Trends in Analytical Chemistry 105 (2018) 352–359

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

journal homepage: www.elsevier.com/locate/trac

Molecular modeling and prediction accuracy in Quantitative Structure-Retention Relationship calculations for chromatography

Ruth I.J. Amos [a, *], Paul R. Haddad [a], Roman Szucs [b], John W. Dolan [c], Christopher A. Pohl [d]

[a] *Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences-Chemistry, University of Tasmania, Private Bag 75, Hobart 7001, Australia*
[b] *Pfizer Global Research and Development, Sandwich, UK*
[c] *LC Resources, 1795 NW Wallace Rd., McMinnville, OR 97128, USA*
[d] *Thermo Fisher Scientific, Sunnyvale, CA, USA*

ARTICLE INFO

Article history:
Available online 2 June 2018

Keywords:
QSRR
Geometry optimization
Descriptor generation
Chiral descriptors
Feature selection
Retention prediction
Similarity
Non-Targeted Analysis
Chromatography

ABSTRACT

Quantitative Structure-Retention Relationship (QSRR) methodology is a useful tool in chromatography of all kinds, allowing the prediction of analyte retention time and providing insight into the mechanisms of separation. The prediction of retention is useful in reducing method development time and identifying analytes in Non-Targeted Analysis. The varying methods used for geometry optimization, descriptor calculation, feature selection, and model generation in many different QSRR settings are investigated and compared. It is found that the method of geometry optimization and descriptor selection is of less importance than the chromatographic similarity of compounds in the training sets used for model building in order to reduce the error of the model.

1. Introduction

Quantitative Structure-Activity Relationship models are well known and utilized in the chemical and biological sciences. In these mathematical models, a relationship between the physico-chemical characteristics of a molecule and a particular activity is described – often a biological activity as in the case of drug design. Since the 1980s the same methodology has been applied to separation science, producing Quantitative Structure-Retention Relationships (QSRR) models where retention is related to the characteristics of molecules [1]. Two very comprehensive earlier reviews of QSRR can be found in the works of Heberger [2] and Kaliszan [3].

QSRR has been applied to many types of chromatography including Reversed Phase Liquid Chromatography (RPLC) [4–9], Hydrophilic Interaction Liquid Chromatography (HILIC) [10–14], Ion Exchange Chromatography (IC) [15–17], Thin Layer Chromatography (TLC) [18–20], Gas Chromatography [21–23], and others. As a technique it is useful for the prediction of the retention of analytes, and for help in understanding the mechanisms of retention in the different techniques.

QSRR can be used to predict the retention of a target compound either as an end in itself, or in order to gain greater understanding of the mechanisms of retention. A training set of analytes with known retention times (t_R) under a specific set of chromatographic conditions is required to build a mathematical model which relates defined characteristics of the training set analytes to their t_R values and this model is then applied to predict the retention of a target compound of known chemical structure. Researchers use QSRR models to predict t_R, but the models can also predict retention factor (k), log k_w (where the log k values are extrapolated to a 0% organic mobile phase) and Retention Index (RI). Measured t_R values as dependent variables give a preferred outcome as the back transformation of log k to t_R increases the error of the predicted value [24].

The characteristics of the training set of analytes utilized in these models are defined using mathematical descriptors derived from chemical structures that depict the physico-chemical aspects of the analytes as quantitative values [25]. To provide the information for calculating two- and three-dimensional descriptors, a method for optimizing the geometry of each analyte is required so that the molecule is represented correctly. These methods vary from very simple 2D maps to complex quantum calculations and

[*] Corresponding author.
E-mail address: Ruth.Amos@utas.edu.au (R.I.J. Amos).

https://doi.org/10.1016/j.trac.2018.05.019
0165-9936/© 2018 Elsevier B.V. All rights reserved.

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
BBB	Blood Brain Barrier
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Analysis
DFT	Density Functional Theory
$Ecom_{50}$	the energy needed to produce 50% fragmentation
GA	Genetic Algorithm
GC	Gas Chromatography
HILIC	Hydrophilic Interaction Liquid Chromatography
HOMO	Highest Occupied Molecular Orbital
HSM	Hydrophobic Subtraction Model
k	retention factor
log k	log of retention factor
log k_w	log of retention factor extrapolated to 0% organic phase
LSS	Linear Solvent Strength model
LSER	Linear Solvation Energy Relationship
LUMO	Lowest Unoccupied Molecular Orbital
MAE	Mean Absolute Error
MIA	Multivariate Image Analysis
MLR	Multiple Linear Regression
MM	Molecular Modeling
MS	Mass Spectrometry
NTA	Non-Targeted Analysis
PAH	Poly-Aromatic Hydrocarbons
PLS	Partial Least Squares
QSiAR	Quantitative Similarity Activity Relationships
QSRR	Quantitative Structure-Retention Relationships
RI	Retention Index
RMSEP	Root Mean Square Error of Prediction
RPLC	Reversed Phase Liquid Chromatography
SE	Semi-Empirical
SMILES	Simplified Molecular Input Line Entry System
SWATH	Sequential Windowed Acquisition of All Theoretical Fragment Ion
t_R	Retention time

can be performed with or without a correction for the presence of solvent [26].

Once the geometry is optimized the descriptors for that geometry can be calculated. The number of descriptors can exceed 5000 for a single molecule [27], therefore a method of feature selection is necessary to decide which descriptors are important in describing the retention of the molecule [28]. Feature selection is either performed by the scientist using chemical knowledge, or it is performed by a statistical package with machine learning capability to exclude some descriptors and utilize others in the final function for retention [28]. The retention of the molecule is then described as a function of the descriptors using a regression method. Multiple Linear Regression (MLR) is the simplest method and produces a linear function where the descriptors included in the function have coefficients giving some idea of the importance of those descriptors, and whether they lead to an increase or decrease in retention. This makes MLR the regression method of choice when the QSRR is being used for mechanistic understanding [18,29]. However, it has been observed that the use of a greater number of descriptors gives a better prediction of retention, and MLR can suffer when the number of descriptors exceeds the number of objects [15]. In addition MLR does not cope well with collinearity [24,30] so another method of regression needs to be utilized.

Partial least squares (PLS) regression takes advantage of the collinearity of independent variables [24,30]. The mechanistic understanding for PLS regression is diluted compared to MLR. However, methods such as variable importance to projection (VIP) analysis, or an analysis of the beta coefficient of the PLS help with this issue [10]. Other examples of regression methods include back propagation artificial neural networks [4] and support vector machine regression [5,31] though these are less popular methods.

There are many methods for reporting the accuracy of prediction in QSRR – Mean Absolute Error (MAE), Root Mean Square Error of Prediction (RMSEP), the correlation coefficient R or the square of R, to name a few [32]. In this article as many error reports as possible have been converted to the RMSEP scaled to retention time [33]:

$$RMSEP\% = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{yi-\widehat{yi}}{yi}\right)^{2}}{n}} \times 100(\%)$$

where y*i* and ŷ*i* are the observed and predicted retention times and *n* is the number of test analytes.

In some instances, this conversion has been impossible due to a lack of information in the source document, so in those cases the reported error measure is utilized.

As computational resources increase, the use of *in silico* methods to add value to chemical research is also increasing. While computational predictions must always be compared to the experimental outcomes, the use of QSRR is an arrow in our quiver that should not be overlooked when investigating separation science.

2. Modeling methods

2.1. Global and local models

The mathematical models used for QSRR can be global or local models. In a global model the complete dataset of analytes is divided into a training set (70%–90% of analytes) and a test set (the remaining analytes) [6,19,21]. A single mathematical model is built using the training set, and that model is then applied to the test set for external validation [6,21]. The dataset can also be divided into three sets: calibration, prediction, and test sets [34] (alternatively these divisions can be called training, validation, and test sets [35]). In this case the calibration set is used to build the mathematical model, the prediction set deals with any overfitting generated, and the test set is used as external validation [34]. For all modeling it is necessary that the test sets have no role in the building of the original model so that they are a true external validation of the model [35].

The disadvantage of global models is that the model is not specific to the target analyte and can contain noise as well as useful information, leading to larger errors. Local models are utilized to reduce noise in the modeling. In some cases, the words ‘local model’ merely indicate the division of the whole data set into segments such as acids, bases, and neutral compounds, where the segments are each divided into training and test sets, and a single model is created pertaining to each segment [6]. However, the most effective form of a local model is to create a novel mathematical model for the prediction of the retention time of each unique analyte [6,10–12,36–38]. In this case, the analyte is utilized as a target, a small training set of around five molecules is selected from the database and used to create the mathematical model, then the t_R of the target is predicted. It has been found that errors of t_R prediction decrease with the use of local models as the training sets are more specific to the target analyte and do not consider unnecessary factors [10–12,36–38].

2.2. Descriptors and geometry optimization

Descriptors can be simple (e.g. the number of oxygen atoms in the molecule), or complex (e.g. the autocorrelation of lag 2 weighted by I-state). Some descriptors only need the molecule to be defined by a Simplified Molecular Input Line Entry System string (SMILES), however if the descriptor is a two-dimensional (2D) descriptor then a 2D map of the molecule is needed, and similarly for 3D descriptors where a three dimensional structure of the molecule is required.

The steps of geometry optimization are summarized in Fig. 1. Initially the geometry is drawn out and converted to a two-dimensional map. Alternatively the SMILES string can be created or downloaded from a database, such as PubChem. After that if three-dimensional geometry optimization is required there are several choices from a simple empirically based 3D geometry to optimization using high level quantum calculations. Many different programs are able to perform geometry optimization.

Some descriptors, such as Abraham solute descriptors, can be measured experimentally and the values utilized in a QSRR equation [39]. However, experimental values are not available for all analytes and therefore calculated descriptors are necessary to supplement or replace experimental descriptors. Some different options for software used for calculating molecular descriptors are listed in Table 1 along with studies that have utilized these options. This is not an exhaustive list and it is very difficult to discern which of these descriptor calculation methods is the best. It is likely that some methods are more suited to particular applications, or that more than one descriptor method is necessary to build an accurate model.

While it is important to make use of technological advances such as SE or DFT calculations, it is a misuse of time if the descriptors used for retention prediction are unable to take 3D geometries or charge states into account [23]. For example, the most complex Dragon descriptors are optimized using molecular modeling technology [25] and therefore combining Dragon with quantum geometry optimization such as DFT does not increase the accuracy of the model. A comparison of DFT and Dragon descriptors was performed by predicting retention indices for polyaromatic hydrocarbons (PAHs) in gas chromatography (GC). In the final models only one parameter was utilized and the mean absolute error was the same regardless of whether that parameter was a quantum descriptor or a topological descriptor [22].

The use of a single method for descriptor calculation and geometry optimization may be too simplistic. Rouille et al. have utilized a variety of SE and DFT methods to predict the retention times of PAHs in reversed-phase liquid chromatography (RPLC) [56]. They focused on descriptors such as isotropic polarizability, dipole moment, submolecular polarity, and a topological polarity index. Interestingly, the authors compared calculated values for polarizability with known experimental values for unsubstituted PAHs and found that the SE methods (AM1, PM3, PM6 and PM7) needed correction (and were sometimes unreliable even with the correction) and that the B3LYP/6-31 + G(d,p) calculation was more accurate due to the diffuse functions involved. Similarly, DFT or PM6 or PM7 were necessary for the dipole moment calculation, whereas DFT significantly overestimated values of subpolarity. The best results were found when using a retention time prediction equation made from DFT calculations of polarizability and the topological index or from using AM1 or PM3 models using polarizability, topological index, and submolecular polarity [56]. A summary of different methods and software packages for geometry optimization can be found in Table 2.

2.3. Solvent correction

The earliest use of solvent corrections for QSRR studies, according to our knowledge, was in 2015 where the COSMO [64] correction in Molecular Orbital PACkage (MOPAC) [65] was utilized for an investigation of the effect of chaotropic salts in hydrophilic interaction liquid chromatography (HILIC) [31]. As many forms of chromatography are conducted using a solvent as the mobile phase it is reasonable that the effects of solvent should be included in the calculations. However, one study compared complex solution calculations (DFT with corrections for water as solvent), with simpler and less time-consuming methods (descriptors from ACDlabs) and found no great increase in accuracy with calculation time [41].

A study from our group utilized DFT calculations corrected for water (as the solvent) for three QSRR studies of HILIC systems. The RMSEP was as low as 3% for the best predicted stationary phase model, with an average RMSEP% over all stationary phases of 5.46%.

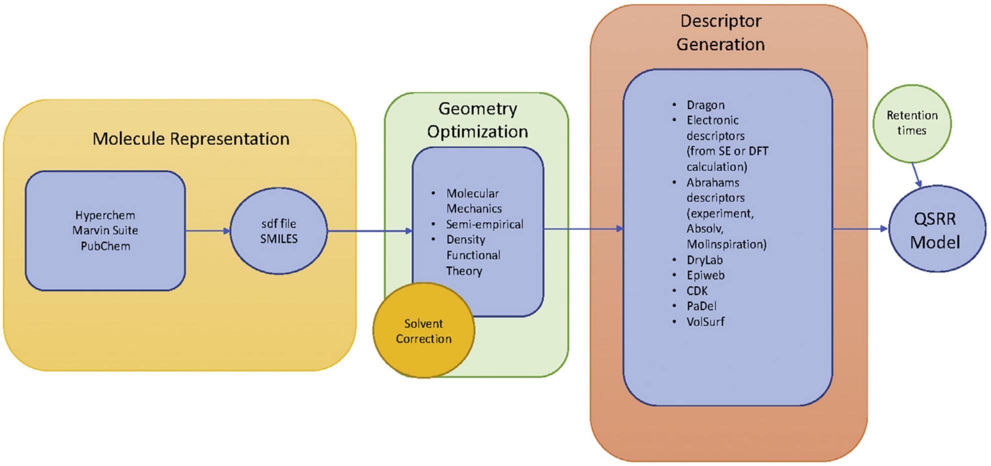

Fig. 1. Methods for preparation of input for QSRR models. For information about studies utilizing these methods or websites please see Table 1.

Table 1
Examples of software for descriptor calculation and selected references outlining the use of these descriptors in QSRR calculations.

Descriptor software	Address	Comments and uses	References
Absolv Webboxes	ACD/Labs, Advanced Chemistry Development, ON, Canada	Estimating experimental descriptors	[7,40]
Advanced Chemistry Development	Talete, Milan, Italy	Building models, predicting protonation, optimizing using semi-empirical (SE) methods, and calculating descriptors	[18,24,41,42]
Cambridge Soft Corporation	Perkin Elmer Inc	Building models, predicting protonation, optimizing using semi-empirical (SE) methods, and calculating descriptors	[42,43]
ChemAxon	Budapest, Hungary	Building models, predicting protonation, optimizing using semi-empirical (SE) methods, and calculating descriptors	[18,42,44]
Chemistry Development Kit	https://cdk.github.io/	Open source collection of modular Java libraries	[13,45]
Dragon	Talete, Milan, Italy	Over 5000 descriptors, from simple to complex, calculated using a wide variety of methods	[4,27,29,41]
DryLab	Molnar Institute, Berlin, Germany	Lipophilicity descriptors	[29,46]
Epiweb 4.1	United States Environmental Protection Agency	Estimating experimental descriptors	[47,48]
Molecular Modeling Program Plus	MMP Plus	Descriptors related to quantum calculations of the geometry	[20]
Molinspiration Cheminformatics	molinspiration.com	Estimating experimental descriptors	[24,42,48,49]
PaDel	http://padel.nus.edu.sg/software/padeldescriptor	Open source. Calculates molecular descriptors and fingerprints	[4,31,50,51]
Schrödinger Suite	Schrödinger, LLC, New York, NY	Building models, predicting protonation, optimizing using semi-empirical (SE) methods, and calculating descriptors	[48,52]
Semi-empirical or Density Functional Theory calculations		Quantum descriptors calculated using various methods	[20,22,41]
VolSurf+	Molecular Discovery Ltd., Hertfordshire, UK	Volume and surface effects	[6,53]
winMolconn v2.1	Hall Associates Consulting: Quincy, MA, 2011	Structure information representation descriptors	[54,55]

However, this low error may be due to more factors than just the geometry optimization method. When a global model (with solvent correction) utilizing all molecules in the database as the training set was used the error was 46.55%. The use of clustered training sets for local models reduced the error to 27.10% and the use of a dual filter to further optimize the clustering was the final factor that led to such a low error in t_R prediction [10–12].

To clarify this issue we compared geometry optimization methodology from MM, to SE, DFT and finally DFT with solvent correction to see if a reduction in error was obtained [26]. The findings showed no change in the mean absolute error for any of the methods, possibly because the majority of the error in each calculation was due to the error inherent in the calculation of the Dragon descriptors utilized for all QSRR models [26].

2.4. Chiral descriptors

Many of the descriptors utilized for QSRR models are unable to distinguish between chiral compounds. This becomes important when working with biochemicals and therefore chiral descriptors of various kinds have been developed. Two-dimensional chirality descriptors use connectivity indices or topological maps and add a chirality correction such as −1 for S or +1 for R [66]. Chiral corrections can also be added to atom-pair descriptors [67]. Some 3D descriptors are able to distinguish between isomers with limited success by taking into account the polar volume of the molecule or the fractional accessible surface area due to atoms with partial charges [4] or the flexibility and globularity of the molecule [39]. Comparative molecular field analysis (CoMFA) [68], as well as similar 3D and 4D methods e.g. CoMSIA [69], and 3D QSiAR [70], require exhaustive conformational analysis and a spatial alignment of the molecules to describe the chirality. These methods therefore do not work as well with high throughput methods and are used more for QSAR than QSRR.

A novel method for model creation was able to distinguish between chiral molecules as well as giving a very low t_R error [71]. Multivariate image analysis (MIA) is a purely mathematical method using the position of pixels to describe the molecule. The MIA model gave a RMSEP of 1.8% for a set of chiral 1-(2-naphthyl)-1-ethanol ester derivatives, distinguishing between (*R*) and (*S*) isomers. As there were no polarity or charge descriptors this method did not give insight into the mechanism of retention, however it was very successful for retention time prediction [71].

2.5. Feature selection

Often when using MLR models the descriptors are chosen by the researcher according to their prior chemistry knowledge [14,39] although to perform this successfully requires a situation where the retention mechanism is understood fully. Because this is rarely the case, on most occasions some form of statistical selection of the most important descriptors is therefore employed [9,72]. This process is known as feature selection. PLS regression allows inclusion of a greater number of descriptors [30] but it has been shown that PLS with some form of descriptor selection that excludes unnecessary or misleading descriptors gives more accurate results [28] and that a combination of feature selection methods improves prediction accuracy and reduces the amount of descriptors used for the model [28].

Buszewski et al. have published a series of studies comparing conventional MLR and PLS regression [17] to the use of PLS with some machine learning or artificial intelligence (AI) such as an

Table 2
Different methods and software packages for geometry optimization.

Name of method	References	URL/address
Hyperchem professional software	[18,20,29,57,58]	Hypercube, Gainesville, USA
Marvin suite	[31,42]	ChemAxon, Budapest, Hungary
PubChem	[13,24]	pubchem.ncbi.nlm.nih.gov
Molecular mechanics	[4,20,31,57,58]	
Semi-empirical	[5,8,18–20,23,28,29,31,34,36,56,57,59–62]	
Density Functional Theory	[11,22,26,41,56,63]	
Solvent correction	[10–12,26,31,41]	

Table 3
Examples of feature selection methods utilized for QSRR.

Feature selection method	Overview of process	Use
Ant colony optimization	Based on the behavior of real ants ACO solves optimization problems by mimicking the process ants use to find the shortest path to food. A benefit to ACO as opposed to ANN or GA is that it can adjust to a change in conditions in real time.	[14]
Artificial neural networks	Inspired by the animal brain, ANNs consist of nodes called neurons and connections called synapses. The strength of the synapse connection is weighted and the weights increase or decrease as the system learns. ANNs have layers and some of the neuron layers are hidden.	[2,9,12,20,46,52,53,68,70–76]
Best first	Starts with a single descriptor that is compares with the output and then performs a backward and forward search to find a better solution, adding a single descriptor each time.	[12]
Competitive adaptive reweighted sampling	Iterative process where variables with small coefficients and variables with less frequency of selection are discarded and the rest fed into the next iteration.	[6]
Genetic algorithm	Data are encoded in chromosomes. A pair of chromosomes is subjected to cross-over and mutation to produce offspring. Offspring with better fitness replace the parents.	[4,6,11,12,19,22–26,51,69]
Greedy stepwise	Starts from an arbitrary point and performs a forward search through the descriptor matrix stopping when the addition of a descriptor results in a decrease in the accuracy of the solution.	[12]
Iteratively retaining informative variables	All variables are mutated and the importance of each variable is determined by comparing error before and after mutation. Only informative variables are kept for the next iteration.	[6]
Least Absolute Shrinkage and Selection Operator (LASSO)	A method for both variable selection and regression. Plots the mean squared error against a parameter encoding the number of descriptors. The U-shaped graph decreases at first but then increases with overfitting. The minimum gives the best number of descriptors.	[43,77]
Linear forward	Is an extension of Best First using a restricted number of descriptors rather than a single descriptor.	[12]
Random forest	A number of decision trees makes up the forest where decision trees are prone to overfitting, Random Forests average the decision trees and reduce the overfitting to the training set.	[36]
Uninformative variable elimination	Uses Monte-Carlo sampling, augments the variable matrix with a noise matrix, then uses a mean to standard deviation threshold to remove variables irrelevant to the response.	[6]
Variable iterative space shrinkage approach	Models are created and those with the lowest error are used to calculate weights of all variables. Variables with larger weights have a greater possibility of survival through the iterations.	[6]

artificial neural network (ANN) or GA [16], and finally adding fuzzy logic [15]. The RMSEP decreased from 18.7% to 7.1% with the addition of AI but the inclusion of fuzzy logic did not make a significant improvement. The authors suggested that with greater computational power allowing faster calculation, the improvement achieved by using fuzzy logic will be greater as more descriptors could be used in the modeling process [15].

Table 3 summarizes some of the important feature selection methods used in QSRR.

3. Retention prediction

One use of QSRR models is to predict the retention of a target molecule using only the chemical structure of that analyte, and mathematical models trained by a set of molecules where the t_R is known. This would lead to a reduction in experimental time and expense as *in silico* methods could be utilized in place of trial and error experimental methods for method development [28]. In order for a prediction to be useful, the error in t_R prediction needs to be lower than ±5% relative to the experimental elution time [73] and the total acceptable error is dependent on the complexity of the chromatogram – a more complex chromatogram needs a lower relative error in prediction to be accurate.

A simple and straightforward method for QSRR (one might say the standard starting point) is elegantly detailed in the work of Cheng and Zhang studying aromatic components of red raspberry. The different components, 47 in all, were modeled using DFT, and seven molecular descriptors (both quantum and physico-chemical) were calculated. A global MLR model was utilized for t_R prediction and interpreted to show that steric factors, the total energy gap between HOMO and LUMO, and the melting point were important factors in retention. The error in the test predictions was 6.6% [21].

As t_R is predicted using a set of chromatographic conditions that is unique to the training set, the prediction of t_R is generally only valid for that set of chromatographic conditions and is not able to be extended to other conditions [13,24]. However, many researchers use porting equations to update expected retention and recalibrate columns [19] and also to compare different types of retention, such as micellar and hydro-organic chromatography [74] and this porting can also be applied to QSRR equations. Eugster et al. states that using an equation to change kinetic parameters such as flow-rate, gradient span, and slope, could allow transfer to other chromatographic conditions [24]. In addition, QSRR has been utilized in RPLC to predict, not the retention time, but the coefficients for the Hydrophobic Subtraction Model (HSM):

$$\log\alpha \equiv \log\left(\frac{k}{k_{EB}}\right) = \eta'\mathbf{H} - \sigma'\mathbf{S}^{*} + \beta'\mathbf{A} + \alpha'\mathbf{B} + \kappa'\mathbf{C} \quad (1)$$

where α is the chromatographic selectivity, k is the retention factor of the solute, and k_{EB} is the retention factor of ethylbenzene, and the coefficients $\eta'\mathbf{H}$ represent hydrophobic interactions, $\sigma'\mathbf{S}^{*}$, steric resistance, $\beta'\mathbf{A}$, hydrogen-bond acidity, $\alpha'\mathbf{B}$, hydrogen-bond basicity, and $\kappa'\mathbf{C}$, ionic interactions of the analyte and column respectively. The use of this model allows transfer of data to any of the stationary phases where the HSM coefficients are known [6]. QSRR has also been used in ion-exchange chromatography to predict the coefficients in the linear solvent strength (LSS) model:

$$\mathrm{Log}\ k = a - b\log\left[\mathrm{E}^{y-}\right] \quad (2)$$

where $[E^{y-}]$ is the molar concentration of the eluent competing ion and a and b values are respectively the intercept and slope [19]. The prediction of the a and b values allows the prediction of retention under any eluent concentration. Data independent acquisition such as SWATH has also been utilized where the MS/MS fragmentation patterns are used for peak identification independent of retention time [7].

3.1. Applicability domains and similarity

Authors often apply QSRR modeling to a particular set of molecules. Some authors desire their models to be as wide ranging as possible and so choose a diverse group of compounds [5,75]. Others might choose pharmaceutical compounds that cross the Blood Brain Barrier (BBB) [76], or compare compounds used for dyeing [77], or for herbicide properties [23], or as painkillers [59]. Another author might be investigating e.g. styryl lactones [42] or aromatic compounds [8] and therefore the molecules chosen would all have similar backbones. The choice of training set has a large effect on the outcome of the modeling process as has been shown by past Applicability Domain studies [77–79]. To get the best t_R prediction for a target molecule, the training set of molecules must be similar to the target analyte [10,19,37,38]. While many errors for t_R prediction are between 20% and 30%, training sets where the molecules are similar can reduce the error dramatically [10,19].

One such study utilized a dataset of sartans that was comprised of only six compounds but was extended to a training set of 70 using varied experimental conditions. Using the LSER equation, the retention time of the sartans was predicted, with excellent accuracy (RMSEP = 3.2%) [80]. It is interesting to note that the LSER is often denigrated for lack of accuracy [2,39] and therefore the excellent error value may be due to the similarity in the training set. However, this approach is not always successful and while the importance of an external test set cannot be overemphasized, the test set must be similar to the training set to get the benefits of similarity [79]. Chen et al. [75] addressed the problem of similarity by removing compounds from the training set that were ‘outliers’ in terms of their error in the training set. However, the compounds in the test set having the largest errors had similar structures to the outliers that had been removed from the training set. It is possible that division of the diverse compounds in the study into smaller and more similar training and test sets would have led to lower errors. Another form of clustering or compound classification is to apply an external analysis such as the seven main classes in the Dictionary of Natural Products [81]. This was applied during QSRR work for NTA of natural products giving greatly improved results with the Q^2 of the test set increasing from 0.88 to 0.92 [24].

If molecules are chromatographically similar, they will interact comparably with the mobile and stationary phases and will be eluted at similar times. Chromatographic similarity, found by comparing *k* values for molecules, has been shown to give very low errors in local models for retention prediction [11,26,37,38,82]. This method goes further than Applicability Domains as it uses similarity of compounds to form an individual training set for each target compound. However, in practice the retention time of the target compound is not known, therefore chromatographic similarity has to be estimated by some other method. Taraji et al. [11](investigating HILIC systems) and Park et al. [37](working with ion-exchange chromatography) have used dual filters to estimate the chromatographic similarity of compounds and therefore make sensible local model training sets for novel test compounds with average errors of 10.5% (down from 46% for a global model) and 6.0% respectively [11,37].

4. QSRR and quality by design (QbD) principles

A technique that combines both the compound classification theory, and the transfer of the QSRR model to other mobile and stationary phases is the combination of QSRR and Design of Experiments (DoE) according to QbD principles. This combination of techniques allows a design space for the separation to be determined, and an optimal region for separation to be found using the smallest possible number of experiments [12,83,84]. Software such as DryLab [85] and ChromSword [86] routinely use DoE principles to predict a robust design space for separations. Muteki et al. utilized compound classification, QbD, and QSRR to predict new compound targets under known conditions, predict known targets under new conditions, and optimize separation conditions in the areas of super-critical fluid chromatography (SFC) and RPLC in a foundational study [83]. Taraji et al. have performed similar work in the area of HILIC, finding optimal conditions for separation with excellent agreement between experimental and predicted retention [12].

5. Non-Targeted Analysis

QSRR and retention prediction is beginning to be applied to the area of metabolomics and NTA. NTA is used to identify compounds of interest from a metabolomics study, the difficulty being the identification of a compound having a known retention time and exact mass, due to the likelihood of having compounds with the same molecular formula but a different arrangements of atoms (and therefore different retention times) in biological systems [87]. Ideally, standards for the suspected compounds are synthesized and retention times and fragmentation patterns compared, however, this can be unsustainable for many laboratories due to time and expense [87]. QSRR can be used to remove false positive compounds of the exact same mass as the target compound by predicting their retention to be out of the range of the target's retention time. This requires a certain level of accuracy because the wider the confidence interval for the predicted retention time, the fewer false positives can be removed [87].

Several different methods have been utilized to calculate the confidence interval for removal of false positives. One group added the standard deviation to the results to give a window [13]. Another provided results in a statement that 93% of the standards were within 35% of predicted retention times [87] and a third gave a predicted retention window of 4 min [7]. The latter approach used the prediction as a ranking tool rather than an identification tool and used other parameters like MS/MS fragmentation to aid in identification [7].

Hall et al. used a combination of QSRR and Collision Induced Dissociation spectra prediction as well as a prediction of the energy needed to produce 50% fragmentation ($Ecom_{50}$) as their combination of identification factors. A combination of the three identifiers reduced a possible field of 315 molecules to 11 possibilities in a human serum sample, and helped to unambiguously differentiate among three structurally similar isomers [88]. Utilizing a training set of 1955 molecules and an ANN model, Hall et al. was then able to increase the applicability domain and make good predictions of RI for a test set of 202 new compounds showing the importance of large databases for QSRR modeling of real world problems [54].

6. Future perspectives

While the use of QSRR can decrease the speed of analytical method development, as a technique it has some distance to go in terms of prediction accuracy, and it is unlikely that it will ever be able to replace experiment completely. As can be seen from the work above, QSRR techniques combined with experimental methods can produce high levels of accuracy and the use of QSRR is a worthy addition to separation science. The use of MIA in the prediction of retention for chiral compounds shows that some ‘thinking outside the box’ may help in the identification of novel techniques for retention prediction [71]. Widespread adoption of QSRR will probably not occur until the problem of application to multiple chromatography conditions is addressed.

As it stands, QSRR is a useful technique, however one drawback is the lack of large retention databases allowing similarity searching and local models to be built. It is to be hoped that the scientific community will be willing to share data to allow more accurate retention prediction modelling.

7. Conclusions

QSRR is a powerful technique allowing the prediction of retention and reducing method development time in chromatography. It has been applied to every area of chromatography with varying levels of success. QSRR is still becoming established as a technique, with no specific methods of geometry optimization, descriptor calculation, feature selection, or model generation emerging as the preferred approaches. It is to be hoped that some methods will be optimized and routinely used by researchers as continuous efforts are made to improve QSRR methodology.

Funding

The authors acknowledge the Australian Research Council for the financial support of this research by an ARC Linkage Projects grant (LP120200700).

References

[1] R. Kaliszan, Quantitative Structure-chromatographic Retention Relationships, in: Chemical Analysis Series, vol. 93, Wiley, New York, 1987.

[2] K. Heberger, Quantitative structure-(chromatographic) retention relationships, J. Chromatogr. A 1158 (2007) 273–305.

[3] R. Kaliszan, QSRR: quantitative structure-(chromatographic) retention relationships, Chem. Rev. 107 (2007) 3212–3246.

[4] T.B. Oliveira, L. Gobbo-Neto, T.J. Schmidt, F.B. Da Costa, Study of chromatographic retention of natural terpenoids by chemoinformatic tools, J. Chem. Inf. Model. 55 (2015) 26–38.

[5] M. Goodarzi, R. Jensen, Y. Vander Heyden, QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 910 (2012) 84–94.

[6] Y. Wen, M. Talebi, R.I.J. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, P.R. Haddad, Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model, J. Chromatogr. A 1541 (2018) 1–11.

[7] T. Bruderer, E. Varesio, G. Hopfgartner, The use of LC predicted retention times to extend metabolites identification with SWATH data acquisition, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 1071 (2017) 3–10.

[8] R. Kaliszan, K. Osmialowski, S.A. Tomellini, S.H. Hsu, S.D. Fazio, R.A. Hartwick, Non-empirical descriptors of sub-molecular polarity and dispersive interactions in reversed-phase HPLC, Chromatographia 20 (1985) 705–708.

[9] R. Kaliszan, M.A. van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure-retention relationships, J. Chromatogr. A 855 (1999) 455–486.

[10] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures, J. Chromatogr. A 1486 (2017) 59–67.

[11] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Use of dual-filtering to create training sets leading to improved accuracy in quantitative structure-retention relationships modelling for hydrophilic interaction liquid chromatographic systems, J. Chromatogr. A 1507 (2017) 53–62.

[12] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Rapid method development in hydrophilic interaction liquid chromatography for pharmaceutical analysis using a combination of quantitative structure-retention relationships and design of experiments, Anal. Chem. 89 (2017) 1870–1878.

[13] M. Cao, K. Fraser, J. Huege, T. Featonby, S. Rasmussen, C. Jones, Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics, Metabolomics 11 (2015) 696–706.

[14] C. West, E. Auroux, Deconvoluting the effects of buffer salt concentration in hydrophilic interaction chromatography on a zwitterionic stationary phase, J. Chromatogr. A 1461 (2016) 92–97.

[15] S. Ukic, M. Novak, A. Krilic, N. Avdalovic, Y. Liu, B. Buszewski, T. Bolanca, Development of gradient retention model in ion chromatography. Part III: fuzzy logic QSRR approach, Chromatographia 78 (2015) 889–898.

[16] S. Ukic, M. Novak, A. Vlahovic, N. Avdalovic, Y. Liu, B. Buszewski, T. Bolanca, Development of gradient retention model in ion chromatography. Part II: artificial intelligence QSRR approach, Chromatographia 77 (2014) 997–1007.

[17] S. Ukic, M. Novak, P. Zuvela, N. Avdalovic, Y. Liu, B. Buszewski, T. Bolanca, Development of gradient retention model in ion chromatography. Part I: conventional QSRR approach, Chromatographia 77 (2014) 985–996.

[18] K. Ciura, A. Rutecka, P. Kawczak, J. Nowakowska, Quantitative structure-retention relationship modeling of the retention behavior of selected antipsychotic drugs in normal-phase thin-layer chromatography, J. Planar Chromatogr. – Mod. TLC 30 (2017) 225–230.

[19] S.H. Park, P.R. Haddad, M. Talebi, E. Tyteca, R.I.J. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model, J. Chromatogr. A 1486 (2017) 68–75.

[20] N.R. Stevanovic, D.S. Peruskovic, U.M. Gasic, V.R. Antunovic, A.D. Lolic, R.M. Baosic, Effect of substituents on prediction of TLC retention of tetradentate Schiff bases and their Copper(II) and Nickel(II) complexes, Biomed. Chromatogr. 31 (2017) e3810.

[21] L.P. Cheng, X.L. Zhang, QSRR study on GC retention time of aromatic components in red raspberry wine, Adv. Mater. Res. 781–784 (2013) 1434–1438.

[22] J.C. Drosos, M. Viola-Rhenals, R. Vivas-Reyes, Quantitative structure-retention relationships of polycyclic aromatic hydrocarbons gas-chromatographic retention indices, J. Chromatogr. A 1217 (2010) 4411–4421.

[23] S. Asadpour, M. Chamsaz, M.J. Haron, Application of MLR, PLS and artificial neural networks for prediction of GC/ECD retention times of chlorinated pesticides, herbicides and organohalides, Res. J. Pharm. Biol. Chem. Sci. 3 (2012) 850–860.

[24] P.J. Eugster, J. Boccard, B. Debrus, L. Breant, J.-L. Wolfender, S. Martel, P.-A. Carrupt, Retention time prediction for dereplication of natural products (CxHyOz) in LC-MS metabolite profiling, Phytochemistry 108 (2014) 196–207.

[25] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2000.

[26] R.I.J. Amos, E. Tyteca, M. Talebi, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl, Benchmarking of computational methods for creation of retention models in quantitative structure-retention relationships studies, J. Chem. Inf. Model. 57 (2017) 2754–2762.

[27] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon software: an easy approach to molecular descriptor calculations, MATCH 56 (2006) 237–248.

[28] M. Talebi, G. Schuster, R.A. Shellie, R. Szucs, P.R. Haddad, Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography, J. Chromatogr. A 1424 (2015) 69–76.

[29] P. Szatkowska-Wandas, M. Koba, G. Smolinski, J. Wandas, QSRR and QSAR studies of antitumor drugs in view of their biological activity prediction, Med. Chem. 12 (2016) 592–600.

[30] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (1984) 735–743.

[31] J. Colovic, A. Malenovic, M. Kalinic, S. Eric, A. Vemic, Investigation into the phenomena affecting the retention behavior of basic analytes in chaotropic chromatography: Joint effects of the most relevant chromatographic factors and analytes' molecular properties, J. Chromatogr. A 1425 (2015) 150–157.

[32] A. Golbraikh, A. Tropsha, Beware of q2!, J. Mol. Graphics Modell. 20 (2002) 269–276.

[33] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Error measures in quantitative structure-retention relationships studies, J. Chromatogr. A 1524 (2017) 298–302.

[34] H. Karimi, H. Noorizadeh, A. Farmany, A QSRR modeling of hazardous psychoactive designer drugs using GA-PIS and L-M ANN, ISRN Chromatogr. 2012 (2012) 838432–838440.

[35] D.L.J. Alexander, A. Tropsha, D.A. Winkler, Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models, J. Chem. Inf. Model. 55 (2015) 1316–1322.

[36] S.H. Park, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan, Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Relationships for retention prediction. III Combination of Tanimoto similarity index, logP, and retention factor ratio to identify optimal analyte training sets for ion chromatography, J. Chromatogr. A 1520 (2017) 107–116.

[37] S.H. Park, M. Talebi, R.I.J. Amos, E. Tyteca, P.R. Haddad, R. Szucs, C.A. Pohl, J.W. Dolan, Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Relationships for retention prediction. II Use

of Tanimoto similarity index in ion chromatography, J. Chromatogr. A 1523 (2017) 173–182.
[38] E. Tyteca, M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad, Towards a chromatographic similarity index to establish localized Quantitative Structure-Retention Models for retention prediction: use of retention factor ratio, J. Chromatogr. A 1486 (2016) 50–58.
[39] S. Khater, M.-A. Lozac'h, I. Adam, E. Francotte, C. West, Comparison of liquid and supercritical fluid chromatography mobile phases for enantioselective separations on polysaccharide stationary phases, J. Chromatogr. A 1467 (2016) 463–472.
[40] Absolv. http://www.acdlabs.com/products/percepta/predictors/absolv/ (Accessed January 2018).
[41] L. Kubik, P. Wiczling, Quantitative structure-(chromatographic) retention relationship models for dissociating compounds, J. Pharm. Biomed. Anal. 127 (2016) 176–183.
[42] M.Z. Karadzic, D.M. Loncar, S.Z. Kovacevic, L.R. Jevric, S.O. Podunavac-Kuzmanovic, G. Benedekovic, I. Kovacevic, V. Popsavin, A comparative study of chromatographic behavior and lipophilicity of selected natural styryl lactones, their derivatives and analogues, Eur. J. Pharm. Sci. 105 (2017) 99–107.
[43] CambridgeSoft PerkinElmer. http://www.cambridgesoft.com/ (Accessed January 2018).
[44] ChemAxon. https://chemaxon.com/ (Accessed January 2018).
[45] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, The chemistry development kit (CDK): an open-source Java library for chemo- and bioinformatics, J. Chem. Inf. Comput. Sci. 43 (2003) 493–500.
[46] DryLab. http://molnar-institute.com/drylab/ (Accessed February 2018).
[47] US EPA, Estimation Programs Interface Suite™ for Microsoft® Windows, V. 4.1, United States Environmental Protection Agency, Washington, DC, USA, 2018.
[48] S. Segan, I. Opsenica, M. Zlatovic, D. Milojkovic-Opsenica, B. Solaja, Quantitative structure retention/activity relationships of biologically relevant 4-amino-7-chloroquinoline based compounds, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 1012–1013 (2016) 144–152.
[49] Molinspiration. http://molinspiration.com (Accessed January 2018).
[50] PaDel Descriptor Software. http://www.yapcwsoft.com/dd/padeldescriptor/ (Accessed January 2018).
[51] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474.
[52] Schrodinger Software. https://www.schrodinger.com/suites (Accessed January 2018).
[53] G. Cruciani, P. Crivori, P.A. Carrupt, B. Testa, Molecular fields in quantitative structure-permeation relationships: the VolSurf approach, J. Mol. Struct. THEOCHEM 503 (2000) 17–30.
[54] L.M. Hall, D.W. Hill, K. Bugden, S. Cawley, L.H. Hall, M.-H. Chen, D.F. Grant, Development of a reverse phase HPLC retention index model for nontargeted metabolomics using synthetic compounds, J. Chem. Inf. Model. 58 (2018) 591–604.
[55] winMolconn, version 1.2.2.1, http://www.molconn.com/products.html (Accessed May 2018), Hall Associates Consulting: Quincy, MA, 2011.
[56] G. Rouille, C. Jaeger, F. Huisken, T. Henning, R. Czerwonka, G. Theumer, C. Boerger, I. Bauer, H.-J. Knoelker, Quantitative structure-retention relationships for polycyclic aromatic hydrocarbons and their oligoalkynyl-substituted derivatives, ChemistryOpen 6 (2017) 519–525.
[57] S. Studzinska, B. Buszewski, Different approaches to quantitative structure-retention relationships in the prediction of oligonucleotide retention, J. Sep. Sci. 38 (2015) 2076–2084.
[58] S. Studzinska, S. Bocian, L. Siecinska, B. Buszewski, Application of phenyl-based stationary phases for the study of retention and separation of oligonucleotides, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 1060 (2017) 36–43.
[59] J. Ghasemi, S. Saaidpour, QSRR prediction of the chromatographic retention behavior of painkiller drugs, J. Chromatogr. Sci. 47 (2009) 156–163.
[60] H. Noorizadeh, A. Farmany, M. Noorizadeh, Application of GA-PLS and GA-KPLS calculations for the prediction of the retention indices of essential oils, Quim. Nova 34 (2011) 1398–1404.
[61] H. Noorizadeh, M. Noorizadeh, A.S. Mumtaz, QSRR analysis of capacity factor of nanoparticle compounds, J. Saudi Chem. Soc. 18 (2014) 183–189.
[62] S. Khodadoust, N. Armand, S. Masoudi, M. Ghorbanzadeh, A QSRR study of liquid chromatography retention time of pesticides using linear and nonlinear chemometric models, J. Chromatogr. Sep. Tech. 3 (2012), 1000149/1000141-1000149/1000147.
[63] O. Lamarche, J.A. Platts, A. Hersey, Theoretical prediction of partition coefficients via molecular electrostatic and electronic properties, J. Chem. Inf. Comput. Sci. 44 (2004) 848–855.
[64] F. Eckert, A. Klamt, Fast solvent screening via quantum chemistry: COSMO-RS approach, AIChE J. 48 (2002) 369–385.
[65] J.J.P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, 2016. http://openmopac.net/.
[66] H.P. Schultz, E.B. Schultz, T.P. Schultz, Topological organic chemistry. 9. Graph theory and molecular topological indices of stereoisomeric organic compounds, J. Chem. Inf. Comput. Sci. 35 (1995) 864–870.
[67] A. Golbraikh, D. Bonchev, A. Tropsha, Novel chirality descriptors derived from molecular topology, J. Chem. Inf. Comput. Sci. 41 (2001) 147–158.
[68] R.D. Cramer 3rd, D.E. Patterson, J.D. Bunce, Recent advances in comparative molecular field analysis (CoMFA), Prog. Clin. Biol. Res. 291 (1989) 161–165.
[69] G. Klebe, Comparative molecular similarity indices analysis. CoMSIA, Perspect. Drug Discov Des. (1998) 87–104, 12/13/14.
[70] H. Kubinyi, F.A. Hamprecht, T. Mietzner, Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrixes, J. Med. Chem. 41 (1998) 2553–2564.
[71] H. Barfeii, Z. Garkani-Nejad, A comparative QSRR study on enantioseparation of ethanol ester enantiomers in HPLC using multivariate image analysis, quantum mechanical and structural descriptors, J. Chin. Chem. Soc. 64 (2017) 176–187.
[72] M. Szultka-Mlynska, B. Buszewski, Chromatographic behavior of selected antibiotic drugs supported by quantitative structure-retention relationships, J. Chromatogr. A 1478 (2016) 50–59.
[73] T. Baczek, R. Kaliszan, Predictive approaches to gradient retention based on analyte structural descriptors from calculation chemistry, J. Chromatogr. A 987 (2003) 29–37.
[74] J.R. Torres-Lapasio, M.J. Ruiz-Angel, M.C. Garcia-Alvarez-Coque, M.H. Abraham, Micellar versus hydro-organic reversed-phase liquid chromatography: a solvation parameter-based perspective, J. Chromatogr. A 1182 (2008) 176–196.
[75] X. Chen, H.-D. Li, F.-Q. Guo, J. Yan, D.-S. Cao, Y.-Z. Liang, QSRR study on flavor compounds of diverse structures on different columns with the help of new chemometric methods, Chromatographia 76 (2013) 241–253.
[76] K. Ciura, M. Belka, P. Kawczak, T. Baczek, J. Nowakowska, The comparative study of micellar TLC and RP-TLC as potential tools for lipophilicity assessment based on QSRR approach, J. Pharm. Biomed. Anal. 149 (2018) 70–79.
[77] D. Beiknejad, M.J. Chaichi, M.H. Fatemi, QSRR study of organic dyes by multiple linear regression method based on genetic algorithm (GA-MLR), Prog. Color Color. Coat. 9 (2016) 195–206.
[78] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, Chemom. Intell. Lab. Syst. 152 (2016) 18–33.
[79] A. Tumpa, M. Kalinic, P. Jovanovic, S. Eric, T. Rakic, B. Jancic-Stojanovic, M. Medenica, Theoretical models and QSRR in retention modeling of eight aminopyridines, J. Chromatogr. Sci. 54 (2016) 436–444.
[80] J. Golubovic, A. Protic, B. Otasevic, M. Zecevic, Quantitative structure-retention relationships applied to development of liquid chromatography gradient-elution method for the separation of sartans, Talanta 150 (2016) 190–197.
[81] J. Buckingham, Dictionary of Natural Products on DVD, CVC Press, Boca Raton, Florida, USA, 2012.
[82] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Chemometric-assisted method development in hydrophilic interaction liquid chromatography: a review, Anal. Chim. Acta 1000 (2018) 20–40.
[83] K. Muteki, J.E. Morgado, G.L. Reid, J. Wang, G. Xue, F.W. Riley, J.W. Harwood, D.T. Fortin, I.J. Miller, Quantitative structure retention relationship models in an analytical quality by design framework: simultaneously accounting for compound properties, mobile-phase conditions, and stationary-phase properties, Ind. Eng. Chem. Res. 52 (2013) 12269–12284.
[84] P. Wiczling, R. Kaliszan, How much can we learn from a single chromatographic experiment? A Bayesian perspective, Anal. Chem. 88 (2016) 997–1002.
[85] R. Kormany, J. Fekete, D. Guillarme, S. Fekete, Reliability of simulated robustness testing in fast liquid chromatography, using state-of-the-art column technology, instrumentation and modelling software, J. Pharm. Biomed. Anal. 89 (2014) 67–75.
[86] F. Vogel, Application of ChromSword software for automatic HPLC method development and robustness studies. Separation of Terbinafine and impurities, Chromatogr Today 6 (2014) 3–8.
[87] D.J. Creek, A. Jankevics, R. Breitling, D.G. Watson, M.P. Barrett, K.E.V. Burgess, Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: improved metabolite identification by retention time prediction, Anal. Chem. 83 (2011) 8703–8710.
[88] L.M. Hall, L.H. Hall, T.M. Kertesz, D.W. Hill, T.R. Sharp, E.Z. Oblak, Y.W. Dong, D.S. Wishart, M.-H. Chen, D.F. Grant, Development of Ecom$_{50}$ and retention index models for nontargeted metabolomics: identification of 1,3-Dicyclohexylurea in human serum by HPLC/mass spectrometry, J. Chem. Inf. Model. 52 (2012) 1222–1237.