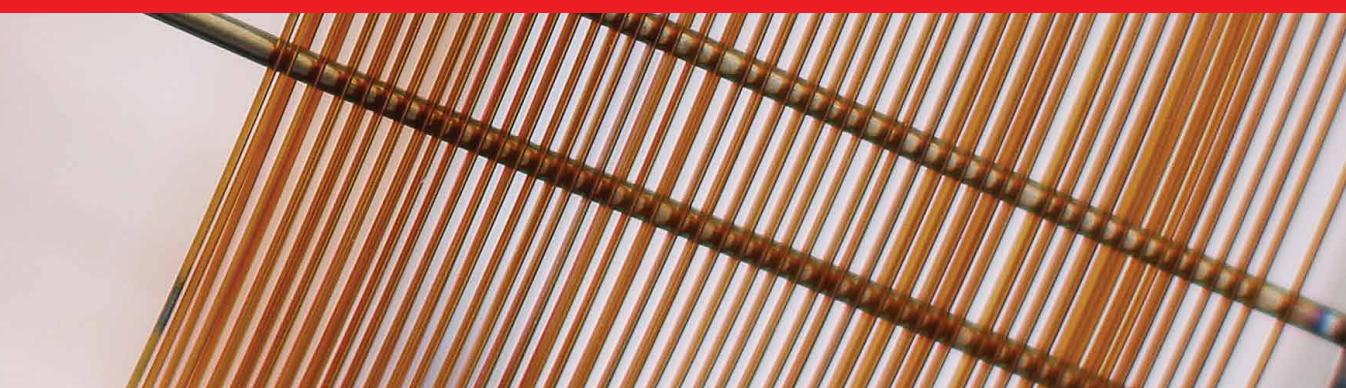
IntechOpen

Novel Aspects of Gas Chromatography and Chemometrics

Edited by Serban C. Moldoveanu, Victor David and Vu Dang Hoang



Novel Aspects of Gas Chromatography and Chemometrics

Edited by Serban C. Moldoveanu, Victor David and Vu Dang Hoang

Published in London, United Kingdom

Novel Aspects of Gas Chromatography and Chemometrics
http://dx.doi.org/10.5772/intechopen.102270
Edited by Serban C. Moldoveanu, Victor David and Vu Dang Hoang

Contributors
Nam-Ky Nguyen, Tung-Dinh Pham, Mai Phuong Vuong, Stella Stylianou, Biljana Otašević, Jovana Krmar, Nevena Đajić, Bojana Svrkota, Jevrem Stojanović, Ana Protić, Zhigang Hao, Vivian Liu, Jake Salerno, Yu Wang, Mania Bankova, Long Pan, Erwin Rosenberg, Bernhard Klampfl, Robert D. Müller, Serban C. Moldoveanu, Victor David, Robert Owen Bussey III, Vu Dang Hoang

First published in London, United Kingdom, 2023 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 5 Princes Gate Court, London, SW7 2QJ, United Kingdom

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Novel Aspects of Gas Chromatography and Chemometrics
Edited by Serban C. Moldoveanu, Victor David and Vu Dang Hoang
p. cm.
Print ISBN 978-1-80356-836-2
Online ISBN 978-1-80356-837-9
eBook (PDF) ISBN 978-1-80356-838-6

Meet the editors

Dr. Serban C. Moldoveanu graduated from the University of Bucharest, Romania, with a Ph.D. in chemistry and an MS in mathematics. He held a variety of teaching positions at the University of Bucharest before emigrating to the United States in 1983, where he taught at the University of Georgia, the University of Louisville, and Mercer University. He also has extensive industrial experience working as an analytical chemist. His research focuses on various aspects of chromatography and sample preparation. He is the author of more than 150 original papers, 12 books, a number of book chapters, and several patents. He is a member of the editorial boards of the *Journal of Analytical Methods in Chemistry* and *Frontiers in Chemistry.*

Dr. Vu Dang Hoang received his Ph.D. in pharmaceutics from the University of Strathclyde, UK, in 2005. He has been lecturing in the Faculty of Analytical Chemistry and Drug Testing at Hanoi University of Pharmacy, Vietnam, since 2007, and became an associate professor in drug quality control in 2015. His expertise is in the physicochemical characterization of topical drug delivery systems and chemometrics-based methods for the analysis of drugs in pharmaceutical dosage forms and biological fluids. He also researches the integration of the rigor of quantum chemical calculations to gain insight into molecular mechanisms. He has authored more than 50 papers and edited four books on analytical chemistry.

Professor Victor David graduated from the Faculty of Chemistry, Bucharest, Romania. Between 2007 and 2019 he was the head of the Analytical Chemistry Department at the University of Bucharest, where he is now Emeritus Professor. His main field of research is separation science (theory and applications). His list of publications includes six books for Elsevier and ten for Romanian publishing houses, 12 chapters in books and encyclopedias, and 150 papers in ISI journals. From 2017-2020 he was Associate Editor of the *Journal of Liquid Chromatography and Related Technologies.* Now he is a member of the editorial boards of *Biomedical Chromatography, Molecules, Journal of Chemistry, Journal of Essential Oil-Bearing Plants,* and *Revue Roumaine de Chimie*. He is co-editor of *Analytical Liquid Chromatography – New Perspectives* (IntechOpen, 2022).

Contents

Preface

Although gas chromatography and chemometrics are both mature fields of analytical chemistry, progress is continually being made in these important fields. New technologies, new methods, and new applications have been frequently reported in peer-reviewed literature, in manufacturer catalogs, and on the internet. A number of such novel aspects are presented in this book.

The introductory chapter describes more recent developments in gas chromatography, and also the utility of chemometrics approaches in gas chromatographic analysis. Chapter 2 describes the principle of negative thermal gradient chromatography, the advantages of this technique, and its applicability. Chapter 3 presents the main characteristics and utility of portable gas chromatography/mass spectrometry systems, and discusses some specific applications. Chapter 4 describes various sampling procedures used to make flavor and fragrance samples amenable to gas chromatographic analysis. Chapter 5 discusses various new applications of gas chromatography in the analysis of biotics and xenobiotics, such as volatile compounds of biological origin, components of biological fluids, drug metabolites, and toxicants. Chapter 6 considers the use of conference matrices as an alternative to other types of screening experiments used in chemometrics to separate key variables from those that are unimportant in large sets of influential parameters. Chapter 7 looks at quantitative structure-retention relationship (QSRR) models for liquid chromatography method development.

The goal of this book is to increase understanding of the subject by including the most recent information described in a unified form by specialists. The book is addressed to a large audience, including analytical chemists in general, either working on applications or lecturing in analytical chemistry.

Serban C. Moldoveanu
R.J. Reynolds Tobacco Co.,
Winston-Salem NC, USA

Vu Dang Hoang
Hanoi University of Pharmacy,
Hanoi, Vietnam

Victor David
University of Bucharest,
Bucharest, Romania

Chapter 7

QSRR Approach: Application to Retention Mechanism in Liquid Chromatography

Jovana Krmar, Bojana Svrkota, Nevena Đajić, Jevrem Stojanović, Ana Protić and Biljana Otašević

Abstract

One-factor-at-a-time experimentation was used for a long time as gold-standard optimization for liquid chromatographic (LC) method development. This approach has two downsides as it requires a needlessly great number of experimental runs and it is unable to identify possible factor interactions. At the end of the last century, however, this problem could be solved with the introduction of new chemometric strategies. This chapter aims at presenting quantitative structure–retention relationship (QSRR) models with structuring possibilities, from the point of feature selection through various machine learning algorithms that can be used in model building, for internal and external validation of the proposed models. The presented strategies of QSRR model can be a good starting point for analysts to use and adopt them as a good practice for their applications. QSRR models can be used in predicting the retention behavior of compounds, to point out the molecular features governing the retention, and consequently to gain insight into the retention mechanisms. In terms of these applications, special attention was drawn to modified chromatographic systems, characterized by mobile or stationary phase modifications. Although chromatographic methods are applied in a wide variety of fields, the greatest attention has been devoted to the analysis of pharmaceuticals.

Keywords: liquid chromatography, machine learning algorithms, molecular descriptors, QSRR model building and validation, analyte's retention predictions

1. Introduction

One of the most widely applied analytical techniques in a broad variety of application areas is high-performance liquid chromatography (HPLC). It stands out due to its high precision, efficacy, and robustness. Despite its undeniably good aspects, the susceptibility of analyte's retention to a diversity of experimental setup parameters makes HPLC method development a time-consuming and expensive process. Unfortunately, the selection of an appropriate combination of chromatographic conditions related to both a stationary and a mobile phase, as a starting point for the analysis of a particular drug chemical entity, is often done using a trial-and-error approach [1].

At the same time, one of the major goals of contemporary chromatographic analysis is to efficiently identify optimal working conditions for a better success rate in the method development. Luckily, a tailored, pragmatic approach denoted as quantitative structure–retention relationship (QSRR) modeling was introduced [1, 2]. With the assistance of computerized statistical methods, QSRR models are supposed to mathematically relate the molecular structural properties with the chromatographic response of a drug generated within a set of defined experimental conditions. The molecular structure encodes its physicochemical information in the form of numerical quantities denoted as molecular descriptors. This approach offers great assistance in understanding the analyte's chromatographic behavior and enables the discovery of physicochemical processes involved. As expected, statistical QSRR studies are, therefore, recognized as a supreme chemometric approach leading to the timely enhanced, high-quality separation, and efficient analytical method development [1, 2].

QSRR models are commonly associated with the retention prediction of a new and non-analyzed compound. However, QSRR models are much more useful since they are applied in revealing the molecular descriptors with the greatest retention predictive potential as well as in revealing the mechanisms that govern the separation in a specific chromatographic system on a molecular level. Based on a reliable QSRR model that accounted for different sets of chromatographic data within the same type of stationary phase (e.g. reversed-phase (RP)), the quantitative comparison of chromatographic columns can be achieved [1, 3]. The additional value of the same data refers to the direction where to look for a chromatographic column with equivalent performance and orthogonal selectivity as well as to upgrade chromatographic performances that are the most responsible for retention parameters inclusive of a short overall run time [4]. Besides all the aforementioned, many authors assert that the retention in an HPLC system, especially in RP- and micellar chromatographic systems, can be closely related to the biological activity of a drug. This can be understandable in terms of a compound's lipophilicity and pKa value because its chromatographic distribution between stationary and mobile phases is highly similar to its bodily distribution between the cell membrane and intracellular or extracellular fluids. As a result, the chromatographic data can be related to the description of biological processes of drug absorption, distribution, and excretion as well as drug-receptor interactions. Looking at the QSRR study within these wider frames, this approach can be used as a valuable in silico method for the prediction of the analyte's lipophilicity and biological activity of potentially new drug molecules. In such a manner, the utilization of less effective experimental methods and animal models can be reduced [3].

Because of their wide applicability, the QSRRs methodologies have been quite extensively studied over the past two decades. The first article, in which Tamf and Kamlet mention QSRR in a similar context known nowadays, dates from 1977. However, an intense interest in this topic has arisen over the last two decades after the work of Roman Kaliszan [5]. The first theory used to describe chromatographic retention was the theory of linear free-energy relationships (LFER), according to which the analyte's retention parameters reflect the free-energy changes associated with the chromatographic distribution [6]. In that regard, a chromatographic column is recognized as a "free-energy transducer," which translates the chemical structure differences of compounds into quantitative differences in the retention parameters. In order to provide the proper knowledge about a chromatographic system, a relatively large set of reliable input data, coming from a group of structurally heterogeneous

compounds and retention data, is needed. The early introduced QSRR models were based on a priori selection of a small set of structural descriptors derived from a molecular formula or a molecular graph reflecting physicochemical properties. These sets of structural descriptors are well known to chemists since they originate from the accepted theories of chromatographic separation and the interpretation of fundamental intermolecular interactions [7]. Since the representation of the separation process solely in terms of intermolecular interactions is questionable, an approach based on linear solvation energy relationships (LSER) was introduced by Abraham [6, 7]. He pointed out a new notably expanded set of molecular descriptors indicating the difference in interactions as a consequence of the solvent properties of the mobile and stationary phase. In parallel with these considerations, to provide reproducible quantitative input chromatographic data, two main methodological directions may be observed in the literature. The retention data can be determined under the same experimental condition or by varying chromatographic conditions, such as mobile phase compositions, flow rates, column temperatures, etc. [1]. The latter approach, also known as mixed QSRR modeling, is found to be advantageous. It enables the recognition of patterns in analyte's retention changes within observed experimental ranges of chromatographic parameters and consequently an in-depth understanding of complex chromatographic systems. In addition to proper input data, statistically significant and physically meaningful QSRR modeling finally relies on solid mathematical analysis. The usual technique for the mathematical description of correlations between all gathered data is multiple linear regression (MLR). However, the advances in liquid chromatography (LC) and an increase in the amount of chromatographic data generated over time make the conduction of a QSRR study difficult to handle traditionally. In that regard, QSRR models have shifted from a priori selection of simple descriptors and traditional regression analysis to the generation of a large pool of molecular descriptors and machine learning algorithms (MLAs) based on linear and/or nonlinear regression analysis [1]. For the sake of obtaining chemically valid interpretations, useful and reliable QSRR models demand a selection of the most informative and predictive descriptors among often mutually correlated ones. Therefore, the need for suitable selection techniques for input information data emerged accompanied by QSRR model validation strategies used to evaluate model prediction performance [2]. High-performance calculations at all the stages have made the process of LC method development more efficient and sustainable. They have also improved the fundamental knowledge of the separation processes. In accordance with numerous benefits, the anatomy of the QSRR modeling is reviewed below in conjunction with the guidance of modern requirements and tendencies.

2. QSRR workflow: a detailed walkthrough

2.1 Molecular descriptors

The power of QSRR comes from the characterization of compounds via molecular descriptors (MDs) that depict the physicochemical information of molecules in a numerical manner. The concept of MDs has come a long way in the last 50 years as it witnessed constant progress in computational chemistry. The accompanying advances in hardware enabled the calculation of over 5000 descriptors for a single molecule [8, 9]. Depending on the classification criteria, molecular descriptors can be divided into several groups. Some descriptors are obtained experimentally, while the others

are purely theoretical. According to the data type, molecular descriptors can be Boolean, integer, real, vector, etc. According to the structural dimensions (D), on the other hand, molecular descriptors can range from 1D to 6D [10, 11]. Based on the references outlining their application in QSRR studies, the extensively used descriptor-calculation software are AlvaDesc [12–14], Dragon [15–19], Molinspiration Cheminformatics [20, 21]. and Chem3D Ultra [19]. The latest and freely available software, PaDEL-Descriptor [14, 18] and Mordred [22, 23], allow MDs to be computed under open science practice representing a valuable addition to the palette of commercial software. For more simple descriptors (e.g. compositional or topological descriptors), a simplified molecular input line entry system string (SMILES) or a 2-D map can be used to represent molecules under study. If descriptors give more information, molecular geometry has to be determined prior calculation process. The accuracy of the most descriptors subsequently depends on the method used to build a 3-D molecular structure. Given a variety of the computational methods used for optimizing the geometry of analytes for QSRR studies and the availability of resources, researchers can opt to perform empirical force field methods (e.g., molecular mechanics), semi-empirical optimization (e.g., AM1, PM3), or sophisticated *ab initio* calculations (e.g., Hartree-Fock and Density Functional Theory) [24, 25]. In an interesting study, Amos et al. investigated how different levels of theory for structure optimization contributed to the QSRR outcome. The sum of ranking differences (SRD) showed that a fast and rational method of structure optimization shared the results with time-consuming and expensive calculations in terms of the final accuracy of the QSRR model. Moreover, the solvent correction did not reduce the mean absolute error of QSRR predictions. The authors carefully explained these unexpected findings in the context of an error inherent in the Dragon descriptor calculation process [26].

2.2 Feature selection

A small set of predictors (i.e., input variables or factors) with well-known physicochemical meaning can be pre-defined when modeling separation in systems with fully elucidated retention mechanisms. For complex chromatographic modes, such as micellar liquid chromatography (MLC) and mixed-mode liquid chromatography (MMLC), a priori attribute selection may compromise the accuracy of QSRR predictions making it a poorly acceptable strategy for retention modeling [27, 28]. Alternatively, a large set of independent variables can be formed; the most significant attributes can be extracted from it and used to build a model for retention time prediction. Clearly, the predictive ability of these models depends on the efficiency of the mathematical algorithm used to select predictor variables [7, 29].

The choice of the most informative features for a particular regression problem poses one of the main challenges in machine learning (ML). Determining the appropriate method of variables selection, in this regard, has been an interesting topic in a broad range of domain applications, including studies for which the datasets with hundreds or thousands of attributes become available along with the development of molecular modeling software. Faced with plenty of noisy and irrelevant features, contemporary QSRR studies call for variable selection without exception. The purpose of variable selection methods is to handle space dimensionality by discarding the features that are redundant and irrelevant in predicting endpoint values. A feature is irrelevant if it is unpredictive for the dependent variable or response. A reduction is needed if it is highly correlated with other features. The adoption of feature selection techniques ultimately

avoids overfitting, improves a model's predictive power, and enhances an understanding of the underlying patterns preserved in data. A decreased computational burden placed on modeling techniques as well as easier data visualization happens to be additional benefits associated with feature selection techniques [29–31].

In a typical MLA-empowered QSRR pipeline, a minimal feature subset is determined after the pre-processing of raw data and before the modeling. Among various techniques, MLR, genetic algorithm (GA), and Relief method have been quite eagerly used in QSRR studies [31, 32]. Other important feature selection methods are least absolute shrinkage and selection operator (Lasso), artificial neural network (ANN), and random forest (RF) [24, 27, 33]. The last two algorithms will be discussed later in the text, as part of Section 2.4.

2.2.1 Multiple linear regression

MLR founds a linear relationship between a dependent variable and two or more independent variables (regular attributes). It is basically the extension of the Ordinary Least Squares (OLS) method.

The general MLR model can be written using Eq. (1):

$$y_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \ \ldots \ + \beta_n x_{jn} + \beta_0 + \varepsilon \tag{1}$$

where y_j is a dependent variable, x_j are independent variables, β_n are slope coefficients for each predictor, β_0 is an intercept, and ε refers to a model's error term.

In the OLS method, the slope coefficients that minimize the loss function come from Eq. (2):

$$\sum_{j=1}^{k}\left(y_j - \hat{y_j}\right) = \sum_{j=1}^{k}\left(y_j - \left(\beta x_j + \beta_0\right)\right)^2 \tag{2}$$

The use of MLR makes sense only if: a) there is a linear relationship between predictors and dependent variable, b) the correlation between variables is not too high, c) the instances are chosen randomly from the population, and d) the residuals are normally distributed. MLR estimator is burdened with a great variance, especially in the cases, where the number of attributes approaches the number of observations [27].

2.2.2 Least absolute shrinkage and selection operator

Lasso regression is one of the most popular regularization methods for selecting significant independent variables. The concept of regularization has been introduced to avoid overfitting in MLR modeling. In brief, the regularization refers to adding a "penalty" term to the best model built upon a training dataset and to achieve a smaller variance and control the influence of the predictor variables over the response. In Lasso regression, this is done by penalizing the absolute value of the magnitude of coefficients (Eq. (3)).

$$LASSO = \sum_{J=1}^{K}\left(y_j - \sum\nolimits_n x_{jn}\beta_n\right)^2 + \lambda\sum_{k=1}^{p}|\beta_k| \tag{3}$$

In Eq. (3), λ is the tuning parameter that controls the amount of shrinkage. If λ is large, the slope coefficients are penalized highly toward 0, and more features are eliminated. If λ is 0, all features are considered and the residual sum of squares criterion is applied. As λ increases, the bias increases. Otherwise, the variance increases [27, 34].

2.2.3 Genetic algorithm

GAs are methods that generate a solution for optimization and search problems by simulating the mechanism of natural selection and the survival of the fittest. In the initial stage, the GA creates a random population of chromosomes. Each chromosome, usually represented by a binary string, encodes a potential solution to the problem under study. In the case of feature selection, individual chromosomes make up a random subset of variables, where the presence or absence of a variable in the chromosome is denoted by 1 or 0, respectively. Using individuals in the current generation, the GA creates a sequence of new populations. To achieve this goal, the algorithm first evaluates each chromosome of the current population by determining its fitness value. The fittest individuals are selected to pass their genes to the next generation. Offsprings are, in fact, produced by subjecting the selected parents to crossover (gene exchange) and mutation (gene change in individuals). In addition, some of the population's members with the best fitness values are chosen as elite children and added directly to the next population. The subsequent generation is formed after children with inherited good characters replace the current population of parents. The GA loops until one of the stopping criteria is met (e.g. a predefined number of generations). The flowchart (**Figure** 1) outlines the main GA steps.

In terms of the prediction accuracy of constructed QSRR models, the GA showed superiority in selecting the most relevant features compared to other variable selection methods [18, 35]. Lately, the GA has been used for the non-polynomial hard problem of feature selection [36], the selection of molecular descriptors for localized QSRR models [37], and the development of a QSRR model intended to improve the structural annotation of triterpene metabolites in an LC-HRMS system [38].

2.2.4 ReliefF

In this method, each attribute is assigned a relevance weighting according to its ability to distinguish between class labels. Attributes with weight above the user-defined threshold τ are considered significant and included in the set of selected features. The underlying principle is that the instances belonging to the same class should be closer than those of different classes. The algorithm cycles over j training cases (R_i) that are chosen by the user. First, n dimensional weight vector W of zeros is initialized. Then, the target instance R_i is selected at random and the distances between it and its two nearest neighbors, namely, *nearestHit* (the closest instance with the same class) and *nearestMiss* (the closest instance with the opposite class) are calculated. Feature weight W is updated so that more weight is assigned to attributes that distinguish an instance from neighbors of different classes (Eq. (4)). After j cycles, each element of the weight vector is divided by j, giving rise to the relevance vector [27, 30, 31].

$$W_i = W_i - (R_i - nearestHit)^2 + (R_i - nearestMiss)^2 \quad (4)$$

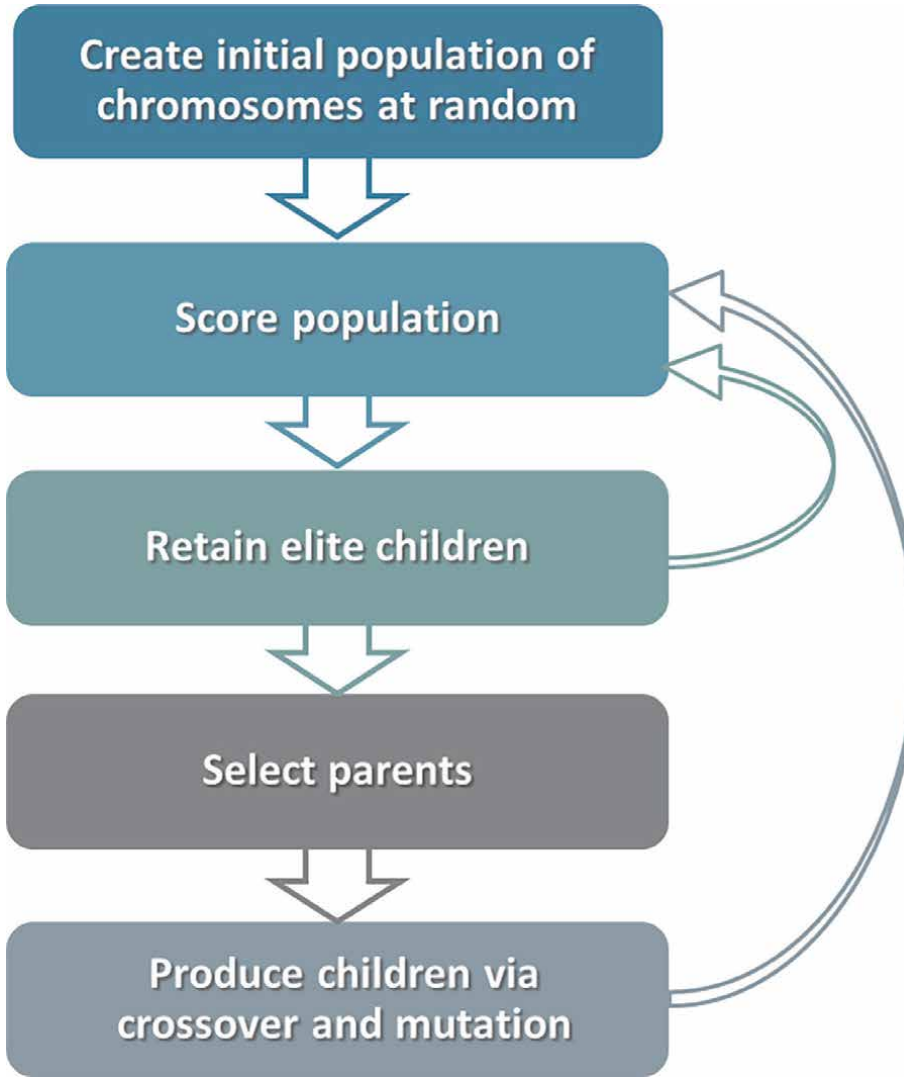


Figure 1.
GA flowchart.

Originally, the Relief algorithm has been intended for classification problems and could be fooled by an insufficient number of cycles. Nowadays, it has been adapted for predicting continuous decision variables and as such is being used for QSRR studies (e.g. to predict retention parameters). The differences between Relief, ReliefF, and RReliefF are presented in detail in [39].

2.3 Response transformation

When implementing supervised algorithms, it is a good practice to examine data distribution. The distortion of the symmetry of normal distribution around its mean is denoted as skewness. A general impression of skewness can be gained by drawing a histogram or computing the skewness coefficient. If the distribution's shape has one peak and a long tail on the right side of the curve, the distribution is positively skewed.

In contrast, the distribution has a negative skew if a long tail is on the left side of the curve. In numerical terms, the skewness for a normal distribution is (approximately) 0. Negative coefficients are related to negative skewness and vice versa. The coefficient values between – 0.5 and + 0.5 indicate moderately skewed data, and if they are less than −1 or greater than + 1, the distribution is highly skewed. A highly skewed dataset can contaminate a model's predictive performance because the algorithm has to deal with scattered endpoints at extreme values. In the case of right-skewed data, for instance, MLAs are likely to predict points with lower values better than those with high values. Therefore, skewed distribution is one of the major obstacles to the application of MLAs to real-world data and should be addressed prior to the modeling. A common strategy for dealing with skewed variables is to transform them. Logarithmic, square root, and cube root transformations are recommended when data follow the power-law distribution, while in the opposite case, it is better to opt for square, higher powers, or cube root transformations [27, 40, 41].

2.4 Model building techniques

The choice of regression technique for correlating molecular descriptors and chromatographic conditions with a chromatographic parameter has a huge impact on the performance of any QSRR model. Due to its simple and explainable character, MLR received considerable attention in mechanistic research long ago [24]. However, if researchers amass vast troves of data and cannot make sense of it in a reasonable amount of time, the process is the main candidate for modeling through more sophisticated MLAs. MLAs fall under the umbrella of artificial intelligence and can process and understand data faster. These algorithms learn to resolve issues by drawing firm conclusions from observation data they are supplied with. Along with improvements in technology and computing power, QSRR can take advantage of machine learning in a fundamental and practical manner. By acknowledging nonlinearity in LC data, MLAs play an important role in the accuracy of property predictions. However, no currently available MLA can deliver optimal performance for every modeling task. A variety of MLAs should be used before selecting a particular regression technique. The common MLAs are ANNs, support vector regression (SVR), and ensemble methods [42].

2.4.1 Artificial neural networks

ANN is a series of machine learning algorithms that mimic the process of natural thinking by making experience-based decisions. Modeled on the human brain, the ANN refers to a massive composition consisting of some primitive processing elements (i.e. artificial neurons). Most operative neural nets are constructed by grouping neurons into layers. An individual neuron might be connected to several nodes in the layer beneath, i.e., above it. Data passing through layers in only one direction makes up a feedforward neural net (or multi-layer perceptron). Apart from the layers, the main components of ANN include the adaptive coefficients –weights, assigned to each of the connections between the layers, as well as the transfer functions, which convert received raw data into output. The transfer functions, learning rules, and architecture itself define the behavior of each ANN [43]. When a neural net is being trained, all of the weights are first randomly assigned to synapses between neurons. Then the input-output pairs of data are fed to the net in an attempt to train an algorithm to recognize the underlying patterns between variables. This strategy pertains to the process of

supervised learning. In a supervised feedforward backpropagation algorithm, the training is performed by comparing the processed signals with the desired outputs and adjusting the inputs' weights until the margin of error is minimal. Herein, the weights are updated in the steepest descent fashion. Higher weights are attributed to the inputs that contribute the most to achieving the right target [44, 45].

Neural nets are a valuable tool for analytical R and D due to the ability to learn nonlinear relationships encountered between predictors and dependent variables in most corresponding systems. Contemporary applications of ANN in the pharmaceutical sciences are broad, ranging from interpretation of analytical data to drug design. Over the past decade, there has been an impressive increase in the number of publications on QSRR studies that used ANN as a modeling technique. In particular, the single-hidden layer neural nets provided a satisfactory level of prediction accuracy [46–51]. After the improvement in computer power and the rise of big data, ANNs began to flourish in the form of deep learning (DL) algorithms [52–55]. Deep neural networks are the ones that have more than one hidden layer. With each additional layer, the DL algorithm can model increasingly complex relationships. As compared to other ML techniques, ANN architecture is characterized by great flexibility and can process raw data and automatically extract a set of the most informative features. Unfortunately, the DL is not free of limitations; in general, these algorithms are data-hungry and require massive training sets. The question to be raised, in that respect, is whether the analytical domain can provide big data without losing valuable resources [56, 57]?

2.4.2 Support vector regression

SVR is another promising machine learning algorithm that acknowledges the nonlinearity in data. It is built on the principles of statistical learning and the concept of constructing a line (or hyperplane in high-dimensional space) that fit the data. Among an infinite number of possible solutions, SVR finds a hyperplane with the greatest distance to the nearest training instances. Finding such a hyperplane is based on minimizing the l2-norm of the coefficient vector, w (Eq. (5)), while the absolute error between the target y_i and predicted values are set to be less than or equal to a specified margin, ε (Eq. (6)).

$$\min \frac{1}{2}\|w\|^2 \tag{5}$$

$$\left|y_i - w_i x_i\right| \leq \varepsilon \tag{6}$$

In Eq. (6), x_i is the i-th input point in the input space (a feature) and w_i is its associated coefficient. The maximum error ε is tuned to gain the predictive ability of the built model satisfactorily.

For the endpoints that reside outside the ε-tube, deviation from the margin is represented by the slack variable, ξ. Term C is added to penalize these points in comparison with those either above or below the hyperplane. With respect to these deviations, the objective function and its constraint are given in Eqs. (7) and (8), respectively.

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{7}$$

$$\left|y_i - w_i x_i\right| \leq \varepsilon + |\xi_i| \tag{8}$$

The SVR hyperplane is constructed after the inputs are mapped into a space of higher dimension(s) than the original using the kernel function (e.g., polynomial, splines, radial basis function, etc.). Then, using a simple linear function, the SVR helps predict the target value. By projecting the optimal hyperplane back into the input space, it takes on a nonlinear form. Due to its remarkable generalization ability, the SVR has gained popularity in QSRR studies [31, 44, 55, 58–63]. In most publications, the empirical performance of SVR matches with or is considerably better than the performance of other MLAs studied.

2.4.3 Ensemble learning algorithms

In ensemble learning, algorithms with high bias or too much variance (so-called weak learners) are merged to produce the most popular result. The underlying idea of aggregating predictions is to create a much more accurate and robust model. Bagging (also known as bootstrap aggregation) and boosting are the most prominent classes of ensemble methods [64].

Weak learners that are used widely in ensemble learning are decision trees (DTs). DTs are nonlinear machine learning techniques that can handle either regression or classification tasks. They are simple, intuitive, and can deal with missing values and large datasets with elegance. The classification and regression tree (CART), introduced in 1984, is a typical DT algorithm [65]. It is presented as a tree-shaped diagram containing a set of nodes and branches growing downwards. This topology gives the idea of a binary and hierarchical algorithm that adopts the recursive partitioning method. It is an iterative procedure that seeks to find the best split (the best splitting feature and the best input data) at each step. Performance metrics, e.g., Gini index, information gain, or error rate, are utilized to assess the quality of the split. Fundamentally greedy nature and poor ability to cope with the penalties on tree complexity (while growing the tree) are the main disadvantages of the top-down approach. Pruning is done to prevent an overfitting phenomenon [66].

2.4.4 Random forest

RF was introduced as a DT-based ensemble in 1984 [65]. It is a collection of unpruned DTs (grown to the maximum extent) that are trained by the bagging method. In bagging, base models are grown on bootstrapped subsets of the data and the individual predictions of all base models are averaged to get the final output. As a result, the ensemble model has less variance than its building elements. While sharing the main idea with bagging, the RF adopts an additional level of randomness – each node of each tree deliberately takes into account only a random subset of features (e.g., the square root of a number of descriptors) for the splitting procedure. An RF model benefits from this tactic in terms of efficacy. In addition, it is important to mention that the internal validation is built into the forest growth. According to the concept of bootstrapping, some of the data are omitted from the samples intended for tree growth, while the others are repeated in the samples. The former is denoted as Out-Of-Bag (OOB) data. Given the fact that the OOB sample is not included in the tree fitting, it is used to estimate the model error. Usually, it makes up to one-third of the available data, while the other two-thirds of the data is used for training. In order to achieve a small OOB error, it is necessary to optimize the number of base models and the size of a subset of features [66, 67].

In QSRR studies, the RF algorithm is readily used as a modeling technique [42, 55, 66, 68] as well as a feature selection method [69, 70]. The latter is due to the ability of the algorithm to quantify the importance of variables under study. The importance of each feature is determined by observing a change in prediction error when the OOB set for that feature is permuted (and the other features are kept constant).

2.4.5 Gradient boosted trees

Gradient boosted trees (GBT) is an extremely powerful ensemble algorithm based on boosting and gradient descent approach. Unlike the bagging, which combines weak learners in parallel, the boosting merges base models linearly. The focus here is especially on shallow DTs that have low variance and high bias.

A correlation between base models (arising from the same data) is precluded by an incremental change of the training set. This is done by assigning weights to each example. Initially, all weights are set to be equal and the first decision tree is trained on the original dataset. Accurately predicted instances have their weights decreased, while the others have their weights increased. The trees that enter the ensemble in subsequent iterations are thus applied to the reweighted data and their goal is to correct the errors made by the previous model. Boosting, which decreases the bias of individual base models, is viewed as one of the groundbreaking concepts introduced in ML over the last decades. The GBT algorithm minimizes a loss function via a gradient descent procedure. The predictive power of the GBT ensemble correlates with the number of base models and the size of learning rate. A larger ensemble will very quickly over-fit, while a combination of too few DTs might lead to poor predictive performance. Lower values of learning rate (a parameter that controls the length of incremental step) may resolve the problem of overfitting, but a prolonged convergence toward the solution can place a lot of computational burden on the model in question [71, 72]. Due to the ability to create highly accurate QSRR models (and the fact that it quite often outperforms many other regression algorithms), the GBT is popular in analytical R&D. The successor to the gradient boosting, regularized gradient boosting (i.e., XGBoost), is increasingly used to provide state-of-the-art solutions to many LC challenges as it yields improved generalization capabilities and better avoids over-fitting [27, 32, 42, 73, 74].

2.5 QSRR model validation

It seems that it is feasible to build mathematical models that fit the data very well. But, there is still a possibility that it may happen due to chance correlations or overfitting. In that case, the models are not considered appropriate for their intended application. Therefore, proper statistical validation of models is of great importance in QSRR studies. The two main concepts, denoted as internal and external validation, will be discussed herein. The internal validation procedures include leave-one-out (LOO) and leave-many-out (LMO) cross-validation (CV), y-randomization, and bootstrapping.

2.5.1 Leave-one-out cross-validation

LOO-CV is performed by excluding each sample (compound) once and building a model with the remaining data and predicting the value of the response for the eliminated sample. Due to the presence of repetitive cutting data set activities, the

LOO is also known as rotation estimation and jack-knife validation method. This approach indicates that the eliminated sample serves as a temporary test set taken from the overall training set. Each cycle of this repetitive procedure is followed by calculating the differences between experimentally observed response values and estimated (predicted) ones by the model. These values are afterward included in Eqs. (9) and (10) corresponding to the root mean square error of CV (*RMSECV*) and the cross-validated correlation coefficient (Q^2), respectively. Finally, the model predictive performances are inspected by the values of the root mean square error of calibration (*RMSE*, Eq. (11)) and the overall CV correlation coefficient value, calculated for the whole original dataset as the average value of Q^2 from each CV cycle [27]. The value of overall Q^2 is usually greater than that of individual Q^2, though a large difference between them (overall Q^2 greater by 25%) indicates that the model suffers from overfitting [75].

$$RMSECV = \sqrt{\frac{\sum \left(y_{experimental} - y_{predicted}\right)^2}{n-1}} \tag{9}$$

$$Q^2 = 1 - \frac{\sum \left(y_{experimental} - y_{predicted}\right)^2}{\sum \left(y_{experimental} - \left\langle y_{experimental} \right\rangle\right)^2} \tag{10}$$

$$RMSE = \sqrt{\frac{\sum \left(y_{experimental} - y_{predicted}\right)^2}{n}} \tag{11}$$

In the aforementioned Eqs. (9)–(11), n stands for the number of samples in the dataset, $y_{experimental}$ are the values of experimentally observed responses and $y_{predicted}$ are the responses calculated (theoretically predicted) based on the built model calculated either from the data used from model development (in case of Q^2 and *RMSECV*) or the original dataset (in case of *RMSE*). The brackets $\langle\ \rangle$ are used to point out the use of the average values of experimentally obtained responses.

2.5.2 Leave-many-out cross-validation

To perform the LMO-CV, the initial dataset is divided into blocks of samples; afterward, each block is eliminated once from the model building in each cycle in a similar manner as applied in the LOO-CV. The prediction of response is made for the block under consideration. It should be noted that the blocks may consist of the same number of constituents, but that is not an obligation. The LMO may also have different validation cycles. In that respect, as an example, the original dataset can be divided into 10 parts indicating that each data block accounts for 10% of the data, 10 validation cycles are needed and the respective method may be referenced as 10-fold CV as well. In comparison with the LOO, this procedure is more time-effective. The validation metrics are similar to those presented for the LOO-CV with adjustments in relation to a sample or a block of samples. It is worth mentioning that the same appropriate adjustments must be implemented in the used equations. Also, since there is no truly new compound under consideration within none of the variations of the internal validation procedures, it is advisable to perform as many as possible internal validation tests for the final justification that the model is of good quality, relevant,

and suitable for its intended use. This recommendation especially stands in the case of the modeling based on small datasets where any omission of data from the original dataset may lead to the inability to perform the modeling procedure at all [75, 76].

2.5.3 y-randomization

Y-randomization is used to ensure the robustness of the developed model. For example, it can check whether there is a molecular descriptor statistically well correlated with the response value y; but, in reality, there is no cause-effect relationship originating from the physical and/or chemical meaning of a molecular descriptor and the respective retention measurement. The model validation is performed by keeping the so-called X matrix with original unchanged descriptors while the vector of the response values y is randomized or scrambled. Since the new models are built based on the same input dataset but associated with changed (false) responses, it is expected that they are of poor quality as reflected by the values of Q^2 and overall Q^2. Kiralj and Ferreira proposed a detailed overview of the possible Q^2 and overall Q^2 values and their interrelationships according to which the chance correlation may be inspected [77].

2.5.4 Bootstrapping

Bootstrapping procedure suggests the random splitting of a complete dataset into training and test sets several times and the building of respective models afterward. While in the LOO and LMO procedures each sample is excluded from the modeling only once, in the bootstrapping there is an equal chance for a sample to be eliminated once, several times, or even never. The corresponding Q^2 and overall Q^2 validation metrics are calculated and expected to be of high values as well as to oscillate around the real values or values obtained from the LOO-CV of a real model. It should be noted that this validation procedure is affected by a number of splits or resampling as well as the structure or similarity between the training and test sets [77].

2.5.5 External validation

The predictive power of a QSRR model is evaluated by the external validation, with *model blind* samples (compounds), meaning that these samples were not previously *seen* by the model or used for model development. Therefore, the extraction of an external validation test set from the original data is required, and a proper selection of the size and type of these data is of crucial importance for a successful validation process. Usually, this subset covers 15–25% of the original dataset [78, 79]. Although the external validation test set is a golden standard of the QSRR models' prediction properties, there is a concern about the relatively small size of the external test set in comparison with the LOO- or LMO-CV where the whole dataset acts as a test set in some moment. At the same time, the consideration of the similarity between the training and external validation subsets is of utmost importance by means of similar variable ranges and distribution. The common trend indicates that a greater similarity between these subsets leads to a decrease in prediction errors [4]. Finally, more than one splitting of the dataset into modeling and external validation test sets is also advisable [80].

The statistical parameters for model evaluation include the multiple coefficients of determination of external validation, also called *predictive* R^2, and the root mean

square error of prediction (*RMSEP*). The values of these parameters can be calculated using Eqs. (10) and (11), taking into consideration that all data correspond to the external test set solely. Another valuable indicator of the model's predictive performance is the Pearson correlation coefficient of prediction (R), which is used to reflect a correlation existing between the experimentally observed responses and the responses predicted by the model. It is expected from the value of R to be maximally close to 1. The parameter R can be calculated by Eq. (12) [77].

$$R = \frac{\sum\left(\left(y_{experimental} - \left\langle y_{experimental}\right\rangle\right)\left(y_{predicted} - \left\langle y_{predicted}\right\rangle\right)\right)}{\sqrt{\sum\left(y_{experimental} - \left\langle y_{experimental}\right\rangle\right)^2}\sqrt{\sum\left(y_{predicted} - \left\langle y_{experimental}\right\rangle\right)^2}} \tag{12}$$

2.5.6 Detection of sources of prediction errors

Apart from the inspection of the statistical parameters computed from the respective validation procedure, to assure the quality and practical usefulness of QSRR models, it is worth getting insight into the possible sources of prediction errors (residuals). In that respect, besides the calculation of *RMSE,* which uses the same units as the response, it is useful to express it in percentages. By analyzing the value of *RMSE (%)*, the magnitude of the prediction error concerning the mean of actual experimentally observed values is clearer for understanding. Another benefit of this is the possibility to detect outliers i.e., the samples for which the predicted values are too distant from the mean of the experimentally observed values. The outliers differ significantly from all other observations due to the exceptional chemical nature or chromatographic behavior of a compound and may occur in the test dataset as well as in the training dataset. Since their values lie outside the overall usually normal distribution of a dataset, it is quite obvious that the outliers can cause serious problems when it comes to the development of reliable and statistically stable QSRR models. Based on the number of outliers and the intensity of their distinction from other data points in a dataset (soft or influential outliers), the model predictive ability and/or model statistical stability may be brought into question [44]. It is recommended that the outliers should be removed from a dataset before proceeding with model development and analyzed for the origin of possible errors [77, 78, 81]. For the sake of building models of suitable quality, various methods for outlier detection immerged among which some are based on visual analysis of scatter plots, histograms, Box plots, and the others on the calculation of Z-score and interquartile range. More sophisticated methods propose so-called acceptable error windows and unambiguous cut-off limits for applicability domain margins while considering the chemical structural diversity of compounds in a dataset, standardized residuals of predictions and a specific leverage (structural) value of each compound (OTrAMS method), a standard deviation of predictive residuals and a mean of predictive residuals (Monte-Carlo sampling method). More detailed information on the use of later outlier detection methods was provided by Aalizadeh et al. [59].

2.5.7 Definition of model applicability domain

In addition to the statistical model assessments, the predictive power of a robust and validated QSRR model must be expressed in terms of the applicability domain. The model interpretability is affected by its characteristics as well. This domain refers

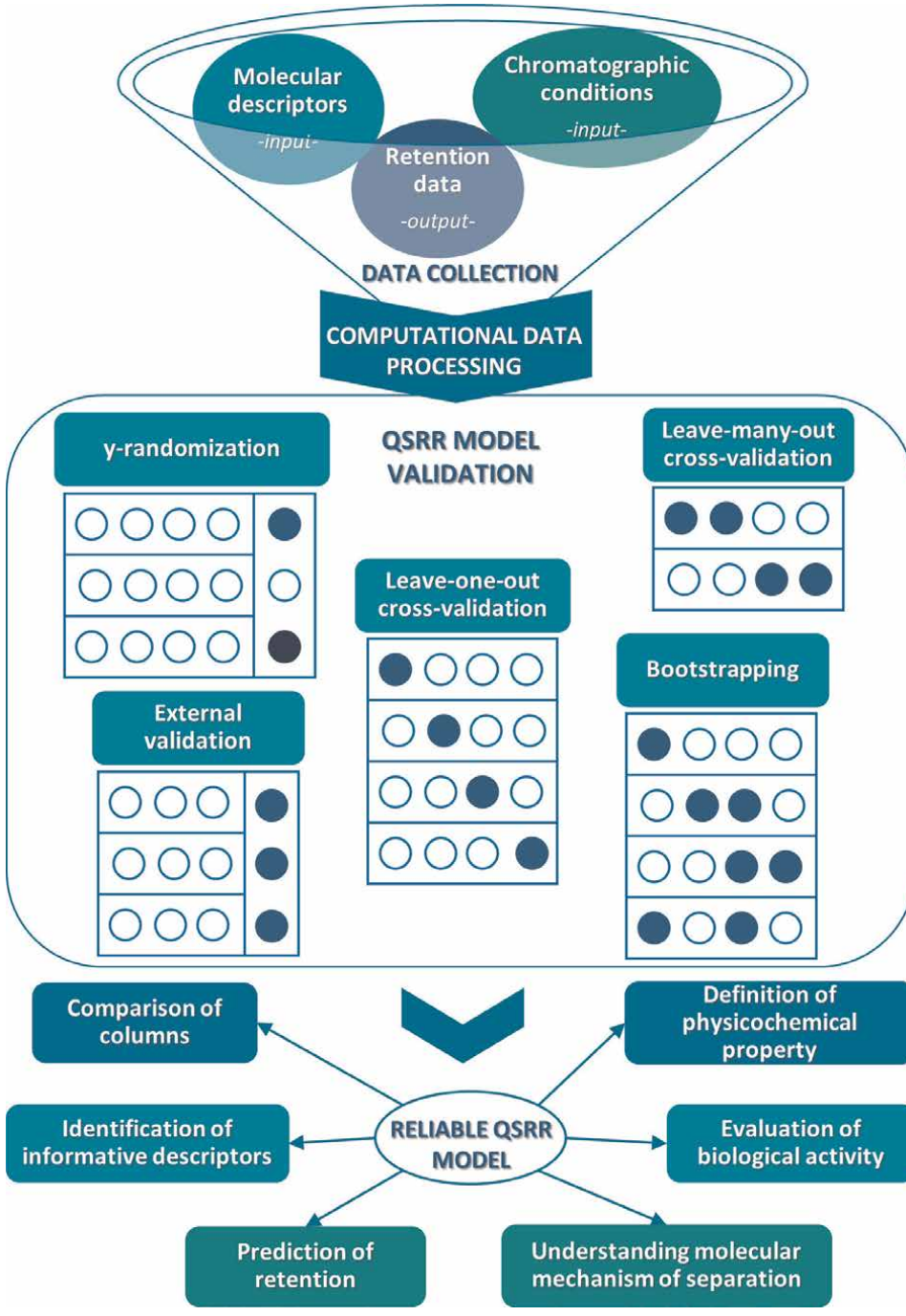


Figure 2.
QSRR workflow.

to a theoretical space defined by a range of the molecular descriptors of compounds used for model training purposes and respective chromatographic conditions as well as a range of the modeled responses. It is obvious that the applicability domain strongly reflects the physicochemical and structural properties and chromatographic

behavior of compounds from a training set. In order to make the best response predictions, the training set must be similar to the target molecule [24, 44, 75]. The aim of narrowing the space for making predictions actually serves to avoid unjustified and inaccurate model extrapolations. The dedicated approaches for the definition of applicability domain based on the range of response variables or the range in the descriptor space (geometrical methods and distance-based and probability density distribution-based methods) were thoroughly described by Roy et al. [82, 83]. As the issue is closely related to the applicability domain, the same authors elaborate the strategy for a proper selection of data to be introduced in the training and/or test dataset out of the original dataset as well [83]. It is perfectly reasonable to state that the lack or poorly conducted selection of compounds increases modeling errors and calls into question the success in all predefined QSRR modeling goals or application areas.

After summing all the previous considerations into a graphical presentation, the QSRR flowchart may look like the one in **Figure** 2.

3. Application of statistical QSRR model in complex HPLC techniques

The application of the QSRR approach is directly driven by its definition. As the QSRR represents a mathematical relationship between molecular retention behavior and its properties inherent in molecular structures (molecular descriptors), they are primarily used to predict the retention behavior of molecules omitted during model development. In addition, it can be used to single out important features, by which the retention behavior is governed and it is possible to gain insight into the retention mechanisms. It can also be applied for stationary phase characterization or their comparison in terms of separation characteristics [5]. In some cases, they can provide drug or xenobiotics classification or an assessment of their bioactivity [2]. By incorporating experimental parameter values into a QSRR model, their application can be expanded on HPLC method development and optimization [84].

Since various highly adaptable mathematical tools are suitable for structuring statistical QSRR models, the QSRR approach shows compatibility with a broad spectrum of HPLC properties. Although it has a place in the modeling of conventional unimodal HPLC, which was discussed in more detail by Haddad et al. [84], it is also a valuable tool in the case of defining more complex HPLC systems. Complicated molecular retention patterns are often generated from mobile or stationary phase modification. Taking into consideration such HPLC system modification, the predictive abilities of QSRR can not only reduce experimental requirements but also provide a deeper insight into the retention mechanism. The following section is not a comprehensive literature review but rather a demonstration of the beneficial properties of QSRR used for characterizing complex HPLC systems applied for the analysis of small molecule substances.

3.1 QSRR approach for HPLC with mobile phase modifications

Increasing the retention of poorly retained analytes in RP-HPLC is often achieved by modifying mobile phase properties. The addition of modifiers can provoke changes in the retention behavior by imposing an additional equilibration process.

3.1.1 Ion-interaction chromatography

Compromised retention of basic solutes can be promoted by introducing ion-interaction agents into the mobile phase. Ion interaction chromatography (IIC) involves a series of equilibration processes between chromatographic phases and analytes, which necessitates the understanding of the separation process [85, 86]. An IIC system, with added chaotropic salts, was assessed by Čolović et al. [63]. A mixed QSRR-SVR model was developed based on the retention data of 34 analytes as independent variables were selected i.e., three mobile phase parameters (concentration of NaPF6, pH, and acetonitrile content) and four molecular descriptors (Branching index EtaB with ring correction relative to molecular size (ETA_EtaP_B_RC), calculated octanol/water partition coefficient (XlogP), 3D topological distance-based autocorrelation – lag 9/weighted by polarizabilities (TDB9p) and radial distribution function – 045/weighted by relative polarizabilities (RDF45p) descriptor). The importance of analytes' steric effects and voluminosity were indicated by ETA_EtaP_B_RC, while XlogP implied the significance of hydrophobicity, which was in line with the RP retention mechanism. However, TDB9p and RDF45p indicated the participation of electrostatic interactions during the retention process. Thus, the hypothesis on the complementarity of the analytes' electronic structure and the electrical bilayer created in the stationary phase was supported.

3.1.2 Micellar liquid chromatography

In MLC, a modification of mobile phase features is attained by adding surfactants. When surfactants are present at a concentration above the critical micellar concentration, micelle formation occurs. Surfactant molecules can coat the stationary phase as the absorbed monolayer. Moreover, surfactant interaction with both analyte and stationary phase implies the presence of secondary equilibration. Thus, the exploration of the MLC retention process is challenging [85, 87]. A QSRR-MLR modeling approach was performed by Ramezani et al. for testing anthraquinones. These authors linked molecular descriptors (partition coefficient calculated from hydrophobic fragmental constants (logP), Geary autocorrelation of lag 8 weighted by van der Waals volume descriptor (GATS8v), the mean topological charge index which represented the effect of analyte charge in the MLC separation (JGI4), and descriptors based on *3D molecule representation of structures based on electron diffraction theory* (3D-MoRSE), namely 3D-MoRSE descriptor of signal 27 (Mor27m) and 3D-MoRSE descriptor of Moran autocorrelation of lag 7 (MATS7md)) and empirical factors of six organic modifiers to anthraquinones' retention time. It was concluded that the retention behavior is significantly influenced by the modifier's logP values, as well as by the mass, molecular weight, and van der Waals volume, in addition to the topological charge [63].

Complementation of the available knowledge on MLC was attained by Krmar et al.; numerous mixed QSRR models were developed using different types of algorithms. Not only was the GBT identified as the most suitable but also the most significant properties relevant for the separation of aripiprazole and impurities were extracted. QSRR models, in addition to MDs, contained experimental parameter values (concentration of non-ionic surfactant Brij L23, pH, and the content of ACN) in line with the Box-Behnken design. Steric effects and dipole-dipole interactions were identified to be the most important thermodynamic molecular parameters relevant for retention behavior [27].

3.1.3 Cyclodextrin-modified liquid chromatography

Shifting the analytes' retention behavior in RP-HPLC can also be provoked by adding cyclodextrins (CD) to the mobile phase. Molecular retention patterns are modified because of CD-analyte complex formation, in addition to the adsorption process of CD on the stationary phase surface [85].

Maljurić et al. developed a QSRR-ANN model for the retention property analysis of risperidone, olanzapine, and related impurities in a CD-modified RP-HPLC system. The values of MSs, complex association constants, and chromatographic factors were used in the model. The most influential molecular descriptors and complex association constants were polarizability (POL), solvent-excluded volume (SEV), octanol/water partition coefficient (logP), dipole-dipole energy (DEN), binding energy (BE), electrostatic energy (EE), and unbound energy (UE) [48]. In a later study, a developed model was employed for determining a change in retention factor, the stability constants, and thermodynamic parameters of complex formation [88].

Another QSRR-ANN model for revealing separation processes in a CD-modified RP-HPLC system was developed by Đajić et al. The experimental parameters were acetonitrile percentage, aqueous phase pH, β-CD concentration, and column temperature. The most important molecular descriptors were identified as radial distribution function – 075/weighted by mass (RDF075m), signal 04/weighted by mass (Mor04v), and CATS2D positive-lipophilic at lag 08 (CATS2D_08_PL). It was found that the molecular size, shape, and lipophilicity of analytes significantly affect their retention. The retention behavior is also governed by the size and lipophilicity of the added CDs as it determines the structural agreement with the tested analytes [89].

3.2 QSRR for HPLC with unconventional stationary phases

Non-straightforward retention behavior resulting from the application of an unconventional stationary phase can be defined similarly as in the previous examples. As the QSRR successfully reveals additional interactions shaped by mobile phase modifiers, it can also expose multiple retention mechanisms provided by the stationary phase.

3.2.1 Immobilized artificial membrane chromatography

The characteristics of the stationary phase used in immobilized artificial membrane (IAM) chromatography are in line with its structure based on phosphatidylcholine residues covalently bound to silicon dioxide. In this way, the column mimics a phospholipid membrane monolayer and exhibits biomimetic properties [90].

In the research of Ciura et al., the general conclusions about the molecular retention mechanisms of isoxazolone on an IAM chromatographic system were derived from a QSRR model. The purpose of this research was to assess isoxazolone derivatives' affinity toward phospholipids. The model was developed using differential evolution combined with partial least squares regression (PLS). Molecular descriptors carrying the information referred to van der Waals volume as well as those defined based upon the weighted holistic invariant *molecular* theory *(*WHIM), geometry, topology, and atom-weights assembly theory (GETAWAY), and *3D molecule representation of structures based on electron diffraction theory* (3D MORSE), stood out as descriptors of importance, carrying the information related to molecular size, shape, symmetry, and atomic distribution. However, polarizability related and descriptors

based on chemically advanced template search theory (CATS) were omitted despite being important for lipophilicity determination. The interpretation of these results led to a conclusion about the insufficient binding of isoxazolone derivatives to phospholipid molecules [90].

In another study, Buszevski et al. tried to gain insight into the biological activity of 30 flavonoids using IAM chromatographic analysis. The GA-PLS algorithm was used for QSRR model development. The conclusion about retention mechanisms was made upon quantum chemical descriptors, indicating that hydrophobic forces, dispersion effect, and electrostatic interactions govern the retention behavior of flavonoids in IAM chromatographic separation [36].

3.2.2 Mixed-mode liquid chromatography

A promising application of QSRR models has also been shown by the explanation of MMLC, where multiple functionalities in charge of providing different intermolecular interactions are integrated into a single stationary phase.

Obradović et al. developed QSRR models to characterize an MMLC system in which RP and hydrophilic interaction (HILIC) modes participate equally. Forty-three substances, serotonin, and imidazole receptor ligands were tested. Interestingly, separate QSRR models for four different types of responses were developed. The retention factor in pure eluents and the turning point for modality shifting were used as selected outputs. For characterizing the partition process in the RP mode, atomic mass, lipophilicity, and intermolecular hydrophobic interactions were proved to be important. The partition process in the HILIC modality was characterized by lipophilicity, distribution of ionic forms, and electrostatic properties. Adsorption, on the other hand, was driven by molecular geometry, electronegativity, polarizability, van der Waals volume, and atomic mass of the tested analytes. For the turning point and modality expressions, distribution of ionic forms, hydrogen bonding properties, and electronic properties, as well as atomic mass, were significant [91].

Russo et al. used a QSRR model developed by PLS in combination with block relevance (BR) to detect retention mechanisms provided by the arginine stationary phase. Due to the diverse interaction ability of the stationary phase, analytes with diverse ionization capacity (neutral, acids, and bases) were selected. It was noticed that the analyte's size and hydrogen donor capacity were important for the retention of neutral substances. For acidic molecules, descriptors calculated with VolSurf+ software and VS+ descriptors, did not describe the electric charge well enough; the MLR strategy was used for confirmation of the electrostatic background of acidic analytes' retention. Also, with the constructed QSRR model, it is possible to recognize the turning point for modality shifting. The basic substances did not show a sufficient degree of retention, so it was not possible to qualitatively define the retention mechanisms involved in their separation [92].

3.3 Future perspectives

With the use of adequate mathematical tools for linking input variables (both molecular descriptors and experimental parameters) with suitable responses, the statistical approach of QSRR modeling does not recognize limitations regarding the type of HPLC system that needs to be characterized. For this reason, it is considered that especially mixed QSRR models can significantly improve the understanding and

development of HPLC methods when complex retention patterns are present due to a possible reduction of the requirements for experimental work.

4. Conclusion

It can be concluded from the literature that QSRR models have been widely applied in chromatographic science, this topic is, therefore, of great interest to researchers in different scientific areas. This chapter has presented the QSRR models with structuring possibilities in detail, the importance of molecular descriptors, and machine learning algorithms selection, as well as different approaches to conducting these important tasks. It can be also used as a guideline when choosing internal and external validation approaches to apply in the consideration of their main advantages and disadvantages. Special attention was put into disclosing the most important QSRR model applications, by pointing out the possibilities of investigating modified HPLC systems that are of great interest to analysts working with different kinds of compounds.

Acknowledgements

This work was supported by the Ministry of Education, Science and Technological Development, the Republic of Serbia through a grant agreement with the University of Belgrade – Faculty of Pharmacy No: 451-03-68/2022-2114/200161.

Conflict of interest

We declare that there is no conflict of interest.

Author details

Jovana Krmar, Bojana Svrkota, Nevena Đajić, Jevrem Stojanović, Ana Protić and Biljana Otašević*
Faculty of Pharmacy, University of Belgrade, Belgrade, Republic of Serbia

*Address all correspondence to: biljana.otasevic@pharmacy.bg.ac.rs

References

[1] Put R, Vander HY. Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure–retention relationships. Analytica Chimica Acta. 2007; **602**(2):164-172. DOI: 10.1016/j.aca.2007.09.014

[2] Héberger K. Quantitative structure–(chromatographic) retention relationships. Journal of Chromatography. A. 2007;**1158**(1): 273-305. DOI: 10.1016/j.chroma.2007.03.108

[3] Kaliszan R. Chapter 17 - Quantitative structure property (Retention) relationships in liquid chromatography. In: Fanali S, Haddad PR, Poole CF, Schoenmakers P, Lloyd D, editors. Liquid Chromatography. Amsterdam: Elsevier; 2013. pp. 385-405. DOI: 10.1016/B978-0-12-415807-8.00017-1

[4] Muteki K, Morgado JE, Reid GL, Wang J, Xue G, Riley FW, et al. Quantitative structure retention relationship models in an analytical quality by design framework: Simultaneously accounting for compound properties, Mobile-phase conditions, and stationary-phase properties. Industrial and Engineering Chemistry Research. 2013; **52**(35):12269-12284. DOI: 10.1021/ie303459a

[5] Kaliszan R. QSRR: Quantitative structure-(chromatographic) retention relationships. Chemical Reviews. 2007; **107**(7):3212-3246. DOI: 10.1021/cr068412z

[6] Kaliszan R. Quantitative structure-retention relationships (QSRR) in chromatography. In: Wilson ID, editor. Encyclopedia of Separation Science. Oxford: Academic Press; 2000. pp. 4063-4075. DOI: 10.1016/b0-12-226770-2/01911-6

[7] Bodzioch K, Durand A, Kaliszan R, Bączek T, Vander HY. Advanced QSRR modeling of peptides behavior in RPLC. Talanta. 2010;**81**(4):1711-1718. DOI: 10.1016/j.talanta.2010.03.028

[8] Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON software: An easy approach to molecular descriptor calculations. MATCH Communications in Mathematical and in Computer Chemistry. 2006;**56**: 237-248

[9] Golubović J. Application of Artificial Neural Networks in Building Models to Predict Retention Behaviour and Intensity of Mass Spectrometric Response in the Analysis of the Selected Azoles and Sartans by High Performance Liquid Chromatography. Belgrade: University of Belgrade - Faculty of Pharmacy; 2016

[10] Mao J, Akhtar J, Zhang X, Sun L, Guan S, Li X, et al. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. iScience. 2021; **24**(9):103052-103052. DOI: 10.1016/j.isci.2021.103052

[11] Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim: Wiley VCHVerlag GmbH; 2000. p. 667. DOI: 10.1002/9783527613106

[12] Szucs R, Brown R, Brunelli C, Heaton JC, Hradski J. Structure driven prediction of chromatographic retention times: Applications to pharmaceutical analysis. International Journal of Molecular Sciences. 2021;**22**(8):3848

[13] Si-Hung L, Izumi Y, Nakao M, Takahashi M, Bamba T. Investigation of supercritical fluid chromatography retention behaviors using quantitative structure-retention relationships. Analytica Chimica Acta. 2022;**1197**: 339463-339463. DOI: 10.1016/j.aca.2022.339463

[14] Rojas C, Aranda JF, Pacheco Jaramillo E, Losilla I, Tripaldi P, Duchowicz PR, et al. Foodinformatic prediction of the retention time of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-orbitrap. Food Chemistry. 2021;**342**: 128354-128354. DOI: 10.1016/j.foodchem.2020.128354

[15] Park SH, Haddad PR, Talebi M, Tyteca E, Amos RIJ, Szucs R, et al. Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model. Journal of Chromatography. A. 2017;**1486**:68-75

[16] D'Archivio AA, Maggi MA, Ruggieri F. Modelling of UPLC behaviour of acylcarnitines by quantitative structure–retention relationships. Journal of Pharmaceutical and Biomedical Analysis. 2014;**96**: 224-230. DOI: 10.1016/j.jpba.2014.04.006

[17] Akbar J, Iqbal S, Batool F, Karim A, Chan KW. Predicting retention times of naturally occurring phenolic compounds in reversed-phase liquid chromatography: A quantitative structure-retention relationship (QSRR) approach. International Journal of Molecular Sciences. 2012;**13**(11): 15387-15400. DOI: 10.3390/ijms131115387

[18] Oliveira TB, Gobbo-Neto L, Schmidt TJ, Da Costa FB. Study of chromatographic retention of natural terpenoids by chemoinformatic tools. Journal of Chemical Information and Modeling. 2015;**55**(1):26-38. DOI: 10.1021/ci500581q

[19] Dobričić V, Nikolic K, Vladimirov S, Čudina O. Biopartitioning micellar chromatography as a predictive tool for skin and corneal permeability of newly synthesized 17β-carboxamide steroids. European Journal of Pharmaceutical Sciences. 2014;**56**:105-112. DOI: 10.1016/j.ejps.2014.02.007

[20] Filipic S, Elek M, Nikolic K, Agbaba D. Quantitative structure-retention relationship Modeling of the retention behavior of guanidine and imidazoline derivatives in reversed-phase thin-layer chromatography. JPC Journal of Planar Chromatography Modern TLC. 2015;**28**(2):119-125. DOI: 10.1556/jpc.28.2015.2.6

[21] Karadžić Banjac M, Jevrić L, Kovačević S, Podunavac-Kuzmanovic S. Retention data from Normal-phase thin-layer chromatography in characterization of some 1,6-Anhydrohexose and D-Aldopentose derivatives by QSRR method. Journal of Liquid Chromatography and Related Technologies. 2015;**38**:1044-1044. DOI: 10.1080/10826076.2015.1012521

[22] Naylor BC, Catrow JL, Maschek JA, Cox JE. QSRR Automator: A tool for automating retention time prediction in Lipidomics and metabolomics. Metabolites. 2020;**10**(6):237

[23] Wang YT, Yang ZX, Piao ZH, Xu XJ, Yu JH, Zhang YH. Prediction of flavor and retention index for compounds in beer depending on molecular structure using a machine learning method. RSC Advances. 2021;**11**(58):36942-36950. DOI: 10.1039/D1RA06551C

[24] Amos RIJ, Haddad PR, Szucs R, Dolan JW, Pohl CA. Molecular modeling and prediction accuracy in quantitative structure-retention relationship calculations for chromatography. TrAC, Trends in Analytical Chemistry. 2018; **105**:352-359. DOI: 10.1016/j.trac.2018.05.019

[25] Bálint D, Jäntschi L. Comparison of molecular geometry optimization methods based on molecular descriptors. Mathematics. 2021;**9**(22):2855

[26] Amos RIJ, Tyteca E, Talebi M, Haddad PR, Szucs R, Dolan JW, et al. Benchmarking of computational methods for creation of retention models in quantitative structure–retention relationships studies. Journal of Chemical Information and Modeling. 2017;**57**(11):2754-2762. DOI: 10.1021/acs.jcim.7b00346

[27] Krmar J, Vukićević M, Kovačević A, Protić A, Zečević M, Otašević B. Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure - retention relationships modelling in micellar liquid chromatography. Journal of Chromatography. A. 2020;**1623**: 461146-461146. DOI: 10.1016/j.chroma.2020.461146

[28] Talebi M, Schuster G, Shellie RA, Szucs R, Haddad PR. Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography. Journal of Chromatography. A. 2015;**1424**:69-76. DOI: 10.1016/j.chroma.2015.10.099

[29] González M, Terán C, Saíz-Urra L, Teijeira M. Variable selection methods in QSAR: An overview. Current Topics in Medicinal Chemistry. 2008;**8**:1606-1627. DOI: 10.2174/156802608786786552

[30] Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics. 2018;**85**:189-203. DOI: 10.1016/j.jbi.2018.07.014

[31] Goodarzi M, Jensen R, Vander HY. QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions. Journal of Chromatography B. 2012;**910**:84-94. DOI: 10.1016/j.jchromb.2012.01.012

[32] Hancock T, Put R, Coomans D, Vander Heyden Y, Everingham Y. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. Chemometrics and Intelligent Laboratory Systems. 2005; **76**(2):185-196. DOI: 10.1016/j.chemolab.2004.11.001

[33] Mizera M, Talaczyńska A, Zalewski P, Skibiński R, Cielecka-Piontek J. Prediction of HPLC retention times of tebipenempivoxyl and its degradation products in solid state by applying adaptive artificial neural network with recursive features elimination. Talanta. 2015;**137**:174-181. DOI: 10.1016/j.talanta.2015.01.032

[34] Daghir-Wojtkowiak E, Wiczling P, Bocian S, Kubik Ł, Kośliński P, Buszewski B, et al. Least absolute shrinkage and selection operator and dimensionality reduction techniques in quantitative structure retention relationship modeling of retention in hydrophilic interaction liquid chromatography. Journal of Chromatography. A. 2015;**1403**:54-62. DOI: 10.1016/j.chroma.2015.05.025

[35] Wen Y, Amos RIJ, Talebi M, Szucs R, Dolan JW, Pohl CA, et al. Retention index prediction using quantitative structure–retention relationships for improving structure identification in nontargeted metabolomics. Analytical Chemistry. 2018;**90**(15):9434-9440. DOI: 10.1021/acs.analchem.8b02084

[36] Buszewski B, Žuvela P, Sagandykova G, Walczak-Skierska J, Pomastowski P, David J, et al. Mechanistic chromatographic column characterization for the analysis of flavonoids using quantitative structure-retention relationships based on density functional theory. International Journal of Molecular Sciences. 2020;**21**(6):2053

[37] Park SH, De Pra M, Haddad PR, Grosse S, Pohl CA, Steiner F. Localised quantitative structure–retention relationship modelling for rapid method development in reversed-phase high performance liquid chromatography. Journal of Chromatography. A. 2020; **1609**:460508-460508. DOI: 10.1016/ j.chroma.2019.460508

[38] Yan P, Wang L, Li S, Liu X, Sun Y, Tao J, et al. Improved structural annotation of triterpene metabolites of traditional Chinese medicine in vivo based on quantitative structure-retention relationships combined with characteristic ions: AlismatisRhizoma as an example. Journal of Chromatography B. 2021;**1187**: 123012-123012. DOI: 10.1016/j.jchromb. 2021.123012

[39] Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning. 2003;**53**(1):23-69. DOI: 10.1023/A: 1025667309714

[40] Pawellek R, Krmar J, Leistner A, Djajić N, Otašević B, Protić A, et al. Charged aerosol detector response modeling for fatty acids based on experimental settings and molecular features: A machine learning approach. Journal of Cheminformatics. 2021;**13**(1): 53-53. DOI: 10.1186/s13321-021-00532-0

[41] Chen J, Tang YY, Fang B, Guo C. In silico prediction of toxic action mechanisms of phenols for imbalanced data with random Forest learner. Journal of Molecular Graphics & Modelling. 2012;**35**:21-27. DOI: 10.1016/j.jmgm. 2012.01.002

[42] Gupta S, Basant N, Mohan D, Singh KP. Room-temperature and temperature-dependent QSRR modelling for predicting the nitrate radical reaction rate constants of organic chemicals using ensemble learning methods. SAR and QSAR in Environmental Research. 2016; **27**(7):539-558. DOI: 10.1080/1062936X. 2016.1199592

[43] Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis. 2000;**22**(5): 717-727. DOI: 10.1016/s0731-7085(99) 00272-1

[44] Žuvela P, Macur K, Jay Liu J, Bączek T. Exploiting non-linear relationships between retention time and molecular structure of peptides originating from proteomes and comparing three multivariate approaches. Journal of Pharmaceutical and Biomedical Analysis. 2016;**127**: 94-100. DOI: 10.1016/j.jpba.2016.01.055

[45] Fatemi MH, Ghorbanzad'e M, Baher E. Quantitative structure retention relationship Modeling of retention time for some organic pollutants. Analytical Letters. 2010;**43**(5):823-835. DOI: 10.1080/00032710903486294

[46] Mozafari Z, Arab Chamjangali M, Arashi M, Goudarzi N. QSRR models for predicting the retention indices of VOCs in different datasets using an efficient variable selection method coupled with artificial neural network modeling: ANN-based QSPR modeling. Journal of the Iranian Chemical Society. 2022; **19**(6):2617-2630. DOI: 10.1007/s13738-021-02488-2

[47] Noorizadeh H, Farmany A, Narimani H, Noorizadeh M. QSRR using evolved artificial neural network for 52 common pharmaceuticals and drugs of abuse in hair from UPLC–TOF-MS. Drug Testing and Analysis. 2013;**5**(5):320-324. DOI: 10.1002/dta.309

[48] Djajić N, Golubović J, Otašević B, Zecevic M, Protić A. Quantitative structure –retention relationship modeling of selected antipsychotics and their impurities in green liquid chromatography using cyclodextrin mobile phases. Analytical and Bioanalytical Chemistry. 2018;**410**: 2533-2550. DOI: 10.1007/s00216-018-0911-3

[49] Golubović J, Protić A, Otašević B, Zečević M. Quantitative structure–retention relationships applied to development of liquid chromatography gradient-elution method for the separation of sartans. Talanta. 2016;**150**: 190-197. DOI: 10.1016/j.talanta.2015.12.035

[50] D'Archivio A, Maggi M, Ruggieri F. Artificial neural network prediction of multilinear gradient retention in reversed-phase HPLC: Comprehensive QSRR-based models combining categorical or structural solute descriptors and gradient profile parameters. Analytical and Bioanalytical Chemistry. 2014;**407**: 1181-1190. DOI: 10.1007/s00216-014-8317-3

[51] Dobričić V, Savić J, Nikolic K, Vladimirov S, Vujić Z, Brborić J. Application of biopartitioning micellar chromatography and QSRR modeling for prediction of gastrointestinal absorption and design of novel β-hydroxy-β-arylalkanoic acids. European Journal of Pharmaceutical Sciences. 2017;**100**: 280-284. DOI: 10.1016/j.ejps.2017.01.023

[52] Parinet J. Predicting reversed-phase liquid chromatographic retention times of pesticides by deep neural networks. Heliyon. 2021;**7**(12):e08563-e08563. DOI: 10.1016/j.heliyon.2021.e08563

[53] Ju R, Liu X, Zheng F, Lu X, Xu G, Lin X. Deep neural network pretrained by weighted autoencoders and transfer learning for retention time prediction of small molecules. Analytical Chemistry. 2021;**93**(47):15651-15658. DOI: 10.1021/acs.analchem.1c03250

[54] Pasin D, Mollerup CB, Rasmussen BS, Linnet K, Dalsgaard PW. Development of a single retention time prediction model integrating multiple liquid chromatography systems: Application to new psychoactive substances. Analytica Chimica Acta. 2021;**1184**:339035-339035. DOI: 10.1016/j.aca.2021.339035

[55] Randazzo GM, Bileck A, Danani A, Vogt B, Groessl M. Steroid identification via deep learning retention time predictions and two-dimensional gas chromatography-high resolution mass spectrometry. Journal of Chromatography. A. 2020;**1612**: 460661-460661. DOI: 10.1016/j.chroma.2019.460661

[56] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discovery Today. 2018;**23**(6):1241-1250. DOI: 10.1016/j.drudis.2018.01.039

[57] Adadi A. A survey on data-efficient algorithms in big data era. Journal of Big Data. 2021;**8**(1):24-24. DOI: 10.1186/s40537-021-00419-9

[58] Ciura K, Pastewska M, Ulenberg S, Kapica H, Kawczak P, Bączek T. Chemometric analysis of bio-inspired micellar electrokinetic chromatographic systems – Modelling of retention mechanism and prediction of biological properties using bile salts surfactants. Microchemical Journal. 2021;**167**: 106340-106340. DOI: 10.1016/j.microc.2021.106340

[59] Aalizadeh R, Nika MC, Thomaidis N. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. Journal of Hazardous Materials. 2019;**363**:275-288. DOI: 10.1016/j.jhazmat.2018.09.047

[60] Riahi S, Pourbasheer E, Ganjali MR, Norouzi P. Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. Journal of Hazardous Materials. 2009;**166**(2): 853-859. DOI: 10.1016/j.jhazmat.2008.11.097

[61] Zhang X, Zhang X, Li Q, Sun Z, Song L, Sun T. Support vector machine applied to study on quantitative structure–retention relationships of polybrominated diphenyl ether congeners. Chromatographia. 2014;**77**: 1387-1398. DOI: 10.1007/s10337-014-2735-4

[62] Song M, Breneman CM, Bi J, Sukumar N, Bennett KP, Cramer S, et al. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. Journal of Chemical Information and Computer Sciences. 2002;**42**(6): 1347-1357. DOI: 10.1021/ci025580t

[63] Čolović J, Kalinić M, Vemić A, Eric S, Malenović A. 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. Journal of Chromatography. A. 2015; **1425**:150-157. DOI: 10.1016/j.chroma.2015.11.027

[64] Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q. Boosting: An ensemble learning tool for compound classification and QSAR Modeling. Journal of Chemical Information and Modeling. 2005;**45**(3):786-799. DOI: 10.1021/ci0500379

[65] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. 1st ed. New York: Routledge; 1984. p. 368. DOI: 10.1201/9781315139470

[66] Goudarzi N, Shahsavani D, Emadi-Gandaghi F, Chamjangali MA. Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons. Journal of Chromatography. A. 2014; **1333**:25-31. DOI: 10.1016/j.chroma.2014.01.048

[67] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random Forest: A classification and regression tool for compound classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences. 2003;**43**(6): 1947-1958. DOI: 10.1021/ci034160g

[68] Wang C, Skibic MJ, Higgs RE, Watson IA, Bui H, Wang J, et al. Evaluating the performances of quantitative structure-retention

relationship models with different sets of molecular descriptors and databases for high-performance liquid chromatography predictions. Journal of Chromatography. A. 2009;**1216**(25): 5030-5038. DOI: 10.1016/j.chroma.2009.04.064

[69] Yang JJ, Han Y, Mah CH, Wanjaya E, Peng B, Xu TF, et al. Streamlined MRM method transfer between instruments assisted with HRMS matching and retention-time prediction. Analytica Chimica Acta. 2020;**1100**:88-96. DOI: 10.1016/j.aca.2019.12.002

[70] Goudarzi N, Shahsavani D. Application of a random forests (RF) method as a new approach for variable selection and modelling in a QSRR study to predict the relative retention time of some polybrominateddiphenylethers (PBDEs). Analytical Methods. 2012;**4**: 3733-3738. DOI: 10.1039/c2ay25484k

[71] Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. In: Hastie T, Tibshirani R, Friedman J, editors. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer New York; 2009. pp. 337-387. DOI: 10.1007/978-0-387-84858-7_10

[72] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning. 1999;**36**(1): 105-139. DOI: 10.1023/A:1007515423169

[73] Bouwmeester R, Martens L, Degroeve S. Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction. Analytical Chemistry. 2019;**91**(5):3694-3703. DOI: 10.1021/acs.analchem.8b05820

[74] Liapikos T, Zisi C, Kodra D, Kademoglou K, Diamantidou D, Begou O, et al. Quantitative structure retention relationship (QSRR) modelling for analytes' retention prediction in LC-HRMS by applying different machine learning algorithms and evaluating their performance. Journal of Chromatography B. 2022;**1191**: 123132-123132. DOI: 10.1016/j.jchromb.2022.123132

[75] Veerasamy R, Rajak H, Jain A, Sivadasan S, Christapher PV, Agrawal R. Validation of QSAR models - strategies and importance. International Journal of Drug Design and Discovery. 2011;**2**: 511-519

[76] Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H. Comparative studies on some metrics for external validation of QSPR models. Journal of Chemical Information and Modeling. 2012;**52**(2): 396-408. DOI: 10.1021/ci200520g

[77] Kiralj R, Ferreira M. Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application. Journal of the Brazilian Chemical Society. 2008;**20**:770-787. DOI: 10.1590/S0103-50532009000400021

[78] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Molecular Informatics. 2010;**29**(6–7):476-488. DOI: 10.1002/minf.201000061

[79] Parinet J. Prediction of pesticide retention time in reversed-phase liquid chromatography using quantitative-structure retention relationship models: A comparative study of seven molecular descriptors datasets. Chemosphere. 2021;**275**:130036-130036. DOI: 10.1016/j.chemosphere.2021.130036

[80] Roy K, Ambure P, Aher RB. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models?

Chemometrics and Intelligent Laboratory Systems. 2017;**162**:44-54. DOI: 10.1016/j.chemolab.2017.01.010

[81] Taraji M, Haddad PR, Amos RIJ, Talebi M, Szucs R, Dolan JW, et al. Error measures in quantitative structure-retention relationships studies. Journal of Chromatography. A. 2017;**1524**: 298-302. DOI: 10.1016/j.chroma.2017.09.050

[82] Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems. 2015;**145**:22-29. DOI: 10.1016/j.chemolab.2015.04.013

[83] Roy K, Kar S, Das RN. Statistical methods in QSAR/QSPR. In: Roy K, Kar S, Das RN, editors. A Primer on QSAR/QSPR Modeling: Fundamental Concepts. Cham: Springer International Publishing; 2015. pp. 37-59. DOI: 10.1007/978-3-319-17281-1_2

[84] Haddad PR, Taraji M, Szücs R. Prediction of analyte retention time in liquid chromatography. Analytical Chemistry. 2021;**93**(1):228-256. DOI: 10.1021/acs.analchem.0c04190

[85] Djajić N, Krmar J, Rmandić M, Rašević M, Otašević B, Zečević M, et al. Modified aqueous mobile phases: A way to improve retention behavior of active pharmaceutical compounds and their impurities in liquid chromatography. Journal of Chromatography Open. 2022; **2**:100023-100023. DOI: 10.1016/j.jcoa.2021.100023

[86] Cecchi T, Passamonti P. Retention mechanism for ion-pair chromatography with chaotropic reagents. Journal of Chromatography. A. 2009;**1216**(10): 1789-1797. DOI: 10.1016/j.chroma.2008.10.031

[87] Ramezani A, Yousefinejad S, Shahsavar A, Mohajeri A, Absalan G. Quantitative structure-retention relationship for chromatographic behaviour of anthraquinone derivatives through considering organic modifier features in micellar liquid chromatography. Journal of Chromatography A. 2019;**1599**:46-54. DOI: 10.1016/j.chroma.2019.03.063

[88] Djajić N, Otašević B, Malenović A, Zecevic M, Protić A. Quantitative structure retention relationship modeling as potential tool in chromatographic determination of stability constants and thermodynamic parameters of β-cyclodextrin complexation process. Journal of Chromatography. A. 2020;**1619**: 460971-460971. DOI: 10.1016/j.chroma.2020.460971

[89] Djajić N, Petković M, Zečević M, Otašević B, Malenović A, Holzgrabe U, et al. A comprehensive study on retention of selected model substances in β-cyclodextrin-modified high performance liquid chromatography. Journal of Chromatography. A. 2021; **1645**:462120-462120. DOI: 10.1016/j.chroma.2021.462120

[90] Ciura K, Fedorowicz J, Žuvela P, Lovrić M, Kapica H, Baranowski P, et al. Affinity of antifungal Isoxazolo[3,4-b] pyridine-3(1H)-ones to phospholipids in immobilized artificial membrane (IAM) chromatography. Molecules. 2020;**25**(20):4835

[91] Obradović D, Oljačić S, Nikolić K, Agbaba D. Investigation and prediction of retention characteristics of imidazoline and serotonin receptor ligands and their related compounds on mixed-mode stationary phase. Journal of Chromatography. A. 2019;**1585**:92-104. DOI: 10.1016/j.chroma.2018.11.051

[92] Russo G, Vallaro M, Cappelli L, Anderson S, Ermondi G, Caron G. Characterization of the new CelerisTM arginine column: Retentive behaviour through a combination of chemometric tools and potential in drug analysis. Journal of Chromatography. A. 2021; **1651**:462316-462316. DOI: 10.1016/j.chroma.2021.462316