A MULTI-TARGET QSRR APPROACH TO MODEL RETENTION TIMES OF SMALL MOLECULES IN RPLC

Priyanka Kumari [a,b,*], Thomas Van Laethem [a,b], Diane Duroux [c], Marianne Fillet [b], Phillipe Hubert [a], Pierre-Yves Sacre [a], Cedric Hubert [a,**]

a Department of Pharmacy, Laboratory of Pharmaceutical Analytical Chemistry, University of Liege (ULiege), CIRM, Quartier Hopital (B36 Tower 4), Avenue Hippocrate, 4000 Liege, Belgium
b Laboratory for the Analysis of Medicines, University of Liege (ULiege), CIRM, Quartier Hopital (B36 Tower 4), Avenue Hippocrate, 4000 Liege, Belgium
c ETH AI Center, OAT X11, Andreasstrasse 5, 8092 Zürich

Keywords: Algorithm adaptation; Molecular descriptors; Multi-target QSRR; Multitask learning; Problem transformation; Random Forest; Regression chain; Reverse Phase Liquid Chromatography; Clinical Biochemistry; Spectroscopy; Drug Discovery; Pharmaceutical Science; Analytical Chemistry

Abstract

Quantitative structure-retention relationship models (QSRR) have been utilized as an alternative to costly and time-consuming separation analyses and associated experiments for predicting retention time. However, achieving 100 % accuracy in retention prediction is unrealistic despite the existence of various tools and approaches. The limitations of vast data availability and time complexity hinder the use of most algorithms for retention prediction. Therefore, in this study, we examined and compared two approaches for modelling retention time using a dataset of small molecules with retention times obtained at multiple conditions, referred to as multi-targets (five pH levels: 2.7, 3.5, 5, 6.5, and 8 at gradient times of 20 min of mobile phase). The first approach involved developing separate models for predicting retention time at each condition (single-target approach), while the second approach aimed to learn a single model for predicting retention across all conditions simultaneously (multi-target approach). Our findings highlight the advantages of the multi-target approach over the single-target modelling approach. The multi-target models are more efficient in terms of size and learning speed compared to the single-target models. These retention prediction models offer two-fold benefits. Firstly, they enhance knowledge and understanding of retention times, identifying molecular descriptors that contribute to changes in retention behaviour under different pH conditions. Secondly, these approaches can be extended to address other multi-target property prediction problems, such as multi-quantitative structure Property(X) relationship studies (mt-QS(X)R).

. Introduction

In the field of analytical chemistry, chromatographic separation has emerged as a powerful technique for separating and analysing complex mixtures. Extensive studies are conducted using various analytical techniques to gain a deeper understanding of the analytes present in a given sample, among which chromatography plays a prominent role. Retention time, a fundamental chromatography parameter, is a critical indicator of an analyte's behaviour within the chromatographic system and holds vital information for its separation and identification. It is often determined through a trial-and-error process, which can be time- consuming and expensive, especially when retention times need to be determined at multiple conditions. In the case of Reverse Phase Liquid Chromatography(RPLC), a widely studied type of chromatography, retention time(tR) can be influenced by various factors. These factors include pH, column type, mobile phase composition, and other variables encountered in various chromatographic techniques. As a result, accurately determining the retention time requires conducting multiple experiments to account for these variables effectively. This can become cost-prohibitive, particularly in high-throughput screening applications. An alternative way of retention evaluation is computational methods using quantitative structure retention relationship models (QSRRs) [1,2]. QSRR is an advanced approach that establishes a statistical relationship between various attributes, such as chemical, physical, and physicochemical properties, and the data associated with the structure of molecules, commonly known as structure-derived descriptors [3]. By carefully selecting appropriate molecular descriptors and utilizing statistical modelling methodologies, a QSRR model can be developed that is both statistically robust and stable [4].

The field of QSRR has undergone significant advancements, progressing from basic linear regression models to sophisticated machine learning algorithms, including algorithms like GA-PLS [5], Bayesian Ridge Regression, Extreme Gradient Boosting Regression, Support Vector Regression etc. [6]. Traditionally, each study in QSRR has employed a single task or single-targeting approach, wherein a separate model is constructed for each response or target in regression studies. In recent studies [7,8], researchers have delved into mixed Quantitative Structure-Retention Relationship (QSRR) models. However, these models predominantly depend on descriptors for target prediction, employing multiple algorithms and feature engineering. However, this approach overlooks the fact that a single molecule can elicit different responses under varying chemical environments and experimental conditions during separation. Consequently, this creates challenges related to multitasking. None of the previous studies have addressed this issue in retention prediction, where multiple experimental targets or responses are considered in the data, thereby neglecting the correlation between these targets. The time and cost required for modelling can vary significantly depending on the number of targets. Employing single-target approaches in QSRR models would not be time and cost-effective when multiple targets need to be predicted. Conversely, multi-target models would be more suitable in such cases.

While multitasking models have been utilized in other fields for activity prediction [9,10], lipophilicity [11], toxicity [12], brain penetration [13], and more, the chromatography field has primarily overlooked their potential application. Some studies have explored multi-output regression in fields like real-time train arrival time prediction [14], ecological modeling [15], gas-phase kinetic rate constants prediction of chemicals [16], and chemometrics to infer concentrations of several analytes from multivariate calibration [17]. However, to our knowledge, none of the previous works have addressed the challenge of incorporating target relationships, including various retention times under varied conditions, into retention prediction models. Therefore, in this study, we aimed to explore different approaches to QSRR modelling for a comprehensive analysis.

In the literature, two methods of multi-target modelling have been reported [18]: (1) the problem transformation method and (2) the algorithm adaptation method.

- **Problem transformation method**: The problem transformation method involves converting the original multi-output regression problem into one or more single-output regression sub-problems, which can be solved using traditional single-output regression algorithms. Several techniques fall under this approach, including the *Independent Model (IM)*, where each output variable is modelled independently using separate single-output regression models. The input features train each model separately, independently predicting each output variable. Another technique is the *Transformation-Based (TB) approach*, where the multi-output regression problem is transformed into a series of single-output regression problems by combining the input features with transformation functions. Separate single-output regression models are then trained for each output variable using these transformed features. An example of this approach is the chaining or regressor chain method [19].
- **Algorithm adaptation method**: The algorithm adaptation method involves modifying existing single-output regression algorithms to handle multiple output variables directly. This is a Multi Task Learning (MTL), where a single model is trained to predict multiple output variables jointly by optimizing a standard objective function that considers all the output variables simultaneously. The idea behind this approach is that the model can leverage the dependencies between the output variables to improve overall prediction performance.

To summarize, relying on a single-target approach-based model may not be sufficient for retention prediction models in real-world scenarios. Although creating separate models for each response variable is an option, it can be time-consuming and less accurate.

Therefore, multi-output-multi-target prediction models, known as the "mt-QSRR modelling" approach, can be a more efficient alternative [20]. The practical utility of mt-QSRR models can be effectively extended and comprehended within the context of analytical method development, particularly for emerging pharmaceutical products. In such scenarios, where the "analytical quality by design" framework is followed [21], the implementation of the design of experiments (DoE) becomes imperative to establish a design space. This design space ensures that the chromatographic method exhibits desirable properties, including robustness in the face of experimental parameters [22]. However, conducting numerous laboratory experiments to identify optimal experimental conditions for the DoE can be time-consuming and resource-intensive. To address this challenge, the initial screening phase can be conveniently performed in silico utilizing one mt-QSRR model, even if their accuracy may not be exceptional. By employing these models, a range of parameters can be selected, significantly streamlining the subsequent experimental optimization DoE [23]. This allows for the identification of the most favourable separation and robustness conditions through practical experimentation not only in analytical chemistry but in other pharmaceutical and biomedical analysis as well [24–26].

In this study, we have compared the model performance of QSRR models based on single-target learning over multi-target learning(mt- QSRR) using retention data gathered for five pHs. Multi-target learning approach offers several advantages over single-target retention prediction[27,28], including considering interdependencies between targets, reducing computational burden by using a single model, improving model interpretability, and training on larger datasets to enhance generalization and reducing overfitting[29,30]. Multi-target QSRR models(mt-QSRR) can significantly advance quantitative structure retention prediction and holds promise for applications in drug discovery, environmental analysis, and other fields where accurate and efficient retention times are critical for chromatographic separations.

Materials and methods

2.1. PROBLEM DEFINITION

For a given data set P containing feature and target couple (x, y) with $\mathrm{x} \in \mathrm{X}$, the input vector and $\mathrm{y} \in \mathrm{Y} = \mathrm{Y}_1 \times \ldots \times \mathrm{Y}_\mathrm{n}$ the target vector. Denote with $yi \in Yi\ the\ i'th$ component of y.

Hence, the mt-QSRR model can be defined as: $\mathrm{Y}_\mathrm{n} = \mathrm{f}(\mathrm{X})$

In **single-target approach:** A learner learns from a data set $P = \{(x, yi)\}$, with $yi \in Yi$ a scalar variable, a function *f_i*: *X* → *Y_i* such that $\sum_{(x,y_i)\in P} L_i(fi(\mathrm{x}), y_i)$ is minimized, with *L_i* some loss function over *Y_i*.

In **multi-target approach:** A learner learns from a data set $P = \{(x, y_i)\}$, with $y \in Y$ an n-dimensional vector, a function $F: X \rightarrow Y$ such that $\sum_{(x,\ y)\in S} L(F(\mathrm{x}),\ \mathrm{y})$ is minimized, with L a loss function over *Y*. In this study, we have checked if the multi-target learner performs better than a single-target learner by checking for any (*x*, *y*), drawn randomly from the population, on average, $L(\mathrm{F}(x), y) < \sum_i L_i(f_i(x), y_i)$

2.2. DATASET

The dataset used in this study was taken from [31], which consists of retention time observed for small pharmaceutical compounds reported in minutes. The data were acquired in RPLC mode at five different pH conditions- 2.7,3.5,5.0,6.5,8.0 with a gradient elution of 0–95 % of methanol in 20 min. The column, flow rate and temperature specification are mentioned in [31,32]. The dataset encompasses compounds with a diverse range of molecular weights, spanning from 46.005 to 454.611 g/mol. The efficacy and usefulness of a model rely heavily on the dataset it is trained on. Therefore, during the data collection process, we prioritised including a diverse range of molecules with varying pKa. This allowed us to capture different trends in retention times as the pH of the analysis increased. Four distinct types of data trends were observed, as depicted in Figs. 1− 3 in the supplementary file.

The training data included various molecule types, with the majority falling into Cases 1 and 3 with 37 % and 33 % of total compounds, while Case 2 with 26 % and a smaller portion belonged to Case 4 with 4 % of the total number of compounds used for modelling (Figure 3 in supplementary file). The retention time showed a strong correlation(in terms of r) across five different pH conditions(Figure 4 supplementary file). Therefore, employing a modelling strategy that considers multiple experimental responses simultaneously and leverages the correlation between the modelled endpoints becomes crucial.

This study used observed retention times at five pH conditions as targets for QSRR modelling. The targets, all with a gradient time of 20 min, are denoted as follows- tR_2.7 for pH 2.7, tR_3.5 for pH 3.5, tR_5.0 for pH 5.0, tR_6.5 for pH 6.5, and tR_8.0 for pH 8.0.

2.3. MOLECULAR DESCRIPTORS

In this study, we employed constitutional, topological, and geometrical descriptors as numerical characteristics to analyze the chemical structures. A total of 225 descriptors were calculated using the RDKit software, which was then utilized to develop models for predicting compound retention based on their physicochemical properties.

Some of these descriptors were aligned with the parameters used in LSER theory, a concept initially applied in retention prediction models [33,34]. LSER theory focuses on the linear solvation energy relationship, which relates solute retention to its solute-solvent interactions. These descriptors capture the specific solvation effects and improve the accuracy of retention prediction models. The remaining descriptors were included to provide additional meaningfulness to the model and enhance its predictive capabilities. The RDKit package, version 2015, was utilized to compute these descriptors derived from the chemical structures [35].

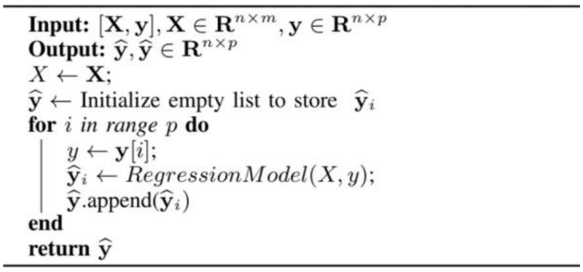
2.4. DATA CLEANING AND PREPROCESSING

Compounds with less than 2 min retention times were classified as non-retained and removed from the dataset. Before modeling, the training data was standardized using a zero mean and unit variance approach. Additionally, the descriptors of the test molecules were standardized using the mean and standard deviation of the training samples.

2.5. QSRR MODELLING

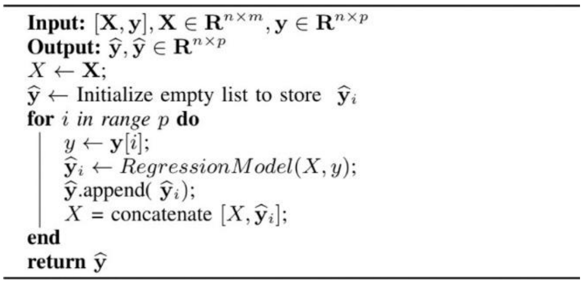
Considering the given data description, our objective was to predict multiple continuous targets (responses) for new test samples based on a set of independent variables. Two approaches were used in this study (Fig. 1) to predict the retention times: the single-target and multi-target regression approaches, which are explained in Section 2.1. The problem transformation and algorithm adaptation methods for retention predictions were employed to check this differentiation. The problem transformation method converts the multi-output regression problem into one or more single-output regression sub-problems. Two ways of modelling were tested for this method- IM and RC(Regressor Chain) methods corresponding to Model1 and Model2, respectively (shown as a red dotted box). On the other hand, the algorithm adaptation method involves modifying existing single-output regression algorithms to handle multiple output variables directly.

Both the RC(Model2) and MTL(Model3) models can handle target correlations but not the Independent model(Model1) that utilizes a multioutput regressor function to build the model. The pseudo algorithms for the three methods are described in Figs. 2, 3 and 4, respectively.



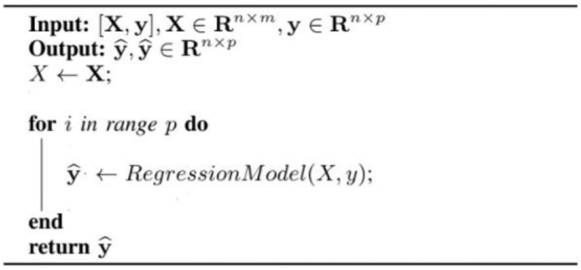
```
Input: $[\mathbf{X},\mathbf{y}], \mathbf{X} \in \mathbf{R}^{n\times m}, \mathbf{y} \in \mathbf{R}^{n\times p}$
Output: $\widehat{\mathbf{y}}, \widehat{\mathbf{y}} \in \mathbf{R}^{n\times p}$
$X \leftarrow \mathbf{X}$;
$\widehat{\mathbf{y}} \leftarrow$ Initialize empty list to store $\widehat{\mathbf{y}}_i$
for $i$ in range $p$ do
    $y \leftarrow \mathbf{y}[i]$;
    $\widehat{\mathbf{y}}_i \leftarrow RegressionModel(X, y)$;
    $\widehat{\mathbf{y}}$.append($\widehat{\mathbf{y}}_i$)
end
return $\widehat{\mathbf{y}}$
```

Fig. 2. Algorithm1: Pseudoalgorithm for DirectMultioutput Regressor (single- target approach used for Model1).



```
Input: $[\mathbf{X},\mathbf{y}], \mathbf{X} \in \mathbf{R}^{n\times m}, \mathbf{y} \in \mathbf{R}^{n\times p}$
Output: $\widehat{\mathbf{y}}, \widehat{\mathbf{y}} \in \mathbf{R}^{n\times p}$
$X \leftarrow \mathbf{X}$;
$\widehat{\mathbf{y}} \leftarrow$ Initialize empty list to store $\widehat{\mathbf{y}}_i$
for $i$ in range $p$ do
    $y \leftarrow \mathbf{y}[i]$;
    $\widehat{\mathbf{y}}_i \leftarrow RegressionModel(X, y)$;
    $\widehat{\mathbf{y}}$.append( $\widehat{\mathbf{y}}_i$);
    $X$ = concatenate $[X, \widehat{\mathbf{y}}_i]$;
end
return $\widehat{\mathbf{y}}$
```

Fig. 3. Algorithm2: Pseudoalgorithm for RegressorChain method (single-target approach used for Model2).



```
Input: $[\mathbf{X},\mathbf{y}], \mathbf{X} \in \mathbf{R}^{n\times m}, \mathbf{y} \in \mathbf{R}^{n\times p}$
Output: $\widehat{\mathbf{y}}, \widehat{\mathbf{y}} \in \mathbf{R}^{n\times p}$
$X \leftarrow \mathbf{X}$;

for $i$ in range $p$ do

    $\widehat{\mathbf{y}} \leftarrow RegressionModel(X, y)$;

end
return $\widehat{\mathbf{y}}$
```

Fig. 4. Algorithm3: Pseudoalgorithm for Algorithm adaptation (multi-target approach used for Model3).

This study focuses on applying a single-target approach (Model1 and Model2) and a multi-target approach (Model3) approaches to deal with the challenges associated with predicting the retention time of small molecules based on multivariate data. The high number of descriptors relative to the compounds used for modelling introduces the possibility of multicollinearity. To address this issue, we employ specific algorithms, with a focus on random forest (RFR) regression method [16,36], which allows for the analysis of multivariate and megavariate data while mitigating the risk of overfitting. By utilizing random forest for retention time prediction, we effectively prevent overfitting and create a robust and reliable model that generalizes well to unseen data. This is achieved through the ensemble nature of the random forest, coupled with feature randomization, bootstrapping, regularization, and out-of-bag (OOB) error estimation. In our analysis, we developed three models using the sklearn library in Python. For Model 1, we utilized a multioutput wrapper around RFR (Random Forest Regressor). Model 2, on the other hand, employed a regressor chain around RFR. Lastly, for Model 3, we directly used the RFR function available from the sklearn.ensemble module. All models were constructed using hyperparameter values as such: n_estimators=100, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=1.0. We employed the impurity based feature importance to calculate the variable importance. This allowed us to identify the descriptors that had the most impact on the predictive performance of the models.

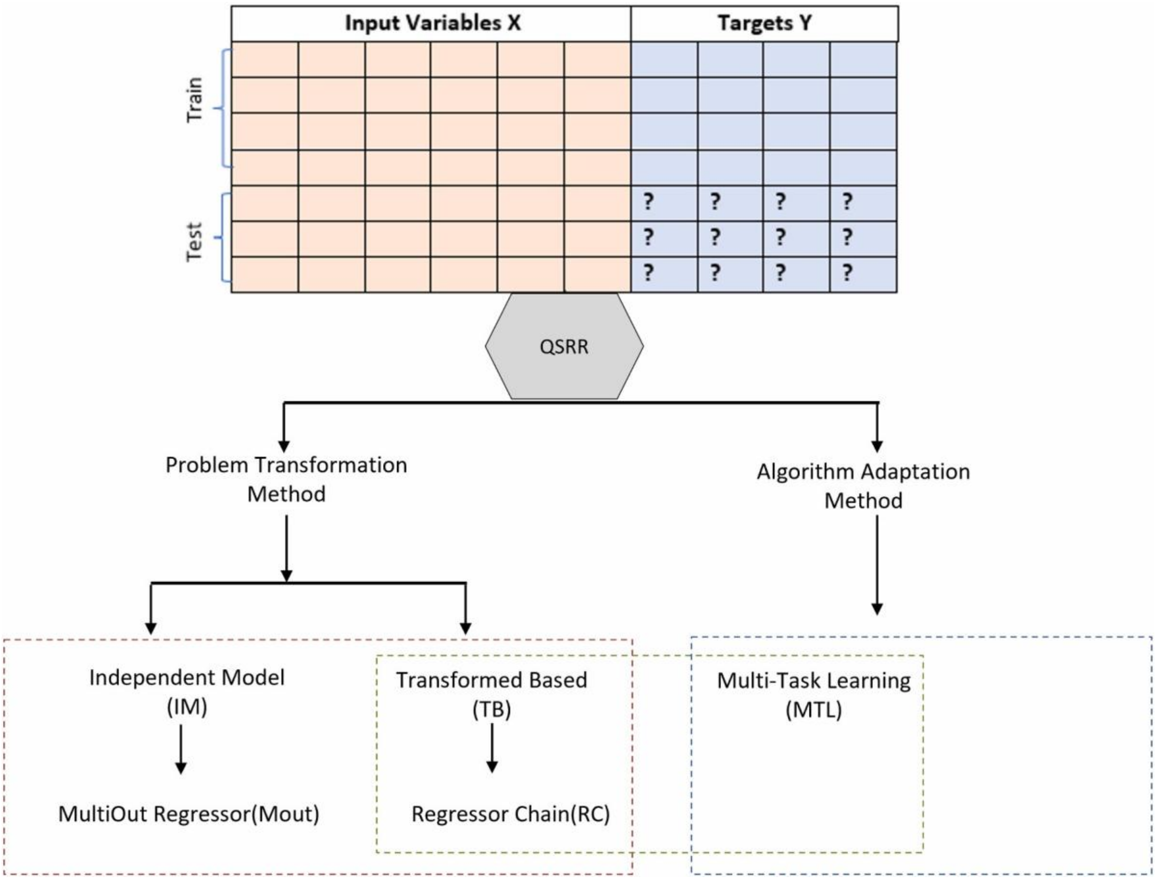


Fig. 1. Different approaches of mt-QSRR models were implemented in this study. Red dotted box: Sequential multiple-output prediction methods with a single-target approach. Blue dotted box: Multi-output simultaneous prediction using a single model approach. Green dotted box: Modeling methods that consider the relationship of the target variable.

2.6. MODEL VALIDATION AND EVALUATION

The developed mt-QSRR model underwent rigorous validation procedures to ensure its accuracy and reliability. Both internal and external validation methods were employed, following a similar approach as outlined in [32], in order to minimize prediction errors across multiple target compounds. For the external validation, a dataset comprising ten compounds was carefully selected based on their diverse trends in observed retention time and chemical nature. This selection ensured that the model's performance was evaluated across a wide range of chemical properties, enhancing its applicability and robustness. By assessing the model's predictive capabilities on this external dataset, its generalizability and ability to handle various compound types were thoroughly assessed. Internal validation, on the other hand, was conducted using a 10-fold cross-validation technique. This method involved dividing the dataset into ten subsets of roughly equal size. The model was trained on nine subsets while utilizing the remaining subset for testing. This process was repeated ten times, each subset serving as the test set once. By performing cross-validation, the model's performance was assessed on multiple iterations, enhancing the credibility of its predictive capabilities. To evaluate the performance of the mt-QSRR model quantitatively, external validation performance measures were calculated. These measures were expressed in terms of the average root mean square error (aRMSE), as shown in Eq. (1), and the average coefficient of determination (aR^2), as shown in Eq. (2). These performance metrics provided a comprehensive assessment of the model's predictive accuracy and its ability to explain the variance in the observed retention times across multiple target compounds. By averaging the performance metrics over all the individual models, a consolidated evaluation was obtained, enabling a comparative analysis between single-target and multi-target prediction approaches. Furthermore, the individual model with the best performance was selected, and its predictions were compared against the corresponding observed values. This visual representation of the model's performance allowed for a more intuitive understanding of its predictive capabilities.

Formulas for calculating RMSE and R^2 for multi-target regression approach:

$$aRMSE = \frac{1}{d}\sum_{i=1}^{d}\sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(Y_i^{(l)} - \widehat{Y}_i^{(l)}\right)^2} \tag{1}$$

$$aR^2 = \frac{1}{d}\sum_{i=1}^{d}\left[1 - \frac{\sum_{l=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{l=1}^{n}(Y_i - \bar{y})^2}\right] \tag{2}$$

In the above-mentioned equations, Y and $\hat{Y}$ represent the observed and predicted retention times of unseen test data respectively, 'n' denotes the number of test molecules while 'd' represents the number of targets which in this study corresponds to five pH conditions.

2.7. SIGNIFICANCE TEST FOR PERFORMANCE DIFFERENCES

To assess whether the differences in performance are statistically significant, we employed the corrected Friedman test [37,38]. The Friedman test is a non-parametric test for multiple hypotheses testing. The algorithms were ranked according to their performances for each dataset separately. The best-performing algorithm was ranked 1, the second 2, and so on. In the situation where there were equal ranks, average rank was used. The Friedman test is based on two assumptions: The *nK*-variate random variables are mutually independent, i.e., the results within one row do not influence the results within the other rows (Tables 2 and 3). The second hypothesis is that the data can be meaningfully ranked. Friedman's test statistic is:

$$T = \frac{12}{nK(K+1)}\sum_{k=1}^{K} R_k^2 - 3n(K+1)$$

where K is the number of models, $R_k = \sum_{i=A}^{n} R_{ik}$ is the sum of the ranks for model k over the n parameters. Under the null hypothesis, the statistic T has an asymptotic Chi-squared distribution with $K - 1$ degrees of freedom. At the α level of significance, the null hypothesis is rejected if $T_1 \geq \chi^2_{K-1;1-\alpha}$, where $\chi^2_{K-1;1-\alpha}$ is the (1-α) quantile of the Chi-squared distribution with $K - 1$ degrees of freedom.

Table 1
Performance measures of each model based on combined prediction(average) of log tR.

Parameters	Model1	Model2	Model3
RMSE-train	0.15	0.15	0.14
RMSE-test	0.15	0.17	0.15
R^2-train	0.74	0.74	0.77
R^2-test	0.7	0.71	0.78

Table 2
Analysis of models for mt-QSRRs based on RMSE for individual targets.

Parameters	Model1 [rank]	Model2 [rank]	Model3 [rank]
tR_2.7	0.09 [2]	0.09 [2]	0.09 [2]
tR_3.5	0.06 [1]	0.12 [3]	0.08 [2]
tR_5.0	0.17 [2]	0.16 [1]	0.18 [3]
tR_6.5	0.20 [2]	0.24 [3]	0.18 [1]
tR_8.0	0.22 [2]	0.25 [3]	0.20 [1]

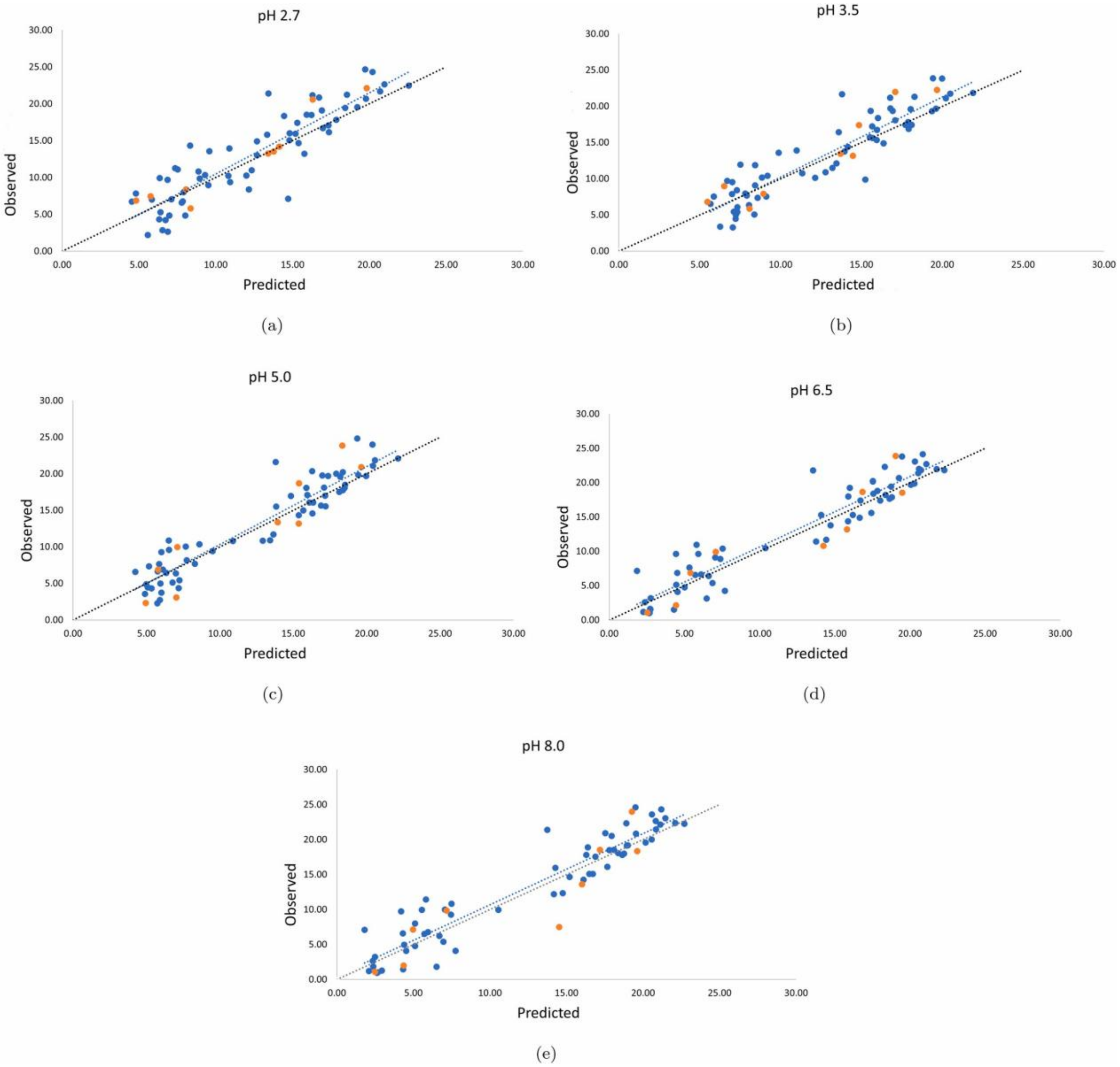
Table 3
Analysis for models for mt-QSRR based on R^2 for individual targets.

Parameters	Model1 [rank]	Model2 [rank]	Model3 [rank]
tR(2.7)	0.75 [3]	0.77 [2]	0.79 [1]
tR(3.5)	0.82 [2]	0.63 [3]	0.85 [1]
tR(5.0)	0.73 [2]	0.76 [1]	0.71 [3]
tR(6.5)	0.77 [2]	0.70 [3]	0.81 [1]
tR(8.0)	0.72 [2]	0.66 [3]	0.76 [1]

3. Results and discussion

3.1. DATA CHARACTERIZATION

The multivariate dataset considered in this study comprised the experimental retention times (in minutes) of diverse small pharmaceutical compounds having varied molecular weights and retention times. The high correlation of retention values across all pH levels (Figure 4 supplementary file) underscores the importance of employing QSRR models that leverage this relationship for predicting retention times.



(a)

(b)

(c)

(d)

(e)

Fig. 5. Plots of Observed retention time(tR) Vs. Experimental retention time(tR) from Model 3 (tR is back transformed in Minutes) for(a) pH 2.7, (b)pH 3.5 (c)pH 5.0 (d)pH 6.5,(d)pH 8.0. Blue points: train, orange points: test, fit line: blue dotted, Regular line: Black dotted.

3.2. MULTI-TARGET QSRR MODELLING AND VALIDATION

This study focuses on studying pH's influence on the retention behaviour of small molecules in RPLC. Here, we attempted to develop an mt-QSRR model for simultaneous prediction of multiple targets that are retention times (retention times at five pH of diverse pharmaceutical compounds). All the targets were experimentally observed as the dependent variables, and the considered compounds' molecular descriptors were calculated computationally as the predictor variables. The optimal model was established by utilizing a training set of 61 compounds. For the most effective model, a set of the top five descriptors was identified using Gini importance, also called mean decrease impurity. While additional descriptors do make a contribution, their importance is comparatively lower. The leading descriptor among them is "MolLogP," which signifies the octanol-water partition coefficient. Other noteworthy descriptors include "LogD," "VSA-Estate5," "SMR- VSA3," and "QED." The model was validated internally using a 10-fold CV and externally with the test set (n = 9). Model1 and Model2 represented prediction from MultiOutput regression and regressor chain methods, and Model 3 as the Algorithm Adaptation method. The performance measures of three mt-QSRR models are given in (Tables 1, 2 and 3). Table 1 displays the performance results based on the average root mean square error (RMSE) computed across all the targets using equations (1) and (2). The models captured 66–85 per cent of the variance in the test data (Table 3 and Fig. 5). A high variance explained by a model implies that the majority of the information present in the data has been encompassed. Moreover, all the developed mt-QSRR models exhibited significantly low RMSE values (< 0.1) for both observed and predicted log values of the target in the test data (Table 2).

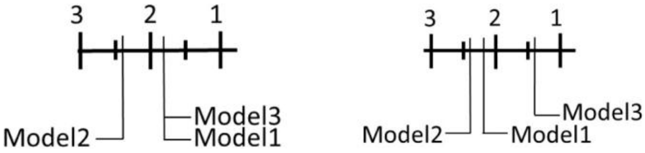


Fig. 6. Average rank of the models based on the RMSE (left) and R^2 (right).

The regressor chain method (Model 2) performed poorly in comparison, suggesting that the effectiveness of chaining methods depends on the specific case. If the initial model's error is high, it may continue to increase with each subsequent target prediction. RMSE provides a measure of the average error in forecasting the dependent variable. The comparable RMSE values between the training and test sets indicate the usefulness of the algorithm adaptation method (model3-MTL)mt-QSRR model. Algorithm adaptation methods have been particularly advantageous in scenarios where the tasks exhibit notable commonalities. They utilize an inductive transfer approach for enhancing generalization in machine learning by leveraging the domain-specific knowledge present in the training data of related tasks. Better performances of this method can be considered effective for simultaneous prediction of multiple retention times due to the regularization it enforces by demanding an algorithm to excel in correlated retention times with given five pHs, surpassing the regularization achieved by uniformly penalizing complexity to prevent overfitting. Significantly, the mt-QSRR model, which predicts multiple retention times simultaneously, demonstrates comparable performance to the single-target QSRR models, highlighting the significance of evaluating performance disparities (see 3.3). Additionally, the time needed for modelling consistently remained lower for mt-QSRR compared to predicting individual targets separately. In the single-target approach, each step had to be repeated multiple times based on the number of targets, whereas this repetition is unnecessary in the mt-QSRR modelling approach.

The newly introduced mt-QSRR model exhibits the potential to efficiently generate variations in retention time for diverse chemical compounds across multiple pH values. This offers the advantage of reduced effort and time.

3.3. COMPARISON OF THE MODELS

The comparison of the models based on their RMSE and R^2 are presented in Fig. 6. On the axis, the algorithms are plotted according to their average rank across analyses. Note that for each analysis, the best model is ranked 1 and the worse is ranked 3. The corresponding radar plots are presented in Fig. 7 as an alternative visualization of the ranks of the models for each analysis separately. In the radar plots, the lower the area in the coloured lines, the better. Overall, Figs. 6 and 7 show that Model 1 and Model 3 perform better than Model 2 based on the RMSE, and Model 3 performs best based on the R^2. Hence, we recommend Model 3 for similar analyses. We used the Friedman test to detect whether the differences in performances of mt-QSRR models are statistically significant. The Friedman test concluded that the difference in the performance of these algorithms is not statistically significant (p-value > 0.05).

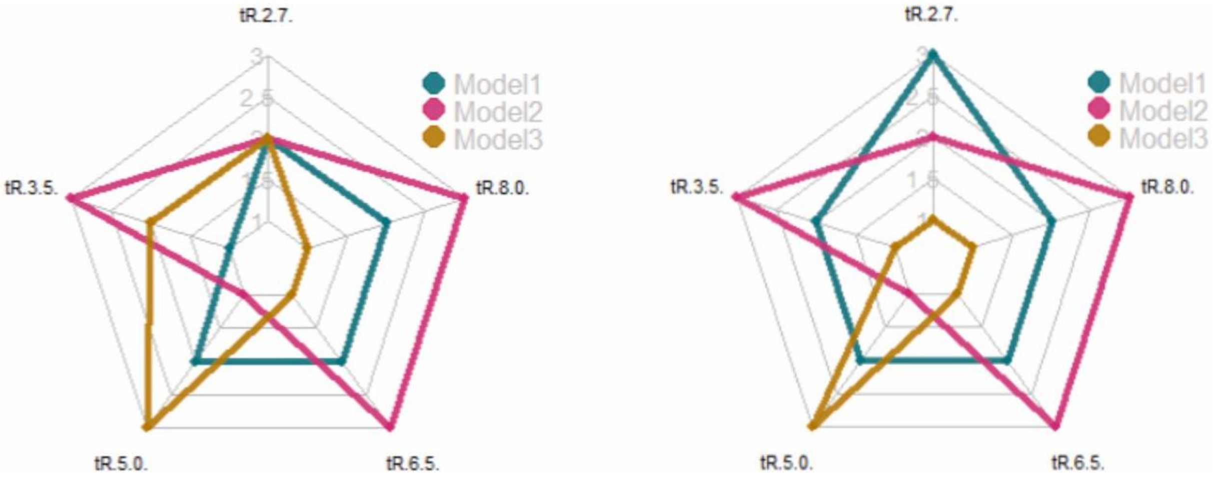


Fig. 7. Per-model rank based on the RMSE (left) and R^2 (right).

Conclusion

This study has successfully developed multiple multi-target QSRR (mt-QSRR) models involving a comparison between two modelling approaches: single-target and multi-target regression. The primary goal was to predict the retention times (tR) of a diverse range of structurally small molecules under various reversed-phase liquid chromatography (RPLC) conditions. The retention time prediction capabilities of the mt- QSRR model were assessed using three distinct methods. However, despite employing these diverse strategies, no statistically significant distinctions were observed in their predictive performance. The performance of the mt-QSRR models within our dataset indicated a reduction in efficiency as pH levels increased. Particularly, the regressor chain method exhibited higher root mean square error (RMSE), suggesting that retention prediction errors accumulate as they progress from lower to higher pH levels. One of the notable advantages of multi- target models is their interpretability in terms of the relationship between features and retention time variations with pH. Unlike single- target models, where descriptor importance varies per target specificity, the mt-QSRR model provides transparent insights into the pertinent input variables for predicting specific groups of response variables. Based on their performance, the optimal mt-QSRR model identified in this study highlighted five pivotal structural features: MolLogP, VSA- Estate5, LogD, SMR-VSA3, and QED. These descriptors encompass the molecular partition coefficient, molecular surface area, distribution coefficient state index, and drug-likeness. These attributes are crucial in accounting for the diverse retention times observed for the considered small molecules across varying pH levels. In summary, our findings underscore the potential of mt-QSRR models as a more effective and efficient predictive strategy compared to constructing separate models for each target. Adopting the mt-QSRR approach holds the promise of streamlining efforts and reducing time and computational costs while simultaneously assessing the effective separation of molecules within the RPLC setup. Lastly, it is imperative to acknowledge that the test set encompasses a limited number of molecules, leading to an incomplete representation of the explored chemical space. As a result, the outcomes presented in this study are preliminary in nature.

Funding

This research was funded by FWO/FNRS Belgium EOS-programme, grant number 30897864 "Chemical Information Mining in a Complex World".

CRediT authorship contribution statement

Priyanka Kumari: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing- original draft, Visualization. **Thomas Van Laethem**: Data curation, Writing- review & editing. **Diane Duroux**: writing- review & editing, Formal analysis. **Marianne Fillet**: Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition. **Phillipe hubert**: Conceptualization, Methodology, Resources, Writing- review & editing, Supervision, Funding acquisition. **Pierre-Yves Sacre**: Conceptualization, Methodology, Resources, Writing- review & editing, Supervision, Project administration. **Cedric Hubert**: Conceptualization, Methodology, Resources, Writing- review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jpba.2023.115690.

References

[1] F. Gritti, Anal. Chem. 93 (2021) 5653.

[2] G. Sagandykova, B. Buszewski, TrAC Trends Anal. Chem. 141 (2021), 116294.

[3] E.N. Muratov, J. Bajorath, R.P. Sheridan, I.V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I.I. Baskin, A. Varnek, A. Roitberg, et al., Chem. Soc. Rev. 49 (2020) 3525. [4] L. Zhao, W. Wang, A. Sedykh, H. Zhu, ACS Omega 2 (2017) 2805.

[5] R.I. Amos, E. Tyteca, M. Talebi, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl, J. Chem. Inf. Model. 57 (2017) 2754.

[6] C. Zisi, I. Sampsonidis, S. Fasoula, K. Papachristos, M. Witting, H.G. Gika, P. Nikitas, A. Pappa-Louisi, Metabolites 7 (2017) 7.

[7] B. Svrkota, J. Krmar, A. Proti´c, B. Otaˇsevi´c, The secret of reversed-phase/weak cation exchange retention mechanisms in mixed-mode liquid chromatography applied for small drug molecule analysis, J Chromatogr A 1690 (2023), 463776.

[8] K. Jovana, V. Milan, K. Ana, P. Ana, Z. Mira, O. Biljana, 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 1623 (2020), 461146.

[9] Z. Zhao, J. Qin, Z. Gou, Y. Zhang, Y. Yang, J. Biomed. Inform. 108 (2020), 103484.

[10] A. de la Vega de Leon, B. Chen, V.J. Gillet, J. Chemin. 10 (2018) ´ 1.

[11] E.B. Lenselink, P.F. Stouten, J. Comput. - Aided Mol. Des. 35 (2021) 901.

[12] B. Sharma, V. Chenthamarakshan, A. Dhurandhar, S. Pereira, J.A. Hendler, J. S. Dordick, P. Das, Sci. Rep. 13 (2023) 4908.

[13] S. Hamzic, R. Lewis, S. Desrayaud, C. Soylu, M. Fortunato, G. Gerebtzoff, R. Rodríguez-P´erez, J. Chem. Inf. Model. 62 (2022) 3180.

[14] K. Tiong, Z. Ma, and C.-W. Palmqvist, 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)(IEEE, 2022)793–798.

[15] D. Kocev, S. Dˇzeroski, M.D. White, G.R. Newell, P. Griffioen, Ecol. Model. 220 (2009) 1159.

[16] N. Basant, S. Gupta, Atmos. Environ. 177 (2018) 166.

[17] A.J. Burnham, J.F. MacGregor, R. Viveros, Chemom. Intell. Lab. Syst. 48 (1999) 167.

[18] H. Borchani, G. Varando, C. Bielza, P. Larranaga, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 5 (2015) 216.

[19] G.T.W.G. Spyromitros-Xioufis, Eleftherios, I. Vlahavas, Mach. Learn. 104 (2016) 55.

[20] Piccart, B. (2012). Algorithms for Multi-Target Learning (Algoritmes voor het leren van multi-target modellen).

[21] K. Muteki, J.E. Morgado, G.L. Reid, J. Wang, G. Xue, F.W. Riley, J.W. Harwood, D. T. Fortin, I.J. Miller, Ind. Eng. Chem. Res. 52 (2013), 12269.

[22] P.R. Haddad, M. Taraji, R. Szucs, Anal. Chem. 93 (2020) 228.

[23] M. Taraji, P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Anal. Chem. 89 (2017) 1870.

[24] K. Ciura, P. Kawczak, K.E. Greber, H. Kapica, J. Nowakowska, T. Baczek, J. Pharm. Biomed. Anal. 176 (2019), 112767.

[25] P. Kawczak, T. Baczek, Open Chem. 10 (2012) 570.

[26] M. Zapadka, M. Kaczmarek, B. Kupcewicz, P. Dekowski, A. Walkowiak, A. Kokotkiewicz, M. Łuczkiewicz, A. Bucinski, J. Pharm. Biomed. Anal. 164 (2019)´ 681.

[27] H. Linusson, Multi-output random forests, (2013).

[28] M. Breskvar, S. Dˇzeroski, IEEE Access 9 (2021), 10509.

[29] D. Kuznar, M. Mozina, and I. Bratko, Proceedings of the 1st workshop on learning from multi-label data (2009)61–68.

[30] Z. Han, Y. Liu, J. Zhao, W. Wang, Control Eng. Pract. 20 (2012) 1400.

[31] T. Van Laethem, P. Kumari, B. Boulanger, P. Hubert, M. Fillet, P.-Y. Sacr´e, C. Hubert, Molecules 27 (2022) 8306.

[32] P. Kumari, T. Van Laethem, P. Hubert, M. Fillet, P.-Y. Sacre, C. Hubert, Molecules´ 28 (2023) 1696.

[33] R. Kaliszan, Liquid Chromatography, Elsevier, 2017, pp. 553–572.

[34] R. Kaliszan, Chem. Rev. 107 (2007) 3212.

[35] G. Landrum, Release 1 (2013) 4.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., J. Mach. Learn. Res. 12 (2011) 2825.

[37] M. Friedman, J. Am. Stat. Assoc. 32 (1937) 675.

[38] D.G. Pereira, A. Afonso, F.M. Medeiros, Commun. Stat. - Simul. Comput. 44 (2015) 2636.