TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

OMV

Diplomarbeit

Kováts Retention Index Prediction for Gas Chromatography of Jet Fuel Components

ausgeführt am Institut für Chemische Technologien und Analytik

der Technischen Universität Wien

unter der Leitung von

Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Erwin Rosenberg

durch

Armig Kabrelian, BSc

[redacted]

Wien, September 2023

_____

Armig Kabrelian

Acknowledgments

Firstly, I would like to express my deepest gratitude to my professor and supervisor Erwin Rosenberg for granting me the chance to join his research group, for his guidance and belief in me. This project was truly inspired through his unconditional support.

I would like to extend my sincere thanks to Noemae Lim, MSc. and Dipl.-Ing. Bernhard Klampfl for supporting me through my work by sharing their knowledge and expertise and by offering their valuable time whenever I needed it. I want to mention also that I had the pleasure of working and collaborating with a fantastic team at our project partner OMV.

Furthermore, I am very grateful to the Austrian Research Promotion Agency (FFG) for funding this project under project number FO999892264 (‘Max-Power-to-Jet’) as well as to OMV Downstream GmbH for co-funding this work.

I would like to express my appreciation to my family for their unconditional love and to my fiancé for keeping my motivation high during this process.

This thesis is dedicated to my late father Dr. Vasken Kabrelian, who always raised me on his motto: “Where there is a will, there’s a way”.

Abstract

The global air traffic, whether as commercial flights or private jets, accounted for 10% of greenhouse gas emissions in 2019. Therefore, the interest has grown to produce sustainable aviation fuels (SAF) that reduce CO_2 emissions. However, to choose potential candidates for SAF production, several tests have to be conducted that investigate the properties of SAF as specified by the American Society for Testing and Materials (ASTM). However, these tests are extremely expensive and time-consuming.

Therefore, the concept of predicting fuel properties has become very important to save time and cost. This work focuses particularly on predicting the Kováts retention index – a standardized measure of gas chromatographic retention – on columns of different stationary phases. The Kováts retention index (RI) is of interest due to the fact that it can be used as a characteristic, substance-specific parameter to distinguish different hydrocarbon isomers within fuels. Since chemical databases do not contain the retention indices (RIs) of all possible compounds and the RIs of isomers are sometimes misassigned, this work attempts to develop a quantitative structure-property relationship (QSPR) model to predict Kováts retention indices.

The data set used in this study consisted of almost 400 compounds from different classes (alkanes, alkenes, cycloalkanes, aromatics, alcohols, acids, aldehydes, ketones and esters). The retention indices (RIs) have been mainly collected from the PubChem database for standard non-polar (DB1), semi-standard non-polar (DB5) and polar wax (PEG) columns. Furthermore, 266 different molecular descriptors (MD) were obtained from the online chemical database that describe the structure and shape of molecules.

To build the model, the data set was split into a training and a test set and was pre-processed by Pareto Scaling. The training set was used to train the model with the following regression methods: Partial Least Square (PLS), and Support Vector Machine Regression (SVM-R). Venetian Blinds were used as a cross validation method and tested on the test set.

The results show that the SVM-R as a non-linear model was better to correctly predict RIs of different compound classes on different columns. For DB1 the SVM model reaches a precision (RMSECV) of 12.8 RI units at a correlation R^2(CV) of 0.999 and a prediction precision (RMSEP) of 12.2 RI units at a correlation R^2(Pred) of 0.999. For DB5: RMSECV of 19.4 at R^2(CV) of 0.999 and RMSEP of 13.7 at R^2(Pred) of 0.998 was obtained. For PEG: RMSECV of 24.5 at R^2(CV) of 0.997 and RMSEP of 24.0 at R^2(Pred) of 0.997 was reached.

SUBJECT AREA: Chemometrics

KEYWORDS: Gas chromatography, Kováts Retention Index, Quantitative structure-property relationship, Modelling, Isomers

Kurzfassung

Der globale Flugverkehr, ob als kommerzielle Flüge oder Privatjets, trägt durch seine Emissionen mit etwa 10% zum Treibhauseffekt bei. Um die CO_2-Emissionen zu reduzieren, haben nachhaltige Flugkraftstoffe (*sustainable aviation fuels* - SAFs) großes Interesse geweckt. Um jedoch potenzielle Kandidaten für die SAF-Produktion zu finden, müssen zahlreiche Tests durchgeführt werden, die unterschiedliche SAF Kraftstoffeigenschaften untersuchen. Diese Tests beruhen auf ASTM-Standards . Jedoch sind diese Tests teuer und zeitaufwändig. Daher hat das Konzept der Vorhersage von Kraftstoffeigenschaften große Bedeutung gewonnen, da damit Zeit und Kosten gespart werden [1].

Diese Arbeit konzentriert sich insbesondere auf die Vorhersage der Kováts-Retentionsindices auf Säulen verschiedener stationärer Phasen. Der Kováts-Retentionsindex (RI) ist von Interesse, weil er als charakteristische, stoffspezifische Größe zur Unterscheidung verschiedener Kohlenwasserstoffisomere verwendet wird. Da chemische Datenbanken keine Retentionsindizes (RIs) aller möglichen Verbindungen enthalten und die RIs vieler Isomere im Chromatogramm falsch zugeordnet werden, wird im Zuge dieser Arbeit versucht, ein quantitatives Struktur-Eigenschafts-Beziehungsmodell (*quantitative structure property relationship* - QSPR) zu entwickeln, um die Kováts Retentionsindices vorherzusagen.

Der für diese Arbeit verwendete Datensatz bestand aus fast 400 Verbindungen verschiedener Stoffgruppen (Alkane, Alkene, Cycloalkane, Aromaten, Alkohole, Säuren, Aldehyde, Ketone und Ester). Die Retentionsindizes (RIs) wurden hauptsächlich aus der PubChem Datenbank für unpolare Standardsäulen (DB1), unpolare semi-Standardsäulen (DB5) und polare Säulen (PEG) gesammelt. Darüber hinaus wurden 266 verschiedene molekulare Deskriptoren (MD) aus der chemischen Online-Datenbank bezogen, die die Struktur und Form von Molekülen beschreiben.

Zur Erstellung des Modells wurde der Datensatz in Trainings- und Testsets aufgeteilt und durch Pareto-Skalierung vorverarbeitet. Das Trainingsset wurde verwendet, um das Modell mit folgenden Regressionsmethoden zu trainieren: Partial Least Square (PLS), Support Vector Machine Regression (SVM-R). Die Venetian Blind-Methode wurde zur Kreuzvalidierung verwendet und auf das Testset angewendet.

Das Ergebnis zeigt, dass der SVM-R als nichtlineares Modell besser geeignet ist, um RIs der Verbindungen verschiedener Stoffgruppen auf verschiedenen Säulen korrekt vorherzusagen. Für DB1 erreicht das SVM-Modell eine Präzision (RMSECV) von 12,8 RI-Einheiten bei einer Korrelation R^2(CV) von 0.999 und eine Vorhersagepräzision (RMSEP) von 12,2 RI-Einheiten bei einer Korrelation R^2(Pred) von 0.999. Für DB5: ist der RMSECV von 19,4 bei R^2(CV) von 0.999 und RMSEP von 13,7 bei R^2(Pred) von 0.998. Für PEG: beträgt RMSECV 24,5 bei R^2(CV) von 0.997 und RMSEP von 24,0 bei R^2(Pred) von 0.997.

FACHGEBIET: Chemometrie

SCHLAGWORTE: Gaschromatographie, Kováts Retentionsindex, Quantitative Struktur-Eigenschafts-Beziehungen, Modellierung, Isomere

List of Abbreviations

AED	Atomic emission detector
ANN	Artificial neural network
CI	Chemical ionization
CNN	Conventional neural network
DB1	Code of standard non-polar column produced by J&W (now: Agilent); 100% poyldimethylsiloxane
DB5	Code of Semi-standard non-polar column produced by J&W (now: Agilent); 5%-polydiphenyl-/95%-polydimethylsiloxane
DC-710	Code of standard non-polar column produced by J&W (now: Agilent); 50% polydiphenyl-/50%polydimethylsiloxane
ECD	Electron capture detector
EI	Electron ionization
ETA	extended topochemical atom
FAME	Fatty acid methyl ester
FID	Flame ionization detector
GC	Gas chromatograph, gas chromatography
GCxGC	Two-dimensional comprehensive GC
He	Helium
HPLC	High performance liquid chromatography
LC	Liquid chromatography
LOO	Leave one out
LV	Latent variable
LWR	Locally weighted regression
MAE	Mean absolute error
Max	Maximum
MD	Molecular descriptor
MLR	Multiple linear regression

Min	Minimum
MS	Mass spectrometer, mass spectrometry
m/z	Mass to charge ratio
OV-7	Code of non-standard non-polar column originally produced by Ohio Valley Specialty Co.; 20% polydiphenyl-/80% Polydimethylsiloxane
OV-25	Code of non-standard non-polar column originally produced by Ohio Valley Specialty Co.; 50%-polydiphenyl-/50%-polydimethylsiloxane
OV-225	Code of non-standard non-polar column originally produced by Ohio Valley Specialty Co.; 50% cyanopropylmethyl-/50%- phenylmethylpolysiloxane
PAH	Polycyclic aromatic hydrocarbon
PC	Principal component
PCR	Principal component regression
PEG	Polyethylene glycol
PLS	Partial least square
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
RBF	Radial basis function
RDF	Radial distribution function
RI	Retention index
RMSE	Root mean square error
RMSEC	Root mean square error for calibration
RMSECV	Root mean square error for cross validation
RMSEP	Root mean square error for prediction
RSD	Relative standard deviation
RT	Retention time
SAF	Sustainable aviation fuel
SE-30	Code of non-standard non-polar column. 100%-Polydimethylsiloxane (not crosslinked)
SEC	Standard error of calibration

Silar-5CP Code and trade name of non-standard semi-polar column. 50%-Cyanopropyl-/50% Phenylmethylpolysiloxane

SEP Standard error of prediction

SVM-R Support vector machine regression

TCD Thermal conductivity detector

TLC Thin layer chromatography

VSA Van der Waals surface area

XE-60 Code of non-standard semi-polar column. 50% Cyanopropylmethyl/50% Dimethylpolysiloxane

Table of Contents

1. Objective

The traditional and currently still most common source of jet fuel is of fossil origin, however with increasing interest to produce sustainable aviation fuel (SAF). In order to find potential candidates for SAF, several tests have to be conducted, which are very expensive and highly time-consuming. To save time and cost, chemometric models have been developed that can predict fuel properties [1–3]. To predict properties, a GCxGC chromatogram can be used as data input to provide substance-specific, quantitative information and molecular descriptors (MDs), combined with a quantitative structure-property relationship (QSPR) approach to derive relevant properties from the chemical composition (refer to figure 1). It is the aim of this thesis to develop a QSPR model with the help of MDs to predict the Kováts retention index on columns of different stationary phases.

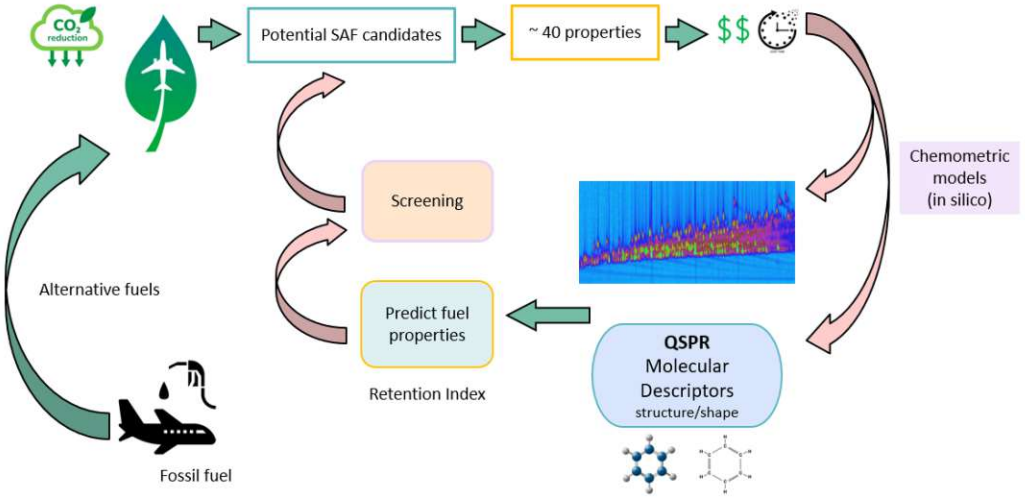


Figure 1: Illustration of the objective of this thesis.

2. Introduction

2.1. Chromatography

Michael Tswett observed for the first time the separation of coloured bands of chlorophyll pigments while applying the pigment solution to a column filled with $CaCO_3$ [4,5]. This technique was first described in the German botanical journal in 1906 [6].

The separation of colours was then named as “Chromatography” (derived from Greek language meaning ‘to write colour’).

However, the technique of chromatography was further developed by Archer John Porter Martin and Richard Laurence Millington Synge and in 1952 they were awarded the Nobel Prize in Chemistry for developing and establishing the principles of partition chromatography. This innovation encouraged the development of several more chromatographic methods such as high-performance liquid chromatography (HPLC), thin layer chromatography (TLC) and gas chromatography (GC) which is the main topic of this thesis and will be discussed in section 2.2 Gas Chromatography [7,8].

Any chromatographic separation is based on the distribution (partitioning) of the sample between the mobile phase and stationary phase. Stationary phase can be a solid or a liquid, whereas the mobile phase can be a liquid or a gas. The mobile phase carries the sample mixture through the stationary phase where the different constituents of the sample are separated based on the different intermolecular interaction between the analyte and the stationary phase. Most important factors affecting on the chromatographic separation are adsorption, partition and affinity between the analyte, mobile phase and the stationary phase. Because of these differences some analytes will elute faster or slower than others, thus resulting in different retention times [5,7].

Chromatography in general can be classified into two groups, column chromatography and planar chromatography, depending on how the stationary phase and mobile phase come into contact. Column Chromatography means the stationary phase is held in a column and the mobile phase runs through by gravity or pressure (e.g. liquid chromatography LC and gas chromatography GC). Whereas, planar chromatography means the stationary phase is lying flat and the mobile phase moves through by capillary action (e.g. thin layer chromatography TLC). The stationary phase can have different polarities, and based on that the chromatography can be classified either as normal phase (where the stationary phase is more polar than the mobile phase) or reversed phase (where the stationary phase is less polar than the mobile phase).

2.2. Gas Chromatography

The modern gas chromatography (GC) was invented in 1952 by James and Martin [9]. The main focus at that time was to separate amino acids. But now due to its high sensitivity and fast analysis, GC has become indispensable for many different scientific fields, providing qualitative and quantitative analysis. The instrumentation for gas chromatography mainly consists of a carrier gas system, injector, gas chromatographic column, detector and data processing unit. The carrier gas (mobile phase), is a

chemically inert gas, which does not influence the selectivity of the separation (e.g. helium). The injector helps to introduce the sample to the GC column. The basis of components separation in GC is by partitioning between two different phases, the stationary phase and the mobile phase (typically, He as the carrier gas) [15]. The separation in GC depends on the column properties such as different column dimensions (length, diameter, stationary phase film thickness) and different polarities (standard non-polar – 100% dimethylpolysiloxane, semi standard non-polar – 5%-Phenyl-/95% methylpolysiloxane or polar – polyethylene glycol). Based on the interaction between the column and the molecules, different compounds will elute from the column at different times (retention times). If coupled to a mass spectrometer as a detector, this allows the MS to ionize and detect the molecules separately [10–13].

Choosing the right temperature program in GC will help achieve an effective and reliable separation since the column temperature is one of the most decisive parameters. Hence, there are two different temperature modes, isothermal separation mode where the column is operated at a constant temperature and temperature programmed mode where a predetermined temperature program is applied [14].

Once the analyte exits the column (elution), it will reach the detector where a chromatogram will be generated, which is a plot of signal intensity versus elution time (retention time). For a successful chromatographic separation the detector should fulfil the following characteristics: High sensitivity, high selectivity, fast response time and being non-destructive. There are many different detectors which are developed for sensitive quantification of analytes such as flame ionization FID, electron capture ECD, thermal conductivity TCD, atomic emission AED, mass spectrometer MS, etc. [13].

Most frequently used detectors with GC are the FID, TCD and MS. In FID once the analyte mixture elutes the column, it will combust in hydrogen/air where ions will form and be detected by the electrode. The generated signal is proportional to the amount of organic carbon in the mixture. FID is the most commonly used as GC detector due to the fact that it is very robust and easy to use with high sensitivity and low noise, it has even high sensitivity for organic compounds, however it cannot detect water or carbon dioxide and is destructive (destroys the sample) [15]. TCD consists of double channel system with electrical heated filaments. The amount of heat loss is a function of the thermal conductivity of the carrier gas flowing through the cell. When the analyte flows through the detector it will change the gas composition and resulting in a change of conductivity [16]. Although TCD is less sensitive than FID, it is non-destructive and has an advantage over FID due to the fact that it can see and detect analytes that the FID does not see (e.g., permanent gases) [15].

Nowadays, GC is often coupled with mass spectrometry (MS) as information-rich detector. GC-MS is considered as a versatile analytical method due to its reproducibility, high resolution and capability of structural elucidation, it can separate mixtures of volatile and semi-volatile compounds with high selectivity and sensitivity [17]. The MS detects the molecules by ionizing them in a high vacuum system, using electron ionisation (EI) or chemical ionisation (CI) techniques. Because of the high-vacuum, MS requires an interface to direct the analytes from the GC to the MS (Refer to figure 2). The generated ions are then accelerated in a magnetic or electric field and sorted according to their m/z (mass to charge ratio), producing the mass spectrum. MS

data is generated as a chromatogram indicating the compound quantity as a function of retention time (elution time from the column) [18].

In 1991, the first comprehensive two-dimensional gas chromatogram (GCxGC) was reported by Liu and Phillips [19]. A typical GCxGC system consists of two GC columns with different dimensions and the both columns differ in their selectivity (e.g. selectivity of column 1 is based on volatility whereas column 2 is based on polarity). Multidimensional GC has gained interest over time to improve the separation of co-eluting compounds and for enrichment of trace components.

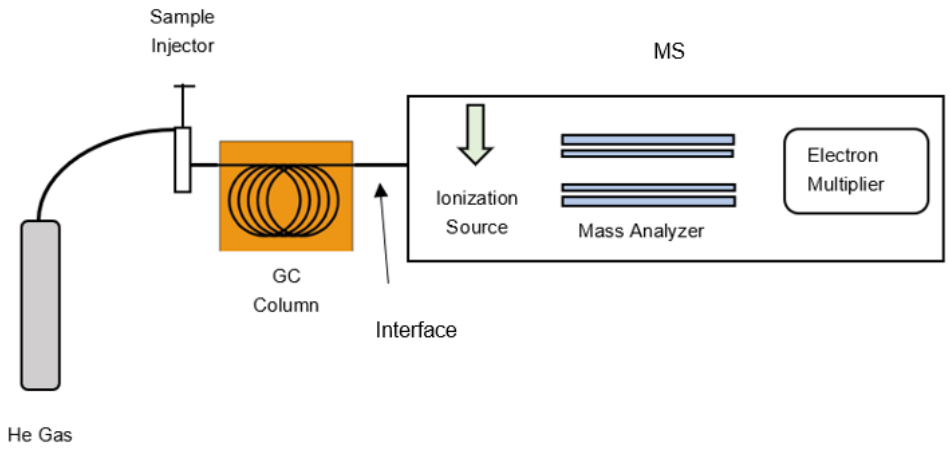


Figure 2:Schematic representation of GC-MS.

2.3. Chromatographic separation

When sample molecules pass through the GC column, they will spend different lengths of time in the stationary phase and in the mobile phase. The time that is required for an unretained solute to reach the detector is called the gas hold-up time (t_0). The unknown compound retention time (t_R) is the time between injection and when the maximum of the peak signal reaches the detector and ($t_R^{'}$) is the adjusted retention time of the unknown compound (refer to figure 3). The adjusted retention time is calculated by [20]:

$$t_R^{'} = t_R + t_0 \quad \text{.....(1)}$$

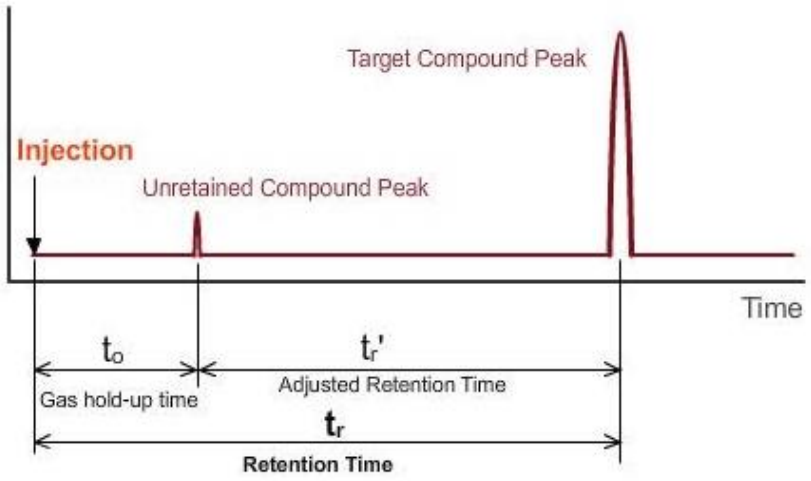


Figure 3: Schematic explanation of retention time [21].

The retention time is influenced by many factors such as temperature, stationary phase, column length etc. Hence, the retention time is not ideal for peak identification. And in order to overcome the problem of expressing retention data in a more general way and to enable the transfer of measured results from one laboratory to another, Kováts introduced the retention index system (RI) in 1958 [22]. In which a homologous series of *n*-alkanes were used as reference peaks in isothermal GC conditions [23].

2.4. Retention Index System

With the beginning of 1980s, the retention index has become a trending subject, where a great number of studies on retention index systems were published to improve their reliability, from which several linear relationships between the retention index and other fundamental properties (e.g. boiling point, melting point, carbon number and molecular weight) were derived [23,24].

Nevertheless, only recently it was recognized how useful RI is, as an independent additional parameter, in supporting the identification of unknown compounds in complex mixtures. Most important applications of RIs is to confirm correct identification of chemical compounds, filtering GC-MS false-positive identifications and the identification of isomers with similar mass spectra. Since mass spectral information alone is not fully reliable in assigning the identity of isomers [25].

Moreover, according to literature the retention parameters can be related simply to the thermodynamic partition coefficient between the gas and the liquid phase [26], and this led to the presumption that physico-chemical properties (such as boiling point) of a molecule are related to its chromatographic retention data (Kováts retention index) [27]. Some studies even used this information to predict physico-chemical properties of organic compounds with the help of mathematical relationships between the retention indices and physico-chemical properties [28].

2.4.1. RI in isothermal and temperature-programmed system

Retention indices (RIs) can be calculated for isothermal and temperature-programmed conditions, but it must be mentioned that the RI values differ between the two systems [23].

In isothermal condition, the column temperature is maintained constant during the analysis. This works best for samples with narrow boiling point distribution. In contrast to the isothermal, the heating rate and the temperature in the temperature-programmed operation can be adjusted during the analysis and this method is well suited for separating a mixture of a broad boiling point range. And during the analysis, components with low boiling points are separated first at low temperature and by increasing the temperature high boiling components are separated.

In 1958 Kováts proposed the first retention index system [22] which was considered for isothermal conditions, for which *n*-alkanes (*n*-C_nH_{2n+2}) were used as reference compounds to define the RI scale under isothermal GC conditions [29]:

$$RI = 100 * nC \text{.....(2)}$$

where nC is the number of carbon atoms of *n*-alkanes [e.g., ethane C_2H_6 has RI of 200]. To obtain the RI of compounds different than *n*-alkanes in isothermal conditions, the retention time of the unknown compound should be normalized to the retention times of adjacently eluting *n*-alkanes before and after the unknown compound [20,29]:

$$RI = 100 * \{k + [log(\frac{t'_{R,x}}{t'_{R,k}}) / log (\frac{t'_{R,k+1}}{t'_{R,k}})]\} \text{.....(3)}$$

Where k is the number of carbon atoms of the *n*-alkane eluting before the unknown compound, $t'_{R,x}$ is the adjusted retention time (RT) of the unknown compound, $t'_{R,k}$ is the RT of the *n*-alkane eluting before the unknown compound, $t'_{R,k}$ is the RT of the *n*-alkane eluting after the unknown compound [29].

In 1963 Van den Dool and Kratz derived the RI formula for linear temperature-programmed system [29–31]:

$$RI = 100 * [k + \frac{(t_{R,x} - t_{R,k})}{(t_{R,k+1} - t_{R,k})}] \text{.....(4)}$$

Where k is the number of carbon atoms of the *n*-alkane eluting before the unknown compound, $t_{R,x}$ is the RT of the unknown compound, $t_{R,k}$ is the RT of the *n*-alkane eluting before the target, $t_{R,k+1}$ is the RT of the *n*-alkane eluting after the unknown compound. The linear temperature-programmed conditions are very useful for separating complex mixtures like petroleum crude oils that contains constituents with a broad range of boiling points [29]. Moreover, obtaining reliable retention index data is often arguable, due to the different heating programs. Thus temperature-programmed conditions are usually less, not more reliable than the isothermal conditions.

Nevertheless, the retention index system is based on the incremental structure and retention relationship of the eluting compound, this means that the retention index system should not be limited to the use of *n*-alkanes as standards. Therefore, there are many other suggested systems than Kováts, such as the fatty acid methyl esters (FAMEs) and polynuclear aromatic hydrocarbons (PAHs) [23,32]. For FAME the RI is calculated by the formula [33]:

$$I = 100 * [k + \frac{(T_x - T_k)}{(T_{k+1} - T_k)}] \ldots\ldots\ldots\ldots (5)$$

Where, k is the number of carbon atoms of the FAME before the peak of the unknown compound, T_x the elution temperature of the unknown compound, T_k the elution temperature of the FAME before the unknown compound and T_{k+1} is the elution temperature of the FAME after the unknown compound.

And in the PAH retention index system, instead of using the retention times of alkanes, it uses the retention times of selected polycyclic aromatic hydrocarbons with the number of rings [32].

There were also other systems which were based on the equation proposed by Kováts such as the standard retention index suggested by Robinson and Odell in 1971 [34], where the reference parameter is not a retention time but boiling point of the unknown analytes and the reference standards.

The need to introduce different RI systems was due to the development of different element-specific detectors such as the electron capture detector (ECD) and the flame photometric detector (FPD) that are not sensitive to *n*-alkanes and therefore they need different standard homologous. For example, the series of chloroalkanes and bromoalkanes can be used with ECD, sulphur-containing molecules e.g. dialkylsulphides can be used with FPD [23]. Therefore, several RI systems were suggested for specific applications. To mention further, RIs generated in one system can be recalculated in the Kováts RIs system. Moreover, Kováts RI is usually the most suggested and most frequently used system, due to following reasons: *n*-alkane standards form a homologous series increasing one carbon at a time, many of them are usually available commercially in high purity, the first 20 linear alkanes already cover a wide range of boiling points, and they have a linear relationship between the carbon atom number of the *n*-alkanes and the logarithm of the corrected retention time [23].

2.4.2. RI Application for Phase Constant Determination

After the RI system was introduced by Kováts, new interests have been raised to understand and characterize different stationary phases. Hence, in 1966 Rohrschneider proposed in his work to investigate the macroscopic level of physico-chemical properties – by modelling the retention of known test compounds on the respective stationary phase [35]. Rohrschneider tried to characterize stationary phases of GC columns by studying the retention index differences (ΔI) of five different test

compounds on squalene (non-polar) stationary phase and on the stationary phase to be characterized, at 120 °C [35].

$$\Delta I = I_{characterized} - I_{squalane}\ [23]\ldots\ldots\ldots\ldots(6)$$

The used test compounds were the following: Benzene, ethanol, 2-butanone, nitromethane and pyridine.

Rohrschneider suggested further that RI differences $\Delta I_{i,j}$ for a solute i on stationary solvent phase j, can be expressed as a summation of five terms, each consisting of solute-specific and solvent-specific factors [35]:

$$\Delta I_{i,j} = a_i x_j + b_i y_j + c_i z_j + d_i u_j + e_i s_j \ \ldots\ldots\ldots\ldots(7)$$

Or more generally [36]:

$$\Delta I_{i,j} = \sum_{k=1}^{n} a_{i,k}\, x_{k,j} \ \ldots\ldots\ldots\ldots(8)$$

Where, $a_{i,k}$ is the solute-specific factor (for test compounds) and $x_{k,j}$ is the solvent-specific factor (for stationary phase). The $x_{k,j}$ factors are obtained experimentally, whereas, $a_{i,k}$ are calculated with the intention of minimising the sum of the squared errors for each substance on the stationary phase, meaning sum of the errors should be zero. Rohrschneider chose these five standard solutes to explain the different interactions of the solutes with different stationary phases [23,35,37,38] (refer to table 1).

Table 1: Interactions of different solutes with the stationary phase.

Test compounds	Interactions
Benzene	π – π, aromatic
Ethanol	Hydrogen bonding for alcohols
2-butanone	Proton acceptor – ketones, aldehydes
Nitromethane	Dipole – dipole interactions
Pyridine	Strong proton acceptor – acid character of column

Later on, McReynolds expanded the five compounds list that was suggested by Rohrschneider and included 10 compounds to characterize the columns even better. Benzene, 1-butanol, 2-pentanone, nitropropane, pyridine, 2-methyl-2-pentanol, 1-iodobutane, 2-octyne, 1,4-dioxane and *cis*-hidrindane. The first five compounds are either the same compounds as Rohrschneider used or homologs of Rohrschneider's compounds [39]. The purpose of McReynolds constants is to more comprehensively

describe retention properties and to give information on phase polarity, which can be used to rank the polarity of different stationary phases [40]. For example, if the McReynolds constants are high for all these 10 test compounds, then this signifies that the characterized (investigated) stationary phase is polar. If the McReynolds constants are not equally high for all these 10 test compounds then this indicates different degrees of polarity of the stationary phase.

For instance, a classical non-polar phase such as squalene, has a polarity number of zero (low value of McReynolds constants). On the other hand, a 50% phenyl-methyl polysiloxane stationary phase is considered moderately polar and has a polarity number of around 20, while polyethylene glycol phase, considered as one of the most polar stationary phases, has a polarity number of around 52 [41].

The McReynolds constants enable to compare polarities of different stationary phases. For instance, benzene has $X' = 0$ on squalene column, $X' = 0.16$ on 100% dimethyl polysiloxane stationary phase and $X' = 3.22$ on polyethylene glycol stationary phase [23,37,38]. Although the McReynolds system gives the advantage of determining and comparing polarities of different columns, it has still some limitations. Hence, the combination of both Rohrschneider and McReynolds data would help develop this research area.

2.4.3. Retention Index Prediction

The concept of Rohrschneider and McReynolds stimulated the idea of predicting retention indices. And the inspiration to develop predictive relationships was due to several reasons; To generate reliable retention data, or to predict retention index of an analyte whose RI is unavailable for study, and also to extrapolate valuable physical-chemical properties of an analyte that could be correlated through its retention index [23].

Despite the fact that several databases contain large amounts of RI data of different chemical compounds, there is still a lack of generally available retention data for many different compounds. Therefore, in the absence of experimental retention data, predicted retention data could be very helpful in confirming correct identification and avoiding false-positive identification [42].

A general approach for retention index prediction depends on generating topological, geometric and electronic molecular descriptors, which are then used to predict retention index using different regression methods.

The topological indices are indispensable for the study of the relationships between molecular structure and chromatographic retention data. As a consequence, they have been widely used to correlate retention indices with several important properties such as boiling point, molecular polarizability, van der Waals volumes [27].

The main used RI predictive approaches are the following: Quantitative structure property relationship (QSPR) approach with molecular descriptors (MD), deep learning approaches and non-learning methods based on functional group increments [43,44]. Hence, the section 2.5. QSPR Predictive Modeling will cover different approaches that

were used in various literature to predict RIs. However, this work mainly focuses on the QSPR approach with MD.

2.5. QSPR Predictive Modeling

The field of quantitative structure–activity relationships (QSARs) modelling was developed by Corwin Hansch [45], and deals with creating a model that relates the chemical structure of a compound (descriptor) to its activity. This modelling was further extended to the physical and chemical properties of the compound (melting point, boiling point or retention index) and was then called quantitative structure–property relationship (QSPR). QSPRs are found to be important complementary tools in computational chemistry to predict a variety of physicochemical properties for the purpose of industrial processes optimization. They have gained remarkable interest due to their fast and inexpensive computation. The prediction by this method relies on the use of descriptors, which derive information from the molecular structure. The descriptors encode numeric information about molecular topology, geometry and electronic features. Consequently, these values are used to build an accurate predictive model [43,44].

2.5.1. Modeling Techniques from Literature

This section discusses different predictive models for retention index, which were published over the years by different research groups.

The group of Biancolillo & D'Archivio [2] used the model of quantitative structure-property relationships (QSPR) to predict retention indices (RIs) of 90 saturated esters that were experimentally collected (isothermal program at 150 °C) on seven different stationary phases: 100%-dimethylpolysiloxane (SE-30), 20% diphenyl-80% dimethylpolysiloxane (OV-7), 50% diphenyl-/50% dimethylpolysiloxane (DC-710), 50% diphenyl-/50%-dimethylpolysiloxane (OV-25), 50% cyanopropylmethyl-/50% dimethylpolysiloxane (XE-60), 50% cyanopropylmethyl-50%- henylmethylpolysiloxane (OV-225) and 50% cyanopropyl-50% phenylmethylpolysiloxane (Silar-5CP) [46,47].

Subsequently, 613 molecular descriptors (MDs) were computed using the Dragon software [48], of which 439 were describing the solute and 174 the stationary phases, with no preliminary selection of the MD. The used MDs belonged to several sub-blocks such as: Constitutional (molecular composition of a molecule), Topological (graph representation of a molecule), Connectivity (arrangement of the atoms in the molecule), Geometrical (molecular geometry) and etc. [49,50]. The QSPR model was trained with Partial Least Square (PLS) regression and was validated with "leave-one-out" 10-fold cross-validation. Eventually, Covariance Selection was used to investigate which descriptors contributing mostly to the model and to filter the least important ones. This feature selection simplifies the system and the interpretation but not the prediction. Any model is validated via Root Mean Square Error (RMSE), known as the standard deviation error in calculation, which can be calculated via the following equation [51]:

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \quad \text{.....(9)}$$

Where, *n* is the number of objects (90 esters in this case), y_i is the observed dependent variable (experimentally collected RIs in this case) and $\hat{y}_i$ is the calculated dependent variable (RIs calculated via the PLS model in this case).

In the work of Biancolillo & D'Archivio [2] the validated results were reported as Root Mean Square Error in Cross Validation (RMSECV) and Root Mean Square Error in Prediction (RMSEP). These should always be reported, since for any prediction model, we need two different data sets; one for training the model (cross validation) and one for testing the prediction (test prediction). RMSECV and RMSEP are also calculated using equation 9, the only difference is the y_i and $\hat{y}_i$ data, for RMSECV is used from the cross validation data set and for RMSEP is used from the test prediction data set. RMSECV and RMSEP of the different stationary phases are presented in table 2. The values stand for the standard deviation error of RIs for cross validation and for test prediction, hence they have retention index units:

Table 2: Model precision of the different stationary phases [2].

Column	RMSECV [RI units]	RMSEP [RI units]
OV-225	20.6	24.4
OV-25	15.5	23.3
OV-7	15.1	17.0
SE-30	15.4	26.6
Silar-5CP	14.9	24.0
XE-60	13.4	17.3
DC-710	18.2	11.2

D'Archivio et al. [3] in an earlier work also showed the use of a QSPR model for the same 90 saturated esters on columns of different polarities (SE-30, OV-7, DC-710, OV-25, XE-60, OV-225 and Silar-5CP), but this time the model was trained via multi-linear regression (MLR) and artificial neural network regression (ANN). The used descriptors consisted of constitutional and topological descriptors, walk and path counts (which is the number of walks of any length in the graph representation), information indices (which take into account the indices of neighborhood symmetry) and connectivity indices (arrangement of the atoms in the molecule) [49,52]. To validate the model leave-one-out cross-validation was used. After comparison between the two approaches, MLR and ANN provide similar predictive performance for similar stationary phases. However, when compared between different polarity columns, ANN prediction becomes better than that MLR. This is especially noticeable when dissimilarity between stationary phase composition grows. This could be explained due to the fact that ANN data treatment overcomes problems related with collinearity of column descriptors which cannot be overcome by MLR. Table 3 summarizes the standard error of calibration (SEC) and standard error of prediction (SEP) of MLR and ANN [3]:

Table 3: Model precision presented with related standard errors (SEC and SEP) in RI units of the different stationary phases with different regression models [3].

MLR-based model			ANN-based model		
Column	SEC	SEP	Column	SEC	SEP
SE-30	13.9	23.3	SE-30	8.3	9.6
OV-7	15.3	13.3	OV-7	8.1	8.3
DC-710	15.2	12.9	DC-710	8.3	7.0
OV-25	15.2	12.9	OV-25	7.5	11.0
XE-60	15.4	12.0	XE-60	7.5	11.6
OV-225	15.3	12.8	OV-225	7.9	9.1
Silar-5CP	12.9	25.4	Silar-5CP	7.8	12.2

The QSPR model was also used in the work of Katritzky et al [53], to predict RIs of 178 methylalkanes (mono-, di-, tri-, tetramethylalkanes) that are produced by insects. The RIs were measured on a non-polar DB1 column with a temperature program of 60 - 320°C. The chosen descriptors belong to the sub-block of topological and geometrical descriptors, these descriptors were found to be very relevant in predicting the RIs of *iso*-alkanes, due to their high coding capability in representing the chemical structures effectively. As the test compounds are all belonging to the same compound class (isomeric alkanes) of same polarity, there was no need to introduce molecular descriptors that describe differences in their electronic structure or in their polarity (which would have been necessary if the test compounds belonged to different compound classes). CODESSA [54] was used to compute 302 descriptors, boiling point was also used as a physicochemical descriptor combined with other structure-based descriptors. The model was validated by leave-one-out cross validation using an external set of 30 methyl-branched alkanes. The predicted retention indexes on non-polar DB1 column had an average error of 4.6 RI units, correlation coefficient R^2 of 0.9585 and standard deviation of 5.8 RI units. The used regression was MLR with four-descriptor equation that might be the reason for the relatively poor correlation coefficient R^2 of 0.9585.

On the other hand the group of Matyushin et al. [55] has compared two predictive models for RI, a deep convolutional neural network model and a molecular descriptor approach with functional groups contributions, to experimentally measured RI data. The models were tested for almost 20,000 compounds containing essential oils, metabolites, flavors that were obtained from several databases (PubChem, NIST, Adams, Golm, Flavors) [56–60]. The RI values were recorded on standard non-polar and semi-standard non-polar columns. For the NN approach, the compounds were converted to SMILES and used as input data. The results of different approaches were compared in terms of percentage of correct compound identification (refer to table 4).

Table 4: Percentage of correct compound identification is reported for each used method [55]:

Used method for compound identification	Percentage of correct identification (F%)
Experimental	92.3
Calculated using functional groups contribution	83.9
Calculated using NN	86.4
Only MS	82.1

Only MS means that the compound was identified without RI system, but using only mass spectra data, which scores 82.1% of correct identification; the identification is correct only for 82.1% of compounds. Hence, F% suggests that using NN approach with RI prediction allows to increase compound correct identification rate from 82.1% (MS detection alone) to 86.4%. Furthermore the Root Mean Square Error (RMSE) was reported for NN to range from 66.9 – 144.8 [RI units] (for the different databases) and RMSE of MD with functional group contribution in the range of 101.1 – 282.0 [RI units] (for the different databases). However, the RMSE values are considered quite high when compared to other literature values (also used NN model), where the RI prediction error was within the range of ± 20 units [3]. A (small) width of this range is important to be useful for compound confirmation [2,3].

Furthermore, the NN approach quoted in this paper [55] does not encode information about stereochemistry and geometric isomerism, which are very important to distinguish different isomers, might have a negative consequence on the modelling ability.

Also a similar approach was found in the paper of Kireev, Osipenko, Mallard, Nikolaev and Kostyukevich [61], where they compare deep learning approaches (trained on a NIST RI database) to a non-learning method based on functional group increments. For this model 4397 different compounds were used from the Organisation for the Prohibition of Chemical Weapons chemical analysis database. Retention index values within the range of 488 – 3309 were retrieved for the DB5MS column, where the average RI values of the non-polar (DB1) and the average RI values of the semi-standard non-polar (DB5) column were used separately. The deep learning model was composed of 1D-CNN and 2D-CNN (Convolutional neural network), for which SMILES string was used as an input. The non-learning model was based on the theory that a RI difference should be preserved between two pairs of molecules that differ by the alkyl chain and between two pairs of molecules that differ by scaffold. However, for this method three molecules with known RI and relevant substituents should be given in advance. The models were validated using MAE mean absolute error in fitting (calculated on test set):

$$MAE = \frac{\sum_i |y_i - \hat{y}_i|}{n} \quad \text{.....(10)}$$

Where, n is the number of objects, y_i is the observed dependent variable (experimentally collected RIs) and $\hat{y}_i$ is the calculated dependent variable (RIs calculated via the model).

The results of 1D-CNN (deep learning approach) and increment-based (non-learning method) are summarized in Table 5.

Table 5: Performance of different approaches reported with MAE RI units [61]:

Test Set	1D-CNN	Increment-Based
Methyl - Phosphonofluoridates	35	4.0
Ethyl - Phosphonofluoridates	11	1.8
Propyl - Phosphonofluoridates	52	3.4

Although deep learning methods are very powerful and state of art, the results indicate that non-learning method significantly enhances the RI predictions specially when modeling molecular properties of structurally similar compounds (homologues and isomers).

A different study tried to predict Kováts retention indices of essential oils using molecular descriptors with two different regression methods; linear and non-linear, multiple linear regression (MLR) and support vector machine (SVM) respectively [62]. The dataset consisted of 340 essential oils obtained experimentally from the work of Babushok et al. [63]. The RIs were collected from DB1, DB5, and PEG columns. A total of 184 molecular descriptors were used that belonged to different blocks (topological, geometrical, constitutional, and hybrid descriptors). The model was validated only externally with a test set and not internally. The prediction performance is reported in table 6. Hence, the results indicate that SVM as a non-linear method has a better precision in predicting Kováts retention indices when compared to MLR.

Table 6: Precision of the different models; Root Mean Square Error (RMSE) used for the training set and Root Mean Square Error of Prediction (RMSEP) used for the test prediction set [62].

Model	RMSE [RI units]	RMSEP [RI units]
MLR	56.55	56.99
SVM	44.62	53.60

2.6. Chemometrics

The concept of chemometrics is based on the use of mathematical and statistical methods to obtain relevant information. The processing of enormous data which is generated by the analytical instruments requires the use of chemometric methods. The initial purpose of chemometrics is to convert complicated mathematical methods into simple practicable version to apply for particular applications, such as the optimization of chromatographic separations or prediction of various properties (boiling point, melting point, retention indices, etc.). One of the major applications of chemometrics is the development of QSAR and QSPR for analytical and chemical purposes [64]. Typical steps of any chemometric modelling comprise data input (acquisition &

sourcing), pre-processing of the data, method calculation by regression methods, method validation and finally checking the accuracy and robustness of the model. If it is found non-acceptable all the previous steps have to be optimized (figure 4).

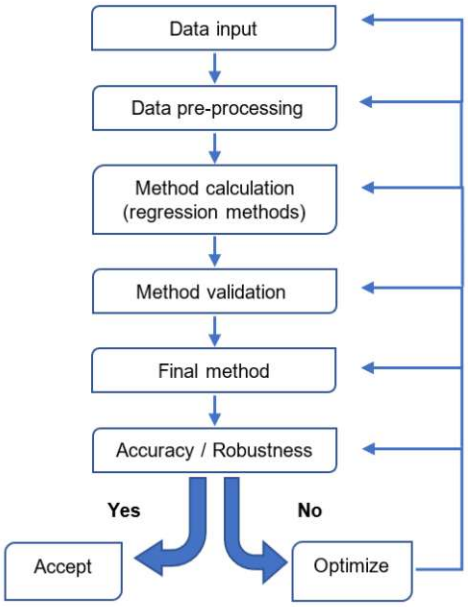


Figure 4: Steps of chemometric model building.

2.6.1. Pre-processing

Data pre-processing is a crucial step before any machine learning regression methods can be applied, because the algorithm of machine learning heavily depends on the input data needed or available to solve a particular problem [65]. In the case of QSPR modelling, several molecular descriptors are used, covering a wide range of numerical values (e.g. molecular weight descriptor can have a range of [16−400], polarity descriptor of [0−40] and connectivity index of [0−0.4]. When no scaling is implemented, descriptors with large numerical values will dominate the model, thus it will be difficult to determine the relative contribution of each descriptor to the QSPR model. Consequently, this compromises the statistical validity of the model. Therefore, to avoid this kind of problem, data pre-processing is recommended [43]. However, during pre-processing there is the danger of losing important information if inappropriate strategies are used. An example of a pre-processing used within this work is the Pareto scaling, which scales the data by dividing each variable by the square root of the standard deviation $\frac{x1}{\sqrt{\sigma}}$, so that each variable has variance equal to 1 (refer to figure 5). This ensures that datasets with both large and small(er) variability are brought to a comparable level of variability and thus the (large) variance of one parameter will not dominate the model. This kind of scaling does not change the original raw data drastically, but instead retains important data from being lost [64].

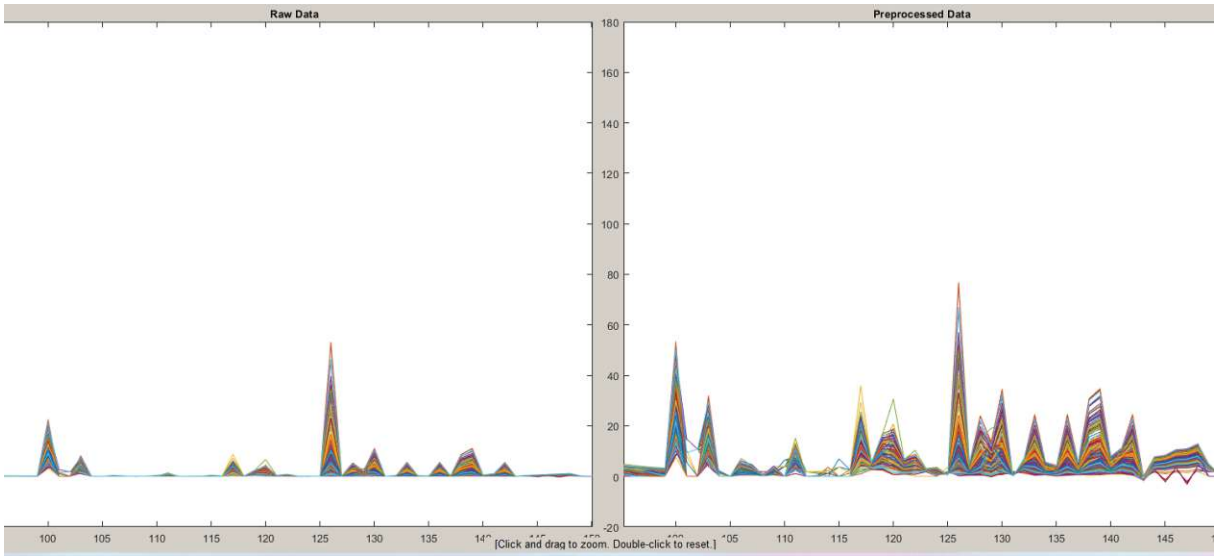


Figure 5: Raw data before and after pareto scaling [66].

2.6.2. Regression Methods

Regression methods are statistical methods used to estimate relationships between a dependent variable and one or more independent variables. They can be used for modelling a relationship between the dependent and independent variables. Regression analysis comprises linear, multiple linear, and non-linear methods. Most commonly used forms of regression models are the linear and multiple linear regression. However, some cases cannot be described by simple linear relationships. Therefore, non-linear regression is used for data in which the dependent and independent variables have a non-linear relationship. This section covers the following methods: Multiple Linear Regression (MLR), Partial Least Square regression (PLS) and Principal Component Regression (PCR) as linear regression, Locally Weighted Regression (LWR), Support Vector Machine regression (SVM), and Neural Network (NN) regression as non-linear regression [64,67].

Multiple Linear Regression (MLR): MLR assumes that there is a linear relationship between the dependent variable Yi and independent Xi variables. It uses Xi to predict Yi with the equations in the form of:

$$Yi = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad \ldots\ldots\ldots\ldots(11)$$

where a is the y-intercept, and b_1 to b_n are the regression coefficients of the first to the last independent variable, respectively. However, for this method to function, it requires that the number of Yi specimens be greater than the number of Xi predictors; also there must not be any significant correlation between the used Xi values. To overcome the limitations of MLR, PLS or PCR can be implemented instead [67,68].

Partial Least Squares (PLS): PLS is one of the primarily used chemometric tools, it can be used to predict linear and non-linear relationships between Xi and Yi. The Xi variables that show a high correlation with the Yi response are given a higher relative significance because they will have more effect in the prediction. Moreover, to predict

Yi variables, PLS reduces the dimensionality of Xi variables by identifying latent variables (LVs) which are orthogonal to each other [64,67]. This concept assists PLS to fully explain the maximum relationship between Xi and Yi. The following equation in figure 6 is used for PLS, where $\boldsymbol{Y}$ is the matrix of dependent variables, $\boldsymbol{X}$ is the matrix of independent variables, $\boldsymbol{B}$ is the matrix of regression parameters, and residuals are the differences between measured and predicted Yi data:

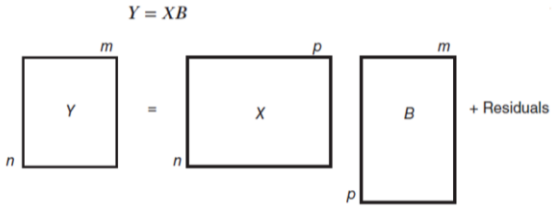


Figure 6: PLS matrix equation [67].

Although PLS is very suitable to model multiple outcome variables (when there is multicollinearity among Xi), PLS has limitations when the data set contains strong non-linear relationships, causing the need to switch to advanced non-linear regression methods.

Principal Component Regression (PCR): PCR regression is very similar to PLS, in the sense of reducing the dimensionality of the variables used in the model. Instead of using the variables directly, it transforms them into principal components (PC) with smaller dimension and this transformation is shown in figure 7. PCR is more sensitive to a systematic error in the predictor values Xi than in the response Yi. The principal components are chosen in a way that they describe as much of the variation in Xi as possible. One major advantage of PCR is that it overcomes the problem of multicollinearity. Hence, when there is high correlation between the predictor variables, MLR fails to deliver a reliable prediction; consequently PCR replaces MLR [67,69,70].

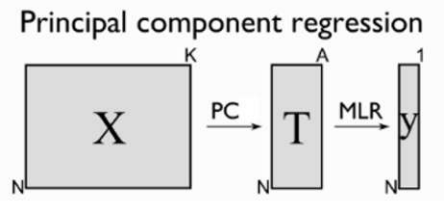


Figure 7: PCR dimension reduction, where a raw data matrix *X* is reduced to a smaller data matrix *T* [70].

Locally Weighted Regression (LWR): LWR is a non-linear memory-based regression, meaning it stores the training data and uses it every time a new prediction is made.

LWR is a so-called local model due to the fact that It chooses a test point x from the training data, and gives more focus (weightage) to the points which are near to x (refer to figure 8). Because, the points which are close to x are a good way to estimate the value of the point to be predicted [71]. LWR approximates a non-linear response by a

linear function on a small (local) scale [72], the points are weighted by proximity to the x and a regression is then computed using the weighted points.

LWR model is assisted with PLS or PCR, it copes well with noise and corrupted data, and is usually used when PLS fails to predict extreme non-linear relationships between dependent and independent variables.

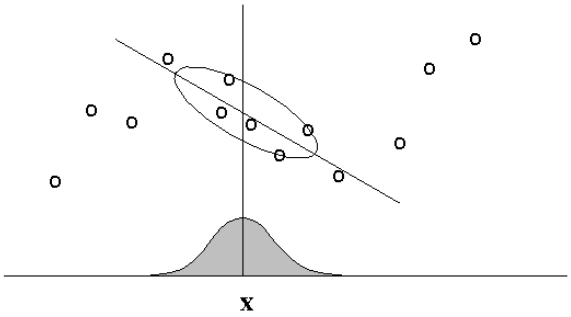


Figure 8: Graphic representation of how LWR works by focusing only on the points in close proximity to x and uses these to compute the model [72].

Support Vector Machine regression (SVM): SVM is another form of a non-linear regression it is a very powerful technique to predict non-linear responses and is considered to be a hybrid of MLR and LWR. SVM, which is basically a machine learning method, was first developed by Vapnik [73]. The model comprises of a number of support vectors (which are the data set) and non-linear model coefficients. The regression can be calculated by two different functions: Linear kernel or radial basis function (RBF) kernel. Kernel is a mathematical function that returns the inner product between two points in a space and is called as the “kernel trick”, it is used to classify different data. Kernel functions are very common functions in machine learning for data analysis.

Linear kernel:

The linear SVM, classifies different data by using maximum margin classifiers that construct decision surfaces called hyper-planes, and by maximizing the margin between two classes it supports the hyper-planes. Figure 9 shows a linear kernel SVM with the decision surface (maximum margin), support vectors, hyper-planes and margin (the distance between hyperplane and support vectors) [74].

Radial basis function (RBF) kernel:

The easiest way to group similar data is with a straight line, however sometimes it is impossible to do so. So when the data that is not linearly separable, the radial basis function (RBF) kernel comes into play. By using the kernel ‘trick’ it projects the data on a new dimension (transforms the data into 3D or higher dimensional space) where the data set can be separated linearly via a linear plane [73] (shown in figure 10).

While the complicated functions behind SVM can be calculated mathematically, it can still be considered as a black-box method, meaning there is a lack of access to the internal workings and parameters of functions, which makes it difficult to interpret the

results. Following are a representation of the different functions for linear and RBF kernels which are the most used [73–75].

$$Linear\ kernel: K(x_1, x_2) = x_1 . x_2 \ \ldots\ldots\ldots\ldots (12)$$

$$RBF\ kernel: K(x_1, x_2) = exp(\frac{-||x_1 - x_2||^2}{2\sigma^2}) \ \ldots\ldots\ldots\ldots (13)$$

Where, $K(x_1, x_2)$ is the kernel function, x_1, x_2 are the support vectors, $||x_1 - x_2||$ is the Euclidean distance between x_1, x_2 and σ represents the kernel width.

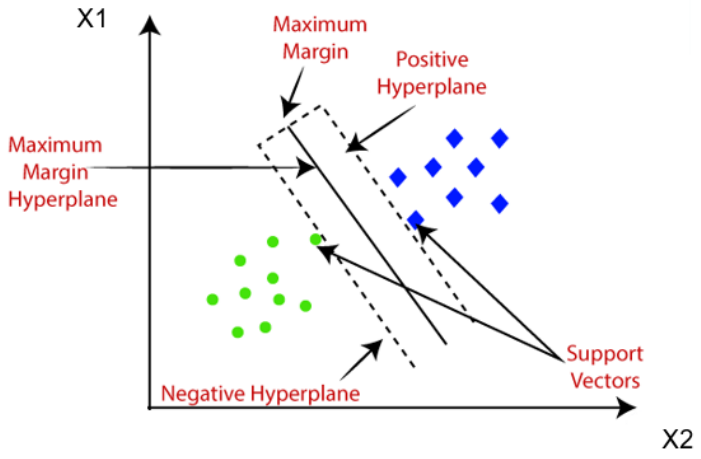


Figure 9: Schematic demonstration of linear SVM decision surface (maximum margin), support vectors and hyper-planes [76].

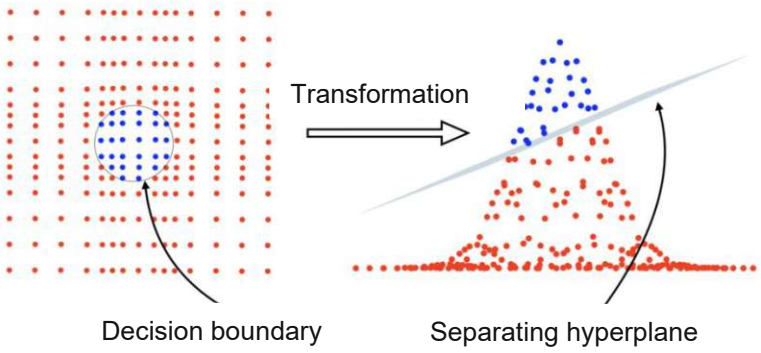


Figure 10: Schematic demonstration of RBF SVM transforming the data into higher dimensions [77].

Neural Network regression (NN): NN is also a non-linear regression and a black-box method. In fact, a NN is generated in a way to imitate the operations of brain neurons. Analogous to the brain neuron, neural networks have many layers of artificial neurons linked to each other. A typical NN consists of a first input layer that receives the input variables, a second hidden layer where the mathematical operations take place, and a third output layer (see figure 11 & 12) [64]. Neural Networks have gained significant interest especially in cases of pattern recognition. Unlike PLS and PCR, NNs do not deduce any initial mathematical relationship between the input and output data. Therefore, NNs are flexible tools in modelling complex relationships. However, this

model has its challenges, for example it doesn't encode information about stereochemistry and geometric isomerism and such subtle modifications of the input data could cause misleading outputs (e.g. small amount of noise would cause NN to misclassify the data) [67].

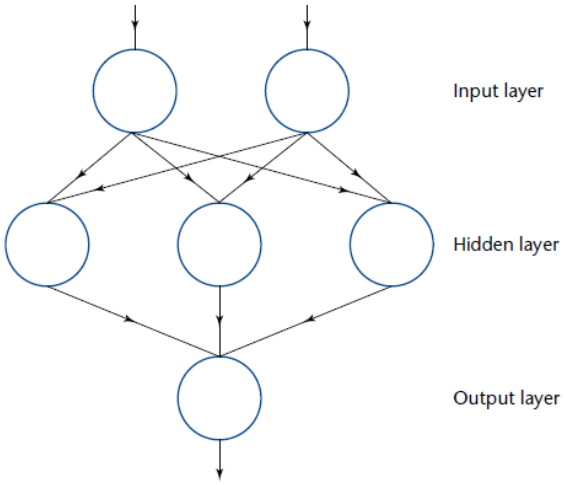


Figure 11: An example of a neural network [67].

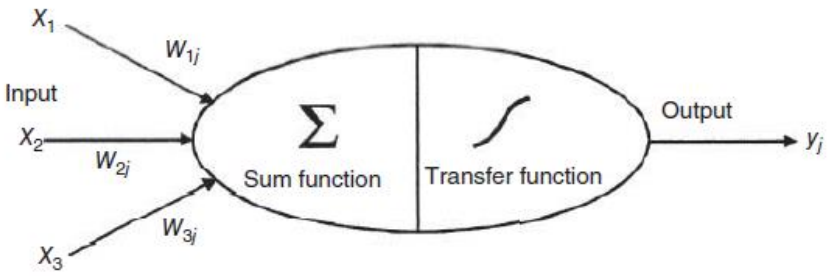


Figure 12: Example of neural operation [64].

2.6.3. Model validation

A requirement in any application field is that an analytical method is suitable for its intended purpose and this is statistically demonstrated. Validation is specially crucial in case of a new method whose viability needs to be tested. Also in case of a method transfer when transferring an analytical method from one laboratory to another, it will only be acceptable if it is properly validated. The same applies in chemometric modelling where the validation is critical to establish a reliable and robust predictive model. Most commonly used validation parameters in chemometric modelling, to assess the precision of the model and to evaluate it, are the following [64,67,78]:

1. Root mean square error (RMSE), which is the standard deviation of the prediction errors (residuals); $\mathrm{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$

2. Root mean square error of cross validation (RMSECV), which is obtained from the cross-validated data (internal validation); $\mathrm{RMSECV} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$

3. Root mean square error of prediction (RMSEP), which is obtained from using an external test data set (external validation); $\mathrm{RMSEP} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$

Where, n is the number of objects, y_i is the observed dependent variable (experimentally collected RIs in this case) and $\hat{y}_i$ is the calculated dependent variable (RIs calculated via the PLS model in this case).

Cross-validation: Is an internal validation of a model which measures the predictive accuracy by removing each time a random subset from the dataset, then constructing the model using the remaining objects in the dataset, subsequently applying the resulting model to the removed objects. This way, the model is tested with objects that were not used to build the model. There are usually four different cross-validation methods (figure 13), varying with respect to how the different objects are selected from the dataset [79]:

1. Venetian Blinds: Removes every n^{th} object from the data set to revalidate
2. Contiguous Blocks: Removes a block of different objects from the data set to revalidate
3. Random Subsets: Removes the objects randomly from the data set to revalidate
4. Leave-One-Out (LOO): Removes each single object at a time from the data set to revalidate

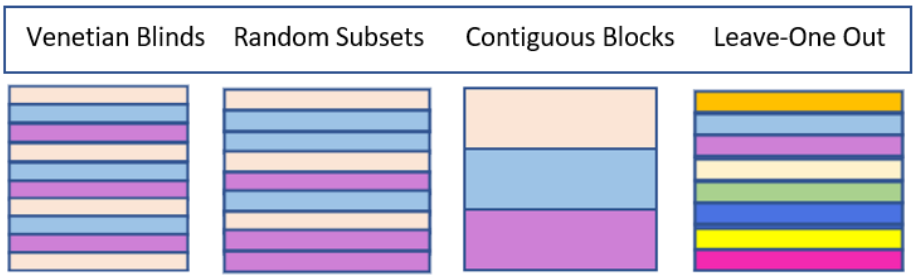


Figure 13: Illustration of different forms of cross-validation where colors signify the different objects within the dataset; in Venetian Blind the colors signify that every 3rd object is selected for the validation; in Random Subset the objects are selected in random order; in Contiguous Block the data is selected block-wise; and in LOO every single object is selected at a time [79].

External validation: Is usually used when a large dataset is available. For this purpose the initial dataset should be split into a training set (this is used to train the model) and a test set which is not included in the training set. Consequently, the test set will be used to externally validate the model.

Moreover, it must be mentioned that there is no exact criterion to obtain the best chemometric model. Hence, to obtain the optimum predictive performance of any model, all the steps of a model building from figure 4 have to be re-optimized until reaching a satisfactory result. Also outliers should be taken into consideration, since they are objects that cannot be described / explained by the model and therefore must be eliminated. This is done by testing the cross validation residuals against the input Y (RI) variables (figure 14).

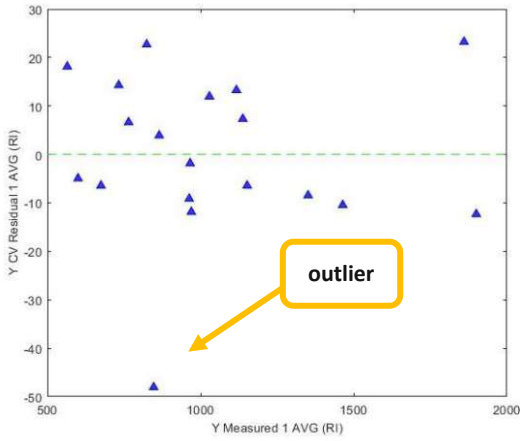


Figure 14: Example of an outlier.

3. Experimental Part

3.1. Dataset

The whole dataset used in this work consists of almost 400 compounds from different classes (alkanes, alkenes, cycloalkanes, aromatics, alcohols, acids, aldehydes, ketones and esters) and with different isomers. All of the retention indices (RIs) have been collected from the PubChem database [60], for standard non-polar (DB1), semi-standard non-polar (DB5) and standard polar (wax / PEG) columns. In the case of multiple data for the same compound, the values were averaged. Furthermore, the data was statistically evaluated based on relative standard deviation (RSD), minimum and maximum, checked for consistency and cleansed, if necessary. Furthermore, RIs of esters on different columns OV-7, DC-710, OV-25, XE-60, OV-225, Silar-5CP were collected from the work of Biancolillo and D'Archivio [2] and were used to evaluate the performance of the model on the mentioned different columns (table 7 summarizes the composition of different columns used in this work). Moreover, the data of 400 compounds was split into ~75% training set and ~25% test set (refer to supplementary materials), training set was used to train the model and the test set was used to externally validate the model.

Table 7: List of used columns with their composition.

Column type	Stationary phase	Polarity [80]
DB1	100% poyldimethylsiloxane	Non-polar
DB5	5%-polydiphenyl-/95%-polydimethylsiloxane	Non-polar
OV-7	20% polydiphenyl-/80% Polydimethylsiloxane	Slightly-polar
DC-710	50% polydiphenyl-/50%polydimethylsiloxane	Mid-polar
OV-25	50%-polydiphenyl-/50%-polydimethylsiloxane	Mid-polar
XE-60	50% Cyanopropylmethyl/50% Dimethylpolysiloxane	Mid-polar
OV-225	50% cyanopropylmethyl-/50%- phenylmethylpolysiloxane	Mid-polar
Silar-5CP	50%-Cyanopropyl-/50% Phenylmethylpolysiloxane	Highly-polar
PEG	Polyethylene glycol	Highly-polar

3.2. Molecular descriptors

Furthermore, 266 different molecular descriptors (MD) were obtained from the online chemical database (free online software), which are a set of real numbers encoding information about the compound under study such as molecular weight, hydrogen bond donor/acceptor, double bonds, hybridization, cyclic or linear system, polarizability, etc. and these information can be linked to experimental values of a molecule [81]. MDs can encode information of a molecule on different representation levels such as 1D, 2D, 3D (see figure 15).

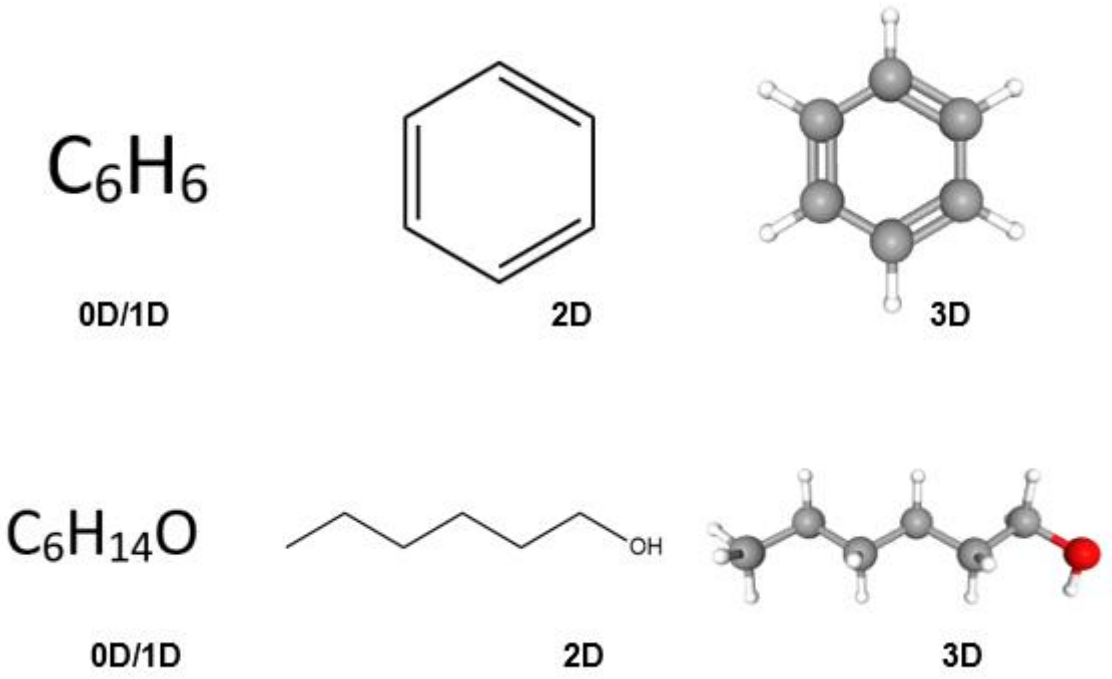


Figure 15:1D, 2D and 3D representations of benzene and hexanol molecules.

The used descriptors within this thesis belong to several sub-blocks mentioned in table 8 and for further information refer to the supplementary materials. These descriptors were specifically chosen based on literature research [2,3,53,62,82–84].

Table 8: Main sub-blocks of used molecular descriptors [49,50,52].

MD sub-blocks	Encodes the following information
Constitutional indices	Molecular weight, number of atoms etc.
Topological indices	Molecular structure and connectivity
ETA indices	Electronegativity of an atom
Functional group counts	Which and how many functional groups are present in a molecule
Charge descriptors	Partial charges of an atom
Geometrical descriptors	Spatial coordinates of atoms in a molecule
Walk and path counts	Where the molecule starts and ends
2D matrix-based descriptors	Degree of branching, the neighboring atoms in terms of electronic & steric effects, flexibility
Ring descriptors	Aromatic ratio and number of rings in a molecule
P_VSA-like descriptors	Van der Waals surface area
3D matrix-based descriptors	Surface properties in contact with a solvent or stationary phase
Edge adjacency indices	Connectivity as in how the edges of a molecule are connected
RDF descriptors	Average distribution of atoms around any given atom within the molecule (it gives the coordination number of a molecule)
WHIM descriptors	It gives the 3D (x,y,z)-atomic coordinates of a molecule
Getaway descriptors	Entropy of a molecule
Burden eigenvalues	Searches for chemical similarities between molecules

3.3. Model calculation and validation

All calculations were done in MATLAB (version R2019b). For each column type (DB1, DB5 and PEG) the model was generated separately. Firstly, for the model calculation, training set was used containing same compounds for DB1, DB5 and PEG with same Xi (MD) data but with different Yi (RI) values for each column (DB1, DB5 and PEG) respectively. Secondly, suitable regression analysis was chosen (PLS or SVM). Thirdly, Xi and Yi data were pre-processed by Pareto Scaling so that each variable has a variance of one and none would dominate the model. Fourthly, The generated model was internally validated using Venetian Blinds (10 splits and 1 sample per split) to predict model performance (refer to figure 16). And finally, the model was calculated and the results were evaluated by Root mean square error of cross validation (RMSECV).

Once the generated model is accepted, the model can be now externally validated using an external test set which is not included in the training set. The predictive ability of the model can be determined by the two metrics: Root mean square error of cross validation (RMSECV) obtained from internal validation and root mean square error of prediction (RMSEP) obtained from external validation.

The exact same steps were repeated to generate the model with different regressions: partial least square (PLS) regression assuming a linear relationship between descriptors and retention indices. Locally weighted regression (LWR) and support vector machine (SVM) regression assuming a non-linear relationship between descriptors and retention indices.

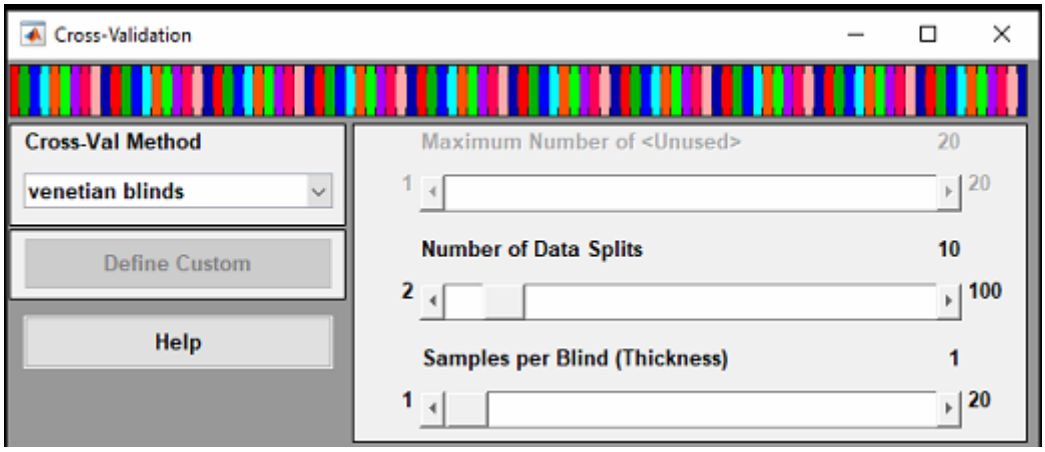


Figure 16: Demonstration of MATLAB Venetian Blinds cross validation, where the number of data splits means every 10th object is removed from the set and revalidated, and thickness means one single object at each time (Different colors explain the data split and the thickness).

3.4. Data Quality

Before plotting the data input in our models, we analyzed the quality of the Pubchem data and we noticed that the reported RIs on Pubchem are from different sources and belong to two different RI systems, Lee retention index and Kováts retention index. An example of this problem when using the complete dataset is demonstrated in figures 17 and 18.

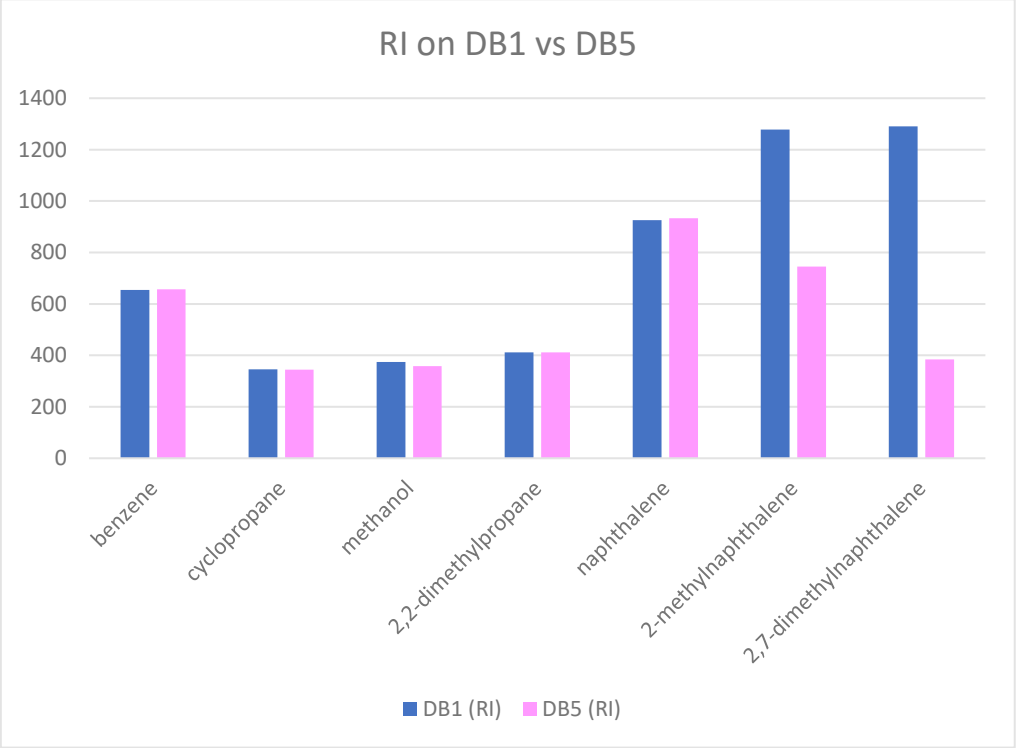


Figure 17: Pubchem averaged RIs of different compounds eluting on DB1 versus on DB5 columns [60].

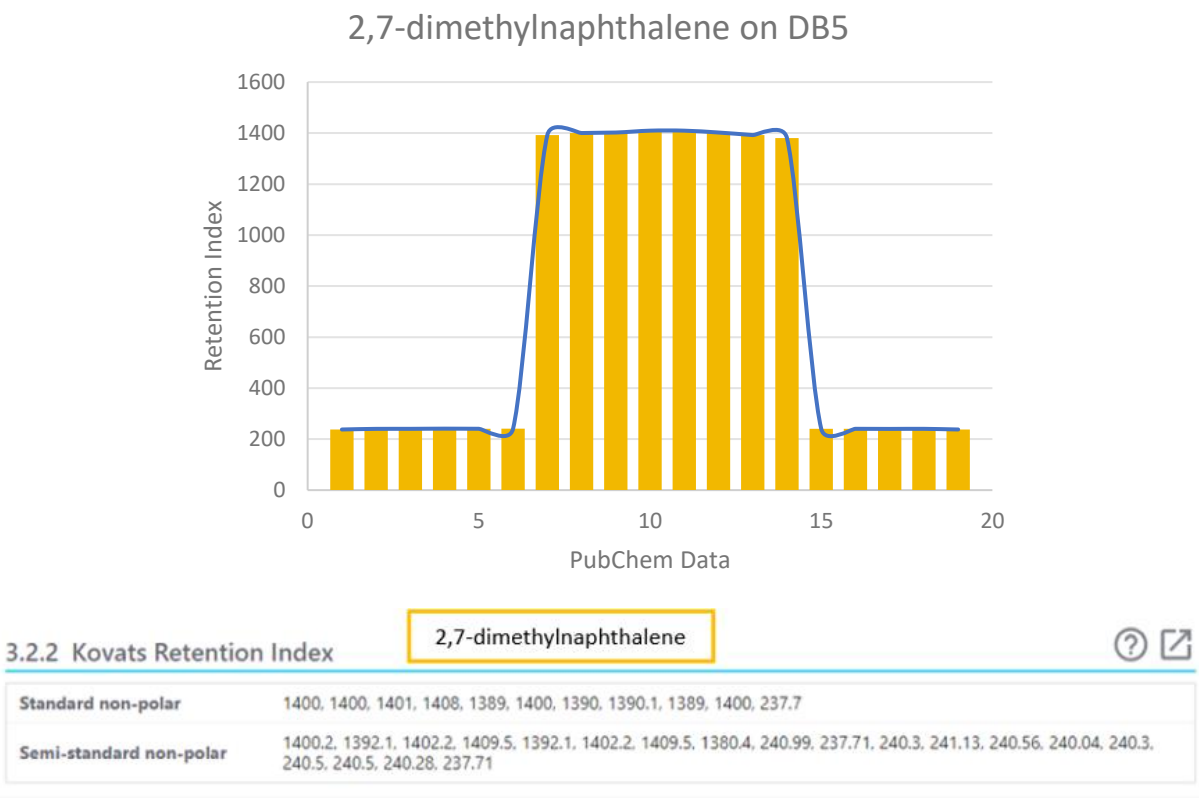


Figure 18: Differently reported RIs of 2,7-dimethylnaphthalene eluting on DB5 (semi-standard non-polar) column according to Pubchem [60].

As DB1 and DB5 columns are not very different in terms of composition, therefore the RI values of any compound should be comparable on both columns. However, we see a big difference between RIs of 2-methylnaphthalene and 2,7-dimethylnaphthalene on DB1 compared to DB5. But knowing that Naphthalene has 10 carbon atoms, it means that the Kováts RI should at least have the value of 1000 RI units, thus the average values of RIs on DB5 of 745 (for 2-methylnaphthalene) and 385 (for 2.4-dimethylnaphthalene) are not reliable. These lower values can only be explained by the (inappropriate) averaging of RI values from two different RI systems, namely the Lee RI system (which is defined in a different way) and the Kováts RI systen, thus producing so erratic numeric values [85].

For demonstration purposes we report in table 12 the mean, standard deviation, RSD, maximum and minimum values of the reported RIs from Pubchem of some compounds. To mention, table 12 does not show the values for all compounds, but demonstrates only fraction of the analyzed data (for further information lookup the supplementary materials).

Table 12: Statistical tests on RIs from Pubchem mainly on DB5 (RSD = relative standard deviation, Max = maximum, Min = minimum).

Compound	Mean	Standard deviation	RSD	Max	Min	RI (Pubchem)
ethane	200	0	0%	200	200	200
octadecane	800	707	89%	1800	297	1800, …, 297
cyclopropane	353	15	4%	367	331	331, 367, 349
methyl 2-methylpentanoate	836	32	4%	867	804	804, 867
1,2-dimethylnaphthalene	729	593	81%	1462	236	1451, 250, 249
2-methylnaphthalene	745	537	72%	1318	216	1318, …, 216
2,7-dimethylnaphthalene	385	572	149%	1410	238	1410, …, 238

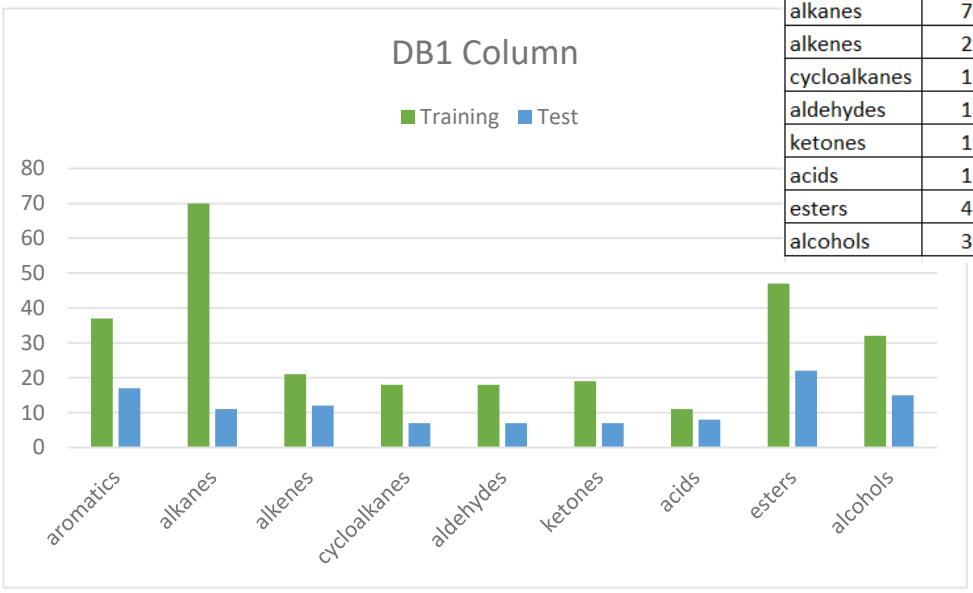
The RSD value is a measure that indicates the presence of possible errors or outliers within the data, and to accept any data it should have low RSD values. An example of extremely high RSD value is noticeable for naphthalenes (aromatics) and octadecane. Hence, the reason for this observation is the presence of RI values from different RI systems. Higher values belong to the Kováts system whereas lower values belong to the Lee system. So by eliminating the lower values (Lee RIs) we were able to obtain RSD values between 0 – 1 %. However, in the example of Cyclopropane and Methyl-2-methylpentanoate we see an error of 4%, even though there are only few data available but the RSD is high. Since we could not determine which RI should be accepted or rejected. Therefore, we included these data in our input despite having RSD value of 4%.

On the other hand, there is also the possibility that the RI data could sometimes be misassigned and DB1 values could be confused for DB5 values. It cannot be excluded that this error in the Pubchem data could be a source of prediction errors in our simulated models.

4. Results and discussion

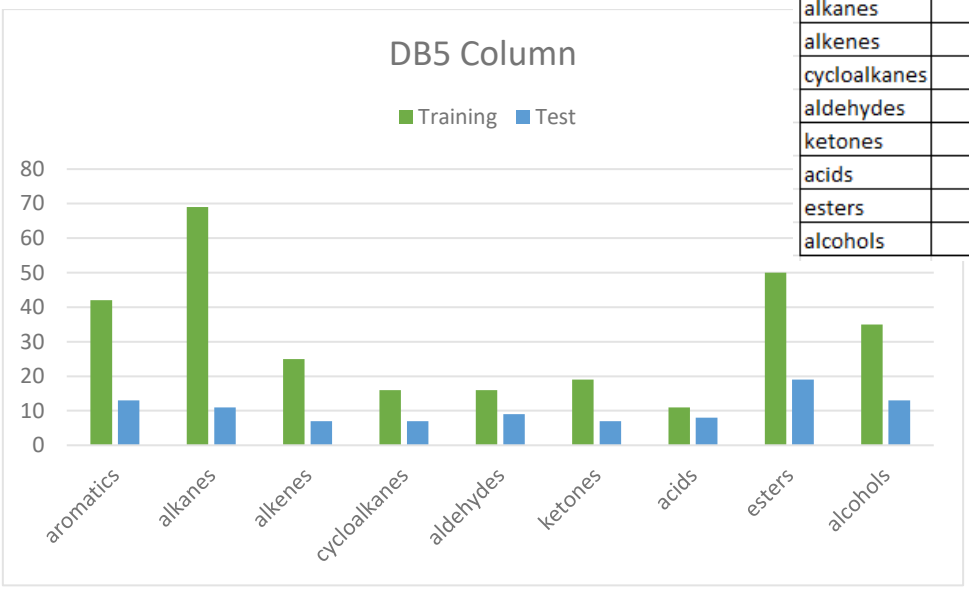
For any predictive modelling task, it is recommended to (initially) assume that there is a linear relationship between the dataset and to start the modelling with the simplest regression form, such as multiple linear regression (MLR). However, when MLR fails due to the limitations (refer to section 1.5.2 Regression methods), partial least square (PLS) can be used to calculate the model, and if PLS also fails to predict a linear relationship between the dataset, one has to shift to using a non-linear regression such as locally weighted regression (LWR) or support vector machine (SVM).

The retention indices (RIs) of different compounds have been collected on DB1, DB5 and PEG columns from Pubchem [60]. The data distribution among the different compound classes within training and test sets for each column is demonstrated in figures 19 – 21 and RI frequency within training and test sets for each column is demonstrated in figure 22. From following figures we can notice that the data is not homogenously distributed between different compound classes and also not all compounds have RI data on all three columns in Pubchem [60].



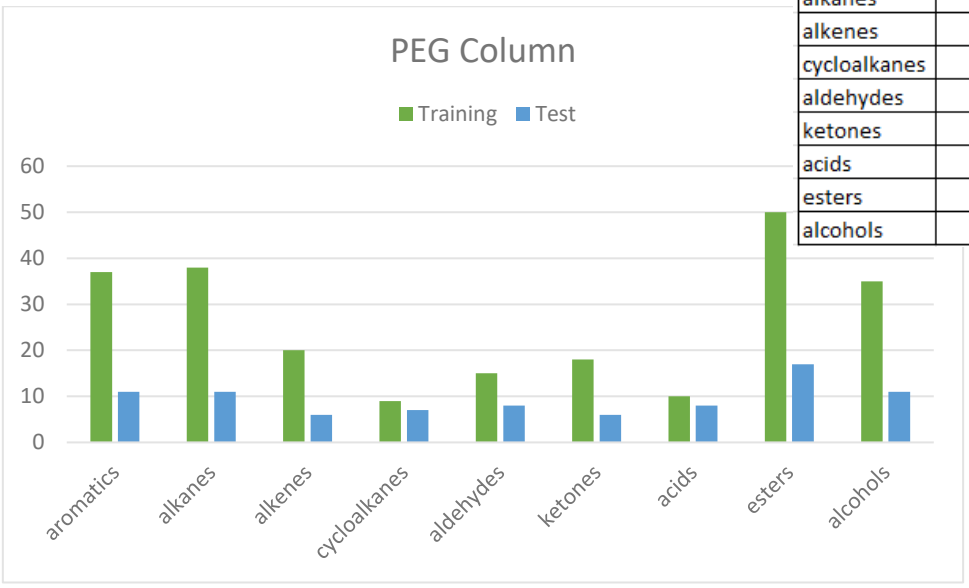
Class	Training	Test	%
aromatics	37	17	46
alkanes	70	11	16
alkenes	21	12	57
cycloalkanes	18	7	39
aldehydes	18	7	39
ketones	19	7	37
acids	11	8	73
esters	47	22	47
alcohols	32	15	47

Figure 19: DB1 data distribution among the different compound classes within training and test sets, and % in the table stands for the percentage of test set compared to training set.



Class	Training	Test	%
aromatics	42	13	31
alkanes	69	11	16
alkenes	25	7	28
cycloalkanes	16	7	44
aldehydes	16	9	56
ketones	19	7	37
acids	11	8	73
esters	50	19	38
alcohols	35	13	37

Figure 20: DB5 data distribution among the different compound classes within training and test sets, and % in the table stands for the percentage of test set compared to training set.



Class	Training	Test	%
aromatics	37	11	30
alkanes	38	11	29
alkenes	20	6	30
cycloalkanes	9	7	78
aldehydes	15	8	53
ketones	18	6	33
acids	10	8	80
esters	50	17	34
alcohols	35	11	31

Figure 21: PEG data distribution among the different compound classes within training and test sets, and % in the table stands for the percentage of test set compared to training set.

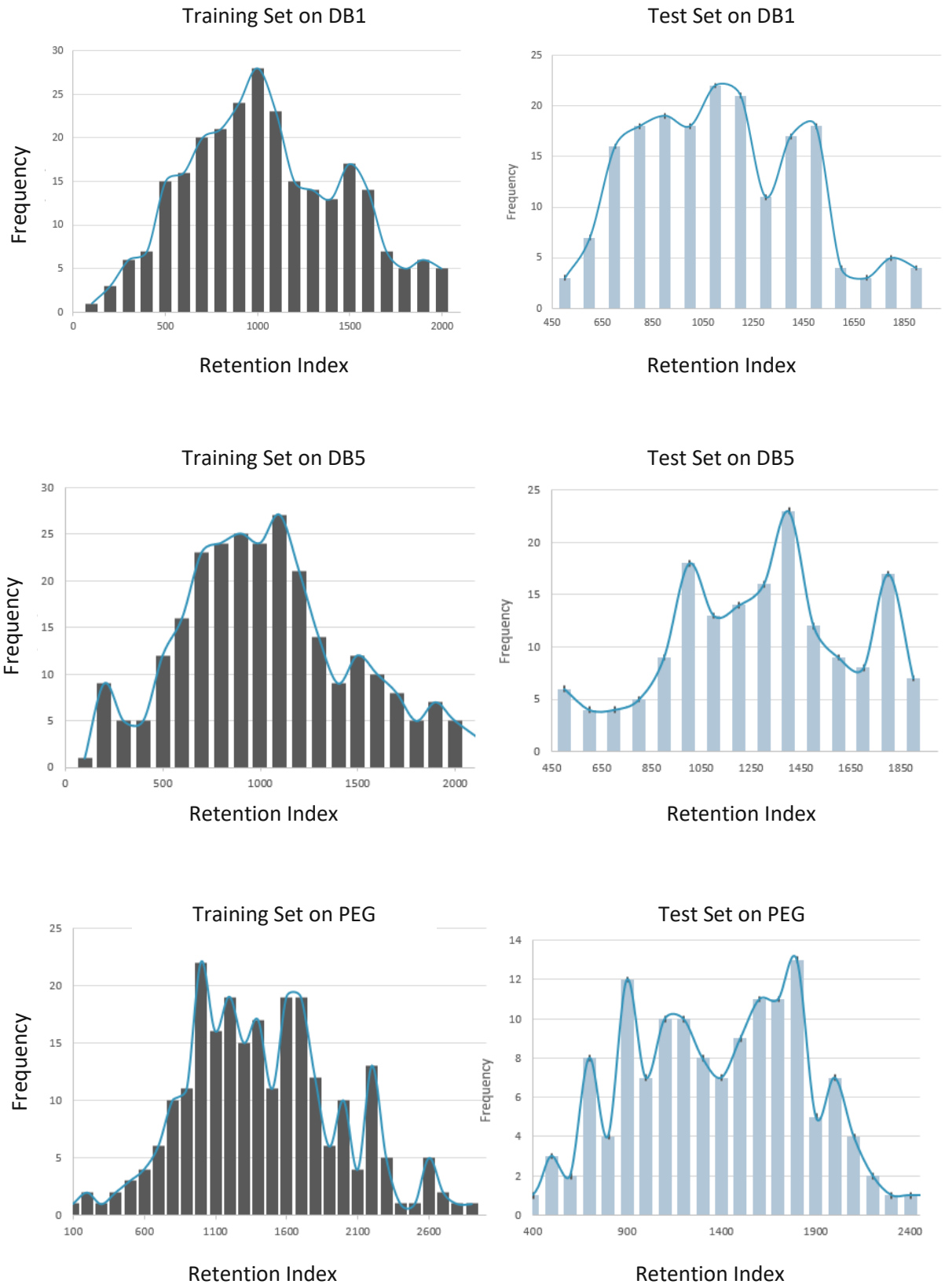


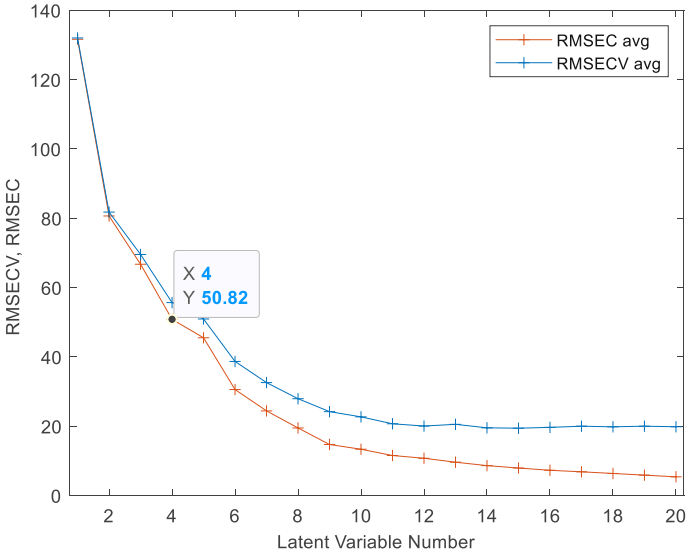
Figure 22: RI Distribution within training and test sets.

4.1. PLS regression for DB1

Since our dataset includes high numbers of molecular descriptors (refer to supplementary materials), so we decided to start the modelling task with PLS regression. The interest for modelling RI on DB1 column, is due to the fact that DB1 column is a non-polar column and has therefore the tendency to separate analytes based on their volatility. Hence, RI predicted on DB1 column has high correlation with boiling point of a compound (proportional to boiling point) and could eventually be used to acquire the boiling point of compounds.

Firstly, retention indices (RIs) were used as Yi data, and 266 MDs were used as Xi variables in MATLAB. Nevertheless, it was necessary to use all of these MDs due to the fact that our dataset contains 9 different compound classes with different functionalities and isomers, hence eliminating any MD is critical and could affect the prediction abilities of the model.

The entire dataset was split into 75% training and a 25% test set. The data was pre-processed by Pareto Scaling so that each variable has a variance of one and none would dominate the model. The training set was used to train the model with PLS regression using 4 latent variables (LVs) which has the optimal root mean square error of calibration (figure 23). The model was cross-validated using Venetian Blinds (10 splits and 1 sample per split, refer to figure 16), the performance of the model was evaluated by RMSECV (root mean square error of cross validation). After training and evaluation of the model, the test set was used to assess the final performance of the model via RMSEP (root mean square error of prediction). The results of PLS regression for DB1 are reported in figures 24 – 26.



Number LVs: 4 Auto Select

Percent Variance Captured by Model (* = suggested)

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV avg	
1	89.57	89.57	98.80	98.80	131.99	
2	4.50	94.07	0.75	99.55	81.752	
3	2.95	97.02	0.14	99.69	69.525	
4	0.64	97.66	0.13	99.82	55.639	current*
5	1.35	99.02	0.04	99.86	50.914	
6	0.25	99.26	0.08	99.94	38.659	
7	0.16	99.42	0.02	99.96	32.559	
8	0.09	99.52	0.01	99.97	27.906	
9	0.05	99.56	0.01	99.99	24.176	
10	0.06	99.63	0.00	99.99	22.67	
11	0.04	99.67	0.00	99.99	20.678	
12	0.06	99.73	0.00	99.99	20.021	
13	0.03	99.76	0.00	99.99	20.51	
14	0.02	99.78	0.00	99.99	19.503	
15	0.02	99.80	0.00	100.00	19.397	
16	0.02	99.82	0.00	100.00	19.638	
17	0.01	99.83	0.00	100.00	19.975	
18	0.01	99.84	0.00	100.00	19.799	
19	0.01	99.85	0.00	100.00	19.974	
20	0.01	99.86	0.00	100.00	19.818	

Figure 23: Plot of RMSEC versus Latent variables (LV) in PLS regression.

Figure 23 shows a summary of which LV should be chosen, that could describe the model at best. In this example MATLAB automatically selects LVs of 4 (LV of 4 means the dimensionality of Xi variables are reduced to 4 variables).

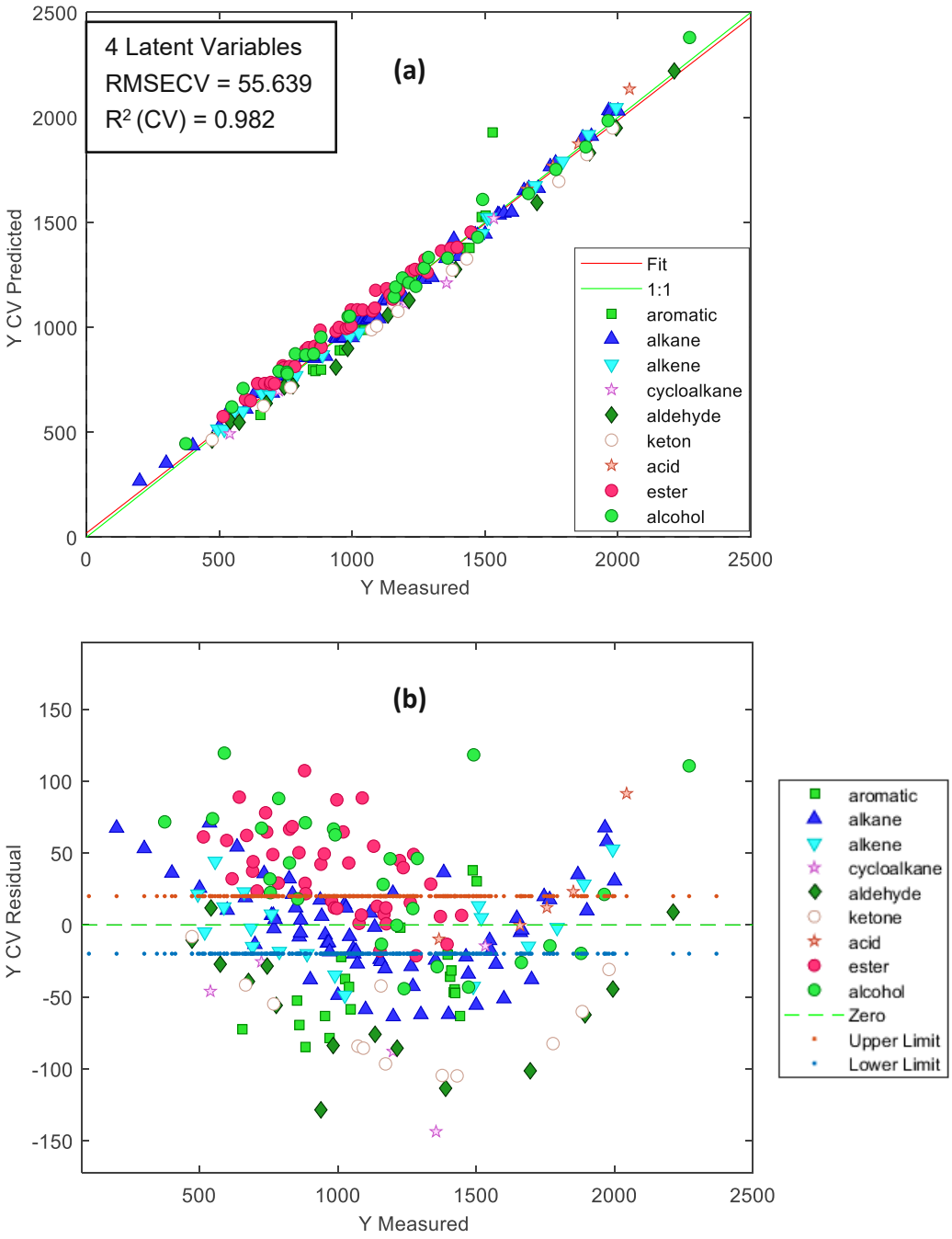


Figure 24: (a) PLS calibration plot with training set on DB1 column. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin (upper and lower error tolerance limits).

From figure 24 we can summarize that, in (a) the calibration error of the PLS model for DB1 is quite high with average RMSECV of 55.6 and correlation coefficient R^2 (CV) of 0.982. This is justified by referring to literature values of RIs, where it is states that predicted RI error can be acceptable within the range of ± 20 units [2,3]. It is even visible in (b) that the majority of the predicted RIs of different compound classes are outside the ± 20 units margin, where a cycloalkane has the highest error of -150 (meaning the predicted value is 150 RI units lower than the actual RI value).

Furthermore, the plots in figure 25 give further explanation of the PLS model. (a) shows Hotelling versus Q residuals plot which is a measure of the variation in each sample within the model, it indicates how far each sample is from the center (scores = 0) of the model. Hotelling score of 97.66% means that this much of the data variation is explained by PLS. The compounds within the threshold of 1 (closer to the center) are better explained than outside the threshold of 1 which are poorly understood/explained. And (b) shows that LV1 can explains 89.57% of the data, by showing the trend of increasing molecular weight of the components (starting with ethane ending with nonadecane – marked with purple color), whereas LV2 explains 4.50% of the data by distinguishing the different compound classes.

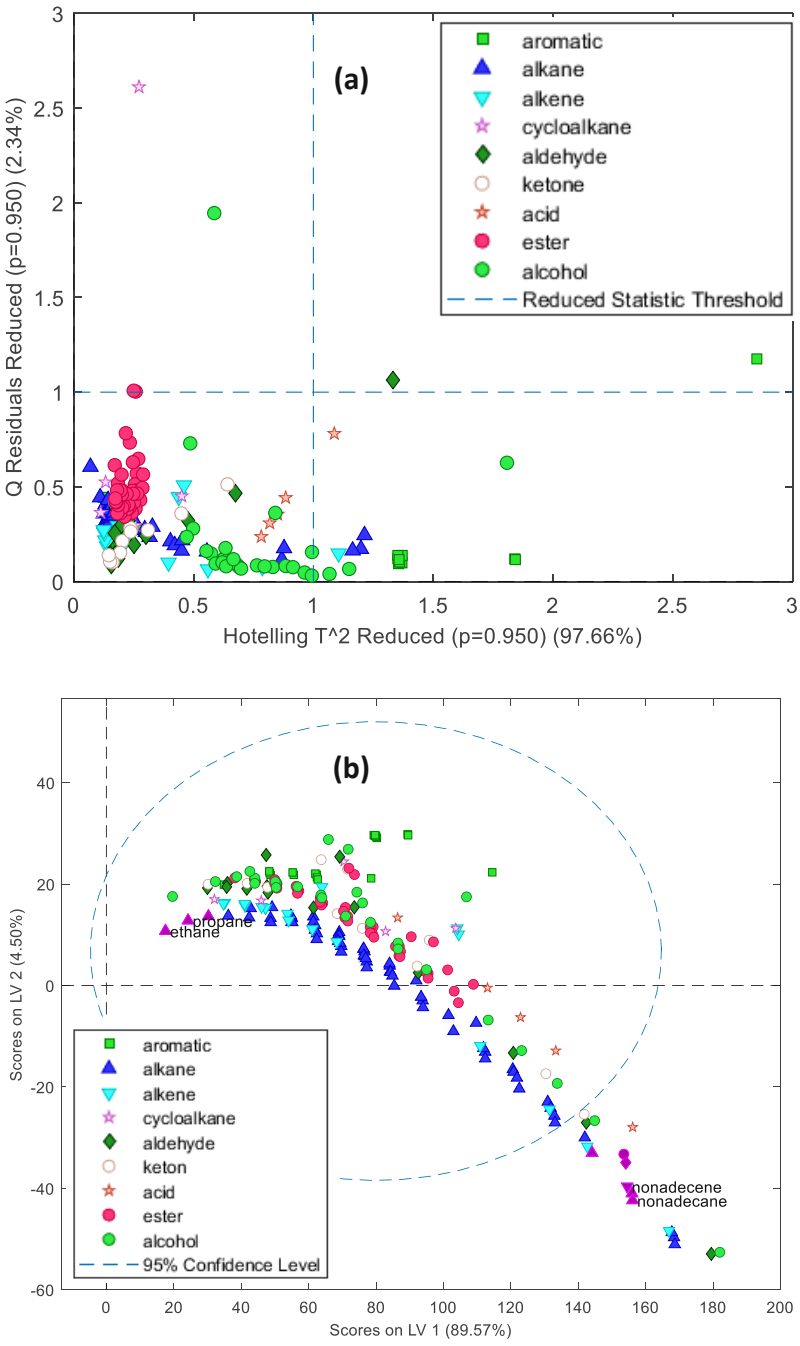


Figure 25: (a) Hotelling versus Q residuals plot. (b) Scores plot LV1 versus LV2, it indicates how much of the data is explained by the respective LV.

The trained model is then used to predict the test set, and the results are reported in figure 26.

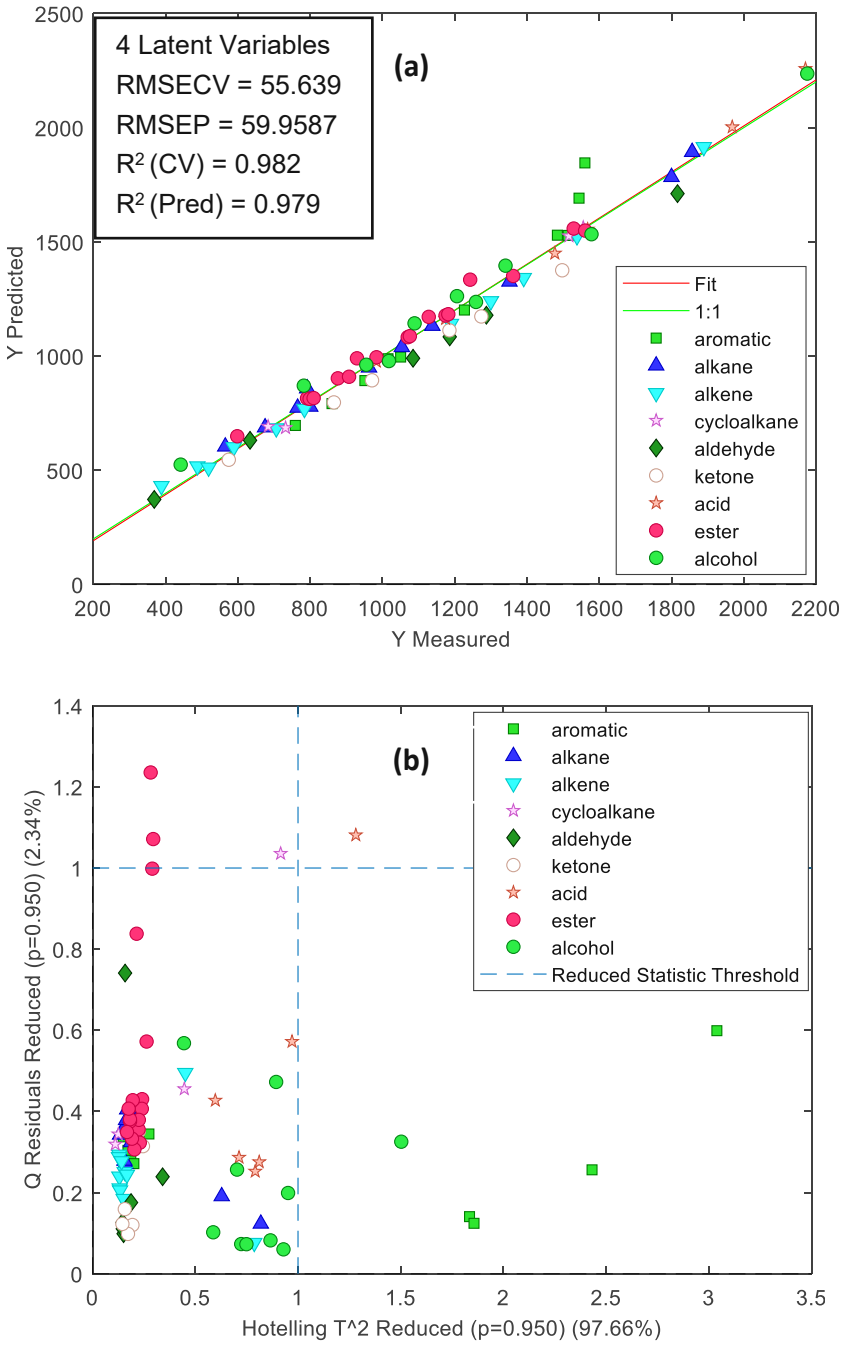


Figure 26: (a) RI prediction on DB1 with PLS using external test set. (b) Hoteling versus Q residuals plot represents a measure of the variation in each sample within the model.

The prediction of the PLS model for test set on the DB1 column has poor results with RMSEP of 60.0 and correlation coefficient R^2 (Pred) of 0.979 (refer to figure 26 (a)). And the Hoteling versus Q residuals plot in (b) shows that some aromatics, esters and cycloalkanes are poorly explained by the model.

From these results we concluded that PLS as a linear regression method is not reliable for predicting RIs on DB1 column, RMSECV and RMSEP having high prediction error of around 55.6 and 60.0 respectively, which is too high when compared to literature values that suggest RI error should be within ± 20 RI units, in order to be useful for

compound prediction or confirmation [2,3]. This could mean that our dataset is not satisfactory modelled by PLS model. Hence, the PLS model was not further tested on DB5 and PEG columns since it failed to deliver satisfactory results for DB1 column.

With these findings we were motivated to explore predictive models with a non-linear regression method such as locally weighted regression (LWR) and support vector machine (SVM) in order to predict non-linear relationships between our retention indices (RIs) and molecular descriptors (MDs).

4.2. LWR regression for DB1

The dataset is treated exactly as described in section 4.1. The only difference is that, for the regression method, LWR was used with 5 principal components (same concept as LV, where Xi variables are reduced to 5 variables) and 5 local points (which are in close proximity to point of interest x). And the regression is then computed using these 5 weighted points. The model is validated in the same order as PLS and the results of calibration and prediction are reported in figures 27 and 28, respectively.

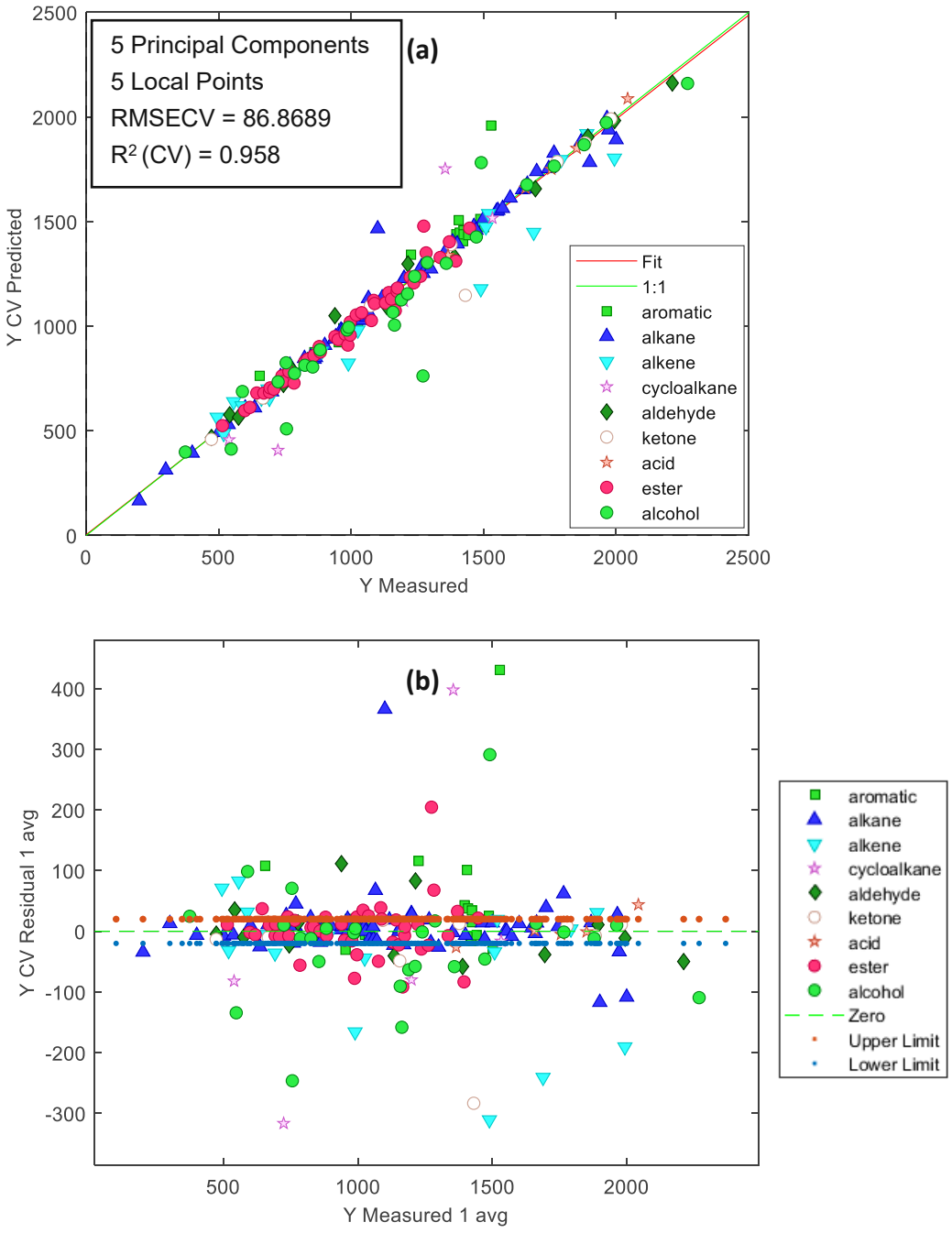


Figure 27: (a) LWR calibration plot with training set on DB1 column. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin (upper and lower error tolerance limits).

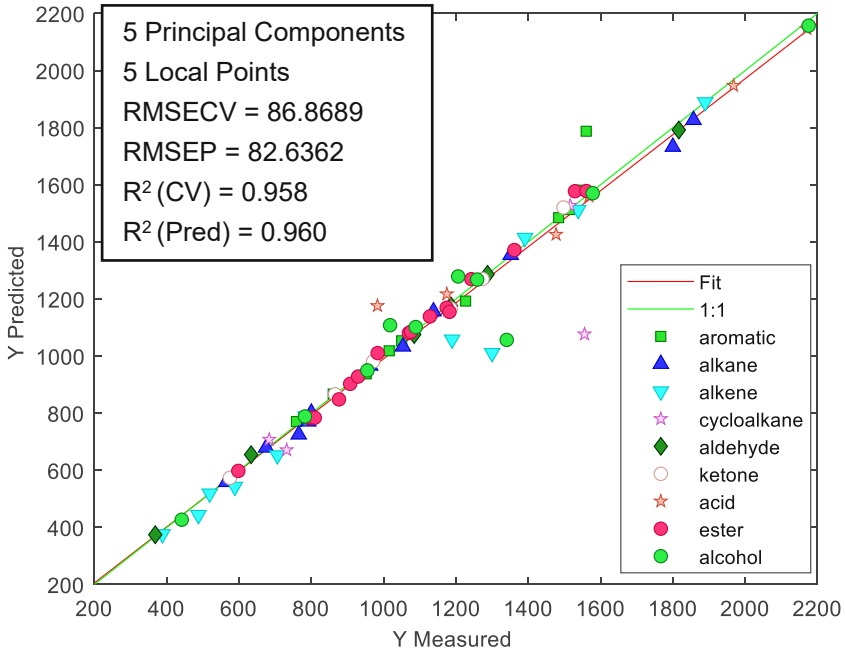


Figure 28: RI prediction on DB1 with LWR using external test set.

From the results in figure 27 and 28 we concluded that LWR performed worse than PLS in predicting RIs on DB1 column, with RMSECV and RMSEP of 86.9 and 82.6 respectively, and R2 (CV) of 0.958 and R2 (Pred) of 0.960. Based on figure 27b the big deviation between the predicted and measured RI values is specially noticeable for couple of alcohol, cycloalkane, alkene and alkane compounds, which could mean that the LWR is not able to explain and model data with big variability and outliers and thus the model was not further tested on DB5 and PEG columns. For the next trial SVM was selected as a machine learning regression to generate the model.

4.3. SVM regression for DB1

For this part the same training set and MDs (from PLS regression) were used to generate the model but this time using SVM regression. The data was also pre-processed by Pareto Scaling. The model was internally cross-validated using the same Venetian Blinds as PLS and externally validated using the test set. The performance of the model was evaluated by RMSECV and the results are reported in figure 29.

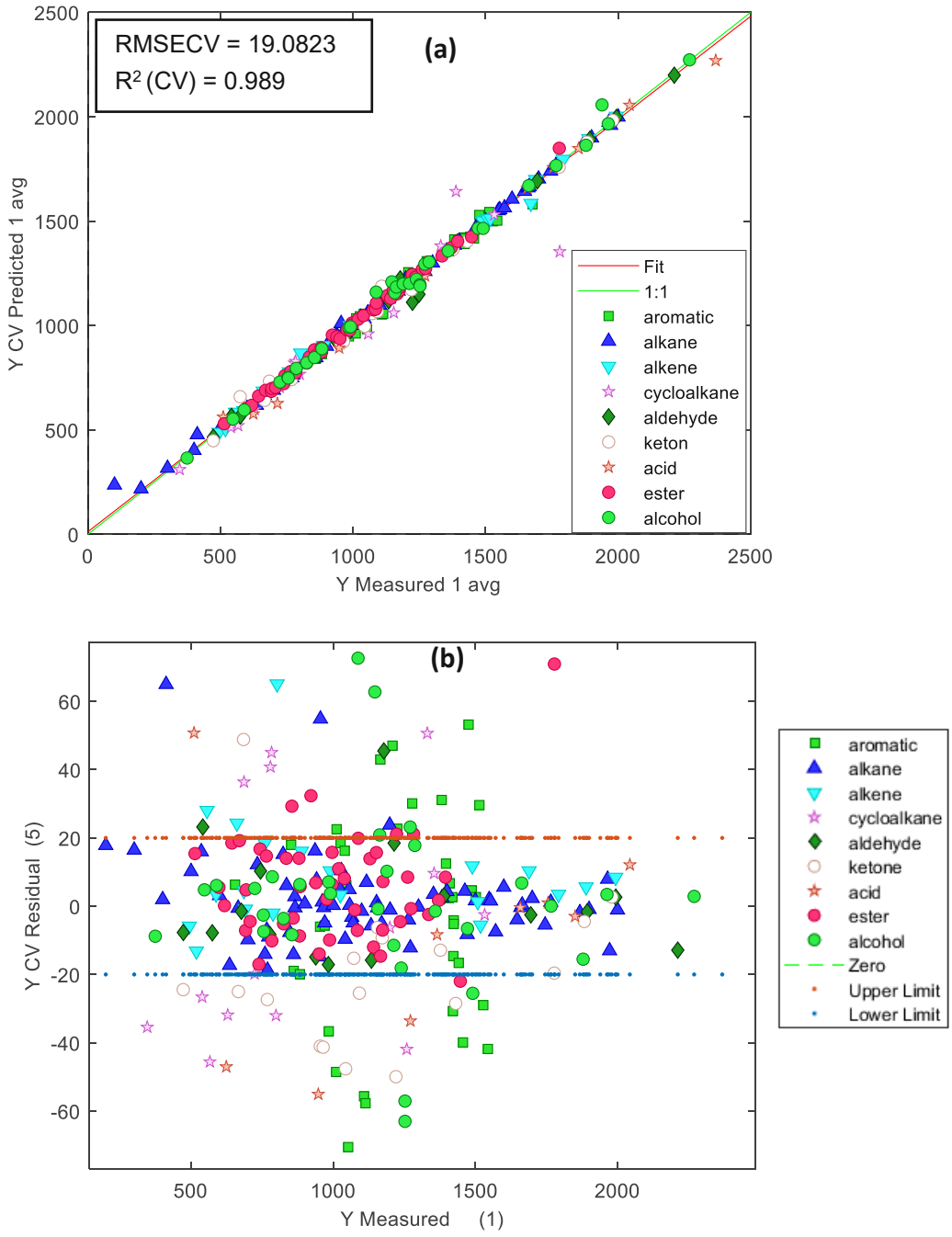


Figure 29: a) SVM calibration plot with training set on DB1 column. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin (upper and lower error tolerance limits).

From figure 29 we can summarize that, in (a) the calibration error of the SVM model for DB1 is obviously lower than with PLS and LWR with an average RMSECV of 19.1 and correlation coefficient R^2 (CV) of 0.989. It is visible in (b) that the majority of the predicted RIs of different compound classes are within the ± 20 units margin, with highest error of around ± 60 (meaning the predicted value is 60 RI units lower or higher than the actual RI value). Components from mostly aromatic, alcohol, cycloalkane and ketone classes are crossing the ± 20 units margin, and only very few components from other compound classes do so, too.

However, it is reported in the literature that predicted RI error can be acceptable within the range of ± 20 units [2,3]. And referring to section 3.4. Data Quality, the data source of Pubchem is not reliable, Therefore, to reduce the errors in our model we had to remove RIs that were of poor quality and some outliers which were not perfectly predictable (data outside the ± 20 units margin are seen as outliers), in order to improve the RMSECV.

The data elimination is reported in figure 30 and table 9), and SVM model was recalculated with the rest of the data. The results are reported in figure 31.

Table 9: It demonstrates how many components were present within each compound class before and after the data elimination and the remaining data percentage.

Number of data points			
Class	Before outlier removal	After removal	% of data points retained
aromatic	37	33	89
alkane	70	52	74
alkene	21	16	76
cycloalkane	18	14	78
aldehyde	18	14	78
ketone	19	15	79
acid	11	9	82
ester	47	38	81
alcohol	32	25	78

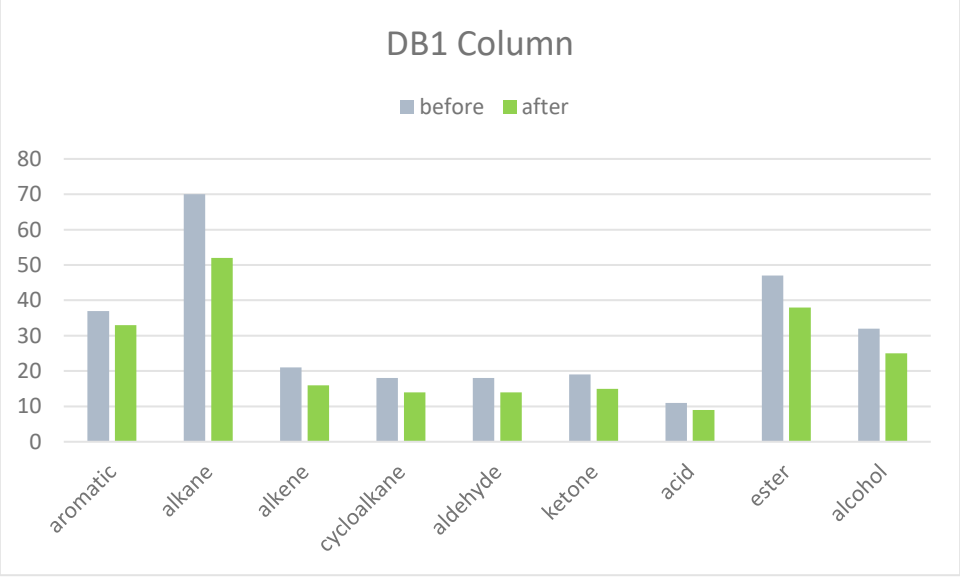


Figure 30: Plotted data from table 9 before and after data elimination per each compound class.

From table 9 we can see that the remaining data percentage for all compound classes is almost similar (data removal was more or less to the same extent) and none of the compound classes was particularly affected by this, except alkanes (having the lowest

percentage of remaining data ~74%). This could mean that the model has a bit difficulties predicting RIs of alkanes on DB1 column.

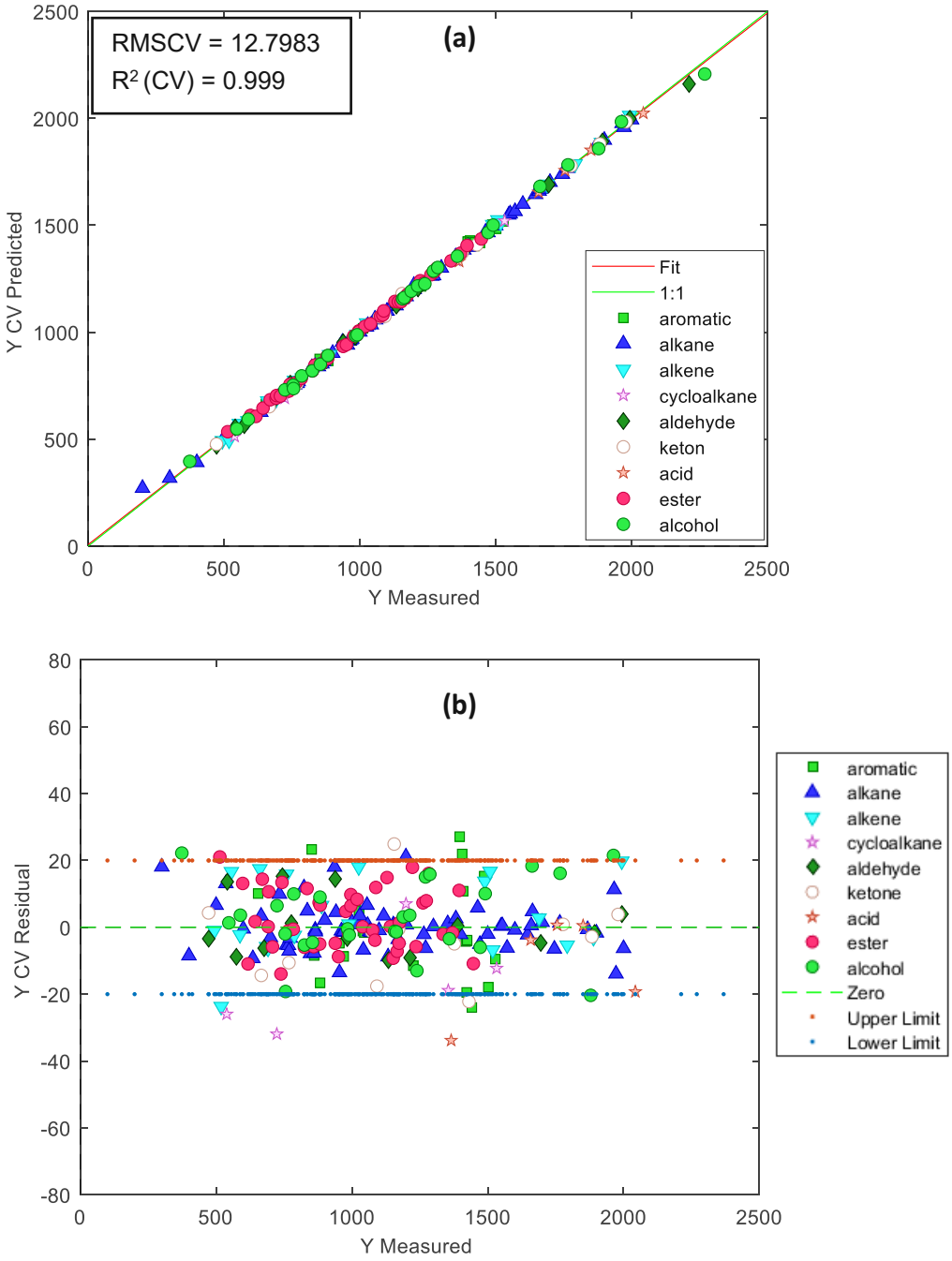


Figure 31: a) SVM calibration plot on DB1 column after data elimination from the training set. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin.

We can conclude from figure 31 (a) that the model shows significant improvement once outliers from figure 29 (b) are removed from the dataset. Application of the SVM model results, after outlier elimination, in an average RMSECV of 12.8 and correlation coefficient R^2 (CV) of 0.999. It is also visible in figure 29 (b) that almost all of the predicted RIs are within the ± 20 units margin with some minor exceptions for aromatics and cycloalkanes which are a bit above/below the limits.

Thus, the current SVM model on DB1 was accepted according to literature values [2,3] and further tested for prediction on the external test set (refer to figure 32).

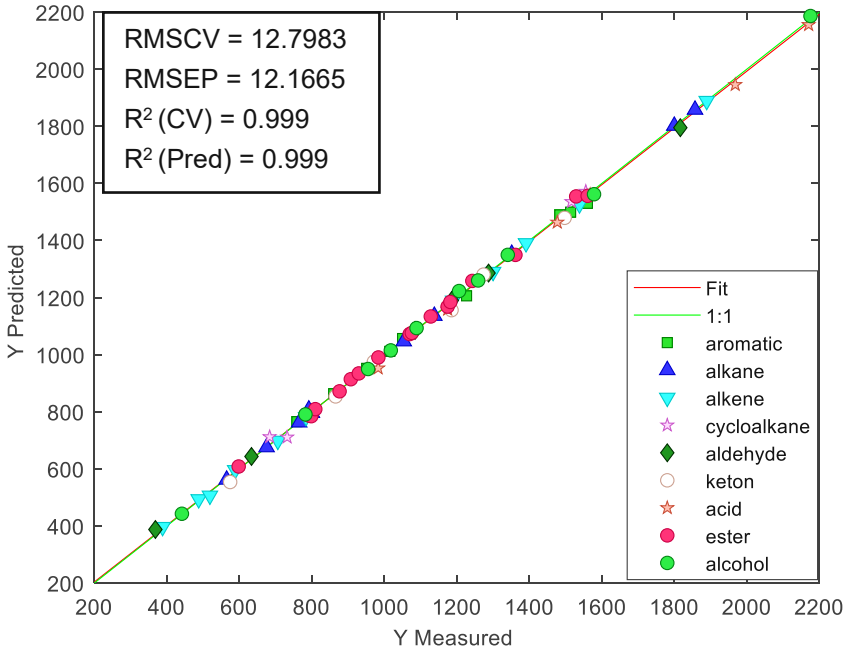


Figure 32: RI prediction on DB1 with the final SVM model using external test set.

The result in figure 32 with RMSEP = 12.2 is very close to RMSECV = 12.8 with R^2 (CV) = 0.999 and R^2 (Pred) = 0.999. Since average RMSECV and RMSEP errors are very close and within the ± 20 units margin, this gives us the conformation that SVM model for the prediction of RI values on a DB1 column is acceptable.

4.4. SVM regression for DB5

For modelling on DB5 column the same approach was used as in section “4.3. SVM regression for DB1”. The only difference to DB1 model is that Yi data (RIs) which is specific for DB5 column in this case. The data was also pre-processed by Pareto Scaling. The model was internally cross-validated using the Venetian Blinds and externally validated using the test set. The performance of the model was evaluated by RMSECV and the results are reported in figure 33.

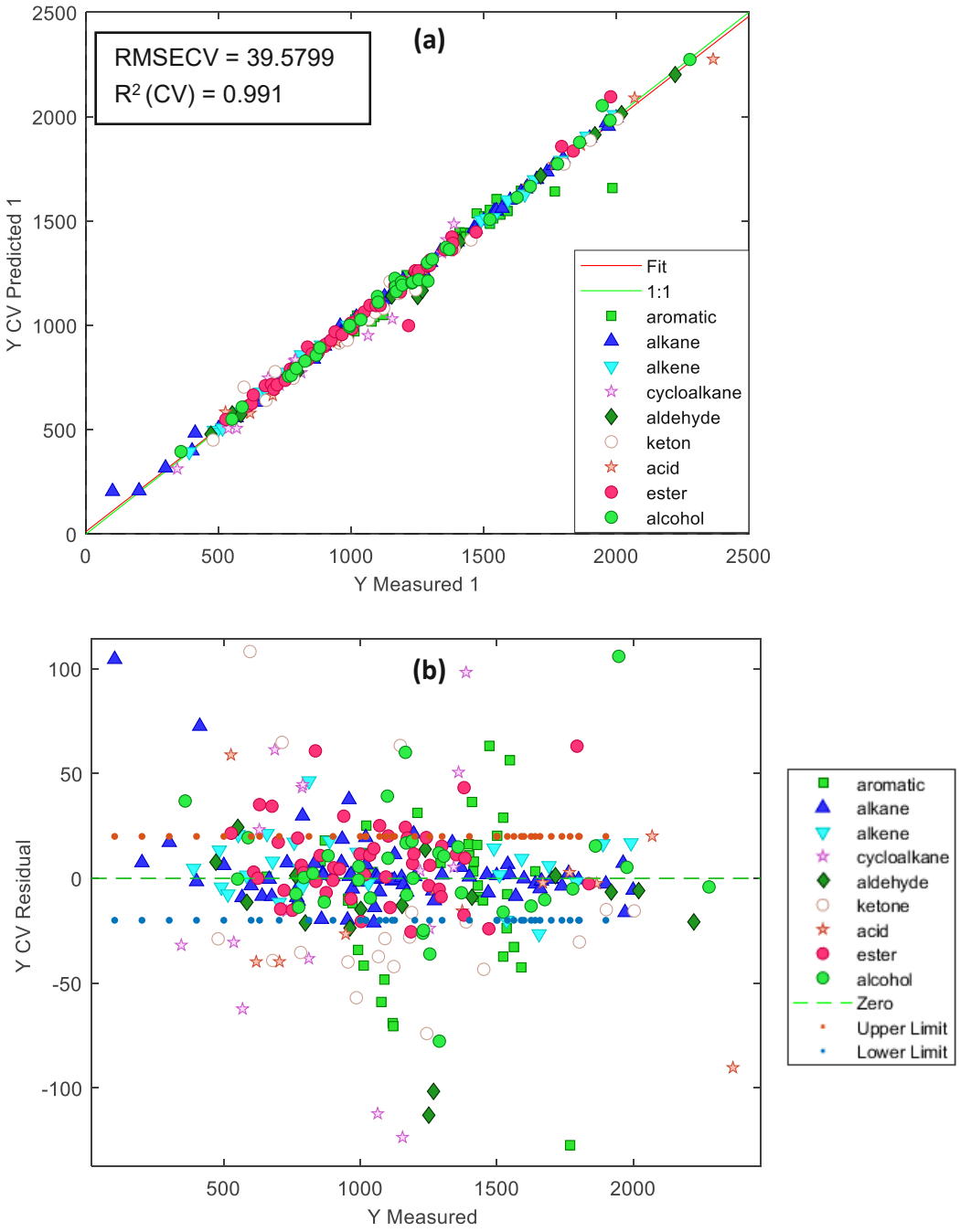


Figure 33: (a) SVM calibration plot with training set on DB5 column. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin.

From figure 33 we can summarize that, in (a) the calibration error of the SVM model for DB5 is a bit higher than for DB1, with an average RMSECV of 39.6 and correlation coefficient R^2 (CV) of 0.991. It is visible in (b) that the predicted RIs of different compound classes are crossing the ± 20 units margin, with highest error of around ± 100 (meaning the predicted value is 100 RI units lower or higher than the actual RI value). Components from mostly aromatic, alcohol, cycloalkane, aldehyde, ester and ketone classes are crossing the ± 20 units margin. So we were following the concept of DB1 model and assuming the components outside the limits as outliers since they cannot be correctly predicted by the model. Hence, all the components outside those

limits were eliminated from the dataset (data elimination is reported in figure 34 and table 10), and SVM model was recalculated with the rest of the data. The results are reported in figure 35.

Table 10: It demonstrates how many components were present within each compound class before and after the data elimination and the remaining data percentage.

Number of data points			
Class	Before outlier removal	After removal	% of data points retained
aromatic	42	28	67
alkane	69	67	97
alkene	25	23	92
cycloalkane	16	6	38
aldehyde	16	13	81
ketone	19	12	63
acid	11	9	82
ester	50	43	86
alcohol	35	29	83

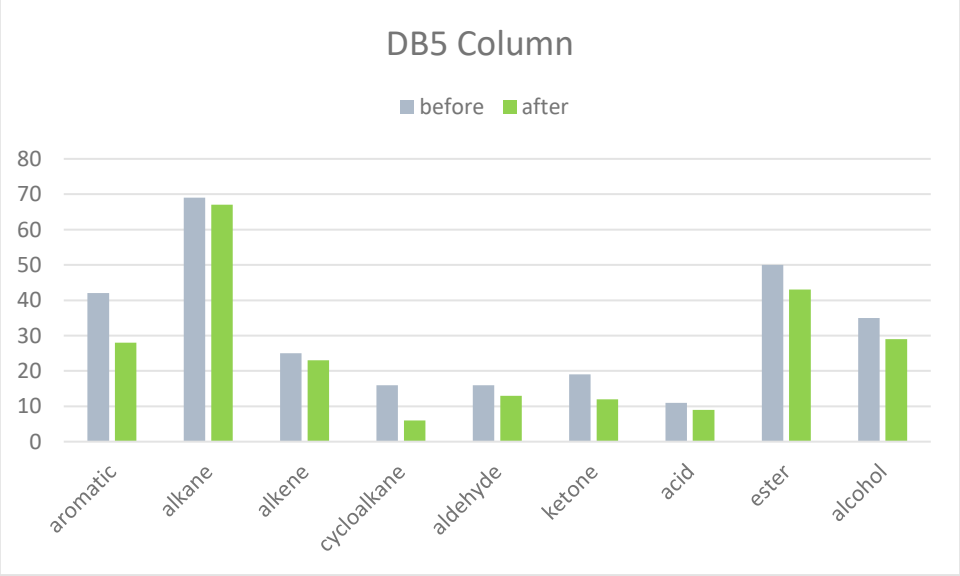


Figure 34: Plotted data from table 10 before and after data elimination per each compound class.

In table 10, the remaining data percentage is very different for each compound class. And for DB5 column more data was eliminated than for DB1 (refer to table 9). For instance, cycloalkane class is significantly affected where only 38% of the original data remained in the dataset and the rest were eliminated. Aromatic class is also affected by the data removal (with 67% data remaining). This could mean that the used descriptors were not fully able to explain polar interactions. Furthermore, it is noticeable that alkanes and alkenes are barely affected by the data removal. Hence, when compared with table 9, we notice that alkanes are better predicted on DB5 column than on DB1 column. This could be explained by the fact that *n*-alkanes on

Pubchem [60] have only a single RI value on all columns, meaning they might have been measured and reported mostly on DB5 column composition (which is the most common used stationary phase in analytical laboratories).

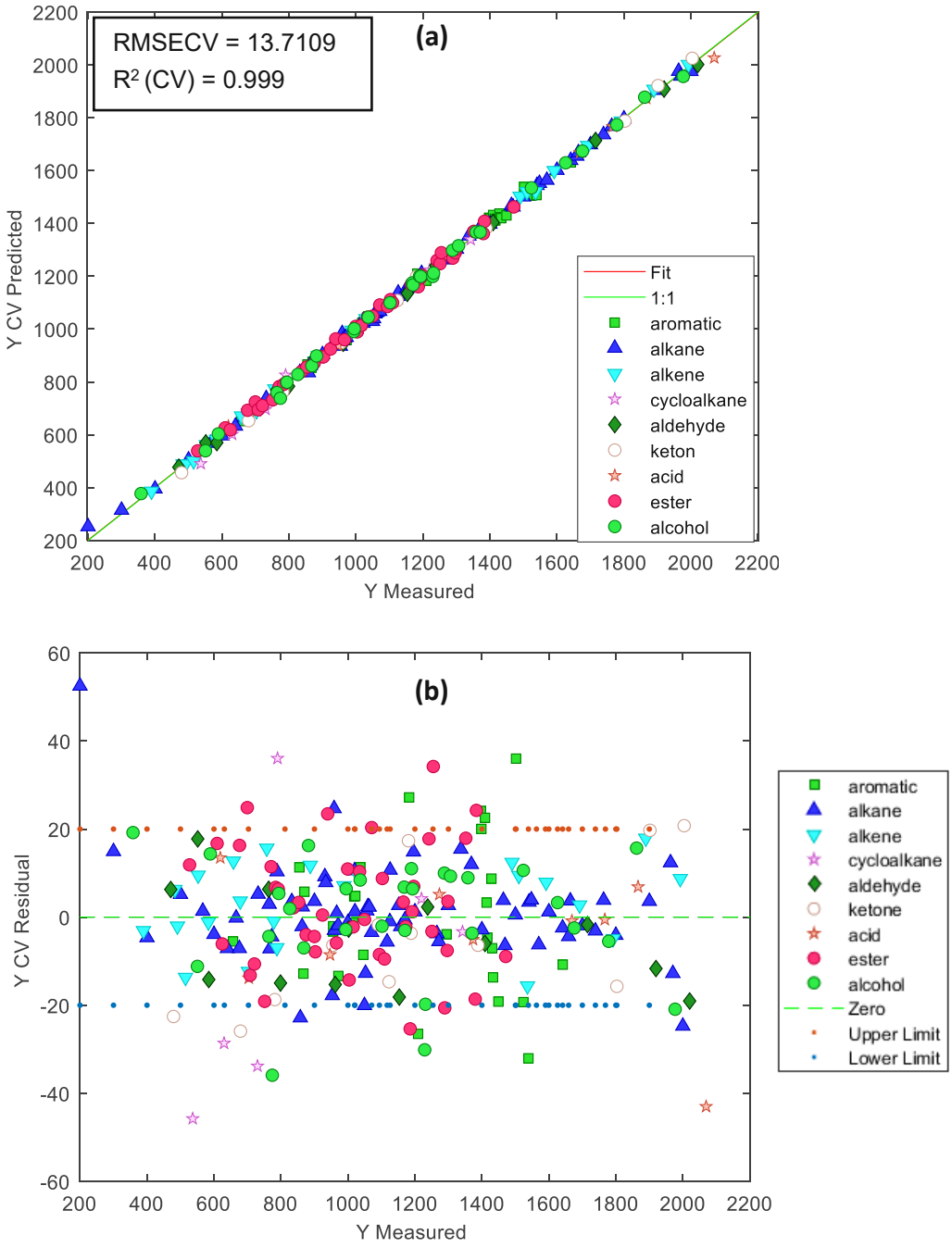


Figure 35: a) SVM calibration plot on DB5 column after data elimination from the training set. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin.

We can conclude from figure 35 (a) that the model improves once outliers from figure 33 (b) are removed from the dataset. SVM model after outlier elimination resulting with an average RMSECV of 13.7 and correlation coefficient R2 (CV) of 0.999. It is also visible in 33 (b) that almost all of the predicted RIs are within the ± 20 units margin with

some minor exceptions for alcohols, aromatics, esters and cycloalkanes which are a bit above/below the limits.

Thus, the current SVM model on DB5 was accepted according to literature values [2,3] and further tested for prediction on the external test set (refer to figure 36).

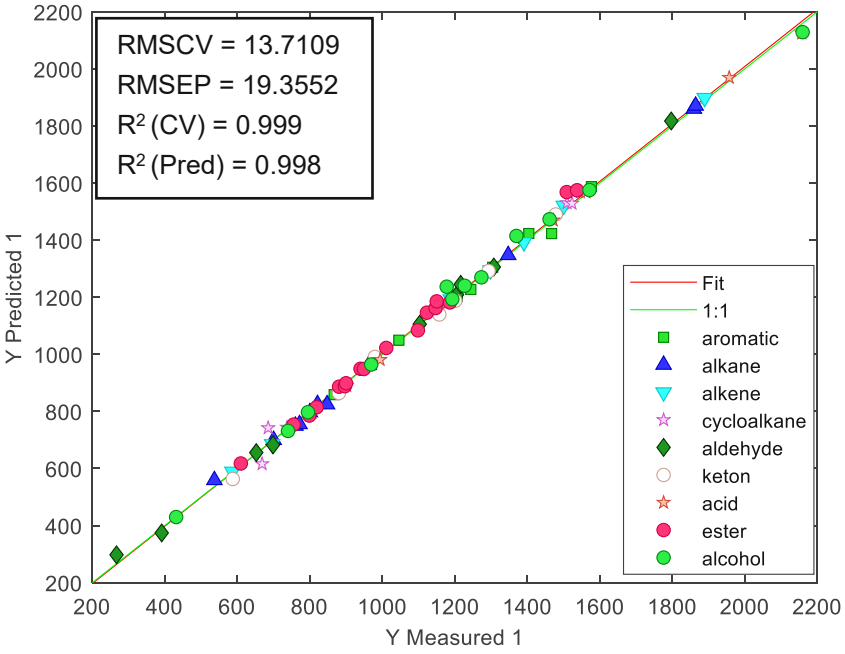


Figure 36: RI prediction on DB5 with the final SVM model using external test set.

The result in figure 36 with RMSEP = 13.7 is somewhat close to RMSECV = 19.4 with R^2 (CV) = 0.999 and R^2 (Pred) = 0.998. Since average RMSECV and RMSEP errors are not too far away from each other and within the ± 20 units margin, this gives us the conformation that SVM model on DB5 is acceptable.

4.5. SVM regression for PEG

Likewise for modeling on PEG column the same approach was used as in section “4.3. SVM regression for DB1”. But this time the Yi data (RIs) is specific for PEG column. The data was pre-processed, internally and externally validated as before. The performance of the model was evaluated by RMSECV and the results are reported in figure 37.

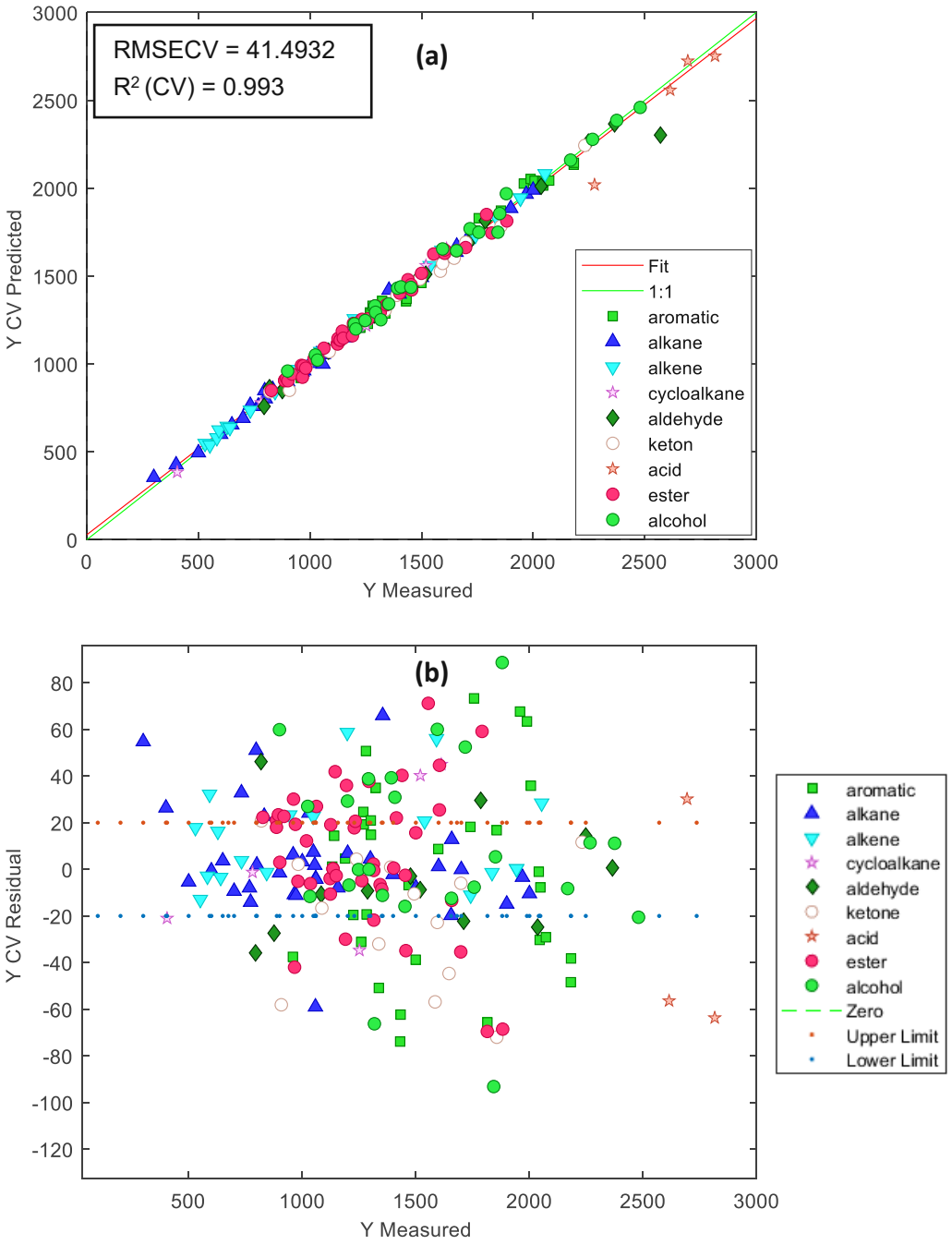


Figure 37: a) SVM calibration plot with training set on PEG column. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin.

As demonstrated in figure 37 we can summarize that, in (a) the calibration error of the SVM model for PEG is comparable to the error on DB5, with an average RMSECV of 41.5 and correlation coefficient R^2 (CV) of 0.993. It is visible in (b) that the predicted RIs of different compound classes are crossing the ± 20 units margin, with highest error of around -100 (meaning the predicted value is 100 RI units lower than the actual RI value). Components from mostly aromatic, alcohol, cycloalkane, ester and ketone classes are crossing the ± 20 units margin. So the same concept is applied for PEG and the components outside the limits are assumed as outliers and were eliminated

from the dataset (data elimination is reported in figure 38 and table 11), and SVM model was recalculated with the rest of the data. The results are reported in figure 39.

Table 11: It demonstrates how many components were present within each compound class before and after the data elimination and the remaining data percentage.

Number of data points			
Class	Before outlier removal	After removal	% of data points retained
aromatic	37	17	46
alkane	38	33	87
alkene	20	17	85
cycloalkane	9	4	44
aldehyde	15	9	60
ketone	18	9	50
acid	10	0	0
ester	50	38	76
alcohol	35	17	49

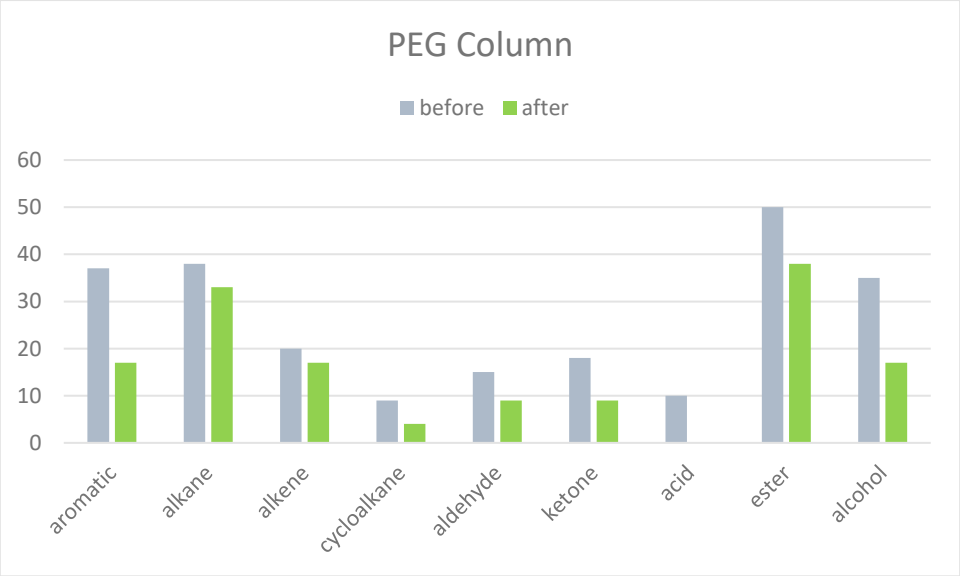


Figure 38: Plotted data from table 11 before and after data elimination per each compound class.

Table 11 tells us a lot about the remaining data percentage from all compound classes. Acids are drastically affected by this where the whole class was removed from the model. Acids, cycloalkanes, alcohols and ketones show also significant data reduction. The least affected classes are the alkanes, alkenes and esters. This could mean that the used descriptors were not fully able to explain strong polar interactions of these certain classes on PEG column. And when compared with the data in table 9 and 10 (for DB1 and DB5), a larger fraction of data was excluded for PEG column.

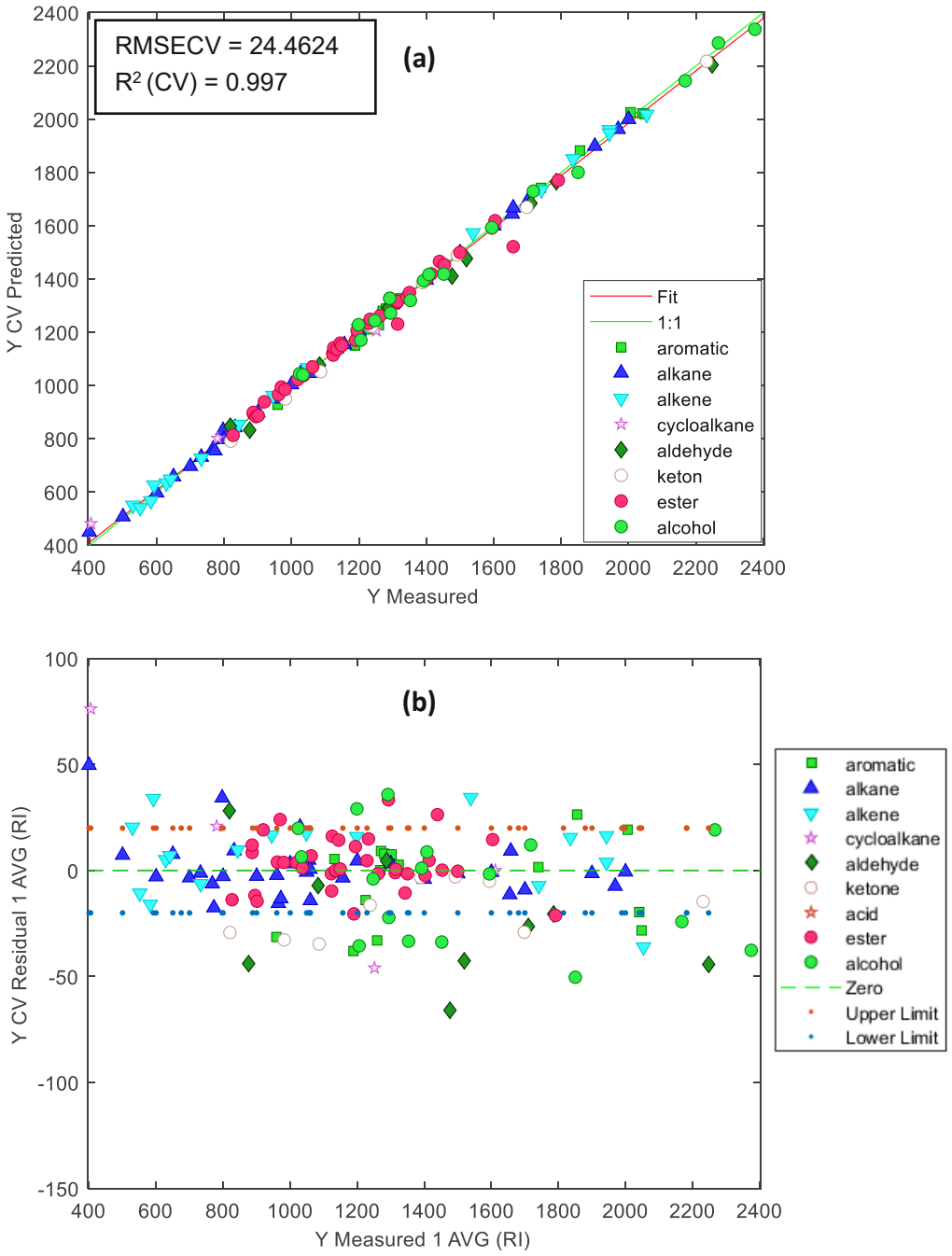


Figure 39: a) SVM calibration plot on PEG column after data elimination from the training set. (b) Residuals plot of predicted RIs versus measured RIs with ± 20 units margin.

We can conclude from figure 39 (a) that the model shows an improvement once outliers from figure 37 (b) are removed from the dataset, but on the cost of losing data (e.g. acid class). SVM model after outlier elimination resulting with an average RMSECV of 24.5 and correlation coefficient R2 (CV) of 0.997. In figure 37 (b) we can notice that the predicted RIs of some compounds are still crossing the ± 20 units margin e.g. aldehydes, alcohols and cycloalkanes. Even though the SVM model for PEG is not as good as for DB1 and DB5, we still accepted it and further tested it for prediction on the external test set (refer to figure 40).

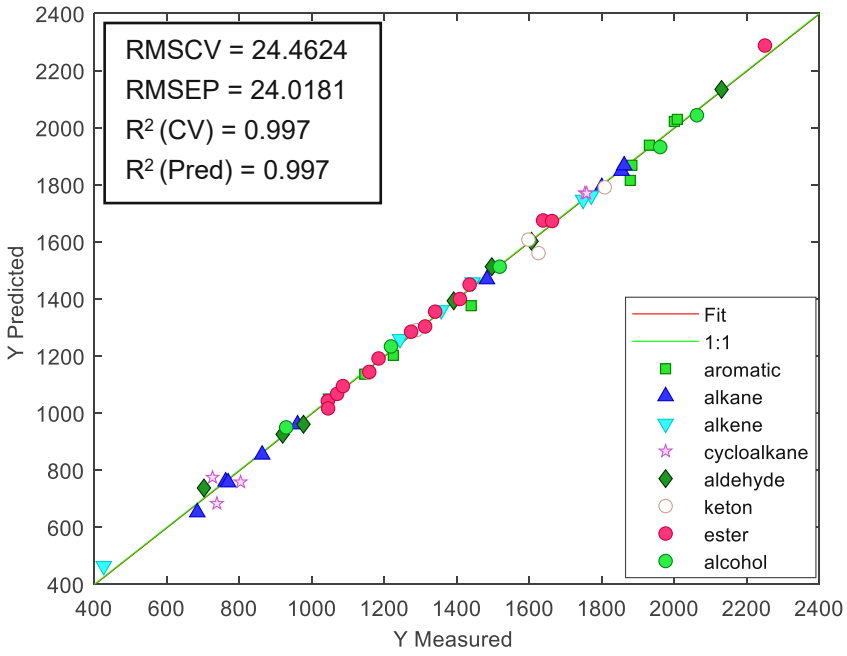


Figure 40: RI prediction on PEG with the final SVM model using external test set.

The result in figure 40 with RMSEP = 24.0 is close to RMSECV = 24.5 with R^2 (CV) = 0.997 and R^2 (Pred) = 0.997. The average RMSECV and RMSEP errors are a bit outside the ± 20 units margin, meaning that the SVM prediction model performs a bit poorly for PEG column. And further data elimination would only mean losing essential data and overfitting the model which is not recommended.

Nevertheless, we were curious to test our SVM model on experimentally obtained RIs values of ester class on different column compositions: OV-7, DC-710, OV-25, XE-60, OV-225 and Silar-5CP (compositions explained in table 7). The RIs of around 90 saturated esters were obtained from literature [2,47] and were calibrated using the same steps for SVM model (4.3. SVM regression for DB1). Due to the small dataset availability, the model was only calibrated but not tested. The results are shown in figure 41.

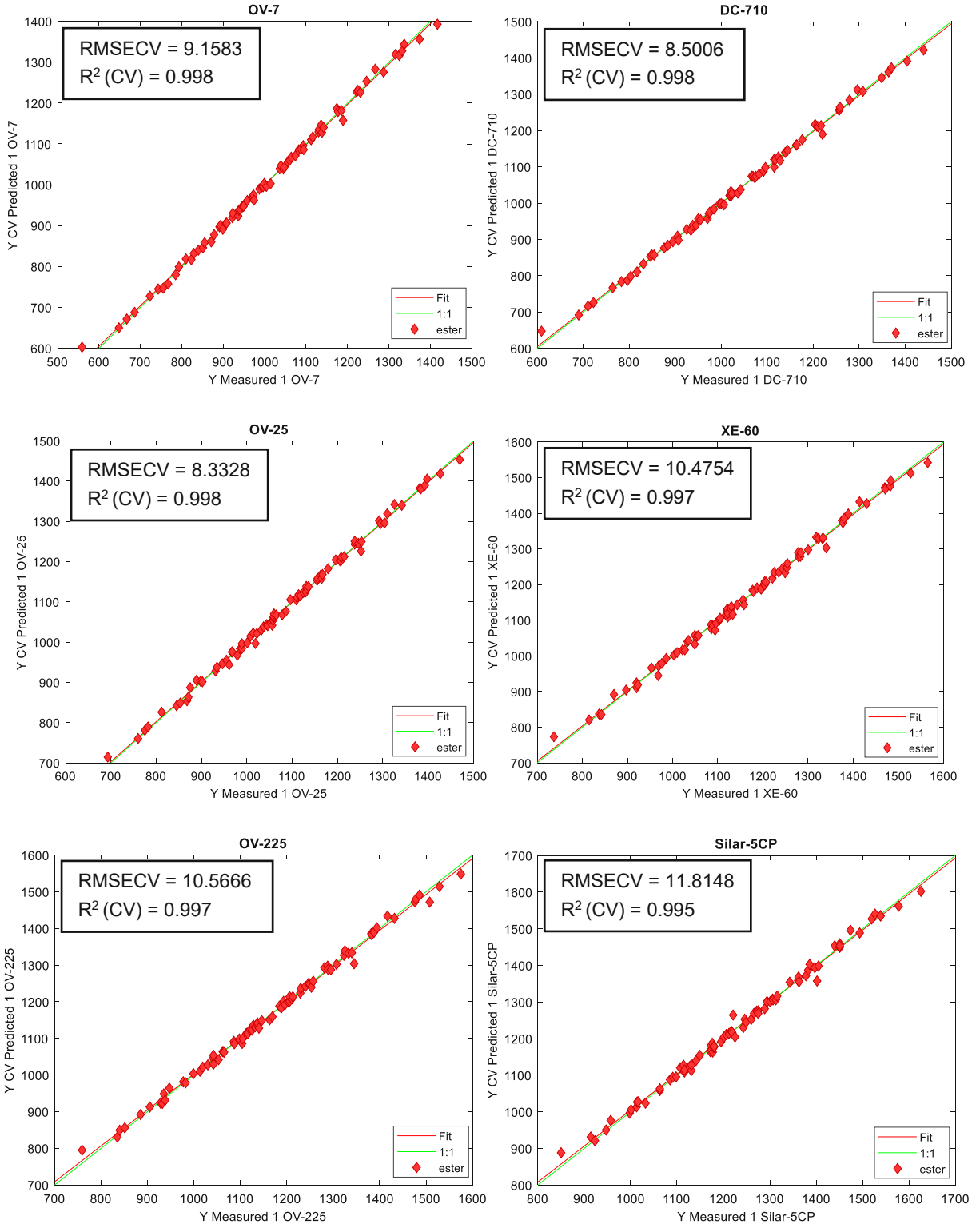


Figure 41: SVM model calibration for RI prediction of esters on different columns.

The results of figure 41 indicate an excellent prediction performance for esters (RMSECV of the different columns ranging between 8 – 12). However, it is quite remarkable, despite the fact that different stationary phases have been used, the prediction of the RI values is still very good compared to the simulations on the DB1, DB5, PEG columns.

At this point, the effect of the data quality is specially emphasized when we compare the SVM model on DB1, DB5 and PEG to the SVM model for the esters [2,47] on different columns. In the case of DB1, DB5 and PEG the models improved once the data (outliers) outside the ± 20 units margin were removed. In contrast to this, in the esters case (figure 41), the model had an excellent performance (RMSECV between 8 – 12) without the need of removing any data.

This observation could be interpreted in two ways: Firstly, the ester data was from one single consistent source and not from different sources of unknown quality as is the case for Pubchem data. Secondly, the data consists of only one single compound class (esters), therefore it makes it easier to predict a single class than having 9 different compound classes.

Hence, this concludes that if the input data is of good quality, so is the output data.

5. Conclusion and Outlook

Several studies have been published where molecular descriptors have been used to predict GC RIs, but most of them restrict the study on certain compound classes and on certain column compositions. However, this thesis was able to derive a quantitative structure-property relationship model to predict GC retention indices of alkanes, alkenes, cycloalkanes, aromatics, alcohols, acids, aldehydes, ketones and esters for around 400 compounds obtained from the Pubchem database [60] on the most commonly used stationary phases DB1, DB5, PEG. For the purpose of prediction, 266 molecular descriptors were obtained from the chemical online database OCHEM [81] to describe the topology, geometry, electronic states, etc. of these different compounds. The use of a single descriptor can capture only part of the property of interest, which is not satisfactory in our case, since we need to describe very different compound classes and on different columns.

To generate the model three different regression methods were used, PLS, LWR and SVM which were compared to predict RIs of different compounds. Based on the quality of the collected data and by comparing the models we arrived to the conclusion that the data of the Pubchem database is comprehensive, however, not always of consistently high quality, and must be used only with critical attention. Otherwise, errors and outliers in the input data may deteriorate the model's predictive abilities, as the descriptors are not able to explain irregular data variations. Under these conditions it has become evident that the SVM model is more robust towards outliers and towards input data errors when compared to PLS and LWR. Therefore, SVM performs better in predicting retention indices of different compound classes (refer to chapter 4).

To summarize, the SVM model performs well for DB1 (RMSECV of 12.8) and DB5 columns (RMSECV of 19.4) after the data removal, but a bit poorerfor predicting RIs on a PEG column (RMSECV of 24.5). Moreover, the data removal did not affect the DB1 data as much as it affected the DB5 and PEG data. For DB5 mostly affected classes were cycloalkanes and aromatics. And for PEG the most affected were cycloalkanes, alcohols, ketones and acids, nevertheless the acid class was completely lost during data treatment. This could mean that the used descriptors were not fully able to explain strong polar interactions of these compound classes on more polar stationary phases. Meaning the descriptors are not specific enough to the column composition and to its interaction with the compound.

Nonetheless, to further improve the current model these different factors should be taken into account:

- Quality of the input data (having a single consistent data source)
- Molecular descriptors specific for the column and for certain interactions with the column
- Data distribution (number of compounds and RI frequency within each compound class should be in equal proportions)

It is expected that by improving these factors the prediction ability of the PLS and LWR models would also improve.

6. References

[1] Feldhausen, J., Bell, D. C., Yang, Z., Faulhaber, C., Boehm, R., and Heyne, J. "Synthetic Aromatic Kerosene Property Prediction Improvements with Isomer Specific Characterization via GCxGC and Vacuum Ultraviolet Spectroscopy." Fuel, Vol. 326, 2022, p. 125002. https://doi.org/10.1016/j.fuel.2022.125002.

[2] Biancolillo, A., and D'Archivio, A. A. "Transfer of Gas Chromatographic Retention Data among Poly(Siloxane) Columns by Quantitative Structure-Retention Relationships Based on Molecular Descriptors of Both Solutes and Stationary Phases." Journal of Chromatography A, Vol. 1663, 2022, p. 462758. https://doi.org/10.1016/j.chroma.2021.462758.

[3] D'Archivio, A. A., Maggi, M. A., and Ruggieri, F. "Cross-Column Prediction of Gas-Chromatographic Retention Indices of Saturated Esters." Journal of Chromatography A, Vol. 1355, 2014, pp. 269–277. https://doi.org/10.1016/j.chroma.2014.06.002.

[4] Ettre, L. S., and Sakodynskii, K. I. "M. S. Tswett and the Discovery of Chromatography I: Early Work (1899–1903)." Chromatographia, Vol. 35, Nos. 3–4, 1993, pp. 223–231. https://doi.org/10.1007/BF02269707.

[5] Sayed, M. A. A Review of Chromatography: Principles, Classification, Applications. DOI: 10.13140/RG.2.2.22113.43361.

[6] Tswett, M.: Physikalisch-chemische Studien Über Das Chlorophyll. Die Adsorptionen." Berichte der Deutschen Botanischen Gesellschaft, Vol. 24, No. 6, 1906, pp. 316–323. https://doi.org/10.1111/j.1438-8677.1906.tb06524.x.

[7] Patel, M. "Review Article: Chromatography Principle and Applications." Human Journals, Vol. 13, No. 4, 2018.

[8] "The Nobel Prize in Chemistry 1952". https://www.nobelprize.org. |Accessed 18 Juni 2023|.

[9] James, A. T., and Martin, A. J. P. "Gas-Liquid Partition Chromatography: The Separation and Micro-Estimation of Volatile Fatty Acids from Formic Acid to Dodecanoic Acid." Biochemical Journal, Vol. 50, No. 5, 1952, pp. 679–690. https://doi.org/10.1042/bj0500679.

[10] Grob, R. L. Theory of Gas Chromatography. In Modern Practice of Gas Chromatography, John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 23–63.

[11] Masucci, J. A., and Caldwell, G. W. Techniques for Gas Chromatography/Mass Spectrometry. In Modern Practice of Gas Chromatography, John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 339–401.

[12] Flanagan, R. J., Cuypers, E., Maurer, H. H., and Whelpton, R. Fundamentals of Analytical Toxicology. Analytical Chemistry. DOI:10.1002/9781119122357, 2020.

[13] Forgács, E., and Cserháti, T. Gas Chromatography. In Food Authenticity and Traceability, Elsevier, 2003, pp. 197–217. DOI: 10.1533/9781855737181.1.197.
[14] Davis, J. M., Pompe, M., and Samuel, C. “Justification of Statistical Overlap Theory in Programmed Temperature Gas Chromatography: Thermodynamic Origin of Random Distribution of Retention Times.” Analytical Chemistry, Vol. 72, No. 22, 2000, pp. 5700–5713. https://doi.org/10.1021/ac000613u.
[15] Harshal D. P., Chandrabhan B Patil, Vikas V. Patil, Pankaj S. Patil, Amol R. Pawar. A Brief Review on Gas Chromatography. Asian Journal of Pharmaceutical Analysis. 2023; 13(1):47-2. doi: 10.52711/2231-5675.2023.00008
[16] Gawale, D. S., Jaiswal, R. S., Chavhan, R. P., Chaudhari, D. D., Jaiswal, N., and Patil, D. A. “A Short Review on Gas Chromatography.” International Journal of Pharmaceutical Research and Applications, Vol. 7, No. 6, 2022.
[17] Beale, D. J., Pinu, F. R., Kouremenos, K. A., Poojary, M. M., Narayana, V. K., Boughton, B. A., Kanojia, K., Dayalan, S., Jones, O. A. H., and Dias, D. A. “Review of Recent Developments in GC–MS Approaches to Metabolomics-Based Research.” Metabolomics, Vol. 14, No. 11, 2018, p. 152. https://doi.org/10.1007/s11306-018-1449-2.
[18] Lovestead, T. M., Urness, K. N. Gas Chromatography/Mass Spectrometry. In Materials Characterization, ASM International, 2019, pp. 235–241. DOI: 10.31399/asm.hb.v10.a0006664.
[19] Liu, Z., and Phillips, J. B. “Comprehensive Two-Dimensional Gas Chromatography Using an On-Column Thermal Modulator Interface.” Journal of Chromatographic Science, Vol. 29, No. 6, 1991, pp. 227–231. https://doi.org/10.1093/chromsci/29.6.227.
[20] Akporhonor, E. E. Temperature-Programmed and Isothermal Kovats Retention Indices in Gas Chromatography. The University of Manchester Institute of Science and Technology. [Doctoral thesis], Manchester UK, 1985.
[21] https://Www.shimadzu.com/an/Service-Support/Technical-Support/Analysis-Basics/Gcms/Fundamentals/Retention/Retentiontime_parameters.Html#:~:Text=We%20call%20the%20length%20of,Called%20the%20adjusted%20retention%20time.|Accessed 13 April 2023|.
[22] Kovats, E. “Gas-Chromatographische Charakterisierung Organischer Verbindungen Teil 1 : Retentionsindices Aliphatischer Halogenide, Alkohole, Aldehyde Und Ketone.” Vol. 41, No. 7, 1958. https://doi.org/https://doi.org/10.1002/hlca.19580410703.
[23] Zellner, B. d’Acampora, Bicchi, C., Dugo, P., Rubiolo, P., Dugo, G., and Mondello, L. “Linear Retention Indices in Gas Chromatographic Analysis: A

Review.” Flavour and Fragrance Journal, Vol. 23, No. 5, 2008, pp. 297–314. https://doi.org/10.1002/ffj.1887.
[24] Royal Society of Chemistry, Analytical Methods Committee. Analyst 1980; 105: 262–273.” https://doi.org/https://doi.org/10.1002/ffj.1887C.
[25] MacNamara, K., Ochiai, N., Sasamoto, K., Hoffmann, A., and Shellie, R. “AromaOffice: Application of a Novel Linear Retention Indices Database to a Complex Hop Essential Oil.” GERSTEL Application Note No. 183, 2016.
[26] Chapter 3 Fundamentals of the Chromatographic Process The Thermodynamics of Retention in Gas Chromatography. 1988, pp. 55–92.
[27] Santiuste, J. M., and Takács, J. M. “Relationships Between Retention Data of Benzene and Chlorobenzenes with Their Physico-Chemical Properties and Topological Indices.” Chromatographia, Vol. 58, Nos. 1–2, 2003, pp. 87–96. https://doi.org/10.1365/s10337-003-0013-y.
[28] Wang, Y. H., and Wong, P. K. “Correlation Relationships between Physico-Chemical Properties and Gas Chromatographic Retention Index of Polychlorinated-Dibenzofurans.” Chemosphere, Vol. 50, No. 4, 2003, pp. 499–505. https://doi.org/10.1016/S0045-6535(02)00491-5.
[29] Zenkevich, I. Kovats Retention Index System. Encyclopedia of Chromatography, Vol. 1. ISBN: 9781420084597.
[30] van Den Dool, H., and Kratz, P.D. “A Generalization of the Retention Index System Including Linear Temperature Programmed Gas—Liquid Partition Chromatography.” Journal of Chromatography A, Vol. 11, 1963, pp. 463–471. https://doi.org/10.1016/S0021-9673(01)80947-X.
[31] Majlát, P., Erdös, Z., and Takás, J. “Calculation and Application of the Retention Indices in Programmed-Temperature Gas Chromatography.” Journal of Chromatography A, Vol. 91, 1974, pp. 89–103. https://doi.org/10.1016/S0021-9673(01)97888-4.
[32] Lee, M. L., Vassilaros, D. L., and White, C. M. “Retention Indices for Programmed-Temperature Capillary-Column Gas Chromatography of Polycyclic Aromatic Hydrocarbons.” Analytical Chemistry, Vol. 51, No. 6, 1979, pp. 768–773. https://doi.org/10.1021/ac50042a043.
[33] Hérent, M.-F., De Bie, V., and Tilquin, B. “Determination of New Retention Indices for Quick Identification of Essential Oils Compounds.” Journal of Pharmaceutical and Biomedical Analysis, Vol. 43, No. 3, 2007, pp. 886–892. https://doi.org/10.1016/j.jpba.2006.09.005.
[34] Robinson, P. G., and Odell, A. L. “A System of Standard Retention Indices and Its Uses the Characterisation of Stationary Phases and the Prediction of

Retention Indices.” Journal of Chromatography A, Vol. 57, 1971, pp. 1–10. https://doi.org/10.1016/0021-9673(71)80001-8.
[35] Rohrschneider, L. “Eine Methode Zur Charakterisierung von Gaschromatographischen Trennflüssigkeiten.” Journal of Chromatography A, Vol. 22, 1966, pp. 6–22. https://doi.org/10.1016/S0021-9673(01)97064-5.
[36] P. Souter. “Calculation of Rohrschneider Constants.” Journal of Chromatography, 1974, pp. 231–236. https://doi.org/https://doi.org/10.1016/S0021-9673(00)85733-7.
[37] Berthod, A., Zhou, E. Y., Le, K., and Armstrong, D. W. “Determination and Use of Rohrschneider-McReynolds Constants for Chiral Stationary Phases Used in Capillary Gas Chromatography.” Analytical Chemistry, Vol. 67, No. 5, 1995, pp. 849–857. https://doi.org/10.1021/ac00101a010.
[38] Rohrschneider, L. “Explanatory Coefficients for Stationary Phases in Gas Chromatography from McReynolds Phase Constants.” Chromatographia, Vol. 38, Nos. 11–12, 1994, pp. 679–688. https://doi.org/10.1007/BF02269621.
[39] Rajkó, R., Körtvélyesi, T., Sebők-Nagy, K., and Görgényi, M. “Theoretical Characterization of McReynolds’ Constants.” Analytica Chimica Acta, Vol. 554, Nos. 1–2, 2005, pp. 163–171. https://doi.org/10.1016/j.aca.2005.08.024.
[40] Merck. The Retention Index System in Gas Chromatography: McReynolds Constants. , 1999.
[41] https://www.chromatographyonline.com/View/Stationary-Phase-Selectivity-Chemistry-behind-Separation-0.|Accessed 13 August 2023|.
[42] Babushok, V. I. “Chromatographic Retention Indices in Identification of Chemical Compounds.” TrAC Trends in Analytical Chemistry, Vol. 69, 2015, pp. 98–104. https://doi.org/10.1016/j.trac.2015.04.001.
[43] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz’min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., and Tropsha, A. “QSAR Modeling: Where Have You Been? Where Are You Going To?” Journal of Medicinal Chemistry, Vol. 57, No. 12, 2014, pp. 4977–5010. https://doi.org/10.1021/jm4004285.
[44] Yousefinejad, S., and Hemmateenejad, B. “Chemometrics Tools in QSAR/QSPR Studies: A Historical Perspective.” Chemometrics and Intelligent Laboratory Systems, Vol. 149, 2015, pp. 177–204. https://doi.org/10.1016/j.chemolab.2015.06.016.
[45] Hansch, C. “Quantitative Structure-Activity Relationships and the Unnamed Science.” Accounts of Chemical Research, Vol. 26, No. 4, 1993, pp. 147–153. https://doi.org/10.1021/ar00028a003.

[46] LIU, F., LIANG, Y., CAO, C., and ZHOU, N. “QSPR Study of GC Retention Indices for Saturated Esters on Seven Stationary Phases Based on Novel Topological Indices.” Talanta, Vol. 72, No. 4, 2007, pp. 1307–1315. https://doi.org/10.1016/j.talanta.2007.01.038.
[47] Lu, C., Guo, W., and Yin, C. “Quantitative Structure-Retention Relationship Study of the Gas Chromatographic Retention Indices of Saturated Esters on Different Stationary Phases Using Novel Topological Indices.” Analytica Chimica Acta, Vol. 561, Nos. 1–2, 2006, pp. 96–102. https://doi.org/10.1016/j.aca.2005.12.058.
[48] Talete Srl. DRAGON 6.0 for Windows (Software for Molecular Descriptor Calcu-Lations), 2015 http://Www.Talete.Mi.It..|Accessed 5 November 2023|.
[49] Dragon Descriptors: https://vcclab.org/Lab/Indexhlp/Dragon_descr.Html.|Accessed 18 March 2023|.
[50] List of Descriptors: https://docs.ochem.eu/x/DAAoAQ.Html.|Accessed 18 March 2023|.
[51] Gramatica, P. On the Development and Validation of QSAR Models. 2013, pp. 499–526. DOI: 10.1007/978-1-62703-059-5_21.
[52] Roberto Todeschini, and Viviana Consonni. Molecular Descriptors for Chemoinformatics. 2009. DOI: 10.1002/9783527628766.
[53] Katritzky, A. R., Chen, K., Maran, U., and Carlson, D. A. “QSPR Correlation and Predictions of GC Retention Indexes for Methyl-Branched Hydrocarbons Produced by Insects.” Analytical Chemistry, Vol. 72, No. 1, 2000, pp. 101–109. https://doi.org/10.1021/ac990800w.
[54] Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Version 2.0 Reference Manual, University of Florida: Gainesville, Florida, 1994.
[55] Matyushin, D. D., Sholokhova, A. Yu., and Buryak, A. K. “A Deep Convolutional Neural Network for the Estimation of Gas Chromatographic Retention Indices.” Journal of Chromatography A, Vol. 1607, 2019, p. 460395. https://doi.org/10.1016/j.chroma.2019.460395.
[56] Adams, R. P. “Identification of Essential Oil Components by Gas Chromatography/Mass Spectrometry.” Allured publishing corporation, Vol. 456, 2007.
[57] Jennings, W. Qualitative Analysis of Flavor and Fragrance Volatiles by Glass Capillary Gas Chromatography. Elsevier, 2012.
[58] NIST: https://webbook.nist.gov/chemistry/name-ser/.|Accessed 10 November 2022|.
[59] Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A. R., and

Steinhauser, D. “GMD@CSB.DB: The Golm Metabolome Database.” Bioinformatics, Vol. 21, No. 8, 2005, pp. 1635–1638. https://doi.org/10.1093/bioinformatics/bti236.

[60] Pubchem: Https://Pubchem.Ncbi.Nlm.Nih.Gov/.|Accessed 24 July 2023|.

[61] Kireev, A., Osipenko, S., Mallard, G., Nikolaev, E., and Kostyukevich, Y. “Comparative Prediction of Gas Chromatographic Retention Indices for GC/MS Identification of Chemicals Related to Chemical Weapons Convention by Incremental and Machine Learning Methods.” Separations, Vol. 9, No. 10, 2022, p. 265. https://doi.org/10.3390/separations9100265.

[62] Noviandy, T. R., Maulana, A., Sasmita, N., Suhendra, R., Irvanizam, I., Muslem, M., Idroes, G., Yusuf, M., Sofyan, H., Abidin, T., and Idroes, R. “The Prediction of Kovats Retention Indices of Essential Oils at Gas Chromatography Using Genetic Algorithm-Multiple Linear Regression and Support Vector Regression.” Journal of Engineering Science and Technology, Vol. 17, 2022, pp. 306–326.

[63] Babushok, V. I., Linstrom, P. J., and Zenkevich, I. G. “Retention Indices for Frequently Reported Compounds of Plant Essential Oils.” Journal of Physical and Chemical Reference Data, Vol. 40, No. 4, 2011, p. 043101. https://doi.org/10.1063/1.3653552.

[64] Otto, M. Chemometrics: Statistics and Computer Application in Analytical Chemistry. 2016.

[65] Holzinger, A. Big Data Calls for Machine Learning. In Encyclopedia of Biomedical Engineering, Elsevier, 2019, pp. 258–264. DOI: 10.1016/B978-0-12-801238-3.10877-3.

[66] MATLAB R2019b. Unvollständig ! Ergänzen !

[67] James N. M., Jane C. M., Robert D. M. Statistics and Chemometrics for Analytical Chemistry. 2018.

[68] Wise, B., Gallagher, N., Bro, R., Shaver, J., Windig, W., and Koch, S. “Chemometrics Tutorial for PLS_Toolbox and Solo.” Eigenvector Research, Inc, 2006, pp. 102–159.

[69] Jolliffe, I. T. “A Note on the Use of Principal Components in Regression.” Applied Statistics, Vol. 31, No. 3, 1982, p. 300. https://doi.org/10.2307/2348005.

[70] Dunn, K. Process Improvement Using Data. https://Learnche.Org/Pid/PID.Pdf. 381–383.

[71] LWR: https://Beginningwithml.Wordpress.Com/2018/07/02/5-Locally-Weighted-Linear-Regression/.|Accessed 26 April 2023|.

[72] Schaal C. S. and Atkeson C. G. “Robot Juggling: Implementation of Memory-Based Learning.” In IEEE Control Systems. 1994, Vol. 14, Issue 1, pp. 57-71. https://doi.org/10.1109/37.257895.

[73] Vapnik, V. N. The Nature of Statistical Learning Theory. Springer New York, New York, NY, 1995.
[74] Sebtosheikh, M. A., Motafakkerfard, R., Riahi, M. A., Moradi, S., and Sabety, N. “Support Vector Machine Method, a New Technique for Lithology Prediction in an Iranian Heterogeneous Carbonate Reservoir Using Petrophysical Well Logs.” Carbonates and Evaporites, Vol. 30, No. 1, 2015, pp. 59–68. https://doi.org/10.1007/s13146-014-0199-0.
[75] Üstün, B., Melssen, W. J., Oudenhuijzen, M., and Buydens, L. M. C. “Determination of Optimal Support Vector Regression Parameters by Genetic Algorithms and Simplex Optimization.” Analytica Chimica Acta, Vol. 544, Nos. 1–2, 2005, pp. 292–305. https://doi.org/10.1016/j.aca.2004.12.024.
[76] Saini, A. Support Vector Machine(SVM): A Complete Guide for Beginners. https://Www.Analyticsvidhya.Com/Blog/2021/10/Support-Vector-Machinessvm-a-Complete-Guide-for-Beginners/.|Accessed 16 June 2023|.
[77] The Intuition behind Kernel Methods: https://Www.Hashpi.Com/the-Intuition-behind-Kernel-Methods.|Accessed 10 March 2023|.
[78] Eigenvector. Evaluating Models: Hating on R-Squared. https://Eigenvector.Com/%EF%BF%BCevaluating-Models-Hating-on-r-Squared/.|Accessed 13 March 2023|.
[79] Eigenvector. Cross Validation. https://Wiki.Eigenvector.Com/Index.Php?Title=Using_Cross-Validation.|Accessed 13 March 2023|.
[80] Column Polarities: https://Mz-at.de/Downloads/Agilent_gc_csg.Pdf.|Accessed 25 April 2023|.
[81] Online Chemical Database: https://Ochem.Eu/Home/Show.Do.|Accessed 15 November 2022|.
[82] Joudaki, D., and Shafiei, F. “QSPR Models to Predict Thermodynamic Properties of Cycloalkanes Using Molecular Descriptors and GA-MLR Method.” Current Computer-Aided Drug Design, Vol. 16, No. 1, 2020, pp. 6–16. https://doi.org/10.2174/1573409915666190227230744.
[83] Olivero, J., Gracia, T., Payares, P., Vivas, R., Díaz, D., Daza, E., and Geerlings, P. “Molecular Structure and Gas Chromatographic Retention Behavior of the Components of Ylang-Ylang Oil.” Journal of Pharmaceutical Sciences, Vol. 86, No. 5, 1997, pp. 625–630. https://doi.org/10.1021/js960196u.
[84] Danishuddin, and Khan, A. U. “Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design.” Drug Discovery Today, Vol. 21, No. 8, 2016, pp. 1291–1302. https://doi.org/10.1016/j.drudis.2016.06.013.

[85] Babushok, V. I., and Linstrom, P. J. “On the Relationship Between Kovats and Lee Retention Indices.” Chromatographia, Vol. 60, Nos. 11–12, 2004, pp. 725–728. https://doi.org/10.1365/s10337-004-0450-2.

7. Appendix

Table 12: Information to the PLS model for DB1 exported from MATLAB:

Model Type	Xi and Yi Training	Preprocess-ing	Xi and Yi Test	Preprocess-ing	Cross validation	LV compo-nents
PLS calculated in MATLAB [66]	266 by 272	Pareto (Sqrt Std) Scaling	266 by 107	Pareto (Sqrt Std) Scaling	Venetian Blinds 10 splits and 1 sample per split	4

Table 13: Information to the LWR model for DB1 exported from MATLAB:

Model Type	Xi and Yi Training	Preprocess-ing	Xi and Yi Test	Preprocess-ing	Cross validation	Principal Compo-nents	Local Points
LWR calculated in MATLAB [66]	266 by 272	Pareto (Sqrt Std) Scaling	266 by 107	Pareto (Sqrt Std) Scaling	Venetian Blinds 10 splits and 1 sample per split	5	5

Table 14: Information to the SVM model for DB1 exported from MATLAB:

Model Type	Xi and Yi Training	Preprocess -ing	Xi and Yi Test	Preprocess -ing	SVM type	SVM kernel type	SVM optimal parameters	SVM: number of SVs	Cross validation
SVM calculated in MATLAB [66]	266 by 272	Pareto (Sqrt Std) Scaling	266 by 107	Pareto (Sqrt Std) Scaling	epsilon-SVR	radial basis function	cost = 100 epsilon = 0.1 gamma = 0.00031623	189	Venetian Blinds 10 splits and 1 sample per split

Table 15: Information to the SVM model for DB5 exported from MATLAB:

Model Type	Xi and Yi Training	Preprocess -ing	Xi and Yi Training	Preprocess -ing	SVM type	SVM kernel type	SVM optimal parameters	SVM: number of SVs	Cross validation
SVM calculated in MATLAB [66]	266 by 283	Pareto (Sqrt Std) Scaling	266 by 94	Pareto (Sqrt Std) Scaling	epsilon-SVR	radial basis function	cost = 100 epsilon = 0.1 gamma = 0.00031623	166	Venetian Blinds 10 splits and 1 sample per split

Table 16: Information to the SVM model for PEG exported from MATLAB:

Model Type	Xi and Yi Training	Preprocess -ing	Xi and Yi Training	Preprocess -ing	SVM type	SVM kernel type	SVM optimal parameters	SVM: number of SVs	Cross validation
SVM calculated in MATLAB [66]	266 by 232	Pareto (Sqrt Std) Scaling	266 by 85	Pareto (Sqrt Std) Scaling	epsilon-SVR	radial basis function	cost = 100 epsilon = 0.1 gamma = 0.00031623	155	Venetian Blinds 10 splits and 1 sample per split