Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

TRANSFER LEARNING APPROACH TO MULTITARGET QSRR MODELING IN RPLC

Priyanka Kumari,[1,3] Madureira Sanches Ribeiro Guilherme[2], Pratyush Choudhary[2], Thomas Van Laethem[1,3], Marianne Fillet[3], Phillipe Hubert[1], Pierre Yves Sacre[1], and Cedric Hubert[1]*

1. *Department of Pharmacy, Laboratory of Pharmaceutical Analytical Chemistry, CIRM, Liège, Belgium 4000*
2. *Northwestern University, Chicago, Illinois 60208, United States*
3. *Department of Pharmacy, Laboratory for the Analysis of Medicines, CIRM, Liège, Belgium 4000*

Keywords:

Chromatography

Computational modeling

Layers

Molecular modeling

Molecules

Abstract

QSRR is a valuable technique for the retention time predictions of small molecules. This aims to bridge the gap between molecular structure and chromatographic behavior, offering invaluable insights for analytical chemistry. Given the challenge of simultaneous target prediction with variable experimental conditions and the scarcity of comprehensive data sets for such predictive modelings in chromatography, this study introduces a transfer learning-based multitarget QSRR approach to enhance retention time prediction. Through a comparative study of four models, both with and without the transfer learning approach, the performance of both single and multitarget QSRR was evaluated based on Mean Squared Error (MSE) and R2 metrics. Individual models were also tested for their performance against benchmark studies in this field. The findings suggest that transfer learning based multitarget models exhibit potential for enhanced accuracy in predicting retention times of small molecules, presenting a promising avenue for QSRR modeling. These models will be highly beneficial for optimizing experimental conditions in method development by better retention time predictions in Reversed-Phase Liquid Chromatography (RPLC). The reliable and effective predictive capabilities of these models make them valuable tools for pharmaceutical research and development endeavors.

Introduction

In the field of analytical chemistry, precise prediction of retention times is indispensable because it underpins the successful execution of various analytical methods and techniques, allowing researchers to obtain reliable and meaningful data. However, traditional experimental methods involve running multiple experiments under different conditions to obtain retention time data and, hence, can be cumbersome and expensive. Quantitative structure-retention relationship (QSRR) modeling, a

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

fundamental tool in chromatographic sciences, offers solutions to address these challenges. (1,2) QSRR models aim to correlate molecular descriptors with chromatographic retention times or factors, providing insights into analyte behavior under various chromatographic conditions. (1,3) They offer accurate and cost-effective alternatives to traditional experimental approaches by leveraging the relationship between a compound's molecular structure and its retention times. (1,4) Through these techniques, valuable insights can be gained into molecular behavior in chromatographic systems, advancing analytical chemistry across diverse fields. (5,6) In the past decade, significant advancements have been made in QSRR models, including statistical models, machine learning algorithms, and descriptor calculations.4? Multiple Linear Regression (MLR) (7,8) and Partial Least Squares (PLS) (9,10) can model linear relationships in the data. However, advanced machine learning (ML) methods, which are capable of handling nonlinear relationships and managing large and diverse data sets, have transformed QSRR modeling. These advanced ML methods include Random Forest (RF), Support Vector Regression (SVR) (both linear and nonlinear), Gradient Boosting Machines (GBM), and Artificial Neural Networks (ANN). (11−15) Beyond ML, Bayesian approaches have gained popularity in QSRR modeling, notably because of their handling of model uncertainty. Bayesian multilevel frameworks incorporate prior knowledge, explicitly quantifying uncertainty and offering a range of possible outcomes. (16,17) This probabilistic nature makes Bayesian models more flexible and generalizable across various chromatographic conditions and analytes compared to traditional QSRR models and ML techniques, which usually provide point estimates without inherent uncertainty measures. The ways of modeling also vary. Traditional methods, such as single-target retention prediction, often struggle with complex problems and situations, requiring a great deal of time and resources. This highlights the need for robust and flexible QSRR models that can predict retention times. While QSRR models are great for single-target predictions (one model for predicting retention time at one condition), they struggle with prediction of multiple retention times under a multitude of conditions at once. (18) Such models, which are also known as multitarget QSRR models, have not been fully explored in scientific studies. Multitarget QSRR has the potential to simultaneously correlate retention times of small molecules observed at varied experimental parameters such as variations in mobile phase compositions (pH, solvent, strength, buffer concentrations, etc.) and molecular descriptors (MDs) with the chromatographic behavior of molecules. (19,20) Researchers have been finding ways to get around this data scarcity. One method is data augmentation, where by artificially increasing the size and diversity of the data set, the models can capture a broader range of retention time variations. (21) While data augmentation increases the quantity of data, it does not necessarily improve the quality of the original data. If the original data set contains errors or biases, simply augmenting it may amplify these issues, and this can affect the accuracy of target predictions. Hence, along with this transfer learning could be a promising solution, (22) where knowledge from related areas is applied to fill in data gaps, making it possible to build more robust models even with scarce data. These innovative approaches open new doors for QSRR modeling, making it more versatile and effective in predicting retention times under various conditions.

TRANSFER LEARNING APPROACH

Transfer learning (TL) in deep learning consists of transferring the knowledge learned from a source domain Ds to a target domain Dt. (22,23) A domain can be defined as D = X,P(X)} where X is the feature

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

space, and P(X) represents the marginal distribution for X = [x1, x2, ..., xn] where xi represents a feature of X. If we learn a task Ts = {Y, f(.)} where Y denotes a label space and f(.) denotes a decision function, TL aims to improve the learning of a decision function in Dt for a different but related task Tt by using f(., θ). The nature of the difference between the domains or between the tasks can be used to categorize different transfer learning settings. The main known types of transfer learning include: Instance-based Transfer Learning that moves specific data points from a source to a similar target task, focusing on bridging marginal probability distribution differences. (24) Feature-based Transfer Learning, on the other hand, builds new feature representations to minimize domain differences while preserving local structure. (25) Relational Knowledge Transfer Learning targets the transfer of interdomain relationships and structures. (26,27) Lastly, parameter-based or model-based approaches are used in this study. This involves adapting entire models or model parameters from a source to a target task, leveraging prelearned knowledge for fine-tuning. (28)

SINGLE-TARGET AND MULTITARGET PREDICTION

In a feed-forward neural network, the primary role of the final layer is to synthesize the features extracted from preceding layers to produce the output. (29) This process can be mathematically represented as follows:

$$\hat{y} = h^L = \sigma(W^L \cdot h^{L-1} + b^L)$$

Here, L signifies the layer index, with W^L being the weight matrix that connects the units from layer L – 1 to layer L, and b^L represents the bias term for layer L. The function σ denotes the activation function, which could be ReLU, (30) LeakyReLU, (31) Tanh, (32) or any other suitable activation function. (33,34) For single-target prediction architectures, the output layer h^L consists of a single unit. This design implies that the network aims to predict a single response variable, such as the retention time of a compound, in QSRR modeling. The network's structure is optimized to focus on accurately predicting this singular outcome based on the input molecular descriptors.

Conversely, multitarget networks are designed with N units in the output layer, represented by a vector h^L of size N. This configuration allows the network to predict multiple response variables simultaneously. For example, in QSRR modeling, this could mean predicting the retention times of a compound under various experimental conditions. (19) Each unit in the output layer corresponds to a different target variable, enabling the network to capture and predict a broader spectrum of chromatographic behaviors based on the same set of input molecular descriptors.

Materials and Methods

In this research, we focused on pH variation as a key experimental parameter with the goal of predicting retention times at various pH levels using QSRR modeling. Within this context, models predicting the retention time of small molecules in reversed-phase liquid chromatography (RPLC) at a single pH level are defined as single-target prediction models. In contrast, models capable of predicting retention times across multiple pH levels are classified as multitarget prediction models. The

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

METLIN(SMRT) data set was downloaded from its Figshare repository. (35) The retention time of nearly all molecules falls within two distinct intervals: 0–2 min and 8–25 min. Molecules with low retention times (Retention time $\leqslant$2 min) were excluded from the SMRT data set, resulting in a total of approximately 77 thousand molecules. The other data sets (RIKEN (36,37)and LPAC (38)), which was our in-house data set, were used for testing purposes. This study did not generate any data. The LPAC data set, containing only 96 compounds, poses a data scarcity issue, making it challenging for any advanced QSRR modeling. However, this data set size is common in QSRR since the acquisition of reference chromatographic retention times requires long and costly experiments. Therefore, obtaining their retention time requires exploring approaches such as transfer learning. Techniques like transfer learning are specifically designed to address these challenges, allowing models to be effectively trained on smaller data using a pretrained model and, hence, improve their performance. The SMRT and Riken data sets had only one experimentally observed retention time; consequently, we employed it in our study exclusively for single-target prediction modeling. Conversely, the LPAC data sets offered five retention times to be predicted (at pH 2.0, pH 3.5, pH 5.0, pH 6.5, and pH 8.0);hence, this was used for multitarget modeling as well.

MOLECULAR DESCRIPTOR CALCULATION

Physicochemical descriptors were used to compare the two approaches (transfer learning and without transfer learning, single-target, and multitarget retention prediction approaches). Physicochemical descriptors were calculated (total 210) using the RDKit package, version 2015. (39)

MODEL ARCHITECTURE

In recent developments within the field of analytical chemistry, particularly in retention time prediction for small pharmaceutical compounds, our study has incorporated advanced deep learning techniques to enhance the accuracy and efficiency of compound separation processes in reversed-phase liquid chromatography (RPLC). By focusing on traditional molecular physicochemical data, we aim to enhance the prediction of retention times for small pharmaceutical compounds, employing a multilayer perceptron (MLP) to address this complex challenge.

The general workflow of the model architecture is shown in Figure 1. At the heart of our approach is an MLP consisting of four hidden layers with configurations of 1000, 500, 200, and 100 units, respectively. The adoption of the LeakyReLU activation function (eq 1) in each layer is a key feature, designed to prevent the issue of dying units commonly associated with the ReLU function. By allowing a small negative slope for negative inputs, LeakyReLU mitigates the vanishing gradient problem, facilitating more effective learning. To further enhance the model's ability to generalize, a dropout layer precedes the output layer, reducing the risk of overfitting.

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

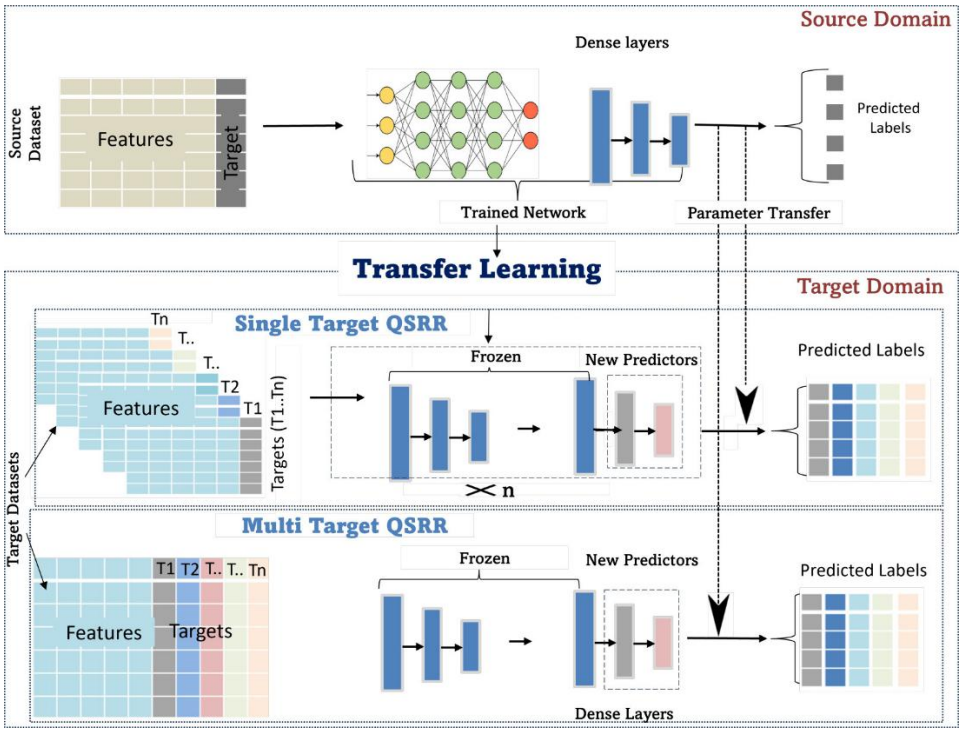

Figure 1. Architecture of QSRR modeling based on the Transfer Learning approach.

TRAINING AND FINE-TUNING

The MLP model was initially pretrained on the extensive SMRT data set. This foundational training phase was crucial for establishing a robust baseline from which the model could be fine-tuned to adapt to specific characteristics of smaller data sets. The fine-tuning process involved model selection based on mean squared error loss, adjusting the model for either single-target or multitarget prediction modes. In the single-target mode, the model treats each target variable independently, enhancing the specificity of the predictions. Conversely, in multitarget mode, all retention times are predicted simultaneously, offering a comprehensive view of the data's predictive landscape. In the single-target settings, the loss was computed individually for each target, and then the averaged loss was reported and used, whereas in the multitarget modeling, the loss was directly computed using the five targets simultaneously by measuring the squared L2 norm between each element in the input and target.

As illustrated in Figure 2, the Multi-Layer Perceptron (MLP) was initially trained using the SMRT data set, which is notably large. This size advantage allows for its division into training, validation, and testing subsets, allocating 80%, 10%, and 10% of the total data (randomly), respectively. Such distribution ratios are standard practice for data sets of substantial size, particularly when the model requires tuning of hyperparameters. The primary purpose of the training and validation sets is to facilitate model selection, which involves determining the optimal number of hidden layers, the number of neurons in each layer, and the dropout rate to prevent overfitting. (40,41) To assess the performance of the most effective model configuration, it underwent a retraining process. This process involved combining the training and validation sets for a comprehensive training phase, followed by an evaluation of the separate test set to measure its predictive accuracy. Subsequent to this initial training phase, the model underwent fine-tuning adjustments for application to smaller data sets.

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

Model selection also involved the intricate process of deciding which layers to freeze or unfreeze and setting the appropriate dropout rates. This decision-making process utilized leave-one-out cross-validation (LOOCV) on 80% of the data set to ensure the selection of the most effective model configuration. Ultimately, the performance of the model configuration that excelled in the LOOCV process was evaluated on a test set, comprising 20% of the original data set, to validate its effectiveness and generalization capability. In this study, five models were constructed, with their respective abbreviations detailed in Table 1.

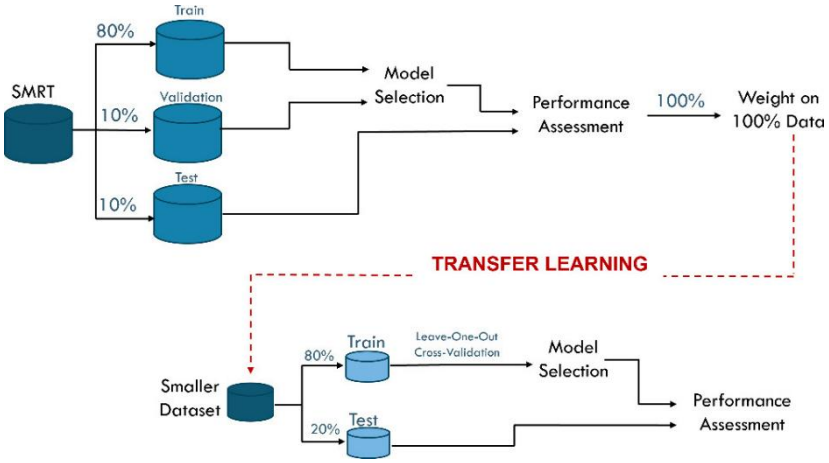

Figure 2. A simple schematic overview of model training using physicochemical descriptors.

Table 1. Summary of Model Abbreviations

Category	Abbreviation	Description
single-target models	M1_WTL	No TL
	M2_TL	With TL
multitarget models	M3_WTL	No TL
	M4_TL	With TL
models tested on SMRT data	M5_WTL	No TL
	M6_TL	With TL

Through this meticulous approach of transfer learning enhanced multitarget QSRR, this study seeks to enhance the efficiency of compound separation processes, thereby advancing the capabilities of RPLC methodologies

Evaluation Metrics.

$$\mathrm{MSE} = \sum_{i=0}^{N} \frac{(\hat{y}_i - y_i)^2}{N}$$

$$\mathrm{MAPE/MRE} = \frac{1}{N} \sum_{i=0}^{N} \frac{|y_i - \hat{y}_i|}{y_i}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

where yi and $\hat{y}\,i$ are the ground truth and the predicted value, respectively, for sample i, and N is the total number of samples.

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

MODEL INTERPRETATION WITH SHAP VALUES

Interpreting models is crucial for accurate predictions. Often, complex models, such as deep neural networks, provide better predictions but are difficult to interpret. In this study, to improve the interpretation of transfer-learned models, we have used SHAP values for the best performing models. SHAP values give each feature a score, indicating its importance for a particular prediction. If we consider f(x) the prediction made by a model given an input x and E[f(x)] is the expected value of the target variable, or in other words, the mean of all predictions (mean(model.predict(X))), then the sum of all SHAP values (one for each feature) must be equal to E[f(x)] – f(x) for a given observation. Thus it is possible to derive when computing the mean SHAP value for each feature on all observations how each feature impacts the model's predictions overall. The Python's shap package has been used to calculate the SHAP (SHapley Additive exPlanations) values (42) for every feature to plot the summary.

Results and Discussion

This study introduces a variety of strategies for QSRR modeling, facilitating the selection of approaches for predicting the retention time of new test molecules. These methods aim to enhance the accuracy and generalizability of QSRR models, particularly when dealing with data sets that are insufficient or include multiple targets to be predicted.

MODEL PERFORMANCES

This study utilized physicochemical descriptors as input to construct the DNN models to investigate their respective impacts on retention time predictions through four different strategies (Tables 1, 2). It is worth noting that physicochemical descriptors are widely employed in retention time prediction due to their ability to encapsulate comprehensive compound information. Multiple deep learning architectures such as 1D and 2D CNNs (43−45) and Graph Neural Networks (GNNs), including Graph Convolutional Networks (GCNs) and Relational Graph Convolutional Networks (RGCNs), have been frequently used in the recent past. (37,46) GNN models offer advanced capabilities for QSRR modeling by capturing the intricate molecular topology and features directly from graph representations of compounds. However, these models are associated with very high computational complexity and are resource-intensive. In comparison to physicochemical descriptors, GNN-based models would require high computational resources for training due to the complex operations on graph structures, especially for large molecular data sets and multiple targets. Additionally, the preprocessing of molecules into graph representations and the tuning of network parameters for optimal performance can be more complex and time-consuming.

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

Table 2. Model Performances[a]

	LPAC		
Models	**MSE**	R^2	**Time (Min)**
single-target Models			
M1_phys_WTL	59.08	–0.35	0.14
M2_phys_TL	16.19	0.64	0.13
multitarget Models			
M3_Phys_WTL	60.83	–0.38	0.05
M4_Phys_TL	15.15	0.66	0.09

[a] Model abbreviations are elaborated in Table 1.

In our investigation, we assessed the flexibility of DNN model architectures based on physicochemical features, tailoring them to single- and multitarget prediction tasks with comparative analyses by employing transfer learning approaches in situations of scarce data availability. The hyperparameters and the model's layers were fine-tuned based on the scores given on the leave-one-out cross-validation (Supplementary Table S1). Our analysis delineated distinct performance trajectories for each modeling approach (Table 2).

For the LPAC data set, Model M4 (multitarget with Transfer Learning) demonstrated the best performance in terms of accuracy, as indicated by the lowest MSE (15.15) and the highest R2 value (0.66) among all models. The implementation of TL resulted in significant accuracy improvements: a decrease in MSE by 42.89 min (from 59.08 to 16.19) and 45.68 min (from 60.83 to 15.15) and an increase in R2 from −0.35 to 0.64 and from −0.38 to 0.66 for single-target and multitarget models, respectively. The models were also compared with the performance of a conventional ML model like SVM (Supplementary Table S3). The MSE and R2 of the SVM model came out to be 48.45 and 0.14, respectively, which are worse than all transfer learned models. This suggests that the application of Transfer Learning significantly enhanced the model's predictive accuracy and its ability to explain the variance in the data set. A decrease of MSE from 16.19 to 15.15 min from M2 (Single-target with Transfer Learning) to M4 (multitarget with Transfer Learning) emphasizes the suitability of the multitarget approach of QSRR, which inherently handles more complex prediction tasks by predicting multiple outputs simultaneously. Overall, the Multitarget models benefit significantly from the transfer learning approach, resulting in a lower MSE and higher R2 values, showcasing the superiority of transfer learning in these cases.

On comparison of M1 with M3 and M2 with M4, it can be clearly seen that the multitarget model performs better than single-target settings. Predicted versus observed retention times for each case are plotted (Figure 3). It can be clearly observed that the transfer learning approach has provided the predictions for targets especially at a higher pH. In Figure 3, Models M2 and M4, which incorporate Transfer Learning given their closer alignment with the identity line, indicate a higher prediction accuracy compared to M1 and M3. Model M1 exhibits the greatest deviation from the ideal, with

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

points scattered far from the line, indicating lower predictive accuracy. Model M3, while better than M1, still shows substantial deviation. After comparing the best performing model (M4) for every target, it can be seen that Target 5 (retention time at pH 8.0) has less scattered points and closeness to the identity line.

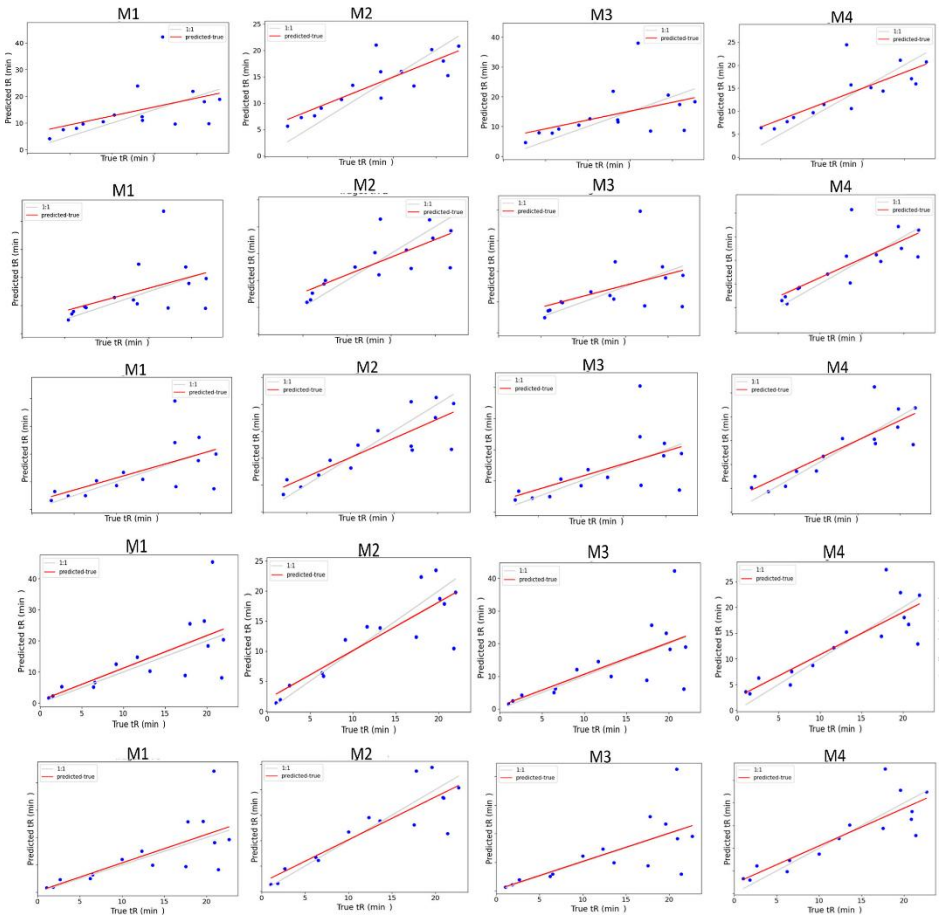

Figure 3. Predicted vs observed retention time (in minutes) for M1 to M4 for the LPAC data set. Rows represent targets: Row 1: Target tR 1 (retention times at pH 2.7), row 2: Target tR 2 (retention times at pH 3.5), row 3: Target tR 3 (retention times at pH 5.0), row 4: Target tR 4 (retention times at pH 6.5), and row 5: Target tR 5 (retention times at pH 8.0). The X-axis represents observed retention time, and the Y-axis represents predicted retention time in minutes. (Red line–fit line and gray line–identity line).

The model's performances were compared on two groups of compounds: Group one, with compounds having the same ionization state (and therefore lipophilicity) throughout all pH levels, and Group two, with compounds whose ionization state (and therefore their lipophilicity) changes over pH. This comparison was conducted using a test set to observe the effect on compounds with lipophilicity changing over pH and those having stable lipophilicity (Table S2). The proposed model's prediction errors do not vary significantly among the two groups (Figure S1). The outlier (Veramapil) in the second group requires further investigation to understand its high variability, but this behavior does not seem to be related to its lipophilicity change over pH.

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

TIME COMPARISON

As evidenced by Table 2, the computational time was marginally affected between M1 to M2 and M3 to M4, demonstrating that TL's advantages in model accuracy do not substantially impact modeling efficiency. When comparing single-target to multitarget models, multitarget models provide better time savings, highlighted by quicker execution times over single-target modeling which is 0.05 min by M3. This analysis underscores the benefits of applying TL in enhancing model performance without compromising on time efficiency and suggests a balanced consideration between single-target and multitarget approaches based on data set characteristics and computational constraints.

PERFORMANCE COMPARISON ON TEST DATA WITH BENCHMARK STUDIES

In our study, we compared the performance of our models against established benchmarks. (37,47−49) The comparison was conducted using the RIKEN data set, which contains 851 data points. We set aside 20% (170 data points) for testing and used the remaining 80% (681 data points) for training and validation. The results, as highlighted in Table 3 and Figure 4, show that the M5_WTL model achieves impressive results across both examined data sets. Specifically, for the SMRT data set, M5_WTL achieves a Mean Relative Error (MRE) of 0.07 and an R2 score of 0.78 which is similar to other better performing models in the list, while for the RIKEN data set, it records a comparable MRE of 0.14 and an R2 of 0.75. These figures demonstrate that M5 is not only comparable to advanced models, but in some cases, such as with Random Forest (RF) and Gradient Boosting (GB), it even outperforms them. Similarly, the M6_TL model showcases notable performance on Riken data sets with an MRE of 0.14 with a good R2 of 0.85, positioning it superiorly in comparison to various other models. These values mark the M6 model, which utilizes a Transfer Learning approach, as a standout, particularly for the RIKEN data set.

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

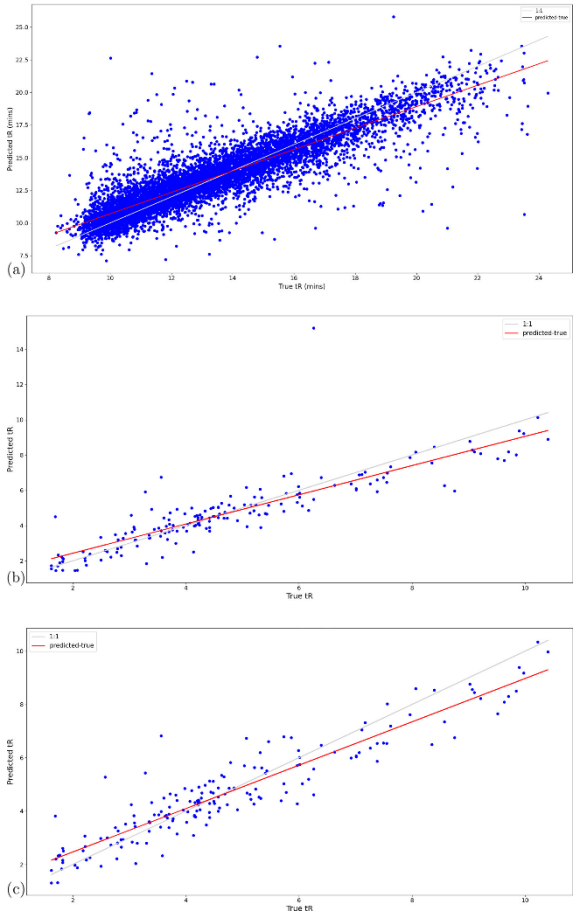

Figure 4. Plot for Predicted vs Observed retention time (min) for Models M5 and M6 for the (a) SMRT data set (M5/M6), (b) Riken Data set by M5(WTL), and (c) Riken Data set by M6(TL). (Red line–fit line and gray line–identity line).

Table 3. Comparison of Model Performances with Benchmarks[a]

	SMRT		**RIKEN**	
Models	**MRE**	R^2	**MRE**	R^2
Pretraining (SMRT)-Proposed model	0.07	0.78	-	-
GCN (37)	0.04	0.89	0.14	0.76
RGCN (37)	0.04	0.89	0.14	0.79
MLP (37)	0.05	0.84	0.10	0.56
RF (37)	0.07	0.78	0.19	0.69
SVM (37)	0.06	0.82	0.18	0.76
AB (37)	0.07	0.76	0.19	0.68
GB37	0.15	0.40	0.19	0.70

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

	SMRT		RIKEN	
Models	**MRE**	R^2	**MRE**	R^2
MPNN (47)	-	0.87	-	0.59
GIN (49)	-	0.87	-	0.48
GNN-RT (50)	0.05	0.85		
M5-WTL	-	-	0.14	0.77
M6-TL	-	-	0.14	0.85

[a] Units from references have been converted from seconds to minutes.

When comparing M5 and M6 (Figure 4 (b, c)) for the RIKEN data set–where M6 utilizes Transfer Learning while M5 does not, analysis of the coefficient of determination (R2) indicates that M6_TL outperforms M5_WTL. This comparison highlights the efficacy of Transfer Learning in improving model performance.

Overall, models M5 and M6 exhibit strong and comparative predictive performance when compared with benchmark models. Their low MRE values indicate their ability to make accurate predictions, while the high R2 scores demonstrate their efficacy in explaining the variance in the data. M6, in particular, stands out with its remarkable R2 score of 0.85 on the RIKEN data set, surpassing the performance of many other models. This implies that transfer learning holds great promise for applications in the field of analytical chemistry, potentially outperforming established models and providing valuable insights. Further investigations and real-world applications of these models are certainly needed.

MODEL INTERPRETATION BASED ON SHAP SUMMARY PLOTS

Understanding the feature importance is critical for optimizing RPLC methods and can provide insights into the molecular characteristics that are most influential under different chromatographic conditions. This knowledge is valuable for method development in RPLC, allowing for a better prediction of retention times and more efficient separations. SHAP values are crucial in this analysis. Summary SHAP plots in supplementary Figures 2-6 and corresponding selected features and their rankings (Table 4) illustrate the interpretation of the transfer learnt multitarget QSRR models for every target (Model 4). It presents the importance of the top 20 molecular descriptors in terms of average impact on model output (the effects on predicted retention time). Lower SHAP values indicate a lower effect of the descriptor, while higher SHAP values indicate high effects of the descriptor. From the plots, it can be observed that specific features, like fr_methoxy, IpC, fr_ether, HallkierAlpha, and TPSA, remain consistently important across all pH levels. These features play a fundamental role in the retention mechanism in liquid chromatography, regardless of the pH. However, the study also identifies features like SLogP_VSA,MolLogP, NHOHCount, PMI3, and NOCount whose importance

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

varies with pH levels. This variability suggests that certain molecular interactions, such as ionization and lipophilicity, may be more relevant under specific conditions. For instance, TPSA, where the ionization state of the molecule is affected by the pH, which would directly influence the molecule's interaction with the aqueous phase and MolLogP, as a measure of lipophilicity, might be more influential at pH levels where the analyte's lipophilic components are less ionized and more likely to interact with the hydrophobic stationary phase. The study also notes varying trends in the importance of features like SlogP_VSA, PEOE_VSA, and RadiusOfGyration in the LPAC data set, indicating that the solute's physicochemical properties, such as lipophilicity, hydrogen bonding capability, and molecular size, differently affect retention times at varying pH levels.

Table 4. Summary of SHAP Values for M4 (Top 10 Features)[a]

Rank	T1	T2	T3	T4	T5
1	fr_Methoxy	f_Methoxy	fr_Methoxy	fr_methoxy	fr_Methoxy
2	Ipc	Ipc	Ipc	ipc	ipc
3	fr_ether	fr_ether	fr_ether	fr_ether	fr_ether
4	HallKierAlpha	HallKierAlpha	HallKierAlpha	TPSA	HallKierAlpha
5	TPSA	TPSA	TPSA	HallKierAlpha	TPSA
6	SlogP_VSA5	SlogP_VSA5	Chi1n	SlogP_VSA11	SlogP_VSA11
7	fr_C_O_noCOO	fr_C_O_noCOO	SlogP_VSA11	Chi1n	Chi1n
8	Chi1n	Chi1n	MolLogP	InertialShapeFactor	InertialShapeFactor
9	MolLogP	MolLogP	SlogP_VSA5	MolLogP	MolLogP
10	PMI3	PMI3	InertialShapeFactor	SlogP_VSA5	PMI3
11	InertialShapeFactor	InertialShapeFactor	fr_C_O_noCOO	NHOHCount	SlogP_VSA5
12	RadiusofGyration	NHOHCount	PMI3	NumAliphaticCarbocycles	PMI3
13	NHOHCount	NumAliphaticCarbocycles	NHOHCount	NumHDonors	NumHDonors
14	NumDonors	SlogP_VSA11	NumHDonors	PMI3	NumAliphaticCarbocycles

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

Rank	T1	T2	T3	T4	T5
15	SlogP_VSA11	NumHDonors	RadiusofGyration	fr_C_O_noCOO	fr_C_O_noCOO
16	NumAliphatic Carbocycles	RadiusofGyration	NumAliphatic Carbocycles	RadiusofGyration	RadiusofGyration
17	PEOE_VSA2	PEOE_VSA1	PEOE_VSA1	PEOE_VSA1	PEOE_VSA1
18	PMI2	PEOE_VSA2	PMI2	fr_NH2	fr_NH2
19	PEOE_VSA1	PMI2	fr_NH2	PMI2	PMI2
20	Chi4n	fr_NH2	PEOE_VSA2	NOCount	NOCount

[a] T1, T2, T3, T4, and T5 are retention times at pH 2.7, 3.5, 5.0, 6.5, and 8.0, respectively.

An important point to note here is that these findings, derived from SHAP summary plots, offer general insights into the factors influencing RPLC retention times differentially with varying targets. A more detailed chemical analysis and domain-specific expertise would be required for precise interpretations, which fall beyond the scope of this study.

Conclusion

In conclusion, this study provides valuable insights into the field of retention time prediction modeling for analytical chemistry. We explored the application of different strategies, including the utilization of physicochemical descriptors and the power of deep learning and transfer learning in single-target and multitarget settings, to enhance the accuracy and generalizability of QSRR models. Our analysis was conducted on our in-house data set utilizing four different models. One of the key findings of this study is the significance of transfer learning in the context of QSRR modeling. It was observed that the application of transfer learning consistently improved the performance of QSRR models, resulting in a lower Mean Squared Error (MSE) and higher coefficient of determination (R2) values. For analytical chemists, working on multitarget retention time prediction settings can be a better approach that can provide insight about the molecule's interplay at varying targets. Such models can enhance model performance while reducing the training time. Moreover, our study showed a good and comparable performance of models with other benchmark studies in the field and demonstrated a strong predictive performance with the Transfer Learning approach, in particular, outperforming many other classical ML based models, suggesting its potential for applications in data-driven tasks in analytical chemistry.

This study also highlights the significance of understanding molecular features in QSRR modeling for RPLC, offering crucial insights in terms of SHAP values for optimizing these models. It emphasizes the importance of certain features that maintain their significance across different pH levels while also pointing out how the relevance of other features can vary under diverse conditions. This highlights the complex relationship between the molecular characteristics and their chromatography responses,

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

suggesting the need for advanced analytical tools and specialized knowledge to develop more accurate and efficient QSRR models. Furthermore, the model performances underline the importance of aligning QSRR modeling strategies with the specific objectives and the characteristics of the data set, such as the availability of molecules for training and testing.

By increasing our understanding of chromatographic processes and supporting the search for new QSRR modeling techniques, this research aims to improve the predictability and operational efficiency of method development in RPLC.

Data Availability

The references are cited in the section 'Materials and Methods' for data accessibility, and the codes are available at https://github.com/pkc533/Transfer-Learning_MTQSRR.git.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00608.

Author Contributions

Priyanka Kumari: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing- original draft, Visualization. Madureira Sanches Ribeiro Guilherme: Formal analysis, Writing- review and editing, Visualization. Pratyush Choudhary: Writing- review and editing, Formal analysis. Thomas Van Laethem: Data curation, Writing- review and editing. Marianne Fillet: Conceptualization, Methodology, Resources, Writing - review and editing, Supervision, Funding acquisition. Phillipe Hubert: Conceptualization, Methodology, Resources, Writing- review and editing, Supervision, Funding acquisition. Pierre-Yves Sacré: Conceptualization, Methodology, Resources, Writing- review and editing, Supervision, Project administration. Cédric Hubert: Conceptualization, Methodology, Resources, Writing- review and editing, Supervision, Project administration, Funding acquisition.

Published in : Journal (année), vol. xx, n°xx, pp. xx–xx
DOI: 10.xxxxxxxxxx
Status : Postprint (Author's version)

Funding

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

Notes

The authors declare no competing financial interest.

References

1. 1Kaliszan, R. QSRR: quantitative structure-(chromatographic) retention relationships. Chem. Rev. 2007, 107, 3212– 3246, DOI: 10.1021/cr068412z
2. 2Taraji, M.; Haddad, P. R.; Amos, R. I.; Talebi, M.; Szucs, R.; Dolan, J. W.; Pohl, C. A. Chemometric-assisted method development in hydrophilic interaction liquid chromatography: A review. Analytica chimica acta 2018, 1000, 20– 40, DOI: 10.1016/j.aca.2017.09.041
3. 3Kaliszan, R. Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography. J. Chromatogr. A 1993, 656, 417– 435, DOI: 10.1016/0021-9673(93)80812-M
4. 4Kaliszan, R. Liquid Chromatography; Elsevier, 2017; pp 553– 572.
5. 5Gritti, F. Perspective on the future approaches to predict retention in liquid chromatography. Anal. Chem. 2021, 93, 5653– 5664, DOI: 10.1021/acs.analchem.0c05078
6. 6Rojas, C.; Duchowicz, P. R.; Tripaldi, P.; Diez, R. P. Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. Journal of Chromatography A 2015, 1422, 277– 288, DOI: 10.1016/j.chroma.2015.10.028
7. 7Hancock, 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, 185– 196, DOI: 10.1016/j.chemolab.2004.11.001
8. 8Ciura, K. Modeling of small molecule's affinity to phospholipids using IAM-HPLC and QSRR approach enhanced by similarity-based machine algorithms. Journal of Chromatography A 2024, 1714, 464549, DOI: 10.1016/j.chroma.2023.464549
9. 9Put, R.; Daszykowski, M.; Baczek, T.; Vander Heyden, Y. Retention prediction of peptides based on uninformative variable elimination by partial least squares. J. Proteome Res. 2006, 5, 1618– 1625, DOI: 10.1021/pr0600430
10. 10Ukić, Š.; Novak, M.; Žuvela, P.; Avdalović, N.; Liu, Y.; Buszewski, B.; Bolanča, T. Development of gradient retention model in ion chromatography. Part I: conventional QSRR approach. Chromatographia 2014, 77, 985– 996, DOI: 10.1007/s10337-014-2653-5

11. 11Ciura, K.; Kovačević, S.; Pastewska, M.; Kapica, H.; Kornela, M.; Sawicki, W. Prediction of the chromatographic hydrophobicity index with immobilized artificial membrane chromatography using simple molecular descriptors and artificial neural networks. Journal of Chromatography A 2021, 1660, 462666, DOI: 10.1016/j.chroma.2021.462666
12. 12Bouwmeester, R.; Martens, L.; Degroeve, S. Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction. Anal. Chem. 2019, 91, 3694– 3703, DOI: 10.1021/acs.analchem.8b05820
13. 13Goudarzi, 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 polybrominated diphenylethers (PBDEs). Analytical Methods 2012, 4, 3733– 3738, DOI: 10.1039/c2ay25484k
14. 14Zhang, J.; Zheng, C.-H.; Xia, Y.; Wang, B.; Chen, P. Optimization enhanced genetic algorithm-support vector regression for the prediction of compound retention indices in gas chromatography. Neurocomputing 2017, 240, 183– 190, DOI: 10.1016/j.neucom.2016.11.070
15. 15Sun, M.-x.; Li, X.-h.; Jiang, M.-t.; Zhang, L.; Ding, M.-x.; Zou, Y.-d.; Gao, X.-m.; Yang, W.-z.; Guo, D.-a.others A practical strategy enabling more reliable identification of ginsenosides from Panax quinquefolius flower by dimension-enhanced liquid chromatography/mass spectrometry and quantitative structure-retention relationship-based retention behavior prediction. Journal of Chromatography A 2023, 1706, 464243, DOI: 10.1016/j.chroma.2023.464243
16. 16Wiczling, P.; Kamedulska, A. Comparison of Chromatographic Stationary Phases Using a Bayesian-Based Multilevel Model. Anal. Chem. 2024, 96, 1310– 1319, DOI: 10.1021/acs.analchem.3c04697
17. 17Wiczling, P.; Kamedulska, A.; Kubik, Ł. Application of Bayesian Multilevel Modeling in the Quantitative Structure–Retention Relationship Studies of Heterogeneous Compounds. Anal. Chem. 2021, 93, 6961– 6971, DOI: 10.1021/acs.analchem.0c05227
18. 18Mazraedoost, S.; Žuvela, P.; Ulenberg, S.; Baczek, T.; Liu, J. J. Cross-column density functional theory–based quantitative structure-retention relationship model development powered by machine learning. Anal. Bioanal. Chem. 2024, 416, 2951– 2968, DOI: 10.1007/s00216-024-05243-7
19. 19Kumari, P.; Duroux, D.; Fillet, M.; Sacre, P. Y.; Hubert, C.others A multi-target QSRR approach to model retention times of small molecules in RPLC. J. Pharm. Biomed. Anal. 2023, 236, 115690, DOI: 10.1016/j.jpba.2023.115690
20. 20Svrkota, B.; Krmar, J.; Protić, A.; Otašević, B. The secret of reversed-phase/weak cation exchange retention mechanisms in mixed-mode liquid chromatography applied for small drug molecule analysis. Journal of Chromatography A 2023, 1690, 463776, DOI: 10.1016/j.chroma.2023.463776
21. 21Magar, R.; Wang, Y.; Lorsung, C.; Liang, C.; Ramasubramanian, H.; Li, P.; Farimani, A. B. AugLiChem: data augmentation library of chemical structures for machine learning. Machine Learning: Science and Technology 2022, 3, 045015, DOI: 10.1088/2632-2153/ac9c84
22. 22Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. Journal of Big data 2016, 3, 9, DOI: 10.1186/s40537-016-0043-6

23. 23Pan, S. J.; Yang, Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering 2010, 22, 1345– 1359, DOI: 10.1109/TKDE.2009.191
24. 24Bursac, P.Instance-based transfer learning for soil organic carbon estimation. Front. Environ. Sci. 2022, 10, 1003918, DOI: 10.3389/fenvs.2022.1003918
25. 25Sevani, N.A Feature-based Transfer Learning to Improve the Image Classification with Support Vector Machine. International Journal of Advanced Computer Science and Applications 2023, DOI: 10.14569/IJACSA.2023.0140632
26. 26Mooney, R. J. Transfer Learning by Mapping and Revising Relational Knowledge. Advances in Artificial Intelligence - SBIA 2008. Berlin, Heidelberg, 2008; pp 2– 3.
27. 27Wang, D.; Li, Y.; Lin, Y.; Zhuang, Y. Relational knowledge transfer for zero-shot learning. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence . 2016; pp 2145– 2151.
28. 28Mudrakarta, P. K.; Sandler, M.; Zhmoginov, A.; Howard, A. K for the price of 1. Parameter efficient multi-task and transfer learning. International Conference on Learning Representations . 2019.
29. 29LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. nature 2015, 521, 436– 444, DOI: 10.1038/nature14539
30. 30Agarap, A. F. Deep learning using rectified linear units (relu). arXiv preprint . arXiv:1803.08375. 2018.
31. 31Maniatopoulos, A.; Mitianoudis, N. Learnable leaky relu (LeLeLU): An alternative accuracy-optimized activation function. Information 2021, 12, 513, DOI: 10.3390/info12120513
32. 32Lau, M. M.; Lim, K. H. Review of adaptive activation function in deep neural network. 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES) . 2018; pp 686– 690.
33. 33Sharma, S.; Sharma, S.; Athaiya, A. Activation functions in neural networks. IJEAST 2020, 04, 310– 316, DOI: 10.33564/IJEAST.2020.v04i12.054
34. 34Dubey, S. R.; Singh, S. K.; Chaudhuri, B. B. Activation functions in deep learning: A comprehensive survey and benchmark. Neurocomputing 2022, 503, 92, DOI: 10.1016/j.neucom.2022.06.111
35. 35Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. The METLIN small molecule dataset for machine learning-based retention time prediction. Nat. Commun. 2019, 10, 5811, DOI: 10.1038/s41467-019-13680-7
36. 36Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D. K.; Fiehn, O. Retip: retention time prediction for compound annotation in untargeted metabolomics. Analytical chemistry 2020, 92, 7515– 7522, DOI: 10.1021/acs.analchem.9b05765
37. 37Kensert, A.; Bouwmeester, R.; Efthymiadis, K.; Van Broeck, P.; Desmet, G.; Cabooter, D. Graph convolutional networks for improved prediction and interpretability of chromatographic retention data. Anal. Chem. 2021, 93, 15633– 15641, DOI: 10.1021/acs.analchem.1c02988
38. 38Van Laethem, T.; Kumari, P.; Hubert, P.; Fillet, M.; Sacré, P.-Y.; Hubert, C. A pharmaceutical-related molecules dataset for reversed-phase chromatography retention time prediction built on combining pH and gradient time conditions. Data in Brief 2022, 42, 108017, DOI: 10.1016/j.dib.2022.108017
39. 39Landrum, G. Rdkit documentation, Release 2013; pp 1– 4.

40. 40Prechelt, L. Neural Networks: Tricks of the Trade. Springer, 2002; pp 55– 69.
41. 41Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 2014, 15, 1929– 1958
42. 42Lundberg, S. M.; Lee, S.-I. In Advances in Neural Information Processing Systems 30; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765– 4774.
43. 43Fedorova, E. S.; Matyushin, D. D.; Plyushchenko, I. V.; Stavrianidi, A. N.; Buryak, A. K. Deep learning for retention time prediction in reversed-phase liquid chromatography. Journal of Chromatography A 2022, 1664, 462792, DOI: 10.1016/j.chroma.2021.462792
44. 44Matyushin, D. D.; Buryak, A. K. Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning. IEEE Access 2020, 8, 223140– 223155, DOI: 10.1109/ACCESS.2020.3045047
45. 45Zhong, S.; Hu, J.; Yu, X.; Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. Chemical Engineering Journal 2021, 408, 127998, DOI: 10.1016/j.cej.2020.127998
46. 46Yang, Q.; Ji, H.; Fan, X.; Zhang, Z.; Lu, H. Retention time prediction in hydrophilic interaction liquid chromatography with graph neural network and transfer learning. Journal of Chromatography A 2021, 1656, 462536, DOI: 10.1016/j.chroma.2021.462536
47. 47Osipenko, S.; Nikolaev, E.; Kostyukevich, Y. Retention time prediction with message-passing neural networks. Separations 2022, 9, 291, DOI: 10.3390/separations9100291
48. 48Fedorova, E. S.; Matyushin, D. D.; Plyushchenko, I. V.; Stavrianidi, A. N.; Buryak, A. K. Deep learning for retention time prediction in reversed-phase liquid chromatography. Journal of Chromatography A 2022, 1664, 462792, DOI: 10.1016/j.chroma.2021.462792
49. 49Kwon, Y.; Kwon, H.; Han, J.; Kang, M.; Kim, J.-Y.; Shin, D.; Choi, Y.-S.; Kang, S. Retention Time Prediction through Learning from a Small Training Data Set with a Pretrained Graph Neural Network. Anal. Chem. 2023, 95, 17273– 17283, DOI: 10.1021/acs.analchem.3c03177
50. 50Yang, Q.; Ji, H.; Lu, H.; Zhang, Z. Prediction of liquid chromatographic retention time with graph neural networks to assist in small molecule identification. Anal. Chem. 2021, 93, 2200–2206, DOI: 10.1021/acs.analchem.0c04071