

**MACHINE LEARNING MODEL FOR PREDICTING SHIFTS IN GAS
CHROMATOGRAPHY RETENTION TIMES CAUSED BY THE
CHANGE OF EXPERIMENT'S PARAMETERS AND CONDITIONS**

BY

Hussein Al-Adwan

A Thesis Presented to the
COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

Materials Science and Engineering

May 2025

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES

This thesis, written by **Hussein AL-Adwan** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **Master of Science in Materials Science and Engineering**.

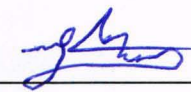


Dr. Nayef Mesnad Alsaifi
Department Chairman

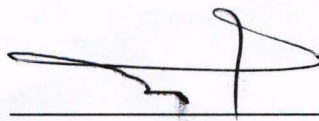


Dr. Suliman S. Al-Homidan
Dean of Graduate Studies

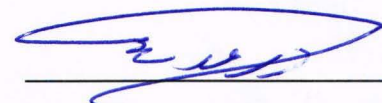
Date:



Dr. Abduljabar Alsayoud
(Advisor)



Dr. Mohammed F. Alkhater
(Member)



Dr. Qasem Drmash
(Member)



© Hussein AL-Adwan

2025

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to **King Fahd University of Petroleum and Minerals (KFUPM)** for providing the facilities, resources, and supportive research environment that made this work possible.

I am particularly thankful to the **Department of Materials Science and Engineering** for their guidance and encouragement throughout my graduate studies.

My sincerest thanks go to my thesis advisor, **Dr. Abduljabar Al-Sayoud**, for his invaluable mentorship, insightful discussions, and unwavering support.

I am also deeply appreciative of my committee members, **Dr. Qasem Drmosh** and **Dr. Mohammed F. Alkhater**, for their thoughtful feedback, constructive criticism, and for taking the time to review and improve this work.

Table of Contents

ABSTRACT	ix
ملخص الرسالة	Error! Bookmark not defined.
Chapter 1	1
Introduction	1
1.1 Application Background: Gas Chromatography	2
1.1.1 Components Detection.....	3
1.1.2 Chromatograms and Calibration	4
1.2 Methodology Background: Machine Learning	7
1.2.1 Supervised Machine Learning	8
1.2.2 Unsupervised Machine Learning: Isolation Forest	12
1.2.3 Preprocessing and Postprocessing Techniques	12
1.3 Problem Definition and Solution Methodology	15
1.3.1 Problem Definition.....	15
1.3.2 Methodology	15
Chapter 2	19
Literature Review	19
2.1 Liquid Chromatography Studies	19
2.2 Multidimensional Chromatography	20
2.3 Gas Chromatography Studies	21
Chapter 3	24
Results and Discussion	24
3.1 Data Cleaning	24
3.2 Monte Carlo Cross Validation	27
3.3 Correlation Results	29
3.4 Final Models:	32
3.5 Performance and Scope Comparison with Models in Literature:	36
3.6 Using the Model for GC Optimization of Xylene Isomers Separation	37
3.7 QLSR Evaluation and Formula	39

List of Figures

Figure 1.1: Gas Chromatography Scheme.	2
Figure 1.2: Ideal GC Chromatogram.....	4
Figure 1 3: Retention Time Shifts in 2 GC experiments.....	6
Figure 1 4: KNN Algorithm	9
Figure 1 5: Illustration of a Simple DT Algorithm	10
Figure 1.6: Reference Chromatogram	18
Figure 3.1: Samples Distribution over: a) Raw data, b) Cleaned Data.....	25
Figure 3.2: Monte Carlo Validation R2 Convergence Test.....	27
Figure 3.3: Correlation Results between the Features and the Target: a) Absolute Pearson Correlation, b) Distance Correlation.	30
Figure 3 4: Pearson Correlation Matrix.	31
Figure 3.5: Predicted vs. Actual RTshifts for the Validation and Testing Sets: a) KNN, b) RF, c) GBDT, d) ABDT, e) XGB, and f) ABSVR.....	33
Figure 3 6: Validation and Testing Residuals of the ABSVR Algorithm.....	35

List of Tables

Table 1.1: The Model Variables with The Reference Chromatogram Values.	18
Table 3.1: Data Statistics of the Cleaned Data.....	26
Table 3.2: Average Monte Carlo Evaluation Metrics.....	28
Table 3.3: Algorithms' Training, Testing, and Validation Metrics.....	34
Table 3.4: Comparison of Different Proposed RT Prediction Models of Different Chromatographic Methods.	37

List of Abbreviations:

GC-MS: Gas Chromatography-Mass Spectroscopy.

RT: Retention Time.

QSRR: Quantitative Structure Retention Relationships.

SMILES: Simplified Molecular Line Entry System Strings.

GRNN: Generalized Regression Neural Networks.

ANN: Artificial Neural Networks.

MAE: Mean Absolute Error.

GNN: Graph Neural Network.

RMSE: Root Mean Squared Error.

RF: Random Forest.

GBDT: Gradient Boosting Decision Tree.

ABDT: Adaptive Boosting Decision Tree.

XGB: eXtreme Gradient Boosting.

KNN: K-Nearest Neighbors.

ABSVR: Adaptive Boosting Support Vector Regression.

QLSR: QLattice Symbolic Regression.

ABSTRACT

Abstract:

Gas Chromatography (GC) is a versatile separation technique that is widely used in many applications that require sample characterization and determination of constituents. However, GC characterization is not directly possible unless it is coupled with other techniques (e.g., mass spectroscopy) or referencing methods, such as retention indices, are used. This work provides a solid machine learning workflow that allows for accurate retention time predictions, which will facilitate the characterization process and allow GC optimization for isomers separation. Six different algorithms were trained on 0.7 of the manually compiled 608 data points that were gathered from RESTEK company chromatograms and published literature, covering 73 different GC settings and 61 compounds ranging from C1 to C12 hydrocarbons. The chosen model, namely the Adaptive Boosting Support Vector Regression, outperformed tree-based ensemble methods and showed high accuracies and consistency across the validation and testing sets with R^2 score of 0.992 -0.993.

Full Name : Hussein Neymer AlSheikh Mejhem AlAdwan

Thesis Title : MACHINE LEARNING MODEL FOR PREDICTING SHIFTS IN
GAS CHROMATOGRAPHY RETENTION TIMES CAUSED BY
THE CHANGE OF EXPERIMENT'S PARAMETERS AND
CONDITIONS

Major Field : Materials Science and Engineering

Date of Degree : May 2025

ملخص الرسالة

الاسم الكامل: حسين نمر الشيخ مجحم العدوان

عنوان الرسالة: توقع وقت الانبثاق الخاص بالجزئيات في تجربة الغاز كروماتوغرافي بناء على مواصفات الجهاز وظروف التجربة.

التخصص: هندسة وعلم المواد

تاريخ الدرجة العلمية: مارس 2025

الكروماتوغرافيا الغازية (GC) هي تقنية فصل متعددة الاستخدامات تُستخدم على نطاق واسع في العديد من التطبيقات التي تتطلب توصيف العينة وتحديد مكوناتها. ومع ذلك، لا يمكن إجراء التوصيف عبر GC بشكل مباشر ما لم ترتبط بتقنيات أخرى (مثل مطيافية الكتلة) أو تُستخدم طرق مرجعية مثل مؤشرات الاحتفاظ. يوفر هذا العمل سير عمل متيناً يعتمد على التعلم الآلي لتمكين التنبؤ الدقيق بأزمنة الاحتفاظ، مما يسهل عملية التوصيف ويسمح بتحسين ظروف GC لفصل الأيزومرات. تم تدريب ست خوارزميات مختلفة على 70% من أصل 608 نقطة بيانات جُمعت يدوياً من كروماتوغرامات شركة RESTEK والأدبيات المنشورة، شملت 73 إعداداً مختلفاً لـ GC و61 مركباً من الهيدروكربونات تتراوح من C1 إلى C12. وقد تفوق النموذج المختار، وهو الانحدار بدعم المتجهات المعزز تكيفياً (Adaptive Boosting SVR)، على الأساليب التجميعية المعتمدة على الأشجار وأظهر دقة عالية واتساقاً ممتازاً عبر مجموعات التحقق والاختبار، حيث بلغ معامل التحديد R^2 ما بين 0.99 و0.992.

Chapter 1

Introduction

Separation techniques aim to achieve sample characterization by splitting the substance into one or more of its components. Gas Chromatography (GC), as one of these techniques, has proven its efficiency across a wide range of applications that extend from tasks requiring relatively low analytical precision, such as SARA analysis (Saturates, Aromatics, Resins, Asphaltenes) of crude oil samples [1], to more critical and sensitive applications, e.g., characterization processes in the pharmaceutical industry [2]. This chapter will be divided into three sections: first, the “application background: gas chromatography” in which the GC principles and challenges will be discussed in detail. Second, the “methodology background: machine learning”, shows the theory of machine learning and the most effective algorithms used in this work. Lastly, the “problem definition and solution methodology”, where a view of the problem is presented, emphasizing the scope of benefits that the proposed solution will provide.

1.1 Application Background: Gas Chromatography

Separation techniques are widely used to reveal the components of unknown samples. GC, as one of the techniques under this category, is a powerful tool that has been used for over half a century, which, as the name suggests, differentiates between the components based on their separation behavior. The conventional GC main components can be reduced as follows: an input unit for inserting the sample, a heating system to vaporize the sample (and keep it vaporized), the column which is the separation unit, a stationary phase that lies along the inner wall of the column and works as a separation activator, and finally, the detection unit which detects the eluted, and supposedly already separated, components ([Figure 1.1](#)).

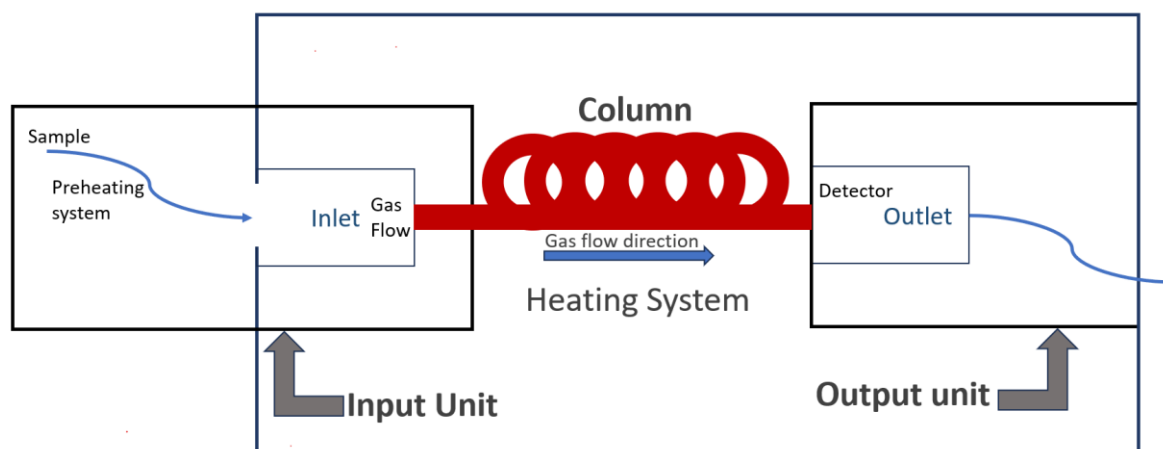


Figure 1.1: Gas Chromatography Scheme.

The carrier gas serves as a sample transporter; in other words, it carries the vaporized sample from the inlet to the detector through the column. The one attribute the carrier gas must possess is its ability to have minimum interaction with both the sample components and the stationary phase, thus, inert gases are usually the better choice. To sum up the process, first, the sample is evaporated through the preheating system; after that, the carrier gas drags the evaporated sample from the inlet into the column, where the components will be separated due to the difference in the mitigation effect.

1.1.1 Components Detection

When the components reach the outlet, they will be subsequently detected at their specific time ([Figure 1.2](#)) [3]. The difference in separation behavior is highly dependent on the interactions between the sample components and the stationary phase. For example, a polar stationary phase will cause a more mitigating effect on a polar component than its effect on a nonpolar component, therefore, the nonpolar component will be eluted first with a specific retention time (RT). The RT can be defined as the time required for a sample to reach the detector and be detected, which is a function of the experimental conditions and the component type [4].

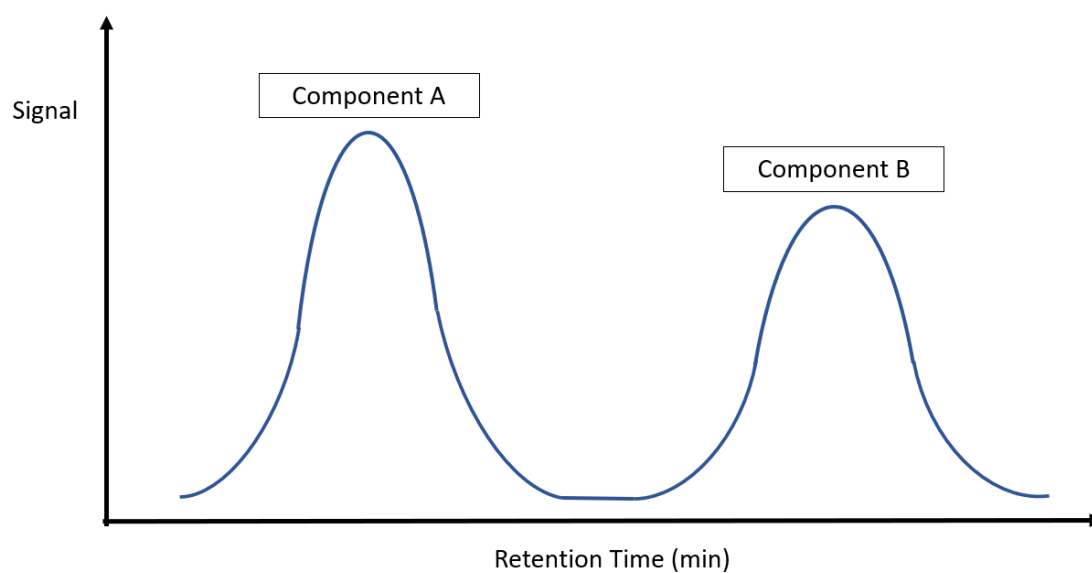


Figure 1.2: Ideal GC Chromatogram.

1.1.2 Chromatograms and Calibration

Chromatograms for the same sample can be different from one GC to another. This difference is called the retention time shift, which depends on the experimental conditions as mentioned before. Even though chromatograms don't provide direct qualitative detection, they still can be used to reveal the components' identity of the samples through standardized measures. For the conventional GCs¹, the identification process (assigning a peak to a specific component) is done according to the RT values. Standardized techniques involve using the retention index which is a relative measure of the RT with respect to the closest previous reference peak. The n-alkanes hydrocarbon peaks are utilized as the reference because of their characteristic high altitude [5]. Different retention indices exist for different experimental settings, more precisely, for different heating programs, for

¹ From this point until the end, GC will refer to the conventional GCs if not specified otherwise.

example, the index increase with time is logarithmic for isothermal separation and linear step-function for temperature-programmed heating [6]. [Figure 1.3](#) represents how RTs can shift from one GC experiment to another, therefore, using the RT values independently in the characterization process is worthless unless we have all possible components chromatograms for each specific GC (calibration), thereby, the RT values of the sample analysis can be compared with chromatograms map to identify each component. Nevertheless, indirect characterization could be made through standardization and referencing processes. Such processes resulted in Kováts retention indices[6] for isothermal experiments and van denDool and Kratz retention indices [4] for temperature-programmed experiments as the most famous methods for describing the GC retention data. These methods are based on normalizing the RT of components of interest to the RT of a reference series of organic compounds (usually n-alkanes). Even though the use of retention indices coupled with standardized experiments allowed for component identification, it is still showing major drawbacks regarding its reliability and generality, as in [7], added to the considerable amount of energy and time wasted for running the referencing experiments. Furthermore, the retention indices are not ideally independent of the GC experiment factors; in fact, the index values vary with changes in the stationary phase type and thickness [8].

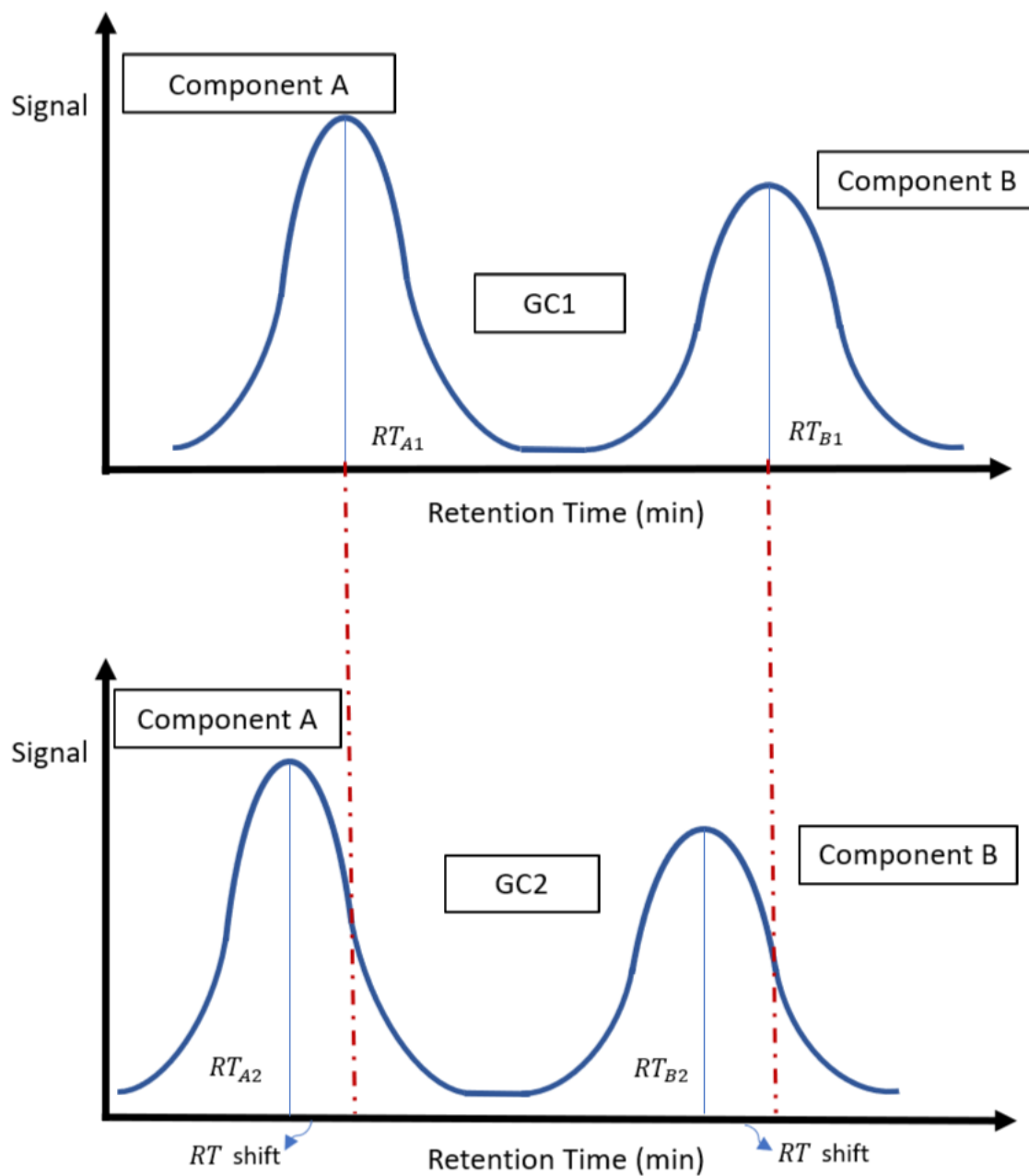


Figure 1.3: Retention Time Shifts in 2 GC experiments.

1.2 Methodology Background: Machine Learning

The field of Machine Learning (ML) was revolutionized by the publication of Geoffrey Hinton et al.'s paper in 2006 [9], where they showed the ability of Deep Learning Neural Networks algorithms to recognize handwritten digits with over 98% accuracy. In the past two decades since then, the use of ML algorithms has spread into almost every field in academic research or even in industry [10]. As a concept, ML is defined as “the field of study that gives computers the ability to learn without being explicitly programmed” (Arthur Samuel, 1959). In a much more recent, but yet verbally opposing, definition, Ethem Alpaydin described ML as “programming computers to optimize a performance criterion using example data or past experience” . However, combining both definitions suggests that the learning process is based on given data, in a process usually referred to as uncovering underlying patterns in datasets. ML algorithms can be classified in a variety of approaches. Based on the learning approach, ML algorithms are generally classified into supervised, unsupervised, and reinforcement learning. Based on the nature of the performed task, they are classified into regression, classification, and anomaly detection. Other categories can be inserted under each categorization method, but these are the most common types. The last way of classifying ML algorithms is based on the simplicity of the model; the more data-related assumptions imposed on the model, the simpler it is. Simpler models are called parametric models. Nonparametric models, on the other hand, employ fewer assumptions, if any, on the model [11].

1.2.1 Supervised Machine Learning

Supervised ML algorithms learn from labeled data, meaning that the input is correlated to the output through the learning process. For this to happen, the output, aka label or target, should be available in the dataset. Supervised ML algorithms can perform both regression tasks, where the output is continuous, and classification tasks, where the output is discrete. They can be parametric, e.g., linear regression, in which a linear relationship between the input and the output is presumed, or nonparametric with no prior assumptions. In ML algorithms, the difference between the prediction and the real value (the given target values) is minimized through the updating. The objective of every supervised ML algorithm is to minimize this difference, which is termed the loss function. The minimization of the loss function via the algorithm operation is called training. In the training process, each algorithm forms its own pathway toward the minimum loss value. Some supervised ML algorithms are presented below.

1.2.1.1 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is one of the simplest ML algorithms. It is distinguished by its fast execution and fair accuracy. The algorithm learns by “the labeled-distance voting”, in which the voting of the closest labels in the input space determines the label of the unknown sample. In the KNN's most basic form, there is no training process, and the predictions occur directly using the positions of the labeled or training data. [Figure 1.4](#) demonstrates how unlabeled data are classified based on their relative position from labeled data using KNN. The number of samples eligible to vote is the K number. The eligibility is defined by how close the training sample is to the unknown sample.

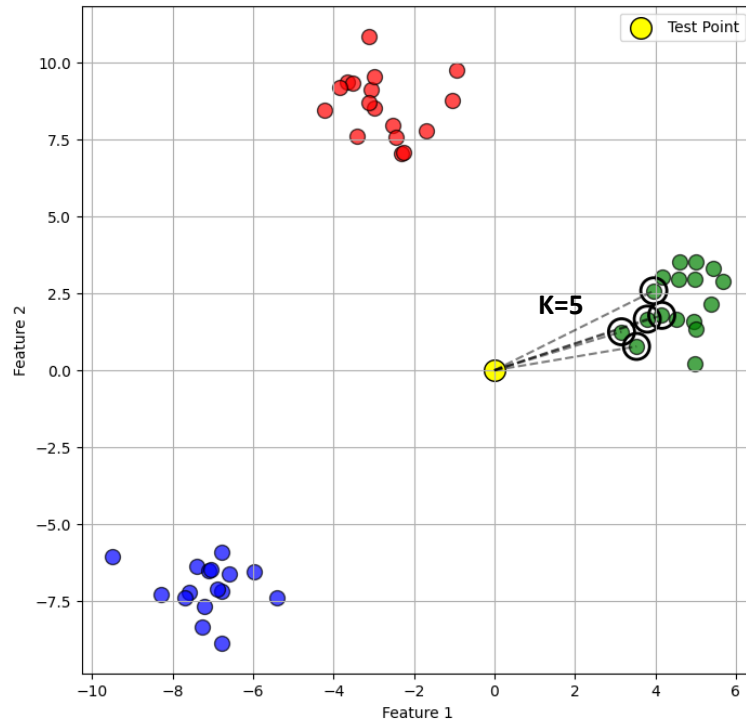


Figure 1.4: KNN Algorithm

1.2.1.2 Decision Trees

Decision Trees (DT) is one of the fundamental ML algorithms that are considered as the essence of ones of the most powerful algorithms available nowadays, namely, the Random Forest and Boosting Decision Trees. Unlike many other algorithms, DT and its derivatives' results can be interpretable, and they can estimate the impact of each feature in the input vector on the output, allowing for a feature importance measure. The algorithm depends on testing the dataset splitting ability of each feature, where at the beginning, the feature that is capable of achieving maximum splitting accuracy will be chosen to be the starting

point of the tree (root node). The splitting is based on a threshold value of the chosen feature, just as the feature is selected based on maximum accuracy, the threshold value is chosen for the same goal. After the first splitting is complete, two datasets will result: one with samples having a lower feature value than the threshold value, and the other with a feature value higher than the threshold value [10], [11]. [Figure 1.5](#) demonstrates a simple DT with each node, after the root node, representing a class in classification tasks or a value in regression tasks. The final nodes in the tree are responsible for the decision and are called leaf nodes.

1.2.1.3 Ensemble Tree Methods

Ensemble ML algorithms refer to models where multiple models participate in the prediction process. These multiple models can be generated by changing the algorithm type

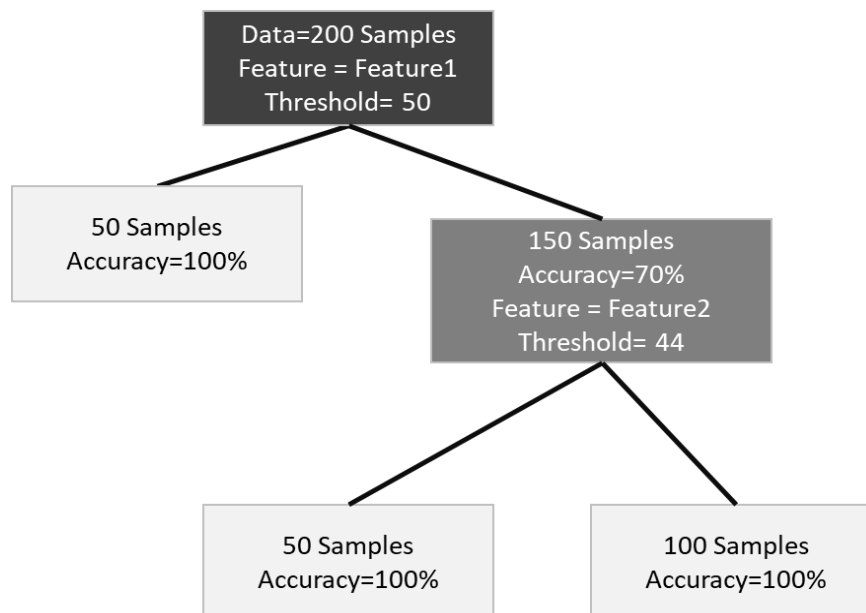


Figure 1.5: Illustration of a Simple DT Algorithm

and training the algorithms on the same dataset, or training the one algorithm on different subsets of the data. Random Forest (RF) and Decision Tree Boosting are ones of the most robust ensemble methods available. They are versatile and show stable performance over classification and regression tasks. RF consists of multiple DTs, where in each DT, a features subset is randomly sampled from the total set of input features to test the splitting. This random sampling adds overall bias to the model, and upgrades the DT to a higher performance level. On the other hand, Boosting Algorithms refer to a set of algorithms that combine low-performance models into higher-performance models (boosting). The boosting process varies from one algorithm to another. For instance, Gradient Boosting DT (GBDT) works by subsequently increasing the number of estimators in the model. The added estimators rectify the previous ones by predicting the residuals made by them and then adding these residuals to the previous estimators' predictions. Variationally, Adaptive Boosting DT (ABDT) follows a different approach by adjusting the weights of both the falsely predicted instances and the weights of the low-performance estimators. The instances with high residuals are given higher loss weights in the subsequent estimator; thereby, the model will focus on them to minimize the overall error. Inversely, the weak estimators are given less weight in the voting process [10], [11].

1.2.1.4 QLattice Symbolic Regression

QLattice Symbolic Regression (QLSR) is a supervised ML learning algorithm employed in Python by the Feyn module, which was developed in recent years by Abzu ApS. [12]. QLSR can generate mathematical formulas that define a claimed deterministic relationship between the input variables and the output. These formulas are generated by combining

different functions, also called interactions, ranging from simple operators such as addition and subtraction to more complex functions like logarithmic or exponential to form different paths based on these combinations, where the path with the highest resulting accuracy is selected.

1.2.2 Unsupervised Machine Learning: Isolation Forest

Unsupervised ML algorithms offer the capability of performing multiple tasks such as data cleaning, clustering for classification, and dimensionality reduction, to name a few. This type of ML algorithms does not require labeled data to learn from; on the contrary, the mentioned tasks are done by leveraging the input space only. In this work, an unsupervised ML algorithm, namely the Isolation Forest, is used for data cleaning. Isolation Forest [13] is an efficient anomaly detection algorithm that performs well with high-dimensional datasets. The algorithm is based on the assumption that anomalies in data are both special in their feature values and rare in their quantity. The final criterion of anomaly is how fast the isolation happens, where the quickness of the isolation is measured by how deep a branch should be for an instance to be isolated. An average of this depth over many trees is used as an anomaly detector in the Isolation Forest algorithm.

1.2.3 Preprocessing and Postprocessing Techniques

1.2.3.1 Data Scaling: Z-score standardization and Min-max Normalization

Most ML algorithms show improved performance when applied to scaled data. The two most common scaling techniques are the Z-score, also known as the standard scaler, and the Min-max normalization. The Z-score for a feature value is equal to the original value minus the mean value of that feature, divided by the standard deviation, whereas min-max normalization transforms each feature value by subtracting the minimum value and then dividing by the range. Each scaling method has its pros and cons; for instance, the min-max normalization is more sensitive to outliers than Z-score standardization and might lead to shrinking the majority of samples because of the presence of some outliers, while the min-max normalization is simpler and easier to transform to a custom range.

1.2.3.2 Correlation Test: Pearson and Distance Correlations

Looking at the correlations between the features and the target before designing an ML learning model can be quite helpful for feature reduction and interpretability. Furthermore, the benefits of correlation tests extend also to post-processing analysis, where they can be used to validate the feature-importance results of ML models, and therefore the model itself. Pearson's coefficient [14] is widely used to capture linear correlations between variables according to the following formula:

$$P(X, Y) = \frac{\sum_{i=1}^k (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^k (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^k (Y_i - \bar{Y})^2}} \dots\dots\dots 1.1$$

Where P is the Pearson correlation coefficient, X_i and Y_i are the instance i value of the two variables, $\bar{X}$ and $\bar{Y}$ are the mean values, respectively. The Pearson correlation score ranges from -1 to 1, where -1 indicates perfect inverse correlation, 0 indicates independence, and 1 indicates perfect proportionality. It is worth mentioning that Pearson

correlation coefficient only accounts for linear correlations; therefore, other nonlinear effects, e.g., exponential or periodic, are undetected [15]. More comprehensive detection of correlations can be achieved through advanced methods such as the distance correlation score [15], [16]. The distance correlation score ranges from 0 to 1, where 0 indicates perfect independence and 1 indicates perfect dependence.

1.2.3.3 Evaluation Metrics

The R^2 score, also known as the coefficient of determination, the Root Mean Squared Error (RMSE), and the Mean Absolute Error (MAE), are among the most common evaluation metrics. The three metrics are defined below:

$$R^2 = 1 - \frac{\sum_i (y_i - y_p)^2}{\sum_i (y_i - \bar{y})^2} \dots\dots\dots 1.2$$

$$RMSE = \frac{1}{n} \sqrt{\sum_i (y_i - y_p)^2} \dots\dots\dots 1.3$$

$$MAE = \frac{1}{n} \sum_i |y_i - y_p| \dots\dots\dots 1.4$$

Where n is the number of predictions, y_i and y_p are the real and predicted values, respectively. The RMSE is well known for being sensitive to outliers, which makes it a better outlier detector than any other metric. Both the MAE and the RMSE are relative to the scale of the data, while the R^2 is more independent of the scale and can be considered as a general fitting metric, where a value of 1 represents a perfect fitting, 0 means that the model is as good as the mean, and negative values mean that the model performs worse

than the mean. To further ensure the models' generalizability, previous studies suggested that Monte Carlo cross-validation should be applied [17]. Monte Carlo cross-validation is a simple technique that aims to show how the model performs when trained on different training sets by changing the random splitting.

1.3 Problem Definition and Solution Methodology

1.3.1 Problem Definition

It was established in Section 1.1.2 that it is not possible to directly reveal the components' identity of a sample via a GC experiment without resorting to referencing processes or introducing more advanced techniques such as mass spectrometry (MS). However, the referencing processes consume energy, time, and materials. On the other hand, GC-MS devices, while being more accurate, are more expensive to purchase with 3 times higher purchase prices as in [18], along with the higher maintenance costs.

1.3.2 Methodology

ML algorithms represent a useful tool that can replace the MS technique or the referencing methods in GC chromatogram analysis. To achieve this target, the retention time values should be given in a way that each molecule has its specific and unique value regardless of the experimental conditions. To tackle this problem, a reference GC is given RT shift values of zero for all compounds, while all other GCs' RT shifts are calculated according to the following equation:

$$RTshifts = RT - RT_{ref}..... 1.5$$

Where RT_{ref} is the retention time of the reference GC for a specific compound, and RT is the retention time of the used GC for the same compound. The RT shift values are used as the target in the ML model; that is, the values will be predicted for any GC experiment, then, equation 1.2 can be employed to predict the RT values of any compound at any GC experiment. The input vector, however, comprises 15 different effective features that can be classified into 3 distinct categories. The first category, the compound-related input, contains two features, namely, the compound ID and the molar mass. The second category is related to instrumental parameters and contains 4 features. This includes the column length and diameter, the stationary phase thickness, and the carrier gas flow rate. The last category is the experimental conditions category, in other words, the heating modes where features are involved based on a 3-steps-temperature program: the starting temperature, the first temperature hold period, the first heating rate, the second temperature, the second temperature hold period, the second heating rate, the third temperature, and the third temperature holding period. In the case of isothermal experiments or experiments involving 2 or 4 temperature steps, the temperature program is systematically and consistently extended or reduced to a 3-step program to maintain model simplicity. If the experiment is isothermal, then the heating rates and the first and second holds are set to 0, where the final hold is the experiment duration. All the temperature steps are given the isothermal temperature value. For 2 temperature steps, the second heating rate and holding time are set to 0, while the second and third temperatures are given the same value. In this case, the third holding time is the one after reaching the second temperature. In the third case where 4 temperature steps are imposed, the third temperature step is ignored and the second heating rate is taken to be the average according to the following equation:

$$\frac{dT}{dt} = \frac{H_2\Delta T_2 + H_3\Delta T_3}{\Delta T} \dots\dots\dots 1.6$$

Where H_2 and H_3 are the second and third heating rates, ΔT_2 and ΔT_3 are the second and third temperature differences. ΔT is the temperature difference between the second and the 4 temperature steps. [Table 1.1](#) shows a summary of the data variables and the reference value of each variable. The data are collected from 2 sources: Restek’s company website, a company that specializes in manufacturing GCs and Liquid Chromatography (LC) devices. The website offers a searchable chromatogram library, from which the data is gathered. The reference chromatogram ([equation 1.2](#)) is shown in [Figure 1.6](#). The cleaned dataset is split into three parts: 0.8 for training, 0.1 as validation set for hyperparameters optimization, and 0.1 as a holdout testing set. The first split is stratified with respect to the “Order” column to ensure that all compounds are involved in the training process in a balanced distribution. Six different algorithms are applied in Python 3.12 environment through Scikit-learn and Feyn modules [19],[12] namely, Random Forest, Gradient Boosting Decision Tree, Adaptive Boosting Decision Tree, eXtreme Gradient Boosting, K-Nearest Neighbors, and Adaptive Boosting Support Vector Regression. (RF, GBDT, ABDT, XGB, KNN, and ABSVR, respectively) through Scikit-learn and the QLattice Symbolic Regression (QLSR) through Feyn. Three evaluation metrics are adopted to evaluate the model: MAE, RMSE, and the R^2 score. Given the complex, size-limited-categorical nature of the input data (each compound considered as a category), Monte Carlo Cross Validation is first applied to the t to get a fair evaluation of the model generalizability [17]. The number of random seeds used is found through validation R^2 score convergence test. After that, all of the previously mentioned algorithms are applied on one random seed to obtain final predictive models.

Table 1.1: The Model Variables with The Reference Chromatogram Values.

Variable	Explanation = Reference Conditions
RTshifts (Label)	Retention time shifts
Order	The compound's identifier; order of elution from the reference chromatogram.
MW	Compound's molar mass.
Codiamm(mm)	The column diameter. = 0.25
Colength(m)	The column length. = 105
FlowRate(m/min)	The carrier gas flow rate. = 2
StaThick(micrometer)	The stationary phase thickness. = 1
1Temp	Initial temperature. = -45 C°
1Hold	First holding time. = 4 min
1IncrRate/min	First heating rate. = 4.5 C°/min
2Temp	Second temperature step. = 125 C°
2Hold	Second holding time. = 0 min
2IncrRate/min	Second heating rate. = 9 C°/min
3Temp	Third temperature step. = 250
3Hold	Third holding time. = 0 min
Polarity	Stationary Phase Polarity. = 0 arbitrary unit

Peaks	t _R (min)	Peaks	t _R (min)	Peaks	t _R (min)
1. Ethylene	7.91	21. 3-Methylpentane	30.70	41. Chlorobenzene-d5 (IS)	43.79
2. Acetylene	8.10	22. 1-Hexene	31.04	42. Ethylbenzene	44.57
3. Ethane	8.32	23. Bromochloromethane (IS)	31.60	43. m-Xylene	44.89
4. Propylene	11.44	24. n-Hexane	31.69	44. p-Xylene	44.93
5. Propane	11.76	25. Methylcyclopentane	33.30	45. Styrene	45.62
6. Isobutane	16.08	26. 2,4-Dimethylpentane	33.41	46. o-Xylene	45.83
7. 1-Butene	18.01	27. Benzene	34.67	47. n-Nonane	46.14
8. n-Butane	18.60	28. Cyclohexane	35.19	48. 4-Bromofluorobenzene*	46.71
9. trans-2-Butene	19.41	29. 1,4-Difluorobenzene (IS)	35.32	49. Isopropylbenzene	46.96
10. cis-2-Butene	20.41	30. 2-Methylhexane	35.44	50. n-Propylbenzene	47.95
11. iso-Pentane	23.79	31. 2,3-Dimethylpentane	35.62	51. m-Ethyltoluene	48.15
12. 1-Pentene	24.80	32. 3-Methylhexane	35.93	52. p-Ethyltoluene	48.23
13. n-Pentane	25.53	33. 2,2,4-Trimethylpentane	36.70	53. 1,3,5-Trimethylbenzene	48.38
14. Isoprene	25.83	34. n-Heptane	37.17	54. o-Ethyltoluene	48.77
15. trans-2-Pentene	26.05	35. Methylcyclohexane	38.52	55. n-Decane	49.32
16. cis-2-Pentene	26.54	36. 2,3,4-Trimethylpentane	39.88	56. 1,2,4-Trimethylbenzene	49.21
17. 2,2-Dimethylbutane	27.76	37. Toluene	40.22	57. 1,2,3-Trimethylbenzene	50.07
18. Cyclopentane	29.50	38. 2-Methylheptane	40.49	58. m-Diethylbenzene	50.67
19. 2,3-Dimethylbutane	29.60	39. 3-Methylheptane	40.89	59. p-Diethylbenzene	50.87
20. 2-Methylpentane	29.86	40. n-Octane	42.09	60. Undecane	51.89
				61. Dodecane	54.02

Figure 1.6:Reference Chromatogram

Chapter 2

Literature Review

As outlined in Chapter 1, this work aims to optimize the GC by using ML algorithms to predict the RT values of different compounds at different GC experimental parameters. Multiple studies were done in an attempt to enhance not only the GC technique using ML but also other chromatographic instruments such as liquid chromatography and high-performance liquid chromatography. In this chapter, a several chromatography-ML studies are overviewed, where the GC related studies mostly utilized GC-MS data and followed Quantitative Structure Retention Relationships (QSRR) [20], yet quite few aimed to directly predict RT values rather than the retention indices and, to the author's knowledge, none provided a complete workflow to use RT values for component identification in conventional capillary GCs.

2.1 Liquid Chromatography Studies

One of the first studies that addressed RT prediction in LC was made by K. Munro et al. [21] where they trained three different ANN algorithms on an LC-high resolution mass spectroscopy data set composed of 166 different compounds related to pharmaceuticals and personal care products in water sports environment with 15 input features generated by Simplified Molecular Line Entry System Strings (SMILES). The best achieved R^2 was 0.9204 by generalized regression neural networks. In another study, E.S. Fedorova, D.D.

Matyushin, I.V. Plyushchenko et al. on reversed phase LC [22]. They developed a deep-learning model for predicting RTs by training ANNs-algorithms on, first, the main dataset of over 80k molecules separated by the same LC setting, then, using 5 more datasets for transfer in a transfer learning process. With this huge dataset, they used 243 different molecular descriptors as input vector. The final model predictive capability was proved by an MAE value of 34.7 seconds. However, it is crucial to consider the range of the target values when judging a model's performance by the MAE which did not exceed 25 minutes. E. Choi, W.J. Yoo, H.-Y. Jang et al. [23] did similar work where their model's scope was focused on LC experiments of dansylated metabolites. The model was trained in a supervised manner on 315 data points that were obtained from a fixed LC column (Zorbax Eclipse C18 column (2.0 mm $\times$ 100 mm, 1.7 μ m, 95 Å)). As in most supervised machine learning studies, the data was divided into a 0.8-0.2 training-testing split. The ANN was applied where 58 SMILES descriptors were used as input features. These descriptors were extracted from 1159 descriptors after using the GenericUnivariateSelect algorithm of the Scikit-learn module. This resulted in a testing MAE of 1.145 min and an adjusted testing R^2 0.967, which proves the model's potential.

2.2 Multidimensional Chromatography

Similar was done for LCs. Veenaas C. et al., and after developing a novel retention index system for two-dimensional GC by referencing to the elution of polyethylene glycol series [24], they investigated the partial least squares algorithm and the federation of local models workflow for predicting four different responses, namely, the first and the second-

dimension RTs, the first-dimension retention index, and their second-dimension retention index[25]. The authors' data consisted of 859 compounds run by a single GC×GC setting. The best first dimension RT prediction is attributed to the partial least squares algorithm with Root Mean Squared Error (RMSE) of 109 seconds for a target range of around 51 minutes.

2.3 Gas Chromatography Studies

To the author's knowledge, the first study that addressed RT predictions is attributed to V.K. Gupta et al. [26], where they followed the QSRR by generating hundreds of numerical descriptors of 64 different fatty acid methyl esters. These descriptors are structure-related parameters that can be utilized as inputs for RT prediction in capillary GC. After eliminating the unvarying descriptors along with the highly correlated ones, a total 172 resulted. The authors used the genetic algorithm to extract the most effective descriptors, which resulted in selecting the gravitational index, the number of cis double bonds, and the number of trans double bonds. Using these three descriptors, the authors trained Artificial Neural Networks (ANN) algorithms with different settings on 49 training data points. The model showed superior accuracy in predicting the testing and validation target values (9 and 6 samples respectively) with testing and validation coefficient of determination scores (R^2) of 0.999 for the best settings. Despite these R^2 values, the model lacks the generality over compounds and GC settings since the training points are obtained from one GC experiment for a very narrow class of compounds. In the last decade, several large-data ML-GC studies started to emerge [8], [27], [28], [29], [30], [31], [32], thanks to the

availability of large GC-MS databases made by NIST. All of these studies focused on predicting the retention indices by leveraging QSRR and SMILES. It is worth mentioning that, in its most recent release, NIST(23) contained more than 492,000 retention index values for 153,000 different compounds [33]. Despite the magnificent impact of the NIST23, the absence of direct RT values and some experiment parameters, such as the carrier gas flow rates, reduced its usability for our approach. Quite recently, three papers were published addressing the problem of RT prediction, 2 of them considered predicting RT for LC experiments, while the other regarded the ion-pair chromatography [34], [35], [36]. An intriguing work was done by M. Zakir et al. [34] as they utilized the variation of the experimental parameters by using a dataset composed of 44 different chromatographic methods and around 10,500 different compounds. They trained a Graph Neural Networks (GNN) model on stratified training data of 90% of the total data. The model is designed to extract the variation of RT values across different chromatographic conditions by sensing the variation in the RT values difference between different compounds after changing the chromatographic conditions, not by directly learning from the experimental parameters input. As a result, while the model showed excellent performance when applied to chromatographic conditions with a large compounds number, it failed to maintain the same performance when used to estimate RT of 24 datasets with less than 100 compounds. The average R^2 and MAE values across these estimations were 0.75 and 1.87 minutes, respectively. Our work aims to take an unusual approach by mainly depending on the GC experimental conditions dimensions to predict the RT values, rather than using detailed molecular descriptors dimensions. We constructed a dataset that is composed of 13 experimental input features and 2 simple molecular descriptors. In addition to predicting

RT values, the proposed methodology can be used for GC design for maximum separation of hardly-separated components. Based on the literature review, our proposed methodology is completely novel, and our results proved its robustness.

Chapter 3

Results and Discussion

3.1 Data Cleaning

The Isolation Forest outliers removal reduced the dataset size from 683 to 608 instances. [Figure 3.1](#) shows the samples' distribution over the compounds before and after the outliers' removal. [Table 3.1](#) demonstrates the statistics of the cleaned data. The final cleaned dataset contained a total of 73 distinct GC experiments that are distinguished by at least one experimental parameter. The distribution of all features is shown in appendix 1.

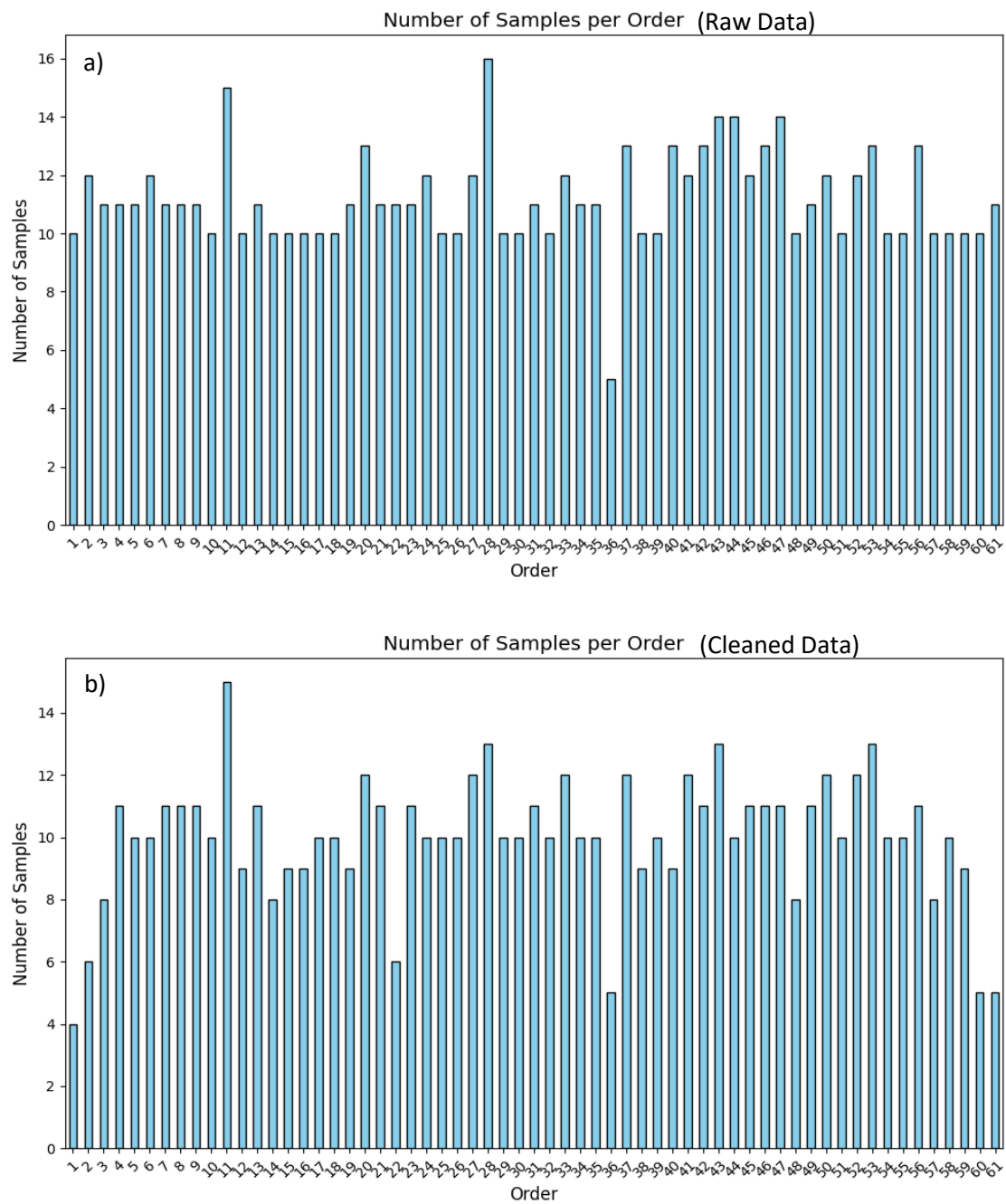


Figure 3.1: Samples Distribution over: a) Raw data, b) Cleaned Data.

Table 3.1: Data Statistics of the Cleaned Data.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Colength(m)	608	58.08	24.94	15.00	30.0	60.0	60.0	105.0
Codiame(mm)	608	0.31	0.15	0.10	0.25	0.25	0.32	0.53
1Temp	608	20.75	35.56	-60.00	0.0	30.0	40.0	85.0
1Hold	608	7.48	6.33	0.0	4.0	6.0	10.0	30.0
1IncrRate/min	608	8.03	7.49	1.0	3.0	5.0	10.0	30.0
2Temp	608	139.27	65.74	50.00	80.0	100.0	200.0	330.0
2Hold	608	0.27	0.94	0.0	0.0	0.0	2.0	40.0
2IncrRate/min	608	10.00	12.25	0.0	0.0	5.0	11.0	40.0
3Temp	608	223.53	39.44	50.00	200.0	225.0	300.0	330.0
3Hold	608	3.13	3.96	0.0	0.0	2.0	5.0	15.0
FlowRate(mL/min)	608	3.90	5.81	0.5	1.2	2.0	4.0	40.0
StaThick(micrometer)	608	2.54	3.89	0.5	1.0	1.4	2.0	20.0
MW	608	96.02	29.54	26.04	72.2	100.2	120.9	175.0
Polarity	608	0.25	0.29	0.0	0.0	0.25	0.5	1.0
Order	608	31.13	17.02	1.0	16.0	31.0	46.0	61.0
RT shifts	608	-15.46	15.19	-46.26	-26.34	-16.36	-5.62	49.43

3.2 Monte Carlo Cross Validation

A simple RF algorithm is used in the Monte Carlo convergence test with respect to the R^2 , resulting in the plot in Figure 3.2. The convergence can be observed at around 260 random seed runs. However, due to the extensive heavy calculations required if this number is used, we selected a much lower number for hyperparameters optimization, which provides a similar average accuracy. The numbers between 15 and 25 seemed to meet this condition; eventually, using 18 random seeds was adopted. After cleaning the data and obtaining the convergence random seed number, hyperparameters optimization was done via grid search based on the lowest Monte Carlo average validation MAE. Since the RMSE penalizes the outliers errors, and considering that some random seeds might produce exceptional residuals, using the average RMSE score would not be the best option to base the hyperparameters selection on. The average Monte Carlo evaluation metrics, denoted by MC, are summarized in Table 3.2.

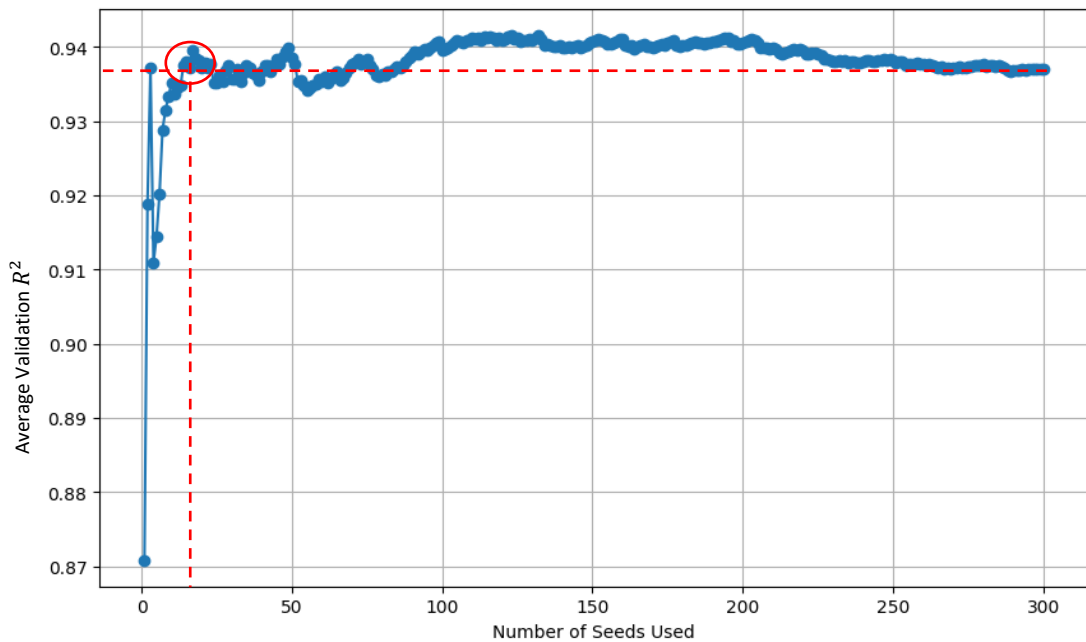


Figure 3.2: Monte Carlo Validation R^2 Convergence Test

Table 3.2: Average Monte Carlo Evaluation Metrics.

Algorithm	Training R^2_{MC}	Validation R^2_{MC}	Testing R^2_{MC}	Testing MAE_{MC}	Testing $RMSE_{MC}$
GBDT	0.999	0.949	0.951	1.245	2.945
ABDT	0.999	0.935	0.938	1.316	3.232
RF	0.992	0.944	0.949	1.467	3.138
XGB	0.998	0.946	0.947	1.391	3.122

The average Monte Carlo evaluation metrics showed that the models can maintain high predictive capability and strong fitting ($R^2 \approx 0.95$ for the best algorithm) and consistent performance over the validation and testing sets. These results allowed for proceeding toward training the models on a single random seed.

3.3 Correlation Results

To see which feature is contributing the most to the shifts in RT across different GCs, Pearson and the distance correlation tests were performed, resulting in correlation scores in Figure 3.3. Before evaluating these results, it should be taken into account that to have representative correlations, the variation in each feature has to be as isolated as possible from the variation of other features to catch the independent effect of each feature on the target, or in other words, to avoid the effect of confounders [37]. Since this is not the case in our limited-size dataset, the association of some important variables with the target might be underestimated. None of the features can be considered as highly correlated with the target value, which justifies the use of nonparametric ML models for accurate RT predictions. Both the distance and Pearson showed agreement regarding the first oven temperature and the column length are the top two most correlated features with the shifts in RTs. The effect of the temperature programs is well studied, as in [38]. While the column length effect on the RT time is obvious, as compounds take a longer time to travel through longer columns. After the first two features, the rank of the most correlated features given by distance correlation starts to deviate from that given by Pearson due to the effects of nonlinear correlations accounted by the distance method. The polarity of the stationary phase is one of the most associated variables with the shifts in RTs, given that it is one of the properties that defines the nature of the interactions between the compounds and the stationary phase, causing the retention effect. Unexpectedly, the carrier gas flow rate showed weak correlation with the target by both methods, where the confounding effect can be proposed as a reasonable interpretation. Except for the Order and the MW (Pearson score =0.9), no significant correlations were found between the features (<0.74 at most).

The Pearson correlation matrix is shown in Figure 3.4. To account for the effect of isomers in the model, we considered keeping both the Order feature as the compound ID and the MW.

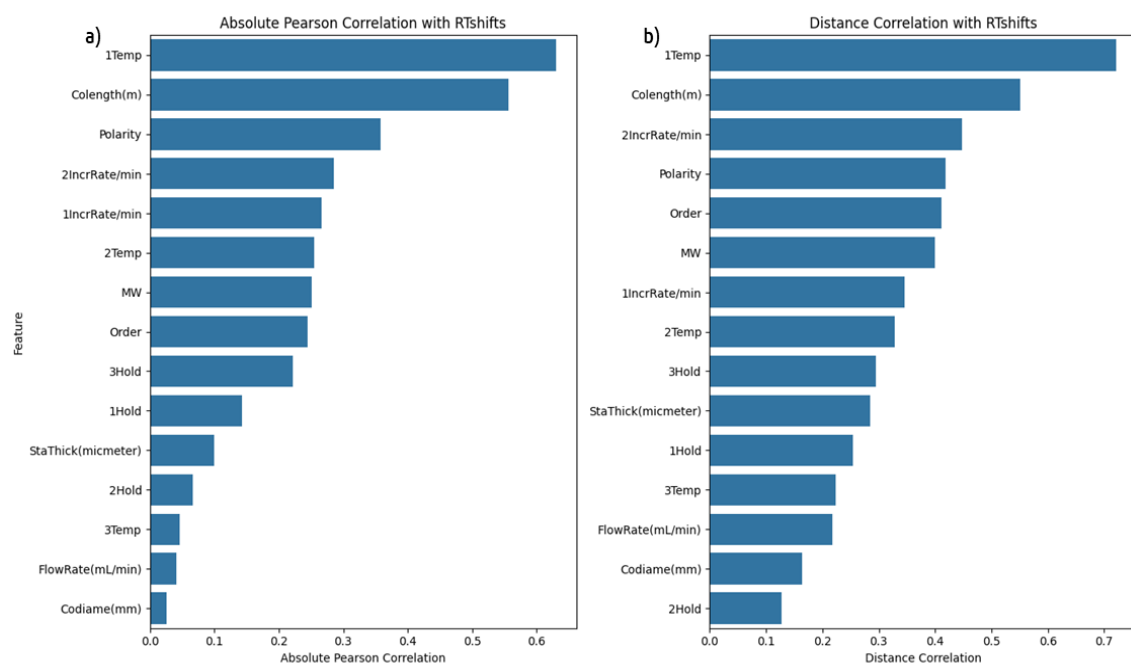


Figure 3.3: Correlation Results between the Features and the Target: a) Absolute Pearson Correlation, b) Distance Correlation.

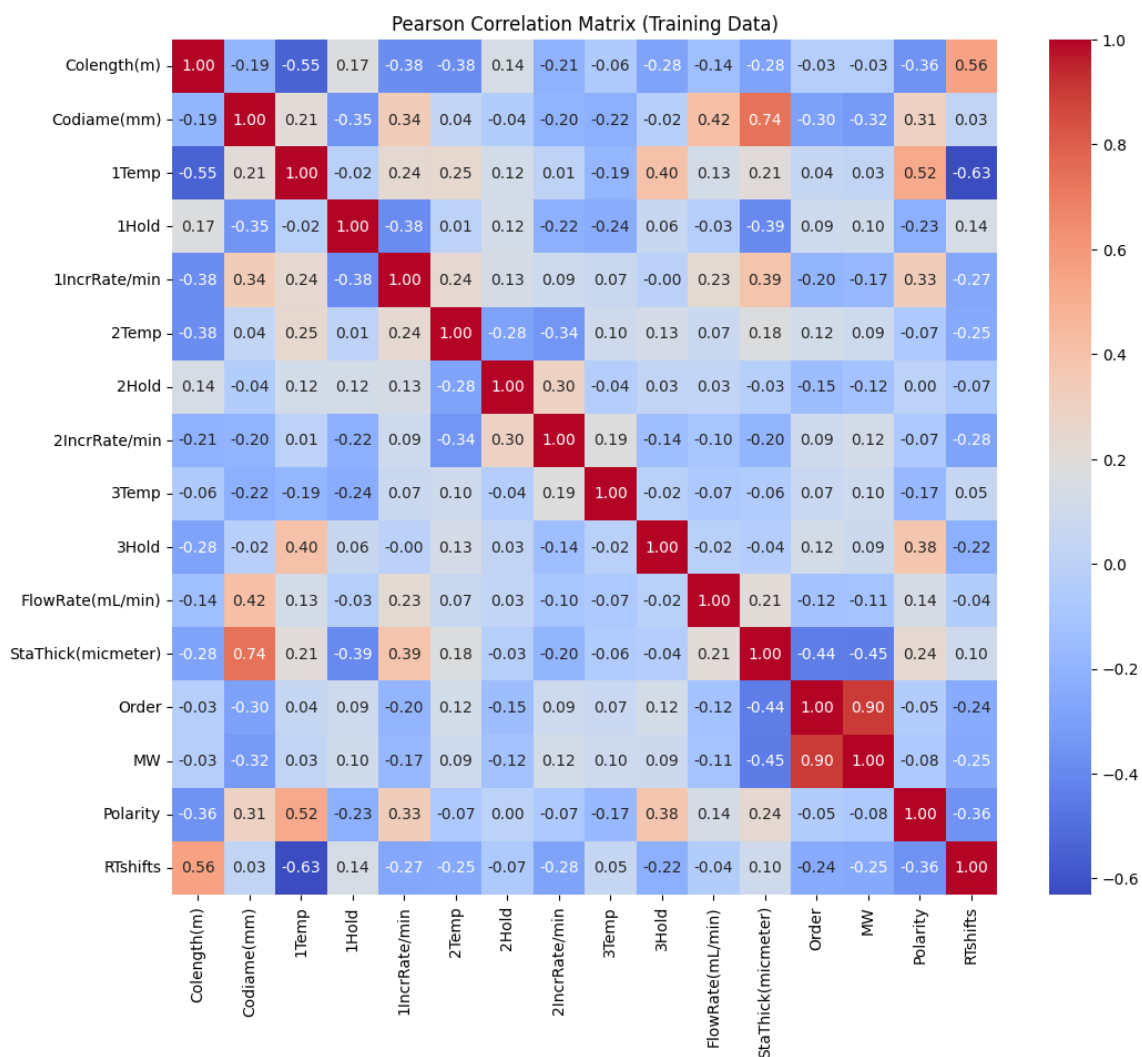


Figure 3.4: Pearson Correlation Matrix.

3.4 Final Models:

Six ML algorithms were applied on the aforementioned basis, namely, RF, GBDT, ABDT, XGB, KNN, and ABSVR. The best settings of the applied algorithms can be found in Appendix 2 based on the highest validation R^2 score where only the non-default parameters are listed. Figure 3.5 shows the models' predictions vs real values. All the used four tree ensemble methods showed quite similar performance with more outlier predictions than the distance-based algorithms, especially the data points between 20 and 40 minutes, where all of these algorithms failed to achieve completely accurate predictions within this range. Nevertheless, these models succeeded in estimating the single testing data point at around 45 minutes, meaning that the predictive capability could cover the whole range of the data. Unlike the tree ensemble methods, the KNN showed better generalization on account of higher bias errors in the high-density scope. This better generalization manifested by the more accurate estimations of the 3 data points in the range between 20 and 40 minutes. Moreover, the ABSVR showed superior performance with less variance and bias errors, and a predictive capability that covers from around -45 to 45 minutes, which is almost the range of the target variable, as in Table 3.1. Table 3.3 shows the training, validation, and testing evaluation metrics of the six algorithms. All the tree-based ensemble methods showed a plain drop from the training prediction to the validation then to the testing ones. The drop in performance over non-training data can be attributed to a slight overfitting, which is further confirmed by the high training R^2 score. On the other hand, the performance mitigation in the test predictions when compared to the validation predictions can be interpreted as knowledge leakage during the tuning process. The high RMSE/MAE ratio indicates that most of the residuals are coming from outlier predictions [39].

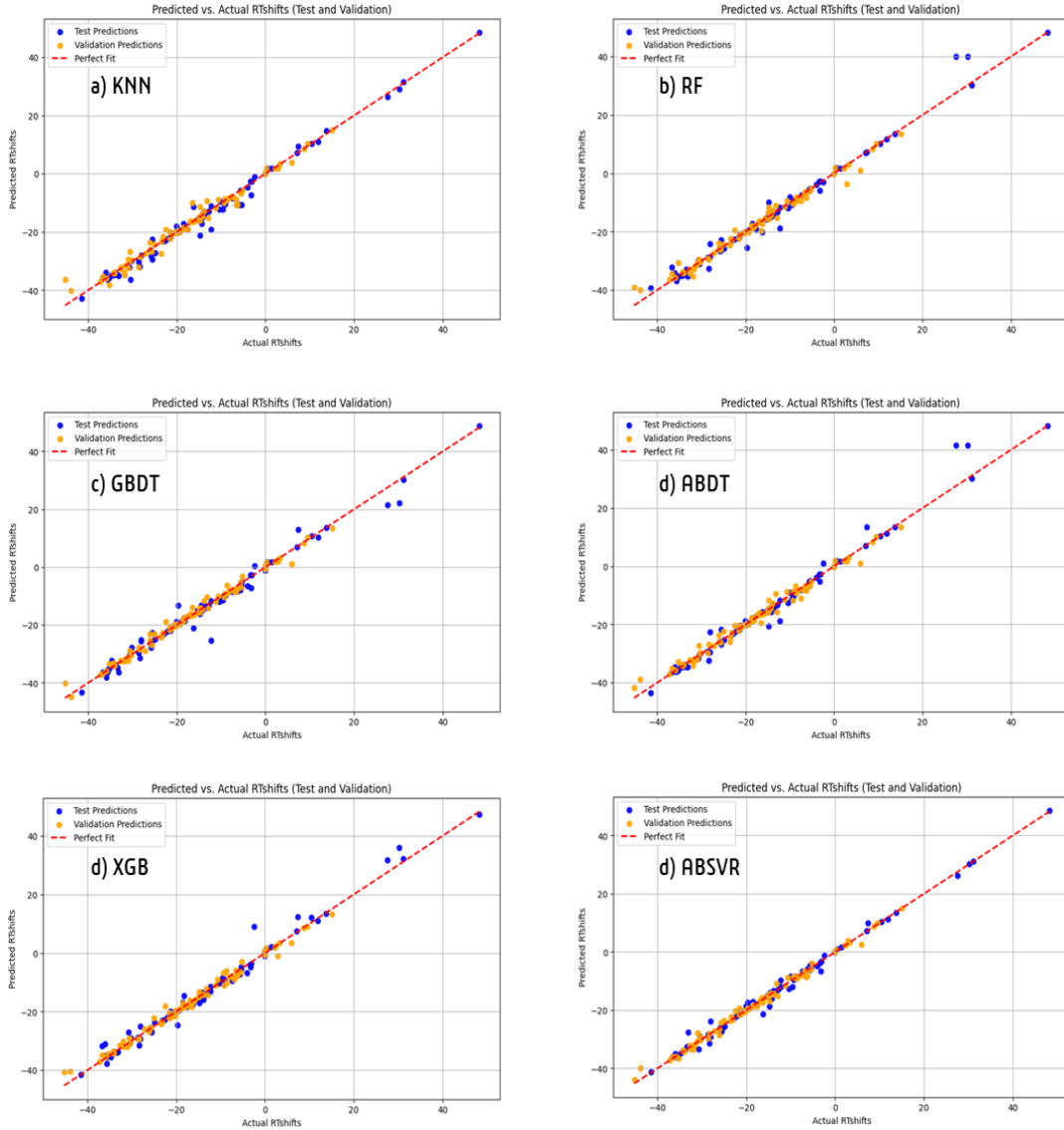


Figure 3.5: Predicted vs. Actual RTshifts for the Validation and Testing Sets: a) KNN, b) RF, c) GBDT, d) ABDT, e) XGB, and f) ABSVR.

Despite the lower MAE values when compared to the other tree methods, KNN showed better generalizability, with improved performance over the testing set when compared to the validation set. Nevertheless, the lack of consistency over the different sets might indicate an overfitting risk. This risk is at minimum if the ABSVR is applied as shown by its high accuracy and consistent performance. With an almost equal R^2 over the three sets,

and an average MAE value of 0.75 minutes over the validation and testing sets, the ABSVR proved high reliability and a clear advantage over other RT prediction models. The low RMSE/MAE ratio indicates less susceptibility to outliers. Therefore, this ABSVR model is selected.

Table 3.3: Algorithms' Training, Testing, and Validation Metrics

Algorithm	Training R^2	Training RMSE	Training MAE	Validation R^2	Validation RMSE	Validation MAE	Testing R^2	Testing RMSE	Testing MAE
KNN	0.999	0.462	0.034	0.979	1.892	1.106	0.985	1.983	1.218
RF	0.999	0.525	0.163	0.985	1.608	0.935	0.981	2.266	1.098
GBDT	0.998	0.601	0.038	0.991	1.272	0.822	0.979	2.379	1.333
ABDT	0.999	0.462	0.036	0.987	1.500	0.970	0.977	2.476	1.158
XGB	0.999	0.468	0.048	0.990	1.342	0.922	0.984	2.095	1.222
ABSVR	0.993	1.249	0.197	0.993	1.068	0.663	0.992	1.423	0.838

The residual plot of the ABSVR model is shown in Figure 3.6. To have a realistic view of the model performance, and since the elution span (the peak width) in capillary GCs typically ranges from 6 to 30 seconds and can reach up to 1 minute depends on the compounds and experimental parameters [40], we counted the number of estimations with error values above 30 seconds, 1 and 2 minutes, and those with error less than 6 and 1 seconds. Out of 183 validation and testing total samples, 21 predictions showed an error of more than 2 minutes, 48 showed with an error above 1 minute, and 70 estimations with an error above half a minute. 47 predictions showed an error below 6 seconds, and 22 exact predictions with an error of less than 1 second.

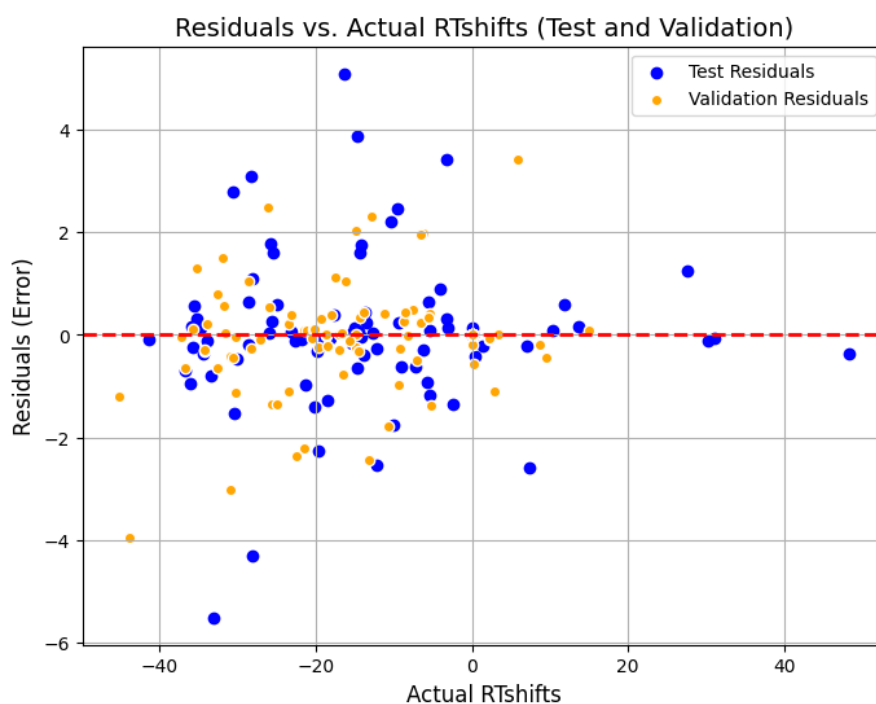


Figure 3.6: Validation and Testing Residuals of the ABSVR Algorithm

3.5 Performance and Scope Comparison with Models in Literature:

A comprehensive comparison of our model results with other models that aimed to predict the retention time, showing the compound and GCs scope, is demonstrated in Table 3.5. Most of the RT prediction models focused on LC rather GC, utilizing the molecular descriptors as the RT estimator. The first model in the table showed remarkable performance across the validation and testing sets in predicting the RT values of GC, but due to the limited testing and validation data points, its reliability is highly questionable. Furthermore, only one GC is considered in the study, which will not serve in a GC optimization process. The other three studies tried to address RT prediction in LCs. The fact that our proposed model showed more accurate performance than LC RT predictive models, despite that LC experiments are usually less susceptible to experimental error noise, proves the method's robustness. This advantage can be attributed to three major merits that distinguish this work: First, the inclusion of explicit experimental parameters in the data features, second, using direct target encoding as molecular ID, namely the MW and the order of elution from the reference GC, third, the prediction of the RT shifts, then transforming these shifts to RT values boosted the performance since the RT shifts are highly correlated not only with the experimental parameters, but also with the compounds identity. However, it is important to emphasize that this work is not offering a final product, given the amount of data; rather, it proposes a work flow for accurate RT predictions following the referencing approach, and it is high likely that merging the molecular descriptors dimensions with the experimental parameters ones will further improve the estimations and allows for accurate RT predictions of untested compounds.

Table 3.4: Comparison of Different Proposed RT Prediction Models of Different Chromatographic Methods.

Algorithm	Chromatographic Methods Generalizability	Compounds Size	Molecular Descriptor	Number of samples/Training R^2	Number of samples/Validation R^2	Number of samples/Testing R^2	Reference
LM-ANN	Limited to 1 GC	64	3 Cheminformatics	49/0.992	9/0.999	6/0.998	2011 [10]
GRNN	Limited to 1 LC	166	15 SMILES	120/0.943	16/0.853	30/0.922	2015 [11]
ANN	Limited to 1 LC	315	58 SMILES	252/0.978 (Adjusted)	-	63/0.967 (Adjusted) ²	2023 [22]
GNN	Limited to 44 LCs	10,500	19 to 324 SMILES	9450/ 0.950	-	2550/0.91, 0.75 ³	2025 [23]
ABSVR	Unlimited to Capillary GCs	61	Order of Elution & MW	426/0.993	92/0.993	91/0.992	This work

3.6 Using the Model for GC Optimization of Xylene Isomers Separation

One more advantage of the proposed model is that it can be used for GC optimization for separating hardly partitioned compounds such as Xylene isomers, thanks to the use of both a compound ID and the MW as molecular descriptors, which allowed the model to sense the difference in RT for equal MW compounds. To test this merit, we did GC screening to find the best GC for separating compounds number 43 and 44, namely, m, p-Xylene, which is barely separated by the reference GC (Only a 2.4-second retention gap). The results showed that a GC experiment with 105 m column length, 0.25 mm column diameter, 40 °C first temperature, 8 minutes first holding period, 1 °C /minute first heating rate, 150 °C second temperature, no second holding, 1 °C /minute second heating rate, 200 °C third temperature, no third holding, 0.5 mL/min carrier gas flow rate, 0.25 stationary phase thickness, and low stationary phase polarity would be the best settings for Xylene

¹ The authors reported an adjusted coefficient of determination that penalizes the model complexity.

² The authors tested their model on two parts of their data, one part with more compounds per subset, which showed higher accuracy, and the other on subsets with fewer compounds, which showed lower accuracy.

isomers separation with a separation gap of around 1.24 minutes. Interestingly, these results agree to a high degree with the outcomes of one of the most m, p-Xylene GC optimization studies, where they suggested a non-polar to low-polarity stationary phase, 0.4 carrier gas flow rate, two-step temperature program with 40 °C initial temperature, 1 minute holding time, 60 °C final temperature, 1 °C /minute heating rate, 0.25 µm stationary phase thickness as an optimum settings for the Xylene isomers partitioning [41]. The column length and diameter were not optimized, and a predetermined (5 m × 0.25 mm) column was chosen, in which the 5 m column helped in reducing the process duration to less than 6 minutes. The reported resolution for these settings is 3.04 with around 1.1 minute RT difference, which is quite close to the predicted model's value, and the slight positive difference can be attributed to the optimization of the column length and the third temperature step. With these outcomes, weeks of work can be done in one or two days of screening over a set of possible combinations of 13 different experimental features.

3.7 QLSR Evaluation and Formula

QLSR is used to obtain a mathematical formula that relates the RT shifts to the experimental parameters. The formula is highly complex and is reported in the supplementary material. The resulting model showed a considerable drop in performance when compared to the ABSVR, but maintained strong generalizability with training and testing R^2 of 0.948 and 0.946, respectively. With an MAE of 2.54 and an RMSE of 3.45 minutes, the model's usability in GC optimization might be limited due to possible magnified prediction by opposite errors for each compound's RT estimation. The mentioned metrics are obtained when there are no restrictions imposed on the complexity of the formula. However, this reflected on the formula shape, which was a hardly solvable and extremely lengthy expression. After restricting the complexity using the "bic", the R^2 dropped to 0.935 for training and 0.934 for testing. The produced formula this time is considerably solvable:

$$\begin{aligned} RTshifts = & 60.0 \tanh(\exp(-0.81(1 - 0.17Order)^2 - 2.0 \exp(-3.7(1 - 0.33 \cdot 1Temp)^2 \\ & (0.23 \cdot 1IncrRate/min + (1.4 \cdot 2IncrRate/min - 1.1) \exp(-3.5(-1Temp - \\ & 0.42)^2 - 2.3(2Hold - 0.62)^2 (0.62Order - 1))^2 - \\ & 0.26)^2 - 0.93(-0.061StaThick(micrometer) - 1)^2) \log((1.9Order + 5.0)^2))) - 41.0 \end{aligned}$$

As can be noticed, the formula ignored some features and mostly focused on the most important features that show more continuous distribution, such as the first temperature, the heating rates, and the Order. Other variables, namely the second holding time, might be added to the model to add some bias and improve the generalizability.

Chapter 4

Conclusion and Recommendations

In addition to the QSLR, six different supervised ML algorithms were trained to predict the shifts in RT, resulting from changing the experimental parameters of the reference GC. These shifts are later transformed into RT values of different compounds and various GC experiments. The use of the order of elution from the reference GC as the compound's ID proved to be efficiently informative to the model and allowed for GC optimization for isomers separation. The best model is the ABSVR, which offered high accuracy and consistency across the testing and validation sets, outperforming tree-based ensemble methods. The model is further tested to predict the best settings for m,p-Xylene separation, the results showed considerable matching with a recently proposed capillary GC for this application. Adding more compounds to the workflow, more GCs, and utilizing deep learning algorithms is highly recommended for future work. Moreover, providing the model with more details about the stationary phase, in addition to thickness and polarity, might also boost the performance. Such details include the porosity and the composition.

References:

- [1] K. K. A. Bissada, J. Tan, E. Szymczyk, M. Darnell, M. Mei, and J. Zhou, "Group-type characterization of crude oil and bitumen. Part I: Enhanced separation and quantification of saturates, aromatics, resins and asphaltenes (SARA)," *Org Geochem*, vol. 95, pp. 21–28, May 2016, doi: 10.1016/j.orggeochem.2016.02.007.
- [2] F. Bernardoni, H. M. Halsey, R. Hartman, T. Nowak, and E. L. Regalado, "Generic gas chromatography flame ionization detection method using hydrogen as the carrier gas for the analysis of solvents in pharmaceuticals," *J Pharm Biomed Anal*, vol. 165, pp. 366–373, Feb. 2019, doi: 10.1016/j.jpba.2018.12.006.
- [3] L. M. Blumberg, "Theory of gas chromatography," in *Gas Chromatography*, Elsevier, 2021, pp. 19–97. doi: 10.1016/B978-0-12-820675-1.00026-5.
- [4] H Van Den Dool and PD Kratz - J, "A generalization of the retention index system head space-solid phase Micro extraction-Gas chromatography," *Chromatogr*, 1963.
- [5] Y. Kalambet, "Data acquisition and integration," in *Gas Chromatography*, Elsevier, 2021, pp. 505–524. doi: 10.1016/B978-0-12-820675-1.00038-1.
- [6] E. Kovats, "Gas-chromatographische charakterisierung organischer verbindungen. teil 1: Retentionsindices aliphatischer halogenide, alkohole, aldehyde und ketone," *Helv. Chim. Acta*, vol. 41, no. 7, pp. 1915–1932, 1958.
- [7] B. d'Acampora Zellner, C. Bicchi, P. Dugo, P. Rubiolo, G. Dugo, and L. Mondello, "Linear retention indices in gas chromatographic analysis: a review," *Flavour Fragr J*, vol. 23, no. 5, pp. 297–314, Sep. 2008, doi: 10.1002/ffj.1887.
- [8] D. D. Matyushin and A. K. Buryak, "Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning," *IEEE Access*, vol. 8, pp. 223140–223155, 2020, doi: 10.1109/ACCESS.2020.3045047.
- [9] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.
- [10] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (3rd ed.)*, 3rd ed. O'Reilly Media, 2023.
- [11] E. Alpaydin, *Introduction to Machine Learning*, 4th ed. Cambridge, MA: MIT Pres, 2020.
- [12] K. R. Broløs *et al.*, "An Approach to Symbolic Regression Using Feyn," Apr. 2021, doi: 10.48550/arXiv.2104.05417.
- [13] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, Dec. 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

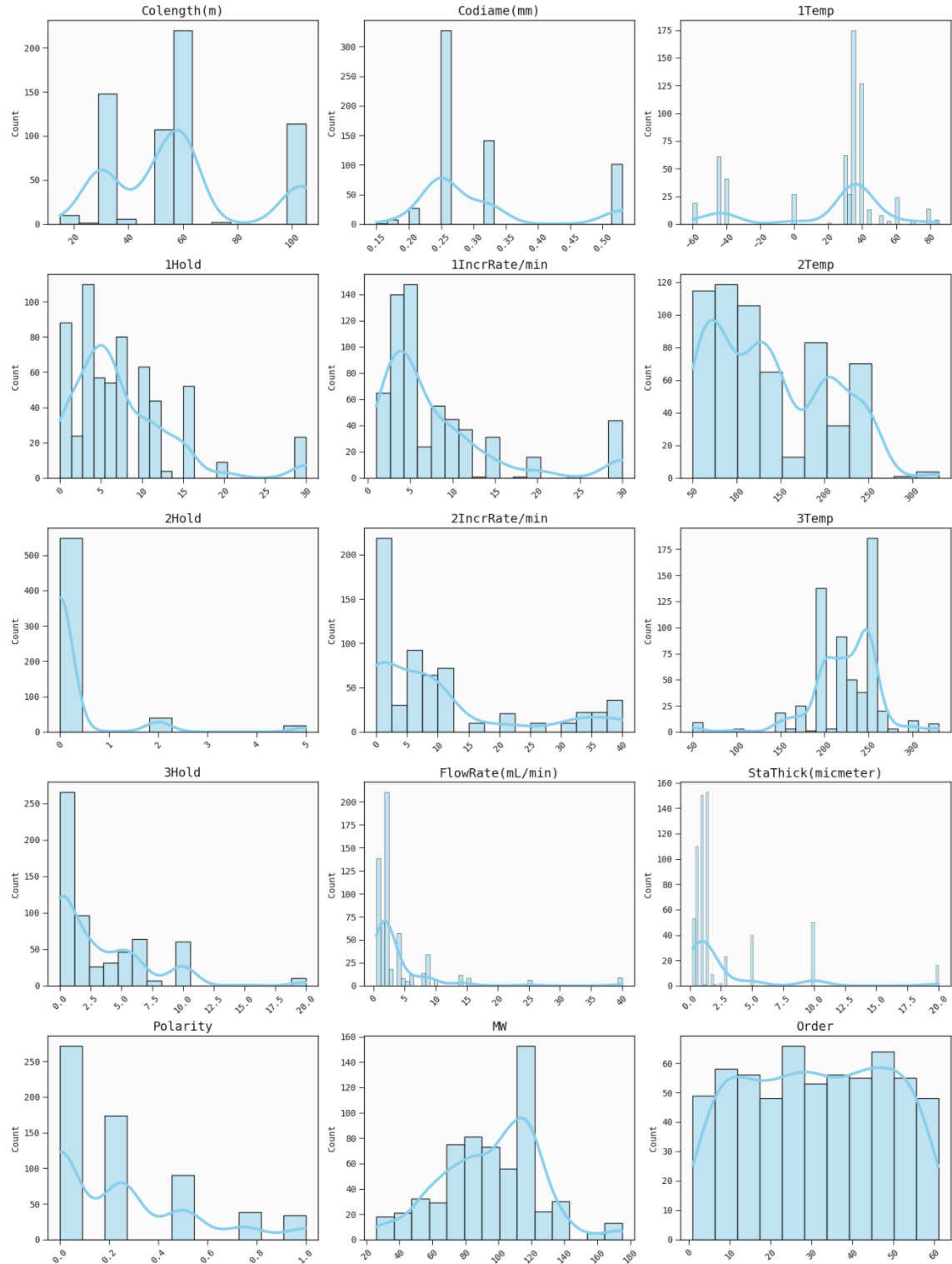
- [14] Z. Dong *et al.*, "Development of a Predictive Model for Carbon Dioxide Corrosion Rate and Severity Based on Machine Learning Algorithms," *Materials*, vol. 17, no. 16, p. 4046, Aug. 2024, doi: 10.3390/ma17164046.
- [15] J. Hou *et al.*, "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, no. 1, p. 81, Dec. 2022, doi: 10.1186/s12859-022-04609-x.
- [16] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, Dec. 2007, doi: 10.1214/0090536070000000505.
- [17] Sebastian Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," 2018.
- [18] "How Much Does A Gas Chromatograph Cost?," axionlabs. Accessed: Jan. 23, 2025. [Online]. Available: <https://axionlabs.com/chromatography-training/how-much-does-a-gas-chromatograph-cost/>
- [19] F. , V. G. , G. A. , M. V. , T. B. , G. O. , B. M. , P. P. , W. R. , D. V. , V. J. , P. A. , C. D. , B. M. , P. M. and D. E. Pedregosa, "Scikit-learn: Machine Learning in {P}ython," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [20] K. Héberger, "Quantitative structure–(chromatographic) retention relationships," *J Chromatogr A*, vol. 1158, no. 1–2, pp. 273–305, Jul. 2007, doi: 10.1016/j.chroma.2007.03.108.
- [21] K. Munro, T. H. Miller, C. P. B. Martins, A. M. Edge, D. A. Cowan, and L. P. Barron, "Artificial neural network modelling of pharmaceutical residue retention times in wastewater extracts using gradient liquid chromatography-high resolution mass spectrometry data," *J Chromatogr A*, vol. 1396, pp. 34–44, May 2015, doi: 10.1016/j.chroma.2015.03.063.
- [22] E. S. Fedorova, D. D. Matyushin, I. V. Plyushchenko, A. N. Stavrianidi, and A. K. Buryak, "Deep learning for retention time prediction in reversed-phase liquid chromatography," *J Chromatogr A*, vol. 1664, p. 462792, Feb. 2022, doi: 10.1016/j.chroma.2021.462792.
- [23] E. Choi, W. J. Yoo, H.-Y. Jang, T.-Y. Kim, S. K. Lee, and H. Bin Oh, "Machine learning liquid chromatography retention time prediction model augments the dansylation strategy for metabolite analysis of urine samples," *J Chromatogr A*, vol. 1705, p. 464167, Aug. 2023, doi: 10.1016/j.chroma.2023.464167.
- [24] C. Veenaas and P. Haglund, "A retention index system for comprehensive two-dimensional gas chromatography using polyethylene glycols," *J Chromatogr A*, vol. 1536, pp. 67–74, Feb. 2018, doi: 10.1016/j.chroma.2017.08.062.
- [25] C. Veenaas, A. Linusson, and P. Haglund, "Retention-time prediction in comprehensive two-dimensional gas chromatography to aid identification of unknown contaminants," *Anal Bioanal Chem*, vol. 410, no. 30, pp. 7931–7941, Dec. 2018, doi: 10.1007/s00216-018-1415-x.

- [26] V. K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, and S. Agarwal, "Prediction of capillary gas chromatographic retention times of fatty acid methyl esters in human blood using MLR, PLS and back-propagation artificial neural networks," *Talanta*, vol. 83, no. 3, pp. 1014–1022, Jan. 2011, doi: 10.1016/j.talanta.2010.11.017.
- [27] D. D. Matyushin, A. Yu. Sholokhova, and A. K. Buryak, "A deep convolutional neural network for the estimation of gas chromatographic retention indices," *J Chromatogr A*, vol. 1607, p. 460395, Dec. 2019, doi: 10.1016/j.chroma.2019.460395.
- [28] T. Vrzal, M. Malečková, and J. Olšovská, "DeepRel: Deep learning-based gas chromatographic retention index predictor," *Anal Chim Acta*, vol. 1147, pp. 64–71, Feb. 2021, doi: 10.1016/j.aca.2020.12.043.
- [29] A. Anjum, J. Liigand, R. Milford, V. Gautam, and D. S. Wishart, "Accurate prediction of isothermal gas chromatographic Kováts retention indices," *J Chromatogr A*, vol. 1705, p. 464176, Aug. 2023, doi: 10.1016/j.chroma.2023.464176.
- [30] Y. Fan *et al.*, "Deep learning-based method for automatic resolution of gas chromatography-mass spectrometry data from complex samples," *J Chromatogr A*, vol. 1690, p. 463768, Feb. 2023, doi: 10.1016/j.chroma.2022.463768.
- [31] A. Yu. Sholokhova, D. D. Matyushin, and M. V. Shashkov, "Quantitative structure-retention relationships for pyridinium-based ionic liquids used as gas chromatographic stationary phases: convenient software and assessment of reliability of the results," *J Chromatogr A*, vol. 1730, p. 465144, Aug. 2024, doi: 10.1016/j.chroma.2024.465144.
- [32] C. Qu, B. I. Schneider, A. J. Kearsley, W. Keyrouz, and T. C. Allison, "Predicting Kováts Retention Indices Using Graph Neural Networks," *J Chromatogr A*, vol. 1646, p. 462100, Jun. 2021, doi: 10.1016/j.chroma.2021.462100.
- [33] "National Institute of Standards and Technology (NIST). 2023 Release of the NIST EI and Tandem Libraries [Announcement]." Accessed: Mar. 27, 2025. [Online]. Available: https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:asms2023:asms_2023
- [34] M. Zakir, M. A. LeVatte, and D. S. Wishart, "RT-Pred: A web server for accurate, customized liquid chromatography retention time prediction of chemicals," *J Chromatogr A*, vol. 1747, p. 465816, Apr. 2025, doi: 10.1016/j.chroma.2025.465816.
- [35] M. Taniguchi *et al.*, "Optimal machine learning algorithm for prediction model for retention times of plant toxins," *Food Control*, vol. 174, p. 111251, Aug. 2025, doi: 10.1016/j.foodcont.2025.111251.
- [36] J. Samuelsson *et al.*, "Improved workflow for constructing machine learning models: Predicting retention times and peak widths in oligonucleotide separation," *J Chromatogr A*, vol. 1747, p. 465746, Apr. 2025, doi: 10.1016/j.chroma.2025.465746.
- [37] P. C. Austin and L. J. Brunner, "Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses," *Stat Med*, vol. 23, no. 7, pp. 1159–1178, Apr. 2004, doi: 10.1002/sim.1687.

- [38] R. L. White, "Fast GC with a Small Volume Column Oven and Low Power Heater," *Chromatographia*, vol. 69, no. 1–2, pp. 129–132, Jan. 2009, doi: 10.1365/s10337-008-0850-9.
- [39] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim Res*, vol. 30, pp. 79–82, 2005, doi: 10.3354/cr030079.
- [40] K. Krisnangkura and V. Pongtonkulpanich, "Prediction of gas chromatographic peak width in capillary columns at different temperatures, carrier gas flows, column lengths, inside diameters and carbon numbers," *J Sep Sci*, vol. 29, no. 1, pp. 81–89, Jan. 2006, doi: 10.1002/jssc.200500141.
- [41] T. Sun *et al.*, "Efficient gas chromatographic separation of xylene and other aromatic isomers by using pillar[6]arene-based stationary phase," *Anal Chim Acta*, vol. 1251, p. 340979, Apr. 2023, doi: 10.1016/j.aca.2023.340979.

Appendices:

Appendix 1: Variables Distributions



Appendix 2: Algorithms' best settings:

Algorithm	Hyperparameters random_state=2025
KNN	n_neighbors=2, weights="distance"
RF	n_estimators= 8, max_depth = 12, max_features=6, bootstrap=False
GBDT	n_estimators = 500, max_depth=20, learning_rate=0.3, max_features=5, loss="huber"
ABDT	n_estimators=50, max_depth=30, max_features=12, learning_rate= 0.1, splitter="random"
XGB	n_estimators= 500, depth=9, lr=0.2, subsample=0.5
ABSVR	C=10000, epsilon= 0.0001, gamma=0.12, learning_rate=0.0001, n_estimators= 100
QLSR	n_epochs=200, threads=48, max_complexity=30, kind="regression", criterion=None

Vitae

Name	:Hussein Neymer AlAdwan
Nationality	:Jordanian
Date of Birth	:9/14/1998
Email	:huss.aladwan@gmail.com
Current address	:Room 224, Building 814, Student housing, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
Academic Background	:B.Sc. in Materials Engineering, June 2022, BAU, Jordan. M.Sc. in Materials Science and Engineering, June 2025, KFUPM, Saudi Arabia.
Reference	: Dr. Abduljabar Al-Sayoud, sayoudaq@kfupm.edu.sa