Facilitating structural elucidation of small environmental solutes in RPLC-HRMS by retention index prediction

Ardiana Kajtazi, Giacomo Russo, Kristina Wicht, Hamed Eghbali, Frédéric Lynen

PII: S0045-6535(23)01628-4

DOI: https://doi.org/10.1016/j.chemosphere.2023.139361

Reference: CHEM 139361

To appear in: *ECSN*

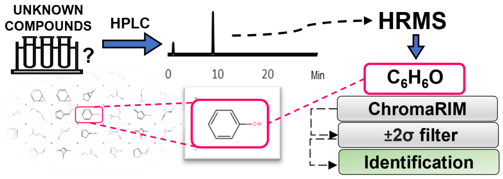
Received Date: 14 March 2023

Revised Date: 6 June 2023

Accepted Date: 26 June 2023

Author Contributions

Ardiana Kajtazi: Conceptualization, Visualization, Methodology, Software, Investigation, Formal analysis, Writing - Original Draft; **Giacomo Russo:** Validation, Writing - Review & Editing; **Kristina Wicht:** Investigation; **Hamed Eghbali:** Supervision; **Frédéric Lynen:** Conceptualization, Visualization, Supervision, Writing - Review & Editing.


UNKNOWN COMPOUNDS
?
HPLC
0
10
20
Min
HRMS
C_6H_6O
OH
ChromaRIM
±2σ filter
Identification

Facilitating structural elucidation of small environmental solutes in RPLC-HRMS by retention index prediction

Ardiana Kajtazi[1], Giacomo Russo[2], Kristina Wicht[1], Hamed Eghbali[3], Frédéric Lynen[1*]

[1]Separation Science Group, Department of Organic and Macromolecular Chemistry, Ghent University, Krijgslaan 281 S4bis, B-9000 Ghent, Belgium

[2]School of Applied Sciences, Sighthill Campus, Edinburgh Napier University, 9 Sighthill Ct, EH11 4BN, Edinburgh, United Kingdom

[3]Packaging and Specialty Plastics R&D, Dow Benelux B.V., Terneuzen, 4530 AA, the Netherlands

*Correspondance: *frederic.lynen@ugent.be*

Abstract

Implementing effective environmental management strategies requires a comprehensive understanding of the chemical composition of environmental pollutants, particularly in complex mixtures. Utilizing innovative analytical techniques, such as high-resolution mass spectrometry and predictive retention index models, can provide valuable insights into the molecular structures of environmental contaminants. Liquid Chromatography-High-Resolution Mass Spectrometry is a powerful tool for the identification of isomeric structures in complex samples. However, there are some limitations that can prevent accurate isomeric structure identification, particularly in cases where the isomers have similar mass and fragmentation patterns. Liquid chromatographic retention, determined by the size, shape, and polarity of the analyte and its interactions with the stationary phase, contains valuable 3D structural information that is vastly underutilized. Therefore, a predictive retention index model is developed which is transferrable to LC-HRMS systems and can assist in the structural elucidation of unknowns. The approach is currently restricted to carbon, hydrogen, and oxygen-based molecules <500 g mol^{-1}. The methodology facilitates the acceptance of accurate structural formulas and the exclusion of erroneous hypothetical structural representations by leveraging retention time estimations, thereby providing a permissible tolerance range for a given elemental composition and experimental retention time. This approach serves as a proof of concept for the development of a Quantitative Structure-Retention Relationship model using a generic gradient LC approach. The use of a widely used reversed-phase (U)HPLC column and a relatively large set of training (101) and test compounds (14) demonstrates the feasibility and potential applicability of this approach for predicting the retention behaviour of compounds in complex mixtures. By providing a standard operating procedure, this approach can be easily replicated and applied to various analytical challenges, further supporting its potential for broader implementation.

KEYWORDS: HPLC-HRMS, Quantitative Structure- Retention Relationship model, In silico prediction, Retention index, Structural elucidation.

1. Introduction

Since the onset of this millennium, the ability of high-resolution mass spectrometry to elucidate the elemental composition of unknown organic molecules has been steadily increasing (Boiteau et al., 2018; De Vijlder et al., 2018). The high mass accuracy in combination with isotope distribution assessment allows now the minimization of a vast number of possible corresponding elemental compositions down to a manageable few. The correct atomic composition can then in many cases fairly easily be obtained based on chemical reasonability, stability, and relevance (De Vijlder et al., 2018). Additional information can and should then also be obtained via MS/MS fragmentation analysis whereby the elemental composition of the daughter ions should be a logical fragment of the parent compound. Because the current instrumentation certainly allows successful implementation of this protocol up to molecular weights of at least 500 g mol^{-1}, the more challenging part of the identification process is largely to be found in the subsequent structural elucidation problem (Boiteau et al., 2018; De Vijlder et al., 2018; Liu et al., 2019).

The identification of previously known molecules via LC-MS can be performed if authentic standards are available and/or if they appear in accessible public databases (such as METLIN, PubChem, and Mass Bank) (Domingo-Almenara et al., 2019; Horai et al., 2010; Wen et al., 2018a). The ensuing identification is then also strongly reliant on mass fragmentography, which can, due to the selectivity of HRMS, allow for the correct identification of the molecule or type of molecule. While this does not exclude the possibility of misidentification due to positional isomer confusion, typically it can be distinguished via chromatographic retention or by ion mobility measurements, whereby further increased reliability is obtained when the information from different separation modes or conditions is combined (Eugster et al., 2014; Kumari et al., 2011).

By contrast, de novo structural elucidation of a priori truly unknown compounds, non-annotated in databases, but for which an elemental composition can be obtained by HRMS, is more problematic (Kumari et al., 2011). This would be typically solved via the combined implementation of various spectroscopic techniques with a particularly strong emphasis on nuclear magnetic resonance spectroscopy (NMR). NMR remains, however, limited due to the at least > 0.1 mg analyte quantity and purity prerequisites, compelling the implementation of multi-repetitive tedious and costly preparative compound fractionation and purification protocols, prior to the spectroscopy (Witting and Böcker, 2020). While this is a well-nurtured approach in chemical or drug development processes, preparative compound purification in life- or environmental sciences are often not feasible due to the too small concentrations and high sample complexity usually involved (Szucs et al., 2021). Additionally, while NMR is the most powerful tool for structural elucidation, the expert nature of the techniques and the lack of specialist-free fully automated structural elucidations algorithms add other hurdles to the challenge (Witting and Böcker, 2020). Another problem with the analysis of unknown solutes is that often they are confused with known related compounds in the databases, whereby obtaining definitive proof of the actual structure is difficult. Therefore, there is a strong need for the development of additional tools allowing to gather structural information of solutes, also when they appear at trace levels that are only detectable by mass spectrometry (Aalizadeh et al., 2021; Boiteau et al., 2018; Cui et al., 2018; Liu et al., 2019).

While the available chromatographic retention information in LC-MS data has been for a long time underused for such purposes, its increased implementation is now gradually emerging to assist in this elucidation process (Gritti, 2023; Zheng et al., 2018). The main challenge therein is that unfortunately chromatographic retention as it is today cannot directly be related to the unambiguous and discrete molecular characteristics hence leading to specific, “easily” understandable, and predictable behavior. This comprising the fragmentation, absorbance or excitation processes observed in mass spectrometry, UV/IR or NMR spectroscopy, respectively (Aalizadeh et al., 2021; Sagandykova and Buszewski, 2021). On the other hand, the retention mechanism of e.g., the reversed phase LC mode is intuitively understood by any chemist. Because also the purity of the stationary phases has concomitantly been improving over time this has led to the many contemporary robust reversed phase methods ubiquitously used in the strictest validated analytical environments (Haddad et al., 2021a). Hence, the composition, structural formula, shape and e.g., the solvation of a molecular structure are all reflected through a particular resulting retention time. The latter can therefore also be considered a molecular characteristic which offers a powerful tool in the search for the structural formula for a given elemental composition.

Much research has been performed with respect to the prediction of molecular retention time for a given structural formula for applications such as swifter method development, suitable column selection and for enhanced compound elucidation within specific compound classes (Meshref et al., 2020; Randazzo et al., 2016a; Wen et al., 2018b; Xu et al., 2023). Various retention models were thereby introduced for Reversed-Phase Liquid Chromatography (RPLC), Hydrophilic Interaction Liquid Chromatography (HILIC), and Ion Chromatography (IC) separation modes (Haddad et al., 2021a; Randazzo et al., 2016a). Such algorithms are also increasingly successfully implemented for the prediction of the retention time of a range of specific groups of analytes, such as lipids (Aicheler et al., 2015; Zheng et al., 2018), steroids (Randazzo et al., 2016a), peptides (Bouwmeester et al., 2021; Dorfer et al., 2018), proteins (Palmblad et al., 2004), and more.

When predicting the chromatographic behavior, the Quantitative Structure-Retention Relationship (QSRR) modelling has often offered a propitious solution in building a promising predictive model (Kaliszan, 1993; Wen et al., 2018a). These mathematical models characterize retention relationships of molecules and have been applied for the aforementioned chromatographic separation techniques, for more than four decades (Amos et al., 2018). In these studies, the model is often used to predict the retention of a target group of compounds to acquire either faster identification and/or greater comprehension of the retention mechanism (Héberger, 2007). The first step involves collecting the experimental retention time of a known training set, such as to be able to build a predictive model that relates retention to the most relevant and broadly applicable molecular characteristics of the training set (Haddad et al., 2021b). Such methods have also been used to predict a variety of molecular characteristics such as retention time (RT) (Ma et al., 2018; Randazzo et al., 2016b; Szucs et al., 2021; Wen et al., 2019; Yang et al., 2021), retention factor (k) (Ruggieri et al., 2005a), logKw (Codesido et al., 2019), logP (Datta et al., 2021), logD (Köhler et al., 2023), and ability to permeate through biological membranes (Russo et al., 2017). Recently the QSRR approach has been increasingly used to prove or disprove the composition of classes of molecules characterized by their modular nature such as peptides or lipids in combination with HRMS/MS (Bouwmeester et al., 2021; Dorfer et al., 2018; Hutchins et al., 2018; Ma et al., 2018; Tiwary et al., 2019). The challenges which have thus far refrained this approach from becoming universally applicable or broadly applied are multifaceted and appear mainly related to standardization and transferability.

On the one hand, unfortunately, much QSRR work has also often been performed on RPLC columns or with chromatographic conditions which are less broadly used. Additionally, the transferability of the resulting retention data to any HPLC instrument type is as important. Considering the notoriously difficult method transfer between different instruments or geographic locations, predictive QSRR models based on retention time or even retention factor are therefore also inherently limited (Haddad et al., 2021a). Additionally, the absence of easily accessible open source information and of fully transferable workflows has also been hindering the development of a gold standard for LC-HRMS based structural elucidation of unknown organic solutes for which an elemental composition has been obtained. Today high-resolution mass spectrometry offers a powerful tool for reasonably reliable prediction of the elemental composition of complete unknowns. Combinations with QSRR then allows translation of the latter into all possible hypothetical structural formulas, for which the corresponding predicted retention (time, factor, or index) can be compared with the experimental retention. This allows removing of a large number of impossible structural formulas for a given retention time.

The proposed research aims to enhance the structural elucidation of unknown environmental solutes with a molecular weight of less than 500 g mol^{-1} (MW<500 g mol^{-1}) that contain carbon, hydrogen, or oxygen atoms. To achieve this, the study presents a novel approach that uniquely combines HRMS and retention information to build a predictive Chromatographic Retention Index Model (ChromaRIM). The transferability of the strategy is maximized through the translation of the retention information into retention indices (RI) on one of the most used stationary and mobile phase combinations, with a gradient spanning the entire elution range. The methodology is tested with known and unknown organic solutes of wastewater treatment relevance.

2. Experimental

2.1 Chemicals and reagents

HPLC grade acetonitrile (MeCN), methanol (MeOH), and ethanol (EtOH) were obtained from Sigma–Aldrich (Steinheim, Germany). Milli-Q grade water (18.2 mΩ cm^{-1}) was purified and deionized in-house by a Milli-Q plus instrument from Millipore (Bedford, USA). Formic acid (FA), 99% purity, was supplied from Sigma–Aldrich (Steinheim, Germany). The 115 neat standard compounds (purity > 98%) were obtained from TCI EUROPE N.V. (Zwijndrecht, Belgium) and Sigma–Aldrich (Steinheim, Germany).

2.2 Sample preparation

Stock solutions of training and test compounds were prepared in concentrations from 1-10 mg mL^{-1} in MeCN, EtOH, and MeOH, depending on their solubility. Once the stock solutions were prepared, they were stored in the fridge or freezer (4 °C/ -18 °C). Standard working solutions were diluted to the concentration of 1-20 µg mL^{-1} in 60:40 (Milli-Q water: Organic solvent) and prepared on the day of analysis.

2.3 Instrumentation and method development

Chromatographic separation was performed on a 1200 series HPLC system (Agilent Technologies, Waldbronn, Germany). The system was constructed out of a 1200 binary pump equipped with a 1200 degasser, a 1200 auto injector, and a 1200 variable wavelength detector (VWD) equipped with a 2 µL microflow cell. RP-LC measurements were performed on a Kinetex Core-shell C18 2.6 µm, 150 x 2.1 mm (Phenomenex, Torrance, CA, USA) with an optimal flow rate of 400 µL min^{-1}. The latter was determined by measuring a reference test mixture isocratically 60:40 (Milli-Q: MeCN) at different flow rates allowing for plate numbers (N) > 27 000. The LC mobile phase, (A) Milli-Q grade water (18.2 mΩ cm^{-1}) and (B) MeCN, were both prepared with 0.1% of FA. Injection volume was 2 µL and the detection for all analytes was recorded at 210 nm, whereas for ketone reference mixture at 280 nm. The column temperature was kept at 30 °C during all analyses. To obtain the most general approach methods were operated from 5-95% (B) in 1) 10 min, 2) 20 min and 3) 40 min followed by re-equilibration with 5% B for the next 10 min. To test the reproducibility and repeatability of the data, these 3 separation methods were performed under the above-listed conditions with random selection of 60 compounds, with differentiation in the linear gradient (Table S2). To generate the predictive model, method 2 was used for further calculations. Full MS (Section S7) was obtained using Q Exactive Orbitrap (Thermo Fisher Scientific). Scan range was 50-500 m/z, Automatic Gain Control (AGC) target was 1e6, Maximum IT was set to 100 ms, and the resolution was 280 000. Detailed ESI parameters for positive and negative mode can be found in Table S10.

2.4 Data collection and molecular descriptor selection

The retention times of all compounds were measured in triplicate and intra- and inter-repeatability were calculated. Subsequently the corresponding RI were calculated according to Kovats RI method usually applied in gradient gas chromatography (Equation SE1). The structures of all compounds were transferred into a Simplified Molecular Input Line Entry System (SMILES) format using ChemDraw and the file was imported as such in the free (of charge) website “Online chemical database” to calculate molecular descriptors of choice using the tool DescriptorsCalculator. A total number of 1879 molecular descriptors were used comprising (2) ALogPS descriptors (Tetko et al., 2005; “Virtual Computational Chemistry Laboratory,” n.d.) and (1877) AlvaDesc v.2.0.14 (Mauri, 2020) from which (198) 2D AlvaDesc descriptors (including constitutional descriptors, Topological indices and P_VSA-like descriptors) and (1677) 3D AlvaDesc descriptors (comprising the following categories: Geometrical descriptors, 3D matrix-based descriptors, 3D autocorrelations, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, Randic molecular profiles, Functional group counts, 3D Atom Pairs, Charge descriptors, Molecular properties, CATS 3D, and WHALES). The value for each descriptor for each solute was calculate via AlvaDesc and exported to Excel.

2.5 QSRR model validation

The QSRR model was calculated using VEGA ZZ 3.2.1.33 (Pedretti et al., 2021), where the experimental RI was a dependent variable and molecular descriptors were the independent variables. Pre-processing of the data was done by normalization min-max feature scaling. The initial screening of descriptors involved two steps: evaluating zero variance and conducting a single-variable regression analysis (Danishuddin and Khan, 2016). Furthermore, by evaluating the variance inflation factor (VIF), collinear descriptors were recognized and those with VIF > 5.00 were disregarded. With the remaining 37 descriptors the best models were calculated with both leave-one-out (LOO) cross-validation and by randomly splitting the dataset into 71:30 pairs of training and test sets in 10 trials. Lastly, the best QSRR model including 7 descriptors was used in the identification of unknown compounds by predicting their

RI. Internal validation was statistically determined with VEGA ZZ (Table S4-S7), and external validation was done by introducing 14 external test compounds. The assessment of applicability domain (AD) was presented in the Williams plot (Figure S4) using standardized residuals and leverages. For the unknowns for which the elemental composition was known, the list of possible structures with the same molecular masses was downloaded from ChemSpider in SDF format, after which molecular descriptors were calculated in the same way. All experimental chromatograms and graphs were processed using OriginPro 9.0 (OriginLab Corporation, Northampton, MA.). Simulation chromatograms were constructed with Microsoft Excel.

3. Results and discussion

This study explores the ability of reversed phase liquid chromatography to confirm or eliminate proposed structures for organic solutes based on their elemental composition. The first phase involves the development of a gradient HPLC methodology that is broadly applicable. This methodology has the potential to be established as a standard approach for chromatography-supported structural elucidation. All retention data is therefore translated towards RI, which are subsequently used to build a QSRR model allowing to accept or reject the retention of structures in a given molecular space. Emphasis is thereby not set on the ability to predict the retention times or indices for specific molecules in the best possible way, but on the capacity of the given model to provide useful and as reliable as possible exclusion or inclusion of structural predictions for C, H, O <500 g mol^{-1} compounds, when compared with the experimental retention of an unknown. In the second part of the work the implementation of the model is rigorously tested. It is thereby shown that it can be used to correctly accept or reject the many hypothetical structural formulas which can be drawn for a given elemental composition. Because the latter quickly leads to an astronomical number of possible structures, it is not realistic with 1D-HPLC to pinpoint only the right structure, but it does offer the ability to remove a vast number of chromatographically impossible structures for a given experimental retention time and atomic composition of an unknown solute. Assuming a robust model is used one can then select the predicted structures, eluting in the range of the experimentally obtained one, as the most probable structures of the true unknown. The latter can then be further refined via conventional exploitation of the MS/HRMS info. The current ChromaRIM approach can serve as platform method for this purpose but can also be considered as a first keystone method in multidimensional approaches whereby each added separation dimension further refines and restrict the search zone.

The work is therefore subdivided into a development section comprising 1) the HPLC method selection/development, 2) selection of the compounds and of the charted molecular space and data collection 3) conversion to retention indices, 4) descriptor selection and attrition and 5) construction of the most suitable QSRR model. The model is then 6) internally assessed and also tested with known environmentally relevant solutes (for additional external consolidation) and finally 7) implemented using the developed model. The general strategy for both the development and implementation is represented in Figure 1A and B.

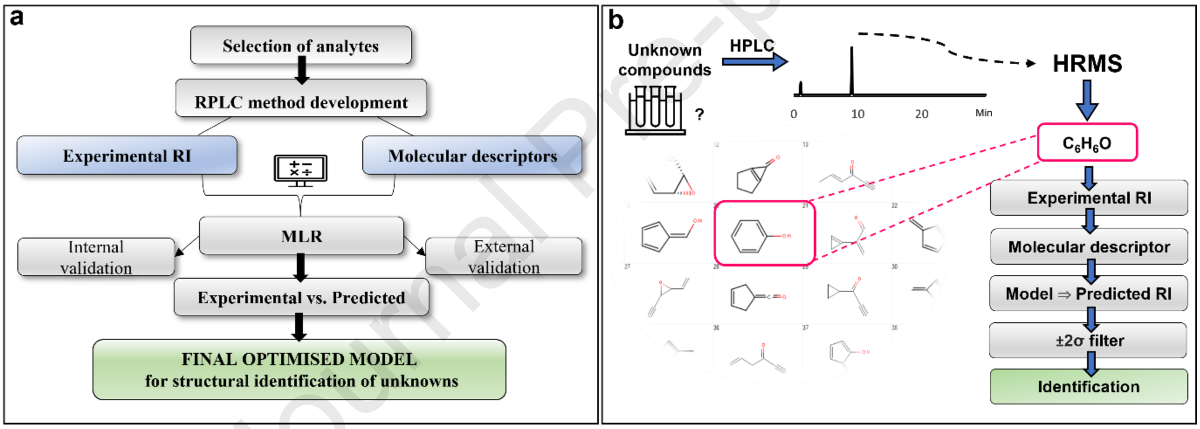


Figure 1. Representation of the workflow applied to develop the model (**a**) and of the proposed implementation by the user (**b**).

3.1 Selecting a generic RPLC method

Proposing “the most” universal RPLC column and method is inherently ambiguous as this depends on the geography or field of application. In this current work the implementation of acetonitrile/water/0.1% formic acid gradients on a core-shell RPLC is proposed for this purpose. RPLC is selected because it is the most broadly applied separation mode (Majors, 2018). A benefit of this mode is that it can retain both neutral compounds as such, or ionized solutes via protonation (for acids) or ion pairing (for bases) when using MS-compatible acidic conditions (e.g., with 0.1% formic acid). Because the versatility of RPLC inherently leads to large differences in polarity of the possible analytes, the use of gradients allowing both retention and elution of all solutes is essential. Acetonitrile is thereby the most suitable choice as it depicts high eluotropic strength, low viscosity, inertness, and excellent MS compatibility. A 150 x 2.1 mm ID core shell-based method was selected because it allows implementation in both HPLC and UHPLC while ensuring easy hyphenation to mass spectrometry. A core shell type of stationary phases was selected as such type is increasingly used, while being highly efficient at a lower pressure drop as compared to full porous particles (González-Ruiz et al., 2015; Tanaka and Mccalley, 2015).

3.2 Selection of compounds as a function of the molecular space of interest

Within a molecular space composed of only C, H and O up to 500 g mol^{-1} a large number of different elemental compositions can occur, leading to billions of possible corresponding structural formulas. Selection of the most representative data set is thereby inherently ambiguous and fraught with challenges. Emphasis was therefore set on the selection of compounds allowing broad coverage of the separation space.

A variety of conventional $C_xH_yO_z$ organic solutes, pharmaceuticals, compounds of environmental concerns were selected for this purpose, such as to cover the molecular space in the best possible way. Van Krevelen plots and ALogPS_logP vs. MW representations were used for this (Figure 2, Table S1). The former illustrates a broad coverage in the amount of unsaturations (from 1 for the ketone ladder compounds to 11 for alizarin) while spanning a fair polarity range reflected through the O/C ratio range from 0 (for e.g., toluene) up to 0.6 for 2,5 dihydroxy-benzoic acid. A reflection of the polarity and hence water solubility is also obtained through visualization of the logP's vs. the MW, where it can be seen that the logP's range from close to 0 up to 7 and in this way e.g., outspan the range of typical pharmaceutical solutes. This also covers the applicability range of gradient RPLC as more polar solutes (saccharides) would barely be retained and more apolar solutes (petrochemical compounds) require stronger elution conditions with less generic solvents. Highly oxygenated or unstable species (saccharides, peroxides, or aldehydes) were avoided due to low compound retention or stability issues involved. Additionally, the expected C, H, O, functional groups were comprised in the dataset (alcohols, carboxylic groups, ketones, esters, aromatic, linear, and branched solutes, etc.).

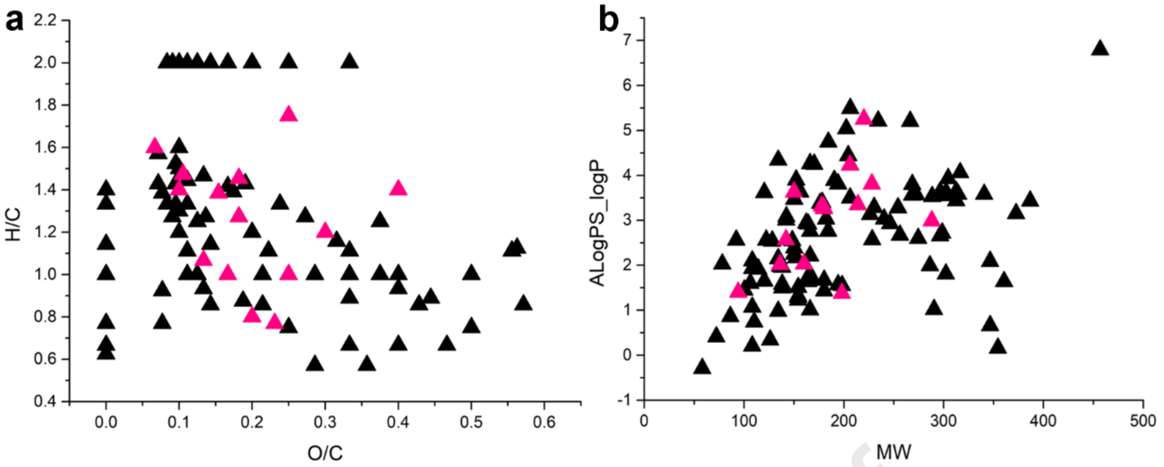


Figure 2. Representation of the (**a**) Van Krevelen plots (H/C vs. O/C ratio's) and of the (**b**) ALogPS_logP vs. MW of the 101 training set (black) and 14 test set (pink).

3.3 Data collection and conversion to retention indices

Although the purpose of this work was to introduce one broadly applicable gradient profile, the retention of 60 solutes was also measured (in triplicate) with 3 gradient profiles spanning the full elution range in 10, 20 or 40 min. (Section S2, Table S2). This to obtain insight in the robustness of the proposed method. The error on the repeatability of the retention times (n=3) was below 1% in all cases (and below 0.1% for 53% of the triplicate analyses) and hence in line with the expectations for HPLC. Subsequently, the data was converted to RI. This such as to allow easier method transfer and instrument independent model implementation. Although, this still imposes usage of the same stationary and mobile phase and to some extent gradient slope, it does allow disconnection from the column dimensions, flow rate, instrument and e.g., connection types used (Rigano et al., 2018). While the use of linear RI is an established approach, strongly supporting the identification of unknowns in gas chromatography, the field of HPLC has been mostly hindered by a lack of standardization on this issue. The latter is partially driven by the aspect that the relationship between RI and the carbon number in HPLC is quasilinear and not rectilinear as in GC (Rigano et al., 2018; Smith et al., 1987; Weitzel et al., 2011). Due the more complex elution process in RPLC in which the compound hydrophobicity is the main, but not the only, parameter controlling the elution, it is challenging, if not impossible, to identify a homologues series of detectable solutes generally depicting a completely linear behaviour over the entire elution range covered by the gradient. Problematic therein is that mere presence of a UVchromophore or API-MS compatible functional group in the calibration series affects linearity and hence limits the broadest possible implementation. Depending on the application in RPLC different types of calibration series have been proposed including alkan-2-ones, alkyl aryl ketones or 1-nitroalkenes (Baker, 1979; Baker and Ma, 1979; Bogusz and Aderjan, 1988; Bogusz and Wu, 1991; Smith, 1982). In this work the former ones are used (from 2-propanone to 2-dodecanone)(Baker and Ma, 1979). This because the alkyl aryl ketones comprise aromatic groups which are complicating the linearity between carbon number and retention. Also, the nitroalkenes are only incrementally useful for mapping the very polar solutes, a zone in which a hydrophobicity based predictive model is anyhow less performant (Baker and Ma, 1979; Bogusz and Aderjan, 1988; Smith, 1982). Because there has also been a lack of standardization in terms of the equation to be used to calculate the RI, the RI vs. carbon number plots were constructed for the alkan-2-ones ladder according to the various possible linearization methods (Figure S1). While none of the plots allows complete linearity it can be seen that over 90% of the plot excellent linearity is obtained and that only in the very low, below 2-butanone, or high retention regime, above 2-nonanone, a deviation is occurring. Considering that additionally linearity is a preferential but not an essential prerequisite for the use of RI, this data illustrates that use of RI in the proposed strategy and in RPLC is certainly a viable approach. Because several equations led to the same degree of linearity and/or the conventional gradient Kovats retention index Equation SE1 led to the highest correlation coefficient (0.97), to simplify the approach the latter was consequently used (Arigò et al., 2021).

3.4 Selection of a model and descriptor types

A QSRR method allows linking the molecular properties of an analyte to the chromatographic retention under given stationary and mobile phase conditions. Both linear models such as multiple linear regression (MLR) or Partial Least Squares regressions (PLS) or, nonlinear models, such as neural networks, have extensively been used for this purpose (Cirera-Domènech et al., 2013). ANN approach can be more flexible for modelling when using both linear and non-linear functions, but compared to MLR, the infrastructure is more complex (Ruggieri et al., 2005b). In the current work MLR modelling was selected as it allows obtaining robust, easy to reproduce via freely accessible software and therefore more transferable, models. A retention relationship (a linear equation) is thereby constructed between a dependent variable (RI), and multiple independent variables, comprising a limited number of molecular descriptors. The identities and weight of the optimal descriptors are selected during the construction of the model. During model usage the actual value of each descriptor is then a priori calculated via software for each structural formula to allow subsequent retention time/index prediction by simple completion of the linear MLR equation. The contemporary availability of over 5000 chemo-informatics based molecular descriptors makes selection of the most suitable ones an increasingly challenging task, whereby models can easily lead to erroneous predictions when descriptor selection is suboptimal. Note that if too many of the available descriptors are used when creating a model, this does not lead to better and higher accuracy, but to overfitting, narrowing down the implementation range of the equation instead of making it generic (Sagandykova and Buszewski, 2021). A variety of 2-dimensional (2D), 3-dimensional (3D) molecular descriptors or other descriptors (such as scaffolds and fingerprint types) can today be directly obtained through online chemical databases. In order to allow selection from the broadest possible and most recent set of molecular descriptors available, in this work they were obtained through the AlvaDesc application. Therein 1879 descriptors were selected in the initial pool providing structural information such as molecular topology, flexibility, geometry.

3.5 Optimized descriptors selection and model construction

The MLR model construction and subsequent descriptor selection was performed through the VEGA ZZ software, which also allowed obtaining up-front model validation information. The initial screening of descriptors involved evaluating zero variance, which means removing any feature that has the same value for all the samples, as it does not add any information for the model. The second step involves conducting a single-variable regression analysis, which helps in identifying the features that are most relevant to the output variable, where poorly correlated ones were excluded ($r^2<0.1$). 1800 of the 1879 descriptors were removed in this way as unable to contribute usefully to a combined MLR model. After conducting an evaluation of the variance inflation (VIF), it was determined that the remaining set contained collinear descriptors, leading to a reduction of the set to 37 relevant descriptors. The chemometrics used for feature selection in this study are commonly employed for high-dimensional data to eliminate descriptors with no variation and identify strongly correlated descriptors, while more advanced methods like principal component analysis, partial least squares, or random forest regression may be needed to capture complex relationships and non-linear interactions between descriptors (Danishuddin and Khan, 2016).

The most suitable MLR model was then obtained through the leave-one-out (LOO) optimization algorithm, including calculation and ranking of the figures of merit of each possible equation. The most excluded compound corresponded to a steroid testosterone undecanoate. The software itself allows models with up to 8 regressors, and all were tested. Finally, the best-optimized model (n-1), depicting a correlation r² = 0.93, was chosen comprising 7 variables, showing the lowest standard error of prediction (SE=58). Another algorithm (r² = 0.90) with 3 regressors was also observed and was suitable for the same purposes of this work but with a slightly larger deviation (Equation SE2). Note that, using high number of regressors can seemingly lead to enhanced models, but this could cause overfitting, and hence lead to a less generically applicable model and errors. Furthermore, using a single global model provides computational efficiency, simplified interpretation and implementation, versatility in handling various analytes, and the ability to identify trends and patterns across multiple analytes, outweighing the potential limitations of using local models and structurally similar training sets. In this way the following Equation 1 was obtained allowing implementation for the predicting the RI values of all possible hypothetically possible structural formulas for a given experimentally observed solute.

$$RI_{predicted} = 489.9565 + 358.6790\, ALogPS_{logP} - 465.4977\, ALogPS_{logS} - 249.3834\, Psi_{i_A} + 465.8030\, Chi_G + 304.7962\, RBN + 150.6071\, TDB06p + 144.8038\, LOC \qquad (1)$$

The model comprises 7 descriptors from which 2 ALogPS (ALogPS_logP and ALogPS_logS) and 5 AlvaDesc (Psi_i_A, Chi_G, RBN, TDB06p, and LOC). Unsurprisingly, a first descriptor selected therein is *ALogPS_logP*, representing the logarithm of the *n*-octanol/water partition coefficient. With a correlation of r^2 value of 0.78, in the single variable regression, it illustrates that indeed the hydrophobic retention on a highly endcapped silica based C18 column is mostly, but not only, based on compound’s lipophilicity. Although, logD might be a more expected solution if other elements such as nitrogen were also comprised, the full protonation of carboxylic groups under the used conditions ensures compound neutrality and hence the same result as one would expect with logD, while allowing use of the simpler, and hence somewhat more robust logP calculations (Dong et al., 2009). *ALogPS_logS*, another descriptor with high correlation value (0.63) represents aqueous solubility of a compound. As expected, more water soluble compounds proved less retained. Although, a collinearity with *ALogPS_logP* could be reasonably expected, statistically this was not the case (VIF<5) (Sun, 2004). *Psi_i_A* (intrinsic state pseudoconnectivity index – type S average), a third descriptor of the model from a group of topological indices, with an r^2 value of 0.62, was also withheld. These 2D descriptors (distance-, degree-, and spectrum-based), also known as connectivity indices, are based on the intrinsic and the electrotopological state values, which have shown beneficial correlations multiple times in literature when building QSAR, QSPR, or QSRR models (Chu et al., 2021; Ling et al., 2019). Furthermore, it was observed for *Chi_G* (Randic-like index from geometrical matrix, a 3D matrix-based descriptor, with r^2=0.34), that with increasing retention, the value drops. This descriptor could assist in distinguishing cyclic molecules (higher values) from more branched ones (lower values), as it contains the information of degree of branching as well as the molecular folding (Eichenlaub et al., 2022). The *RBN* (number of rotatable bonds) parameter, describes the number of any single bonds allowing the free rotation and is related to the size and flexibility of the molecule (Falcón-Cano et al., 2022). *TDB06p* (3D Topological distance based descriptors – lag 6 weighted by polarizability), a 3D autocorrelation type of descriptor, r^2=0.21, describing the shortest length distance between two atoms in a molecule with an emphasis to the polarizability of the molecule. Previous research showed that polarizability of a molecule can highly affect the elution order in RPLC (Andrade-Eiroa, 2011; Klein et al., 2004). Finally, the *LOC* (lopping centric index) descriptor belonging to the group of topological indices (with a correlation of 0.11) was the final descriptor selected in the model. Furthermore, it can represent the molecular branching degree, where the value increases with more branching graphs (Todeschini and Consonni, 2010; Yu, 2019).

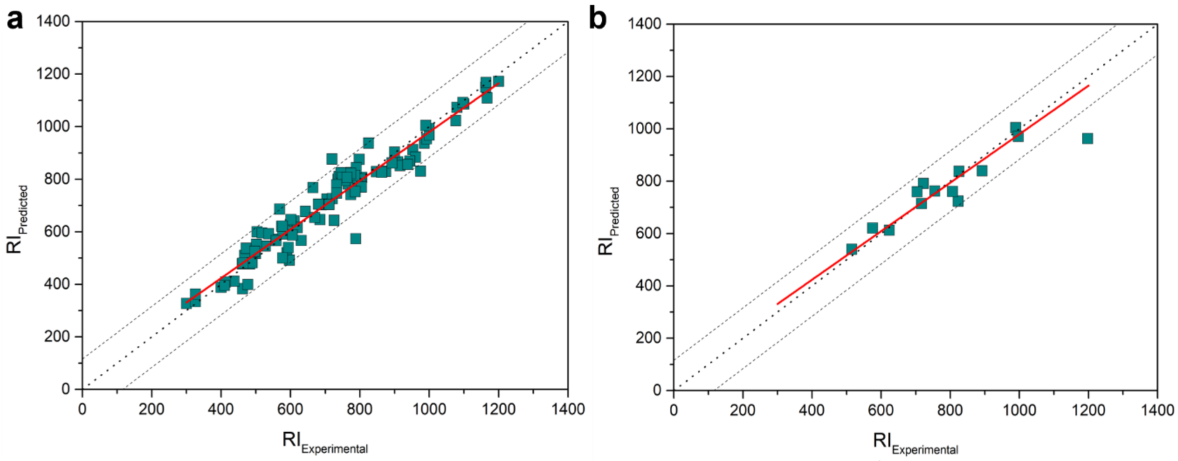


Figure 3. Linear fit displays RI predicted vs. RI experimental (**a**) for training set and (**b**) test set.

3.6 Model performance assessment with known environmentally relevant solutes

In Figure 3a the predicted RI as obtained via Equation 1 is represented versus the experimental RI for all 101 compounds, delivering a plot depicting a correlation of $r^2 = 0.93$. Only 4% of data points did not fit into ±2σ, and 74% fit ±1σ, where σ (SE=58) is the standard deviation of the errors (Table S3). In order to assess the predictive accuracy of the model with unrelated molecules from outside the training set, it was subsequently tested with 14 compounds of environmental and pharmaceutical relevance (Figure S3). An overlay of the thereby obtained predicted and experimental RI is shown in Figure 3b, where obtained results fitted the 95% confidence margin, except one compound, 2,6-di-tert-butyl-4-methyl phenol (BHT). The latter solute was, however, eluting after the latest eluting reference compound from the ketone ladder (2-dodecanone), and is therefore too retained to fall into the applicability range of the developed algorithm. The diverse test set were selected to span the molecular space as represented in Figure 2. As can be seen in Figure 4 representing the experimental vs the predicted retention (Table S8), 13 out of the 14 compounds meet the deviation margin, except for BHT depicting a larger error due to above-mentioned reason. Although, for BHT, the real RI is impossible to calculate, this was done by estimating the elution time of the next ketone elution in order. In general, it can, thus be concluded that for solutes falling into the range for which the model was designed (comprising only C, H, O, MW<500 and eluting between acetophenone and 2-dodecanone) that a predictive deviation ±2σ or ±116 RI is a realistic reliability threshold, which can be used in the structural elucidation work (Section 3.7).

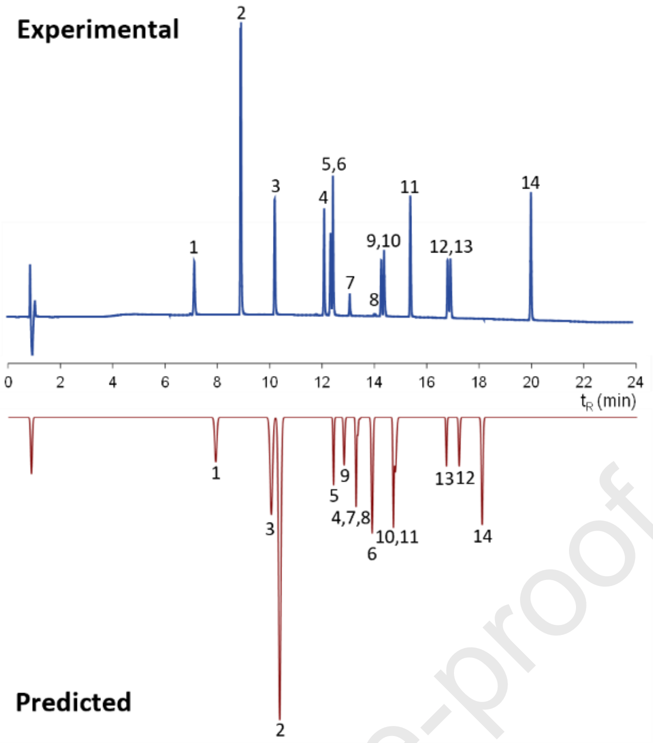


Figure 4. Overlay of the experimentally obtained retention of 14 solutes not used during the model design with the predicted values. Peak identity: 1) phenol; 2) 2,7-dihydroxy naphthalene; 3) toluic acid; 4) propylparaben; 5) 1,3 butanediol diacrylate; 6) bisphenol A; 7) testosterone; 8) trans-2-hexenyl acetate; 9) 3-tert-butyl-4-hydroxyanisole (BHA); 10) 4-ter-butyl benzoic acid; 11) diphenyl carbonate; 12) 4-hexylbenzoic acid; 13) butyl phenyl ether; 14) 2,6-di-tert-butyl-4-methyl phenol (BHT).

3.7 Implementation of the model to assist in de novo structural identification of unknown solutes by RPLC-HRMS

The actual goal of this work is to implement such models to support the structural elucidation process of unknowns for which the elemental compositions and experimental retention times/indices were obtained. The rationale is thereby that the developed model should be able to predict the retention index of every hypothetic structural formula that can be drawn for a given elemental composition, whereby the proposed structures eluting outside the ±2σ margin can be excluded upfront. The challenges therein are the astronomical number of possible structures that can be drawn for a given atomic composition and the current absence of embeddable algorithms allowing both generation of all structures and incorporation in the proposed workflow. Another approach can be the use of the publicly available libraries, which contain a number of possible known structures. While the former strategy is ideally preferential, in order to demonstrate the current possibilities of the model, the proposed hypothetic structures were in this case obtained from the ChemSpider database (Pogliani, 2000).

This approach was tested with all 14 compounds used in section 3.6 and 6 out of 14 were presented in the text below: phenol, propylparaben, diphenyl carbonate, BHA, p-toluic acid and 1,3 butanediol diacrylate (for the remaining 8 solutes see Table S9). These were used as “unknowns”, for which predicted elemental compositions were obtained. Using ChemSpider as a database source, for the phenol case, the elemental composition (C_6H_6O) led to 83 hypothetical structures. When plotting the corresponding RI for the obtained structures and definition of the ± 116 RI error margin zone above and below the experimental RI of the unknown (Figure 5a), it can be seen that only 17% of the proposed structures is eliminated. Specifically, the number of 83 possible structures thereby dropped to 69, one of which indeed corresponded to the predicted RI of phenol. The predicted RI (539) of the correct phenol structure thereby deviated 24 RI from the experimental value (515). While this illustrates that the proposed 1-D HPLC based predictive modelling method cannot on itself allow for sufficient attrition of all the incorrect structures, the proposed tool can be powerful in combination with the other available structural elucidation information. The structure of phenol can therefore, from the shortlisted of 69 solutes, subsequently be obtained via e.g., mass (MS/MS) and UV spectrometric information, via comparison with standards, but also through chemical stability assessments (as most, if not all, of the non-aromatic hypothetic structures are highly reactive). The database delivered 3576 possible structural formulas for the elemental composition ($C_{10}H_{12}O_3$) of propylparaben. It can be seen in Figure 5b that all structures deliver a RI <588 and >820, after applying the model, elimination of the erroneous structures obtained was 42%. The left over indeed comprises the correct structure of propylparaben depicting a ΔRI=55 between the experimental (704) and predicted (759) RI value. In a fully analogous way, all RI for the possible structural formulas corresponding to $C_{13}H_{10}O_3$ are represented in Figure 5c. Subsequent comparison with the experimental RI illustrates that the correct structure of diphenyl carbonate is included in a shortlist comprising only 136 of the 985 structures, corresponding to removal of 86% of the incorrect structures. Furthermore, Figure 5d shows another successful removal of 67% of impossible tentative structures for $C_{11}H_{16}O_2$, where 279 out of 849 remain for identification of BHA.

Somewhat less removal was obtained for p-toluic acid case, $C_8H_8O_2$ (Figure 5e). Out of 582 possible structures generated, 28% can be eliminated. While the predictive accuracy for this solute is good (ΔRI=12) the large number of possible RI in the vicinity of the correct structures leaves the user with (too) many remaining possible solutions. Lastly, for the following elemental composition, $C_{12}H_{18}O_4$, database comprised 2886 possible structural compositions. After applying a model, remaining 917 were left for further identification of 1,3 butanediol diacrylate allowing up to 68% of elimination (Figure 5f).

This limited number of examples illustrates the potential of the approach while proving the concept. A remaining hurdle with easy implementation of the ChromaRIM approach is the need to develop integrated software which can automatically generate all possible structures for a given atomic composition (or link to the public databases), calculate the corresponding descriptor values, generate the corresponding RI and eliminate all impossible (or at least improbable) ones in single automated procedure. While such integrated software is under development, the current work is mainly intending to introduce the principle, workflow, and an already applicable protocol to accept or exclude possible structures using ChromaRIM website (“Home page - ChromaRIM,” n.d.). It should be stressed that with the provided information the reader is already having all the required information to implement the tool for structural elucidation purposes. To assist this process, a user friendly standard operating procedure is therefore added with the supplementary information (section S5) to help the user in implementing the current model. Our website also contains an application allowing automated calculation of the RI based on the provided retention times, which will be further enhanced towards automated library search as this work is further progressing. Ideally in the future such an algorithm could be embedded in the LC-HRMS software for fully automated implementation.

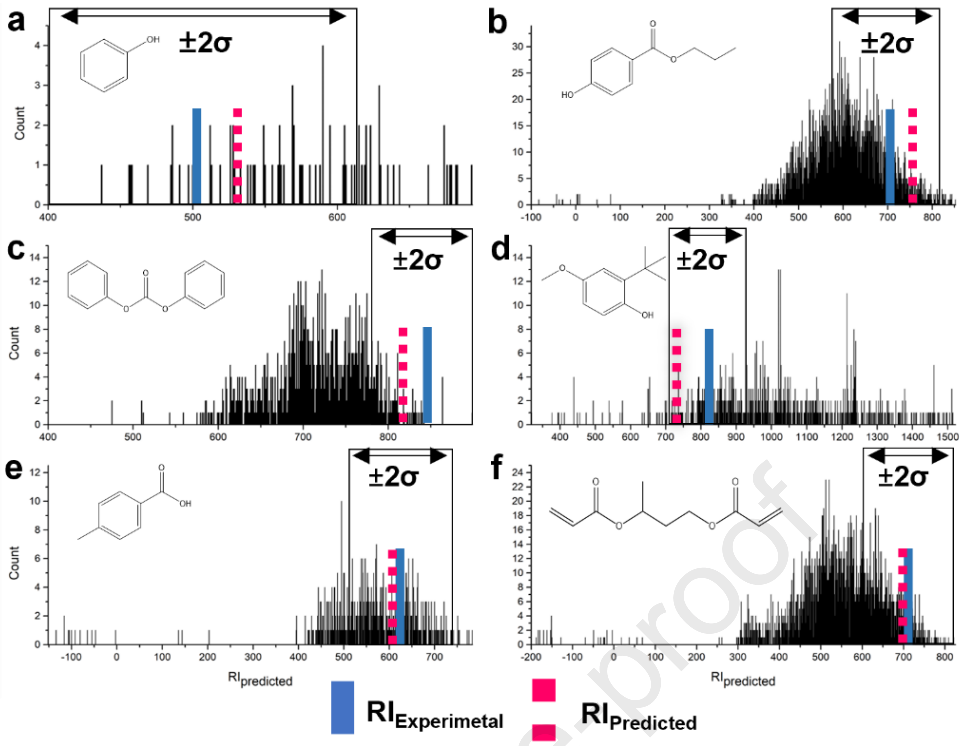


Figure 5. Representation of the calculated RI for all obtained structures from ChemSpider for (**a**) phenol, (**b**) propylparaben, (**c**) diphenyl carbonate, (**d**) 3-tert-butyl-4-hydroxyanisole (BHA), (**e**) p-toluic acid, and (**f**) 1,3 butanediol diacrylate. The zone eliminated by a ±2σ or ±116 RI deviation above and below the experimental value of the elution time (converted to RI) is indicated together with the experimental elution index and the predicted one for the correct structure.

4. Conclusions

In this work a QSRR methodology is developed to assist in the structural elucidation of unknown solutes composed of carbon, hydrogen, and oxygen with a molecular weight of up to a 500 g mol^{-1}. The methodology was specifically developed to be instrument-independent and hence fully and easily transferable and reproducible on any (U)HPLC-HRMS system. For this purpose, the predictive algorithm was developed on a broadly used reversed phase column (Kinetex, core-shell C18) with generic water/acetonitrile + 0.1% formic acid gradients covering the full range in eluotropic strengths. By data conversion to RI (with a ketone ladder) transferability is facilitated. An optimized multiple linear regression-based model was developed based on the retention of 101 training solutes, whereby an initial number of 1879 possible descriptors were screened. The latter were fine-tuned down to 7 remaining most influential descriptors in a linear equation allowing optimal prediction for all training solutes. This offers a model which can effectively be used for the prediction of the RI of unknowns within the predefined separation space. While, due to the sheer number of molecules in the latter is impossible to test the model with all solutes, the accuracy of the latter proved to allow correct RI prediction within a ±2σ range (mostly ±1σ) for all test solutes not included in the training set and eluting within the ketone ladder. This suggest that broad implementation of the model is foreseeable. The applicability of the model is demonstrated through the correct elimination of large fractions of all possible structural formulas for a given elemental composition, effectively simulating the situation one would be confronted with when performing LC-HRMS. In all the six treated examples the model allowed correct elimination of a significant percentage of the incorrect structural formulas, whereby the RI of the correct structure was always within the remaining possible structures. The tool therefore appears applicable to support the identification of unknown C, H, O containing solutes < 500 g mol^{-1}.

Author information

Corresponding Author

* Frédéric Lynen

Tel: +32 (0) 9 264 9606; Fax: +32 (0) 9 264 4998.

E-mail: frederic.lynen@ugent.be

Author Contributions

Ardiana Kajtazi: Conceptualization, Visualization, Methodology, Software, Investigation, Formal analysis, Writing - Original Draft; **Giacomo Russo:** Validation, Writing - Review & Editing; **Kristina Wicht:** Investigation; **Hamed Eghbali:** Supervision; **Frédéric Lynen:** Conceptualization, Visualization, Supervision, Writing - Review & Editing.

Acknowledgments

We thank mag.ing.comp. Marin Kajtazi for assistance in developing the ChromaRIM software. This project has received funding from the European Union’s EU Framework Programme for Research and

Innovation Horizon 2020 under Grant Agreement No 861369. (innoveox.eu). The FWO and the FNRS are acknowledged for funding part of this research through the Excellence of Science grant (30897864).

References

Aalizadeh, R., Alygizakis, N.A., Schymanski, E.L., Krauss, M., Schulze, T., Ibáñ Ez, M., Mceachran, A.D., Chao, A., Williams, A.J., Gago-Ferrero, P., Covaci, A., Moschet, C., Young, T.M., Hollender, J., Slobodnik, J., Thomaidis, N.S., 2021. Development and Application of Liquid Chromatographic Retention Time Indices in HRMS-Based Suspect and Nontarget Screening. Cite This Anal. Chem 93, 11601–11611. https://doi.org/10.1021/acs.analchem.1c02348

Aicheler, F., Li, J., Hoene, M., Lehmann, R., Xu, G., Kohlbacher, O., 2015. Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches. Anal. Chem. 87, 7698–7704. https://doi.org/10.1021/acs.analchem.5b01139

Amos, R.I.J., Haddad, P.R., Szucs, R., Dolan, J.W., Pohl, C.A., 2018. Molecular modeling and prediction accuracy in Quantitative Structure-Retention Relationship calculations for chromatography. TrAC - Trends Anal. Chem. 105, 352–359. https://doi.org/10.1016/j.trac.2018.05.019

Andrade-Eiroa, A., 2011. Reverse-High Performance Liquid Chromatography Mechanism Explained by Polarization of Stationary Phase. CheM 1, 62–79. https://doi.org/10.5618/chem.2011.v1.n1.8

Arigò, A., Dugo, P., Rigano, F., Mondello, L., 2021. Linear retention index approach applied to liquid chromatography coupled to triple quadrupole mass spectrometry to determine oxygen heterocyclic compounds at trace level in finished cosmetics. J. Chromatogr. A 1649, 462183. https://doi.org/10.1016/j.chroma.2021.462183

Baker, J.K., 1979. Estimation of high pressure liquid chromatographic retention indices. Anal. Chem. 51, 1693–1697. https://doi.org/10.1021/ac50047a025

Baker, J.K., Ma, C.Y., 1979. Retention index scale for liquid-liquid chromatography. J. Chromatogr. A 169, 107–115. https://doi.org/10.1016/0021-9673(75)85036-9

Bogusz, M., Aderjan, R., 1988. Improved standardization in reversed-phase high-performance liquid chromatography using 1-nitroalkanes as a retention index scale. J. Chromatogr. A 435, 43–53.

https://doi.org/10.1016/S0021-9673(01)82161-0

Bogusz, M., Wu, M., 1991. Standardized HPLC/DAD system, based on retention indices and spectral library, applicable for systematic toxicological screening. J. Anal. Toxicol. 15, 188–197. https://doi.org/10.1093/jat/15.4.188

Boiteau, R.M., Hoyt, D.W., Nicora, C.D., Kinmonth-Schultz, H.A., Ward, J.K., Bingol, K., 2018. Structure elucidation of unknown metabolites in metabolomics by combined NMR and MS/MS prediction. Metabolites 8. https://doi.org/10.3390/metabo8010008

Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L., Degroeve, S., 2021. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat. Methods 2021 1811 18, 1363–1369. https://doi.org/10.1038/s41592-021-01301-5

Chu, Y.M., Julietraja, K., Venugopal, P., Siddiqui, M.K., Prabhu, S., 2021. Degree- and irregularity-based molecular descriptors for benzenoid systems. Eur. Phys. J. Plus 2021 1361 136, 1–17. https://doi.org/10.1140/EPJP/S13360-020-01033-Z

Cirera-Domènech, E., Estrada-Tejedor, R., Broto-Puig, F., Teixidó, J., Gassiot-Matas, M., Comellas, L., Lliberia, J.L., Méndez, A., Paz-Estivill, S., Delgado-Ortiz, M.R., 2013. Quantitative structure–retention relationships applied to liquid chromatography gradient elution method for the determination of carbonyl-2,4-dinitrophenylhydrazone compounds. J. Chromatogr. A 1276, 65–77. https://doi.org/10.1016/J.CHROMA.2012.12.027

Codesido, S., Randazzo, G.M., Lehmann, F., González-Ruiz, V., García, A., Xenarios, I., Liechti, R., Bridge, A., Boccard, J., Rudaz, S., 2019. DynaStI: A Dynamic Retention Time Database for Steroidomics. Metab. 2019, Vol. 9, Page 85 9, 85. https://doi.org/10.3390/METABO9050085

Cui, L., Lu, H., Lee, Y.H., 2018. Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases. Mass Spectrom. Rev. 37, 772–792. https://doi.org/10.1002/MAS.21562

Danishuddin, Khan, A.U., 2016. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov. Today 21, 1291–1302. https://doi.org/10.1016/j.drudis.2016.06.013

Datta, R., Das, D., Das, S., 2021. Efficient lipophilicity prediction of molecules employing deep-learning models. Chemom. Intell. Lab. Syst. 213, 104309. https://doi.org/10.1016/j.chemolab.2021.104309

De Vijlder, T., Valkenborg, D., Lemière, F., Romijn, E.P., Laukens, K., Cuyckens, F., 2018. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. Mass Spectrom. Rev. 37, 607–629. https://doi.org/10.1002/MAS.21551

Domingo-Almenara, X., Guijas, C., Billings, E., Montenegro-Burke, J.R., Uritboonthai, W., Aisporna, A.E., Chen, E., Benton, H.P., Siuzdak, G., 2019. The METLIN small molecule dataset for machine learning-based retention time prediction. Nat. Commun. 2019 101 10, 1–9. https://doi.org/10.1038/s41467-019-13680-7

Dong, P.P., Ge, G.B., Zhang, Y.Y., Ai, C.Z., Li, G.H., Zhu, L.L., Luan, H.W., Liu, X.B., Yang, L., 2009. Quantitative structure-retention relationship studies for taxanes including epimers and isomeric metabolites in ultra fast liquid chromatography. J. Chromatogr. A 1216, 7055–7062. https://doi.org/10.1016/J.CHROMA.2009.08.079

Dorfer, V., Maltsev, S., Winkler, S., Mechtler, K., 2018. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. J. Proteome Res. 17, 2581–2589. https://doi.org/10.1021/acs.jproteome.7b00836

Eichenlaub, J., Rakowska, P.W., Kloskowski, A., 2022. User-assisted methodology targeted for building structure interpretable QSPR models for boosting CO2 capture with ionic liquids. J. Mol. Liq. 350, 118511. https://doi.org/10.1016/J.MOLLIQ.2022.118511

Eugster, P.J., Boccard, J., Debrus, B., Bréant, L., Wolfender, J.L., Martel, S., Carrupt, P.A., 2014. Retention time prediction for dereplication of natural products (CxHyOz) in LC-MS metabolite profiling. Phytochemistry 108, 196–207. https://doi.org/10.1016/j.phytochem.2014.10.005

Falcón-Cano, G., Molina, C., Cabrera-Pérez, M.Á., 2022. Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches. Pharm. 2022, Vol. 14, Page 1998 14, 1998. https://doi.org/10.3390/PHARMACEUTICS14101998

González-Ruiz, V., Olives, A.I., Martín, M.A., 2015. Core-shell particles lead the way to renewing high-performance liquid chromatography. TrAC Trends Anal. Chem. 64, 17–28. https://doi.org/10.1016/J.TRAC.2014.08.008

Gritti, F., 2023. Perspective on the Future Approaches to Predict Retention in Liquid Chromatography 15, 39. https://doi.org/10.1021/acs.analchem.0c05078

Haddad, P.R., Taraji, M., Szücs, R., 2021a. Prediction of Analyte Retention Time in Liquid Chromatography. Anal. Chem. https://doi.org/10.1021/acs.analchem.0c04190

Haddad, P.R., Taraji, M., Szücs, R., 2021b. Prediction of Analyte Retention Time in Liquid Chromatography. Anal. Chem. 93, 228–256. https://doi.org/10.1021/acs.analchem.0c04190

Héberger, K., 2007. Quantitative structure–(chromatographic) retention relationships. J. Chromatogr. A 1158, 273–305. https://doi.org/10.1016/j.chroma.2007.03.108

Home page - ChromaRIM [WWW Document], n.d. URL https://chromarim.ugent.be/ (accessed 9.27.22).

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, Kenichi, 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, Ken, 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, 703–714. https://doi.org/10.1002/JMS.1777

Hutchins, P.D., Russell, J.D., Coon, J.J., 2018. LipiDex: An Integrated Software Package for High-Confidence Lipid Identification. Cell Syst. 6, 621-625.e5. https://doi.org/10.1016/J.CELS.2018.03.011

Kaliszan, R., 1993. Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography. J. Chromatogr. A 656, 417–435. https://doi.org/10.1016/0021-9673(93)80812-M

Klein, C.T., Kaiser, D., Ecker, G., 2004. Topological Distance Based 3D Descriptors for Use in QSAR and Diversity Analysis. J. Chem. Inf. Comput. Sci. 44, 200–209. https://doi.org/10.1021/CI0256236/ASSET/IMAGES/LARGE/CI0256236F00004.JPEG

Köhler, H.-R., Gräff, T., Schweizer, M., Blumhardt, J., Burkhardt, J., Ehmann, L., Hebel, J., Heid, C., Kundy, L., Kuttler, J., Malusova, M., Moroff, F.-M., Schlösinger, A.-F., Schulze-Berge, P., Panagopoulou, E.I., Damalas, D.E., Thomaidis, N.S., Triebskorn, R., Maletzki, D., Kühnen, U., von der Ohe, P.C., 2023. LogD-based modelling and ΔlogD as a proxy for pH-dependent action of ionizable chemicals reveal the relevance of both neutral and ionic species for fish embryotoxicity and possess great potential for practical application in the regulation of chemicals. Water Res. 235, 119864. https://doi.org/10.1016/j.watres.2023.119864

Kumari, S., Stevens, D., Kind, T., Denkert, C., Fiehn, O., 2011. Applying in-silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. Anal. Chem. 83, 5895–5902. https://doi.org/10.1021/ac2006137

Ling, Y., Klemes, M.J., Steinschneider, S., Dichtel, W.R., Helbling, D.E., 2019. QSARs to predict adsorption affinity of organic micropollutants for activated carbon and B-cyclodextrin polymer adsorbents. Water Res. 154, 217–226. https://doi.org/10.1016/J.WATRES.2019.02.012

Liu, Y., Romijn, E.P., Verniest, G., Laukens, K., De Vijlder, T., 2019. Mass spectrometry-based structure elucidation of small molecule impurities and degradation products in pharmaceutical development. TrAC - Trends Anal. Chem. 121. https://doi.org/10.1016/J.TRAC.2019.115686

Ma, C., Ren, Y., Yang, J., Ren, Z., Yang, H., Liu, S., 2018. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. Anal. Chem. 90, 10881–10888. https://doi.org/10.1021/ACS.ANALCHEM.8B02386/SUPPL_FILE/AC8B02386_SI_001.PDF

Majors, R., 2018. HPLC and UHPLC Columns: Then, Now, Next. LC-GC North Am. 36, 128–132.

Mauri, A., 2020. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Methods Pharmacol. Toxicol. 801–820. https://doi.org/10.1007/978-1-0716-0150-1_32/FIGURES/3

Meshref, S., Li, Y., Feng, Y.L., 2020. Prediction of liquid chromatographic retention time using quantitative structure-retention relationships to assist non-targeted identification of unknown metabolites of phthalates in human urine with high-resolution mass spectrometry. J. Chromatogr. A 1634. https://doi.org/10.1016/j.chroma.2020.461691

Palmblad, M., RAMSTROM, M., BAILEY, C., MCCUTCHENMALONEY, S., BERGQUIST, J., ZELLER, L., 2004. Protein identification by liquid chromatography–mass spectrometry using retention time prediction. J. Chromatogr. B 803, 131–135. https://doi.org/10.1016/j.jchromb.2003.11.007

Pedretti, A., Mazzolari, A., Gervasoni, S., Fumagalli, L., Vistoli, G., 2021. The VEGA suite of programs: An versatile platform for cheminformatics and drug design projects. Bioinformatics 37, 1174–1175. https://doi.org/10.1093/bioinformatics/btaa774

Pogliani, L., 2000. Modeling with molecular pseudoconnectivity descriptors. A useful extension of the intrinsic I-state concent. J. Phys. Chem. A 104, 9029–9045. https://doi.org/10.1021/JP001191V/ASSET/IMAGES/MEDIUM/JP001191VU00015A.GIF

Randazzo, G.M., Tonoli, D., Hambye, S., Guillarme, D., Jeanneret, F., Nurisso, A., Goracci, L., Boccard, J., Rudaz, S., 2016a. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. Anal. Chim. Acta 916, 8–16. https://doi.org/10.1016/j.aca.2016.02.014

Randazzo, G.M., Tonoli, D., Hambye, S., Guillarme, D., Jeanneret, F., Nurisso, A., Goracci, L., Boccard, J., Rudaz, S., 2016b. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. Anal. Chim. Acta 916, 8–16. https://doi.org/10.1016/J.ACA.2016.02.014

Rigano, F., Oteri, M., Russo, M., Dugo, P., Mondello, L., 2018. Proposal of a Linear Retention Index System for Improving Identification Reliability of Triacylglycerol Profiles in Lipid Samples by Liquid Chromatography Methods. Anal. Chem. 90, 3313–3320. https://doi.org/10.1021/acs.analchem.7b04837

Ruggieri, F., D'Archivio, A.A., Carlucci, G., Mazzeo, P., 2005a. Application of artificial neural networks for prediction of retention factors of triazine herbicides in reversed-phase liquid chromatography. J. Chromatogr. A 1076, 163–169. https://doi.org/10.1016/j.chroma.2005.04.038

Ruggieri, F., D'Archivio, A.A., Carlucci, G., Mazzeo, P., 2005b. Application of artificial neural networks for prediction of retention factors of triazine herbicides in reversed-phase liquid chromatography. J. Chromatogr. A 1076, 163–169. https://doi.org/10.1016/J.CHROMA.2005.04.038

Russo, G., Grumetto, L., Szucs, R., Barbato, F., Lynen, F., 2017. Determination of in Vitro and in Silico Indexes for the Modeling of Blood−Brain Barrier Partitioning of Drugs via Micellar and Immobilized Artificial Membrane Liquid Chromatography. https://doi.org/10.1021/acs.jmedchem.6b01811

Sagandykova, G., Buszewski, B., 2021. Perspectives and recent advances in quantitative structure-retention relationships for high performance liquid chromatography. How far are we? TrAC Trends Anal. Chem. 141, 116294. https://doi.org/10.1016/J.TRAC.2021.116294

Smith, R.M., 1982. Alkylarylketones as a retention index scale in liquid chromatography. J. Chromatogr. A 236, 313–320. https://doi.org/10.1016/S0021-9673(00)84880-3

Smith, R.M., Murilla, G.A., Burr, C.M., 1987. Alkyl aryl ketones as a retention index scale with acetonitrile or tetrahydrofuran containing eluents in reversed-phase high-performance liquid chromatography. J. Chromatogr. A 388, 37–49. https://doi.org/10.1016/S0021-9673(01)94464-4

Sun, H., 2004. A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption. https://doi.org/10.1021/ci030304f

Szucs, R., Brown, R., Brunelli, C., Heaton, J.C., Hradski, J., 2021. Structure Driven Prediction of Chromatographic Retention Times: Applications to Pharmaceutical Analysis. Int. J. Mol. Sci. 22, 3848. https://doi.org/10.3390/ijms22083848

Tanaka, N., Mccalley, D. V, 2015. Core−Shell, Ultrasmall Particles, Monoliths, and Other Support Materials in High-Performance Liquid Chromatography. https://doi.org/10.1021/acs.analchem.5b04093

Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V.A., Radchenko, E. V., 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. https://doi.org/10.1007/S10822-005-8694-Y

Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K.K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., Cox, J., 2019. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat. Methods 2019 166 16, 519–525. https://doi.org/10.1038/s41592-019-0427-6

Todeschini, R., Consonni, V., 2010. Molecular Descriptors for Chemoinformatics, Molecular Descriptors for Chemoinformatics. https://doi.org/10.1002/9783527628766

Virtual Computational Chemistry Laboratory [WWW Document], n.d. URL http://www.vcclab.org/ (accessed 9.27.22).

Weitzel, K., Chemie, F., Rev, M.S., Introduction, I., Reference, C., 2011. Bond-Dissociation Energies of Cations — Pushing the. WHO Libr. Cat. Data 221–235. https://doi.org/10.1002/mas

Wen, Y., Amos, R.I.J., Talebi, M., Szucs, R., Dolan, J.W., Pohl, C.A., Haddad, P.R., 2019. Retention prediction using quantitative structure-retention relationships combined with the hydrophobic subtraction model in reversed-phase liquid chromatography. Electrophoresis 40, 2415–2419. https://doi.org/10.1002/elps.201900022

Wen, Y., Amos, R.I.J., Talebi, M., Szucs, R., Dolan, J.W., Pohl, C.A., Haddad, P.R., 2018a. Retention Index Prediction Using Quantitative Structure-Retention Relationships for Improving Structure Identification in Nontargeted Metabolomics. Anal. Chem. 90, 9434–9440. https://doi.org/10.1021/acs.analchem.8b02084

Wen, Y., Talebi, M., Amos, R.I.J., Szucs, R., Dolan, J.W., Pohl, C.A., Haddad, P.R., 2018b. Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model. J. Chromatogr. A 1541, 1–11. https://doi.org/10.1016/j.chroma.2018.01.053

Witting, M., Böcker, S., 2020. Current status of retention time prediction in metabolite identification. J. Sep. Sci. 43, 1746–1754. https://doi.org/10.1002/jssc.202000060

Xu, Z., Chughtai, H., Tian, L., Liu, L., Roy, J.F., Bayen, S., 2023. Development of quantitative structure-retention relationship models to improve the identification of leachables in food packaging using non-targeted analysis. Talanta 253. https://doi.org/10.1016/j.talanta.2022.123861

Yang, Q., Ji, H., Lu, H., Zhang, Z., 2021. Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification. Anal. Chem. https://doi.org/10.1021/acs.analchem.0c04071

Yu, X., 2019. Prediction of Depuration Rate Constants for Polychlorinated Biphenyl Congeners. ACS Omega 4, 15615–15620. https://doi.org/10.1021/ACSOMEGA.9B02072/ASSET/IMAGES/MEDIUM/AO9B02072_M010.GIF

Zheng, S.J., Liu, S.J., Zhu, Q.F., Guo, N., Wang, Y.L., Yuan, B.F., Feng, Y.Q., 2018. Establishment of Liquid Chromatography Retention Index Based on Chemical Labeling for Metabolomic Analysis. Anal. Chem. 90, 8412–8420. https://doi.org/10.1021/acs.analchem.8b00901

Highlights

- Chromatographic Retention Index Model (ChromaRIM) in RPLC-HRMS
- Structural elucidation of small environmental solutes assisted by developed model
- Supporting unknown identification of $C_xH_yO_z$ molecules < 500 Da
- The model implementation was demonstrated with 6 relevant compounds
- Elimination of a significant % of the incorrect structural formulas was achieved

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: