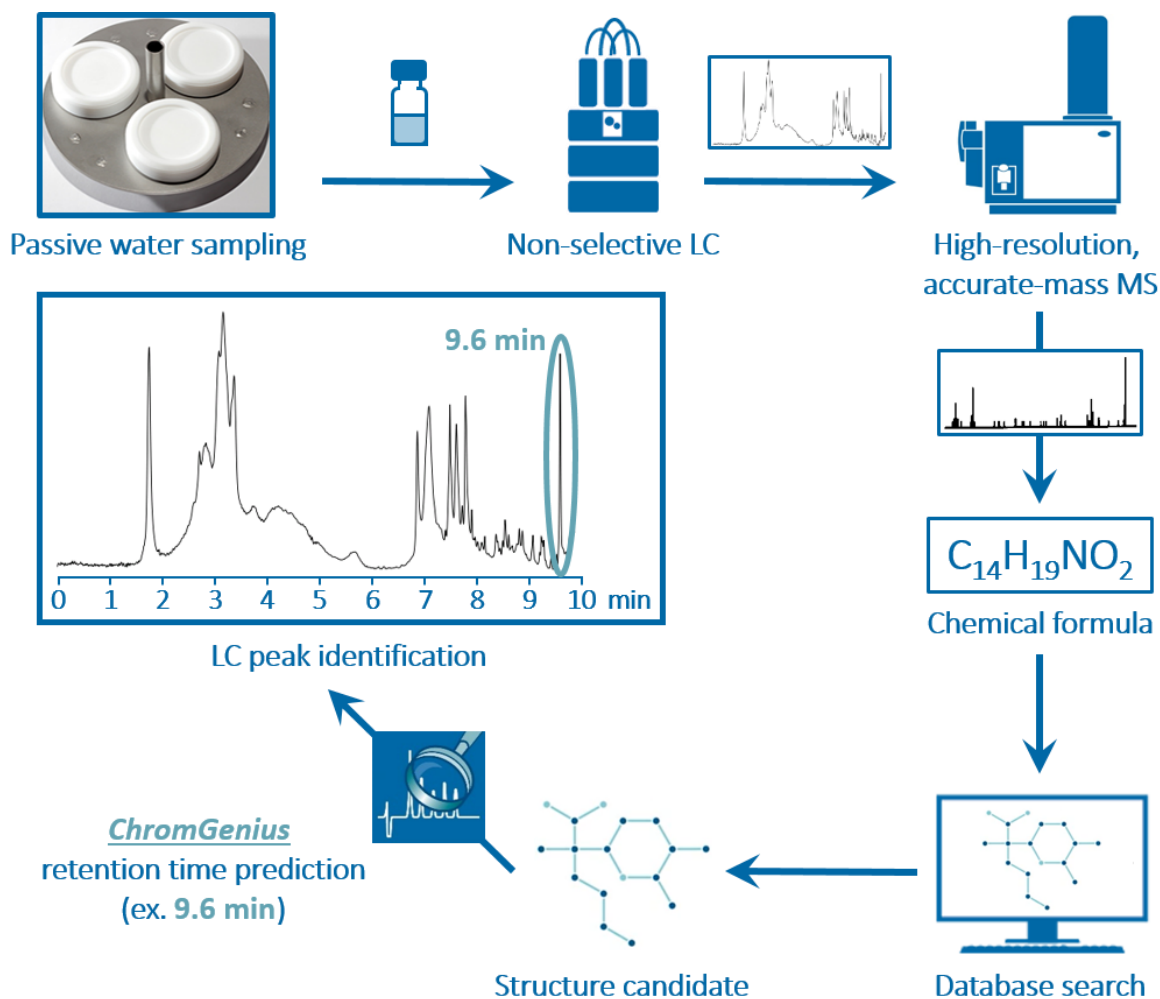
Figure 1. Incorporation of ChromGenius into a typical investigative monitoring framework workflow.

To demonstrate the validity of applying ChromGenius in this investigative monitoring capacity, a retention time prediction training set from the commercial ToxScreener™ database was compiled,[3] which contains accurate mass, retention time, and fragment ion information for 1556 compounds. The testing set represented a subgroup of 103 pharmaceutical compounds from within the larger ToxScreener library. These chemicals were selected based on their presence in a report of the most commonly prescribed pharmaceuticals in the country of Wales, UK during 2014.[4]

For the remaining 1453 training compounds, canonical simplified molecular line entry system strings (SMILES) were downloaded from ChemSpider[5] or created using ACD/ChemSketch, and then imported into ChromGenius. Note that the correlation of measured and estimated training set retention times are displayed in Figure 2, and the resultant R^2 value of 0.8052 provides confidence in applying the model to the testing set. Multiple linear regression was then used to design quantitative structure-retention relationship (QSRR) models for retention time prediction of the 103 testing set compounds based on two different training scenarios:

1. All 1453 training set compounds, referred to as "all compounds fit".
2. A subset of 50 training set compounds exhibiting the greatest structural likeness to the target testing set analyte, referred to as "best 50 compounds fit". ChromGenius is able to generate 2D fingerprints of chemical structures and then calculate similarity scores between fingerprints (scaled from 0 to 1) based on the Tanimoto similarity index.
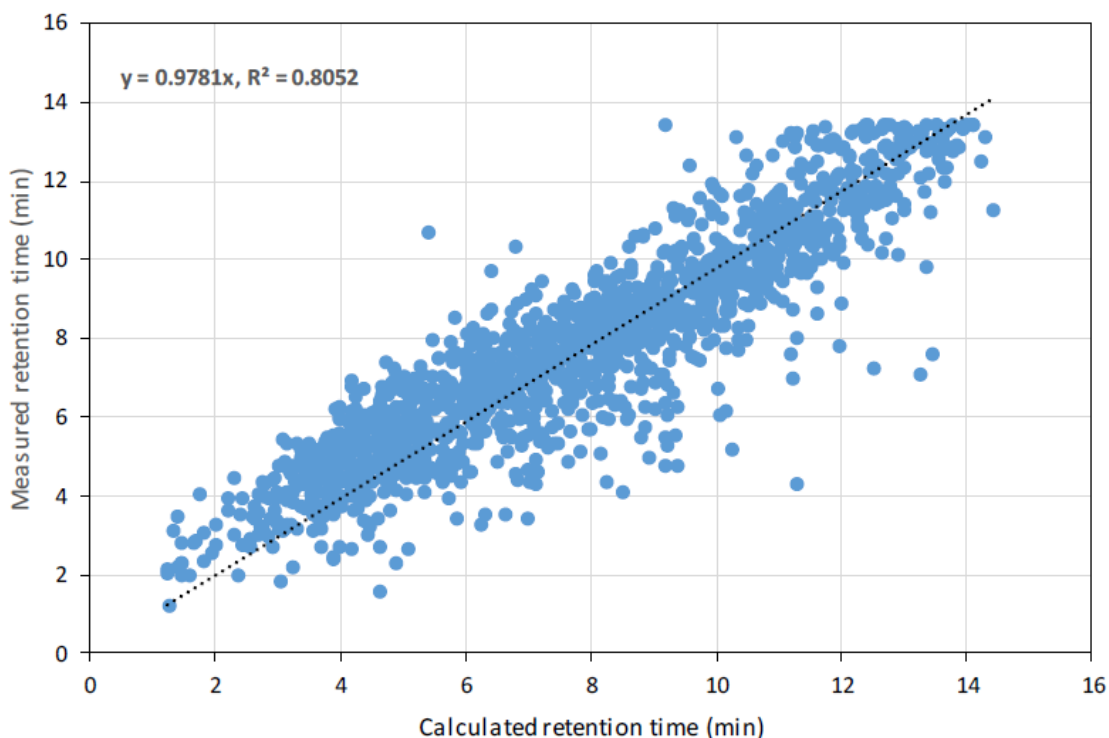

Figure 2. Correlation between measured and estimated UHPLC retention times (min.) for the 1453 compounds in the training set.

Results

Prediction of UHPLC Retention Time Values

The distribution of errors in UHPLC retention times predicted by ChromGenius for the pharmaceutical testing set is depicted in Figure 3, using both all compounds fit and best 50 compounds fit. The ChromGenius prediction algorithm was able to estimate retention times for all 103 testing pharmaceuticals with percentage mean error values of 6.6% and 6.2% for the all compounds fit and best 50 compounds fit, respectively. Moreover, the all compounds fit and best 50 compounds fit approaches generated respective retention time estimates that were within 1 minute of measured values for 70% and 65% of test compounds, and within 2 minutes for 93% and 92% of the dataset.

For the 8 compounds possessing predicted retention times in excess of 2 minutes, ChromGenius structural similarity scores were below 0.8 for all 1453 constituents of the training set, suggesting that a lack of structural analogs was largely responsible for the low-end fit. Conversely, 15 testing set pharmaceuticals exhibited retention time predictions that were within 0.10 minutes of measured values, and were typified by complements of at least 5 training set structural analogs with structural similarity scores of > 0.85. Note that these model performance results are also summarized in Table 1.
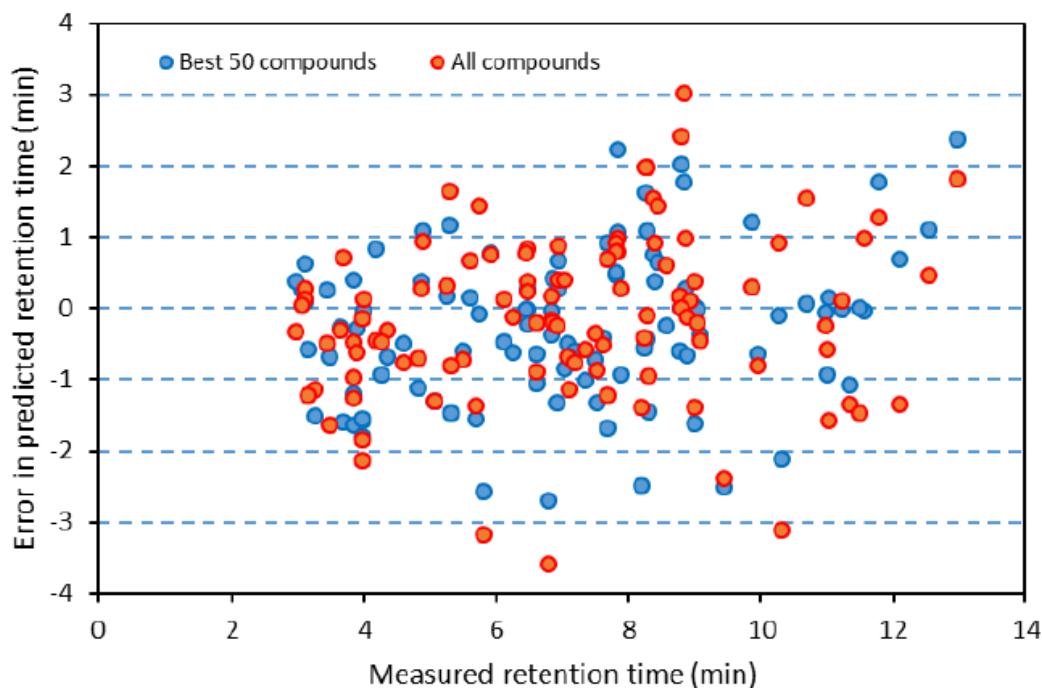
Figure 3. Distribution of errors in estimated UHPLC retention times (min.) for the 103 pharmaceuticals from the testing set, using both the all compounds fit and best 50 compounds fit prediction models.

Table 1. Summary of model performance for ChromGenius retention time predictions.

Prediction model	Mean percent error	Compared to measurements, percent of predictions within…			For predictions outside 2 min		For predictions within 0.10 min	
		1 min	2 min	10%	n	Structural similarity scores	n	Structural similarity scores
All compounds fit	6.6%	70%	93%	54%	8	All < 0.80	15	At least 5 > 0.85
Best 50 compounds fit	6.2%	65%	92%	44%				

Though no guidelines currently exist for classifying compounds via predicted UHPLC retention time data in environmental investigative monitoring, the Society of Toxicological and Forensic Chemistry recommends an acceptable tolerance of 5% for the repeatability of relative retention times in order to identify unknowns.[6] Given this guideline is designed for methodologies where compound reference materials are available, a broader repeatability limit of 10% should be considered acceptable for UHPLC compound identification where reference standards are unavailable or prohibitively expensive. Encouragingly, the retention time prediction algorithm was able to estimate values that were within ± 10% of empirical data for 54% (56/103) and 44% (45/103) of compounds using the respective best 50 compounds fit and all compounds fit, as shown in Table 1.

Conclusion

The current Application Note clearly suggests the potential for ChromGenius to improve environmental investigative monitoring efficiency and effectiveness by demonstrating its UHPLC retention time prediction capabilities. Critically, in order to generate accurate retention time estimates a sufficient number of training structures with high similarity to target compounds are crucial. Of course this is not always possible, especially when investigating emerging contaminants. Nonetheless, the 1453-compound training set was able to predict 103 pharmaceutical UHPLC retention times that were within two minutes of measured values for all but 8 compounds. In fact, 15 compounds exhibited modeled/measured discrepancies that were less than 0.10 minutes. Thus, incorporation of ChromGenius into investigative monitoring approaches could offer significant time and cost savings in emerging pollutant identification, especially because the majority of these chemicals inherently lack reference standards.

To learn more about ChromGenius, please visit www.acdlabs.com/chromgenius/

References

1. European Parliament and the Council of the European Union. 2000. Directive 2000/60/EC - Establishing a framework for Community action in the field of water policy. *Official Journal of the European Communities* L 327. Accessed online 12/05/17: http://ec.europa.eu/environment/water/water-framework/index_en.html.

2. Petrie B, Gravell A, Mills GA, Youdan J, Barden R, Kasprzyk-Hordern B. In Situ Calibration of a New Chemcatcher® Configuration for the Determination of Polar Organic Micropollutants in Wastewater Effluent. Environ Sci Technol. 2016;50(17):9469-78. Accessed online 12/06/17: http://pubs.acs.org/doi/10.1021/acs.est.6b02216

3. Bruker Daltonics. Bruker ToxScreener™ – A Comprehensive Screening Solution for Forensic Toxicology. Accessed online 12/04/17: https://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/Separations_MassSpectrometry/Literature/Brochures/1832909_ToxScreener_Brochure_10-2014_ebook.pdf.

4. The Pharma Letter. 2015. Most prescribed drugs in general practice in 2014 in England and Wales. Accessed online 12/05/17: https://www.thepharmaletter.com/article/most-prescribed-drugs-in-general-practice-in-2014-in-england-and-wales?success_login=1.

5. Royal Society of Chemistry. 2017. Chemspider. Accessed online 12/07/17: http://www.chemspider.com/.

6. Society of Toxicological and Forensic Chemistry. 2009. Guideline for quality control in forensic-toxicological analyses. Accessed online 12/06/17: https://www.gtfch.org/cms/images/stories/files/Guidelines%20for%20quality%20control%20in%20forensic-toxicological%20analyses%20%28GTFCh%2020090601%29.pdf.