University of Liège
Faculty of Medicine
Department of Pharmacy

Different approaches of Quantitative Structure Retention Relationship of small molecules in Liquid Chromatography (QSRR)

Author:
Priyanka Kumari

Supervisor:
Prof. Phillipe Hubert
Co-supervisor:
Dr. Cedric Hubert

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Biomedical and Pharmaceutical Sciences
in the
Laboratory of Pharmaceutical and Analytical Chemistry,
Center for Interdisciplinary Research on Medicines
Department of Pharmacy

Academic Year 2023 - 2024

© 2024 Priyanka Kumari. All rights reserved.

Members of the jury:

Prof. Marianne FILLET (University of Liège, Belgium)	**President**
Dr. Pierre Yves SACRE (University of Liège, LCAP)	**Secretary**
Prof. Eric ZIEMONS (University of Liège, LCAP)	
Prof. Pierre GEURTS (University of Liège, Institut Montefiore)	
Prof. Bruno BOULANGER (Cencora PharmaLex)	**External Member**
Dr. Julien BOCCARD (University of Geneva)	**External Member**
Dr. Cedric HUBERT (University of Liège, LCAP)	**Co-promoter**
Prof. Phillipe HUBERT (University of Liège, LCAP)	**Promoter**

"The woods are lovely, dark and deep, But I have promises to keep, And miles to go before I sleep, And miles to go before I sleep."

Robert Frost, (1874 – 1963)

This thesis is the Dedication to the memory of my father Prof. Bashikant Choudhary.

It was just the beginning of my Ph.D. journey when he left this world, ascending to the heavens above. His absence during those days weighed heavily on my heart. I missed the endless discussions we used to share, discussions that had a unique power to relax from the stresses of the day. It was a time of deep solitude, yet it was his constant motivation, encouragement, and firm belief in me that became the guiding light. His memory and enduring support have been a source of strength, propelling me forward in both my academic pursuits and the journey of life itself.

Acknowledgement

After four years of challenging yet rewarding Ph.D. studies, I emerge transformed and motivated, always striving to get better. There are many people whom I wish to acknowledge for their contributions.
First and foremost I want to thank my supervisor, Professor Phillipe Hubert, and co-supervisor, Dr Cedric Hubert, for giving me the opportunity and the resources to complete this thesis. Their mentorship has not only empowered me to navigate an independent path but also provided unwavering support during my moments of personal and professional need.

I am profoundly grateful to Dr. Pierre Yves Sacre for his invaluable guidance and support. His time, ideas, suggestions, and feedback have significantly contributed to the completion and improvement of my research. Alongside the research pursuits, I have much to learn from his management skills, and his remarkable ability to oversee multitude of projects effectively.

I would like to express my gratitude to my thesis committee members for their annual suggestions and guidance throughout my PhD progress. I am thankful to the jury members for their willingness to assess my work. The Belgium National Fund for scientific Research(FNRS) and Excellence of Science(EOS) are gratefully acknowledged for the research grant making this thesis possible.

Thanks to all members of the LPAC for their invaluable support and the positive atmosphere in the lab. I would like to acknowledge Thomas Van Laethem who offered valuable insights into chemical data and was always open to fruitful discussions while working on the EOS project. Thanks to other past group members and the numerous summer interns who have come through the lab, with whom I have had the pleasure of discussions and exchange of ideas. I am immensely thankful to Murielle Bihain for her invaluable assistance in navigating the labyrinth of administrative tasks.

Special thanks to Prof. Kristel Van Steen and Prof. Meyer. Achieving this stage wouldn't have been possible without the past research opportunities provided by them. I would like to thank Dr. Abhigyan Nath for introducing me to the Machine learning field during my master's research.

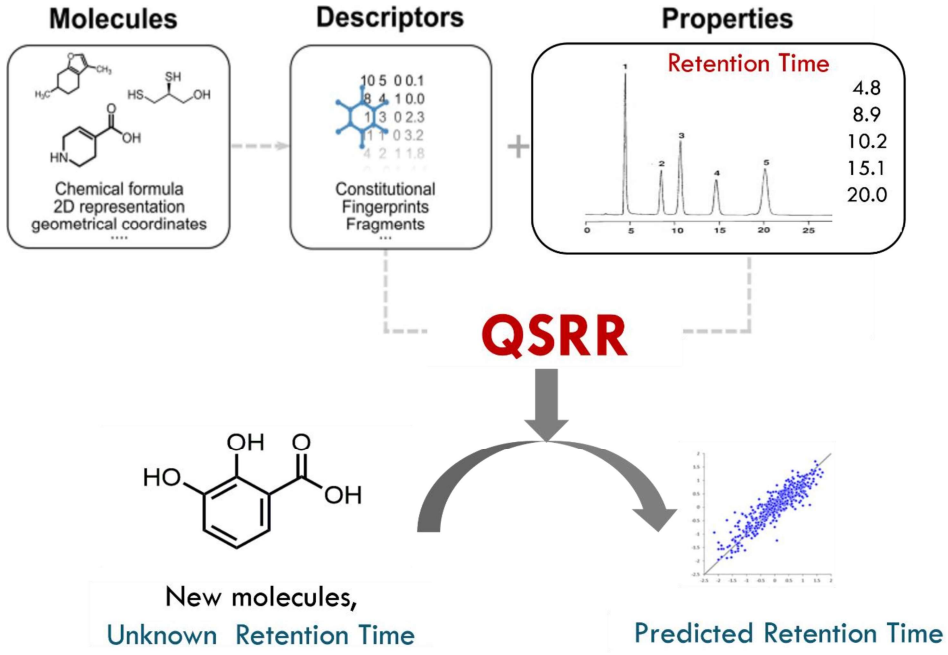
My heartfelt thanks go out to Tant Anna & Tonton Julien and my friends Ragi,Arpita, Archana, Naman, Ratish, Navdeep, and Bonny for creating an environment in Liege that felt like a second home. Thank you Bharti for your assistance in editing my thesis and for always being ready to lend a helping hand.
This thesis wouldn't have been possible without my friend Shivalika, whose strong support has been a reliable anchor. I thank her for the encouragements and patiently listening to all my rants and frustrations.
I would like to thank my friends Kamshat & Henry for their valuable suggestions and constant support. It's because of them that my time in Europe became significantly enjoyable, fun, and stress-free.Additionally, I extend my heartfelt gratitude to them and Kinjal for their invaluable assistance and care during the unfortunate flooding in Liege.

My family has always been strong pillar of strength in everything I do in life. I want to convey my deep gratitude and love to my siblings, Bhaiya (Piyush Choudhary) and Chote Bhaiya (Pratyush Choudhary). Their steadfast support has been a constant companion through the highs and lows of life, guiding me along the path of Ph.D. They have truly been my rock, providing enduring strength in all my personal and professional endeavors.
I am truly grateful for the warmth and encouragement brought into my life by my sisters-in-laws, Neelam and Ria. I consider myself incredibly fortunate to have received their help and support since they became part of our family.
I must also express my appreciation for my seven years old nephew, Pragnay Vatsa, whose infectious smile and witty remarks on my PhD progress always lifted my spirits and motivated me to keep going.I express my sincere gratitude to my late grandmother, whose exemplary resilience and determination greatly influenced me since childhood, instilling in me the values of hard work and excellence.

My dearest mother Lata Choudhary, thank you for always believing in me. your unconditional support and encouragements have been the cornerstone of all my pursuit, without which I could never have envisaged attaining this milestone. You are my strength, and my inspiration!

LABORATORY OF
ANALYTICAL CHEMISTRY
Prof. Phillipe HUBERT
Dr. Cedric HUBERT

Different approaches of Quantitative Structure Retention Relationship of small molecules in Liquid Chromatography (QSRR)



PhD Thesis - 2024

Priyanka Kumari

Thesis submitted to obtain the degree of Doctor of Philosophy in Biomedical and Pharmaceutical Sciences

Thesis abstract

This PhD dissertation explores various Quantitative Structure-Retention Relationship (QSRR) modelling approaches to enhance the method development process in analytical chemistry. By establishing a predictive framework that relates the chemical structure of analytes to their chromatographic retention behaviours, this approach aims to minimize experimental efforts and increase efficiency in developing robust chromatographic methods for pharmaceutical compound separation and analysis. For instance, instead of empirically testing a broad range of conditions, QSRR models enable the prediction of analytes' retention behaviour under diverse experimental conditions based on their molecular structure. This method can significantly decrease the necessity for experimental trials by focusing efforts on conditions most likely to enhance separation. In addressing the challenges that analytical chemists encounter, particularly in retention prediction modelling with varying data availability, this study sets out to bridge the gap in the field. These challenges range from determining a starting point in situations of data scarcity to selecting the optimal modelling strategy when faced with large datasets ready for model training. Further complexities arise in choosing the appropriate modelling approach as experimental variations expand and the nature of the dataset evolves. Recognizing the absence of a clear, definitive strategy for QSRR modelling, this study began with Single Target Retention Prediction Modelling. A detailed QSRR strategy was developed, incorporating a wide range of methods for selecting descriptors and utilizing a variety of regression algorithms, including linear, non-linear, parametric, non-parametric, and ensemble methods, all developed to predict retention times across different pH conditions in Reversed-Phase Liquid Chromatography (RPLC). Each condition, referred to as a target, was analysed individually. By implementing this comprehensive QSRR approach, the study aims to systematically tackle the aforementioned challenges, thereby setting a foundation for future progress in this

area. After exploring single-target QSRR, the research progressed to Multitarget QSRR modelling. This phase compared the accuracy of retention predictions using two different approaches: one that creates separate models for each condition (single-target) and another that uses a unified model to predict retention times across all conditions simultaneously (multitarget). The goal was to find a more efficient way to model multiple target properties at once, potentially making the process quicker and more compact. This has significant implications for improving chromatographic separation methods, offering analytical chemists a valuable tool in their method development efforts. The thesis advances QSRR modelling by incorporating Transfer Learning, to investigate enhancements in both accuracy and model efficiency, particularly when data is scarce. This research delved into employing both physicochemical properties and image-based features of small molecules for QSRR modelling using techniques emerging from advanced Artificial Intelligence, aiming to broaden the methodological framework and improve predictive capabilities. In summary, this thesis offers valuable insights and tools for pharmaceutical research and development. By integrating computational modelling with RPLC, it introduces a systematic approach and various potential strategies for analytical chemists to explore, aiming to predict small molecule separation. This could ultimately lead to optimized compound separation with reduced time and cost expenditures.

Résumé

Cette thèse de doctorat explore diverses approches de modélisation de la Relation Quantitative Structure-Rétention (QSRR) afin d'améliorer le processus de développement de méthodes en chromatographie liquide. En établissant un cadre prédictif qui relie la structure chimique des analytes à leur rétention chromatographique, cette approche vise à minimiser les efforts expérimentaux et à augmenter l'efficacité dans le développement de méthodes chromatographiques robustes pour la séparation et l'analyse de composés pharmaceutiques.

Cette thèse a pour objectif de répondre à divers défis auxquels les scientifiques sont confrontés lors du développement de modèles QSRR. Un premier défi est le choix de l'approche de modélisation appropriée en fonction de la taille de la base de donées disponible et de la nature des données constituant celui-ci. Reconnaissant l'absence d'une stratégie claire et définitive pour la modélisation QSRR, cette étude a commencé par l'établissement d'une stratégie QSRR générique pour la modélisation de la prédiction de rétention dans le cas d'une condition chromatographique unique. Cette stratégie incorpore un large éventail de méthodes pour la sélection des variables et l'utilisation de divers algorithmes de régression, y compris linéaires, non linéaires, paramétriques, non paramétriques et des méthodes d'ensemble, tous développés pour prédire les temps de rétention dans différentes conditions de pH en chromatographie liquide en phase inverse (RPLC). Chaque condition de pH, a été analysée individuellement. En mettant en œuvre cette approche QSRR complète, l'étude vise à aborder systématiquement les défis mentionnés précédemment, jetant ainsi les bases pour des progrès futurs dans ce domaine.

Après avoir exploré le QSRR pour des conditions uniques, la recherche a progressé vers la modélisation QSRR multi-conditions. Cette phase a comparé l'exactitude des prédictions de rétention en utilisant deux approches différentes : l'une qui crée des modèles séparés pour chaque condition (cible unique) et une autre qui utilise un modèle unifié pour prédire les

temps de rétention pour toutes les conditions simultanément (multicibles). L'objectif était de trouver une manière plus efficace de modéliser la rétention à différentes conditions simultanément, rendant potentiellement le processus plus rapide et plus performant. Ceci a des implications significatives pour l'amélioration des méthodes de séparation chromatographique, offrant aux analystes un outil précieux dans leurs efforts de développement de méthode. La modélisation multicible a commencé par l'utilisation d'approches conventionnelles puis a investigué l'intérêt de l'Apprentissage par Transfert en termes de précision et d'exactitude des prédictions, particulièrement lorsque les données sont peu nombreuse. Cette partie du travail s'est penché sur l'utilisation des propriétés physicochimiques et des caractéristiques basées sur l'image de petites molécules pour la modélisation QSRR en utilisant des techniques issues de l'Intelligence Artificielle avancée, visant à élargir le cadre méthodologique et à améliorer les capacités prédictives. En résumé, cette thèse offre des outils pour la recherche et le développement pharmaceutiques. En intégrant la modélisation computationnelle avec la RPLC, elle introduit une approche systématique et diverses stratégies potentielles pour prédire la séparation de petites molécules. L'objectif final est une séparation optimisée des composés tout en réduisant les dépenses en termes de coût et de temps.

Abbreviations

Table 1: List of abbreviations

Abbreviation	Definition
AB	Adaptive Boosting
ABC-PLS	Artificial Bee Colony Partial Least Squares
ACN_Data	ACN_Data dataset name
AD	Applicability Domain
AI	Artificial Intelligence
ANN	Artificial Neural Network
BRR	Bayesian Ridge Regression
CFS	Correlation-based Feature Selection
CNN	Convolutional Neural Network
CV	Cross-validation
DNN	Deep Neural Network
DoE	Design of Experiments
ECFP	Extended Connectivity Fingerprints
FA-PLS	Firefly Algorithm Partial Least Squares
FPA-PLS	Flower Pollination Algorithm Partial Least Squares
GA	Genetic Algorithms
GA-PLS	Genetic Algorithm-Partial Least Squares
GB	Gradient Boosting
GBR	Gradient Boosted Regression
GCN	Graph Convolutional Network
GNN	Graph Neural Network
iPLS	Interval Partial Least Squares
KFold	k-Fold Cross-Validation
LASSO	Least Absolute Shrinkage and Selection Operator
LC	Liquid Chromatography
LFER	Linear Free-Energy Relationship
LOO	Leave-One-Out (cross-validation)
LSS	Linear Solvent Strength model
LSER	Linear Solvation Energy Relationship
MAPE/MRE	Mean Absolute Percentage Error/Mean Relative Error
MCDA	Multi Criteria Decision Analysis
MD	Molecular Descriptor
MIA	Multivariate Image Analysis
ML	Machine Learning

Abbreviation	Definition
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MTL	Multi Task Learning
MT QSRR	Multi-Target QSRR
MTP	Multi-Target Prediction
PLS	Partial Least Squares method
PSO-PLS	Particle Swarm Optimization Partial Least Squares
QbD	Quality by Design
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
QSRR	Quantitative Structure-Retention Relationship
QSTR	Quantitative Structure-Toxicity Relationship
R2	Coefficient of Correlation
RC	Regressor Chain
RF	Random Forest
RFE	Recursive Feature Elimination
RFR	Random Forest Regressor
RGCN	Relational Graph Convolutional Network
RIKEN	RIKEN dataset name
RMSE	Root Mean Squared Error
RPLC	Reversed-Phase Liquid Chromatography
ReLU	Rectified Linear Unit
SELFIES	Self-referencing Embedded Strings
SHAP	SHapley Additive exPlanations
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SMIRKS	SMILES Reaction Transform Language
SMRT	SMRT dataset name
ST QSRR	Single-Target QSRR
STP	Single-Target Prediction
SVR	Support Vector Regression
SYBYL Line Notation	A chemical structure representation format

Abbreviation	Definition
TB	Transformation-Based
TL	Transfer Learning
tR	Retention Time
UVE	Uninformative Variable Elimination
2D-QSAR	Two-Dimensional Quantitative Structure-Activity Relationship
3D-QSAR	Three-Dimensional Quantitative Structure-Activity Relationship

Contents

List of Tables

List of Figures

1 Introduction

1.1 Preamble

This thesis introduces a comprehensive Quantitative structure retention relationship (QSRR) studies by explaining first the QS(X)R, the umbrella term that includes QSRR. This initial discussion connects QSRR to other related concepts. The content is structured into three main parts of QSRR: 1. Structure (S): Discuss the structural properties of the compounds, including how these structures are represented and the various structural descriptors used. 2. Retention (R): Focus on the property to be predicted, namely the retention time, explaining the calculation methods, data collection processes, and other relevant considerations. 3. Relationship (R): Outlines the core of QSRR modeling, covering various methods, including state-of-the-art techniques and workflow detailing every individual steps of QSRR modelling.

1.2 Laying the Foundation of QSRR: QS(X)R

Quantitative structure-(X)property relationship (QSXR) modelling is a scientific approach that involves establishing mathematical correlations between the chemical response of structurally related compounds and quantitative chemical attributes that define their molecular features [2]. This approach aims to develop a formalism that describes the behaviour of chemicals in terms of their physicochemical properties, biological activity, toxicity, or retention data. By leveraging the relationship between the structural characteristics of chemical compounds and their properties, QS(X)R, which stands for Quantitative Structure-(X)-Relationship modelling (Figure 1.1), provides valuable insights into the underlying mechanisms governing chemical behaviour, where 'X' represents a property of interest such as activity, retention, toxicity, etc. The specific nomenclature used for modelling depends on the nature of the response being modelled. For instance, in the quantitative structure-activity relationship (QSAR), the property(X) of interest is activity. In the quantitative structure-toxicity relationship (QSTR), 'X' signifies toxicity, and in the quantitative structure-retention relationship (QSRR), the focus is on retention time(X).

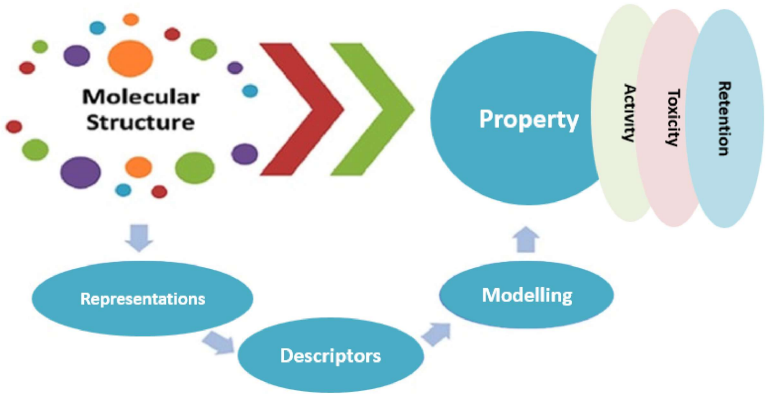


Figure 1.1: A simple schematic overview of QS(X)R

In QS(X)R, mathematical relationships are developed to predict the behaviour of molecules, including new chemicals or hypothetical molecules[15]. The basic formalism of QS(X)R can be mathematically represented as follows:

$$Property\ of\ interest = f(chemical\ attributes)$$

The term *"chemical attributes"* pertains to the characteristics defining the observable behaviour or response (such as activity, toxicity, or retention

time) of the studied chemical compounds. These attributes, also called descriptors, are quantitative information about the chemistry of molecules that can be obtained through experimental analysis or theoretical algorithms, and the responses, also known as targets, are obtained from experiments. QSRR is a method used to understand how changes in one or sometimes more responses (known as Y-variables) can be linked to changes in various factors (called X-variables), with the goal of predicting or explaining these outcomes. The Y-variables are usually dependent on the X-variables (descriptors), which are independent. This technique allows the retention prediction of novel, not yet synthesized compounds, solely from their structural descriptors[16].

1.3 Goals and Application of QSRR

The goal of Quantitative Structure-Retention Relationships (QSRR) in chromatography is to construct a robust predictive model by identifying molecular descriptors that strongly correlate with the retention behaviour of analytes(Figure 1.2). This model is built upon a thorough understanding of the physicochemical properties of molecules and the mechanisms governing their chromatographic separation. It's main goal is to reduce the time and

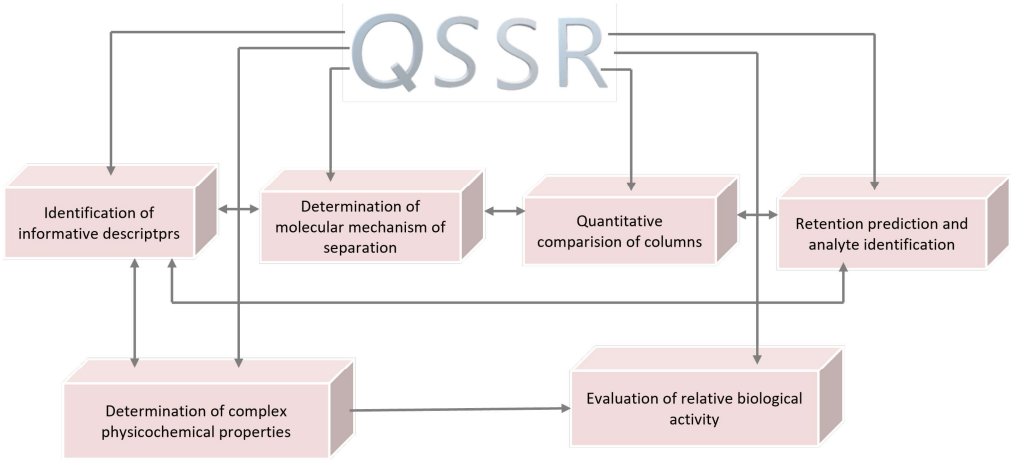


Figure 1.2: Goals of QSRR studies[2]

cost of chromatographic separations and method development. Moreover, the QSRR tool could represent an advantage considering the complexity of chromatographic separations, where several factors can influence the retention time of a compound, such as the type of column, the mobile phase, the temperature, the pressure, and the chemical structure of the compound

itself[17]. Ultimately, QSRR endeavours to enhance the efficiency of chromatographic analysis, facilitating the identification of unknown compounds in complex mixtures[18] and enhancing the throughput and reliability of analytical methods used in scientific research.

1.3.1 Applications of QSRR models

QSRR modelling finds numerous applications in the field of analytical chemistry, particularly in enhancing the efficiency and effectiveness of chromatographic processes. Some applications include:

- *Predicting Retention Times and Method development:* By retention time predictions of compounds, QSRR applications facilitate the optimization of chromatographic separations, significantly saving time and resources during method development. Furthermore, they aid in selecting optimal conditions for separating compounds of interest, including the choice of appropriate column and mobile phase composition, based on the chemical properties of the analytes[19, 20].

- *Analyte Identification:* By predicting retention times based on molecular structure, QSRR models assist in the identification of unknown compounds in complex mixtures. This is useful in various applications including environmental analysis, food safety, and toxicology studies where unknown compounds need to be identified[21, 22].

- *Chemical Property Estimation:* QSRR models can be used to estimate physicochemical properties of compounds, such as lipophilicity, that are important for drug absorption, distribution, metabolism, and excretion (ADME) studies[21]. This application is crucial in pharmaceutical research for predicting drug behaviour in the body.

- *Metabolomics and Proteomics:* In the study of metabolites and proteins, QSRR models facilitate the separation and analysis of complex biological samples, identification of products contributing to advances in life sciences and biomedical research[23, 24].

- *Drug Discovery:* QSRR models are instrumental in drug discovery processes, where they are used to predict the chromatographic behaviour of new chemical entities, thus speeding up the screening and development of potential drug candidates[25].

QSRR can be applied in other fields apart from pharmaceutical sciences such as:

- *Environmental Monitoring:* In environmental chemistry, QSRR models help predict the behaviour of pollutants in chromatographic systems, aiding in the detection and quantification of these substances in environmental samples[26].
- *Food Analysis:* QSRR modelling is used in food chemistry to identify and quantify food components and contaminants, ensuring food safety and quality.
- *Petrochemical analysis:* To predict the retention times of petroleum compounds in crude oil, allowing for more efficient refining processes and product development[27].

1.4 Fundamentals of QSRR

There are mainly three aspects to explore in the QSRR study:

1. Molecular structure or features
2. property to be predicted- Retention time;
3. relationship, the methodology or algorithms to be used for the prediction.

These aspects make up the building blocks of QSRR(Figure 1.3) which are described in more detail below.

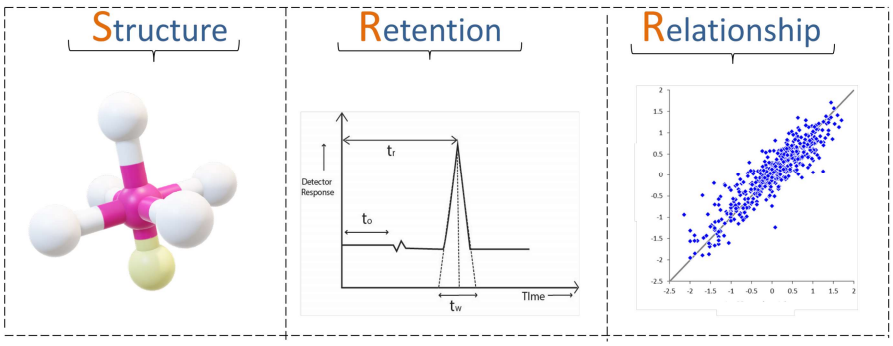


Figure 1.3: QSRR: Building blocks of study

1.4.1 Structure: The basic dogma of chemistry of compounds

QSRR is a powerful tool in computational chemistry that relates the structure-derived properties of molecules to their chromatographic retention behaviour. Such structure-derived properties are known as molecular descriptors. To derive such properties, it is important to represent molecules in a digital format that makes them machine-readable and facilitates retention predictions(Molecular representations). A few examples are shown in Figure 1.6 (a) and (b). There can be multiple formats that can be used: SMILES, SMARTS, SMIRKS, OpenSMILES, SYBYL Line Notation and recently, SELFIES[28]. SMILES is the most frequently used representation and is used in this thesis for the calculation of structure-derived features. These are Simplified Molecular Input Line Entry Systems (SMILES) which translate a chemical's three-dimensional structure into a string of symbols to make them understandable by computer software. However, descriptors derived from SMILES in QSRR studies lack 3D conformational information hence, potentially overlooking critical spatial and dynamic molecular interactions. For example, the SMILE structure of Ibuprofen is:
CC(C)CC1=CC=C(C=C1)C(C)C(=O)O
To understand this notation it is important to know the symbols:

- **Representing Atoms**
  Some examples of atomic symbols and their corresponding SMILES notations:
  - C: methane (CH_4)
  - O: water (H_2O)

  Usually, hydrogen is not shown in SMILES representations.

- **Representing Bonds**:
  - single –
  - double =
  - triple #

Normally single bonds and aromatic bonds do not need to be written in the SMILES notation. Branches are specified by enclosures in parentheses. Other rules are explained in detail in article [29].

Molecular Descriptors There are multiple types of molecular descriptors that can be used for QSRR studies(Figure 1.4), including physicochemical descriptors, categorization based on dimensions, fingerprints, graphs-based descriptors, image-based descriptors, and transformed descriptors such as principal components (PCs).

Physicochemical Descriptors Physicochemical descriptors provide information about the molecular properties of a compound's physical and chemical characteristics [3]. They can be further classified into the following subcategories:

Figure 1.4: Schematic diagram showing multiple descriptors[3]

- Constitutional Descriptors: Capture information about the molecular composition, such as the number of atoms, functional groups, and molecular weight.
- Geometrical Descriptors: Consider the molecular shape and size, including molecular volume, surface area, and molecular diameter.

- Topological Descriptors: Represent the connectivity and arrangement of atoms in a molecule, such as the number of bonds, branching, and cyclicity.
- Quantum Descriptors: Use quantum mechanical calculations to evaluate electronic properties, such as molecular orbitals, electron density, and polarizability.
- Thermodynamics Descriptors: Provide thermodynamic internations between a solute and the thermodynamic systems. Some common thermodynamics-related molecular descriptors for retention predictions include Gibbs Free Energy (ΔG), MolRef(Molar Refractivity, AlogP etc.

Categorization Based on Dimensions: Molecular descriptors can also be categorized based on their dimensions1.5, which represent different levels of structural complexity [4]:
0D Descriptors: Provide global properties of a molecule independent of its spatial arrangement. Examples include the number of atoms, molecular weight, and hydrogen bond acceptors/donors.
1D Descriptors: Capture sequential information, such as molecular connectivity, atom types, and bond types.
2D Descriptors: Consider the planar structure of molecules and include features like topological indices, molecular fingerprints(Described below), and fragment counts.
3D Descriptors: Incorporate three-dimensional information, including molecular shape, chirality, and conformational flexibility.

Fingerprints: Fingerprints are binary or bit-string representations that encode the presence or absence of specific molecular features. They are widely used for structural similarity searching and virtual screening. Common fingerprint types include circular fingerprints, MACCS keys, and extended connectivity fingerprints (ECFP) [30].
Graphs-Based Descriptors: Graph-based descriptors represent a molecule as a mathematical graph, where atoms are nodes and bonds are edges. These descriptors capture molecules' structural and topological features, including connectivity, cycles, and symmetry. Examples of graph-based descriptors are the Wiener index, Randic index, and molecular walk counts [31, 32].
Image-Based Descriptors: Image-based descriptors employ image-processing techniques to analyze molecular structures. They convert 2D chemical structures into grayscale or colour images and then extract relevant features using image descriptors like pixel intensity, texture, and shape. These descriptors can be useful for applications involving machine learning and deep learning in retention time predictions.

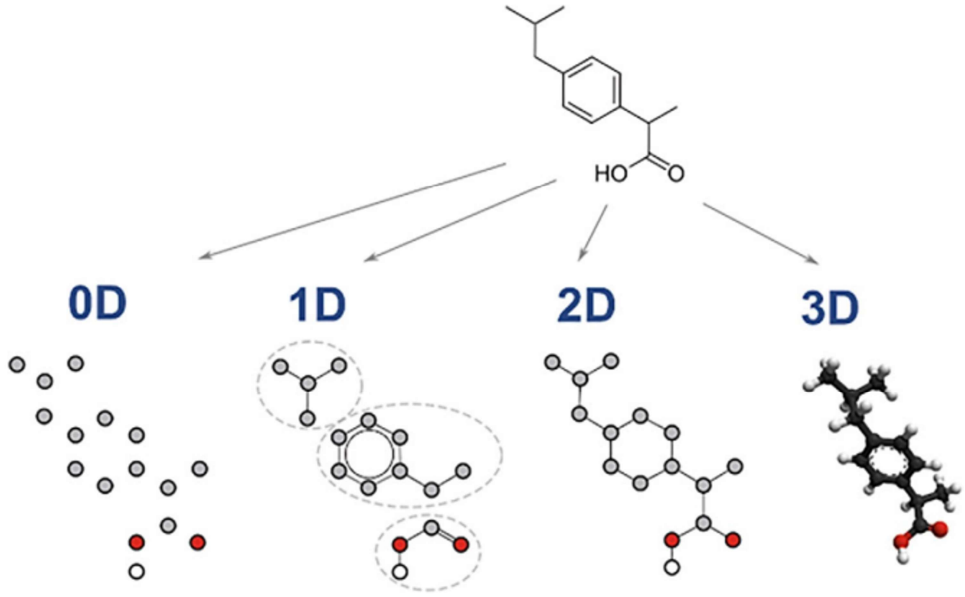


Figure 1.5: Schematic diagram showing multiple descriptors[4]

Transformed Descriptors: Transformed descriptors involve dimensionality reduction techniques to capture the most important information from a large set of descriptor.
Principal Components (PCs): PCs are linear combinations of original descriptors that capture the maximum variance in the dataset. They help reduce the data's dimensionality while retaining most of the relevant information [33].

1.4.2 Property: *Retention Time*

Retention times are the focus of study in QSRR, which refers to the duration analytes take to traverse the column within a chromatographic system, such as HPLC. Before understanding retention times, it is necessary to have an understanding of chromatography.

Table 1.1: Components of Liquid chromatography[1]

Mobile phase	**Stationary phase**	**Sample types**
Liquid	Solid/Liquid	- Liquid samples - Solvent-soluble solid samples

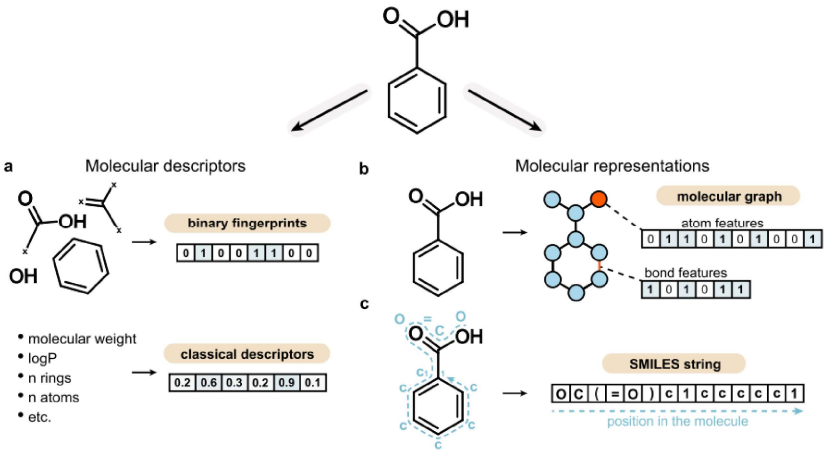


Figure 1.6: QSRR:Building blocks of study[5] (a) Simplified representation of molecular descriptors, which capture predefined molecular features. Traditional molecular descriptors are used in this work. We use the term physicochemical descriptors. (b) Molecular graph, in which atoms are represented as nodes (with corresponding node features) and bonds are represented as edges (with corresponding edge features, if any). (c) SMILES strings, which capture two-dimensional information (atom and bond type and molecular topology) into a string

This section will delve deeper into the topic called chromatography, building upon the fundamental knowledge elucidated in this section. This provides the base for the dataset used in this study.
Chromatography is a powerful separation technique to separate and analyze complex mixtures into their components. It is a potent separation process utilized in a variety of disciplines, including chemistry, biology, and pharmaceuticals.

Basic details about the component of Liquid chromatography are mentioned in Table 1.1.

Principle of RPLC

RPLC is a widely used chromatography technique that separates compounds based on their hydrophobicity or hydrophilicity, which determines their relative affinity for the stationary phase or the mobile phase. The basic principle of RPLC involves the use of a stationary phase consisting of a nonpolar material, such as a hydrophobic functionalised silica, and a mobile phase consisting of a polar solvent, such as water or an organic solvent. The

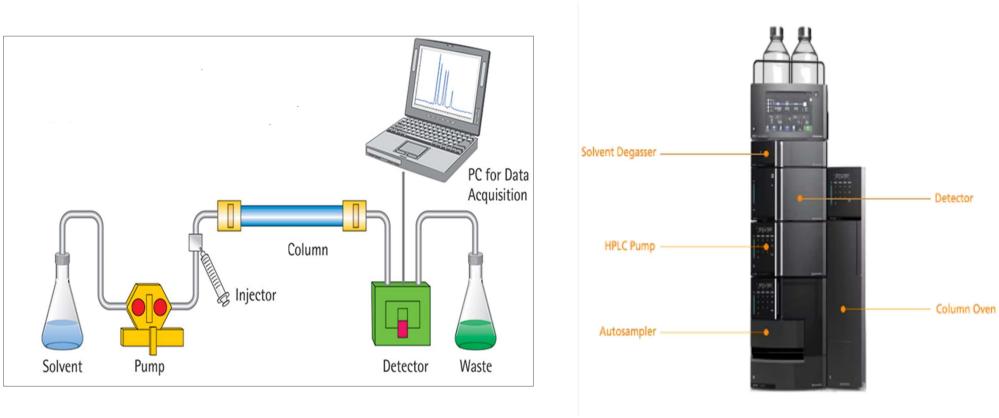


Figure 1.7: Components of chromatography [6]

mixture to be separated is introduced into the mobile phase, which is then pumped through the stationary phase. As the mobile phase passes through the stationary phase, the more hydrophobic or nonpolar components of the mixture tend to adsorb onto the stationary phase and are retained longer, while the more polar or hydrophilic components tend to elute more quickly (Figure 1.9). Adjusting the mobile phase's composition and the stationary phase's properties allows a wide range of components to be separated and detected based on their retention times and/or spectral properties.

Instrumentation of RPLC

The instrumentation of RPLC typically involves a liquid mobile phase, a pump for delivering the mobile phase, an injector for introducing the sample onto the column, a column which is a solid stationary phase, for separation, and a detector for monitoring the eluting compounds (Figure 1.7). The **stationary phase** is typically a silica-based material with hydrophobic bonded phases, such as C18 or C8, which retain nonpolar analytes through hydrophobic interactions Figure 1.8. The **mobile phase** is usually a mixture of water, with or without additives(for pH adjustments) and an organic modifier(typically Acetonitrile(ACN), or Methanol(MeOH), added to modulate the elution of analytes [1].The concept of 'like dissolves like' governs the behavior of a chromatographic column in retaining sample constituents, particularly those that are hydrophobic. These constituents are retained as long as their affinity for the stationary phase is stronger than their affinity for the mobile phase. Conversely, more polar sample constituents tend to elute faster because they are less retained, having a lower affinity for the stationary phase and a higher affinity for the mobile phase. The sample is injected into the column using an **autosampler**, and the mobile phase is pumped through the column. As the mobile phase interacts with the

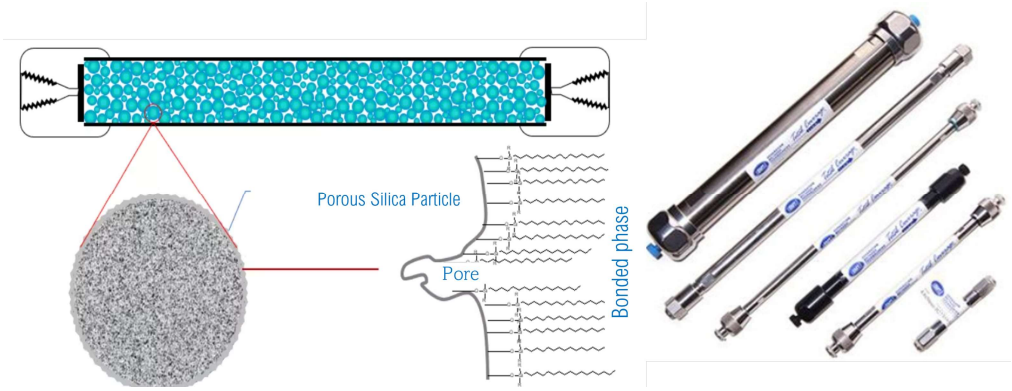


Figure 1.8: Diagram of a traditional column used in RPLC[7, 8]

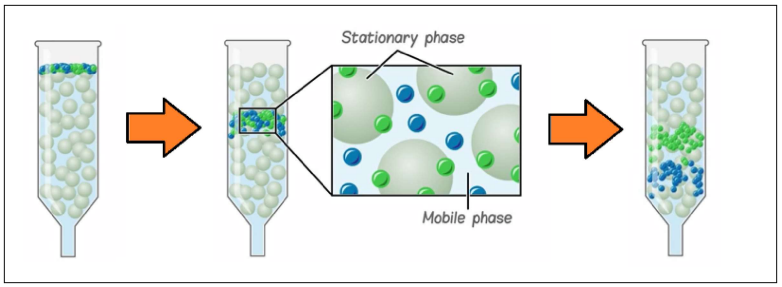


Figure 1.9: Differential elution inside the column [8]

stationary phase, compounds in the sample are separated based on their polarity and hydrophobicity. The eluting compounds are detected by a **detector**. The function of the detector component is to record the quantity and time at which a substance is eluted from the column and plotting the intensity according to time is called chromatogram. [1] There are various types of detectors accessible, depending on the structural features of the substance being analyzed, such as a UV-Vis detector or a mass spectrometer, and their retention times are compared to known standards or reference spectra to identify the compounds.

Comprehending Retention Time

Figure 1.10 shows the schematic diagram of a chromatogram. The detector unit (1.4.2) records signal peaks of separated analytes transported by the mobile phase, and their integrated area under the curve is known as a chromatogram. Each peak can offer both qualitative and quantitative information about the analyte, with the former being conveyed by characteristics such as peak shape, signal intensity, and appearance time in the

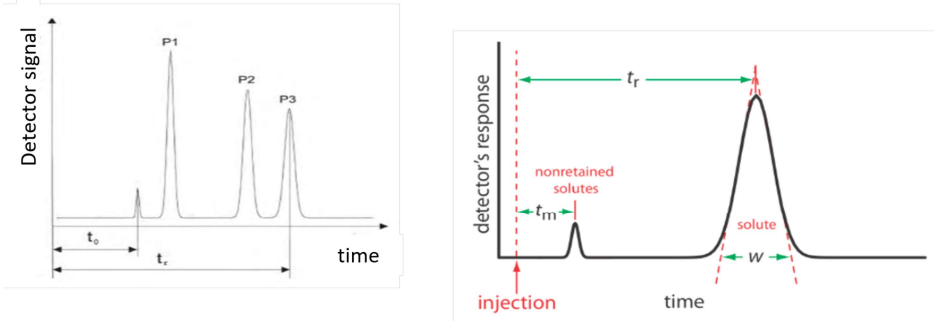


Figure 1.10: Schematic diagram of a Chromatogram. solute's retention time-tR, baseline width-w, and the column's void time-tm for non-retained solutes.[9, 7]

chromatogram.

Retention time t_r

The retention time denotes the duration between the injection and detection of a compound, encompassing the time it spends in both the mobile and stationary phases. As a substance-specific measure, it should yield consistent values when the conditions are identical.

1.4.3 Relationship: *QSRR Modelling*

Quantitative Structure-Retention Relationship (QSRR) modelling refers to a computational approach in cheminformatics and analytical chemistry where mathematical models are developed to establish a quantitative relationship between the chemical structure of molecules and their retention times in chromatographic systems. The goal here is to predict and understand how the structure of a molecule influences its retention time. These models can then be employed to predict the retention times of new compounds based on their structural features. When it comes to QSRR modelling, sometimes the models are built on scarce data. This raises concerns about the reliability on predictions and hence, applicability of these models to new or unknown external compounds. Consequently, implementing (QS(X)R) models required meeting several important validation requirements. REACH, which stands for Registration, Evaluation, Authorization, and Restriction of Chemicals, is a European legislation on chemicals that came into effect in 2007 for this. REACH legislation, along with OECD (Organization for Economic Co-operation and Development (OECD), came with certain requirements

for such predictive modelling, which were named OECD principles. Adhering to these requirements is vital in demonstrating the validity of (QS(X)R) models intended for regulatory use [34].

OECD Principles:

The following principles are encompassed:

1. **Defined end point**: The model should clearly define the specific aspect being predicted.
2. **An unambiguous algorithm**: The model's algorithm should be explicit and leave no room for ambiguity.
3. **A defined domain of applicability**: The model should specify the range of compounds and conditions for which it is applicable.
4. **Appropriate measures of goodness-of-fit**: The model should employ appropriate metrics to evaluate its accuracy and reliability.
5. **A mechanistic interpretation(Optional)**: The model may provide a mechanistic interpretation that explains the relationship between the model descriptors and the predicted endpoint.

In general, these principles furnish users with essential information regarding the predicted endpoint, the algorithm employed by the model, the scope of its applicability, any associated limitations, the model's performance, and an understanding of how the model descriptors are linked to the predicted endpoint. This research adheres to all these principles to ensure that the developed model can be readily applied to new test compounds.

1.5 State of the Art

1.5.1 Evolution of QSRR

The evolution of Quantitative Structure-Retention Relationship (QSRR) modelling has undergone several key stages, each marked by important advancements.

Early Phase (1960s-1980s): During the early phase of QSRR development, the focus was on creating simple models to predict compound retention time in chromatographic techniques. Early QSRR models relied on basic physicochemical parameters like molecular weight, partition coefficient (logP), and boiling point, providing initial insights but with limited predictive power[35, 36].

Middle Phase (1990s-2010s): As technology and computational tools progressed, more sophisticated QSRR techniques emerged. Molecular descriptors, numerical representations of compound structure and properties, became integral to QSRR models. These descriptors encompassed various parameters such as topological indices, fragment-based descriptors, and quantum-chemical descriptors[37, 38, 39]. Incorporating these descriptors into QSRR models led to increased accuracy and predictive capabilities. Another significant development was the utilization of machine learning algorithms in QSRR modelling. These algorithms, including Partial Least Square method (PLS), SVR, RF, ANN analyzed large datasets to identify complex relationships between structural features and retention behaviour [40, 41, 42, 43].

Recent Phase (2010s-Present): In recent years, the availability of chemical databases and advancements in computational power have influenced the evolution of QSRR. Deep learning methods, such as deep neural networks, have become increasingly popular [44, 45, 31, 46]. Data mining and chemoinformatics approaches extract knowledge and patterns from large chemical datasets, aiding in the identification of relevant structural features and the development of more accurate QSRR models[47]. Integration of QSRR with other computational tools, such as molecular docking and molecular dynamics simulations, has expanded its applications [48]. This integration allows researchers to explore the relationship between compound structure, retention, and biological activity, supporting drug discovery and design processes. The complexity of chromatographic processes is profound, as elucidated in Section 1.4. This complexity is attributable to a multitude of factors influencing retention time, including the chromatographic system's design, the constituents of stationary and mobile phases, operational conditions such as temperature and pressure, and the inherent properties of the compounds being analyzed. Despite the high accuracy and efficacy of RPLC, time and cost associated with experimental measures poses a significant challenge. These challenges are not only technical but also involve considerations of cost and time efficiency. To mitigate these issues, extensive research has been dedicated to developing methodologies for chromatographic retention prediction, leading to the establishment of multiple approaches. These approaches can be broadly classified into two main categories[49]:

1. Models designed to predict the retention time of a specific set of solutes under variable chromatographic conditions. These models base their predictions on empirical data derived from previous measurements of the same solute under different conditions, such as varying solvent strengths.

2. Models developed for a specific chromatographic system with the aim of

predicting retention times for new solutes. Such models(including QSRR modelling) utilize retention data from a representative set of substances, all measured under identical chromatographic conditions.

The first category predominantly employs computer-assisted methods, which are frequently utilized during the development of chromatographic methods. A key model in this category is the Linear Solvent Strength (LSS) model[50, 18], which is based on the equation:

$$\log k = \log k_w - S\phi \tag{1.1}$$

where $\log k$ represents the solute retention factor, $\log k_w$ is the logarithm of the retention factor extrapolated to a mobile phase composition with 0% organic modifier, S is a constant specific to the solute and chromatographic system, and ϕ denotes the volume fraction of the organic modifier in the mobile phase. This model, among others in its category, plays a crucial role in computer-assisted chromatographic method development, offering a systematic approach to predicting solute retention under varying conditions [51, 52]. However, LSS models, primarily designed to predict retention times in RPLC based on changes in solvent strength [53], have a limited scope as they may not adequately account for factors like temperature, column properties, or specific solute-stationary phase interactions and operate under the assumption of a linear relationship between solvent strength and retention time, which may not apply to all solute-stationary phase interactions, especially those involving complex mechanisms or unique solute properties [54, 55].
The second category of prediction models encompasses Linear Free-Energy Relationships (LFERs) and Quantitative Structure–Retention Relationships (QSRRs), differentiated primarily by the nature of the molecular parameters they utilize. Based on this differentiation the models can be categorized into two types-

- *Mechanistic models*, which are based on the fundamental physicochemical interactions between the analyte and the stationary phase of the chromatographic system. They aim to describe these interactions using theoretical principles and equations derived from physical chemistry. For example- LFER models. LFER models leverage a set of solvatochromic solute parameters to characterize retention in RPLC[56, 57]. This approach was further refined by Abraham [58], who introduced a more generalized equation incorporating hydrogen-bond descriptors derived from complexation scales, thereby transitioning the nomenclature to Linear Solvation Energy Relationships (LSERs):

$$\log(k) = \log(k_0) + \frac{mV^2}{100} + s\pi + a\alpha^2 + b\beta^2 \tag{1.2}$$

where V_2 represents the analyte molecular volume, π^* denotes the dipolarity/polarizability descriptor, R^2 quantifies the analyte's hydrogen bond donating capability, and α^2 measures the analyte's hydrogen bond accepting potency. The coefficients $\log k_0$, $m/100$, s, a, and b reflect the differences in specific bulk properties between the stationary and mobile phases.

- *Empirical Models*, which are based on statistical or machine learning methods to find correlations between the structure of a molecule and its retention time, without necessarily understanding the underlying physicochemical principles.For example- Insilico QSRR modelling. Molecular descriptors are defined as either the outcome of a logical and mathematical process that transforms chemical information encoded within a molecule's symbolic representation into a useful numerical value (theoretical descriptor) or as the result of a standardized experimental procedure (experimental descriptor) [39]. This approach has inspired the exploration of various chemical factors and computational methods for predicting compound retention times, ranging from basic to complex statistical analyses [59], machine learning techniques [60, 61], and other computer-based strategies, detailed in Sections 1.6 and 1.4.1. Notably, the term QSRRs also encompasses LFERs for the purpose of retention prediction, given that both methodologies utilize a set of molecular descriptors to elucidate retention. However, the designation QSRR is predominantly reserved for models not explicitly classified under LFERs [49] and unlike them which does not need any kinds of experimentation for retention time predictions and hence, predict the retention times in-silico. The primary objective of QSRRs is to formulate a model that accurately describes chromatographic retention within a specific system, thereby enabling the prediction of retention times for new solutes based on the model derived from a representative set of substances[62]. Once a statistically significant and meaningful model is established, it obviates the need for further experimental data to predict retention times for new analytes, significantly streamlining the analytical process.

Over the past decade, the domain of retention predictions via QSRR models has witnessed significant methodological advancements [2, 63]. QSRR modelling establishes a correlation between the chemical structure of compounds and their chromatographic retention times, achieved through a variety of chemometric and computational methods. These encompass both linear and nonlinear statistical models, machine learning algorithms, and the calculation and selection of molecular descriptors [2, 63].
Initial studies, such as those by Put et al. [64], utilized Uninformative Variable Elimination (UVE) coupled with Partial Least Squares (PLS) for the

descriptor selection process. This was followed by the work of Ukić et al. [65] and Chen et al. [66], who demonstrated the superior performance of the full PLS model over UVE-PLS in terms of reducing prediction errors. VolSurfb and 3D molecular descriptors combined with gonane topological weighted fingerprint (GTWF) were used for QSRR modelling based on PLs for steroid identification[18]. Further advancements included the employment of Genetic Algorithms (GA) in conjunction with PLS for more refined variable selection, as employed by Golmohammadi et al. [67], despite not optimizing GA parameters. Traditional methodologies, such as Stepwise Multiple Linear Regression (MLR), were applied in predicting retention factors for specific compounds, highlighting the challenges of underperformance or overfitting in cases with limited data [68]. A significant leap was made by Zuvela et al. [69], who conducted a comprehensive comparative analysis of variable selection techniques in peptide QSRR model development employing nature-inspired optimization algorithms. These included Genetic Algorithms (GA-PLS) [70], Particle Swarm Optimization (PSO-PLS) [71], Artificial Bee Colony (ABC-PLS) [72], Firefly Algorithm (FA-PLS) [73], and Flower Pollination Algorithm (FPA-PLS) [74], compared against Interval PLS (iPLS) [75] and Sparse PLS (sPLS) [76]. Their findings indicated that nature-inspired algorithms outperformed both iPLS and sPLS in predictive accuracy. Moreover, Perisic et al. [77] explored hybrid models that integrate machine learning techniques with quantum chemical calculations to enhance model performance through aggregated predictions. With the advent of Artificial Intelligence (AI), novel QSRR studies have focused on the graph properties of compounds [18]. Ju et al. [78] utilized a deep neural network (DNN) pre-trained with weighted autoencoders and transfer learning for efficient prediction of compound retention times. This was further advanced by Kwon et al. [79] and Kensert et al. [32], who investigated the efficacy of Graph Convolutional Networks (GCN) and transfer learning approaches to enhance molecule retention time (RT) predictions. However, using GNN/GCN for QSRR retention time predictions can be challenging due to the need for large, diverse datasets, computational intensity, complexity in capturing chemical interactions, and issues with model interpretability and generalizability. Advanced statistical approaches, such as Bayesian estimation and multilevel modeling[80, 81], have also been used for chromatographic data analysis, offering a probabilistic framework for model development that accommodates data uncertainty and variability [59, 82]. Finally, Bouwmeester et al. [60] constructed QSRR models using seven machine learning algorithms-[60] built QSRR using seven machine learning algorithms including Bayesian Ridge Regression(BRR)[83], LASSO[84],

ANN[85], Adaptive Boosting(AB)[86], Gradient Boosting(GB)[87], Random Forest (RF)[88] and SVR(Linear and non-linear)[89, 90] across 36 metabolomic datasets, highlighting that no single algorithm universally excels, with performance varying by analyte type or experimental protocol.

1.6 Workflow of QSRR Modelling

The complete workflow [91] of QSRR modelling can be divided into three main steps(Figure 1.11), which are:
(1) Data collection, data preparation and preprocessing and data splitting
(2) Model Development that includes selection of modelling technique, molecular descriptor selection, model training and validation
(3) Model testing and Applicability domain check

1.6.1 Data collection and preprocessing

Data collection:

The first step in QSRR is to collect a dataset of compounds that include target properties, such as retention time, and molecular descriptors, which are structure-derived features in a chromatographic system of interest. This dataset should be representative of the compounds of interest. Retention time can be obtained through experimental measurements. Molecular descriptors can be experimentally determined or calculated through computational models or using some tools or software for example- RDKIt tool([92], Chemaxon [93], ADMET Predictor[94],AlvaDesc[95] etc.

Data Preprocessing

Data preprocessing is a pivotal data mining approach that encompasses the conversion of raw data into a comprehensible format [96]. Real-world data frequently exhibits incompleteness, with missing attribute values, absence of specific attributes of interest, or reliance on aggregate data. Additionally, such data may be characterized by noise, incorporating errors or outliers, as well as inconsistencies in codes or names. The application of data preprocessing serves as an established method for addressing and resolving these inherent issues. The data processing phase encompasses several key procedures, including the identification and handling of missing values, the encoding of categorical data into numerical representations as per requirements, feature scaling and data splitting.

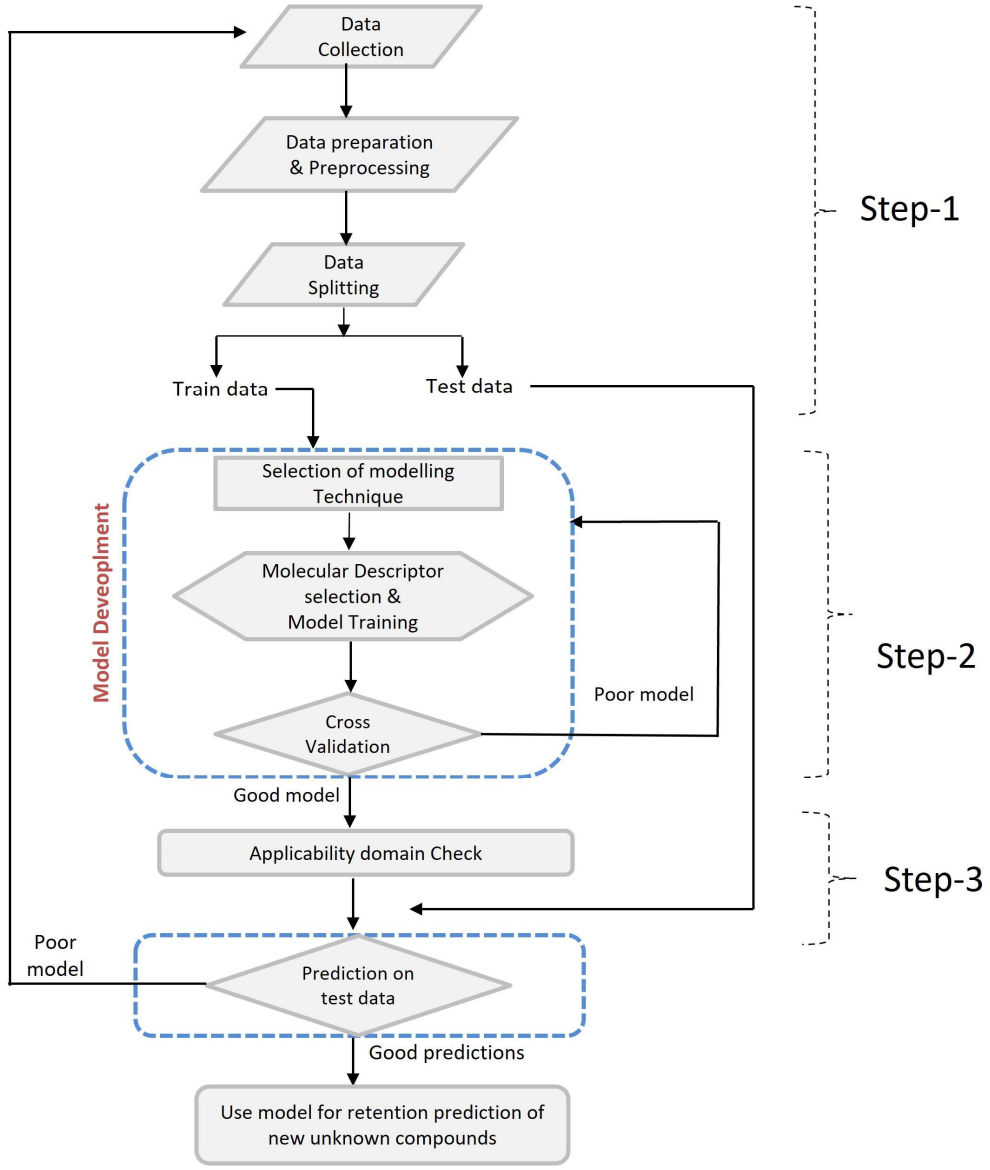


Figure 1.11: Schematic diagram showing QSRR workflow

Handling Missing value: This step involves detecting and managing missing data points in the dataset, such as missing molecular descriptor values. Missing values can be imputed (filled in) or the corresponding records might be removed to ensure the model is trained on complete and accurate data.

Encoding categorical data into numerical:If the QSRR dataset includes categorical data (hypothetical example, types of stationary phases in chromatography), these categories are converted into numerical form since most QSRR models require numerical input. This step is less common in QSRR modeling, as the majority of input features (molecular descriptors) are inherently numerical.

feature Scaling: This involves standardizing or normalizing the range of feature values, such as molecular descriptors, to ensure that no single descriptor disproportionately influences the model due to its scale. Techniques like StandardScaler, MinMaxScaler are some of the mostly used techniques to adjust the features into a comparable scale, enhancing the QSRR model's convergence and performance[97].

- Standardization is a scaling method where the values are centered around the mean with a unit standard deviation. The formula is given by:

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \tag{1.3}$$

  where μ is the mean of the dataset, σ is the standard deviation of the dataset, and $x \in \mathbb{R}$ represents x represents an individual data point or a variable within the dataset being processed..

- The MinMaxScaler scales the features within a specified range (typically 0 to 1), also known as data normalization. Selecting the target range depends on the nature of the data. The formula is given by:

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1.4}$$

  where x represents an individual data point or a variable within the dataset, $\min(x)$ and $\max(x)$ refer to the minimum and maximum values in the dataset, respectively. Here, $x \in [\min(x), \max(x)]$ and $x_{\text{scaled}} \in [0, 1]$.

data splitting: The dataset is divided into training and testing sets (and possibly a validation set) to enable the independent training and evaluation of the QSRR model. This step is essential for assessing the model's predictive performance on unseen data.Typically 70-80% of the total dataset is used in model training and rest of the dataset is used for testing purposes. Randomization is commonly applied for the selection of the training and testing datasets, with the data being shuffled randomly to ensure the model generalizes effectively.

1.6.2 Step 2: Model Development

Selection of modelling technique

This step includes model training and model validation including molecular descriptor selection. After completing the data-cleaning process and identifying crucial features influencing the retention of specific molecules, we proceed to employ prediction models, commonly referred to as regression methods. The model is constructed using the training set and is validated using model validation.

The choice of algorithms for model development is contingent upon the fundamental objective and the characteristics inherent in the available dataset. These methods can span from classical machine learning models to cutting-edge artificial intelligence techniques. The selected algorithms may vary, encompassing linear methods that seek to capture the linear relationship between the target property and chosen descriptors as well as non-linear methods, which excel at capturing non-linear dependencies based on the specific goals and dataset characteristics. Hyperparameter tuning marks the important step in regression models and involves systematically adjusting the model's settings to find the optimal combination that minimizes error and improves predictive accuracy [98]. Grid search and random search are commonly used methods for parameter optimization in classical machine learning. Grid search methodically tests a predefined range of hyperparameter values, with fixed step sizes influenced by domain knowledge, computational resources, preliminary results, and hyperparameter sensitivity, determining the granularity of the search and impacting both the thoroughness and computational cost [99]. Random search samples hyperparameter values from a defined distribution, offering a faster but potentially less exhaustive exploration. Each of these methods balances between exploration of new parameters and exploitation of known good parameters to efficiently optimize model settings[100]. Outlined below are descriptions of some of the most frequently employed machine learning algorithms([101, 102]).

- MLR(Multiple Linear Regression): MLR is a linear regression method that models the relationship between a dependent variable and multiple independent variables. It estimates coefficients to create a linear equation that predicts the dependent variable based on the independent variables.

- Lasso regression: Lasso regression is a linear regression method that uses shrinkage. It Performs L1 regularization, i.e., adds a penalty equivalent to the absolute value of the magnitude of coefficients and encourages sparsity by shrinking some coefficients to zero, effectively selecting important features. It reduces model complexity and overfitting resulting from simple

Linear regression[103].

$$\text{L1 regularization (Lasso)} = \lambda \sum_{j=1}^{p} |\theta_j| \tag{1.5}$$

Here, λ is the regularization parameter, and θ_j represents the coefficients of the features in the model. The symbol $|\cdot|$ denotes the absolute value.

- SVR (Support Vector Regression): SVR is a regression method that uses support vector machines to find a hyperplane that best fits the data points in a continuous space and tries to maximise the margin between the data points and the regression line. It handles both linear and non-linear relationships by using kernel functions which are a set of mathematical functions that help in taking the data as input and transforming it into the required form into a higher-dimensional feature space[104]. Linear, Non-Linear, Polynomial, Radial Basis Function(RBF) and Sigmoid are some of the kernels. Among all, RBF is the mostly used kernel.

- RF(Random Forest): RF is an ensemble method that constructs multiple decision trees using bootstrap samples of the data and random feature subsets. The trees vote to make predictions, and the final prediction is based on the majority vote. RF reduces overfitting and provides robust predictions [105].

- GBR (Gradient Boosting Regression): GBR is an ensemble method that combines multiple weak regression models sequentially. Each model corrects the errors of the previous model, gradually improving the prediction accuracy. It builds a strong predictive model by minimizing the residual errors [40].

- ANN (Artificial Neural Network): ANN is a machine learning model inspired by the structure of the human brain. It consists of interconnected nodes (neurons) organized in layers. Signals flow through the network, and each neuron applies a non-linear activation function to make predictions. ANN is effective for complex, non-linear relationships[106].

- DNN(Deep Neural Network): Deep learning, a subset of machine learning that in turn falls under the broader category of artificial intelligence (AI), serves as a foundational technology for automating tasks and improving accuracy in AI applications. Distinguished from traditional machine learning, deep learning exhibits the capability to handle complex and unstructured data(images and text) without the need for extensive preprocessing. This autonomy in feature extraction diminishes the reliance on human

experts. Employing processes like gradient descent and backpropagation, deep learning continually refines itself to achieve precise predictions. DNN is characterized by multiple hidden layers positioned between the input and output layers [44]. This architecture (Figure- 1.12) enables DNNs to grasp intricate patterns and construct hierarchical representations of data. The training of DNNs involves backpropagation, facilitating the resolution of intricate tasks. MLP which stands for Multi-Layer Perceptron, is a type of artificial neural network. It is a class of feedforward neural networks where information moves in one direction—from the input layer through one or more hidden layers to the output layer[107]. It can be DNN based on increased number of hidden layers.

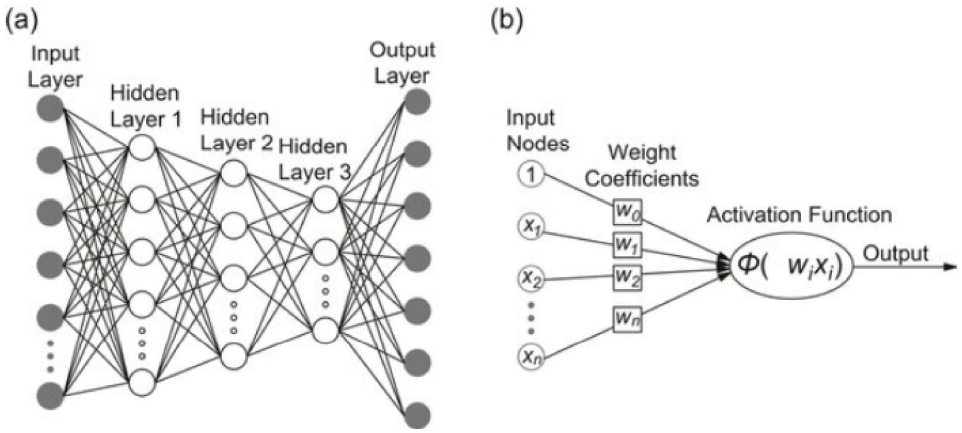


Figure 1.12: Schematic diagrams of deep learning neural network (DNN). (a) The overall structure of DNN. (b) Concept of weight coefficient and activation function. [10]

- *CNN (Convolutional Neural Network)*: Figure-1.13 They belong to the category of discriminative deep architectures, demonstrating satisfactory performance in handling two-dimensional data characterized by a grid-like topology, as exemplified in images and videos [11]

- *Stacking:* Stacking is an ensemble learning technique that involves using a combination of different base learners(Figure 1.14), often heterogeneous ones, in parallel. These weak learners make predictions independently, and their predictions are then used as features for a meta-learner. The meta-learner is trained to learn how to best combine these input predictions to generate a final output prediction. Typically, Linear Regression, Random

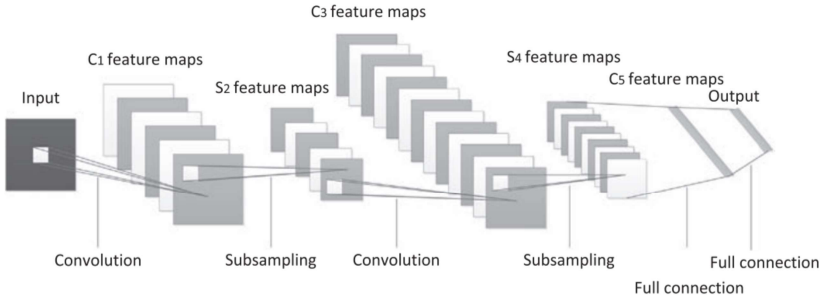


Figure 1.13: Architecture of CNN used for image recognition as an example[11]

Forest, decision trees, or neural networks are employed as meta-models in stacking architectures, with the choice depending on the specific requirements of the task, such as the need for interpretability or prediction power.

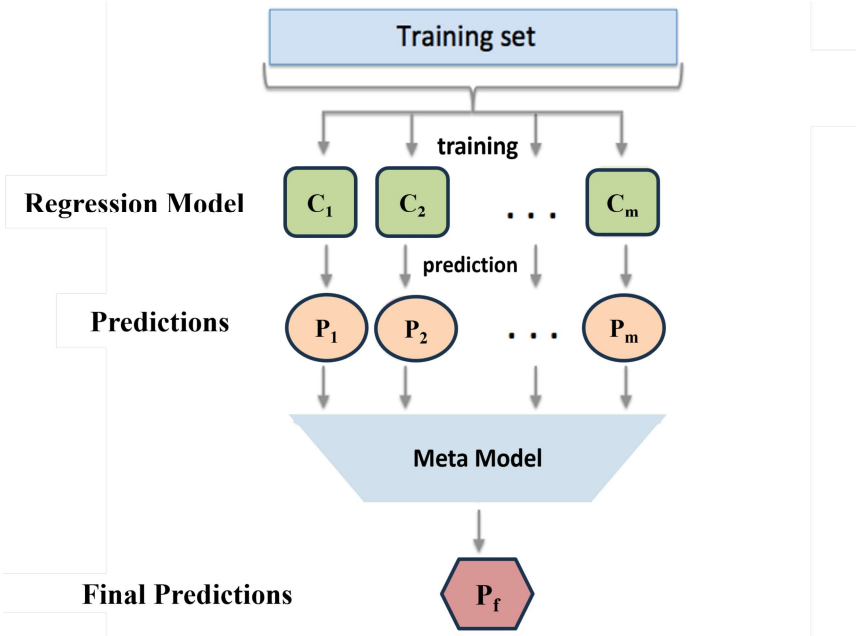


Figure 1.14: Architecture of Stacking[12]

Molecular descriptor calculation and selection:

Given multiple types of molecular descriptors, it becomes important to select only key features which carry important information about the target property and should be selected to be used while modelling. Hence, feature selection methods in QSRR aim to identify the most relevant molecular descriptors or features that contribute significantly to the retention behaviour. It is important to note that the choice of feature selection method depends on various factors, such as the size of the descriptor pool, the sample size, the complexity of the retention behaviour, and the specific goals of the QSRR modelling study. Different methods may yield different subsets of features and varying model performance [3]. Hence, comparing and validating the results obtained from different feature selection techniques is often recommended. Here is a summary of the feature selection methods commonly used in QSRR modelling (Fig 1.15)[3]. Figure 1.16 explains some advantages and disadvantages of all feature selection methods.

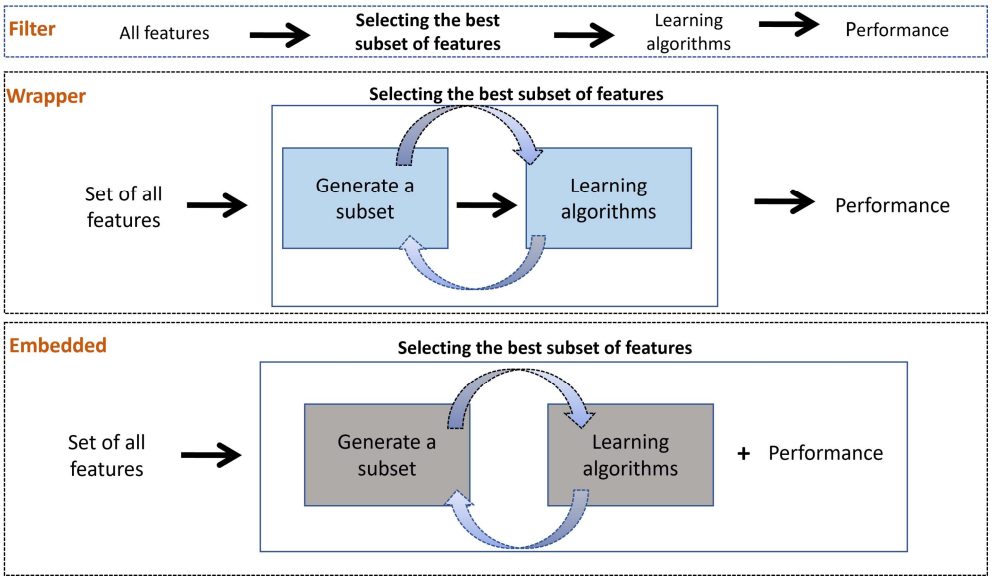


Figure 1.15: Pictorial representation of multiple descriptor selection methods

1. Filter method: Filter method: Analyses each descriptor individually to assess its correlation with the target variable (retention time). Statistical tests such as t-tests or correlation analysis are used to evaluate the significance of the relationship. Descriptors with high correlation or significant p-values are selected as relevant features. Such a method should

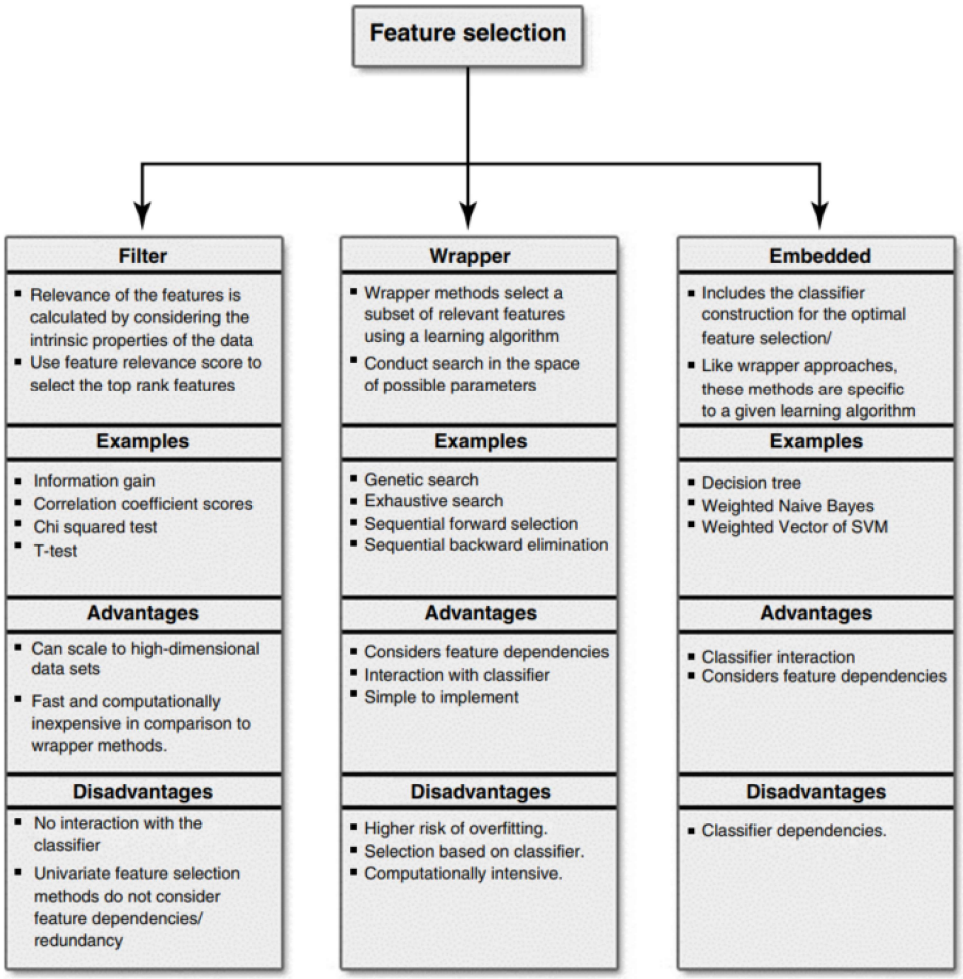


Figure 1.16: Important points of every multiple descriptor selection methods[3]

be used when the dataset is small, and there is a need for a quick and simple feature selection method. Feature selection using filter methods are computationally efficient and can provide insights into the individual relationships between descriptors and the target variable. These methods include techniques such as correlation-based filter(CFS)[108, 109],and ReliefF-based algorithms [110] etc. The CFS algorithm works by first calculating the correlation between each feature and the target variable. Then, it computes the correlation between each pair of features. Next, it selects the subset of features that has the highest correlation with the target variable and the lowest correlation with each other [111].The subset evaluation phase involves computing the merit of a feature subset. This is

done using a heuristic that combines the average correlation of the features with the target (denoted as $\overline{r_{cf}}$) and the average inter-feature correlation ($\overline{r_{ff}}$) [112]. The merit score formula,

$$M(S) = \frac{k \cdot \overline{r_{cf}}}{\sqrt{k + k \cdot (k-1) \cdot \overline{r_{ff}}}}$$

balances relevance and redundancy by promoting high relevance to the target while penalizing feature overlap. In the search strategy phase, features are added incrementally through a greedy forward selection process[113], choosing features that maximize the subset merit. The process concludes when adding additional features no longer improves the merit score, indicating that further additions are either redundant or irrelevant to the target variable.

2. Wrapper methods: Wrapper methods use an external model to evaluate the performance of different feature subsets. These methods typically involve a search algorithm that iteratively evaluates different combinations of descriptors using a performance metric (e.g., cross-validation error or R-squared). Examples of wrapper methods include sequential forward and backward selection [114], genetic algorithms[115], and recursive feature elimination[116].
   These methods provide a thorough search of the feature space, considering the performance of the external model as the selection criterion. RFE for instance, operates by iteratively selecting a subset of features from the initial set in the training data. It does this by fitting the chosen machine learning algorithm, assessing feature importance, and discarding the least significant features. The model is then refitted, and this cycle continues until the desired number of features is reached [117]. Wrapper methods can be computationally expensive, especially when the number of descriptors is high. In such cases, the search process may become infeasible or time-consuming and hence, should be avoided when there is a very large dataset and the feature space is vast.

3. Embedded methods: Embedded methods incorporate feature selection within the model-building process. They aim to find the optimal feature subset while simultaneously training the predictive model. Examples of embedded methods include LASSO (Least Absolute Shrinkage and Selection Operator) and Random Forest, which apply regularization techniques to penalize irrelevant or redundant descriptors during model training. Such methods can handle large descriptor sets effectively and automatically penalize irrelevant or redundant features during model training.

Model Validation

There are multiple methods of model validation techniques. A few of them are LOOCV(Leave One Out Cross Validation), LMOCV(Leave Multiple Out Cross Validation), and Kfold cross-validation. Other types of model validations are Y-randomization and bootstrapping[102, 118]. Each method has its strengths and weaknesses, and the choice depends on the specific dataset and the goals of the analysis. LOOCV (Leave-One-Out Cross-Validation) is a cross-validation method where the dataset is split into training and testing sets, where each observation is used as the testing set once, while the rest are used for training[119]. This process is repeated for every observation, and the performance metrics are averaged to evaluate the model.LMOCV(Leave-Multiple-Out Cross-Validation) is similar to LOOCV, but it leaves a specific number of observations out for testing instead of just one. It is useful when the dataset is large and leaving out every single observation for testing is computationally expensive [120]. In *K-Fold Cross-Validation*[45], the dataset is divided into k equally (or nearly equally) sized folds or subsets. In each of the k iterations, a different fold is used as the validation set, and the remaining $k-1$ folds are combined to form a training set. This method aims to utilize all available data for both training and validation, ensuring that every observation acts as part of a validation set exactly once and part of a training set $k-1$ times[121]. *K-Fold Cross-Validation* is widely used due to its balance between computational efficiency and the thoroughness of the evaluation.

1.6.3 Step 3: Model Testing and Applicability domain check

Model Testing

The models, which have been trained and validated, undergo testing using unseen test data to assess their generalization performance. Various evaluation metrics are employed to check the performance of all developed models. These quantitative measures serve as tools to evaluate the effectiveness of prediction models, providing insights into how well the models perform in tasks related to retention time prediction. The choice of performance metrics depends on the nature of the problem and the goals of the model. Some common performance metrics used in machine learnings[122] are MSE(Mean Squared error), RMSE(Root mean squared error), MRE(mean relative error)/ MAPE(Mean absolute percentage error), MAE(Mean Absolute Error), and R^2(Coefficient of determination). Their mathematical formula is as such:

$$MSE = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n} \tag{1.6}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{1.7}$$

$$MRE/MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \tag{1.8}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1.9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{1.10}$$

where y_i and $\hat{y}_i$ are the ground truth and the predicted value, respectively, for sample i,$\bar{y}$ is the mean(average) of all the observed values (y_i) of the dependant variable in the dataset and n is the total number of sample.

Applicability domain

Applicability Domain (AD) defines the limitations of QSRR models based on structural and response criteria. Reliable predictions are limited to structurally similar chemicals used in model development. Query chemicals within the model's scope are considered interpolated, while those outside are extrapolations. AD ensures higher reliability for predictions within its boundaries, which is crucial for model accuracy. Molecular descriptors also influence AD. Query chemicals differing from the training set's structural limitations are outliers in that chemical space [130]. There are several methods that have been proposed to define the model's applicability. They are summarized as in Table 1.2.

Table 1.2: Summary of Applicability Domain Calculation Methods

AD Approach	**Description**	**Reference**
Range-based	Involves using the range of individual descriptors used for model development. Molecules exceeding these ranges are excluded from the model's AD. Variants include PCA Bounding Box, which considers principal component values.	[34, 123, 124, 125]
Convex Hull	Encloses the training space in the smallest possible convex area. Molecules outside this area are excluded. Not used in high-dimensional data sets due to limitations.	[34, 123, 124]
Distance-based	Includes methods like centroid distance (where distance from the centroid is calculated) and kNN-based strategies (where distances from k nearest neighbors are used). These methods define thresholds beyond which molecules are excluded from the AD.	[34, 123, 124, 125, 126]
Probability Density Function (PDF)	Utilizes kernel methods to estimate the density of points in the descriptor space, setting a threshold to determine AD. Variants include fixed, optimized, and variable Gaussian kernels, as well as adaptive and triangular kernels.	[127, 128]
kNN with Variable Thresholds	Combines kNN distance measures with adaptive kernel methods for density estimation. Molecules closer to training data than a variable threshold are included in the AD.	[129]

2
Objectives

Reversed-Phase Liquid Chromatography (RPLC) is a key technique for separating and analyzing substances in mixtures, fundamentally relying on the interactions between analytes in the mobile phase and a typically hydrophobic stationary phase. These interactions determine the retention time of analytes through the chromatographic column. The method development in liquid chromatography stands as a basic aspect of analytical chemistry, especially within pharmaceutical sciences, drug analysis, bioanalysis, and medicinal chemistry. The core of method development in liquid chromatography is to establish optimal conditions for the effective separation of analytes, which depends on various experimental parameters like chromatography mode, stationary phase type, and mobile phase characteristics, including organic modifiers, additives, pH as well as temperature and pressure. Given the resource-intensive nature of traditional trial and error in method development—marked by significant time and cost—there's a growing interest in using *In-silico* methods to create predictive models that correlate compound properties with chromatographic retention time. Quantitative Structure-Retention Relationships (QSRR) models represent this approach by predicting retention times based on molecular structure, thereby simplifying method development and boosting the efficiency of RPLC analyses. In this direction, this thesis delves into diverse QSRR strategies to improve retention time prediction models, considering experimental variables and their impact on compound retention times, by taking pH variations as one of the example. The profound influence of pH on RPLC analyses underscores the complexity of determining accurate retention times, affecting not only operational efficiency but also the cost-effectiveness of the chromatographic process. Minor pH adjustments can significantly shift the ionization state of analytes altering their interactions and consequently, their retention times. This necessitates extensive method development efforts. The ultimate goal is to offer advancements in separation methods, thereby supporting pharmaceutical and biochemical research through unique and practical analytical strategies by investigating multiple ways to handle varying retention times of a compound with multiple pH. In order to explore the intricate interplay between pH and retention mechanisms for small molecules in RPLC, three main questions (below) have been addressed:

2.1 Objective 1

How to predict retention time at specified condition individually and select the key features responsible for predictions?

The literature offers a wide array of algorithms and regression methods for

modelling retention time predictions. Various strategies and chemical factors, ranging from basic to advanced statistical methods, machine learning, and other computer-based techniques, have been utilized to predict retention time of a compound in a chromatographic system. These approaches leverage chemical structure details from the simplest to the most complex. Yet, no method has emerged as perfect, and there is a lack of a definitive starting point for users beginning QSRR modeling to manage varying retentions across different pH levels. Hence, the first objective of the thesis was to investigate various categories of molecular descriptor selection methods as well as multiple types of prediction models to streamline the QSRR modelling process from start to end that can be tried and that can work in case of small molecules in RPLC at multiple pH(Chapter 4.1)

2.2 Objective 2

Is it feasible to develop a QSRR model that simultaneously predicts retention times across all pH levels effectively, while accounting for the interrelationships among them?

The second objective of this thesis involves a shift from the conventional approach of constructing individual retention prediction models for distinct pH levels. While this method helped identify which descriptor groups were linked to retention time changes at different pH levels, it proved to be time-consuming and unable to utilize the insights from the interrelationships between targets effectively. Hence, the second objective of the thesis was to determine whether it is feasible to model retention times across all pH levels simultaneously in a way that incorporates the relationships between targets(Chapter 4.2).

2.3 Objective 3

What strategies can be implemented to overcome the challenges posed by scarce dataset availability in QSRR modeling, thereby enhancing the accuracy and efficiency of multi-target predictions?

Analyzing analytes using various separation techniques, including RPLC, presents a substantial challenge for computational research due to smaller data availability. These experiments require a significant amount of time to generate and validate high-quality data, often necessitating numerous repetitions. As a result, collecting enough data for analysis in machine learning

or AI-driven studies as used in QSRR modelling, which demands extensive data, becomes a daunting and prolonged task.
In light of this, the third objective of the thesis was to investigate an alternative strategy to overcome the obstacles presented by limited data availability in QSRR modeling(Chapter 4.3).
In summary, the significance of this research lies in its potential to advance the understanding and application of QSRR modeling in chromatography. By focusing on the effects of pH on retention times and leveraging QSRR for predictive purposes, this thesis aims to develop a methodological framework that can be applied to a wide range of analytical challenges. This approach promises to improve the efficiency, cost-effectiveness, and reliability of chromatographic analyses, thereby supporting the broader goals of analytical chemistry in achieving precise, accurate, and efficient analytical outcomes.

3
Material

3.1 Preamble

This chapter explores the different datasets used in the study. It covers five points:
Data Source and Availability, **Data Structure**: gives a summary of the data's structure and its format.
Feature Description: Provides an overview of the variables or features included in the dataset. It contains specifics on the data types for each characteristic as well as information about their meaning and importance.
Target Variable: explains the target variable's characteristics corresponding to every dataset.
Software and Tools: Describes the particular libraries and modelling packages used in the data collection.

3.2 Summary of Datasets

There are four retention time datasets used in this study for QSRR modelling (Table 3.1). These datasets are abbreviated as: Dataset1- small dataset- LPAC, and Dataset 2 -metlin data or SMRT dataset, Dataset3-ACN, Dataset4-Riken. It's important to note that the labels LPAC and ACN for the datasets-1 and dataset-3, have been assigned by us to facilitate easier discussion and precise referencing throughout the thesis. These are not the actual and conventional names published.

It can be seen from Table 3.3, and 3.4 that the multiple targets to be

Table 3.1: Summary of Datasets Used in the Thesis, tR - Retention Time, PC- Physicochemical, IB- Image based descriptors

Feature	**DataSet-1**	**DataSet-2**	**DataSet-3**	**DataSet-4**
Chapters	4.1, 4.2, 4.3	4.3	4.3	4.3
Names	LPAC	Metlin/SMRT	ACN	RIKEN
Purpose	single target, multitarget, transfer learning based qsrr	QSRR modeling for 77K compounds	single target, multitarget, transfer learning qsrr	Pre training for transfer learning model
Source	Data in Brief[131]	Nature Communication [13]	Analytical Chemistry[132]	Analytical Chemistry[32]
Data Structure	97 rows, 239 columns	77K rows, 226 columns	130 images	750 rows, 226 columns
Feature Description	PC, IB	PC, IB	IB	PC
Target Variable	tR at five pH	tR at one pH	tR at five pH	tR at one pH
Tools and Software	RDKit, Chemicalize	RDKit	RDKit, Chemicalize	RDKit

predicted are highly correlated. Advanced machine learning techniques rely heavily on related responses because they enable models to transfer insights from one target to another while minimising the need for large amounts

of data and computational power. This leads to an optimisation of learning processes. Effectively managing these correlations also requires using models that either take into account the combined distribution of targets or constrain the prediction of one target based on the predictions of other. In order to improve prediction accuracy and model dependability, strategies like graphical models [133], Random forests [134], and structured neural networks [135] are especially useful since they can identify and take advantage of inter-target interactions. These modelling methods emphasise how crucial it is to choose the appropriate model architectures and learning frameworks in order to maximise the benefits and address the challenges in highly correlated multi-target scenarios.

Description of Molecular descriptors

There were two types of molecular descriptors used in this thesis. 1) Physicochemical descriptor 2) Image based descriptors

3.2.1 Physicochemical descriptors

Due to the extensive list, descriptions are kept concised. Table 3.2 summarizes the physicochemical descriptors used in this thesis.

3.2.2 Image based descriptors-MIA

Multivariate Image Analysis descriptors/Image-based descriptors for chemical compounds are a relatively newer approach that utilizes the power of visual representation to capture structural information of molecules for retention time predictions (Example of input image is shown in Figure 4.3.7 . The images demonstrate a strong association with retention times and serve as a method for encoding chemical properties[136]. The differences in pixel positions reflect changes in the structure within a related group, thereby accounting for the variance in retention times observed within the series[137].

3.2.3 DataSet-1 LPAC dataset

The dataset is derived from Reversed Phase Liquid Chromatography experiments for small pharmaceutical compounds to build QSRR models with varying pH. This in-house dataset having five retention times of all chemical compounds being studied, which help in analyzing QSRR modelling approaches in Chapter 4. The distribution of retention times of compounds at all five pH are shown in Figure 3.2 and the correlation among the targets are shown in Table 3.3 This dataset was used in the study chapter 4_1,4_2,4_3. Other information is available in Table 3.1.

Table 3.2: Molecular Descriptors from RDKit

Descriptor	**Description**
MolWt	Molecular weight
EState_VSA (1-10)	Electrotopological state indices capturing electronic and surface area aspects
fr_ (specific groups)	Presence of specific chemical groups or motifs
Min/Max(EStateIndex, AbsEStateIndex)	Minimum and maximum electrotopological state indices
PEOE_VSA (1-14)	Electronic distribution based on van der Waals surface areas
SlogP_VSA (1-12), MolLogP	Hydrophobicity indicators based on LogP values and surface areas
VSA_EState (1-10)	Electronic environment's influence on surface areas
logD	Distribution coefficient at specific pH
Asymmetric.atom.count, Atom.count	Count of atoms, including asymmetric ones
BalabanJ, BertzCT, Chi indices, Ipc, Kappa	Topological and connectivity indices
FpDensityMorgan (1-3)	Fingerprint density indicating complexity
HallKierAlpha, Heavy.atom.count, HeavyAtomMolWt	Size, shape, and atom count indicators
Hetero.ring.count, Hydrogen bond counts, NHOHCount, NOCount	Specific atom or feature counts
Num (various types)	Counts of ring types and structural motifs
Polarizability, qed, Ring.count, Rotatable.bond.count, TPSA	Polarizability, drug-likeness, structural, and surface area descriptors
SMR_VSA (specific indices)	Molar refractivity based surface areas

Data distribution

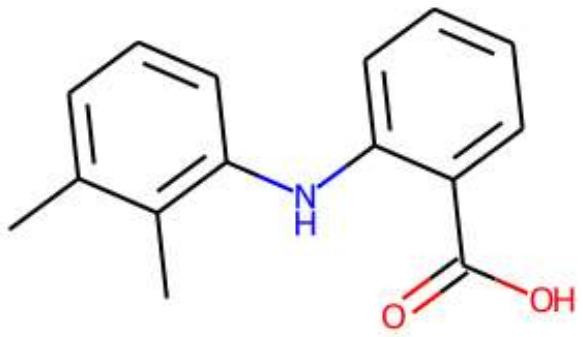
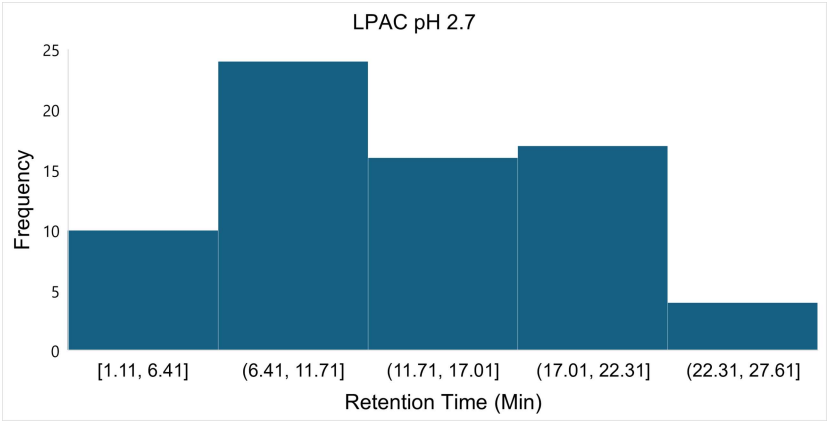
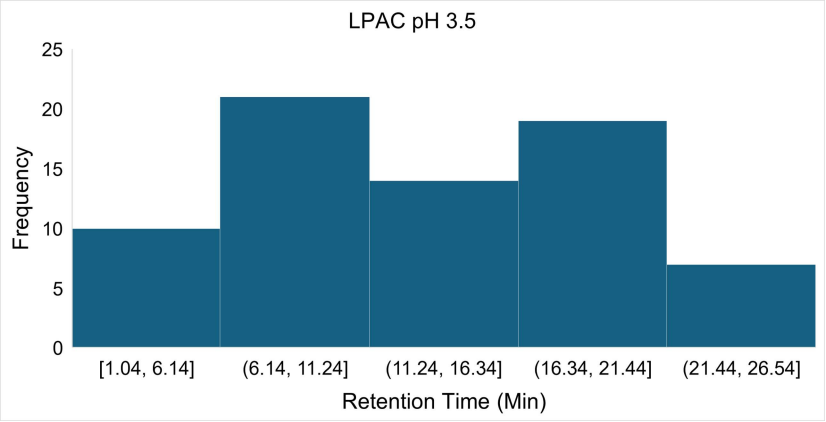
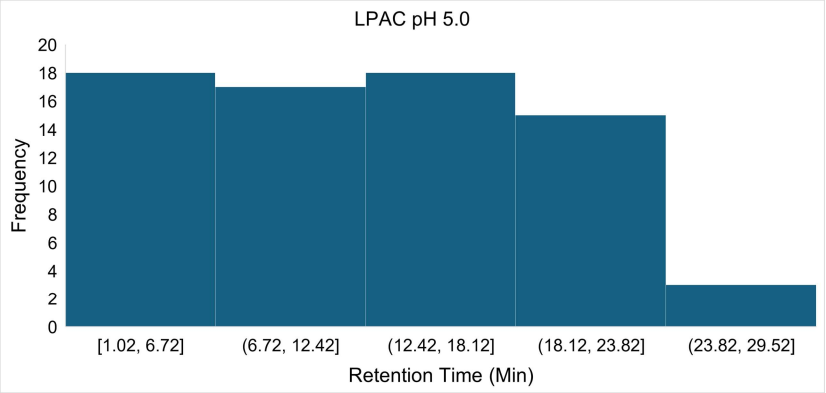
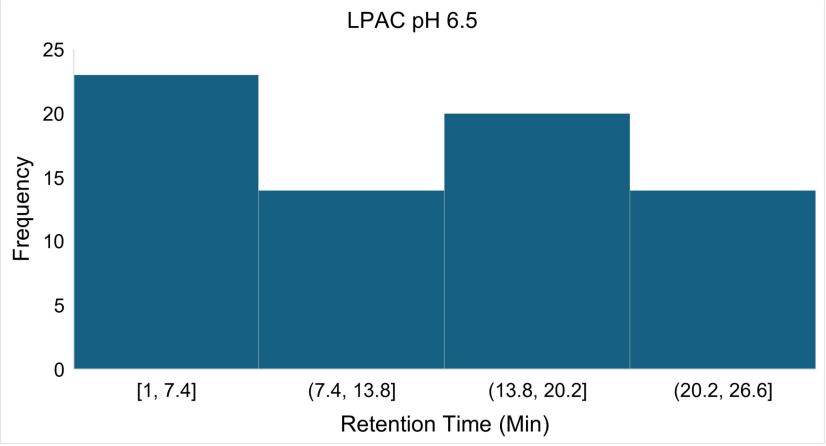


Figure 3.1: Example(3aminobenzoic acid) of an image used as input in CNN model




LPAC pH 3.5
Frequency
25
20
15
10
5
0
[1.04, 6.14]
(6.14, 11.24]
(11.24, 16.34]
(16.34, 21.44]
(21.44, 26.54]
Retention Time (Min)


LPAC pH 5.0
Frequency
20
18
16
14
12
10
8
6
4
2
0
[1.02, 6.72]
(6.72, 12.42]
(12.42, 18.12]
(18.12, 23.82]
(23.82, 29.52]
Retention Time (Min)


LPAC pH 6.5
Frequency
25
20
15
10
5
0
[1, 7.4]
(7.4, 13.8]
(13.8, 20.2]
(20.2, 26.6]
Retention Time (Min)

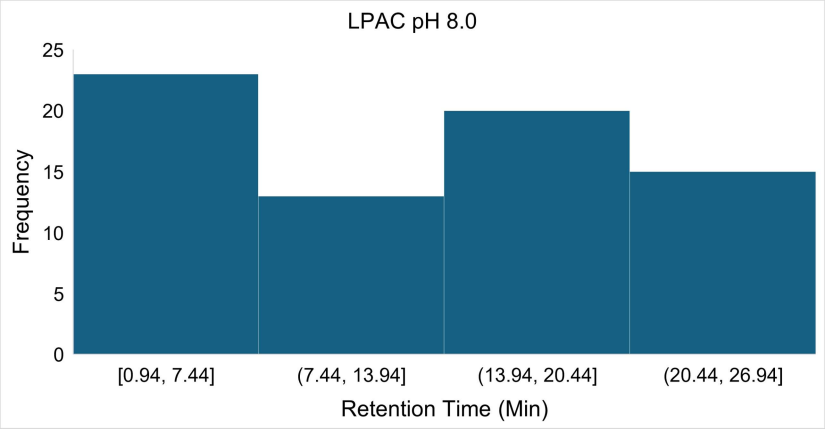


Figure 3.2: Target distributions of Dataset1(LPAC)

Table 3.3: Correlation among targets for dataset1(LPAC)

Conditions	**tR_pH2.0**	**tR_pH3.5**	**tR_pH5.0**	**tR_pH6.5**	**tR_pH8.0**
tR_pH2.0	1.00	0.98	0.92	0.85	0.84
tR_pH3.5	1.00	0.98	0.92	0.85	0.84
tR_pH5.0	0.92	0.96	1.00	0.97	0.96
tR_pH6.5	0.85	0.90	0.97	1.00	1.00
tR_pH8.0	0.84	0.89	0.96	1.00	1.00

- **Data Source and Availability:** The source of the dataset is the published article of our own in the journal Data in Brief[131].
- **Data Structure:** The structure of the dataset is tabular and consists of 97 rows(compounds) and 229 columns(including five targets(retention times at five pH), and molecular descriptors). SMILE structures were used as input for calculating the molecular descriptors.
- **Data format** varied when used in the Single target approach of QSRR(Chapter 4.1) and Multitarget approach of QSRR(Chapter 4.2).
  In the Single target approach- Feature values varied at every pH corresponding to the target pH value. The method of calculation is given in the supplementary file of Chapter 4.1, also in paper [19].
  In the multitarget approach of QSRR, all features were constant at every pH, and hence, a constant set of data matrics with multiple target column was used to build one MT-QSRR model.

- **Feature Description:** Molecular descriptors consisted of physicochemical descriptors having constitutional, topological and geometrical descriptors having 1D, 2D, and 3D information about the compounds and image descriptors.
- **Target Variable:** Retention times at five different pH(ph 2.0, pH 3.5, pH 5.0, pH 6.5 and pH 8.0) were used as the target variable(Figure 3.2).
- **Tools and Software:** The rdKit package available in Python was used for the calculation of both physicochmical and image based molecular descriptors. Chemicalize was used to draw the structures and gather the values of the logD at all pH.

3.2.4 DataSet-2 SMRT

The dataset is derived from Reversed Phase Liquid Chromatography experiments for small pharmaceutical compounds. The dataset is comparatively bigger than the LPAC dataset, which helped build QSRR models in Chapter 4.3. The chemical taxonomy of the data has been shown in Figure 3.3 and the distribution of retention time in the complete dataset is shown in Figure 3.4

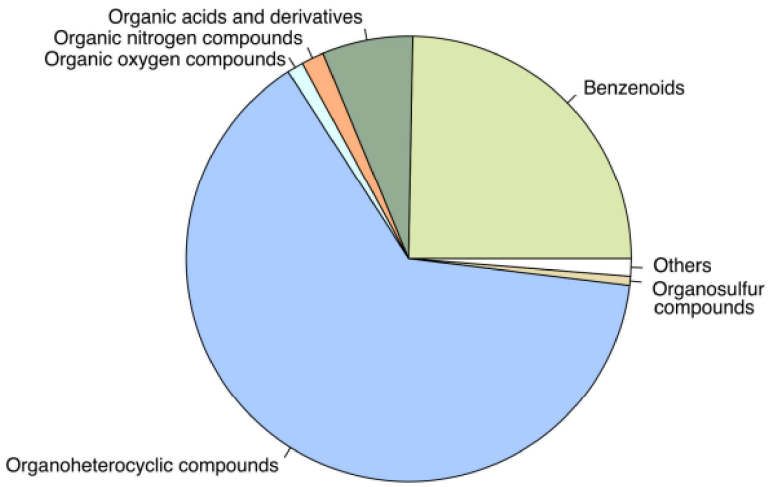


Figure 3.3: Chemical taxonomy of dataset2(METLIN) [13]

- **Data Source and Availability:** The source of the dataset is the published dataset in [13]. It is freely available to download.

- **Data Structure:** The structure of the dataset is tabular and consists of 77k rows(compounds), 225 columns(Descriptors), and one target column for retention time(unit - minute).
- **Feature Description:** Similar to dataset1, dataset 2 also consisted of same physicochemical and image based descriptors.
- **Target Variable:** Retention times at one pH(pH 2.7) . Target distribution is shown in Figure 3.4.
- **Tools and Software:** The rdKit package available in Python was used for the calculation of molecular descriptors.

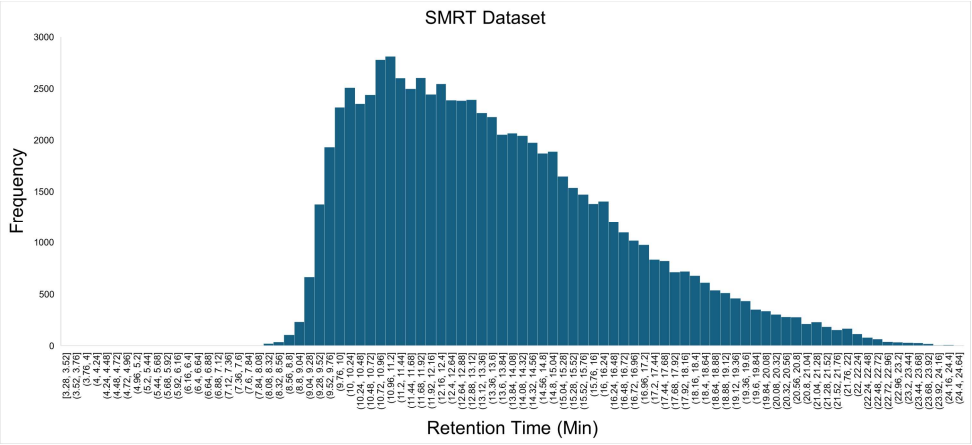
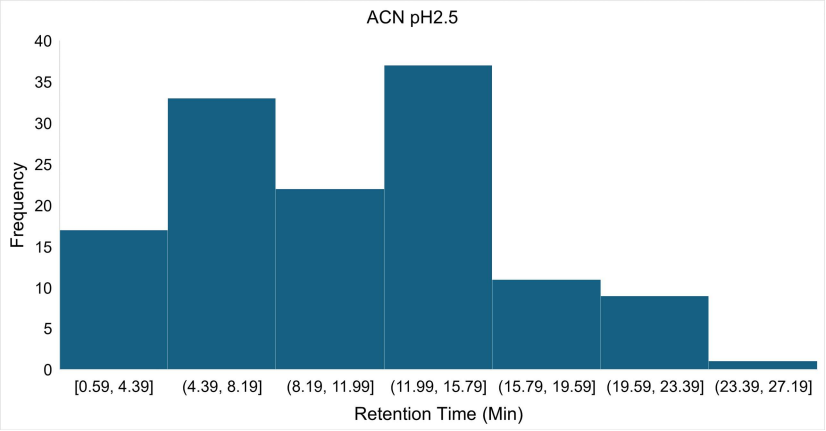


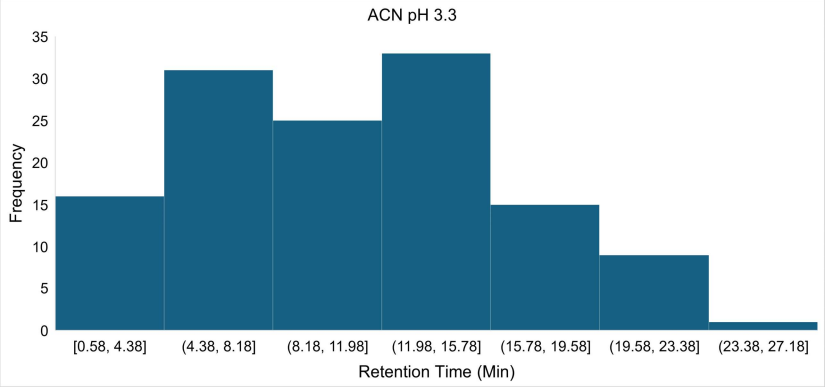
Figure 3.4: Target distribution for dataset2(Metlin)

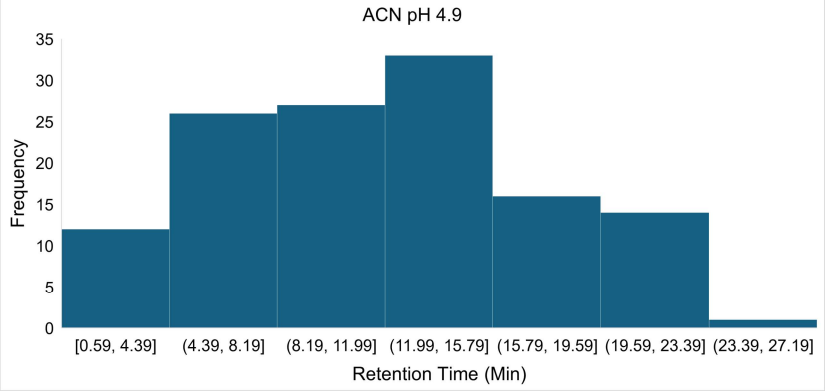
3.2.5 DataSet-3 ACN

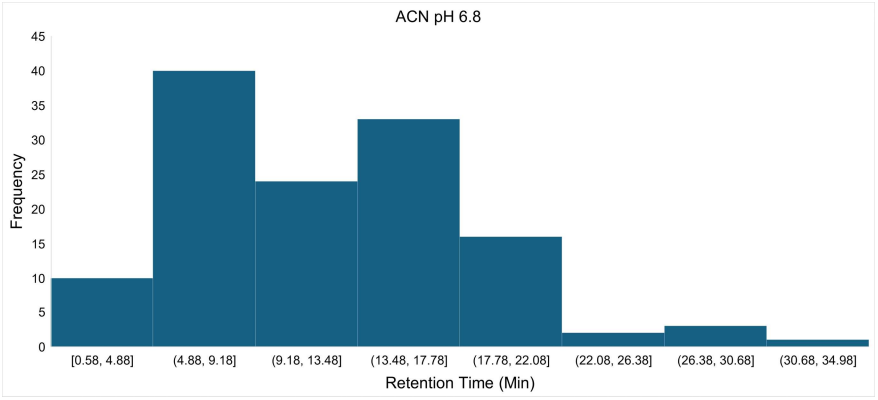
- **Data Source and Availability:** The source of the dataset is the published dataset and freely available to download[132].The data comes from the liquid chromatography experiment performed at temperature 25 degree celsius, with Acetonitrile(ACN) as mobile phase. Out of all(nine pH), we took ACN data at five different pH which are close to the LPAC dataset that is our own inhouse dataset.
- **Data Structure:** Five target columns for retention time of 130 compounds, and 130 images(2D))
- **Feature Description:** Similar to dataset1, dataset 2 also consisted of same physicochemical and image based descriptors.

- **Target Variable:** Retention times at five pH (pH 2.5, pH 3.3, pH 4.9, pH 6.8, pH 8.9). Data distribution is shown in Figure 3.5 and the correlation among them is shown in Table 3.4
- **Tools and Software:** The rdKit package available in Python was used for the calculation of image based molecular descriptors.









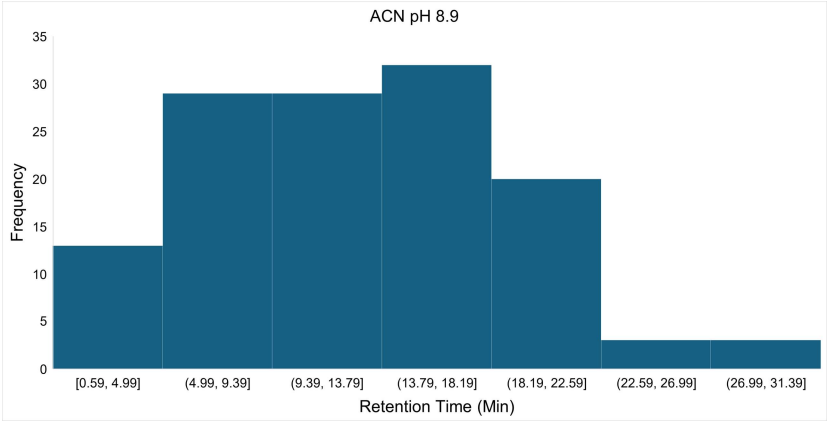


Figure 3.5: Target distributions of Dataset3(ACN)

Table 3.4: Correlation among targets for dataset3(ACN dataset)

Conditions	**tR_pH2.5**	**tR_pH3.3**	**tR_pH4.9**	**tR_pH6.8**	**tR_pH8.9**
tR_pH2.5	1.00	0.97	0.90	0.72	0.61
tR_pH3.3	0.97	1.00	0.91	0.73	0.63
tR_pH4.9	0.90	0.91	1.00	0.90	0.81
tR_pH6.8	0.72	0.73	0.90	1.00	0.95
tR_pH8.9	0.61	0.63	0.81	0.95	1.00

3.2.6 DataSet-4 RIKEN

- **Data Source and Availability:** The primary RIKEN dataset was published in Nature Methods, as cited in Tsugawa et al., 2019 [138]. The dataset was subsequently downloaded and partitioned similarly as by Kensert

et al. for their graph neural network (GNN)-based QSRR modeling [139, 32].

- **Data Structure:** The structure of the dataset is tabular and consists of 852 rows(compounds), 219 columns of molecular descriptors, and one target column for retention time)
- **Feature Description:** Similar to dataset1 and dataset 2, this dataset is consisted of similar physicochemical descriptors.
- **Target Variable:** Retention times at one pH(pH 2.7). Distribution is shown in Figure 3.6.
- **Tools and Software:** The rdKit package available in Python was used for the calculation of physicochemical descriptors.

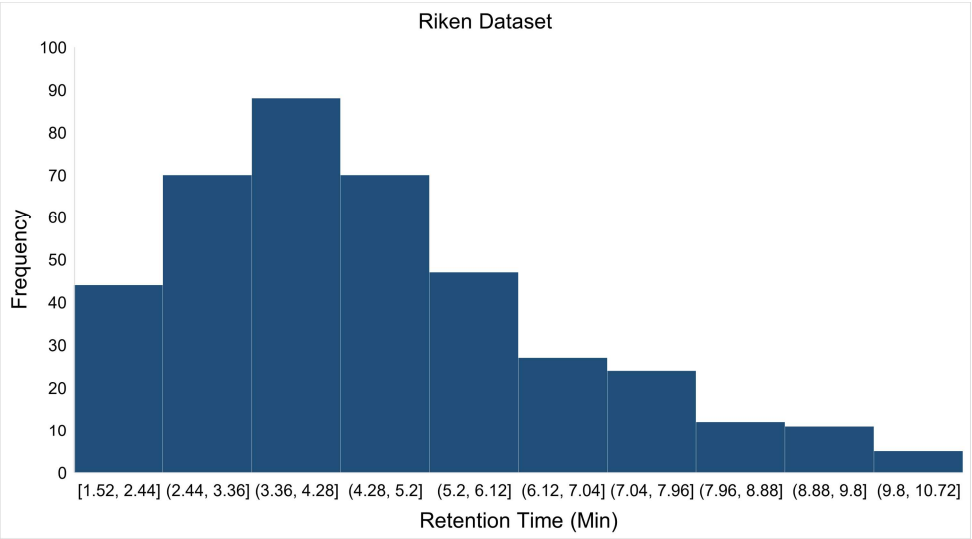


Figure 3.6: Target distribution of dataset4(RIKEN)

4.1 Single Target QSRR

4.1.1 Preamble

The use of different regression techniques to predict retention times in reversed-phase liquid chromatography (RPLC) is covered in detail in this chapter. It focuses on building Quantitative Structure-Retention Relationship (QSRR) models for five pH levels (pH 2.7, 3.5, 5.0, 6.5, and 8.0) utilising a single-target approach. Along with other conventional machine learning techniques, stacking method with MLR as meta learner was innovatively used for predictions. The goal is to improve the QSRR modelling procedure by the incorporation of several modelling techniques. We aim to identify the key molecular descriptors that affect retention times at different pH levels using broad categories of feature selection techniques. Strong and consistent processes are ensured by the modelling strategy's adherence to the OECD requirements. This inquiry is intended to function as a starting point for any QSRR modelling project, which may then be customised as per nature of the targets.
This work has been published in the journal "Molecules" Here is the citation:

> ***Kumari, Priyanka, et al. "Quantitative structure retention-relationship modeling: Towards an innovative general-purpose strategy." Molecules 28.4 (2023): 1696.***

4.1.2 Abstract

Reversed-Phase Liquid Chromatography (RPLC) is a common liquid chromatographic mode used for the control of pharmaceutical compounds during their drug life cycle. Nevertheless, determining the optimal chromatographic conditions that enable this separation is time-consuming and requires a lot of lab work. Quantitative Structure Retention Relationship models (QSRR) are helpful for doing this job with minimal time and cost by predicting retention times of known compounds without performing experiments. In the current work, several QSRR models were built and compared for their adequacy to predict the retention times. The regression models were based on a combination of linear and non-linear algorithms such as Multiple Linear Regression, Support Vector Regression, Least Absolute Shrinkage and Selection Operator, Random Forest, and Gradient Boosted Regression. Models were built for five pH conditions, i.e., at pH 2.7, 3.5, 6.5 and 8.0. In the end, the model predictions were combined using stacking and the performances of each models were compared. The k-Nearest neighbor-based application domain filter was established to assess the reliability of the prediction for further compound prioritization. Altogether, this study can be insightful for analytical chemists working with RPLC to begin with the computational prediction modeling like QSRR to predict the separation of small molecules.

4.1.3 Introduction

Liquid chromatography (LC) is widely used in the context of identification and assay of analytes present in a mixture. Several modes such as normal phase liquid chromatography (NPLC), reversed-phase liquid chromatography (RPLC) or hydro-philic interaction liquid chromatography (HILIC) are available. All these modes are based on the same principle where analytes are present in a liquid mobile phase and are passed through a column containing solid stationary phase under high pressure. The retention time (tR) observed is the time taken by the analyte to travel across the column and is dependent on the difference of interaction of the analyte with mobile and stationary phases at varied conditions. Several experimental parameters may influence these interactions leading to a separation of the compounds. Among these, the composition of the mobile phase (i.e., pH, organic modifier, gradient elution) and the stationary phases must be selected. Given the multiple possibilities, finding an optimal condition for such separation is generally performed on a trial-and-error basis and largely depends on the

researcher's prior knowledge. This, in turn, becomes time and resource-consuming and represents a significant bottleneck of LC analysis in many domains [140]. Quantitative Structure-Retention Relationships (QSRRs) modeling was proposed as an alternative solution to optimize the method development phase[36, 141].

QSRR models are computational models that establish a statistically significant relationship between a chromatographic retention parameter and molecular descriptors, which are numerical quantities carrying physico-chemical information of the molecules [142]. Such prediction models could be applied to any type of separation analysis irrespective of the chromatographic techniques or even the modes of a particular technique. Hence, its application range covers many interesting systems such as TLC [142], GC [143], IC [65], RP-LC [47, 49], and HILIC chromatography modes [144].

QSRR model development not only enlarges the range of applications but also increases the understanding of the separation mechanisms. There are several ways of QSRR modeling including the models based on mechanistic equations [122] or based on machine learning methods. The latter are quite popular because of their efficiency and the availability of multiple algorithms. The support vector (SVR) and Partial Least Square (PLS) models are the most popular options [140, 60, 89, 69, 145, 146, 43], but other types of regression algorithms such as Gradient Boosting Regression (GBR), Random Forest, Neural networks etc., have been successfully applied [40, 147, 148, 149].

Most of the recent machine learning algorithms can be severely limited in accuracy and applicability by the size and nature of the dataset, number and type of descriptors, etc. However, the LC datasets are generally small because of the time and resources needed to build it. Therefore, most modelling strategies imply a feature selection step to avoid overfitting and ensure sparsity of the models since; sparse models being generally more robust. Hence, to achieve it multiple strategies of descriptor selections have been used and shown to have performing differently on different datasets. A feature selection comparison study proposed by Goodarzi et al. showed that models built on descriptors selected by ant colony optimization algorithm coupled with SVR regression could be an excellent alternative for retention prediction modeling [89]. Zuvela, Petar, et al used a PLS regression model built on molecular descriptors selected by a genetic algorithm (GA), particle swarm optimization (PSO), artificial bee colony (ABC), firefly algorithm (FA), and flower pollination algorithm (FPA) [69], whereas Krmar, Jovana, et al compared a combination of linear (MLR) and nonlinear model (SVM) based on a preselected feature set [150]. Pastewska,

Monika, et al. and Ulenberg, Szymon, et al. used Genetic algorithm coupled with MLR(Multiple Linear Regression) for [151, 152]. At the same time, there are models which are based on Bayesian approach that involves using prior knowledge and represented as probability distributions to make retention time predictions. The prior knowledge is combined with experimental data to produce a posterior probability distribution, which provides a prediction of the retention time. The choice of mechanistic descriptors and the form of the prior distributions can have a significant impact on the accuracy of the predictions made by the model[132, 153, 154]. Since QSRR models are computational models, prediction discrepancies are frequent because of overfitting which, in turns, question the reliability of their practical use on new untested chemical compounds. Therefore, it is a good practice to review the model's validity as per Organization for Economic Cooperation and Development [155]. Although few research studies have checked applicability domain [147, 156, 157, 158], it is still very rare where all QSRR models are accompanied with such validations.

When looking at the literature, the proper well-structured strategy to get started with the structure-derived retention modeling, i.e., choice of descriptor set, and the selection of a specific regression algorithm is not clearly defined yet. Most studies are based on the researcher's previous experience or the most cited methods in the literature pool. Hence, a comprehensive generalized overview of the practical strategy when there is a limited dataset which is the most frequent scenario for such separation studies, would benefit to the field of analytical chemistry. Consequently, we propose a strategy that might be used in a variety of cases because of its conception (use of linear, nonlinear algorithms, use of diverse feature selection tools, and applicability domain of the use of selected model). Looking at the current time where deep learning approaches are dominating the ML space, applying them on small dataset is not feasible. Hence, this approach is versatile and useful even on small datasets.

4.1.4 Material and Methods

4.1.4.1 Dataset Collection

The dataset used in this study was built in-house [159] and consists of retention time observed for 98 small pharmaceutical compounds reported in minutes. List of small molecules in the dataset came from [**?** 160, 161]. The compounds were tested for their druglikeliness(following Lipinski's rule) using SwissADME tool [162] and more than 90% of the compounds followed all rules of Lipinski's representing the usefulness of the trained model for other

druglike molecules too. Moreover, the compound selection was done as such that apart from RPLC they could be relevant for other chromatographic modes such as ionic (IC) and hydrophilic interaction (HILIC). Hence, the strategies developed for one mode can be expandable on another. The data was acquired in RPLC mode using a Waters XSelect HSS T3 (100x2.1 mm, 3.5μm) column at 25°C, with flow rate 0.3 ml/min at five different pH conditions- 2.7, 3.5, 5.0, 6.5, and 8.0 with a gradient elution of 0-95% of methanol in 20 minutes time.

4.1.4.2 Molecular Descriptors and their Calculation

Molecular descriptors play an important role in achieving accurate retention prediction. They form the firm basis for any QSRR model. For regression models a set of 1D and 2D descriptors covering physical, chemical and structural properties were calculated for every molecule in the dataset using their SMILE structure taken from PubChem database [163]. The molecular descriptors in this article are calculated taking the ionization state of the compound at the pH of interest into account with the weighted average where the weights are the percentage of distribution of the microspecies at the considered pH. (Described with example in **Supplementary file S1(Appendix)** in appendix). The ionization states were obtained from Chemaxon software and the descriptor values were calculated using RdKit library. An additional descriptor, logD was added in the final molecular descriptor set. The value of this descriptor was calculated by Chemaxon at the value of the pH of interest. Thus, a total of 229 molecular descriptors were computed for each molecule at each pH condition (names of descriptors are mentioned in Supplementary file Table S2 in appendix).

4.1.4.3 Data cleaning and preprocessing

There are five datasets (varying with ph-2.7, 3.5, 5.0, 6.5, 8.0) used in this study. Each dataset consists of 97 rows and 239 columns initially. All compounds with retention times below 2 minutes at all pH were removed. Zero variance descriptors were also removed. Filtered feature names are mentioned in **Table S2(Appendix**). Since our dataset had features with values of different ranges hence, the final dataset was standardized before QSRR modeling. The first step involved mean centering and in second step data values were divided to standard deviation making the variance of variable to 1 and mean 0. Final dataset had 67 compounds for modeling and 10 compounds in external test set at each pH.

4.1.4.4 QSRR modeling with feature selection

The choice of regression techniques for correlating structural descriptors with the analyte's experimental retention time plays crucial role in constructing best and efficient QSRR models. There are no best algorithms defined for such retention predictions. One type of algorithm can work better for one problem but fail to achieve the same level of accuracy on another. The performance of such regression models depends also on quality of the dataset. Out of many available descriptors, there is a high chance of some redundant, noisy or irrelevant features in the starting dataset that can create problems in retention prediction: the curse of dimensionality, overfitting problems, high training time for model construction, and poor generalization ability of built models are amongst a few of them [164]. Therefore, a more systematic strategy adapted for QSRR methods is required to determine the possible preliminary, intermediate, and final steps to achieve the absolute accuracy of the best selected QSRR models.

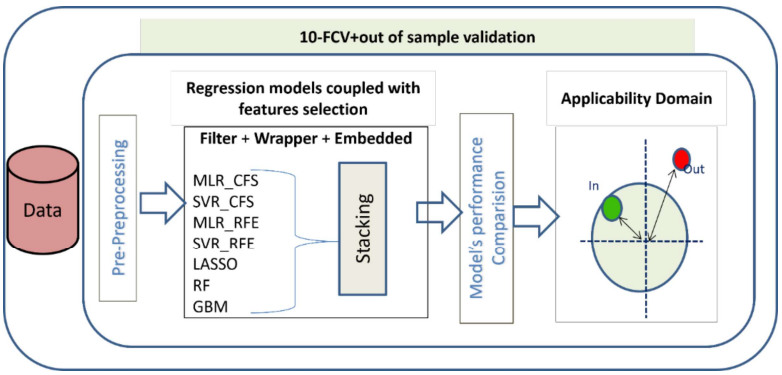


Figure 4.1.1: Workflow describing the steps of QSRR Modeling

Consequently, a well-organized strategy is proposed here (Figure 4.1.1). Five machine learning algorithms-Multiple Linear regression (MLR), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Regression (SVR), Random Forest (RF) and Gradient Boosting Regression (GBR). These algorithms were coupled with three feature selection method (i) filter (correlation-based filter) (ii) wrapper (Recursive Feature Selection methods, RFE) and (iii) embedded methods were compared for their prediction abilities using small molecule data sets.

4.1.4.5 Combining multiple predictions using Stacking

At the end multiple predictions from single were combined using stacking algorithm. Model stacking is an ensemble method that uses a meta learner to club the predictions from single learners and then combine them to get the final predictions [34, 165]. Two level model architecture (Figure 4.1.2) was used to build stacking regressor with a hypothesis that combining individual model's predictions would increase the prediction performance. At level 1, all base learners are built and optimized to get the best individual predictions. At level 2, the meta learner combines the predictions coming from level-1 models. Predictions made on external test data was used to test the stacking model. The simplest and most widely used algorithm (MLR) was chosen as meta regressor.

All models were built using 10-Fold cross validation and RMSE was used as performance metric. The model with top ranking (based on ranking over all data sets with sorted RMSE) was selected as the best algorithm for retention prediction of small molecules in RPLC.

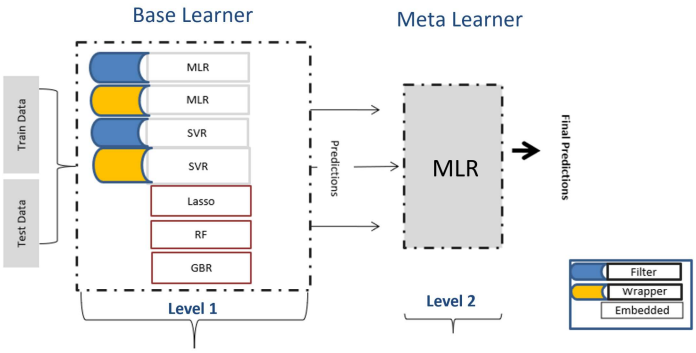


Figure 4.1.2: Architecture of Stacking used in this study

Algorithms

As shown in figure 4.1.2, five algorithms used at level-1. LASSO regression is a type of linear regression that uses shrinkage by applying penalty equal to the absolute value of the magnitude of coefficients (L1 regularization) [166]. The LASSO procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited

for models showing high levels of multicollinearity. SVR with radial basis function kernel (RBF) was used to check the nonlinear dependencies. SVR provides the flexibility to define how much error is acceptable in the model and will find an appropriate line (or hyperplane in higher dimensions) to fit the data. The objective function of SVR is to minimize the coefficients more specifically, the L2-norm of the coefficient vector [167]. The error term is handled in the constraints, where one set the absolute error less than or equal to a specified margin, called the maximum error, ϵ (epsilon).

RF and GBR both are ensemble learning methods and predict by combining the outputs from individual trees (Tree based regressions) [168, 169]. They differ in the way the trees are built: the order and the way the results are combined. The main objective of RF which represents bagging, is to create several subsets of data from training samples chosen randomly with replacement. Every subset data is used to train their individual trees resulting an ensemble of different models. Average of all the predictions from different trees are used for predictions. In contrast to random forest regression, in GBR the learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. Consecutive trees (random sample) are fit and at every step with the goal to improve the performance from the prior tree by applying different weights. Hence, in turn this process converts weak learners into better performing model.

4.1.4.6 Hyperparameter Optimization

To customize and get most out of QSRR models, hyperparameters were configured using grid search that allowed models to be customized for specific task on all the datasets. Optimization was done using 10-Fold cross validation and RMSE was used as performance metric. Grid Search works by defining a search space or hyperparameter values in the form of a grid and evaluate each and every position in that grid. The hyperparameters set with least RMSE were selected to build the prediction models. Grid search built in caret package itself was used for optimized parameter search.

4.1.4.7 Applicability domain

K-Nearest Neighbors (KNN) method has been used to calculate the AD of models. By this method we calculate the distance of query compounds from a defined point within the descriptor space of the training data [170]. In this method, the average Euclidean distances of training molecules are calculated from their k nearest training neighbors. Average distance value corresponding to a user-defined percentile is considered as threshold. Those test compounds that have average distance from their k closest training

neighbors greater than this threshold are reported to be out of the scope of the model's applicability and vice versa. In the present study, a k=5 number of nearest neighbors and a 95th percentile was selected to compute the AD.

4.1.4.8 Model Validation

Model validation step which accounts for the fourth principle of OECD, ensures the predictability, and reliability of the QSRR model to evaluate the credibility of the model's predictions on any new set of data. In the current study, the predictive abilities of QSRR regression models were assessed using 10-fold cross validation and external validation test data. In 10-fold cross validation the compounds in the dataset were randomly divided into 10 partitions of equal size. Nine parts were used for training while the last tenth was used as test set. The process was repeated ten times in such a way that each sample was used exactly once as the test data in each cycle. There are many performance comparison metrices available in the literature to compare the generalization performance of fitted regression models for example- mean absolute error (mae), percentage mean absolute error(%mae), root mean square error(rmse), percentage root mean square error(%rmse), R^2 for evaluating the predictive ability of quantitative structure-retention relationships (QSRR) models [171] but, in the current studyRoot Mean Squared Error (RMSE) and R^2 are used for the same. The reason being that these two are considered an excellent general-purpose error metric for numerical predictions in most of the QSRR studies reported in the literature.

4.1.4.9 Tools and Software used

RDKit library in Python version 2021.09.5 [172] and Chemicalize were used for calculation of molecular descriptor set. Statistical evaluation of the data: preprocessing, feature selection and regression prediction has been performed using Caret package in R version 3.6 [173]. GGplot2 available in R was used for plotting observed versus prediction plots and MS excel was used for plotting bar plots [174]. Applicability Domain toolbox was used for the applicability domain calculation for prediction models [84].

4.1.5 Results and discussion

In this study a simple, clear and well-defined strategy (Figure 4.1.1) for QSRR modelling is proposed which can be referred to use when the new

test molecule structures are known. Seven diverse machine learning algorithms coupled with three methods of feature selections were evaluated for their retention time prediction abilities. The regression algorithms and feature selections were chosen based on the fundamental difference in their working mechanism so that the strategy could give a holistic view of the performances of variety of methods suitable for such predictions. The selected regression algorithms were a combination of linear and nonlinear methods based on single modelling and ensembles too. Ensemble models were a combination of methods that take advantage of bagging or boosting. The molecular descriptor dataset used for all regression models were varied according to the method of feature selections applied on the dataset. Since linear regression modelling could not handle multicollinearity issue hence, they were coupled with feature selection before proceeding for regression prediction. These comparative methods provide the insights about the applicability of varied models with feature sets for users when there will be insufficient or complete lack of domain knowledge or when there will be a need to support expert knowledge to achieve higher prediction performances with given set of descriptors. The dataset associated with each step are as such: preprocessing and feature selection led us to have three types of datasets at each pH: [140] data where features were selected using filter method (e.g.-CFS) [36] data where features were selected using wrapper method (eg. RFE) and [141] data with all features remaining after preprocessing. All datasets were used for regression modeling and their predictive performances were compared in 10-fold cv and on the external test set.

4.1.5.1 Diversity of the dataset

It is expected that more diverse the dataset the better the trained models and their generalization performance on new test set. The diversity of the dataset was checked based on molecular weight and their chemical taxonomy. The molecular weight of the compounds varied from 46 to 456 g/mol. ClassyFire3 [175] was used to obtain a chemical taxonomy of molecules in the dataset using their smile structure(**Supplementary file S3(Appendix)**). Majority of molecules were classified into eight ClassyFire's groups on the level of superclass, namely: benzenoids (40.0%) organoheterocyclic compounds (29%), organic acids and derivatives (17%), homogenous non-metal compounds (5%), nucleosides, nucleotides and analogues (4%), organic oxygen compounds (2%), phenylpropanoids and polyketides (2%) and rest other compounds (1%) such as lipids and lipid molecules.

4.1.5.2 Comparison of feature selection methods

Our data was a high-dimensional QSRR data sets i.e., less data points than number of features. Hence, it was a prerequisite to apply a dimensionality reduction algorithm to make the models computationally less expensive and to improve their prediction performances. In this study there were three feature selection methods used that were coupled with regression prediction. (1) Filter methods: In this method variables were chosen regardless of the model building hence, these are robust and effective in terms of overfitting and computation time respectively. These methods work by estimating a relevance score based on a user-defined threshold to select the best-scoring features such as the correlation with the predictive dependent variable [176]. (2) In wrapper method, which is comparatively computationally expensive and prone to overfitting, exists as a wrapper around the predictive model algorithms and uses the same model to select the best features based on some performance measures for example RMSE in this study [176]. (3) The embedded Method is a mix of both filter and wrapper methods. Here, the feature selection process is embedded in the learning or the model building phase and is done with some penalty on unfavorable features. In other words, these algorithms have an intrinsic strategy of feature selection and overfitting prevention [177, 178, 92]. In this study, all three categories of feature selection methods were analyzed for their performances in accordance with their use in regression models. From Tables 4.1.1, 4.1.2, 4.1.3, 4.1.4, 4.1.5 the algorithm with feature selections embedded (RF and GBR) and wrapper method- RFE performed comparatively better at all pH. It is also interesting to note that the filter method (CFS) and wrapper (RFE), when coupled with non-linear regression methods, perform better than when coupled with linear methods. This could be understood in terms of multicollinearity in the dataset with features. Multicollinearity creates model instability. Better performance of embedded feature selection method could be justified by two arguments: firstly- they consider the interaction between features giving much closer and detailed information about the data pattern and secondly there is no issue of multicollinearity since they apply penalties on correlated features.

4.1.5.3 Important Features

Every microspecies of molecules exist in dynamic equilibrium during the separation process and their retention times varies with changing pH. Therefore, the weighted average of their features is expected to give a more informative and descriptive feature set. Good accurate QSRR models at multiple

Models	CV		External Test	
	RMSECV	R^2	RMSE	R^2
MLR_CFS	0.17	0.71	0.25	0.50
SVR_CFS	0.15	0.78	0.22	0.64
MLR_CFS	0.14	0.83	0.22	0.70
SVR_RFE	0.13	0.83	0.17	0.80
Lasso	0.13	0.84	0.20	0.70
RF	0.13	0.83	0.19	0.76
GBM	0.13	0.81	0.18	0.72
Stack	0.13	0.82	0.25	0.80

Table 4.1.1: Prediction performances of all models at pH 2.7

Models	CV		External Test	
	RMSECV	R^2	RMSE	R^2
MLR_CFS	0.15	0.79	0.34	0.41
SVR_CFS	0.17	0.72	0.25	0.53
MLR_RFE	0.14	0.81	0.30	0.58
SVR_RFE	0.13	0.89	0.21	0.70
Lasso	0.13	0.82	0.22	0.66
RF	0.14	0.81	0.21	0.70
GBM	0.15	0.80	0.24	0.50
Stack	0.12	0.87	0.18	0.77

Table 4.1.2: Prediction performances of all models at pH 3.5

Models	CV		External Test	
	RMSECV	R^2	RMSE	R^2
MLR_CFS	0.15	0.81	0.41	0.42
SVR_CFS	0.19	0.78	0.26	0.63
MLR_RFE	0.15	0.82	0.26	0.64
SVR_RFE	0.14	0.85	0.19	0.83
Lasso	0.13	0.87	0.23	0.71
RF	0.14	0.87	0.22	0.75
GBM	0.14	0.85	0.23	0.69
Stack	0.12	0.87	0.21	0.75

Table 4.1.3: Prediction performances of all models at pH 5.0

Models	CV		External Test	
	RMSECV	R^2	RMSE	R^2
MLR_CFS	0.20	0.76	0.31	0.58
SVR_CFS	0.23	0.73	0.35	0.44
MLR_RFE	0.16	0.87	0.29	0.63
SVR_RFE	0.16	0.88	0.19	0.84
Lasso	0.16	0.81	0.28	0.71
RF	0.15	0.87	0.20	0.84
GBM	0.15	0.88	0.15	0.90
Stack	0.13	0.90	0.18	0.85

Table 4.1.4: Prediction performances of all models at pH 6.5

Models	CV		External Test	
	RMSECV	R^2	RMSE	R^2
MLR_CFS	0.21	0.77	0.26	0.71
SVR_CFS	0.22	0.76	0.29	0.64
MLR_RFE	0.21	0.83	0.22	0.79
SVR_RFE	0.17	0.87	0.15	0.91
Lasso	0.15	0.89	0.30	0.70
RF	0.15	0.86	0.17	0.88
GBM	0.16	0.86	0.15	0.89
Stack	0.14	0.92	0.12	0.93

Table 4.1.5: Prediction performances of all models at pH 8.0

pH can give us information about the most relevant descriptors for the retention times prediction [179]. Better prediction performance of nonlinear models over linear models inferring those nonlinear patterns of molecular descriptors predict retention time relatively well. The following steps were followed to make the maximum inference about the selected features: All the features selected using the filter method and wrapper method and the top 20 features used by prediction models (embedded feature selections) were compared. Mutually inclusive features from all the models were selected as the most essential and representative features for retention time predictions. These selected features are listed in Table **Supplementary file S4(Appendix)**. LogD, MolLogP and PEOE_VSA6 are the most selected features by all the models at every pH. Apart from these, there were other features like- NHOHCount, VSA_Estate are also among the other selected features. The study has been done in reversed-phase liquid chromatography, where the difference in lipophilicity of the compounds causes is the main factor affecting the retention of the molecules. Hence, the selection of descriptors related to lipophilicity exemplifies better feature selections. LogD and MolLogP which are the pH dependent distribution coefficients and octanol-water partition coefficients respectively for every microspecies of a molecule i.e, neutral and ionized both. PEOE_VSA, which represents the partial atomic charge of the molecule, ranges from 1 to 14 based on the partial charge distribution. In PEOE_VSA parameter- PEOE denotes Partial Equalization of Orbital Electronegativities which is a charge calculation method and VSA signifies-Van der Waals Surface Area. It is interesting to note that out of 14; it is PEOE_VSA6 which denotes Van der Waals surface ar-ea having the atomic partial charge is in the range of -0.10 to -0.05, that was selected maximally [180]. "NHOHCount" gives the molecule's NHs and OHs count whereas "polarizability" is a measure of electric dipole or electronic charge dispersion in response to an external electric field. These descriptors can, in principle, distinguish between slight differences in a local region of two globally similar molecules. The use of such information, as given by logD, LogP and the PEOE_VSA descriptors, seems necessary to construct a robust and accurate in silico model from structural information of test compounds. These descriptors are a parameterized representation of the hydrophobicity displayed in all modes of RPLC for separation of varied kinds of analytes.

4.1.5.4 Predictive performance of the different algorithms on all datasets

Performance differences between the different QSRR models were evaluated in terms of RMSE and R^2 on all five datasets. For each data set, all compounds are used in a nested 10CV approach to assess the generalization performance. To validate the model, a separate test set of 10 molecules was used. Every model performance on test set was compared at each condition and reported. Grid search method was used tuning parameters. Tuned parameters for each model at every pH is listed in file **Supplementary file-S10(Appendix)**. The detailed CV results for RMSE and R^2 for each dataset are shown in Tables 4.1.1, 4.1.2, 4.1.3, 4.1.4, 4.1.5, respectively. Mean rank over all data sets (all pH) when the performance was sorted on RMSE was calculated to find the best suitable model for retention time prediction. From the Figure 4.1.3, it is evident that Stacking is the best algorithm and hence, can be used for retention time prediction for small molecules in RPLC setup. Linear models such as MLR (CFS, RFE) and LASSO are not performing very well. Figure 4.1.3 shows how stacking reduces the RMSE of models over other single models. Note that the ensemble methods like RF and GBM performed comparatively better than single models at lower pH i.e., at pH 2.7 and 3.5 emphasizing the fact that ensembling is a better way to fit nonlinear relations in a model. The SVR (nonlinear RBF kernel + RFE) model followed after them, performing well for data sets at extreme pH conditions i.e., at 2.7, 6.5 and 8.0. Stacking performance was comparatively similar to GBM, RF and SVR_RFE at pH 2.7. Apart from one pH, this algorithm performed consistently well throughout the given pH range. The minimum error of prediction was as little as 0.02. The highest prediction error was observed at pH 2.7. These observations support the fact that except in a few circumstances, out of all algorithms, stacking is most likely to show better generalization performance. More explanatory discussion about the performance of feature selection coupled with regression models can be provided using observed versus prediction score plots. The closer the fitted line is to the identity line, the better the model. The predicted and their corresponding experimental retention times for stacking model at all pH are plotted in Figure 4.1.4 and for the rest all models are plotted in **Supplementary files S5, S6, S7, S8, S9(Appendix**. Residuals i.e., the difference between predicted and experimental values for the stacking model plotted at all pH to get a closer look at the predictions (Figure 4.1.6). The residual distributions for all datasets validated the superiority of the stacking model. Note that, to the author's knowledge, the stacking algorithm has never been applied

before for retention time prediction in RPLC.

4.1.5.5 Applicability Domain Check

It is impossible to anticipate the whole universe of compounds when building a single QSRR model. Hence, there is a need to define the model limitations with respect to its structural domain and response space which can further be used to evaluate the ambiguity in the prediction of a given molecule relying on the structural similarity of molecules used in the development of the QSRR model. This structural boundary to determine the subspace of chemical structures for reliable property prediction is defined as applicability domain which is also the third OECD principle [170]. The query chemicals falling under the defined boundaries of the model are considered within the applicability domain and hence, their predictions will be considered reliable. The predictions of the other molecules which are outside the applicability domain won't be trusted. In cases like this study, where several QSRR models have been built for retention prediction of small molecules, the knowledge of applicability domain helps to compare the reliability of prediction by each QSRR model.

A KNN-fix method (section- applicability Domain in material and methods) at a distance of 95% confidence interval was used to define the applicability domain of the QSRR model concerning its structural domain and response space. It is observed that stacking outperformed the rest of the single models; hence, the study of this section was focused on the stacking models only. The error of predictions of all QSRR models for each compound were compared with the distances among features (all features) calculated using the KNN-fix method [181]. It can be seen in Table 4.1.6, Figure 4.1.5 and **Supplementary file-S11(Appendix)**, that the error of prediction at all pH was bad for compound Miconazole which turned out to be out of the applicability domain since its calculated distance was higher than the threshold at every pH.

Prediction performance and hence, the regression line and residual plot was better when plotted (Figures 4.1.5 and 4.1.6) after removing Miconazole from the external test set. Hence, it can be inferred that the retention time prediction of miconazole or any new test compound similar to this cannot be considered reliable. The calculated threshold could serve as a very good measure for filtering new test compounds for retention prediction. Detailed analysis of such behaviour of Miconazole was out of the scope of this study.

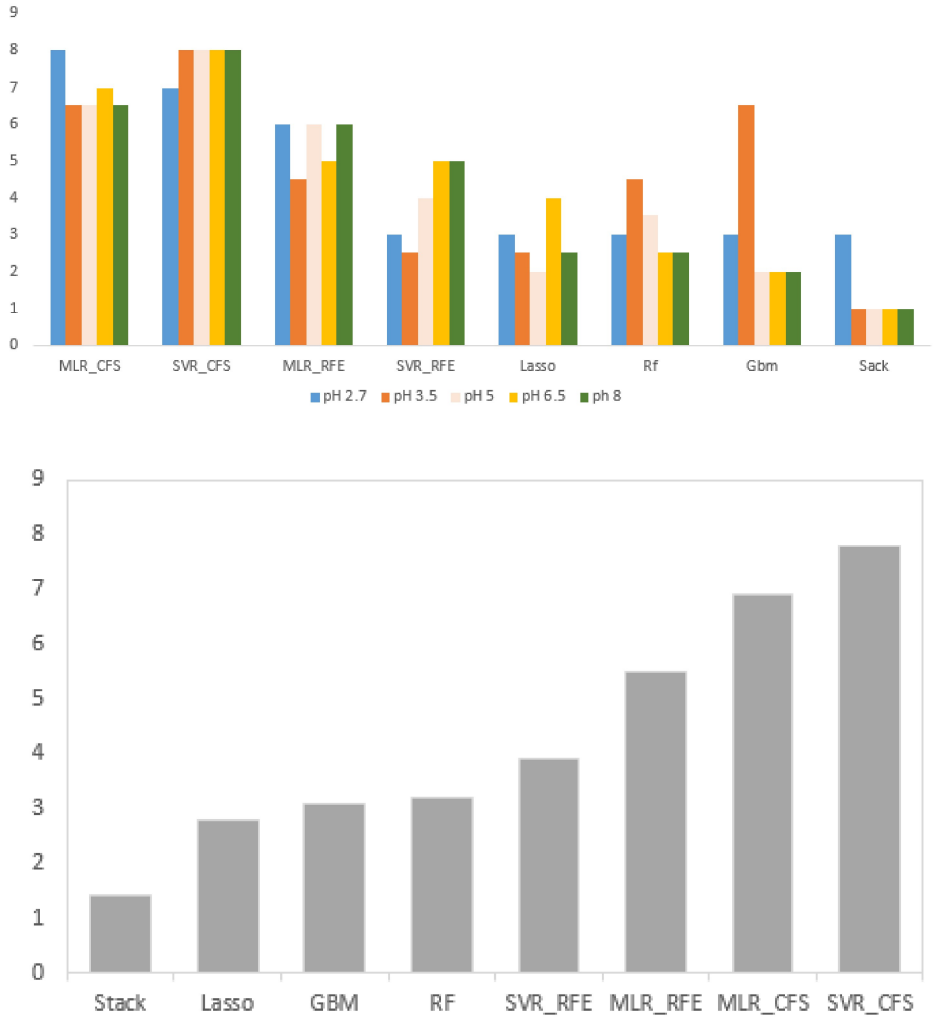


Figure 4.1.3: (a) Rank of every algorithm based on RMSE corresponding every target; (b) The mean rank over all data sets when the performance is sorted on RMSE

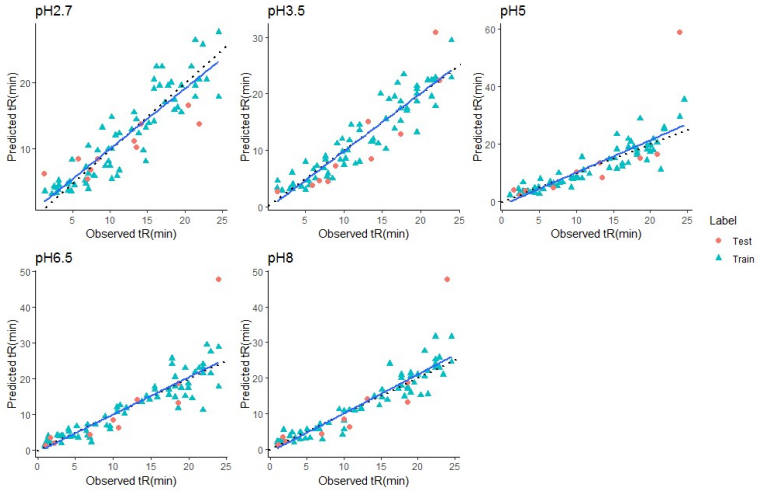


Figure 4.1.4: Predicted vs. Experimental retention times (in Min.) for Stacking model at all pH. (Blue line- Fit, Black dashed line-identity line)

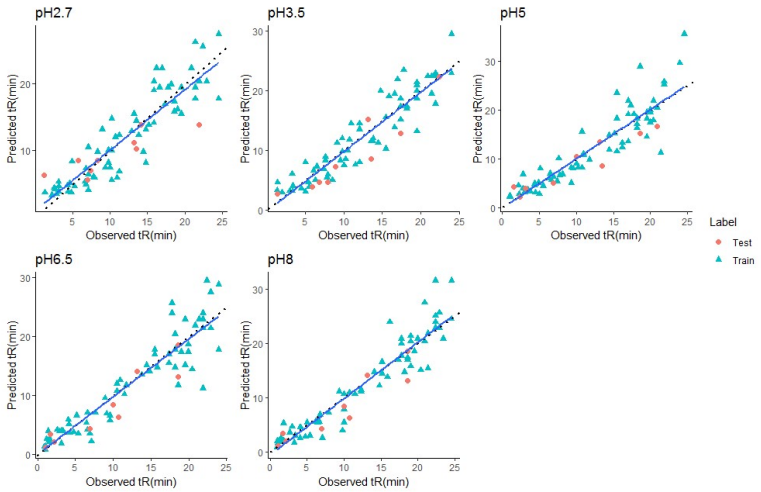


Figure 4.1.5: Predicted vs. Experimental retention times (in min.) for stacking model at all pH- After removing Miconazole (Blue line- Fit line, Black dashed line- identity line)

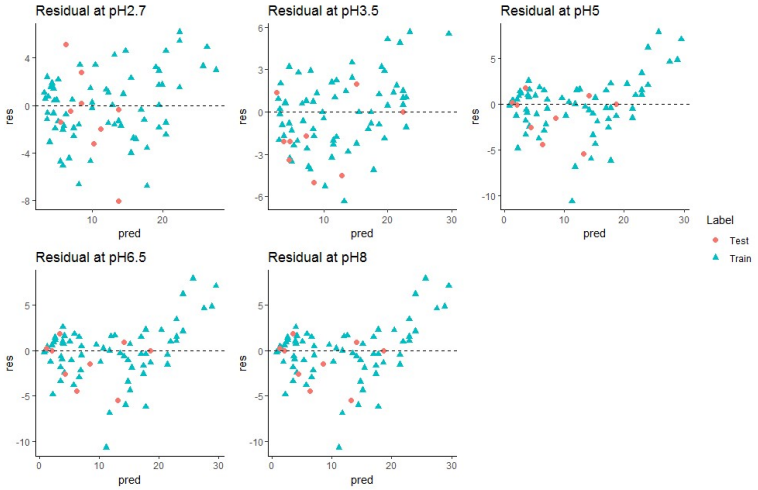


Figure 4.1.6: Residual plots (in Min.) for Stacking model at all pH (without Miconazole)

Table 4.1.6: Applicability domain calculated for each compound in the test set. (Errors in columns 2-6 are the errors of prediction from all the models specific for compounds. Distances in columns 7-11 are their distances calculated using KNN fixed methods). Errors are based on back-transformed retention times (min unit).

Compound	Error pH 2.7	Error pH 3.5	Error pH 5.0	Error pH 6.5	Error pH 8.0	Distance pH 2.7	Distance pH 3.5	Distance pH 5.0	Distance pH 6.5	Distance pH 8.0	Applicability
23dideoxyadenosine	0.49	1.67	0.47	1.49	2.82	13.86	13.57	12.87	13.15	13.35	In
mefenamic acid	8.07	0.00	4.30	5.44	1.60	11.32	11.34	11.43	11.37	11.40	In
cytosine	5.11	1.37	2.62	1.81	1.65	9.51	9.16	8.81	9.06	9.26	In
gallic acid	2.76	2.09	0.11	0.21	0.20	8.39	8.45	8.44	8.48	8.53	In
4aminosalicylic acid	0.19	3.37	0.80	0.05	0.37	5.84	6.20	6.21	6.14	6.21	In
2deoxyguanosine	1.42	2.08	1.91	2.55	0.47	12.77	12.39	12.37	12.64	12.36	In
miconazole	3.82	9.03	34.90	23.87	6.21	21.72	21.81	21.89	21.47	21.52	Out
chlordiazepoxide	0.32	4.50	3.49	0.00	0.84	11.50	11.54	11.70	11.52	11.86	In
4nitrophenol	1.96	4.98	4.98	4.41	1.96	7.66	7.73	7.77	8.05	9.12	In
coumarin	3.26	1.95	0.31	0.94	1.65	7.44	7.49	8.03	8.27	8.50	In
Threshold						15.87	15.86	15.88	15.64	14.99	

4.1.6 Conclusion

Chromatographic separation of small molecules is a complex process and the development of new separation methods may be a long and costly process. QSRR proved to be an alternative solution enabling the selection of pre-optimal conditions based on in silico computations. However, such computational modeling approaches become tricky with increasing number of chromatographic parameters. It is very challenging to use one type of algorithm over others since non-linear relationships between retention properties and the molecular descriptors may be present. The current study attempts to simplify the prediction modeling steps by taking a holistic approach that could be applied to any QSRR modeling for similar chemical compounds. Since structures of compounds play a vital role in deciding separation patterns, the type of molecular descriptors and the way they have been calculated is crucial. The influence of change in pH on structure-derived molecular descriptors gave a deeper and better understanding of molecules being studied and their retention pattern in the RPLC mode. The method of feature selection also affects the retention prediction performances. Stacking could be an excellent approach to combine predictions coming from different models and get better performances. QSRR modeling using a multitarget approach could be an advanced and more convenient way to deal with retention predictions with many experimental conditions. We expect that the current study will provide the initial guiding points for a practical and effective method for analytical chemists working with LC platforms to get an optional working condition and the way to improve the predictive confidence of studies.

4.2 MultiTarget QSRR

4.2.1 Preamble

Building upon the groundwork laid in the previous chapter, which focused on a single-target approach, it becomes apparent that this approach is sub-optimal when considering the intricate relationships among multiple diverse targets. That method did not consider the relationships among targets in the modeling process, resulting in potential information loss. Additionally, our attempt to model each target separately in the previous chapter proved to be time and resource-intensive. In response to these challenges, we introduce a multitarget approach in the current chapter to overcome the limitations of conventional single-target modeling in RPLC. We explored various methods for predicting retention times across multiple experimental condition variations simultaneously log transformed retention times at pH 2.7, pH 3.5, pH 5.0, pH 6.5, and pH 8.0).

This work has been published in the journal "Journal of Pharmaceutical and Biomedical Analysis."

Kumari, Priyanka, et al. "A multi-target QSRR approach to model retention times of small molecules in RPLC." Journal of Pharmaceutical and Biomedical Analysis 236 (2023): 115690.

4.2.2 Abstract

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

Keywords: Reverse Phase Liquid Chromatography, multi-target QSRR, Random Forest, molecular descriptors, Regression chain, Multitask learning, problem transformation, algorithm adaptation

4.2.3 Introduction

In the field of analytical chemistry, chromatographic separation has emerged as a powerful technique for separating and analysing complex mixtures. Extensive studies are conducted using various analytical techniques to gain a deeper understanding of the analytes present in a given sample, among which chromatography plays a prominent role. Retention time, a fundamental chromatography parameter, is a critical indicator of an analyte's behaviour within the chromatographic system and holds vital information for its separation and identification. It is often determined through a trial-and-error process, which can be time-consuming and expensive, especially when retention times need to be determined at multiple conditions. In the case of Reverse Phase Liquid Chromatography(RPLC), a widely studied type of chromatography, retention time(tR) can be influenced by various factors. These factors include pH, column type, mobile phase composition, and other variables encountered in various chromatographic techniques. As a result, accurately determining the retention time requires conducting multiple experiments to account for these variables effectively. This can become cost-prohibitive, particularly in high-throughput screening applications. An alternative way of retention evaluation is computational methods using quantitative structure retention

relationship models(QSRRs) [182, 122]. QSRR is an advanced approach that establishes a statistical relationship between various attributes, such as chemical, physical, and physicochemical properties, and the data associated with the structure of molecules, commonly known as structure-derived descriptors [183]. By carefully selecting appropriate molecular descriptors and utilizing statistical modelling methodologies, a QSRR model can be developed that is both statistically robust and stable [184].

The field of QSRR has undergone significant advancements, progressing from basic linear regression models to sophisticated machine learning algorithms, including algorithms like GA-PLS[185], Bayesian Ridge Regression, Extreme Gradient Boosting Regression, Support Vector Regression etc.,[186]. Traditionally, each study in QSRR has employed a single task or single-targeting approach, wherein a separate model is constructed for each response or target in regression studies. In recent studies [187, 150], researchers have delved into mixed Quantitative Structure-Retention Relationship (QSRR) models. However, these models predominantly depend on descriptors for target prediction, employing multiple algorithms and feature engineering. However, this approach overlooks the fact that a single molecule can elicit different responses under varying chemical environments and experimental conditions during separation. Consequently, this creates challenges related to multitasking. None of the previous studies has addressed this issue in retention prediction, where multiple experimental targets or responses are considered in the data, thereby neglecting the correlation between these targets. The time and cost required for modelling can vary significantly depending on the number of targets. Employing single-target approaches in QSRR models would not be time and cost-effective when multiple targets need to be predicted. Conversely, multi-target models would be more suitable in such cases.
While multitasking models have been utilized in other fields for activity prediction[188, 189], lipophilicity [190], toxicity[191], brain penetration[192], and more, the chromatography field has primarily overlooked their potential application. Some studies have explored multi-output regression in fields like real-time train arrival time prediction[193], ecological modeling[194], gas-phase kinetic rate constants prediction of chemicals[134], and chemometrics to infer concentrations of several analytes from multivariate calibration[195]. However, to our knowledge, none of the previous works has addressed the challenge of incorporating target relationships, including various retention times in varied conditions, into retention prediction models. Therefore, in this study, we aimed to explore different approaches to QSRR modelling for a comprehensive analysis.

In the literature, two methods of multi-target modelling have been reported [196]: (1) the problem transformation method and (2) the algorithm adaptation method.

- **Problem transformation method**: The problem transformation method involves converting the original multi-output regression problem into one or more single-output regression sub-problems, which can be solved using traditional single-output regression algorithms. Several techniques fall under this approach, including the *Independent Model (IM)*, where each output variable is modelled

independently using separate single-output regression models. The input features train each model separately, independently predicting each output variable. Another technique is the *Transformation-Based (TB) approach*, where the multi-output regression problem is transformed into a series of single-output regression problems by combining the input features with transformation functions. Separate single-output regression models are then trained for each output variable using these transformed features. An example of this approach is the chaining or regressor chain method [197].

- **Algorithm adaptation method**: The algorithm adaptation method involves modifying existing single-output regression algorithms to handle multiple output variables directly. This is a Multi Task Learning (MTL), where a single model is trained to predict multiple output variables jointly by optimizing a standard objective function that considers all the output variables simultaneously. The idea behind this approach is that the model can leverage the dependencies between the output variables to improve overall prediction performance.

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

Therefore, multi-output-multi-target prediction models, known as the "mt-QSRR modelling" approach, can be a more efficient alternative [198]. The practical utility of mt-QSRR models can be effectively extended and comprehended within the context of analytical method development, particularly for emerging pharmaceutical products. In such scenarios, where the "analytical quality by design" framework is followed [199], the implementation of the design of experiments (DoE) becomes imperative to establish a design space. This design space ensures that the chromatographic method exhibits desirable properties, including robustness in the face of experimental parameters[200]. However, conducting numerous laboratory experiments to identify optimal experimental conditions for the DoE can be time-consuming and resource-intensive. To address this challenge, the initial screening phase can be conveniently performed in silico utilizing one mt-QSRR model, even if their accuracy may not be exceptional. By employing these models, a range of parameters can be selected, significantly streamlining the subsequent experimental optimization DoE[168]. This allows for the identification of the most favourable separation and robustness conditions through practical experimentation not only in analytical chemistry but in other pharmaceutical and biomedical analysis as well [201, 202, 21]

In this study, we have compared the model performance of QSRR models based on single-target learning over multi-target learning(mt-QSRR) using retention data gathered for five pHs. Multi-target learning approach offers several advantages over single-target retention prediction[203, 204], including considering interdependencies between targets, reducing computational burden by using a single model, improving model interpretability, and training on larger datasets to enhance generalization and reducing overfitting[205, 206]. Multi-target QSRR

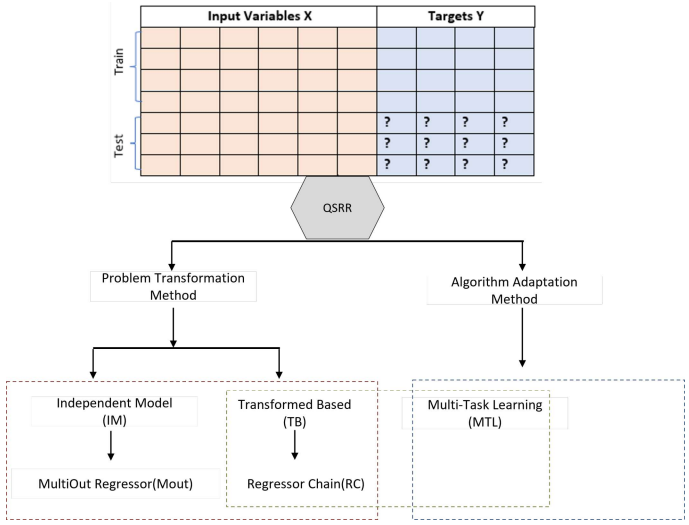


Figure 4.2.1: Different approaches of mt-QSRR models were implemented in this study. Red dotted box: Sequential multiple-output prediction methods with a single-target approach. Blue dotted box: Multi-output simultaneous prediction using a single model approach. Green dotted box: Modeling methods that consider the relationship of the target variable.

models(mt-QSRR) can significantly advance quantitative structure retention prediction and holds promise for applications in drug discovery, environmental analysis, and other fields where accurate and efficient retention times are critical for chromatographic separations.

4.2.4 Materials and methods

4.2.4.1 Problem definition

For a given data set P containing feature and target couple (x, y) with x ∈ X, the input vector and y $\in Y = Y_1 \times \ldots \times Y_n$ the target vector. Denote with $y_i \in Y_i$ the $i'th$ component of y.
Hence, the mt-QSRR model can be defined as:

$$Y_n = f(X)$$

In **single-target approach:** A learner learns from a data set $P = \{(x, y_i)\}$, with $y_i \in Y_i$ a scalar variable, a function $f_i : X \to Y_i$ such that $\sum_{(x,y_i)\in P} L_i(fi(x), y_i)$ is minimized, with L_i some loss function over Y_i.
In **multi-target approach:** A learner learns from a data set $P = \{(x, y_i)\}$, with y ∈ Y an n-dimensional vector, a function F: $X \to Y$ such that $\sum_{(x,y)\in S} L(F(x),y)$ is minimized, with L a loss function over Y . In this study, we have checked if the multi-target learner performs better than a single-target learner by checking for

any (x, y), drawn randomly from the population, on average, $L(\mathrm{F}(x), y) < \sum_i L_i(f_i(x), y_i)$.

4.2.4.2 Dataset

The dataset used in this study was taken from [207], which consists of retention time observed for small pharmaceutical compounds reported in minutes. The data were acquired in RPLC mode at five different pH conditions- 2.7,3.5,5.0,6.5,8.0 with a gradient elution of 0 to 95 % of methanol in 20 minutes. The column, flow rate and temperature specification are mentioned in [207, 19]. The dataset encompasses compounds with a diverse range of molecular weights, spanning from 46.005 to 454.611 g/mol. The efficacy and usefulness of a model rely heavily on the dataset it is trained on. Therefore, during the data collection process, we prioritised including a diverse range of molecules with varying pKa. This allowed us to capture different trends in retention times as the pH of the analysis increased. Four distinct types of data trends were observed, as depicted in Figure **(1-3)** in the supplementary file.

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

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

4.2.4.3 Molecular descriptors

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

Some of these descriptors were aligned with the parameters used in LSER theory, a concept initially applied in retention prediction models [2, 62]. LSER theory focuses on the linear solvation energy relationship, which relates solute retention to its solute-solvent interactions. These descriptors capture the specific solvation effects and improve the accuracy of retention prediction models. The remaining descriptors were included to provide additional meaningfulness to the model and enhance its predictive capabilities. The RDKit package, specifically version 2015, was utilized to compute these descriptors derived from the chemical structures [92].

4.2.4.4 Data cleaning and preprocessing

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

4.2.4.5 QSRR Modelling

Considering the given data description, our objective was to predict multiple continuous targets (responses) for new test samples based on a set of independent variables. Two approaches were used in this study (Figure 4.2.1) to predict the retention times: the single-target and multi-target regression approaches, which are explained in section 4.2.4.1.

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

Figure 4.2.2: Algorithm1: Pseudoalgorithm for DirectMultioutput Regressor(single-target approach used for Model1)

The problem transformation and algorithm adaptation methods for retention predictions were employed to check this differentiation. The problem transformation method converts the multi-output regression problem into one or more single-output regression sub-problems. Two ways of modelling were tested for this method- IDM and RC(Regressor Chain) methods corresponding to Model1 and Model2, respectively (shown as a red dotted box). On the other hand, the algorithm adaptation method involves modifying existing single-output regression algorithms to handle multiple output variables directly.

Both the RC(Model2) and MTL(Model3) models can handle target correlations but not the Independent model(Model1) that utilizes a multioutput regressor function to build the model. The pseudo algorithms for the three methods are described in Figure 4.2.2, 4.2.3 and 4.2.4, respectively:

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

Figure 4.2.3: Algorithm2: Pseudoalgorithm for RegressorChain method(single-target approach used for Model2)

This study focuses on applying a single-target approach (Model1 and Model2) and a multi-target approach (Model3) approaches to deal with the challenges associated with predicting the retention time of small molecules based on multivariate data. The high number of descriptors relative to the compounds used for modelling introduces the possibility of multicollinearity. To address this issue, we employ specific algorithms, with a focus on random forest (RFR) regression method [208, 134], which allows for the analysis of multivariate and megavariate data while mitigating the risk of overfitting. By utilizing random forest for retention time prediction, we effectively prevent overfitting and create a robust and reliable model that generalizes well to unseen data. This is achieved through the ensemble nature of the random forest, coupled with feature randomization, bootstrapping, regularization, and out-of-bag (OOB) error estimation. In our analysis, we developed three models using the sklearn library in Python. For Model 1, we utilized a multioutput wrapper around RFR (Random Forest Regressor). Model 2, on the other hand, employed a regressor chain around RFR. Lastly, for Model 3, we directly used the RFR function available from the sklearn.ensemble module. All models were constructed using the hyperparameter values as such: n_estimators=100, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=1.0. We employed the variable importance calculation for the features to determine the significant molecular descriptors, an inbuilt default function within the RFR algorithm. This allowed us to identify the descriptors that had the most impact on the predictive performance of the models.

4.2.4.6 Model Validation and evaluation

The developed mt-QSRR model underwent rigorous validation procedures to ensure its accuracy and reliability. Both internal and external validation methods

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

for *i in range p* **do**

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

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

Figure 4.2.4: Algorithm3: Pseudoalgorithm for Algorithm adaptation(multi-target approach used for Model3)

were employed, following a similar approach as outlined in [19], in order to minimize prediction errors across multiple target compounds. For the external validation, a dataset comprising ten compounds was carefully selected based on their diverse trends in observed retention time and chemical nature. This selection ensured that the model's performance was evaluated across a wide range of chemical properties, enhancing its applicability and robustness. By assessing the model's predictive capabilities on this external dataset, its generalizability and ability to handle various compound types were thoroughly assessed. Internal validation, on the other hand, was conducted using a 10-fold cross-validation technique. This method involved dividing the dataset into ten subsets of roughly equal size. The model was trained on nine subsets while utilizing the remaining subset for testing. This process was repeated ten times, each subset serving as the test set once. By performing cross-validation, the model's performance was assessed on multiple iterations, enhancing the credibility of its predictive capabilities. To evaluate the performance of the mt-QSRR model quantitatively, external validation performance measures were calculated. These measures were expressed in terms of the average root mean square error (aRMSE), as shown in equation 4.2.1, and the average coefficient of determination (aR^2), as shown in equation 4.2.2. These performance metrics provided a comprehensive assessment of the model's predictive accuracy and its ability to explain the variance in the observed retention times across multiple target compounds. By averaging the performance metrics over all the individual models, a consolidated evaluation was obtained, enabling a comparative analysis between single-target and multi-target prediction approaches. Furthermore, the individual model with the best performance was selected, and its predictions were compared against the corresponding observed values. This visual representation of the model's performance allowed for a more intuitive understanding of its predictive capabilities.

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

$$aRMSE = \frac{1}{d}\sum_{i=1}^{d}\sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(Y_i^{(l)} - \hat{Y}_i^{(l)}\right)^2} \tag{4.2.1}$$

$$aR^2 = \frac{1}{d}\sum_{i=1}^{d}\left[1 - \frac{\sum_{l=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{l=1}^{n}\left(Y_i - \overline{y}\right)^2}\right] \tag{4.2.2}$$

In the above-mentioned equations, Y and $\hat{Y}$ are the observed and predicted retention times of unseen test data, n is the number of test molecules, and d is the number of targets which is five pH conditions in this study.

4.2.4.7 Significance test for performance differences

To assess whether the differences in performance are statistically significant, we employed the corrected Friedman test [209, 210]. The Friedman test is a non-parametric test for multiple hypotheses testing. The algorithms were ranked according to their performances for each dataset separately. The best-performing algorithm was ranked 1, the second 2, and so on. In the situation where there were equal ranks, average rank was used. The Friedman test is based on two assumptions: The n K-variate random variables are mutually independent, i.e., the results within one row do not influence the results within the other rows (Table 4.2.2 and 4.2.3). The second hypothesis is that the data can be meaningfully ranked. Friedman's test statistic is:

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

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

4.2.5 Results and discussion

4.2.5.1 Data characterization

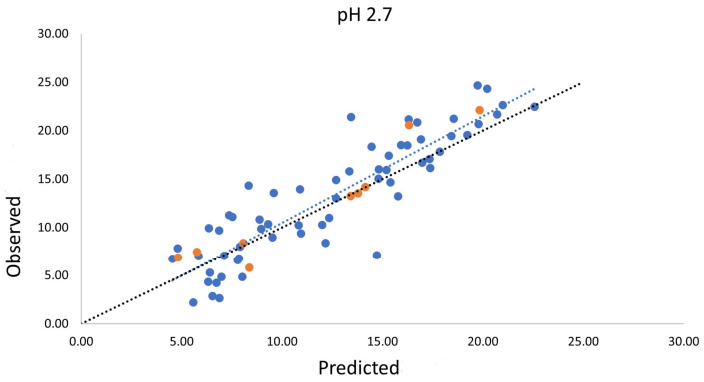
The multivariate dataset considered in this study comprised the experimental retention times(in minutes) of diverse small pharmaceutical compounds having varied molecular weights and retention times. The high correlation of retention values

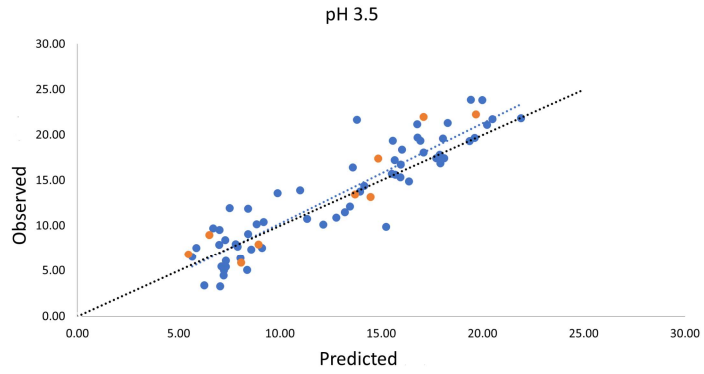
across all pH levels (Figure 4 supplementary file) underscores the importance of employing QSRR models that leverage this relationship for predicting retention times.

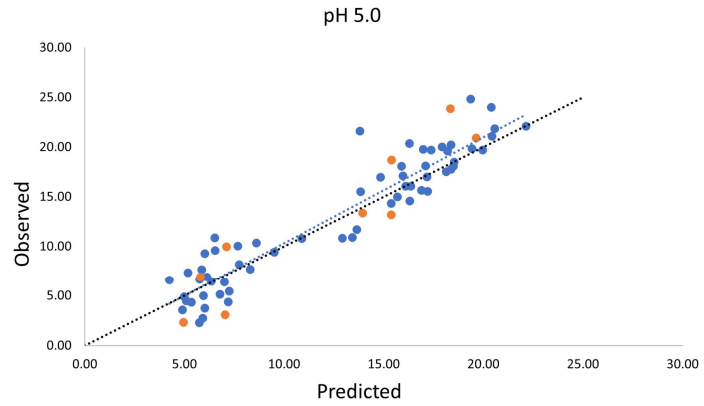
4.2.5.2 Multi-target QSRR modelling and validation

This study focuses on studying pH's influence on the retention behaviour of small molecules in RPLC. Here, we attempted to develop an mt-QSRR model for simultaneous prediction of multiple targets that are retention times (retention times at five pH of diverse pharmaceutical compounds). All the targets were experimentally observed as the dependent variables, and the considered compounds' molecular descriptors were calculated computationally as the predictor variables. The optimal model was established by utilizing a training set of 61 compounds. For the most effective model, a set of the top five descriptors was identified using Gini importance, also called mean decrease impurity. While additional descriptors do make a contribution, their importance is comparatively lower. The leading descriptor among them is "MolLogP," which signifies the octanol-water partition coefficient. Other noteworthy descriptors include "LogD," "VSA-Estate5," "SMR-VSA3," and "QED." The model was validated internally using a 10-fold CV and externally with the test set (n = 9). The parameters used for RFR for the mt-QSRR models were as such: $n\_estimators$ = 100, max_depth = None, $min\_samples\_split$ = 2, $min\_samples\_leaf$ = 1,$min\_weight\_fraction\_leaf$ = 0.0 , $max\_features$ = 1.0, $max\_leaf\_nodes = None$. Model1 and Model2 represented prediction from MultiOutput regression and regressor chain methods, and Model 3 as the Algorithm Adaptation method. The performance measures of three mt-QSRR models are given in (Table 4.2.1, 4.2.2 and 4.2.3). Table I displays the performance results based on the average root mean square error (RMSE) computed across all the targets using equations 1 and 2. The models captured 66-85 per cent of the variance in the test data(Table 4.2.3 and Figure 4.2.5). A high variance explained by a model implies that the majority of the information present in the data has been encompassed. Moreover, all the developed mt-QSRR models exhibited significantly low RMSE values (< 0.1) for both observed and predicted log values of the target in the test data (Table 4.2.2).

The regressor chain method (Model 2) performed poorly in comparison, suggesting that the effectiveness of chaining methods depends on the specific case. If the initial model's error is high, it may continue to increase with each subsequent target prediction. RMSE provides a measure of the average error in forecasting the dependent variable. The comparable RMSE values between the training and test sets indicate the usefulness of the algorithm adaptation method (model3-MTL)mt-QSRR model. Algorithm adaptation methods have been particularly advantageous in scenarios where the tasks exhibit notable commonalities. They utilize an inductive transfer approach for enhancing generalization in machine learning by leveraging the domain-specific knowledge present in the training data of related tasks. Better performances of this method can be considered effective for simultaneous prediction of multiple retention times due to the regularization it enforces


pH 2.7
Observed
Predicted
0.00
5.00
10.00
15.00
20.00
25.00
30.00


pH 3.5
Observed
Predicted
0.00
5.00
10.00
15.00
20.00
25.00
30.00


pH 5.0
Observed
Predicted
0.00
5.00
10.00
15.00
20.00
25.00
30.00

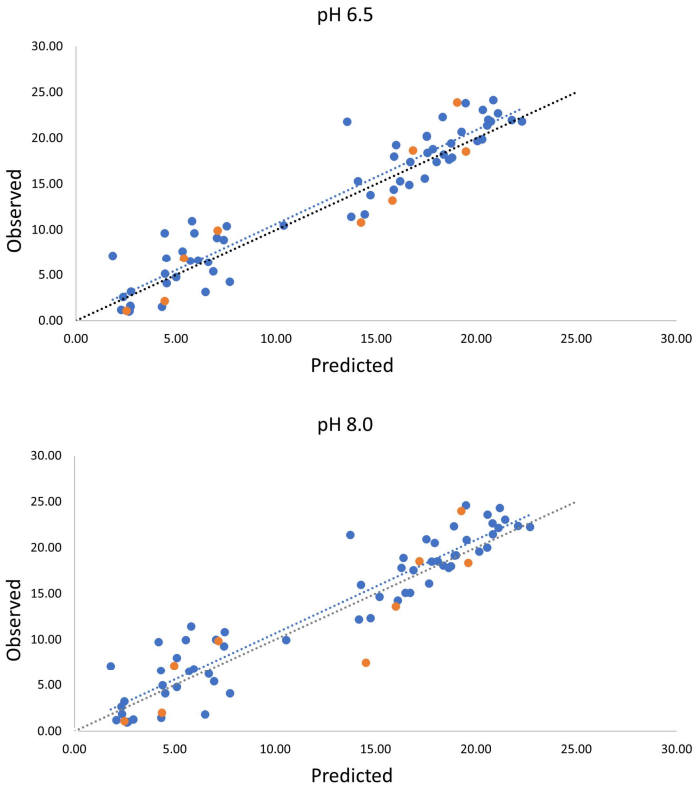


Figure 4.2.5: Plots of Observed retention time (tR) Vs. Experimental retention time (tR) from Model 3 (tR is back transformed in Minutes) for (a) pH 2.7, (b) pH 3.5, (c) pH 5.0, (d) pH 6.5, (e) pH 8.0. Blue points: train, orange points: test, fit line: blue dotted, Regular line: Black dotted

by demanding an algorithm to excel in correlated retention times with given five pHs, surpassing the regularization achieved by uniformly penalizing complexity to prevent overfitting. Significantly, the mt-QSRR model, which predicts multiple retention times simultaneously, demonstrates comparable performance to the single-target QSRR models, highlighting the significance of evaluating performance disparities (see 4.2.5.3). Additionally, the time needed for modelling consistently remained lower for mt-QSRR compared to predicting individual targets separately. In the single-target approach, each step had to be repeated multiple times based on the number of targets, whereas this repetition is unnecessary in the mt-QSRR modelling approach.

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

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

Table 4.2.1: Performance measures of each model based on combined prediction(average) of log tR

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

Table 4.2.2: Analysis of models for mt-QSRRs based on RMSE for individual targets

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

Table 4.2.3: Analysis for models for mt-QSRR based on R^2 for individual targets

Figure 4.2.6: Average rank of the models based on the RMSE (left) and R^2 (right).

Figure 4.2.7: Per-model rank based on the RMSE (left) and R^2 (right).

4.2.5.3 Comparison of the models

The comparison of the models based on their RMSE and R^2 are presented in Figure 4.2.6. On the axis, the algorithms are plotted according to their average rank across analyses. Note that for each analysis, the best model is ranked 1 and the worse is ranked 3. The corresponding radar plots are presented in Figure 4.2.7 as an alternative visualization of the ranks of the models for each analysis separately. In the radar plots, the lower the area in the coloured lines, the better. Overall, Figures 4.2.6 and 4.2.7 show that Model 1 and Model 3 perform better than Model 2 based on the RMSE, and Model 3 performs best based on the R^2. Hence, we recommend Model 3 for similar analyses. We used the Friedman test to detect whether the differences in performances of mt-QSRR models are statistically significant. The Friedman test concluded that the difference in the performance of these algorithms is not statistically significant (p-value > 0.05).

4.2.6 Conclusion

This study has successfully developed multiple multi-target QSRR (mt-QSRR) models involving a comparison between two modelling approaches: single-target and multi-target regression. The primary goal was to predict the retention times (tR) of a diverse range of structurally small molecules under various reversed-phase liquid chromatography (RPLC) conditions. The retention time prediction capabilities of the mt-QSRR model were assessed using three distinct methods. However, despite employing these diverse strategies, no statistically significant distinctions were observed in their predictive performance. The performance of the mt-QSRR models within our dataset indicated a reduction in efficiency as pH levels increased. Particularly, the regressor chain method exhibited higher root mean square error (RMSE), suggesting that retention prediction errors accumulate as they progress from lower to higher pH levels. One of the notable advantages of multi-target models is their interpretability in terms of the relationship between features and retention time variations with pH. Unlike single-target models, where descriptor importance varies per target specificity, the mt-QSRR model provides transparent insights into the pertinent input variables for predicting specific groups of response variables. Based on their performance, the optimal mt-QSRR model identified in this study highlighted five pivotal structural features: MolLogP, VSA-Estate5,

LogD, SMR-VSA3, and QED. These descriptors encompass the molecular partition coefficient, molecular surface area, distribution coefficient state index, and drug-likeness. These attributes are crucial in accounting for the diverse retention times observed for the considered small molecules across varying pH levels. In summary, our findings underscore the potential of mt-QSRR models as a more effective and efficient predictive strategy compared to constructing separate models for each target. Adopting the mt-QSRR approach holds the promise of streamlining efforts and reducing time and computational costs while simultaneously assessing the effective separation of molecules within the RPLC setup. Lastly, it is imperative to acknowledge that the test set encompasses a limited number of molecules, leading to an incomplete representation of the explored chemical space. As a result, the outcomes presented in this study are preliminary in nature.

4.3

Transfer Learning Enhanced mtQSRR

4.3.1 Preamble

In this section, we delve into a sophisticated method that employs artificial intelligence to predict the retention times in Reversed-Phase Liquid Chromatography (RPLC). Our exploration is driven by the challenge of scarce data in RPLC and seeks effective alternative solution.
We conduct an in-depth analysis to evaluate the impact of image-based descriptors and to assess how they measure up against traditional physicochemical descriptors. Furthermore, we scrutinize Quantitative Structure-Retention Relationship (QSRR) modeling techniques, contrasting single target models and multitarget models, as well as exploring the role of Transfer Learning. This examination follows on from the initial strategy outlined in Chapter 4.1, Chapter 4.2 and leading up to the refined strategy presented in this section.

4.3.2 Abstract

QSRR is a valuable technique for retention time predictions of small molecules. This aims to bridge the gap between molecular structure and chromatographic behaviour, offering invaluable insights for analytical chemistry. Given the challenge of simultaneous target prediction with variable experimental conditions and the scarcity of comprehensive datasets for such predictive modellings 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 were evaluated based on Mean Squared Error (MSE) and R^2 metrices. 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 optimising experiemntal 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 endeavours.

4.3.3 Introduction

In the field of analytical chemistry, the 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) modelling, comes with a solution to address these challenges[62, 211]. 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[62, 2]. Through these techniques, valuable insights can be gained into molecular behaviour in chromatographic systems, advancing analytical chemistry across diverse fields [182, 212].Conventional methods, like single-target retention prediction, encounter difficulties when confronted with intricate scenarios, often requiring significant time and resources. Therefore, the imperative lies in the development of resilient and adaptable QSRR models, capable of swiftly and accurately predicting 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. Such models which are also known as multitarget QSRR models, has 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 (EPs) such as variations in mobile phase

compositions(pH, solvent, strength, buffer concentrations etc.) and molecular descriptors (MDs) with the chromatographic behaviour of molecules[213, 187]. A major hurdle in developing such models is the lack of comprehensive data. To establish dependable and resilient models, researchers necessitate extensive and diverse datasets that encompass a broad spectrum of compounds and experimental settings[211, 214]. However, obtaining such datasets can be challenging. Data collection is resource-intensive, and available datasets may not be diverse enough for multitarget predictions[215].

Researchers have been finding ways around this data scarcity. One method is data augmentation, where by artificially increasing the size and diversity of the dataset, the models can capture a broader range of retention time variations.[216]. While data augmentation increases the quantity of data, it doesn't necessarily improve the quality of the original data. If the original dataset 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[14], where knowledge from related areas is applied to fill in data gaps, making it possible to build more robust models even with limited data. These innovative approaches open new doors for QSRR modeling, making it more versatile and effective in predicting retention times under various conditions.

4.3.3.1 Transfer learning approach

Transfer learning (TL) in deep learning consists of transferring the knowledge learned from a source domain D_s to a target domain D_t [14, 217]. A domain can be defined as $D = \{\mathbf{X}, P(\mathbf{X})\}$ where $\boldsymbol{X}$ is the feature space and $P(\boldsymbol{X})$ represents the marginal distribution for $\boldsymbol{X} = [x^1, x^2, ..., x^n]$ where x^i represents a feature of $\boldsymbol{X}$. If we learn a task $T_s = \{Y, f(.)\}$ where Y denotes a label space and $f(.)$ denotes a decision function. TL aims to improve the learning of a decision function in D_t for a different but related task T_t by using $f(.)$. Transfer learning in machine learning focuses on applying knowledge from a source domain to improve performance in a target domain, categorized mainly into homogeneous and heterogeneous strategies. Homogeneous transfer learning addresses differences in the marginal and conditional distributions within the same domain to better adapt models to new tasks[14]. Methods include correcting disparities in either marginal or conditional distributions, or both, to normalize these differences and enhance model accuracy ([218], [219]). Heterogeneous transfer learning, conversely, aims at aligning the input spaces of the source and target domains under the assumption of similar domain distributions [220]. When these distributions are not equivalent, further adaptation techniques are employed to adjust the models appropriately. Within these broader categories, transfer learning can also be segmented by the type of information transferred: instances and features. Instance-based transfer learning involves reweighting source domain instances to align with the target domain's marginal distribution[221]. This approach is particularly effective when the conditional distributions between the domains are consistent([218], [219]). Feature-based transfer learning, on the other hand, includes two main strategies (Figure 4.3.1. The first is asymmetric feature transformation, where features from the source are reweighted or transformed to closely match those of the target domain,

facilitating smoother model adaptation [222]. The second strategy involves identifying a common latent feature space that reduces marginal distribution differences and finds predictive structures beneficial across domains [223]. These strategies collectively enhance the applicability of models across varied domains by leveraging existing knowledge and reducing the necessity for extensive domain-specific data collection and model training.

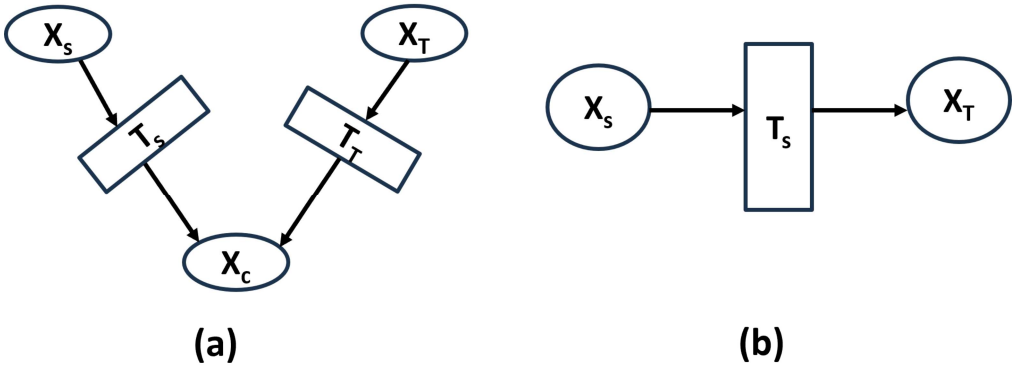


Figure 4.3.1: (a)The symmetric transformation mapping (TS and TT) of the source (XS) and target (XT) domains into a common latent feature space. (b) The asymmetric transformation (TT) of the source domain (XS) to the target domain (XT) [14]

4.3.3.2 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[224]. 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[225], LeakyReLU[226], Tanh[227], or any other suitable activation function[228, 229]. 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[213]. 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.

4.3.4 Materials and methods

4.3.4.1 Data sets

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 METLIN(SMRT) data set downloaded from its Figshare repository[13]. The retention time of nearly all molecules falls within two distinct intervals: 0-2 minutes and 8-25 minutes. Molecules with low retention times (Retention time $\leqslant$ 2 minutes) were excluded from the SMRT dataset, resulting in a total of approximately 77 thousand molecules. The other data sets (RIKEN[139, 32], LPAC[131] was used for testing purposes. The LPAC dataset, containing only 96 compounds, poses a data scarcity issue, making it challenging for any advanced QSRR modelling. Therefore, obtaining their retention time requires exploring approaches such as transfer learning.
The SMRT and Riken datasets had only one experimentally observed retention time consequently, we employed it in our study exclusively for single-target prediction modelling. Conversely, the LPAC datasets 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 modelling as well.

4.3.4.2 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 using RDKit package, version 2015[92].

4.3.5 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 multi-layer perceptron (MLP) to address this complex challenge. The general workflow of the model architecture is shown in Figure 4.3.2. 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 (Equation 4.3.1) 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.

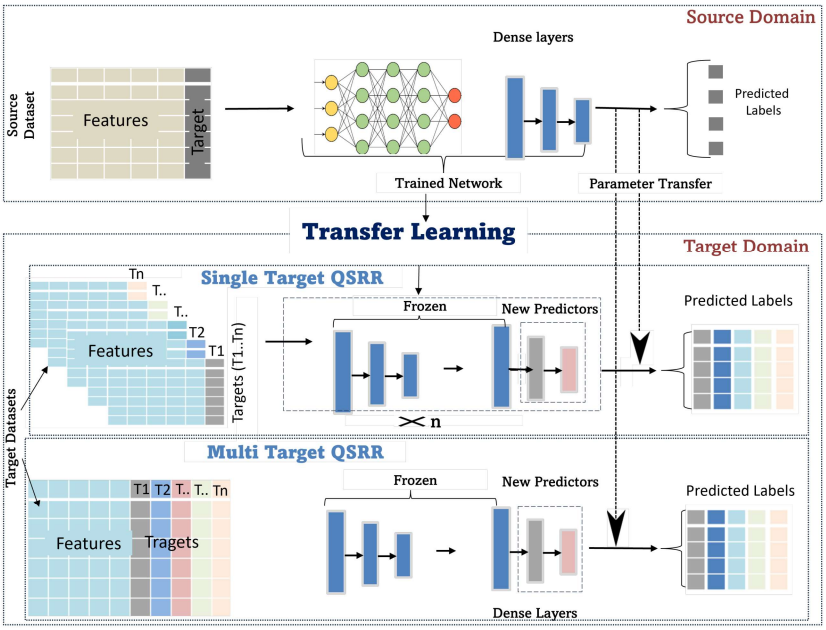


Figure 4.3.2: The architecture of QSRR modelling based on Transfer Learning approach

To further enhance the model's ability to generalize, a dropout layer precedes the output layer, reducing the risk of overfitting.

$$LeakyReLU(x) = max(0, x) + negative\_slope * min(0, x) \tag{4.3.1}$$

4.3.5.1 Training and Fine-Tuning

The MLP model was initially pre-trained on the extensive SMRT dataset. 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 datasets. 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 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 modelling, 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 4.3.3, the Multi-Layer Perceptron

Table 4.3.1: Summary of Model Abbreviations

Category	**Abbreviation**	**Description**
Single Target models	M1_WTL	No TL
	M2_TL	With TL
Multi Target models	M3_WTL	No TL
	M4_TL	With TL, physicochemical descriptors
models tested on SMRT data	M5-WTL	No TL
	M6_TL	With TL

(MLP) was initially trained using the SMRT dataset, which is notably large. This size advantage allows for its division into training, validation, and testing subsets, allocated 80%, 10%, and 10% of the total data(randomly), respectively. Such distribution ratios are standard practice for datasets 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[230, 231]. To assess the performance of the most effective model configuration, it underwent a re-training process. This process involved combining the training and validation sets for a comprehensive

training phase, followed by an evaluation on 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 datasets. 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 dataset 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 dataset, to validate its effectiveness and generalization capability. In this study, five models were constructed, with their respective abbreviations detailed in Table 4.3.1. Through this meticulous approach

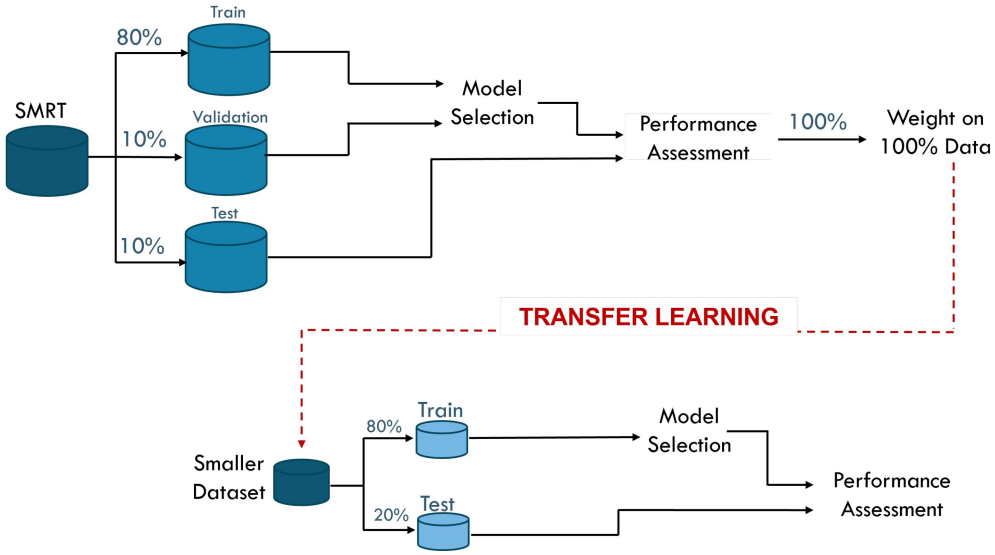


Figure 4.3.3: A simple schematic overview of model training using physicochemical descriptors

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

4.3.5.2 Evaluation metrics

Three commonly used metrices were used for assessing the performance of predictive models:

1. The Mean Squared Error (MSE) is calculated as the sum of the squared differences between the predicted ($\hat{y}_i$) and actual (y_i) values, divided by the total number of samples (N):

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

2. The Mean Absolute Percentage Error (MAPE) or Mean Relative Error (MRE) is a measure of the average magnitude of errors in a set of predictions, relative to the actual values.:

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

3. The Coefficient of Determination (R^2) evaluates the proportion of the variance in the dependent variable (y) that is predictable from the independent variable ($\hat{y}$). It is calculated as 1 minus the ratio of the sum of squared errors of the predicted values to the total sum of squares:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4.3.2}$$

In these equations, y_i represents the ground truth value, $\hat{y}_i$ represents the predicted value, $\bar{y}$ denotes the mean of the ground truth values, and N is the total number of samples.

4.3.5.3 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've used SHAP values for the best performing models. SHAP values give each feature a score, indicating its importance for a particular prediction. Thus, if,

- $f(x)$: This represents the prediction made by the model for an input x and,
- $E[f(x)]$: The expected value (mean) of the predictions across all possible inputs. Often, it is approximated by the average of the model's predictions over a sample or the training set, mean($model.predict(X)$).

Then, based on the SHAP theory, the relationship between SHAP values and the model output can be expressed as:

$$f(x) = E[f(x)] + \sum(\text{SHAP values for each feature})$$

The SHAP values essentially explain the deviation of $f(x)$ from its expected value $E[f(x)]$. Hence, to get the SHAP values the equations can be written as such:

$$f(x) - E[f(x)] = \sum(\text{SHAP values for each feature})$$

Here, the sum of SHAP values for all features explains the difference between the actual prediction $f(x)$ and the expected prediction $E[f(x)]$. 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[232] for every features to plot the summary.

4.3.6 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 datasets that are insufficient or include multiple targets to be predicted.

4.3.6.1 Model Performances

This study involved physicochemical descriptors as input to construct the DNN models to investigate their respective impacts on retention time predictions through four different strategies(Table 4.3.1, 4.3.2). It's 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 [233, 234, 235], and Graph Neural Networks (GNNs), including Graph Convolutional Networks (GCNs) and Relational Graph Convolutional Networks (RGCNs),have been frequently used in recent past[32, 236]. 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 comparision 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 datasets 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.

In our investigation, we assessed the flexibility of DNN model architectures based on physicochemical features, tailoring them to single and multi-target prediction tasks with comparative analyses by employing transfer learning approaches in situations of scarce data availability. Our analysis delineated distinct performance trajectories for each modeling approach Table 4.3.2. For the LPAC dataset, 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 R^2 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 R^2, from -0.35 to 0.64 and from -0.38 to 0.66 for Single Target and MultiTarget models, respectively.This suggests that the application of Transfer Learning significantly enhanced the model's predictive accuracy and its ability to explain the variance in the dataset. A decrease

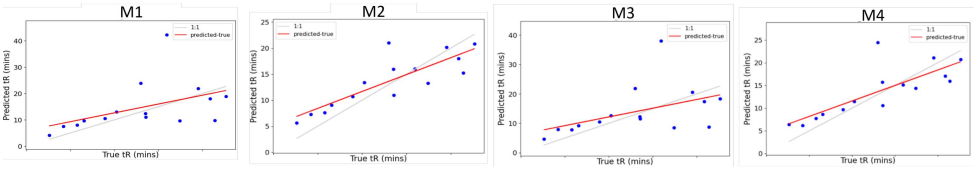
of MSE from 16.19 to 15.15 min from M4(Multitarget with Transfer Learning) to M2(Singletarget with Transfer Learning) emphasizes the suitability of Multitarget approach of QSRR, which inherently handles more complex prediction tasks by predicting multiple outputs simultaneously. Overall, the Multi-target models benefit significantly from the transfer learning approach, resulting in lower MSE and higher R^2 values, showcasing the superiority of transfer learning in these cases. On comparision of M1 with M3 and M2 with M4, it can be clearly seen that multitarget model perform better than Single target settings. Predicted versus observed retention times for each cases are plotted(Figure 4.3.4). It can be clearly observed that transfer learning approach has benefitted the predictions for targets especially at higher pH. In Figure 4.3.4, 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 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 targets it can be seen that Target 5(retention time at pH 8.0) has less scattered points and closeness to the identity line.

4.3.6.2 Time comparison

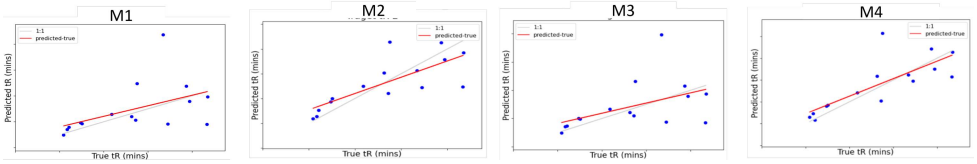
As evidenced by Table 4.3.2, the computational time was marginally affected between M1 to M2 and M3 to M4, demonstrating that TL's advantages in model accuracy do not substantially impact modelling efficiency. When comparing Single Target to MultiTarget models, MultiTarget models provide better time savings, highlighted by quicker execution times over Single target modelling which is 0.05 minutes 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 dataset characteristics and computational constraints.

	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

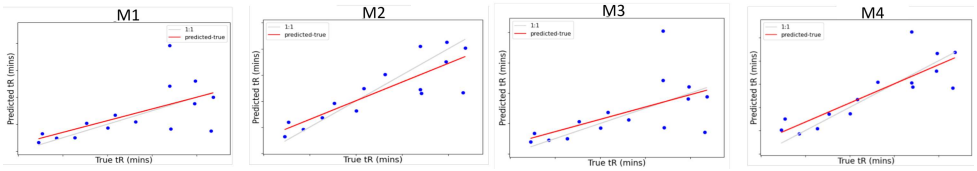
Table 4.3.2: **Model performances.(Model abbreviations are elaborated in Table 4.3.1)**



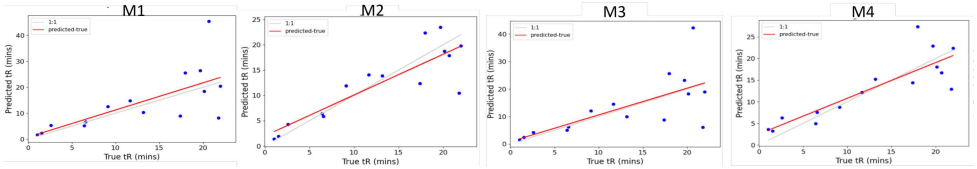
(a) Target 1(Retention times at pH 2.7)



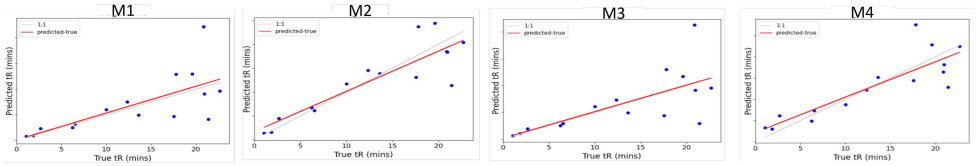
(b) Target 2(Retention time at pH 3.5)



(c) Target 3(Retention time at pH 5.0



(d) Target 4(Retention time at pH 6.5



(e) Target 5(Retention time at pH 8.0)

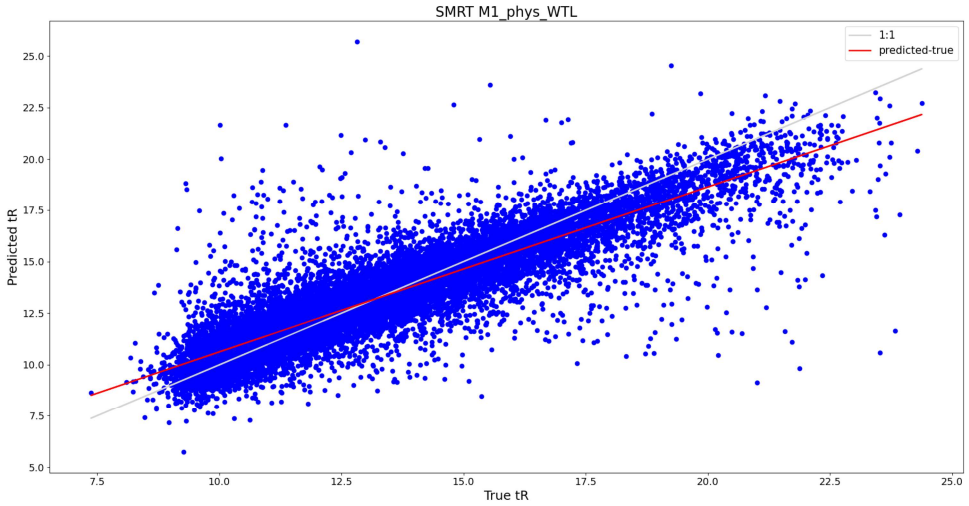
Figure 4.3.4: Plots for Predicted vs. Observed retention time(min) for M1 to M4 for LPAC dataset, X-axis-Observed and Y-axis- Predicted retention time(min)

4.3.6.3 Performance comparison on test data with benchmark studies

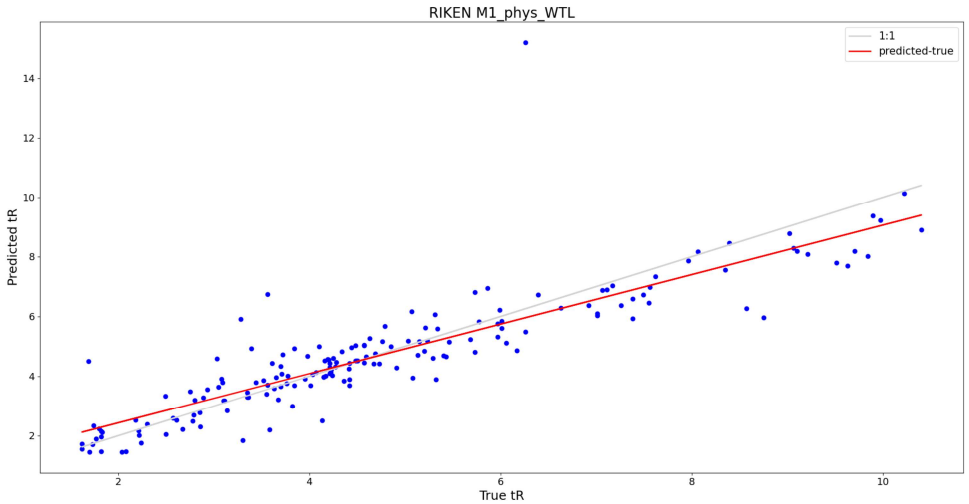
In our study, we compared the performance of our models against established benchmarks cited in Kensert et al.[32]. The results are highlighted in Table 4.3.3. It is important to note that M5 and M6 for SMRT dataset are same, because the data used for building the base model overlaps with the data employed for evaluating the model (transfer learning approach when other datas are used). The values in the table and Figure 4.3.5, show that the M5_M6 model achieves good results across both examined datasets. Specifically, for the SMRT dataset, M5_M6 achieves a Mean Relative Error (MRE) of 0.07 and an R^2 score of 0.78 which is similar to other models in the list like RF, AB, and comparable to other models including MLP in benchmark study and GCN, RGCN, SVM. For the RIKEN dataset, it records a comparable MRE of 0.14 and an R^2 of 0.75. These figures not only put M5 on a competitive footing with, but in some cases, ahead of, advanced models like Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB). Similarly, the M6_TL model showcases notable performance on Riken datasets with MRE of 0.14 with a good R^2 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 dataset. When comparing M5 and M6 (Figure 4.3.5 (b), (c)) for RIKEN dataset—where M6 utilizes Transfer Learning while M5 does not, analysis of the coefficient of determination (R^2) 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 to benchmark models. Their low MRE values indicate their ability to make accurate predictions, while the high R^2 scores demonstrate their efficacy in explaining the variance in the data. M6, in particular, stands out with its remarkable R^2 score of 0.85 on the RIKEN dataset, 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.

4.3.6.4 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 better prediction of retention times and more efficient separations. SHAP values are crucial in this analysis. Summary SHAP plots Figure 4.3.6 and corresponding selected features and their rankings(Table 4.3.4) illustrate the interpretation of the transfer learnt multitarget QSRR models for every target(Model4). It presents the importance of the top 20 molecular



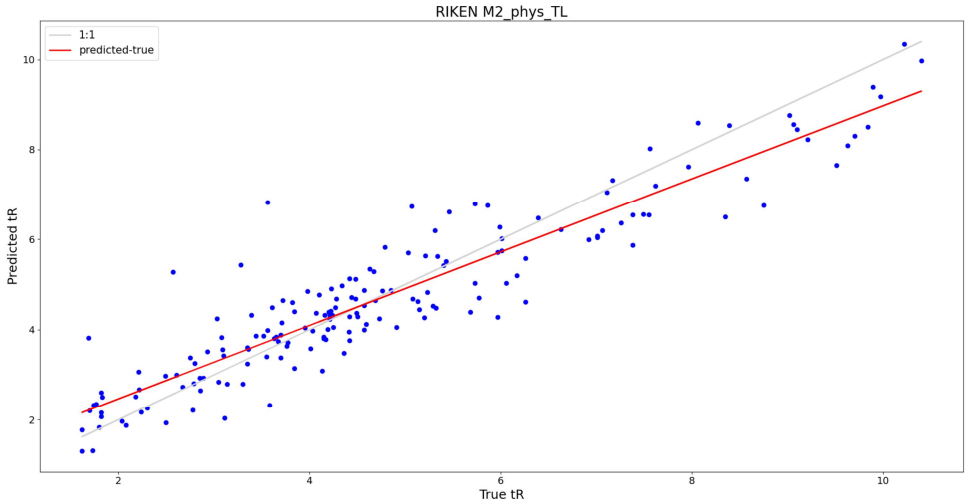
(a) Plot for M5/M6 models on SMRT dataset



(b) Plot for M5(Model WTL) on Riken dataset

Table 4.3.3: **Comparison of Model Performances with Benchmarks**

Models	SMRT		RIKEN	
	MRE	R^2	**MRE**	R^2
GCN	0.04	0.89	0.14	0.76
RGCN	0.04	0.89	0.14	0.79
MLP	0.05	0.84	0.10	0.56
RF	0.07	0.78	0.19	0.69
SVM	0.06	0.82	0.18	0.76
AB	0.07	0.76	0.19	0.68
GB	0.15	0.40	0.19	0.70
M5-WTL	0.07	0.78	0.14	0.77
M6-phys-TL	0.07	0.78	0.14	0.85



(c) Plot for M6(Model TL) on Riken dataset

Figure 4.3.5: Plot for Predicted vs. Observed retention time(min);X-axis-Observed and Y-axis - Predicted retention time

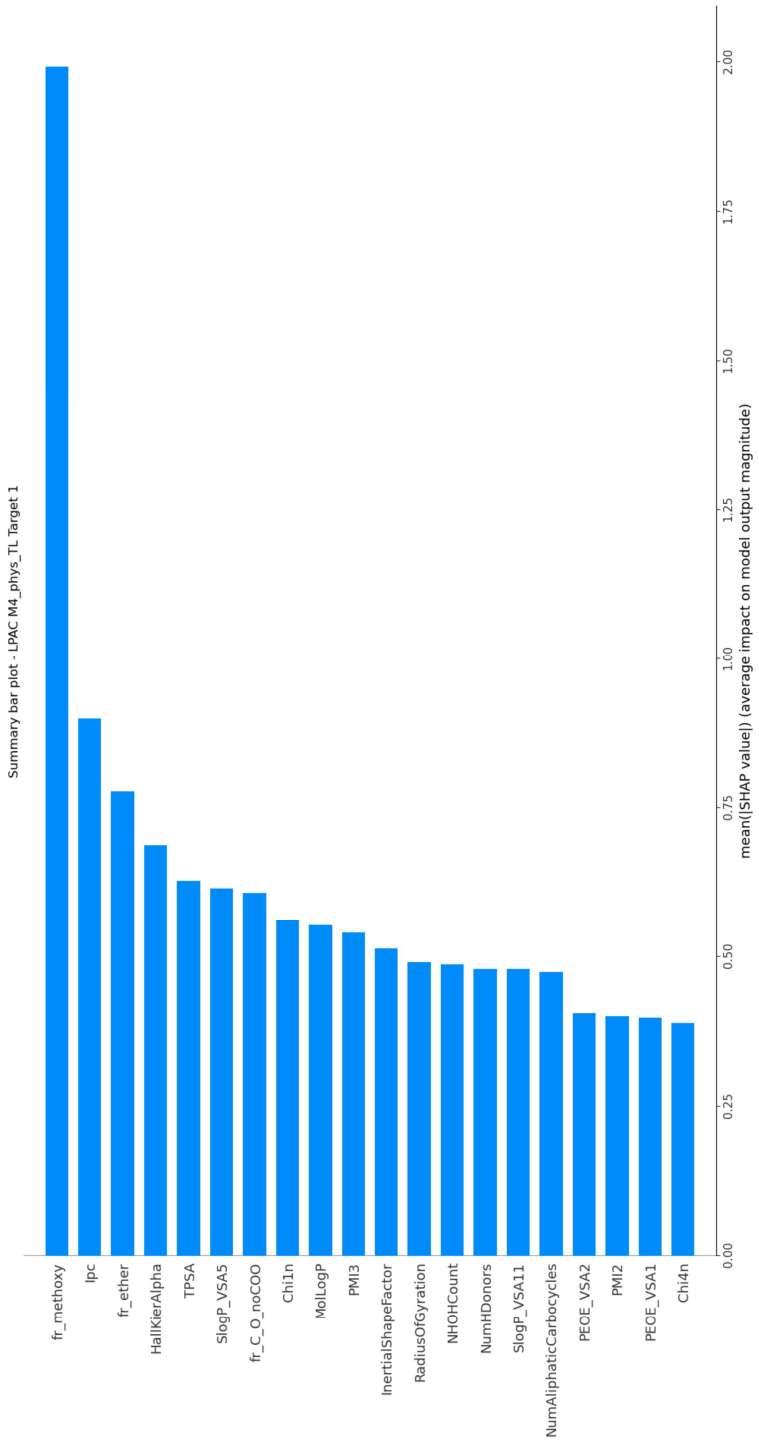
descriptors in terms of average impact on model output magnitude (the effects on predicted retention time). Lower SHAP values indicate lower effect of the descriptor while higher SHAP values indicate high effects of the descriptor. From the plots, it can be observed that specific features, like fr_methoxy, IpC, fr_ether, HallkierAlpha, TPSA remain consistently important across all pH levels. These features play a fundamental role in retention mechanism in liquid chromatography, regardless of the pH. However, the study also identifies features like TPSA

and MolLogP whose importance 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_VSA5, NumHDonors, and RadiusOfGyration in the LPAC dataset, indicating that the solute's physicochemical properties, such as lipophilicity, hydrogen bonding capability, and molecular size, differently affect retention times at varying pH levels.

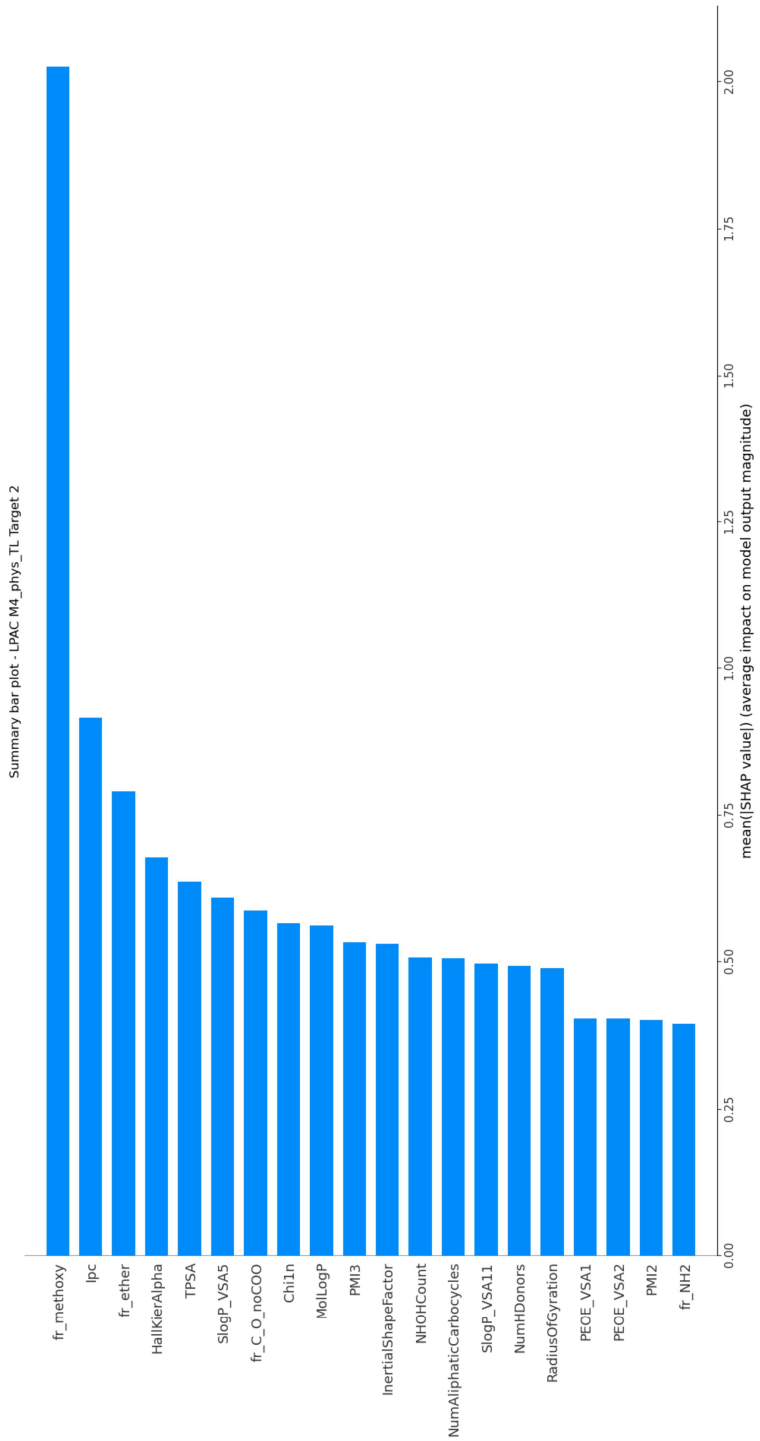
Important point to note here is, that these findings, derived from SHAP summary plots, offer general insights into the factors influencing RPLC retention times differentially with varying targets. A more detailed chemical analysis and domain-specific expertise would be required for precise interpretations,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

Table 4.3.4: **Summary of SHAP values for M4(Top 10 features); ISF- InertialShapeFactor**

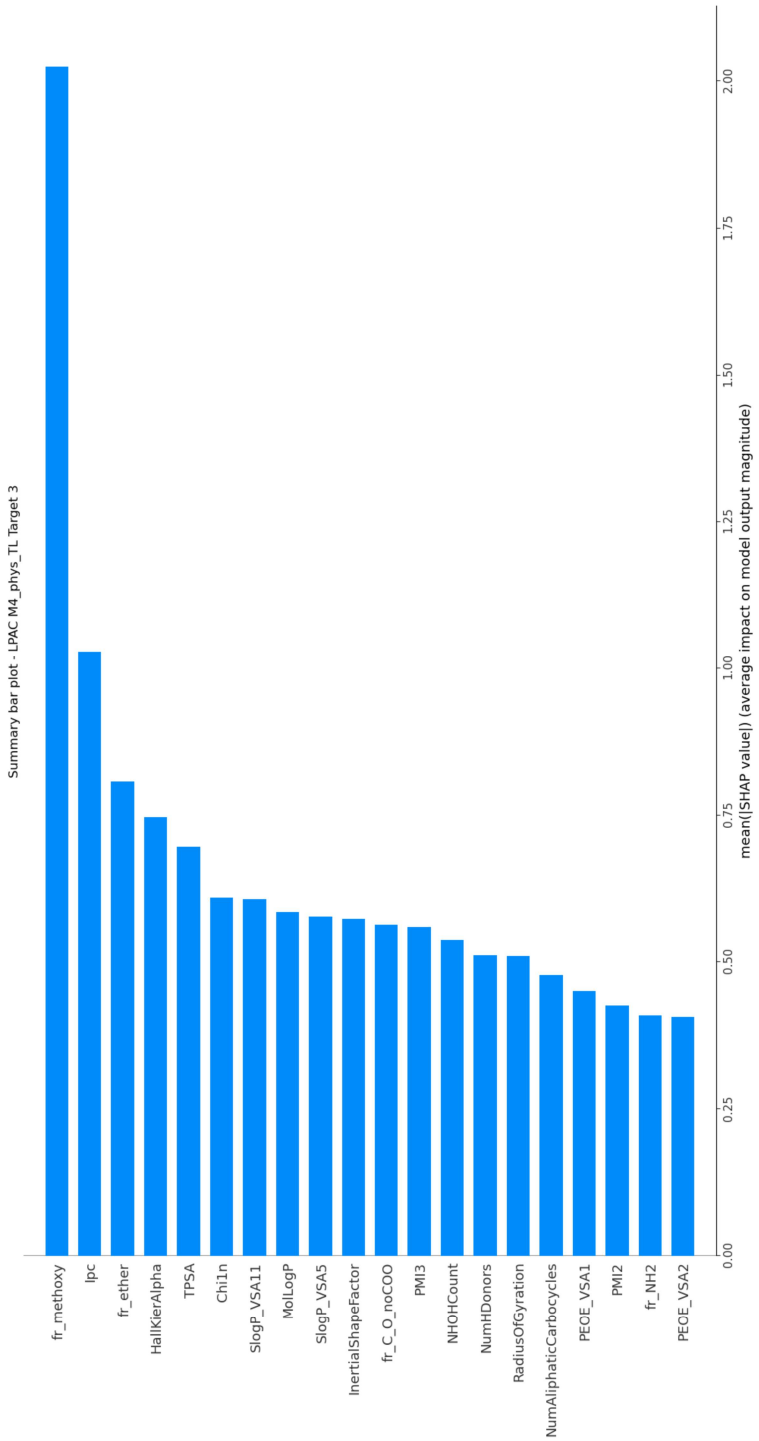
Rank	T1	T2	T3	T4	T5
1	fr_Methoxy	f_Methoxy	fr_Methoxy	fr_methoxy	fr_Methoxy
2	Ipc	Ipc	Ipc	ipc	ipc
3	fr_ether	fr_ether	fr_ether	fr_ether	fr_ether
4	HallKierAlpha	HallKierAlpha	HallKierAlpha	TPSA	HallKierAlpha
5	TPSA	TPSA	TPSA	HallKierAlpha	TPSA
6	SlogP_VSA5	SlogP_VSA5	Chi1n	SlogP_VSA11	SlogP_VSA11
7	fr_C_O_noCOO	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	NumAliphatic Carbocycles	PMI3
13	NHOHCount	NumAliphatic Carbocycles	NHOHCount	NumHDonors	NumHDonors
14	NumDonors	SlogP_VSA11	NumHDonors	PMI3	NumAliphatic Carbocycles
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



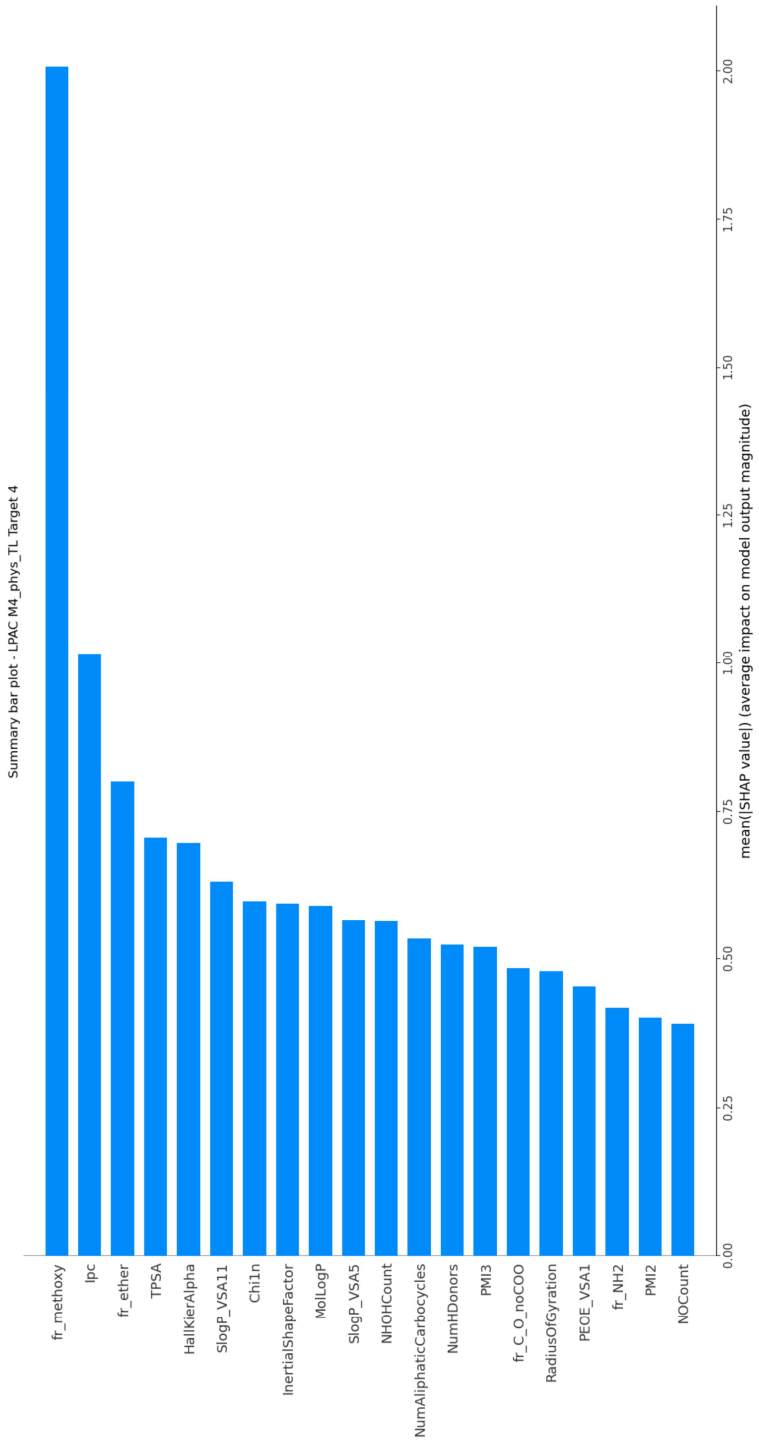
(a) Target1- Retention time at pH 2.7



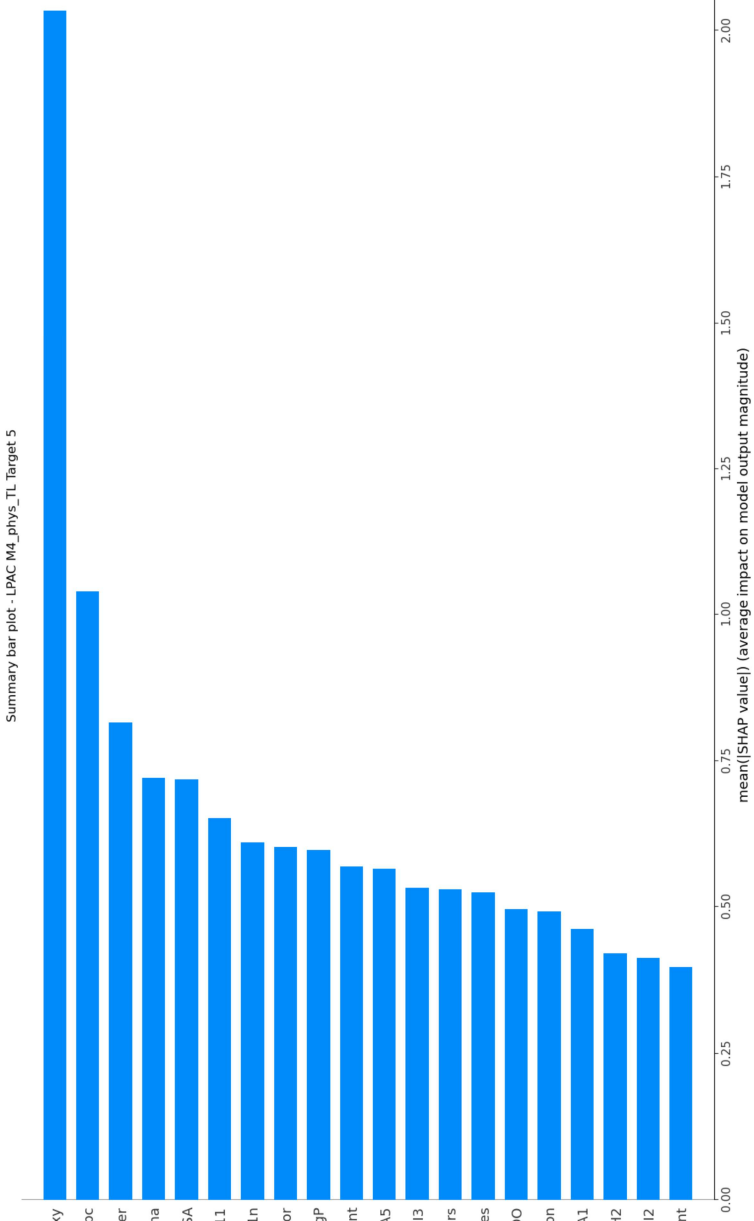
(b) Target2- Retention time at pH 3.5



(c) Target3- Retention time at pH 5.0



(d) Target4- Retention time at pH 6.5



(e) Target5- Retention time at pH 8.0

Figure 4.3.6: SHAP summary plots for LPAC dastaset. Y-axis- Molecular descriptors, X-axis- effects of molecular descriptors on the targets(SHAP values)

.

4.3.7 Conclusion

In conclusion, this study provides valuable insights into the field of retention time prediction modelling 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 methodologies in Single target and multi target settings, to enhance the accuracy and generalizability of QSRR models. Our analysis was conducted on two distinct datasets utilizing four different models. One of the key findings of this study is the significance of transfer learning in the context of QSRR modelling. It was observed that the application of transfer learning consistently improved the performance of QSRR models, resulting in lower Mean Squared Error (MSE) and higher coefficient of determination (R^2) values. For analytical chemists working on multi-target 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 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, outperformed 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 response, 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 dataset, 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.

4.3.8 Transfer Learning Multi-Target QSRR Modeling:Analysis based on MIA descriptors

4.3.8.1 Background

Multivariate Image Analysis descriptors/Image-based descriptors for chemical compounds are a relatively newer approach that leverages the power of visual representation to capture structural information of molecules for retention time predictions.

The images demonstrate a strong association with retention times and serve as a method for encoding chemical properties[136]. The differences in pixel positions reflect changes in the structure within a related group, thereby accounting for the variance in retention times observed within the series[137].

4.3.8.2 Image data processing

In this work, image based descriptors were used to compare the two approaches(Transfer learning and without transfer learning in combination to single target and multi-target retention prediction approaches). SMILE structure was used for every compound to generate the 2D images(Example in Figure 4.3.7) using rdkit.Chem[92]. The 2D-generated images were colored and used as such for CNN models, without any modification. Retention time prediction using image-based descriptors

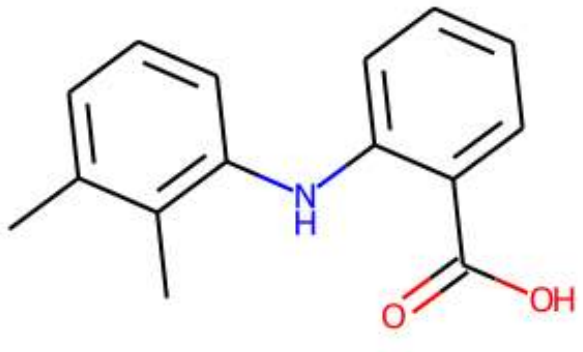


Figure 4.3.7: Example(3aminobenzoic acid) of an image used as input in CNN model

involves several steps, from the generation of molecular images to the training of deep learning models. Below is a more detailed explanation of how these descriptors are used for retention time predictions:

Data Preparation

Start with a dataset of chemical compounds for which experimental retention times are known. For each compound, a molecular representation in the form of a 2D molecular graph was generated. Making images carefully is very important in this

study. So, RDKit was specifically chosen to ensure that all images were reproducible and consistent, while also acting to minimize the level of diversity that could be introduced through the manual drawing of molecules. RDKit enables the representation of molecules as objects derived from SMILES strings, with specific algorithms[237, 238, 239] applied to establish their 2D layouts for visual representation. This involves the Chem module for processing chemical information and the Draw module for the graphical depiction of molecules, where atoms, bonds, and optional annotations are illustrated on a digital canvas.

Image Representation

Convert the molecular representations into image-like formats. This involved creating 2D graphs where atoms and bonds are represented as pixels, converting circular fingerprints into binary images.

4.3.8.3 Model Training:

In case of image data processing for QSRR modelling(Figure 4.3.8), the CNN was pre-trained on the SMRT dataset(Architecture is shown in Figure 4.3.9). Due to the relatively big size of the SMRT dataset, it is possible to split it into train, validation and test sets (70%-15%-15% respectively) . The first two are used for model selection. In this case, it consists of choosing the number of hidden layers, the number of units for each hidden layer and the dropout percentage of the dropout layer. The performance assessment for the best-performing model on the validation set was then obtained by re-training from scratch on the train-validation set, by merging the train and test set and then computing the evaluating metrics on the independent test set. Then, the model was fine-tuned on the other smaller datasets. Model selection, where the frozen/unfrozen layer selection and dropout percentage are determined, was performed using leave-one-out cross-validation (LOOCV) on 85% of the dataset. Finally, the performance of the best model obtained during LOOCV was assessed on the remaining and independent test set containing 15% of the original dataset. The model selection when fine-tuning the model on the smaller datasets was performed using 5-fold cross-validation. Those changes in splitting ratios were caused by the increase of computational requirements when training CNNs with 2D images. Hyperparameters and unfrozen layers for the best performing models, for each dataset, found during cross-validation for the MIA descriptors are shown in Table 4.3.5. Four models were developed for the comparision of different approaches of QSRR settings. Details of models can be found in Table 4.3.6.

4.3.8.4 Results and Discussion

The results presented in the tables provide a detailed overview of the performance of various Quantitative Structure-Retention Relationship (QSRR) models that incorporate Multivariate Image Analysis (MIA) descriptors, with a focus on both single and multi-target models, and the impact of applying Transfer Learning (TL). The models are evaluated based on Mean Squared Error (MSE), R-squared (R^2)

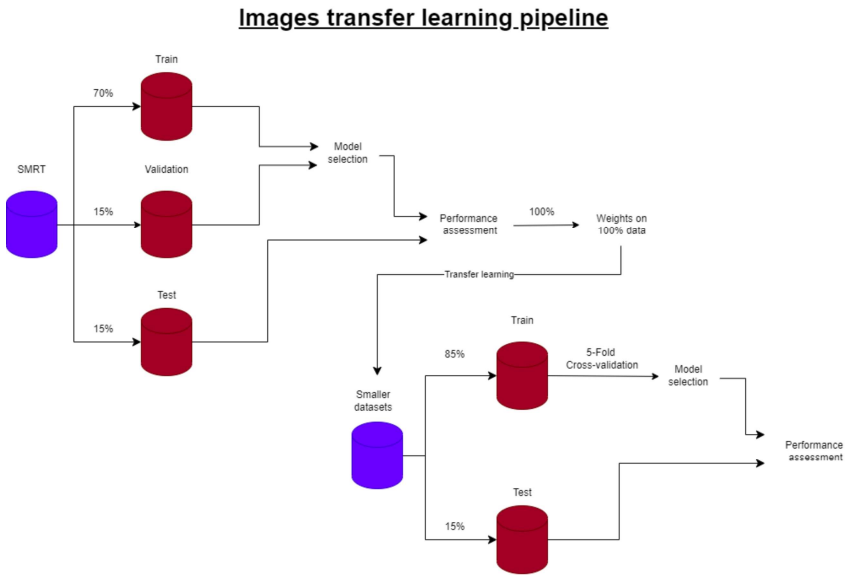


Figure 4.3.8: A simple schematic overview of model training using MIA descriptors

values, and computation time (Time(Min)) across two different datasets, labeled "Small" which is LPAC dataset and "ACN" dataset.

1. MSE and R^2 Values: The MSE and R^2 values offer insights into the models' accuracy and predictability. Lower MSE values and higher R^2 values are desirable, indicating more accurate predictions and a model that accounts for a greater proportion of the variance observed in the data, respectively. For *single target models*, M8_MIA_TL shows an improvement in the R^2 value in the Small dataset, indicating a positive effect of TL on model predictability. However, this model also exhibits a higher MSE, suggesting a discrepancy in prediction accuracy. *In the multi-target* QSRR modelling, both models M9_MIA_WTL and M10_MIA_TL reflect less favorable outcomes, with negative R^2 values in the Small dataset and modest improvements in the ACN dataset. These results imply challenges in the models' ability to accurately predict across multiple targets.

2. Time Analysis: The computational time (Time(Min)) required for model predictions is crucial, especially when processing large datasets or in applications where speed is of the essence. *Single Target Models:* There's a noticeable decrease in prediction time when applying transfer learning (M8_MIA_TL), compared to no TL (M7_MIA_WTL), particularly in the ACN dataset. This suggests that TL not only impacts the predictive performance but also efficiency. *Multi-Target Models:* Similarly, M10_MIA_TL demonstrates a reduction in prediction time compared to M9_MIA_WTL, indicating that TL might offer computational efficiency gains in multi-target settings as well.

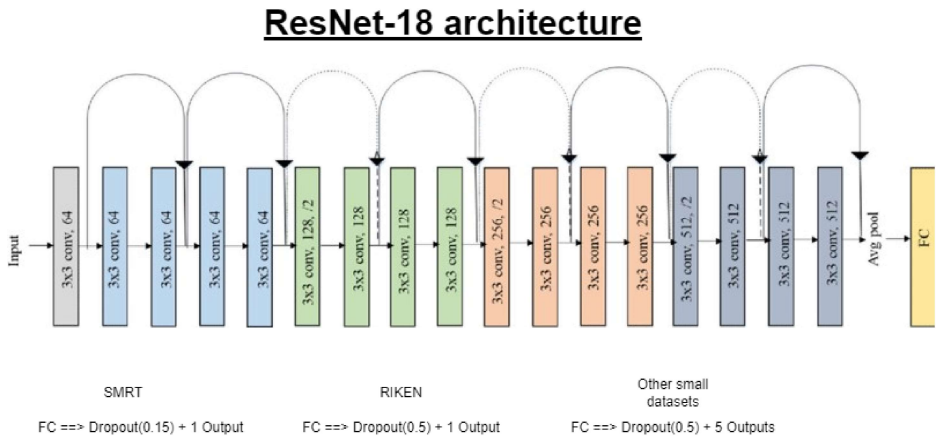


Figure 4.3.9: A Schematic architecture of CNN model used in this study

Dataset	Learning rate	weight decay	unfrozen layers
5-output models			
LPAC	1e-3	0.05	3, 4, FC and batch normalization
ACN	1e-3	0.05	3, 4, FC and batch normalization
Single-output models			
LPAC	1e-3	0.01	4, FC and batch normalization
ACN	1e-3	0.01	4, FC and batch normalization
RIKEN	1e-3	0.01	FC and batch normalization

Table 4.3.5: Hyperparameters and unfrozen layers for each MIA descriptors dataset

4.3.8.5 Conclusion

The results, while not entirely favorable in terms of MSE and R^2 values, provide important insights into the application of transfer learning and the use of MIA descriptors in QSRR modeling. The varied performances across single and multi-target models underscore the complexities of modeling retention times and the potential of transfer learning to improve model fit in certain conditions. The model performances suggest that transfer learning, when applied to models utilizing MIA descriptors, has the potential to enhance the explanatory power of these models, as evidenced by the improvement in R^2 values in certain cases. However, the effectiveness of transfer learning seems to vary depending on the target and the specific nature of the dataset used for modelling.
Despite the innovative approach of incorporating MIA descriptors in QSRR models, the results indicate that achieving high accuracy and model fit remains challenging. The presence of high MSE and negative R^2 values in several models suggests that the relationship between MIA descriptors and retention times may

Table 4.3.6: Summary of Model Abbreviations with MIA descriptors

Category	**Abbreviation**	**Description**
Single Target models	M7_WTL	No TL
	M8_TL	With TL, MIA descriptors
Multi Target models	M9_WTL	No TL
	M10_TL	With TL

	Small			ACN		
Models	MSE	R^2	Time(Min)	MSE	R^2	Time(Min)
Single target Models						
M7_MIA_WTL	43.40	-0.06	7.59	36.82	-0.56	6.43
M8_MIA_TL	73.39	-0.74	5.96	18.35	0.24	5.19
Multitarget Models						
M9_MIA_WTL	61.46	-0.71	0.28	27.63	-0.17	0.13
M10_MIA_TL	52.21	-0.37	0.09	16.21	0.33	0.15

Table 4.3.7: **Model performances for Multivariate Image descriptors (MIA)**

be complex and not fully captured by the current modeling approaches. Hence, all these findings highlight the need for further exploration into optimizing the use of MIA descriptors and transfer learning within QSRR models. Future research could focus on refining the descriptors, exploring alternative modeling techniques, and expanding the dataset like data augmentation which very much possible in case of image data, to improve model performance. Thus, while the results do not showcase high prediction accuracy, they do provide valuable information into the potential and limitations of applying transfer learning based on MIA descriptors in QSRR modeling. These inferences pave the way for future research aimed at enhancing the predictive capabilities of QSRR models in both single and multi-target settings.

5

General Discussion

5.1 General discussion

In the growing field of Quantitative Structure-Retention Relationship modelling, researchers have many methods to choose from. These methods help predict the retention times of compounds in chromatographic processes. This multifaceted area of study, rich in technique and application, aims to bridge the gap between molecular structure and chromatographic behaviour, offering invaluable insights for analytical chemistry. The selection of an optimal QSRR modelling approach, however, poses a significant challenge due to the complexity of molecular interactions and the intricate nature of chromatographic systems. In this direction, this thesis aims to explore three different QSRR modelling methods, which are explained in Chapter 4. These include the Single Target QSRR approach (Chapter 4.1), the Multitarget QSRR approach (Chapter 4.2), and the Transfer Learning-based QSRR approach (Chapter 4.3). The study highlights two critical elements of QSRR modelling: the selection of modelling algorithms and the importance of molecular descriptors. It examines the characteristics and challenges related to their use in this thesis, providing insights that can help in providing the reference point to choose the appropriate approach and their application for the specific problem at hand.

5.1.1 The choice of modelling algorithms

The thesis provides a critical assessment of various algorithms employed across the single and multitarget QSRR scenarios, including classical machine learning algorithms like MLR, SVR, Lasso, GBR, RF, Stacking and advanced artificial intelligence algorithms like DNN. The analysis revealed that algorithm choice significantly affects prediction accuracy. For example, Stacking, an ensemble technique, stands out by combining predictions from multiple models to make a final prediction in single target QSRR models(Chapter 4.1). This method effectively leverages the strengths of various algorithms, improving accuracy, especially when dealing with scarce structural data. By integrating different perspectives on the data, stacking provides a more refined prediction than any single model could. However, despite its potential to enhance retention time prediction accuracy, it can introduce significant setbacks. The added complexity of combining various base models with a meta-model complicates interpretation, making it challenging to extract clear scientific insights, a critical aspect in chromatography studies. Additionally, the increased computational demands for training and deploying stacked models may not align with the constraints of laboratories requiring rapid analysis or those with limited computational resources and expertise at times. On the other hand use of another ensemble algorithm like RF, is a good choice for predicting multiple targets altogether i.e, evident from the results of Chapter 4.2. The reason could be it's algorithm adaptability from single-target prediction (STP) to multiple-targets prediction (MTP) settings where the key difference lies in how the cost of node splits is calculated and how predictions are made.

Table 5.1: Comparison of DNN vs. Classical ML Algorithms

Aspect	**DNN**	**Classical ML**
Computational Resources	Capable of handling large, complex datasets but require significant computational power for training and inference.	More efficient on less powerful machines, suitable for limited resources
Data Requirements	Can learn from large, high-dimensional datasets effectively but require large amounts of data to generalize.	Perform well with smaller datasets but may struggle with high-dimensional or complex data structures.
Model Complexity	Able to model highly complex relationships in data with a risk of overfitting; complexity makes tuning challenging.	Simpler models have lower risk of overfitting and are easier to tune but are limited in capturing complex, non-linear relationships.
Interpretability	Often seen as "black boxes" due to complex structures.	Comparatively higher interpretability, decisions are easier to understand.
Development Time	Development, tuning, and validation can be time-consuming.	Generally faster development cycles with simpler models.

- STP calculates the node cost using the sum of squares of the differences between observed and predicted values. The prediction for a test sample is the weighted average of responses across all trees.
- MTP, uses the sum of squared Mahalanobis distances for its node cost, taking into account the covariance among multiple target variables. Predictions for a test sample involve averaging over the multivariate responses[240].

In the context of this thesis, the utilization of DNN, as discussed in Chapter 4.3, underscores the advancing role of AI in uncovering hidden patterns within data, thereby providing efficient retention prediction methods. This thesis strategically employs classical ML in Chapters 4.1 and 4.2 for modelling on the LPAC dataset, while applying the more complex DNN-CNN models in Chapter 4.3. In this chapter, a model initially trained on a larger dataset is adapted to smaller datasets like LPAC for retention time prediction. The decision to select particular algorithms rests on understanding their strengths and limitations in handling data. The use of DNNs in QSRR modelling offers significant advantages in terms of automatic feature extraction, handling of high-dimensional and complex data, and the ability to model non-linear relationships and target interactions. These characteristics can lead to more accurate, robust, and generalizable models, ultimately enhancing our understanding and prediction of compound retention in chromatographic systems. However, these models come at the cost of needing significant computational power and a vast amount of data. These methods also carry the risk of overfitting and may result in models that are hard to interpret. On the other hand, classical Machine Learning (ML) approaches are valued for their efficiency, especially when computational resources or data are scarce. They are well-suited for smaller datasets and are characterized by their straightforward

and comparatively better interpretable models. Despite these advantages, classical ML techniques may not perform well when faced with the complexity or the sheer volume of data that DNNs can manage. The table (5.1) encapsulates key distinctions, guiding the choice between DNN and classical machine learning. Referring to this comparison can be beneficial for addressing project-specific needs in QSRR modelling. This approach highlights a strategic decision-making process in model selection, balancing dataset size, computational demands, and interpretability requirements.

5.1.2 The choice of Molecular Descriptors in QSRR Approaches

The use of molecular descriptors significantly shapes the effectiveness and efficiency of QSRR modeling techniques in analytical chemistry. These descriptors serve as a fundamental component in the construction and optimization of models for predicting retention times, with each approach leveraging them in unique ways to cater to specific research needs. This thesis utilizes physicochemical descriptors(Chapters 4.1, 4.2, and 4.3) and explores the potential of image-based descriptors(chapter 4.3).

Unlike Multitarget QSRR models that utilize a common set of descriptors for every target, Single-Target QSRR capitalizes on the precision of molecular descriptors to develop highly tailored models for individual targets. By focusing on target-specific descriptors, the single-target QSRR approach ensures a high degree of model accuracy and specificity. This meticulous selection process, however, requires extensive descriptor analysis and multiple model building for every new target, which can be resource and time-intensive.

On the other hand, by selecting descriptors that capture the commonalities across different chemical entities, Multitarget QSRR models can efficiently predict retention times for multiple targets simultaneously through one model only. These strategies require a meticulous selection of descriptors to prevent overgeneralization, ensuring the models' sensitivity to the fine distinctions between targets.

Transfer Learning introduces a novel perspective on the use of molecular descriptors by adapting models developed for one set of targets to new, yet chemically related, targets. This process involves identifying descriptors that are universally applicable across different datasets, enabling the transfer of learned patterns through adjusted weights from pre-trained models. The success of this approach heavily depends on the relevance and adaptability of the low-level and high-level features to both the original and new datasets, highlighting the need for a strategic descriptor selection process that maximizes cross-dataset applicability.

The interpretability of QSRR models, especially those employing advanced AI techniques like deep learning, poses a significant challenge. While these models can achieve high predictive accuracy, they are often considered "black boxes," offering little insight into how molecular descriptors influence retention times. The use of SHAP (Shapley Additive explanations) can aid in interpreting the contributions of different descriptors within such models, providing a deeper understanding of

their roles. However, the application of SHAP in QSRR modeling demands considerable computational resources for bigger data and specific knowledge to analyze and interpret the results effectively. Therefore, it should be used in situations where understanding the intricacies of descriptor contributions is essential, despite the potential challenges in computational demand and interpretation.

5.1.3 Model performances

This thesis primarily employs physicochemical descriptors to construct Quantitative Structure-Retention Relationship (QSRR) models. However, the models sometimes fail to accurately predict retention times, a discrepancy that could arise from multiple sources. A significant factor is the reliance on molecular descriptors generated by online tools and software, such as Rdkit and ChemAxon's Chemicalize. These platforms apply their predictive algorithms to compute descriptor values. Therefore, any small error in these models could propagate and have a significant impact on the QSRR models developed in this study. Additionally, it's quite possible that physicochemical descriptors are not sufficient to capture the entire pattern of dependency of targets on molecular descriptors, necessitating the inclusion of other types of descriptors, such as molecular fingerprints or graph properties or a set of mixed descriptors.
Furthermore, this thesis also investigates the potential of image-based descriptors in QSRR modelling, offering a new dimension of information on molecular retention beyond what traditional physicochemical descriptors can provide. Initially, the integration of image-based descriptors resulted in reduced prediction accuracy compared to models using only physicochemical descriptors, as noted in Chapter 4.3. This reduction in accuracy may be attributed to the absence of molecular conformation optimization for the specific environment employed in the study. This capability for data augmentation with image-based descriptors which was out of the scope of this study, can be particularly advantageous in scenarios where the modelling is constrained by the availability of scarce datasets. By generating synthetic images or modifying existing ones through techniques such as rotation, scaling, and flipping, researchers can artificially expand the dataset, potentially enhancing the model's accuracy and robustness. This unique advantage of image-based descriptors, allowing for the expansion and diversification of training datasets, positions them as a valuable tool in QSRR modelling, especially in exploratory studies or when conventional descriptors fail to capture certain molecular features. However, the utilization of image-based descriptors should be approached with caution. While they offer the potential for improved model performance through data augmentation, the prioritization of accuracy and interpretability remains paramount. Their use is most beneficial when balanced with an understanding of the trade-offs involved, making them a complementary rather than a replacement option for physicochemical descriptors in situations where the small size of the dataset poses a challenge to model development and validation.

5.1.4 Characteristics and Challenges of three approaches

In analytical chemistry, selecting between Single-Target QSRR, Multitarget QSRR, or Transfer Learning approaches depends on the study's specific goals, the characteristics of the dataset, and various constraints. This selection is crucial for the study's success due to the multiple factors that influence variability in retention times. A comparison of different methods for predicting retention times typically reveals their unique strengths and weaknesses. All the points about the three QSRR approaches discussed so far can be summarised as detailed in Table 5.2. This comparison guides us in choosing the most appropriate approach for specific use cases.

- The Single-Target QSRR approach, which creates individual models for each target (e.g. individual pH level), is advantageous for its simplicity and specificity. It allows for tailored optimization and fine-tuning and hence, potentially enhancing model accuracy and specificity for specific datasets for individual targets. However, this method can become time-consuming and less efficient as the number of targets increases, leading to an exponential increase in the number of models that may require significant resources to develop and maintain.
- Multitarget QSRR uses one model to predict retention times across different targets, making modelling more straightforward and efficient. It's especially useful for research involving a broad range of targets because it simplifies the process and requires fewer computational resources than creating a separate model for each target, as is done in Single-Target QSRR. However, implementing Multitarget QSRR can be complex and challenging When dealing with a very large number of targets. These can introduce noise into the model and can make it harder for the model to learn the underlying patterns and hence, difficulty of the model to differentiate between useful information and noise, especially when the outcomes are very much related and influence each other. Despite the challenges, Multitarget QSRR is valued for its efficiency and ability to maintain accurate retention predictions for a variety of conditions simultaneously.
- Transfer Learning offers a versatile solution that is applicable to both Single-Target and Multitarget QSRR approaches. Its strength lies in the ability to leverage existing models trained on one dataset to make predictions for similar molecules in a different dataset, which is especially useful in scenarios with insufficient data availability for model development. This approach can significantly reduce the time and resources needed for model development. However, the accuracy of transfer Learning may be affected if the training and target datasets are significantly different, posing a limitation to its applicability.

Overall, each QSRR modeling approach presents a set of advantages that make them suitable for different research needs. The selection of a modeling approach should therefore consider the specific requirements of the study, balancing the advantages against the potential limitations to achieve the most effective retention time prediction.

Table 5.2: Comprehensive Comparison of QSRR Modeling Strategies

Strategy	Characteristics	Challenges
Single-Target QSRR	• Tailors descriptors to specific target, allowing for high customization and potentially higher accuracy. • Simple to implement for individual parameter variation, facilitating detailed analysis. • Performance may be superior with diverse descriptors for different targets. • High improvement potential with refined target-specific descriptors.	• Time and resource-intensive with an increasing number of targets, requiring separate models. • May compromise performance with more variables and targets. • Requires target specific details of dataset, adding to the complexity of specific modeling adjustments.
Multitarget QSRR	• Efficient for simultaneous predictions across multiple targets with a single model, saving time and resources. • Maintains consistent descriptors, simplifying the modeling process. • Applicable to multiple targets with a simplified process and medium improvement potential.	• Implementation complexity and advanced data preprocessing required. • Could be less accurate for specific target due to uniform descriptors. • Generalizability may affect specificity and overlook target-specific molecular behaviors.
Transfer Learning based QSRR	• Leverages existing models for new data predictions, effective with limited data availability. • Offers flexibility, applicable to both single-target and multitarget scenarios. • Useful for tasks with significant data similarity, enhancing QSRR modeling capabilities.	• Source and target data differences can affect accuracy. • Requires careful source model selection, fine-tuning, and potentially complex implementation. • Additional steps needed for model adaptation and validation.

5.1.5 Application of QSRR strategies

The methodologies discussed in this thesis have applications ranging from broad to specific. In the course of this thesis, they have been utilized to predict the retention of N-nitrosamines (NAs) under ten different conditions, aiming to identify optimal separation conditions regarding matrice endogenous compounds to enable the quality control of these impurities.
To achieve this goal, a single-target QSRR approach was combined with response surface models and Multi Criteria Decision Analysis(MCDA), supported by desirability indexes[207]. This strategy proved effective for many compounds and indicated potential for further improvement for others. Based on these findings, alternative approaches, including multi-target QSRR modeling and mechanistic models can be explored. Such exploration will allow us to compare the effectiveness of mechanistic and empirical models, determine their relative merits, and provide a basis for assessing the strategies' effectiveness and associated risk management.

6

General Conclusion

Overall Conclusion

In the dynamic field of analytical chemistry, scientists constantly face the challenge of accurately analyzing an ever-growing variety of chemical compounds. Among the various techniques available, Reversed-Phase Liquid Chromatography (RPLC) is one of the techniques known for its versatility and effectiveness in separating and quantifying these compounds. However, optimizing RPLC processes to achieve precise, reliable results under varying conditions presents significant hurdles. These challenges include dealing with the complexity of chemical mixtures, and the need for time-efficient and cost-effective methods that maintain high accuracy. Through the strategic application of QSRR, analytical chemists can overcome these pressing challenges, paving the way for advancements in analytical techniques and methodologies.

This thesis successfully addresses the critical challenges in the field of analytical chemistry, particularly in the optimization of RPLC separation methods. However, the analytical landscape is characterized by the complexity of sample matrices, changes in experimental conditions, and the need for the separation and identification of diverse compounds. Consequently, one-size-fits-all QSRR models may not suffice. Hence, by thoroughly investigating various feature selection methods and machine learning algorithms, and their combinations in multiple ways, this research investigates the most effective strategies for retention time predictions that can be applied to changes in experimental conditions.

Throughout the study, a comprehensive examination of molecular descriptor selection methods ranging from filter, wrapper to embedded methods and regression algorithms including many linear and non linear models have been tested to address the first research question i.e,*How can the best feature selection methods and machine learning algorithms be identified for precise retention time predictions.* The research findings indicate that for selecting crucial features, the embedded method proves to be more effective. Additionally, when predicting retention times for a single target via single-target QSRR modelling, stacking emerges as the superior prediction method However, for scenarios involving multiple targets, models like Random Forest stand out due to their algorithm adaptability and ability to accommodate target relationships.

Building on the groundwork laid by the initial objective, the thesis progresses to examine a QSRR model in multitarget settings. This model is designed to predict retention times at various pH levels, showcasing how retention time can vary with different experimental conditions, using pH as a key example. The findings of the second objective address the research question: *"Can a QSRR model be developed to accurately predict retention times across all pH levels, while also understanding the complex relationships between different targets?"*. This part of the research demonstrates the viability of creating a multi-target QSRR model that effectively captures the complex relationships between descriptors and targets. This approach marks a significant advancement over traditional single-target models. By utilizing insights from the dynamic interactions across different pH levels, the model enhances both the efficiency and accuracy of retention time predictions, offering a more comprehensive and effective method for understanding these relationships.

In tackling the issue of scarce data, the third chapter addresses the research question: *"What strategies can be utilized to overcome the challenge of limited data in QSRR modeling, thereby enhancing the accuracy and reliability of retention time predictions in chromatographic analyses?"* This section introduces an innovative strategy aimed at improving the precision and dependability of QSRR modeling. By employing transfer learning and other sophisticated AI techniques, the study successfully navigates the hurdles associated with limited data availability. This approach demonstrates how these methods can significantly refine retention time predictions in chromatographic analyses. The evaluation of both single-target and multi-target prediction methods, along with the application of transfer learning, represents a notable progress in the field. It opens up new avenues for optimizing QSRR models when dealing with limited datasets, offering better insights into effectively enhancing model performance under such conditions.

Collectively, the three chapters of this thesis offer an in-depth examination of various methodologies for retention prediction in RPLC, each providing distinctive insights and addressing specific challenges in this domain. Through a comparative analysis, this work elucidates the suitability of these methods for retention time prediction, aiming to enhance the understanding of their strengths and limitations. This facilitates the development of more precise and reliable retention prediction models for small molecules in RPLC, marking a significant advancement in analytical chemistry. The contributions of this thesis extend beyond methodological innovations, such as the integration of single-target to multitarget models and the application of transfer learning. It establishes a comprehensive framework for QSRR studies, covering everything from feature selection to algorithm choice, thus laying a groundwork for novices in the field. This research not only improves the precision and applicability of QSRR models but also emphasizes the necessity for adaptable and efficient analytical methods to meet the evolving demands of chemical analysis. The advancements presented promise to refine RPLC retention prediction for small molecules in pharmaceutical research. By exploring a range of strategies, this thesis addresses contemporary analytical challenges and establishing a foundation for future innovations. It enriches the existing body of knowledge, thereby facilitating the development of more precise and efficient analytical techniques in RPLC, and the continued exploration and advancement in the field.

7
Perspectives

Perspective

Based on the discussions on the challenges of developing fully accurate retention prediction models, a number of perspectives on future works emerges from this research. These perspectives are summarized below:

Firstly, broadening the types of descriptors used in QSRR modeling represents a crucial step forward. Investigating a wider array of molecular descriptors, such as molecular fingerprints and graph-based representations, is essential for overcoming current limitations and improving model performance. By incorporating a more diverse set of descriptors, we can aim not only to enhance model accuracy but also to deepen our understanding of molecular behavior. This effort emphasizes the importance of innovative methodologies, including the development of hybrid models that combine multiple descriptor types. These models use the best features of each descriptor and hence, can give a detailed view of how compounds interact, with the goal of greatly improving how accurately they can predict the targets.

Secondly, the advancement of our mechanistic interpretation of both single and multi-target QSRR models, including those utilizing deep learning. Applying the fifth OECD principle for QSRR modeling, enhancing our mechanistic insights will not only improve the reliability of predictions but also contribute significantly to the field of cheminformatics by providing clearer links between molecular structure and chromatographic retention mechanisms.

A third future work would be to increase data size for model training. To further enhance the performance and utility of QSRR models, expanding the scope of data collection is essential. Specific methods or new ways of acquiring more data under a variety of conditions can be explored. This effort involves incorporating diverse variables, such as pH variations and gradient times, into experimental designs, yielding a more complex and robust dataset. Such a comprehensive approach can facilitate the development of models that are precise and adaptable to the complex realities of compound separation, paving the way for interdisciplinary collaboration. In this direction ,'Generative AI' which is a boom in the field of AI at present, for instance, can offer a novel way to overcome data scarcity and imbalance through the generation of realistic, synthetic molecular data [241].

A fourth and possible future work would be implementening more advanced AI techniques to extend the usability and transferability of QSRR models. Utilizing techniques like domain adaptive transfer learning [242] which could be an easy extension of the study in Chapter 4.3, we can try to predict the retention times of molecules in different chromatographic modes for example hydrophobic to hydrophillic and vice-versa.

A fifth potential perspective would be to integrate QSRR with DoE and AQbD frameworks. Such combined approach offers an optimal design space, streamlining

chromatographic method development. These models improve predictive capabilities for new compound sets under unseen conditions by using computed retention times of pharmaceutical test analytes to calculate separation selectivity. The flexibility offered by this integrated approach can be particularly advantageous in routine work, where efficient separation of diverse compounds is required, illustrating the interconnectedness of these future directions in advancing QSRR modeling.

Lastly, the possible future extension of this research would be the creation of a comprehensive website. Developing a comprehensive website could serve as a centralized repository for QSRR studies, significantly aiding new researchers in grasping foundational concepts, gathering pertinent literature, and determining suitable research approaches. This platform would provide all necessary information on QSRR aspects—Structure-Molecular Descriptors, Retention Time, and Relationships—alongside retention time prediction tools, thus enhancing efficiency and accessibility in QSRR research

Appendix

A

Appendix

Supplementary Information

Title: Quantitative structure retention-relationship modelling: Towards an innovative general-purpose strategy

Authors: Priyanka Kumari[a,b*], Thomas Van Laethem[a,b], Philippe Hubert[a], Marianne Fillet[b], Pierre-Yves Sacré[a], Cédric Hubert[a*]

a. University of Liège (ULiege), CIRM, Laboratory of Pharmaceutical Analytical Chemistry, Liège, Belgium

b. University of Liège (ULiege), CIRM, Laboratory for the Analysis of Medicines, Liège, Belgium

---

S1: Illustration of molecular descriptor calculation with an example

There were two tools used for molecular descriptor calculation: [1] RdKit [2] Chemicalize

- Values of LogD was calculated using Chemicalize at each pH

Method of calculation of other descriptors:

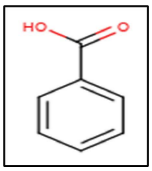
- Step1: Smile strings of compounds into Chemicalize
- Step2: Retrieve Smile strings of all microspecies and their distributions at all pH
- Step3: Calculate molecular descriptors of every microspecies from RdKit
- Step4: Calculate weighted average of molecular descriptors at each pH using formula below

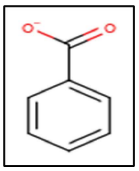
$$FV_{ph} = \frac{\sum_{i=1}^{n} MS_i * D_i}{\sum_{i=1}^{n} D_i}$$

Where , FV_{ph} = Weighted average, MS_i = Descriptor value for microspecies and,

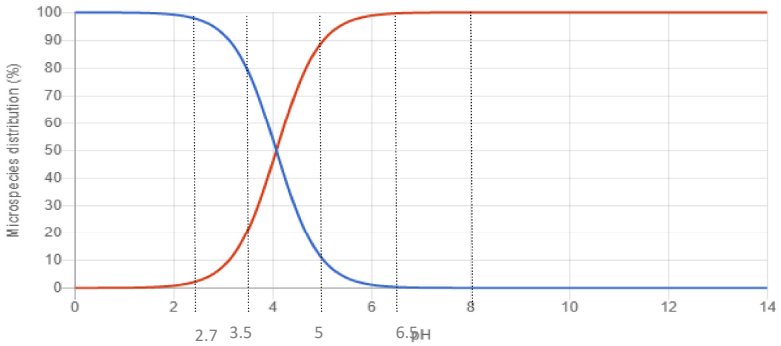
D_i = %Distribution of microspecies, n = no. of microspecies , ph = Specific pH at which final value is being calculated

Step-1





[b]Benzoic Acid (MS1) - C1=CC=C(C=C1)C(=O)O [b]Benzoate(MS2)- MS1C1=CC=C(C=C1)C(=O)[O-]



Step2: Microspecies distribution calculation using Chemicalize

Condition	MS1	MS2
pH2.50	0.97	0.03
pH3.50	0.79	0.21
pH5.00	0.11	0.89
pH6.50	0.00	1.00
pH8	0.00	1.00

Step3: Calculate molecular descriptors of every microspecies from RdKit

Microspecies	PEOE_VSA7
MS1	12.13
Ms2	5.56

Step4: Calculate weighted average of molecular descriptors at each pH :

Final value of PEOE_VSA7

[1] At pH2.7

= (12.13*0.97 + 5.56*0.03)/1

= 11.76 + 0.166

= 11.92

[2] At pH 3.5

= (12.13*0.79 + 5.56*0.21)/1

= 9.58 + 1.16

= 10.74

[3] At pH 5

= (12.13*0.11 + 5.56*0.89)/1

= 1.33 + 4.94

= 6.27

[4] At pH 6.5

= (12.13*0.0 + 5.56* 1.0)/1

=5.56

[5] At pH 8

= (12.13*0.0 + 5.56* 1.0)/1

= 5.56

Final Value of molecular descriptors: -

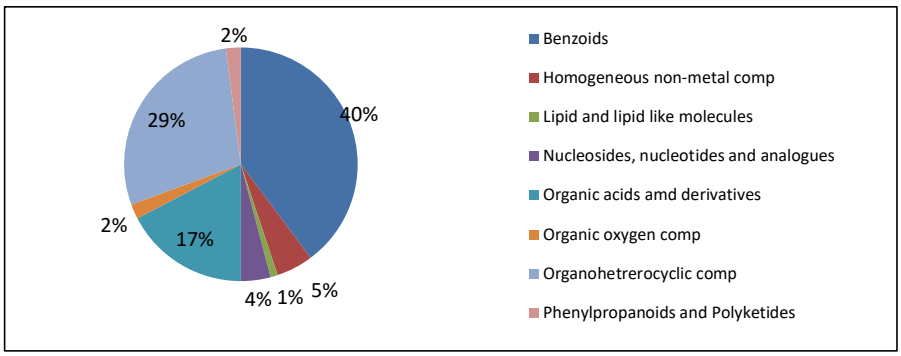
	Final Value
Condition	**PEOE_VSA7**

pH 2.7	11.92
pH 3.5	10.74
pH5	6.27
pH6.5	5.56
pH 8	5.56

S2: Table: Name of features used to start QSRR modeling

MolWt	EState_VSA2	fr_ArN	fr_quatN	MinEStateIndex	PEOE_VSA10	SlogP_VSA3	VSA_EState6
logD	EState_VSA3	fr_aryl_methyl	FractionCSP3	MinPartialCharge	PEOE_VSA11	SlogP_VSA4	VSA_EState7
Asymmetric.atom.count	EState_VSA4	fr_benzene	FSP3	Molar.refractivity	PEOE_VSA12	SlogP_VSA5	VSA_EState8
Atom.count	EState_VSA5	fr_bicyclic	HallKierAlpha	MolLogP	PEOE_VSA13	SlogP_VSA6	VSA_EState9
BalabanJ	EState_VSA6	fr_C_O	Heavy.atom.count	MolMR	PEOE_VSA14	SlogP_VSA7	
BertzCT	EState_VSA7	fr_C_O_noCOO	HeavyAtomMolWt	NHOHCount	PEOE_VSA2	SlogP_VSA8	
Chi0	EState_VSA8	fr_COO	Hetero.ring.count	NOCount	PEOE_VSA3	SMR_VSA1	
Chi0n	EState_VSA9	fr_COO2	Hydrogen.bond.acceptor.count	NumAliphaticHeterocycles	PEOE_VSA6	SMR_VSA10	
Chi0v	FpDensityMorgan1	fr_ether	Hydrogen.bond.donor.count	NumAliphaticRings	PEOE_VSA7	SMR_VSA3	
Chi1	FpDensityMorgan2	fr_halogen	Ipc	NumAromaticCarbocycles	PEOE_VSA8	SMR_VSA5	
Chi1n	FpDensityMorgan3	fr_imidazole	Kappa1	NumAromaticHeterocycles	PEOE_VSA9	SMR_VSA6	

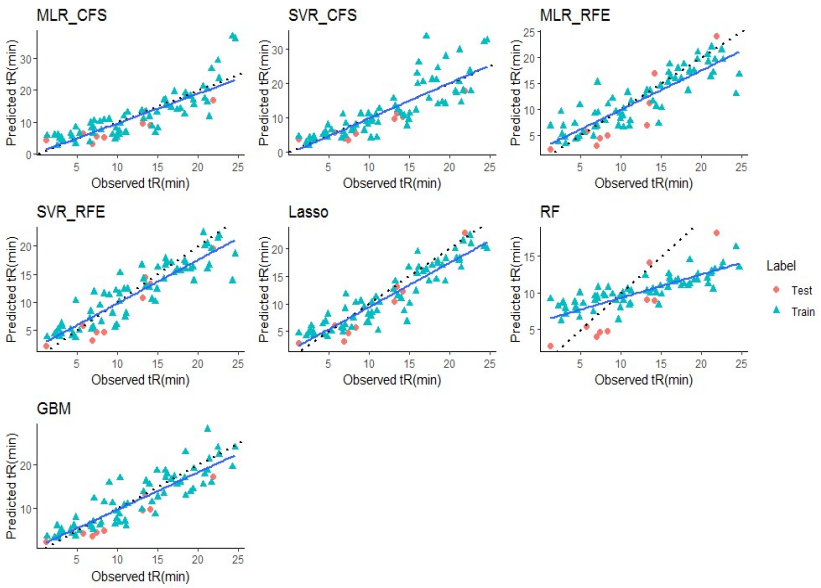
Chi1v	fr_Al_COO	fr_Ndealkylation1	Kappa2	NumAromaticRings	Polarizability	SMR_VSA7	
Chi2n	fr_Al_OH	fr_NH0	Kappa3	NumHAcceptors	qed	SMR_VSA9	
Chi2v	fr_Al_OH_noTert	fr_NH1	LabuteASA	NumHDonors	Ring.count	TPSA	
Chi3n	fr_amide	fr_NH2	MaxAbsEStateIndex	NumHeteroatoms	Rotatable.bond.count	VSA_EState1	
Chi3v	fr_aniline	fr_Nhpyrrole	MaxAbsPartialCharge	NumRotatableBonds	SlogP_VSA1	VSA_EState10	
Chi4n	fr_Ar_COO	fr_para_hydroxylation	MaxEStateIndex	NumSaturatedHeterocycles	SlogP_VSA10	VSA_EState2	
Chi4v	fr_Ar_N	fr_phenol	MaxPartialCharge	NumSaturatedRings	SlogP_VSA11	VSA_EState3	
EState_VSA1	fr_Ar_NH	fr_phenol_noOrthoHbond	MinAbsEStateIndex	NumValenceElectrons	SlogP_VSA12	VSA_EState4	
EState_VSA10	fr_Ar_OH	fr_pyridine	MinAbsPartialCharge	PEOE_VSA1	SlogP_VSA2	VSA_EState5	



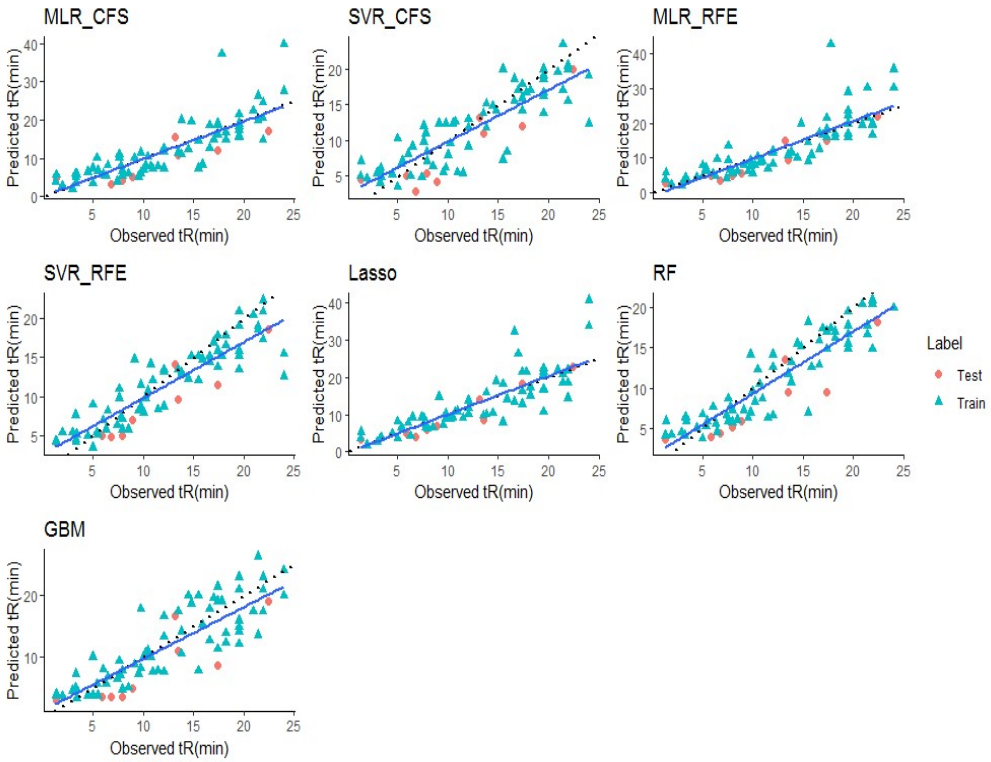
S3: Figure :Chemical taxonomy of the molecules in the dataset

S4: Table: Important features selected by each algorithm on data at each pH

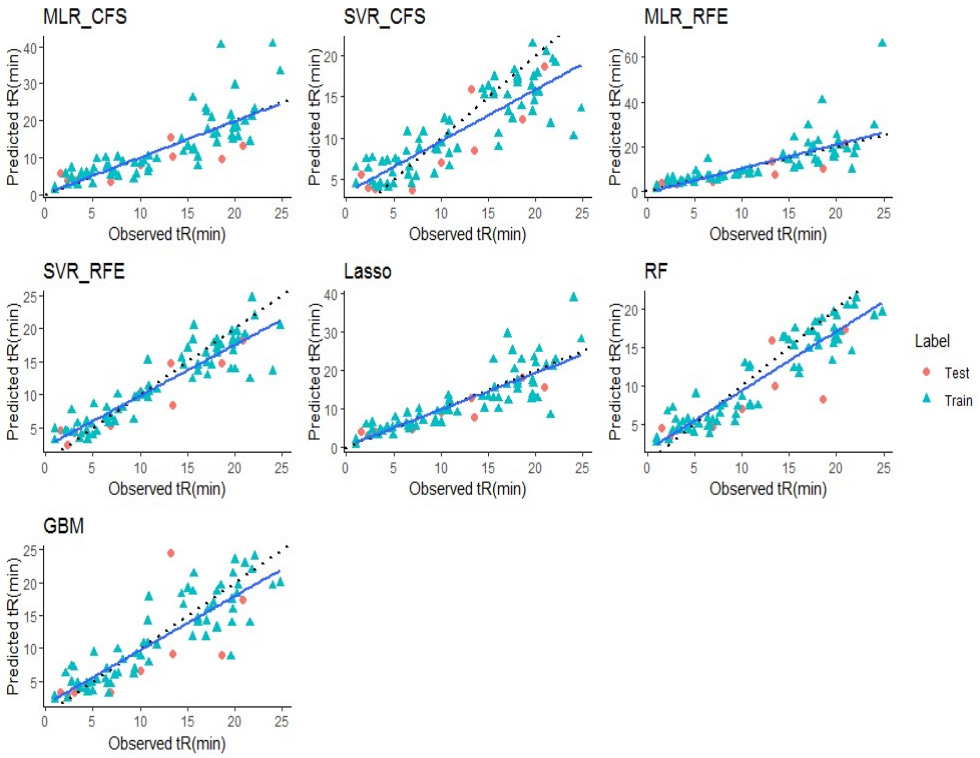
pH	Total	Descriptors
2.7	5	MolLogP, logD , NHOHCount , fr_Ar_NH, PEOE_VSA6
3.5	3	MolLogP, logD, NHOHCount
5	4	Polarizability, MolLogP, logD, PEOE_VSA6
6.5	3	MolLogP,logD,PEOE_VSA6
8	3	MolLogP , logD, PEOE_VSA6



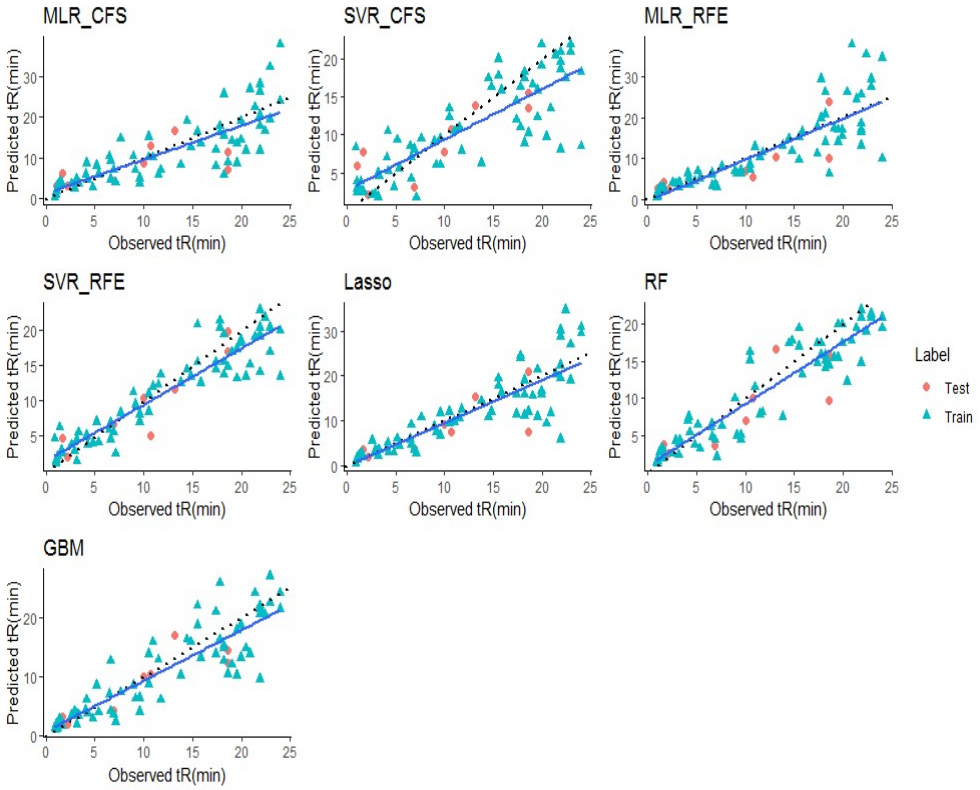
S5: Predicted Vs. Experimental Retention times (in min.) for all models at pH2.7 (Blue line- fit , Black dashed line- identity line)



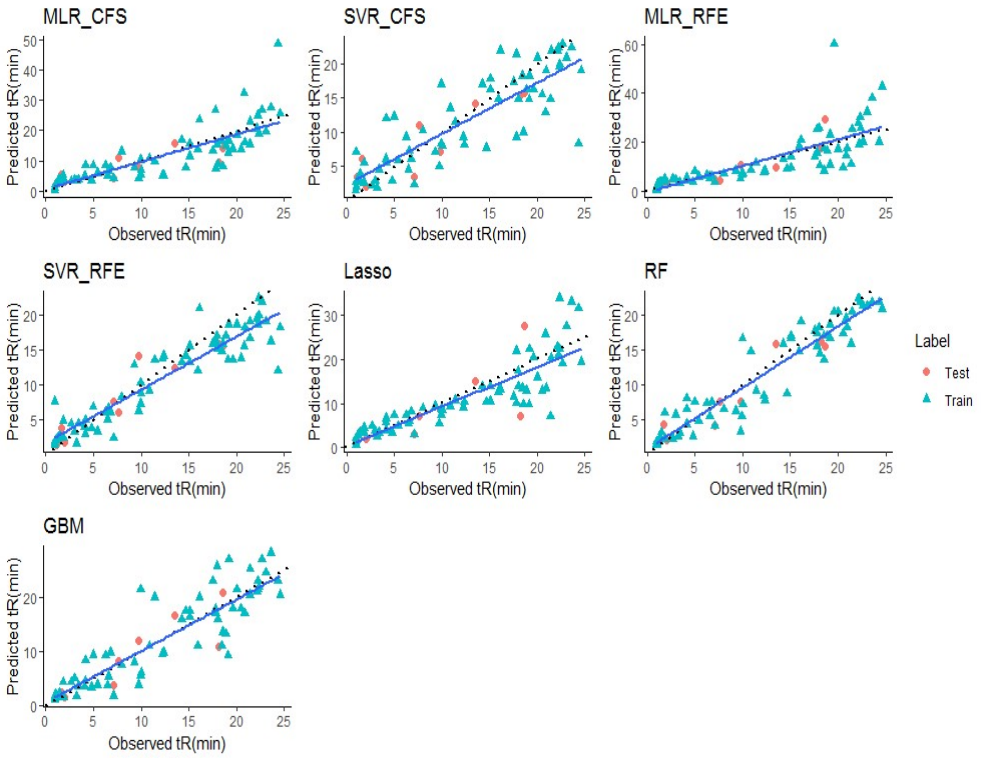
S6: Predicted Vs. Experimental Retention times (in min.) for qsrr models at pH 3.5((Blue line- fit , Black dashed line- identity line)



S7: Predicted Vs. Experimental Retention times (in min.) for qsrr models at pH 5 (Blue line- fit , Black dashed line- identity line)



S8: Predicted Vs. Experimental Retention times for qsrr models at pH 6.5 (Blue line- fit, Black dashed line- identity line)



S9: Predicted Vs. Experimental Retention times for qsrr models at pH 8 (Blue line- fit, Black dashed line- identity line)

S10: Parameters used by prediction models at all pH

Models at pH2.7:

SVR_CFS	Sigma = 0.362	C = 1
SVR_RFE	Sigma = 0.048	C=1
Lasso	Aplha = 1	Lambda = 0.014
RF	Mtry = 73	
GBM	n. trees = 150, Shrinkage = 0.1	Interaction depth =2, n. minobsinnode = 10

Models at pH3.5:

SVR_CFS	Sigma = 0.122	C = 1
SVR_RFE	Sigma = 0.075	C=1
Lasso	Aplha = 0.1	Lambda = 0.048
RF	Mtry = 73	
GBM	n. trees = 150, Shrinkage = 0.1	Interaction depth =2, n.minobsinnode = 10

Models at pH 5:

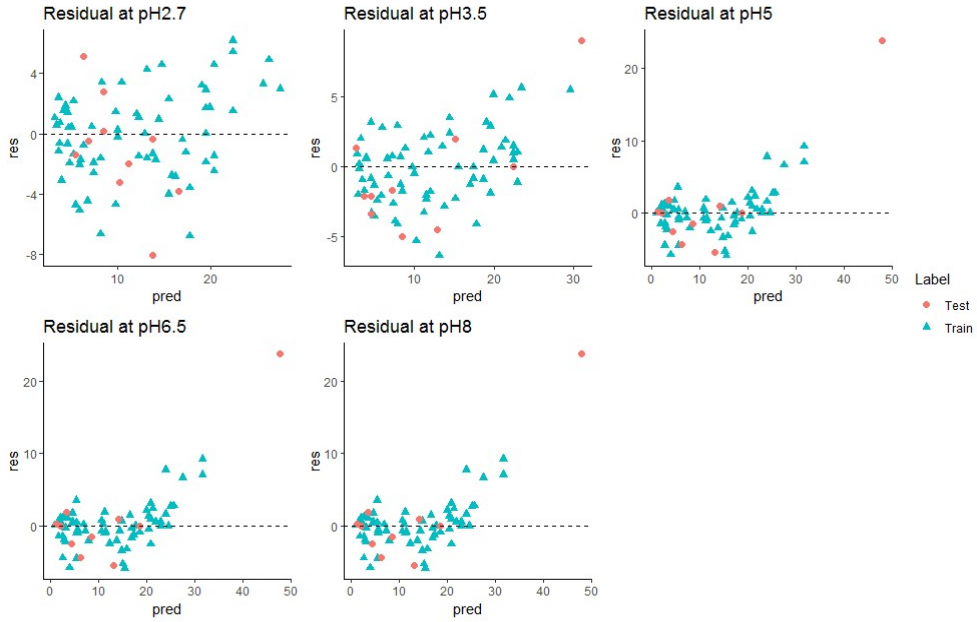
SVR_CFS	Sigma = 0.353	C = 1
SVR_RFE	Sigma = 0.054	C=1
Lasso	Aplha = 0.1	Lambda = 0.058
RF	Mtry = 73	
GBM	n.trees = 150, Shrinkage = 0.1	Interaction depth =2, n.minobsinnode = 10

Models at pH 6.5:

SVR_CFS	Sigma = 0.313	C = 1
SVR_RFE	Sigma = 0.042	C=1
Lasso	Aplha = 0.1	Lambda = 0.02
RF	Mtry = 147	
GBM	n.trees = 150, Shrinkage = 0.1	Interaction depth =3, n.minobsinnode = 10

Models at pH 8:

SVR_CFS	Sigma = 0.319	C = 1
SVR_RFE	Sigma = 0.110	C=1
Lasso	Aplha = 0.1	Lambda = 0.024
RF	Mtry = 147	
GBM	n.trees = 150, Shrinkage = 0.1	Interaction depth =2, n.minobsinnode = 10



S11: Residual plots (in min) for Stacking model for all dataset(with Miconazole)

B

Scientific contributions

Publications in international peer-reviewed journal

1. **Kumari, P.**, Van Laethem, T., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C. (2023). Quantitative Structure Retention-Relationship Modeling: Towards an Innovative General-Purpose Strategy. Molecules, 28(4), 1696.
2. **Kumari, P.**, Van Laethem, T., Duroux D., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C. (2023). A multi-target QSRR approach to model retention times of small molecules in RPLC (Journal of Pharmaceutical and Biomedical Analysis).https://doi.org/10.1016/j.jpba.2023.115690
3. **Kumari, P.**, Pradhan, B., Koromina, M., Patrinos, G. P., & Steen, K. V. (2022). Discovery of new drug indications for COVID-19: A drug repurposing approach. PloS one, 17(5), e0267095.
4. Van Laethem, T., **Kumari, P.**, Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C. (2022). A pharmaceutical-related molecules dataset for reversed-phase chromatography retention time prediction built on combining pH and gradient time conditions. Data in Brief, 42, 108017.
5. Van Laethem, T., **Kumari, P.**, Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C. (2022). A pharmaceutical-related molecules dataset for reversed-phase chromatography retention time prediction built on combining pH and gradient time conditions. Data in Brief, 42, 108017.
6. Raman, R., Antony, M., Nivelle, R., Lavergne, A., Zappia, J., Guerrero-Limón, G., Caetano da Silva, C., **Kumari, P.**, Sojan, J.M., Degueldre, C. and Bahri, M.A., 2024. The Osteoblast Transcriptome in Developing Zebrafish Reveals Key Roles for Extracellular Matrix Proteins Col10a1a and Fbln1 in Skeletal Development and Homeostasis. Biomolecules, 14(2), p.139.
7. **Kumari, P.**, Madureira Sanches Ribeiro G., Choudhary P., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C.(2024). Transfer Learning-Enhanced Multi-Target QSRR Modeling: A solution to low data regime in RPLC.(Under Review)

Oral communication

- International congress

1. Kumari, P., Van Laethem, T., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C. (2022, September). QSRR for small pharmaceutical compounds in RPLC: A machine learning approach. Oral session presented at conference "Chemometrics in Analytical Chemistry(CAC)", Rome, Italy.
2. Kumari, P., Van Laethem, T., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C. (2021, February). Quantitative structure retention relationship of small compounds in Liquid chromatography. Oral session presented at Conference(Virtual) "Chimiométrie", Liège, Belgium

- National Congress

1. Kumari, P., Van Laethem, T., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C (2023, June). A multi-target QSRR(mT-QSRR) approach to model retention time of small pharmaceutical compounds in RPLC Oral session presented at CIRM Day, Liège, Belgium

Poster presentation

- International congress

1. Kumari, P., Van Laethem, T., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C (2024,June). Exploring Diverse Methods for Quantitative Structure-Property Predictions (QS(X)R of Small Molecules. Poster session presented at Summer Innovation Programme(2024), Washington D.C, United States of America.
2. Kumari, P., Van Laethem, T., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C (2021, August). A QSRR modelling of small pharmaceutical compounds in Reverse Phase Liquid Chromatography. Poster session presented at Machine Learning Summer School(MLSS),Virtual event, Taipei, Taiwan.

- National Congress

1. Kumari, P., Van Laethem, T., Hubert, P., Fillet, M., Sacré, P. Y., & Hubert, C (2022, February). Defining a generic approach of structure derived retention prediction for small pharmaceutical compounds in RPLC. Poster session presented at CIRM Day, Liège, Belgium

Bibliography

[1] hitachi hightech.com. Principle and system configuration of hplc (1). URL https://www.hitachi-hightech.com/global/en/knowledge/analytical-systems/hplc/hplc-basics/course1.html.

[2] Roman Kaliszan. Quantitative structure property (retention) relationships in liquid chromatography. pages 553–572, 2017.

[3] Asad U Khan et al. Descriptors and their selection methods in qsar analysis: paradigm for drug design. *Drug discovery today*, 21(8):1291–1302, 2016.

[4] Bogusław Buszewski, Justyna Walczak-Skierska, and Paul R Haddad. Prediction of retention in liquid chromatography. In *Liquid Chromatography*, pages 795–819. Elsevier, 2023.

[5] Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23):5938–5951, 2022.

[6] Hitachi High-Tech Analytical Science. Introduction to high-performance liquid chromatography (hplc) - basics. https://www.hitachi-hightech.com/global/en/knowledge/analytical-systems/hplc/hplc-basics/course1.html. Accessed on January 25, 2024.

[7] www.knauer.net. Hplc basics – principles and parameters. URL https://www.knauer.net/en/Systems-Solutions/Analytical-HPLC-UHPLC/HPLC-Basics---principles-and-parameters.

[8] shimadzu.com. High performance liquid chromatography (hplc) basics. URL https://www.ssi.shimadzu.com/service-support/faq/liquid-chromatography/knowledge-base/hplc-basics/index.html.

[9] chem.libretexts.org. General theory of column chromatography. URL https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Analytical_Chemistry_2.1_%28Harvey%29/12%3A_Chromatographic_and_Electrophoretic_Methods/12.02%3A_General_Theory_of_Column_Chromatography.

[10] Zhirong Cai and Hajime Naruse. Inverse analysis of experimental scale turbidity currents using deep learning neural networks. *Journal of Geophysical Research: Earth Surface*, 126(8):e2021JF006276, 2021.

[11] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[12] Sebastian Raschka. StackingCVClassifier – mlxtend, 2023. URL https://rasbt.github.io/mlxtend/user_guide/classifier/StackingCVClassifier/. Accessed: insert access date here.

[13] Xavier Domingo-Almenara, Carlos Guijas, Elizabeth Billings, J Rafael Montenegro-Burke, Winnie Uritboonthai, Aries E Aisporna, Emily Chen, H Paul Benton, and Gary Siuzdak. The metlin small molecule dataset for machine learning-based retention time prediction. *Nature communications*, 10(1):5811, 2019.

[14] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[15] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer, 2015.

[16] Károly Héberger. Quantitative structure–(chromatographic) retention relationships. *Journal of chromatography A*, 1158(1-2):273–305, 2007.

[17] Georges Guiochon and Claude L. Guillemin. Fundamentals of the chromatographic process: The thermodynamics of retention in gas chromatography. In *For Laboratory Analyses and On-Line Process Control*, volume 42 of *Journal of Chromatography Library*, pages 55–92. Elsevier, 1988.

[18] Giuseppe Marco Randazzo, David Tonoli, Stephanie Hambye, Davy Guillarme, Fabienne Jeanneret, Alessandra Nurisso, Laura Goracci, Julien Boccard, and Serge Rudaz. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. *Analytica chimica acta*, 916:8–16, 2016.

[19] Priyanka Kumari, Thomas Van Laethem, Philippe Hubert, Marianne Fillet, Pierre-Yves Sacré, and Cédric Hubert. Quantitative structure retention-relationship modeling: Towards an innovative general-purpose strategy. *Molecules*, 28(4):1696, 2023.

[20] Paula Beatriz Silva Passarin and Felipe Rebello Lourenço. Enhancing analytical development in the pharmaceutical industry: A doe-qsrr model for virtual method operable design region assessment. *Journal of Pharmaceutical and Biomedical Analysis*, 239:115907, 2024.

[21] Mariusz Zapadka, Mateusz Kaczmarek, Bogumiła Kupcewicz, Przemysław Dekowski, Agata Walkowiak, Adam Kokotkiewicz, Maria Łuczkiewicz, and

Adam Buciński. An application of qsrr approach and multiple linear regression method for lipophilicity assessment of flavonoids. *Journal of Pharmaceutical and Biomedical Analysis*, 164:681–689, 2019.

[22] Caihong Wang, Jinlan Zhang, Caisheng Wu, and Zhe Wang. A multiple-dimension liquid chromatography coupled with mass spectrometry data strategy for the rapid discovery and identification of unknown compounds from a chinese herbal formula (er-xian decoction). *Journal of Chromatography A*, 1518:59–69, 2017.

[23] Qianqian Shen, Wenwen Tao, Yujie Guo, Shijia Wang, Yanfei Wang, Er mei Zheng, Zhongxiu Chen, and Kexian Chen. Quantitative structure-retention relationships of the chromatographic retentions of phthalic acid ester contaminants in foods. *Journal of separation science*, 42(17):2771–2778, 2019.

[24] Philippe J. Eugster et al. Retention time prediction for dereplication of natural products (cxhyoz) in lc–ms metabolite profiling. *Phytochemistry*, 108:196–207, 2014.

[25] Strahinja Z Kovačević, Sanja O Podunavac-Kuzmanović, Lidija R Jevrić, Pavle T Jovanov, Evgenija A Djurendić, and Jovana J Ajduković. Comprehensive qsrr modeling as a starting point in characterization and further development of anticancer drugs based on 17α-picolyl and 17 (e)-picolinylidene androstane structures. *European Journal of Pharmaceutical Sciences*, 93:1–10, 2016.

[26] Hamzeh Karimi, Abbas Farmany, and Hadi Noorizadeh. Chemometrics analysis for investigation of retention behavior of hazardous compounds in effluents. *Environmental monitoring and assessment*, 185:473–483, 2013.

[27] JM Sutter, TA Peterson, and PC Jurs. Prediction of gas chromatographic retention indices of alkylbenzenes. *Analytica chimica acta*, 342(2-3):113–122, 1997.

[28] Miles McGibbon, Steven Shave, Jie Dong, Yumiao Gao, Douglas R Houston, Jiancong Xie, Yuedong Yang, Philippe Schwaller, and Vincent Blay. From intuition to ai: evolution of small molecule representations in drug discovery. *Briefings in bioinformatics*, 25(1):bbad422, 2024.

[29] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[30] Elizaveta S Fedorova, Dmitriy D Matyushin, Ivan V Plyushchenko, Andrey N Stavrianidi, and Aleksey K Buryak. Deep learning for retention time prediction in reversed-phase liquid chromatography. *Journal of Chromatography A*, 1664:462792, 2022.

[31] Qiong Yang, Hongchao Ji, Xiaqiong Fan, Zhimin Zhang, and Hongmei Lu. Retention time prediction in hydrophilic interaction liquid chromatography

with graph neural network and transfer learning. *Journal of Chromatography A*, 1656:462536, 2021.

[32] Alexander Kensert, Robbin Bouwmeester, Kyriakos Efthymiadis, Peter Van Broeck, Gert Desmet, and Deirdre Cabooter. Graph convolutional networks for improved prediction and interpretability of chromatographic retention data. *Analytical Chemistry*, 93(47):15633–15641, 2021.

[33] Yiming Nie, Jia Li, Xinying Yang, Xuben Hou, and Hao Fang. Development of qsrr model for hydroxamic acids using pca-ga-bp algorithm incorporated with molecular interaction-based features. *Frontiers in Chemistry*, 10:1056701, 2022.

[34] Faizan Sahigara, Kamel Mansouri, Davide Ballabio, Andrea Mauri, Viviana Consonni, and Roberto Todeschini. Comparison of different approaches to define the applicability domain of qsar models. *Molecules*, 17(5):4791–4810, 2012.

[35] R Kaliszan and H Foks. The relationship between the rm values and the connectivity indices for pyrazine carbothioamide derivatives. *Chromatographia*, 10(7):346–349, 1977.

[36] R Kaliszan. Correlation between the retention indices and the connectivity indices of alcohols and methyl esters with complex cyclic structure. *Chromatographia*, 10:529–531, 1977.

[37] Esther Forgács and Tibor Cserháti. *Molecular basis of chromatographic separation*. CRC Press, 1997.

[38] Mati Karelson, Victor S Lobanov, and Alan R Katritzky. Quantum-chemical descriptors in qsar/qspr studies. *Chemical reviews*, 96(3):1027–1044, 1996.

[39] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*. John Wiley & Sons, 2008.

[40] Tim Hancock, Raf Put, Danny Coomans, Yvan Vander Heyden, and Yvette Everingham. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic qsrr studies. *Chemometrics and Intelligent Laboratory Systems*, 76(2):185–196, 2005.

[41] MH Fatemi and E Baher. Prediction of retention factors in supercritical fluid chromatography using artificial neural network. *Journal of Analytical Chemistry*, 60:860–865, 2005.

[42] Orsolya Farkas and Károly Héberger. Comparison of ridge regression, partial least-squares, pairwise correlation, forward-and best subset selection methods for prediction of retention indices for aliphatic alcohols. *Journal of chemical information and modeling*, 45(2):339–346, 2005.

[43] Minghu Song, Curt M Breneman, Jinbo Bi, Nagamani Sukumar, Kristin P Bennett, Steven Cramer, and Nihal Tugcu. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of chemical information and computer sciences*, 42(6): 1347–1357, 2002.

[44] Giuseppe Marco Randazzo, Andrea Bileck, Andrea Danani, Bruno Vogt, and Michael Groessl. Steroid identification via deep learning retention time predictions and two-dimensional gas chromatography-high resolution mass spectrometry. *Journal of Chromatography A*, 1612:460661, 2020.

[45] Qi He, Hua Li, Binyan Jin, Wei Li, Bing Shao, and Li Zhang. Qsrr model for identification and screening of emerging pollutants based on artificial intelligence algorithms. *Environmental Pollutants and Bioavailability*, 34(1): 331–337, 2022.

[46] Bingyi Wang et al. Rt-tranformer: Retention time prediction for metabolite annotation to assist in metabolite identification. 2023.

[47] Karolina Bodzioch, Alexandra Durand, R Kaliszan, T Bączek, and Yvan Vander Heyden. Advanced qsrr modeling of peptides behavior in rplc. *Talanta*, 81(4-5):1711–1718, 2010.

[48] Chakshu Vats, Jaspreet Kaur Dhanjal, Sukriti Goyal, Navneeta Bharadvaja, and Abhinav Grover. Computational design of novel flavonoid analogues as potential ache inhibitors: analysis using group-based qsar, molecular docking and molecular dynamics simulations. *Structural Chemistry*, 26:467–476, 2015.

[49] Raf Put and Yvan Vander Heyden. Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure–retention relationships. *Analytica chimica acta*, 602(2):164–172, 2007.

[50] LR Snyder, JW Dolan, and JR Gant. Gradient elution in high-performance liquid chromatography: I. theoretical basis for reversed-phase systems. *Journal of Chromatography A*, 165(1):3–30, 1979.

[51] Roman Kaliszan, Tomasz Bączek, Adam Buciński, Bogusław Buszewski, and Małgorzata Sztupecka. Prediction of gradient retention from the linear solvent strength (lss) model, quantitative structure-retention relationships (qsrr), and artificial neural networks (ann). *Journal of separation science*, 26(3-4):271–282, 2003.

[52] Colin F Poole and Sanka N Atapattu. Analysis of the solvent strength parameter (linear solvent strength model) for isocratic separations in reversed-phase liquid chromatography. *Journal of Chromatography A*, 1675:463153, 2022.

[53] Martin Gilar et al. Utility of linear and nonlinear models for retention prediction in liquid chromatography. *Journal of Chromatography A*, 1613: 460690, 2020.

[54] Colin F Poole and Nicole Lenca. Applications of the solvation parameter model in reversed-phase liquid chromatography. *Journal of Chromatography A*, 1486:2–19, 2017.

[55] Giuseppe Marco Randazzo, Evelyne Vigneau, Philippe Courcoux, Corentin Harrouet, Yves Lijour, Pierre Dardenne, Julien Boccard, and Serge Rudaz. Indirect quantitative structure-retention relationship for steroid identification: A chemometric challenge at "chimiométrie 2016". *Chemometrics and Intelligent Laboratory Systems*, 160:52–58, 2017.

[56] Paul C Sadek, Peter W Carr, Ruth M Doherty, Mortimer J Kamlet, Robert W Taft, and Michael H Abraham. Study of retention processes in reversed-phase high-performance liquid chromatography by the use of the solvatochromic comparison method. *Analytical chemistry*, 57(14):2971–2978, 1985.

[57] Peter W Carr, Ruth M Doherty, Mortimer J Kamlet, Robert W Taft, Wayne Melander, and Csaba Horvath. Study of temperature and mobile-phase effects in reversed-phase high-performance liquid chromatography by the use of the solvatochromic comparison method. *Analytical chemistry*, 58(13): 2674–2680, 1986.

[58] Michael H. Abraham. Scales of solute hydrogen-bonding: Their construction and application to physicochemical and biochemical processes. *Chem. Soc. Rev.*, 22(2):73–83, 1993.

[59] Paweł Wiczling, Łukasz Kubik, and Roman Kaliszan. Maximum a posteriori bayesian estimation of chromatographic parameters by limited number of experiments. *Analytical chemistry*, 87(14):7241–7249, 2015.

[60] R. Bouwmeester, L. Martens, and S. Degroeve. Comprehensive and empirical evaluation of machine learning algorithms for small molecule lc retention time prediction. *Anal. Chem.*, 91:3694–3703, 2019.

[61] Sara M de Cripan, Adrià Cereto-Massagué, Pol Herrero, Andrei Barcaru, Núria Canela, and Xavier Domingo-Almenara. Machine learning-based retention time prediction of trimethylsilyl derivatives of metabolites. *Biomedicines*, 10(4):879, 2022.

[62] Roman Kaliszan. Qsrr: quantitative structure-(chromatographic) retention relationships. *Chemical reviews*, 107(7):3212–3246, 2007.

[63] Roman Kaliszan. Recent advances in quantitative structure–retention relationships. *Handbook of Analytical Separations*, 8:587–632, 2020.

[64] Raf Put, Michal Daszykowski, T Baczek, and Yvan Vander Heyden. Retention prediction of peptides based on uninformative variable elimination by partial least squares. *Journal of proteome research*, 5(7):1618–1625, 2006.

[65] Šime Ukić, Mirjana Novak, Petar Žuvela, Nebojša Avdalović, Yan Liu, Bogusław Buszewski, and Tomislav Bolanča. Development of gradient retention model in ion chromatography. part i: conventional qsrr approach. *Chromatographia*, 77:985–996, 2014.

[66] Da Chen, Bin Hu, Xueguang Shao, and Qingde Su. Variable selection by modified ipw (iterative predictor weighting)-pls (partial least squares) in continuous wavelet regression models. *Analyst*, 129(7):664–669, 2004.

[67] Hassan Golmohammadi, Zahra Dashtbozorgi, and Yvan Vander Heyden. Support vector regression based qspr for the prediction of retention time of peptides in reversed-phase liquid chromatography. *Chromatographia*, 78: 7–19, 2015.

[68] Vladimir Dobricic, Katarina Nikolic, Sote Vladimirov, and Olivera Cudina. Biopartitioningmicellar chromatographyas a predictive tool for skin and corneal permeability of newly synthesized 17β-carboxamide steroids. *Eur. J. Pharm. Sci*, 56:105–112, 2014.

[69] Petar Zuvela, J Jay Liu, Katarzyna Macur, and Tomasz Baczek. Molecular descriptor subset selection in theoretical peptide quantitative structure–retention relationship model development using nature-inspired optimization algorithms. *Analytical chemistry*, 87(19):9876–9883, 2015.

[70] HX Liu, RS Zhang, XJ Yao, MC Liu, ZD Hu, and Bo Tao Fan. Prediction of the isoelectric point of an amino acid based on ga-pls and svms. *Journal of chemical information and computer sciences*, 44(1):161–167, 2004.

[71] Federico Marini and Beata Walczak. Particle swarm optimization (pso). a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 149:153–165, 2015.

[72] Yanhong Lin, Jing Wang, Xiaolin Li, Yuanzi Zhang, and Shiguo Huang. An improved artificial bee colony for feature selection in qsar. *Algorithms*, 14 (4):120, 2021.

[73] Mohammad Goodarzi and Leandro dos Santos Coelho. Firefly as a novel swarm intelligence variable selection method in spectroscopy. *Analytica chimica acta*, 852:20–27, 2014.

[74] Mohamed Abdel-Basset and Laila A Shawky. Flower pollination algorithm: a comprehensive review. *Artificial Intelligence Review*, 52:2533–2557, 2019.

[75] L Nørgaard, A Saudland, J Wagner, J Pram Nielsen, L Munck, and S Balling Engelsen. Interval partial least-squares regression (i pls): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied spectroscopy*, 54(3):413–419, 2000.

[76] Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1):3–25, 2010.

[77] Nada Perisic-Janjic, Roman Kaliszan, Natasa Milosevic, Gordana Uscumlic, and Nebojsa Banjac. Chromatographic retention parameters in correlation analysis with in silico biological descriptors of a novel series of n-phenyl-3-methyl succinimide derivatives. *Journal of Pharmaceutical and Biomedical Analysis*, 72:65–73, 2013.

[78] Ran Ju, Xinyu Liu, Fujian Zheng, Xin Lu, Guowang Xu, and Xiaohui Lin. Deep neural network pretrained by weighted autoencoders and transfer learning for retention time prediction of small molecules. *Analytical Chemistry*, 93(47):15651–15658, 2021.

[79] Youngchun Kwon, Hyukju Kwon, Jongmin Han, Myeonginn Kang, Ji-Yeong Kim, Dongyeeb Shin, Youn-Suk Choi, and Seokho Kang. Retention time prediction through learning from a small training data set with a pretrained graph neural network. *Analytical Chemistry*, 95(47):17273–17283, 2023.

[80] Dennis V. Lindley and Adrian F.M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(1):1–18, 1972.

[81] George C. Tiao and Arnold Zellner. On the bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):277–285, 1964.

[82] Łukasz Kubik, Roman Kaliszan, and Paweł Wiczling. Analysis of isocratic-chromatographic-retention data using bayesian multilevel modeling. *Analytical chemistry*, 90(22):13670–13679, 2018.

[83] Qi Shi, Mohamed Abdel-Aty, and Jaeyoung Lee. A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. *Accident Analysis & Prevention*, 88:124–137, 2016.

[84] Jonas Ranstam and J. A. Cook. Lasso regression. *Journal of British Surgery*, 105(10):1348, 2018.

[85] Donald F. Specht. A general regression neural network. *IEEE transactions on neural networks*, 2(6):568–576, 1991.

[86] Nigel Duffy and David Helmbold. Boosting methods for regression. *Machine Learning*, 47:153–200, 2002.

[87] Richard Zemel and Toniann Pitassi. A gradient-based boosting algorithm for regression problems. In *Advances in neural information processing systems*, volume 13, 2000.

[88] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[89] Mohammad Goodarzi, Richard Jensen, and Yvan Vander Heyden. Qsrr modeling for diverse drugs using different feature selection methods coupled

with linear and nonlinear regressions. *Journal of Chromatography B*, 910: 84–94, 2012.

[90] Julien Boccard, Jean-Luc Veuthey, and Serge Rudaz. Knowledge discovery in metabolomics: an overview of ms data handling. *Journal of Separation Science*, 33(3):290–304, 2010.

[91] Jovana Krmar, Bojana Svrkota, Nevena Đajić, Jevrem Stojanović, Ana Protić, and Biljana Otašević. Correction to: Qsrr approach: Application to retention mechanism in liquid chromatography. In *Novel Aspects of Gas Chromatography and Chemometrics*. IntechOpen, 2022.

[92] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[93] ChemAxon. Chemaxon: Solutions for chemistry & life science. https://docs.ochem.eu/x/IwJr.html, 2023. Accessed: 2024-02-15.

[94] Admet predictor. https://admet-predictor.software.informer.com/. Accessed: 2024-02-15.

[95] Alvadesc descriptors. https://www.alvascience.com/alvadesc-descriptors/. Accessed: 2024-02-15.

[96] Cheng Fan, Meiling Chen, Xinghua Wang, Jiayuan Wang, and Bufu Huang. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, 9:652801, 2021.

[97] Murali Shanker, Michael Y Hu, and Ming S Hung. Effect of data standardization on neural network training. *Omega*, 24(4):385–397, 1996.

[98] Matthias Feurer and Frank Hutter. Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges*, pages 3–33, 2019.

[99] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019.

[100] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[101] Peixun Liu and Wei Long. Current mathematical methods used in qsar/qspr studies. *International Journal of Molecular Sciences*, 10(5):1978–1998, 2009.

[102] Jovana et al. Krmar. Revealing retention mechanisms in liquid chromatography: Qsrr approach. In *Novel Aspects in Gas Chromatography and Chemometrics*. IntechOpen, 2022.

[103] Morten Arendt Rasmussen and Rasmus Bro. A tutorial on the lasso approach to sparse modeling. *Chemometrics and Intelligent Laboratory Systems*, 119: 21–31, 2012.

[104] Sabrina Hamla, Pierre-Yves Sacré, Allison Derenne, Kheiro-Mouna Derfoufi, Ben Cowper, Claire I Butré, Arnaud Delobel, Erik Goormaghtigh, Philippe Hubert, and Eric Ziemons. A new alternative tool to analyse glycosylation in pharmaceutical proteins based on infrared spectroscopy combined with nonlinear support vector regression. *Analyst*, 147(6):1086–1098, 2022.

[105] L Zheng, DG Watson, BF Johnston, Rachael L Clark, Ruangelie Edrada-Ebel, and W Elseheri. A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with pca, support vector machines and random forest data modeling. *Analytica Chimica Acta*, 642 (1-2):257–265, 2009.

[106] Kelly Munro, Thomas H Miller, Claudia PB Martins, Anthony M Edge, David A Cowan, and Leon P Barron. Artificial neural network modelling of pharmaceutical residue retention times in wastewater extracts using gradient liquid chromatography-high resolution mass spectrometry data. *Journal of Chromatography A*, 1396:34–44, 2015.

[107] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.

[108] Mark A Hall. Correlation-based feature selection of discrete and numeric class machine learning. 2000.

[109] N Gopika and A Meena Kowshalaya ME. Correlation based feature selection algorithm for machine learning. In *2018 3rd international conference on communication and electronics systems (ICCES)*, pages 692–695. IEEE, 2018.

[110] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

[111] Mital Doshi. Correlation based feature selection (cfs) technique to predict student perfromance. *International Journal of Computer Networks & Communications*, 6(3):197, 2014.

[112] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[113] Ioannis Tsamardinos, Giorgos Borboudakis, Pavlos Katsogridakis, Polyvios Pratikakis, and Vassilis Christophides. A greedy feature selection algorithm for big data of high dimensionality. *Machine learning*, 108:149–202, 2019.

[114] Kezhi Z Mao. Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):629–634, 2004.

[115] Riccardo Leardi, Riccardo Boggia, and M Terrile. Genetic algorithms as a strategy for feature selection. *Journal of chemometrics*, 6(5):267–281, 1992.

[116] Xue-wen Chen and Jong Cheol Jeong. Enhanced recursive feature elimination. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, pages 429–435. IEEE, 2007.

[117] Xiangyan Zeng, Yen-Wei Chen, and Caixia Tao. Feature selection using recursive feature elimination for handwritten digit recognition. In *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 1205–1208. IEEE, 2009.

[118] Rudolf Kiralj and Márcia Ferreira. Basic validation procedures for regression models in qsar and qspr studies: theory and application. *Journal of the Brazilian Chemical Society*, 20:770–787, 2009.

[119] Robyn Larracy, Angkoon Phinyomark, and Erik Scheme. Machine learning model validation for early stage studies with small sample sizes. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2314–2319. IEEE, 2021.

[120] Moona Emrarian, Mahmoud Reza Sohrabi, Nasser Goudarzi, and Fariba Tadayon. Retention time prediction of polycyclic aromatic hydrocarbons in gas chromatography–mass spectrometry using qspr based on random forests and artificial neural network. *Structural Chemistry*, 32:49–61, 2021.

[121] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[122] Gulyaim Sagandykova and Bogusław Buszewski. Perspectives and recent advances in quantitative structure-retention relationships for high-performance liquid chromatography. how far are we? *TrAC Trends in Analytical Chemistry*, 141:116294, 2021.

[123] Tatiana I Netzeva, Andrew P Worth, Tom Aldenberg, Romualdo Benigni, Mark TD Cronin, Paola Gramatica, Joanna S Jaworska, Scott Kahn, Gilles Klopman, Carol A Marchant, et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52. *Alternatives to Laboratory Animals*, 33(2):155–173, 2005.

[124] Joanna Jaworska, Nina Nikolova-Jeliazkova, and Tom Aldenberg. Qsar applicability domain estimation by projection of the training set in descriptor space: a review. *Alternatives to laboratory animals*, 33(5):445–459, 2005.

[125] Andrew P Worth, Arianna Bassan, Ana Gallegos, Tatiana I Netzeva, Grace Patlewicz, Manuela Pavan, Ivanka Tsakovska, and Marjan Vračko. *The characterisation of (quantitative) structure-activity relationships: preliminary guidance*. Institute for Health and Consumer Protection, Toxicology and Chemical . . . , 2005.

[126] Igor V Tetko, Iurii Sushko, Anil Kumar Pandey, Hao Zhu, Alexander Tropsha, Ester Papa, Tomas Oberg, Roberto Todeschini, Denis Fourches, and Alexandre Varnek. Critical assessment of qsar models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *Journal of chemical information and modeling*, 48(9):1733–1746, 2008.

[127] Faizan Sahigara, Davide Ballabio, Roberto Todeschini, and Viviana Consonni. Defining a novel k-nearest neighbours approach to assess the applicability domain of a qsar model for reliable predictions. *Journal of cheminformatics*, 5:1–9, 2013.

[128] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[129] Leo Breiman, William Meisel, and Edward Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

[130] Faizan Sahigara, Davide Ballabio, Roberto Todeschini, and Viviana Consonni. Assessing the validity of qsars for ready biodegradability of chemicals: an applicability domain perspective. *Current computer-aided drug design*, 10 (2):137–147, 2014.

[131] Thomas Van Laethem, Priyanka Kumari, Philippe Hubert, Marianne Fillet, Pierre-Yves Sacré, and Cédric Hubert. A pharmaceutical-related molecules dataset for reversed-phase chromatography retention time prediction built on combining ph and gradient time conditions. *Data in Brief*, 42:108017, 2022.

[132] Agnieszka Kamedulska, Łukasz Kubik, Julia Jacyna, Wiktoria Struck-Lewicka, Michał J Markuszewski, and Paweł Wiczling. Toward the general mechanistic model of liquid chromatographic retention. *Analytical Chemistry*, 94(31):11070–11080, 2022.

[133] Wen Wang, Wei Zhang, Shukai Liu, Qi Liu, Bo Zhang, Leyu Lin, and Hongyuan Zha. Incorporating link prediction into multi-relational item graph modeling for session-based recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2683–2696, 2021.

[134] Nikita Basant and Shikha Gupta. Multi-target qspr modeling for simultaneous prediction of multiple gas-phase kinetic rate constants of diverse chemicals. *Atmospheric Environment*, 177:166–174, 2018.

[135] Dimitrios Iliadis, Bernard De Baets, and Willem Waegeman. Multi-target prediction for dummies using two-branch neural networks. *Machine Learning*, 111(2):651–684, 2022.

[136] Zahra Garkani-Nejad and Mohammad Ahmadvand. Comparative qsrr modeling of nitrobenzene derivatives based on original molecular descriptors and multivariate image analysis descriptors. *Chromatographia*, 73:733–742, 2011.

[137] Hamideh Barfeii and Zahra Garkani-Nejad. A comparative qsrr study on enantioseparation of ethanol ester enantiomers in hplc using multivariate image analysis, quantum mechanical and structural descriptors. *Journal of the Chinese Chemical Society*, 64(2):176–187, 2017.

[138] Hiroshi Tsugawa, Ryo Nakabayashi, Tetsuya Mori, Yutaka Yamada, Mikiko Takahashi, Amit Rai, Ryosuke Sugiyama, Hiroyuki Yamamoto, Taiki Nakaya, Mami Yamazaki, et al. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nature methods*, 16 (4):295–298, 2019.

[139] Paolo Bonini, Tobias Kind, Hiroshi Tsugawa, Dinesh Kumar Barupal, and Oliver Fiehn. Retip: retention time prediction for compound annotation in untargeted metabolomics. *Analytical chemistry*, 92(11):7515–7522, 2020.

[140] I. François, K. Sandra, and P. Sandra. Comprehensive liquid chromatography: Fundamental aspects and practical considerations—a review. *Anal. Chim. Acta*, 641:14–31, 2009.

[141] R. Kaliszan. Quantitative structure—retention relationships (qsrr) in chromatography. pages 4063–4075, 2000.

[142] K. Ciura, M. Belka, P. Kawczak, T. Bączek, and J. Nowakowska. The comparative study of micellar tlc and rp-tlc as potential tools for lipophilicity assessment based on qsrr approach. *J. Pharm. Biomed. Anal.*, 149:70–79, 2018.

[143] Y. Ren, H. Liu, X. Yao, and M. Liu. An accurate qsrr model for the prediction of the gc×gc–tofms retention time of polychlorinated biphenyl (pcb) congeners. *Anal. Bioanal. Chem.*, 388:165–172, 2007.

[144] K. Goryn´ski, B. Bojko, A. Nowaczyk, A. Bucin´ ski, J. Pawliszyn, and R. Kaliszan. Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds. *Anal. Chim. Acta*, 797:13–19, 2013.

[145] P. Žuvela, K. Macur, J.J. Liu, and T. Bączek. Exploiting non-linear relationships between retention time and molecular structure of peptides originating from proteomes and comparing three multivariate approaches. *J. Pharm. Biomed. Anal.*, 127:94–100, 2016.

[146] Y. Wen, M. Talebi, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, and P.R. Haddad. Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the hydrophobic subtraction model. *J. Chromatogr. A*, 1541:1–11, 2018.

[147] Z. Xu, H. Chughtai, L. Tian, L. Liu, J.F. Roy, and S. Bayen. Development of quantitative structure-retention relationship models to improve the identification of leachables in food packaging using non-targeted analysis. *Talanta*, 253:123861, 2023.

[148] A.A. D'Archivio, M.A. Maggi, and F. Ruggieri. Artificial neural network prediction of multilinear gradient retention in reversed-phase hplc: Comprehensive qsrr-based models combining categorical or structural solute descriptors and gradient profile parameters. *Anal. Bioanal. Chem.*, 407:1181–1190, 2015.

[149] K. Ciura, S. Kovacevic, M. Pastewska, H. Kapica, M. Kornela, and W. Sawicki. Prediction of the chromatographic hydrophobicity index with immobilized artificial membrane chromatography using simple molecular descriptors and artificial neural networks. *J. Chromatogr. A*, 1660:462666, 2021.

[150] J. Krmar, M. Vukićević, A. Kovačević, A. Protić, M. Zečević, and B. Otašević. Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure-retention relationships modelling in micellar liquid chromatography. *J. Chromatogr. A*, 1623:461–146, 2020.

[151] M. Pastewska, B. Bednarczyk-Cwynar, S. Kovac ̌evic ́, N. Buławska, S. Ulenberg, P. Georgiev, H. Kapica, P. Kawczak, T. Bączek, W. Sawicki, and et al. Multivariate assessment of anticancer oleanane triterpenoids lipophilicity. *J. Chromatogr. A*, 1656:462552, 2021.

[152] S. Ulenberg, K. Ciura, P. Georgiev, M. Pastewska, G. ́Slifirski, M. Król, F. Herold, and T. Bączek. Use of biomimetic chromatography and in vitro assay to develop predictive ga-mlr model for use in drug-property prediction among anti-depressant drug candidates. *Microchem. J.*, 175:107183, 2022.

[153] A. Kamedulska, Ł. Kubik, and P. Wiczling. Statistical analysis of isocratic chromatographic data using bayesian modeling. *Anal. Bioanal. Chem.*, 414: 3471–3481, 2022.

[154] P. Wiczling, A. Kamedulska, and Ł. Kubik. Application of bayesian multilevel modeling in the quantitative structure–retention relationship studies of heterogeneous compounds. *Anal. Chem.*, 93:6961–6971, 2021.

[155] M. Ghorbanzadeh, K.I. van Ede, M. Larsson, M.B. van Duursen, L. Poellinger, S. Lucke-Johansson, M. Machala, K. Pencikova, J. Vondracek, M. van den Berg, et al. In vitro and in silico derived relative effect potencies of ah-receptor-mediated effects by pcdd/fs and pcbs in rat, mouse, and guinea pig calux cell lines. *Chem. Res. Toxicol.*, 27:1120–1132, 2014.

[156] C. Sutton, M. Boley, L.M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.*, 11:4428, 2020.

[157] A. Al-Fakih, Z. Algamal, M. Lee, and M. Aziz. A sparse qsrr model for predicting retention indices of essential oils based on robust screening approach. *SAR QSAR Environ. Res.*, 28:691–703, 2017.

[158] M.A. Fouad, E.H. Tolba, A. Manal, and A.M. El Kerdawy. Qsrr modeling for the chromatographic retention behavior of some β-lactam antibiotics using forward and firefly variable selection algorithms coupled with multiple linear regression. *J. Chromatogr. A*, 1549:51–62, 2018.

[159] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, et al. Classyfire: Automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.*, 8:61, 2016.

[160] R. Muthukrishnan and R. Rohini. Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pages 18–20, 2016.

[161] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *J. Chem. Inf. Comput. Sci.*, 43:1947–1958, 2003.

[162] T. Hepp, M. Schmid, O. Gefeller, E. Waldmann, and A. Mayr. Approaches to regularized regression–a comparison between gradient boosting and the lasso. *Methods Inf. Med.*, 55:422–430, 2016.

[163] R. Kaliszan. Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography. *J. Chromatogr. A*, 656:417–435, 1993.

[164] The rdkit 2022.03.1 documentation, 2022. URL https://www.rdkit.org/docs/. Accessed on 1 January 2023.

[165] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, V.V. Kovalishyn, V.V. Prokopenko, and I.V. Tetko. Applicability domain for in silico models to achieve accuracy of experimental measurements. *J. Chemom.*, 24:202–208, 2010.

[166] T. Van Laethem, P. Kumari, P. Hubert, M. Fillet, P.Y. Sacré, and C. Hubert. A pharmaceutical-related molecules dataset for reversed-phase chromatography retention time prediction built on combining ph and gradient time conditions. *Data Brief*, 42:108017, 2022.

[167] Ł. Kubik and P. Wiczling. Quantitative structure-(chromatographic) retention relationship models for dissociating compounds. *J. Pharm. Biomed. Anal.*, 127:176–183, 2016.

[168] Maryam Taraji, Paul R Haddad, Ruth IJ Amos, Mohammad Talebi, Roman Szucs, John W Dolan, and Chris A Pohl. Rapid method development in hydrophilic interaction liquid chromatography for pharmaceutical analysis using a combination of quantitative structure–retention relationships and design of experiments. *Analytical chemistry*, 89(3):1870–1878, 2017.

[169] E. Tyteca, M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, and P.R. Haddad. Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: Use of retention factor ratio. *J. Chromatogr. A*, 1486: 50–58, 2017.

[170] A. Daina, O. Michielin, and V. Zoete. Swissadme: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.*, 7:42717, 2017.

[171] S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, et al. Pubchem substance and compound databases. *Nucleic Acids Res.*, 44:D1202–D1213, 2016.

[172] H. Liu, E.R. Dougherty, J.G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, et al. Evolving feature selection. *IEEE Intell. Syst.*, 20:64–76, 2005.

[173] P. Charoenkwan, W. Chiangjong, C. Nantasenamat, M.M. Hasan, B. Manavalan, and W. Shoombuatong. Stackil6: A stacking ensemble model for improving the prediction of il-6 inducing peptides. *Brief. Bioinform.*, 22: bbab172, 2021.

[174] K. Matlock, C. de Niz, R. Rahman, S. Ghosh, and R. Pal. Investigation of model stacking for drug sensitivity prediction. *BMC Bioinform.*, 19:33, 2018.

[175] M. Awad and R. Khanna. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Springer Nature, 2015.

[176] M.R. Segal. Machine learning benchmarks and random forest regression. Technical report, Center for Bioinformatics and Molecular Biostatistics, San Francisco, CA, USA, 2004.

[177] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Front. Neurorobot.*, 7:21, 2013.

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

[179] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R.C. Team, et al. *Package 'caret'*, 2020. URL https://cran.r-project.org/web/packages/caret/caret.pdf.

[180] H. Wickham, W. Chang, and M.H. Wickham. *Package 'ggplot2'*, 2016. URL https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. Create Elegant Data Visualisations Using the Grammar of Graphics: Version 2.

[181] Milano Chemometrics and QSAR Research Group. Applicability domain toolbox (for matlab), 2013. URL https://michem.unimib.it/download/matlab-toolboxes/applicability-domain-toolbox-for-matlab/. Accessed on 10 October 2022.

[182] Fabrice Gritti. Perspective on the future approaches to predict retention in liquid chromatography. *Analytical Chemistry*, 93(14):5653–5664, 2021.

[183] Eugene N Muratov, Jürgen Bajorath, Robert P Sheridan, Igor V Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I Oprea, Igor I Baskin, Alexandre Varnek, Adrian Roitberg, et al. Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564, 2020.

[184] Linlin Zhao, Wenyi Wang, Alexander Sedykh, and Hao Zhu. Experimental errors in qsar modeling sets: what we can do and what we cannot do. *ACS omega*, 2(6):2805–2812, 2017.

[185] Ruth IJ Amos, Eva Tyteca, Mohammad Talebi, Paul R Haddad, Roman Szucs, John W Dolan, and Christopher A Pohl. Benchmarking of computational methods for creation of retention models in quantitative structure–retention relationships studies. *Journal of Chemical Information and Modeling*, 57(11):2754–2762, 2017.

[186] Chrysostomi Zisi, Ioannis Sampsonidis, Stella Fasoula, Konstantinos Papachristos, Michael Witting, Helen G Gika, Panagiotis Nikitas, and Adriani Pappa-Louisi. Qsrr modeling for metabolite standards analyzed by two different chromatographic columns using multiple linear regression. *Metabolites*, 7(1):7, 2017.

[187] Bojana Svrkota, Jovana Krmar, Ana Protić, and Biljana Otašević. 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*, 1690:463776, 2023.

[188] Zhili Zhao, Jian Qin, Zhuoyue Gou, Yanan Zhang, and Yi Yang. Multi-task learning models for predicting active compounds. *Journal of Biomedical Informatics*, 108:103484, 2020.

[189] Antonio de la Vega de León, Beining Chen, and Valerie J Gillet. Effect of missing data on multitask prediction methods. *Journal of cheminformatics*, 10(1):1–12, 2018.

[190] Eelke B Lenselink and Pieter FW Stouten. Multitask machine learning models for predicting lipophilicity (logp) in the sampl7 challenge. *Journal of Computer-Aided Molecular Design*, 35(8):901–909, 2021.

[191] Bhanushee Sharma, Vijil Chenthamarakshan, Amit Dhurandhar, Shiranee Pereira, James A Hendler, Jonathan S Dordick, and Payel Das. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. *Scientific Reports*, 13(1):4908, 2023.

[192] Seid Hamzic, Richard Lewis, Sandrine Desrayaud, Cihan Soylu, Mike Fortunato, Gregori Gerebtzoff, and Raquel Rodríguez-Pérez. Predicting in vivo compound brain penetration using multi-task graph neural networks. *Journal of chemical information and modeling*, 62(13):3180–3190, 2022.

[193] KahYong Tiong, Zhenliang Ma, and Carl-William Palmqvist. Real-time train arrival time prediction at multiple stations and arbitrary times. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 793–798. IEEE, 2022.

[194] Dragi Kocev, Sašo Džeroski, Matt D White, Graeme R Newell, and Peter Griffioen. Using single-and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecological Modelling*, 220 (8):1159–1168, 2009.

[195] Alison J Burnham, John F MacGregor, and Roman Viveros. Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems*, 48(2):167–180, 1999.

[196] Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larranaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

[197] Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104:55–98, 2016.

[198] Beauregard Piccart. Algorithms for multi-target learning (algoritmes voor het leren van multi-target modellen). 2012.

[199] Koji Muteki, James E Morgado, George L Reid, Jian Wang, Gang Xue, Frank W Riley, Jeffrey W Harwood, David T Fortin, and Ian J Miller. Quantitative structure retention relationship models in an analytical quality by design framework: Simultaneously accounting for compound properties, mobile-phase conditions, and stationary-phase properties. *Industrial & Engineering Chemistry Research*, 52(35):12269–12284, 2013.

[200] Paul R Haddad, Maryam Taraji, and Roman Szücs. Prediction of analyte retention time in liquid chromatography. *Analytical Chemistry*, 93(1):228–256, 2020.

[201] Krzesimir Ciura et al. Application of reversed-phase thin layer chromatography and qsrr modelling for prediction of protein binding of selected β-blockers. *Journal of Pharmaceutical and Biomedical Analysis*, 176:112767, 2019.

[202] Piotr Kawczak and Tomasz Bączek. Recent theoretical and practical applications of micellar liquid chromatography (mlc) in pharmaceutical and biomedical analysis. *Open Chemistry*, 10(3):570–584, 2012.

[203] Henrik Linusson. Multi-output random forests, 2013.

[204] Martin Breskvar and Sašo Džeroski. Multi-target regression rules with random output selections. *IEEE Access*, 9:10509–10522, 2021.

[205] Damjan Kuznar, Martin Mozina, and Ivan Bratko. Curve prediction with kernel regression. In *Proceedings of the 1st workshop on learning from multi-label data*, pages 61–68, 2009.

[206] Zhongyang Han, Ying Liu, Jun Zhao, and Wei Wang. Real time prediction for converter gas tank levels based on multi-output least square support vector regressor. *Control Engineering Practice*, 20(12):1400–1409, 2012.

[207] Thomas Van Laethem, Priyanka Kumari, Bruno Boulanger, Philippe Hubert, Marianne Fillet, Pierre-Yves Sacré, and Cédric Hubert. User-driven strategy for in silico screening of reversed-phase liquid chromatography conditions for known pharmaceutical-related small molecules. *Molecules*, 27(23): 8306, 2022.

[208] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[209] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[210] Dulce G Pereira, Anabela Afonso, and Fátima Melo Medeiros. Overview of friedman's test and post-hoc analysis. *Communications in Statistics-Simulation and Computation*, 44(10):2636–2653, 2015.

[211] Maryam Taraji, Paul R Haddad, Ruth IJ Amos, Mohammad Talebi, Roman Szucs, John W Dolan, and Christopher A Pohl. Chemometric-assisted method development in hydrophilic interaction liquid chromatography: A review. *Analytica chimica acta*, 1000:20–40, 2018.

[212] Cristian Rojas, Pablo R Duchowicz, Piercosimo Tripaldi, and Reinaldo Pis Diez. Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. *Journal of Chromatography A*, 1422:277–288, 2015.

[213] Priyanka Kumari, Diane Duroux, Marianne Fillet, Pierre Yves Sacre, Cedric Hubert, et al. A multi-target qsrr approach to model retention times of small molecules in rplc. *Journal of Pharmaceutical and Biomedical Analysis*, page 115690, 2023.

[214] Angelo Antonio D'Archivio, Andrea Giannitto, and Maria Anna Maggi. Cross-column prediction of gas-chromatographic retention of polybrominated diphenyl ethers. *Journal of Chromatography A*, 1298:118–131, 2013.

[215] Pauric Bannigan, Matteo Aldeghi, Zeqing Bao, Florian Häse, Alan Aspuru-Guzik, and Christine Allen. Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews*, 175:113806, 2021.

[216] Rishikesh Magar, Yuyang Wang, Cooper Lorsung, Chen Liang, Hariharan Ramasubramanian, Peiyuan Li, and Amir Barati Farimani. Auglichem: data augmentation library of chemical structures for machine learning. *Machine Learning: Science and Technology*, 3(4):045015, 2022.

[217] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[218] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

[219] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. ACL, 2007.

[220] Oscar Day and Taghi M Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4:1–42, 2017.

[221] Petar Bursać, Miloš Kovačević, and Branislav Bajat. Instance-based transfer learning for soil organic carbon estimation. *Frontiers in Environmental Science*, 10:1003918, 2022.

[222] Sinno Jialin Pan, James T Kwok, Qiang Yang, et al. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008.

[223] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

[224] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[225] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[226] Andreas Maniatopoulos and Nikolaos Mitianoudis. Learnable leaky relu (lelelu): An alternative accuracy-optimized activation function. *Information*, 12(12):513, 2021.

[227] Mian Mian Lau and King Hann Lim. Review of adaptive activation function in deep neural network. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 686–690. IEEE, 2018.

[228] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316, 2017.

[229] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 2022.

[230] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[231] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[232] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[233] Elizaveta S. Fedorova, Dmitriy D. Matyushin, Ivan V. Plyushchenko, Andrey N. Stavrianidi, and Aleksey K. Buryak. Deep learning for retention time prediction in reversed-phase liquid chromatography. *Journal of Chromatography A*, 1664:462792, 2022. ISSN 0021-9673. doi:https://doi.org/10.1016/j.chroma.2021.462792. URL https://www.sciencedirect.com/science/article/pii/S0021967321009146.

[234] Dmitriy D. Matyushin and Aleksey K. Buryak. Gas chromatographic retention index prediction using multimodal machine learning. *IEEE Access*, 8: 223140–223155, 2020.

[235] Shifa Zhong, Jiajie Hu, Xiong Yu, and Huichun Zhang. 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*, 408:127998, 2021. ISSN 1385-8947. doi:https://doi.org/10.1016/j.cej.2020.127998. URL https://www.sciencedirect.com/science/article/pii/S1385894720341176.

[236] Qiong Yang, Hongchao Ji, Xiaqiong Fan, Zhimin Zhang, and Hongmei Lu. Retention time prediction in hydrophilic interaction liquid chromatography with graph neural network and transfer learning. *Journal of Chromatography A*, 1656:462536, oct 2021. doi:10.1016/j.chroma.2021.462536. URL https://doi.org/10.1016/j.chroma.2021.462536.

[237] Gordon M Crippen, Timothy F Havel, et al. *Distance geometry and molecular conformation*, volume 74. Research Studies Press Taunton, 1988.

[238] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[239] Th Hanser, Ph Jauffret, and Gérard Kaufmann. A new algorithm for exhaustive ring perception in a molecular graph. *Journal of Chemical Information and Computer Sciences*, 36(6):1146–1152, 1996.

[240] Nikita Basant and Shikha Gupta. Multi-target qstr modeling for simultaneous prediction of multiple toxicity endpoints of nano-metal oxides. *Nanotoxicology*, 11(3):339–350, 2017.

[241] Francesca Grisoni, Michael Moret, Robin Lingwood, and Gisbert Schneider. Bidirectional molecule generation with recurrent neural networks. *Journal of chemical information and modeling*, 60(3):1175–1183, 2020.

[242] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.