ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΧΗΜΕΙΑΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ «ΧΗΜΕΙΑΣ»

ΕΙΔΙΚΕΥΣΗ «ΑΝΑΛΥΤΙΚΗΧΗΜΕΙΑ»

ΕΡΕΥΝΗΤΙΚΗ ΕΡΓΑΣΙΑ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

Ανάπτυξη και επικύρωση μοντέλων πρόβλεψης χρόνου ανάσχεσης για την ταυτοποίηση αναδυόμενων ρύπων σε περιβαλλοντικά δείγματαμε μη στοχευμένη σάρωση και τεχνικές φασματομετρίας μαζών υψηλής διακριτικής ικανότητας

REZA AALIZADEH

ΧΗΜΙΚΟΣ

ΑΘΗΝΑ

ΙΟΥΛΙΟΣ 2015

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

FACULTY OF SCIENCE

DEPARTMENT OF CHEMISTRY

GRADUATE PROGRAM «CHEMISTRY»

SPECIALIZATION «ANALYTICAL CHEMISTRY»

RESEARCH WORK OF SPECIALISATION DIPLOMA

Development and validation of wide-scope retention time prediction models to support suspect and non-target screening of emerging contaminants in environmental samples

REZA AALIZADEH

CHEMIST

ATHENS

JULY 2015

ΕΡΕΥΝΗΤΙΚΗ ΕΡΓΑΣΙΑ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

Ανάπτυξη και επικύρωση μοντέλων πρόβλεψης χρόνου ανάσχεσης για την ταυτοποίηση αναδυόμενων ρύπων σε περιβαλλοντικά δείγματα με μη στοχευμένη σάρωση και τεχνικές φασματομετρίας μαζών υψηλής διακριτικής ικανότητας

ΡΕΖΑ ΑΑΛΙΖΑΔΕΧ

Α.Μ.: 11306

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

ΝΙΚΟΛΑΟΣ ΘΩΜΑΪΔΗΣ, ΑναπληρωτήςΚαθηγητής, ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ :

ΚΩΝΣΤΑΝΤΙΝΟ ΣΕΥΣΤΑΘΙΟΥ

ΜΙΧΑΗΛ ΚΟΥΠΠΑΡΗΣ

ΝΙΚΟΛΑΟΣ ΘΩΜΑΪΔΗΣ

ΗΜΕΡΟΜΗΝΙΑΕΞΕΤΑΣΗΣ: 20/07/2015

ΠΕΡΙΛΗΨΗ

Την τελευταία δεκαετία η εφαρμογή της φασματομετρίας μαζών υψηλής διακριτικής ικανότητας συζευγμένη με υγροχρωματογραφία (LC-HRMS) έχει αναπτυχθεί ραγδαία λόγω της ικανότητας της τεχνικής αυτής να ανιχνεύει και να ταυτοποιεί πιθανές ή ύποπτες και άγνωστες ενώσεις στα περιβαλλοντικά δείγματα. Προκειμένου να επιτευχθεί αυτός ο σκοπός, πρέπει να αποκτηθούν οι πληροφορίες της ακριβούς μάζας και του ισοτοπικού προφίλ του ψευδομοριακού ιόντος, να πραγματοποιηθεί αξιολόγηση των φασμάτων MS/MS και ο χρόνος κατακράτησης να είναι ευλογοφανής έτσι, ώστε να επιτευχθεί η επιβεβαίωση της ταυτότητας μιας ένωσης. Στο πλαίσιο αυτό, αναπτύχθηκε μια υπολογιστική μεθοδολογία και τα αντίστοιχα μοντέλα πρόβλεψης για την κατανόηση της συμπεριφοράς του χρόνου ανάσχεσης ενός μεγάλου αριθμού αναλυτών που ανήκουν στην κατηγορία των αναδυόμενων ρύπων. Για το σκοπό αυτό χρησιμοποιήθηκε μια εκτεταμένη βάση δεδομένων που περιέχει την πληροφορία του χρόνου ανάσχεσης για 528 και 303 αναλύτες σε θετικό και αρνητικό ιοντισμό, αντίστοιχα, έτσι ώστε να επιτευχθεί η ανάπτυξη μοντέλων πρόβλεψης χρόνου ανάσχεσης με τη μέγιστη δυνατή περιοχή εφαρμογής (applicability domain). Η βάση δεδομένων διαχωρίστηκε σε ομάδα εκπαίδευσης (training set) και ομάδα ελέγχου (test set) με την τεχνική της συσταδοποίησης των K-κοντινότερων γειτόνων έτσι, ώστε να δομηθούν και να επικυρωθούν τα μοντέλα όσο αφορά την προβλεπτική τους ικανότητα. Το καλύτερο υποσύνολο μοριακών περιγραφέων (molecular descriptors) επιλέχθηκε με τη χρήση γενετικών αλγόριθμων (genetic algorithms), οι οποίοι είναι βασισμένοι σε υπολογιστικά εξελικτικά μοντέλα και μπορούν να επιλέξουν τους πιο αντιπροσωπευτικούς μοριακούς περιγραφείς για όλες τις ενώσεις σε σχέση με το υπό μοντελοποίηση πρόβλημα. Για τη μοντελοποίηση, χρησιμοποιήθηκαν οι εξής χημειομετρικές τεχνικές: πολλαπλή γραμμική παλινδρόμηση (MLR), νευρωνικά δίκτυα (ANNs) και η τεχνική Support Vector Machines (SVM) ώστε να συσχετιστούν τους, επιλεγμένους μοριακούς περιγραφείς με τον πειραματικά προσδιοριζόμενο χρόνο ανάσχεσης. Χρησιμοποιήθηκαν πολλές τεχνικές επικύρωσης, συμπεριλαμβανομένων των ακολούθων: τα κριτήρια Golbraikh-Tropsha, το πεδίο εφαρμογής βασισμένο στην ευκλείδεια απόσταση, ο συντελεστής r^2m, και ο συντελεστής συμφωνικήςσυσχέτισης (concordance correlation coefficient). Τα καλύτερα γραμμικά και μη γραμμικά μοντέλα για κάθε βάση δεδομένων που προέκυψαν χρησιμοποιήθηκαν στην πρόβλεψη του χρόνου

ανάσχεσης πιθανών/ύποπτων ενώσεων έτσι, ώστε να επιτευχθεί εξωτερική αξιολόγηση των μοντέλων. Γενικά, η προτεινόμενη πορεία είναι γρήγορη, αξιόπιστη, ελάχιστα δαπανηρή και μπορεί να εφαρμοστεί για τη μείωση των ψευδώς θετικών ευρημάτων κατά την εφαρμογή μεθόδων σάρωσης με LC-HRMS και την επιτυχή ανίχνευση και ταυτοποίηση άγνωστων ενώσεων σε περιβαλλοντικά δείγματα.

Περιοχή έρευνας: Αναλυτική Χημεία, Χημειομετρία

Λέξεις κλειδιά: χρόνος ανάσχεσης, σάρωση για ύποπτες ενώσεις, μη στοχευμένη ανάλυση, φασματομετρία μαζών υψηλής διακριτικής ικανότητας, μοριακοί περιγραφείς, τεχνική SVM

ABSTRACT

Over the last decade, the application of liquid chromatography - high resolution mass spectroscopy (LC-HRMS) has been growing extensively due its ability to identify a wide range of suspect and unknown compounds in environmental samples. However, certain information such as mass accuracy and isotopic pattern of the precursor ion, MS/MS spectra evaluation and retention time plausibility are needed to confirm its identity. In this context, a comprehensive workflow based on computational tools was developed to understand the retention time behavior of a large number of compounds belonging to emerging contaminants. An extensive dataset was provided, containing information for the retention time of 528 and 303 compounds for positive and negative electrospray ionization mode, respectively, to expand the applicability domain of the developed models. Then, the dataset was split into training and test employing k-nearest neighborhood clustering, so as to build and validate the models' internal and external prediction ability. The best subset of molecular descriptors was selected using genetic algorithms which is based on the evolutionary computations, and could result in representative selection of descriptors. Multiple Linear Regression, Artificial Neural Networks and Support Vector Machines were used to correlate the selected descriptors with the experimental retention times. Several validation techniques were used, including Golbraikh-Tropsha acceptable model criteria's, Euclidean based applicability domain, r^2_m, concordance correlation coefficient values to measure the accuracy and precision of the models. The best linear and non-linear models for each dataset were derived and used to predict the retention time of suspect compounds in a wide-scope survey as the evaluation data set. Overall, the proposed workflow was fast, reliable, cost-effective and can be employed as an effective filtering tool for decreasing false positives of wide-scope HRMS screening of environmental samples.

SUBJECT AREA: Analytical Chemistry, Chemometrics

KEY WORDS: Retention Time, Suspect Screening, Non-target Screening, High Resolution Mass Spectrometry, Molecular Descriptors, Support Vector Machines.

ACKNOWLEDGMENT

I foremost, would like to express my deepest gratitude to Associate Prof. Nikolaos Thomaidis for his guidance, support, encouragements and patience during this research. I also would like to appreciate my family for their patience and continuous support. I deeply thank Ms. Anna A. Bletsou and Dr. Pablo Gago Ferrero for their supports during this project. Finally, I am thankful to all who supported me unconditionally to fulfill my tasks. **I would like also to acknowledge financial support from the State Scholarships Foundation of Greece (I.K.Y).**

CONTENTS

LIST OF FIGURES

LIST OF TABLES

< This page initially left blank>

CHAPTER 1

QSRR AS SCREENING TOOLS

1.1 Target, suspect and non-target screening in reversed phase liquid chromatography-high resolution mass spectroscopy (RP-LC-HRMS)

Over the last decades, thousands of substances with potential risks for human and aquatic life are disposed in the environment. Their rapid and accurate identification is emerged as an important field in both analytical and environmental science. The evolution of high resolution mass spectroscopy coupled with liquid chromatography has opened up a new opportunity for the identification of polar compounds in complex environmental samples. With this technique, many compounds with a great variety of of functional groups and polarities, which are not well identified *via* gas chromatography (GC), can be detected effectively. Identification procedures in LC-HRMS were detailed into three categories including target analysis (with reference standards), suspect screening (with suspected substances based on prior information but no reference standards) and finally non-target screening (no prior information, no reference standards)[1].

1.1.1 Target Analysis

For a successful and full target analysis, a reference standard is required to determine the concentration of target in sample, and also comparing and matching the observed retention time (t_R) and tandem mass spectrum (MS/MS). Target analysis is relied on purchase of reference standard for quantification and confirmation. The use of an isotopic labeled internal standard facilitates the analysis but it is not always available. Target analysis can be performed by following the procedure explained in figure 1.

1.1.2 Suspect screening

Suspect screening with LC-HRMS relies on accurate mass and isotope information for the precursor ion. Compounds that are expected to exist in the samples (suspects or suspected compounds), can be screened using the exact mass of their expected ions in negative ($[M–H]^-$) or positive ($[M+H]^+$) electrospray ionization mode (ESI). Exact mass screening methods are computationally rapid and many masses can be screened in a

given sample, but the risk of false positive results is high. Additional information is needed to reach a tentative identification, apart from the mass accuracy and isotopic fit, such as evaluation of the MS/MS spectra and retention time plaucibility. However, gathering of evidence and conformation of the detected masses is still a time consuming task. Calculating the retention time for the suspects list and comparing it to the observed retention time for the observed peaks could be an efficient filtering tool over the confirmation or rejection of the suspected substances. The general procedure for performing suspect screening in LC-HRMS is shown in figure 1.

1.1.3 Non-target screening

Non-target screening involves m/z ratios (ions, usually called “features”) that are detected in the sample, and there is not any a priori information available for the observed peaks. It is often difficult to fully identify the unknown peaks with no guarantee of a successful outcome. The procedure for the non-target screening is shown in figure 1. First, automated peak detection is used by exact mass filtering from the chromatographic run. Next, elemental formula can be assigned to the exact mass of interest and finally searching the database for hits. Through the validated computational models based on quantitative structure retention time relationship (QSRR), their retention times can be calculated and those matched can be investigated further with MS/MS fragmentation to give the most possible substances.

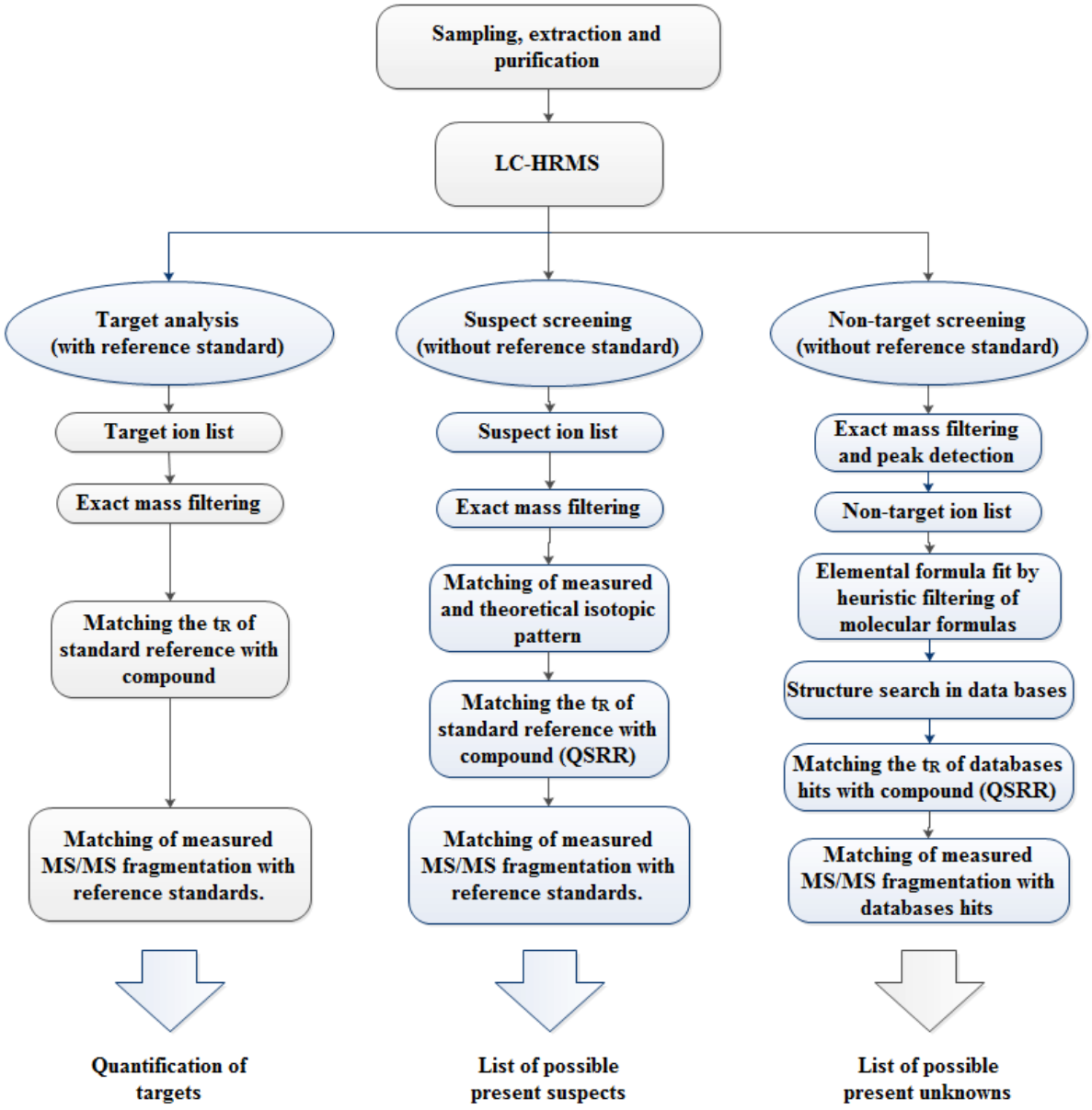


Figure 1: Procedures for target analysis, suspect and non-target screening[1]

1.2 Quantitative structure-retention time relationship (QSRR)

In 1977, the first three publications were published with the aim of finding correlation between chemical structures and their chromatographic behavior which is now called QSRR. Since then, a large number of efforts were made to derive robust mathematical models that not only predict the retention time of compounds, but also explain the chemical

features affecting retention time values. Several good models previously were reported for gas-chromatographic (GC) retention based on chemical features derived from molecular graphs and quantum chemical energy-related[2-4]. Generally, QSRR results for liquid chromatographic (LC) retention data present lower statistical quality than those reported for GC and this is due to the effect of chromatographic conditions such as stationary phase, column type, separation conditions and elution mechanism at different molecular level over retention behavior of compounds[5].Beside the lack of ability for inclusion of these effects to QSRR based models, a certain workflow that enriches the applicability domain of models for application of different type of compounds was not proposed[6]. Use of a data set consisted of large chemical diversity (i.e. increasing the chromatographic effects over retention time values) would also unable the models to find the rational chemical features and thus insufficient interpretations[6]. By growth of chemometics and introduction of new type of molecular features for 3D structure of molecules, capabilities of models were increased to handle dataset with abnormal retention time. Recent advances in both chromatographic science and chemometics caused a revolutionary enhancement of identification and interpretation of results however modeling of retention time in LC-HRMS is still a challenging work due to complexity of chromatographic and instrumental system[7, 8]. It is a need of computational tools such as QSRR to help the identification of unknown substances in the environment[1, 9]. Three major steps should be followed after the preparation of the initial dataset, for a correct modeling:

- Geometry optimization of chemical structures and calculating molecular descriptors[10]
- Molecular descriptors selection and their modeling[11, 12]
- Defining the applicability domain with certain method of outlier detection techniques[13]

These steps are explained further in more details in following sections.

1.2.1 Chemical structures and their geometries

A true knowledge of geometry of molecular structure can provide better interpretation of its stability, interactions with its environment, and several molecular properties such polar surface area, ionization, isomeric states and conformation of compounds can be derived

accurately. Since the development of 3D-based molecular descriptors, it is now so important to optimize a chemical structure before deriving the molecular descriptors so as to distinguish between similar compounds. Optimization of chemical structures can be done with four major methods including:

- Molecular mechanics: Molecular mechanics force field (MM+) is an extension of MM2 force field developed by Allinger and co-workers [14, 15]. This method is designed for small organic molecules and also can be carried out for geometry optimization of peptides. Molecular mechanical force field uses the equations of classical mechanics to describe the potential energy surfaces and physical properties of molecules. One component of a force field is the energy that is originated from compression and stretching a bond. Unlike quantum mechanics, molecular mechanics does not treat electrons explicitly and thus it cannot explain bond formation and bond breaking. This method is also lack of accuracy for a system by which electronic delocalization or molecular orbital interactions plays a major role in determining geometry or properties.
- Semi-empirical methods (AM1): They use a certain number of experimental data throughout the calculation. For example, bond lengths of a specific type will have a fixed value independently of the system (C=C bond will always be taken as 134 pm, for example). This dramatically speeds up computational time, but in general is not very accurate. Usually, semi-empirical methods are used for very big systems, since they can handle large amount of calculations.
- Hartee-Fock (HF): Quantum mechanics calculations use either of two forms of the wave function: Restricted Hartree-Fock (RHF) or Unrestricted Hartree-Fock (UHF). The RHF wave function can be used for singlet electronic states, such as the ground states of stable organic molecules. The UHF wave function is most often used for multiplicities greater than singlets. Hartee-Fock can be performed based on various biases set (Configuration Interaction (CI), Møller-Plesset (MP) perturbation theory) to provide the UV spectra, energy of excited states, breaking of bonds and change of spin coupling. The electronic state of molecules and energy of ground and excited state can be obtained with high accuracy for large organic compounds with many

orbitals in a small energy range. However the major drawback of HF method is the exclusion of electron correlation.

- Density functional theory (DFT): DFT methods are becoming more and more popular because the results obtained are comparable to the ones obtained using Hartree-Fock method, however CPU time is drastically reduced. DFT differs from methods based on HF calculations in the way that it is the electron density that is used to calculate the energy instead of a wave function. DFT can optimize the geometry of large groups of compounds such as nanotubes, semiconductors, and complexes with high accuracy depending on the biases set (B3LYP, PW91, VWN, etc.) that is being used. The application of above methods for different optimization purposes is listed in Table 1.

Table 1: The general techniques for optimization of chemical structures

Task	Molecular mechanics	Semi-empirical	Hartee-Fock	Density functional theory
Geometry (organic)	C	G	G	G
Geometry (metals)	-	G	P	G
Transition-state geometry	-	C	G	G
conformation	G	P	G	G
Thermochemistry	-	P	C	G

G: Good; C: good with cautious application; P: Poor

Polar surface area and energy of equilibrium geometry for a compound () based on above methods were calculated and shown in figure 2 to show how molecular geometry effects the properties of molecule.

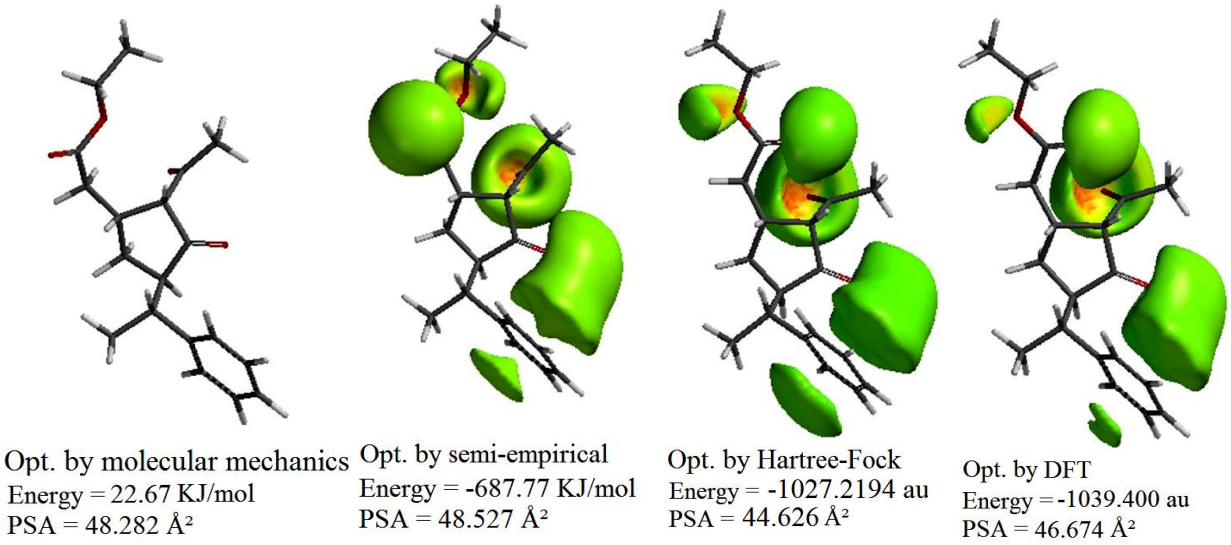


Figure 2: Polar surface area and energy of equilibrium geometry based on different geometry optimization techniques.

As it can be seen, use of quantum mechanics methods (HF, DFT) for obtaining the correct geometry of chemical structures would provide the lowest energy level and thus a stable form of compound rather than that of derived based on molecular mechanics. However, HF and DFT methods are very time consuming and should be used when there is a need for electronic state of molecules.

1.2.2 Molecular descriptors for chromatographic retention

Chemical structures and their properties can be used to get the retention time with acceptable accuracy. Effect of chromatographic conditions such as content of stationary phase and mobile phase over retention time of a compound could help in calculation of molecular descriptors more precisely. In normal phase silica as stationary phase, polar compounds will bond to the stationary phase and thus they will appear at higher retention time and thus in this case, polarizability of compounds should be calculated. Since pH affects the charge of the stationary phase and of the compounds in the first place, a molecular feature that incorporates the pH effects on logD should be considered. Figure 3 shows the chemical structure of silica based stationary phase. For the reversed phase chromatography, since the stationary phase is neutral, charge type descriptors have less effect on interpretation of retention time; The major contribution of charge descriptors are over the ionization pattern of compounds and correspondingly the functional groups. They

are also affecting the LogD values indirectly as pk_a and pH are varying. However hydrophobicity has significant influence for retention time behavior of compounds. Considering these prior information about modeling of retention time, more chemical properties are required to perform successful retention time prediction. Therefore, Various molecular descriptors should be calculated, includingconstitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelation, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, Radial Distribution Function (RDF) descriptors, 3D-MoRSE (3D Molecular Representation of Structure based on Electron diffraction) descriptors, WHIM (Weighted Holistic Invariant Molecular) descriptors, GETAWAY (geometry, topology and atoms-weighted Assembly) descriptors, functional group counts, atom-centred fragments, charge descriptors, molecular properties, 2D binary fingerprints and 2D frequency fingerprints [16-19]. Among the above descriptors, the constitutional descriptors are referring to atomic or molecular properties and are independent of the overall molecular connectivity. These types of descriptors encode the size of molecules and chemical properties. Geometrical descriptors are presenting features of the molecular geometry e.g. distances between particular points on the molecular surface and distances between given chemical groups. Topological descriptors reflect the type and the connection of atoms in the 2D space [20].

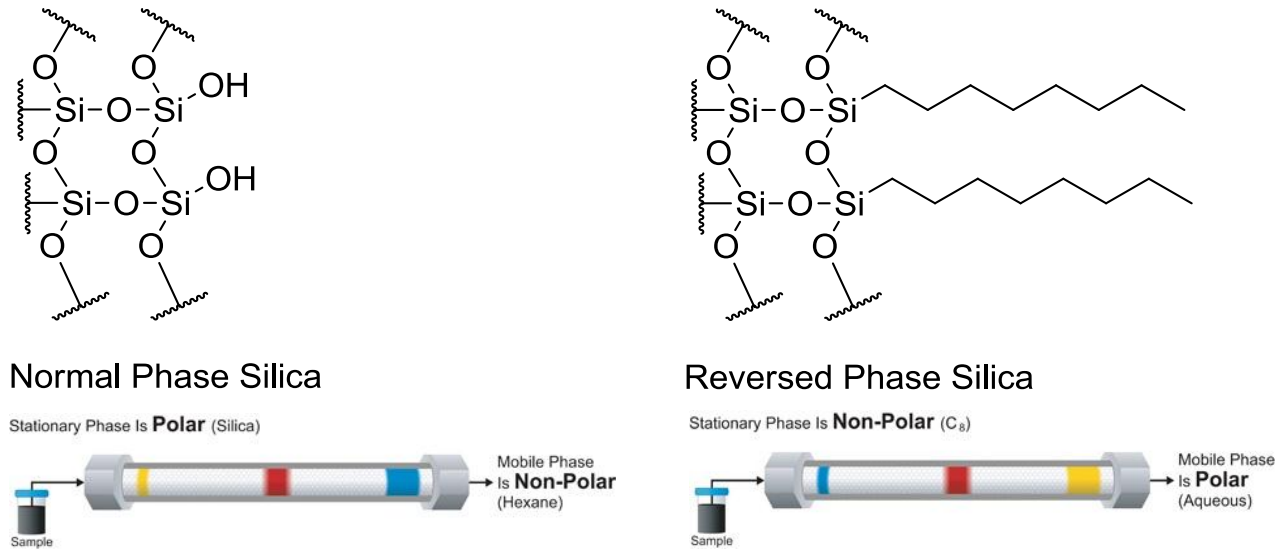


Figure 3: The stationary phases in normal and reversed phase chromatography

1.2.2.1 2D-Molecular descriptors

Several groups of molecular descriptors can be calculated based on 2D chemical structures which are independent to conformation of chemical structure. Some of the important descriptors belonged to the 2D-molecular descriptors are as follows:

- Topological charge indices: topological charge indices were proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule [21].
- Connectivity indices: connectivity indices are among the most popular topological indices and are calculated from the vertex degree δ of the atoms in the H-depleted molecular graph. The Randic connectivity index was the first connectivity index proposed[22]; it is also called connectivity index or branching index which can describe the bond order, intermolecular accessibility[23] and molecular branching [24].
- Molecular properties: these descriptors are representing the properties of chemical structures such as hydrophobicity, molar refractivity, polar surface area, unsaturation index and octanol-water partition coefficient (logP).
- Edge adjacency indices: These descriptors are derived from molecular graph and denoting the bond connectivity and matrix with their representative weighted properties in between graph edges[22].
- Walk and path counts: Atomic path/walk indices are described for each atom as the ratio between atomic path count and atomic walk count for the same length. The number of paths in a molecule is bounded and determined by the molecule diameter, whereas the number of walks is unbounded. However, being interested only in quotients, the walk count is terminated when it exceeds the maximum allowed length of the corresponding path. Molecular path/walk indices are explained as the average sum of atomic path/walk indices of equal length [25]. As the path/walk count ratio is independent of molecular size, these descriptors can be considered as shape descriptors.

1.2.2.1 3D-Molecular descriptors

In contrast to 2D-molecular descriptors, a method of optimization affects the results of 3D-molecular descriptors significantly. Therefore, in case of high similarity and also isomers for group of compounds, it is important to derive the 3D-molecular descriptors for quantitative analysis purposes. Four major groups of descriptors that are largely being used are reported below:

- WHIM descriptors: these are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes [22]. WHIM descriptors are built in such a way as to capture relevant molecular 3D information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. The algorithm consists in performing a Principal Components Analysis (PCA) on the centered Cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighting schemes for the atoms:

$$s_{jk} = \frac{\sum_{i=1}^{A} w_i \left(q_{ij} - \bar{q}_j\right)(q_{ik} - \bar{q}_k)}{\sum_{i=1}^{A} w_i} \qquad (Eq.1)$$

  where s_{jk} is the weighted covariance between the j_{th} and k_{th} atomic coordinates, A is the number of atoms, w_i the weight of the *i*th atom, q_{ij} and q_{ik} represent the *j*th and *k*th coordinate of the *i*th atom respectively, and $\bar{q}$ the corresponding average value.
- RDF descriptors: Radial distribution function in this form meets all the requirements for the 3D structure descriptors. It is independent of the atom number (i.e. the size of a molecule), and is unique regarding the three-dimensional arrangement of the atoms and is also invariant against the translation and rotation of the entire molecule. Additionally, the RDF descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space (e.g. to describe the steric hindrance or the structure ⁄ activity properties of a molecule).
- GETAWAY descriptors: these descriptors have recently been proposed as chemical structure descriptors derived from a new representation of molecular structure, the Molecular Influence Matrix (MIM), denoted by H and defined as follows:

  $$H = M \cdot (M^T \cdot M)^{-1} \cdot M^T \qquad (Eq.2)$$

where M is the molecular matrix consisting of the centered Cartesian coordinates of atoms of a compound in optimized geometry. T is refereeing to transposed matrix. For different types of GETAWAY, H values can be coupled with molecular properties as weight factor to show the effect of molecular properties in specific topological and geometrical region of molecular graph [26].

- 3D-MoRSE descriptors: 3D-MoRSE (Molecular Representation of Structures based on Electronic diffraction) descriptors were introduced in 1996 by Schuur, Selzer and Gasteiger with the motivation for encoding 3D structure of a molecule by a fixed number of variables [27, 28]. Indeed, the most obvious way to present 3D structure is its representation within cartesian or internal coordinates. Simplifying the equations used in electron diffraction studies, the function was calculated as:

$$I(S) = \sum_{i=2}^{N} \sum_{j=1}^{i=1} A_i \, A_j \frac{\sin s r_{ij}}{s r_{ij}} \qquad (Eq.3)$$

where *s* is the scattering parameter, r_{ij} is the Euclidean distance between *i*th and *j*th atoms, *N* is the total number of atoms and A_i and A_j are different atomic properties used as weights. Each term of this function depends on distance and thus may be viewed as a radial basis function itself. Assigning to *s* integer values in the range of 0–31 $Å^{-1}$, 32 values of function 1 can be calculated[29].

1.2.3 Dataset division

1.2.3.1 Principle Components Analysis (PCA)

Supposing x is set of compounds in raw and p is molecular descriptors; the aim of PCA is to derive the variances of the p or correlations between the variables of p. Unless p is small, or the structure is very simple, it will often not be very helpful to simply look at the p variances correlations or covariances. An alternative approach is to look for a few (<<p) derived variables which preserve most of the information given by these variances and correlations or covariances. The main idea about principle component analysis (PCA)[30]is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. The reduction task could be achieved by transforming to a new set of variables termed principal components in which are uncorrelated, and ordered in a way that the first few retain most of

the variation present in all of the original variables. The distribution results of data points can be plotted to see how similar and scattered the chemical structures (score plot) and how the molecular descriptors distributed relative to the molecules (loading plot).

1.2.3.2 K-nearest neighborhood (kNN)

k-nearest neighborhood is a hierarchal clustering technique that separates data by putting them into clusters. In this method, the analysis begins with each case in its own separated cluster and then identical clusters combine each other, this continues until just one cluster left [31, 32]. In order to combine the cluster accurately in each time a measure of similarity between cases is required and this can be achieved by using an appropriate metric [33]. The most used similarity metrics is Euclidean distance[34] which is calculating as follows:

$$d_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2} \qquad (Eq.4)$$

d_{ij} is distance score between two different compounds (x_i and x_j). The results of hierarchical clustering are presented as a dendrogram which can be used to do data mining so accurately.

1.2.4 Molecular descriptors selection

Since there is less information about the parameters that would affect retention time behavior of compounds, there is a need to use variable selection tools for deriving these molecular features.

1.2.4.1 Stepwise variable selection (SW)

Stepwise selection technique was a well-known and simplest method for identifying the right number of variables in data matrix that the procedure includes a regression models for its selection base [35, 36]. The stepwise variable selection technique is performing by forward selection (figure 4) and back elimination rule, where the variable possessed the highest correlation value with response (experimental data) is being selected, and based on the regression model, its regression coefficient is being calculated. Each selected variable (here the molecular descriptor) is then tested using F-test [36-39] to see its significance and contribution to the model, where if it improves the model it is included in

the model. This procedure is called forward selection. However, if the selected variable does not contribute in improvement of the model is excluded from the set of significant variables and is eliminated from the model. This step is called backward elimination step [35, 38]. The two steps were continuing until no further improvement is observed by excluding or including the variables. The only disadvantage of this technique is over-fitting since the selection is based on data fitting. To prevent this problem, cross-validation should be employed to evaluate the predictive ability of the proposed model [37-39].

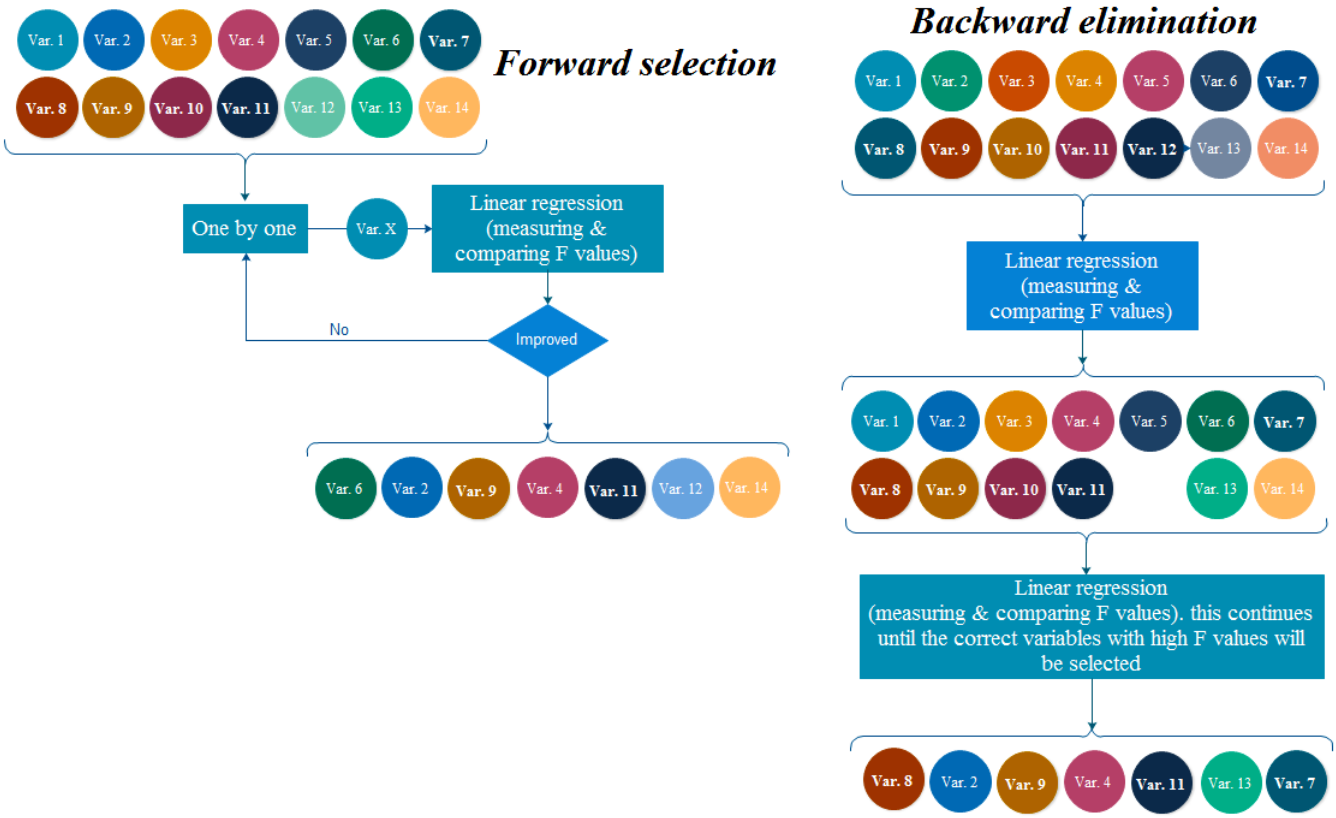


Figure 4: The procedure of forward and backward variable selection for stepwise technique

1.2.4.2 Genetic Algorithms (GAs)

Apart from the stepwise variable selection algorithm, one of the most accomplished techniques for this purpose is genetic algorithms (GAs) [40, 41] which are inspired from

natural evolution concepts by which the fittest species have high chance of survival. The GA technique starts with binary coding of molecular descriptors values for each compound to permit the mathematical treatment of “chromosomes”. “Chromosomes” are randomly selected group of molecular descriptors that the descriptors inside these “chromosomes” are called “genes”. The total number of “chromosomes” is indicating the population (generally lies between 50 and 500) which is depending on the dimension of the problems. These “chromosomes” are evaluated based on the *fitness function* (here is the correlation coefficient-leave one out cross validation (Q^2_{LOO})), so that if chromosomes couldn’t meet the cut off criteria, they are being stopped from spreading for the next generations. Next, the survived “chromosomes” are reproducing new number of population, and the probability level of each “chromosome” is calculated based on its outcomes associated with the taken responses. The best number of “chromosomes” would be selected finally by their higher probability that results in better response. The cross-over technique is then being applied to these “chromosomes” to pair them in a new generation for deriving the most effective “genes” in “chromosomes”. Finally, mutation which causes to impose values that are not tried for each descriptor is being applied to newly derived generation [40]. The reproduction and mutation of “chromosomes” continue until the best number of descriptors in a “chromosome” is selected within the GAs iteration of generation. This process is shown in Figure 5.

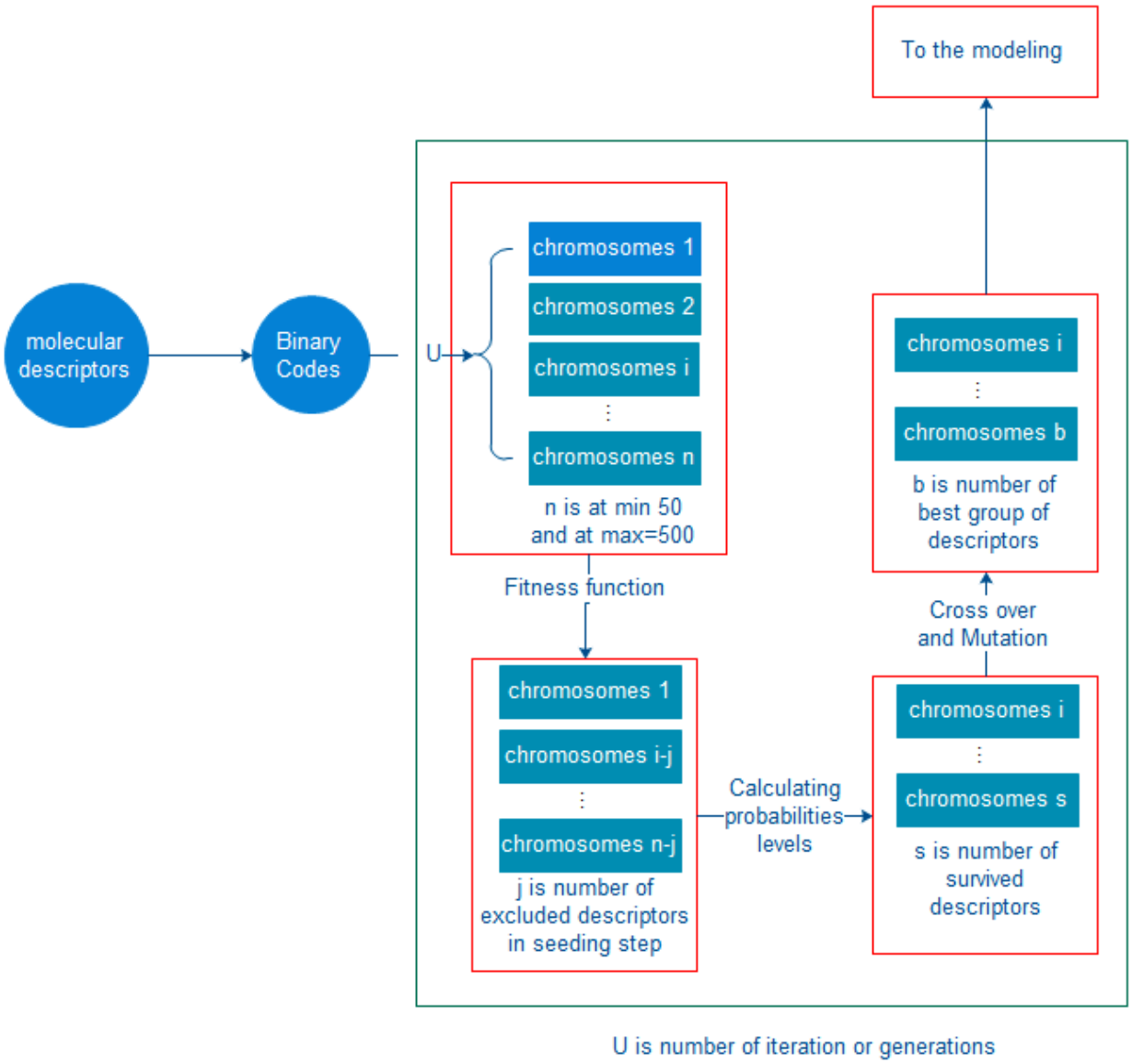


Figure 5: Procedure of Genetic algorithms as variable selection tool

1.2.5 Modeling techniques

1.2.5.1 Multiple Linear Regressions (MLR)

Multiple linear regressions (MLR) method is one of the most used linear models in QSRR. To derive a MLR model, the number of molecules in data set should be five times higher than the number of selected descriptors (the descriptors should be orthogonal). A low number of descriptors is of interest in order to minimize the information overlap in descriptors. In this work, to obtain the best linear model, the statistical parameters (R^2 and

Q^2 values) were considered. The MLR model provided a linear equation which is linking the structural features to the retention times of the compound:

$$Rt = a_0 + b_1x_1 + \cdots + b_nx_n \qquad (Eq.5)$$

where a_0 is the intercept and the b_i is regression coefficients of the selected descriptors x_i .

1.2.5.2 Artificial Neural Networks (ANN)

Artificial neural networks are computational models inspired by the biological nervous system. The feed forward artificial neural network with back-propagation of error algorithm is the most known method to derive an ANN nonlinear model [42, 43]. The input for the model generation is the selected variables (descriptors) based on genetic algorithm's selection. The initial weights were randomly chosen between 0 and 1 [44]. Optimization of the weights and biases is performed based on the resilient back-propagation algorithm[45]. The complex step in performing the ANN model is identifying the correct hidden layers to generate the QSRR model [44]. Generally, a three-layer network with a sigmoidal transfer function can be designed for simple modeling purposes[46]. To obtain the correct nodes in the hidden layers, RMSE values should be considered for both test and training sets, and the nodes with the lower RMSE can be selected as final output[44]. The high number of iterations (20000) would also decrease the error of models. However, in most cases, increasing the iterations would cause to increase the value of standard error of prediction set started and therefore, over-fitting occurs[3]. The increased numbers of iterations have several advantages: the architecture of the generated ANN is correctly designed, and the descriptors that appeared in the model have been effectively selected. In a sample usage, a data set should be divided into three groups using principle component analysis or clustering techniques, separately. In our particular problem, a training set, a validation set and a prediction set for negative and positive ESI with the proportional of 60%:20%:20%, respectively, should be produced. The training and validation sets are for building the predictive model and the prediction set is for evaluating the external prediction accuracy of the generated model [3]. For obtaining the best model, despite the control of RMSE, R^2 and mean percentage deviation (MPD) values for the results of each node analysis, some external statistical analyses should be considered so as to select the number of nodes

correctly. The neural networks can be implemented using Neural Network Toolbox for MATLAB 6.5.

1.2.5.3 Support Vector Machines (SVM)

Support vector machines (SVM) [47] are non-linearly correlating the selected molecular descriptors with the observed retention time values. In SVM, the dataset consists of the molecular features are transferred to high dimensional space using Kernel function leading to handle the non-linear problem by using linear regressions in derived feature space [48]. The general advantages of SVM over conventional neural networks are its capability of avoiding the local minima and automatically derivation of network topology structure. The linear regression in feature space is given below:

$$f(x) = \omega \cdot \phi(x) + b \qquad (Eq.6)$$

where ω and b are the slope and the offset for the regression line, respectively. x is the input dataset and ϕ is the mapping function (kernel) that can map the input dataset in higher dimension. To obtain regressions function (to calculate ω and b), the risk function (ε-insensitive loss function) should be minimized so that the function could be as flat as possible:

$$J_{SVM}(C) = \frac{1}{2}\|\omega\|^2 + C\frac{1}{2}\sum_{i}^{n} L_{\varepsilon}(d_i, y_i) \qquad (Eq.7)$$

$$subjectsto L_{\varepsilon}(d,y) = \begin{cases} |d-y| - \varepsilon, & |d-y| \geq \varepsilon \\ 0 & otherwise \end{cases} \qquad (Eq.8)$$

where $C\frac{1}{2}\sum_{i}^{n} L_{\varepsilon}(d_i, y_i)$ is the empirical error and it is calculated from Eq.7, $\frac{1}{2}\|\omega\|^2$ is termed regularized parameter and ε is the tube (or vector) size. Here, C is a regularization constant which is determining the trade-off between regularization parameter and empirical error. The positive slack variables (ξ and ξ^*) can be amended to Eq. 9as follows:

$$J_{SVM}(\omega, \xi^*) = \frac{1}{2}\|\omega\|^2 + C\sum_{i}^{n}(\xi_i + \xi_i^*) \qquad (Eq.9)$$

Finally, introduction of Largrange multipliers (a_i) and (a_i^*) would result in modification of Eq. 9as below:

$$f(x, a_i^*) = \sum_{i=1}^{n} (a_i - a_i^*) K(x, x_i) + b \qquad (Eq. 10)$$

here K is the kernel function that consisted of linear, polynomial, radial basis function and splines. Here, to develop a SVM model, Gaussian radial basis function (Eq.11) was employed:

$$K(\overline{x_\iota}, \overline{x_J}) = \exp\left(-\gamma \left\| \overline{x_\iota} - \overline{x_J} \right\|^2\right), \overline{x_\iota}\ and \overline{x_J}\ are\ independent\ parameters \qquad (Eq. 11)$$

1.2.6 Validation of the models

1.2.6.1 Internal validation criteria

To evaluate the strengths and goodness of the model, the coefficient of multiple determinations was used. R^2 value calculates the proportion of the variation in the response where obtained as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{I} (y_i - \bar{y})^2} \qquad (Eq. 12)$$

where y_i is the observed property/activity (here is the experimental retention time), $\bar{y}$ is the mean value of the experimental data and $\hat{y}_i$is the calculated retention time. The R^2 value higher than 0.5 and near 1.0 indicates the acceptable predictive ability of the model. Generally, theR^2 value can change (either increased or decreased values) by adding extra variables to the model. Therefore, this problem can be solved considering the adjusted R^2 values (R^2_{adj}):

$$R^2_{adj} = \left[1 - (I - R^2)\left(\frac{I - 1}{I - n - 1}\right)\right]^{1/2} \qquad (Eq. 13)$$

In this equation, the number of calibration objects is (*I*), and the number of the selected descriptors for model is (*n*). The statistical significance of the proposed model can also be given by null hypothesis where this implies that all the descriptors in the model beyond the constant value are required for modeling. To derive the given null hypothesis, comparison of the F-value can be used as follows:

$$F = \frac{(n - k - 1) \sum_{i=1}^{n} (y_i - \bar{y})^2}{k \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (Eq. 14)$$

where *n* is the number of the compounds in the dataset and *k* is the number of descriptors. The higher the F –value becomes, the greater the probability that the equation is significant. Therefore, procedure results in selection of appropriate and relevant descriptors if its null hypothesis rejected by having higher F values. Another important statistical parameter that is used in both linear and non-linear methods to validate the outcome of the derived models is the root mean square error (RMSE), where the lower RMSE value indicates the less error generated by built models, and thus, the model can be accepted for prediction purposes. The RMSE value is calculated as below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (Eq.15)$$

The most important statistical parameter that is showing the validation of models in multiple linear regression modeling (linear regressions) is the cross-validation correlation coefficient which is calculated as leave-one-out compound principle. In every calculation process for obtaining Q_{LOO}^2 value, one of the compounds in the dataset is being excluded from the model and its activity is calculating from the proposed model. This process is continued until all available compounds in the data matrix are excluded once, and their activities are being predicted by the model. Therefore, this technique is a good indicator of the strength of the derived models. A robust model should implement highQ_{LOO}^2 value. This value can be calculated as follows:

$$q_{LOO}^2 = r_{cv}^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2} \qquad (Eq.16)$$

Further, the external predictive ability of the constructed model can be assessed by modified r^2 value (Eq. 17) and the concordance correlation coefficient (Eq. 18) methods evaluating both accuracy and precise[49]. Concordance correlation coefficient (CCC) evaluates the degree to which pairs of observations fall on the 45° line through the origin:

$$r_m^2 = r^2\left(1 - \left|\sqrt{r^2 - r_o^2}\right|\right) \qquad (Eq.17)$$

where r^2 and r_o^2 are squared correlation coefficients between the observed and predicted retention time value of the test set compounds with and without intercept, namely.

$$CCC = \rho C_b \qquad (Eq.18)$$

where ρ is the Pearson correlation coefficient, and measures how far each observation deviates from the best-fit line. Thus, the ρ value is a measure of the precision, and C_b is the bias correction factor which calculates how far the best-fit line deviates from the 45° line through the origin, and therefore, it is a measure of the accuracy.

1.2.6.1 External validation criteria

After the development of the models, it is highly needed to apply methods for evaluating the external predictive ability of the models. There are several external validation methods which can be used. However, there is an important work performed by Tropsha who discussed mainly the importance of the model validation [50]. As discussed, refereeing to Q^2_{LOO} and R^2 values for presenting the predictive ability of a built model is not enough in all cases, and the predictive power of a model can be investigated only based on the prediction results of the test set compounds. Therefore, an accurate and valid model can be established only based on model validation procedure consisted of compounds which were not included in the model development. Tropsha suggested that to simulate the use of QSAR/QSPR models, there should be another set of compounds with known activities/properties that are not included in either training or test sets. Then, by the proposed models, the activities of the built models are being predicted. In general, the size of the external validation set should be about 15%–20% of the entire dataset, and the remaining part of the dataset is called modeling set. Golbraikh and Tropsha acceptable model criteria's can also be a sufficient tool [51] to verify the predictive ability of the developed models. They introduced four conditions for accepting a model, as follows:

- Q^2_{LOO} value must be higher than 0.5
- R^2 value must be higher than 0.6
- $R_0^2 - R_0'^2/R^2 < 0.1$ and $0.85 < K' < 1.15$ or $R^2 - R_0^2/R^2 < 0.1$ or $0.85 < K < 1.15$
- $R_0^2 - R_0'^2 < 0.3$

where R is the correlation coefficient between the predicted and observed values; R_0^2 is the coefficient of determination (correlation of predicted *versus* observed values with an intercept of zero), and $R_0'^2$ is the correlation between observed *versus* predicted values for regressions through the origin; K is the slope and K′ is the slope of the regression lines through the origin [51].

1.2.7 Applicability domain

Outlier detection and defining applicability domain is an important part of the QSRR [52]. It was suggested that in a case of QSRR, it is better to use manual outlier detection process by considering the information of both experimental and chemometrics tools [52]. Application of different automated outlier detection tools could also decrease the inaccurate outlier treatment and cause better data analysis for a large dataset.

1.2.7.1 Williams plot

Williams plot is a robust method, not only to measure the applicability domain of any proposed model, but also to detect the outliers presented in the model [53]. It is based on the leverage and standardized residual values. Leverages can be calculated from the molecular descriptors as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i where h^* = 3(p+1)/n \qquad (Eq. 19)$$

where X is the molecular descriptors matrix, T is an indicator of the training set, x_i is the descriptor vector for each molecule, n is the number of the compounds in the training set, p is the number of the molecular descriptors as modeling variables, and h* is the warning leverage value and it is a cut-off value to show that the chemical structures outside of this value are outliers due to their high dissimilarity of chemical structures [53]. The commonly used cut-off value for standardized residual is ±3δ where it covers 99% of normally distributed data. Compounds which locate outside of this cut-off value will be considered as outliers due to the abnormal response observed (here, wrong retention times).However, compounds outside of the leverage cut-off value but inside the standardized residual limits are considered as good leverages which can be included in the modeling results.

1.2.7.2 Euclidean based applicability domain

Euclidean distance can be measured for training and test set, and, then, the mean distance for the test set compounds, normalized on the mean distance of training set *versus* observed t_R, can be obtained to show how the diversity of chemical structures behaves toward the t_R[54]. Test set compound outside the cut-off value of 1.0 (calculated by

normalization of mean distance of training set), are considered to be outside of the applicability domain of the model, and the training set is not representative for this compound in the used test set.

CHAPTER 2

LITERATURE REVIEW OF DEVELOPED MODELS FOR LC

There are a few number of articles published to explain the retention time behavior of molecules in LC-HRMS. Table 2 presents a short review of the published articles on this topic. Former studies such as No 1 and 10 presented a QSRR model for predicting retention time of some forbidden and anti-doping substances based on optimized geometry of chemical structures. However, the studies are suffering from future applications due to narrow applicability domain as well as lack of outlier detection studies. Several previously reported studies such as No 2, 3, 4, 14 and 9 are also created based on molecular descriptors that were not selected by a validated procedure such as genetic algorithm and thus resulted in the lack of fitness both internally and externally for future applications. Although, the major issues that have been not discussed extensively so far are a quantitative approach for detecting outliers and criteria over the acceptance or rejection of prediction results, there are less number of publications addressing these important issues [52]. The choice of molecular descriptors is also important case but yet difficult task. Application of complex molecular descriptors or use of pK_a or $logK_{ow}$ limits the future of application of QSRR models for newly detected compounds by which the value of these descriptors are not clear (Table 2, No 2, 3 and 4). Apart from these specific points, there has been not any publication with large chemical diversity for prediction of retention time in polar compounds of concerns. In this study, all of these shortages were addressed for predicting retention time accurately.

Table 2: Literature review of research articles for prediction of RT in LC-HRMS

No	Author	Journal information	Details
1	Gorynski K,Bojko B, Nowaczykc A, Bucinski A, Pawliszyn J and Kaliszan R.	Analytica Chimica Acta, (2013) 797: 13– 19 Doi: 10.1016/j.aca.2013.08.025	The models were built based on ALogP, BELe6, R2p and ALogP^2, FDI, BLTA96 with statistical power of R^2= 0.95 for RP-LC and R^2=0.84 for HILIC
2	Miller TH, Musenga A, Cowan D A and Barron LP.	Analytical Chemistry, (2013) 85: 10330−10337, Doi: 10.1021/ac4024878	Descriptors including pKa, Ghose−Crippen and Moriguchi log P (AlogP or MlogP), number of double bonds (nDB), LogD, and number of carbon or oxygen atoms (nC or nO)
3	Munro K, Miller TH, Martins CPB, Edge AM, Cowan DA and Barron LP	Journal of Chromatography A, (2015) 1396: 34–44 Doi:10.1016/j.chroma.2015.03.063	Descriptors including nDB, logD, nO, pK_a and AlogP
4	Bade R, Bijlsma L, Sancho JV and Hernández F	Talanta, (2015) 139: 143–149 Doi:10.1016/j.talanta.2015.02.055	The only descriptor is LogKow with R^2=0.6947
5	Creek D J, Jankevics A, Breitling R, Watson D G, Barrett M P and Burgess K E V.	Analytical Chemistry, (2011) 83: 8703–8710 Doi: 10.1021/ac2021823	Descriptors including log $D_{(pH=3.5)}$, Neg = negative charge at pH 3.5, Pos = positive charge at pH 3.5, Rot = number of rotatable bonds, Phos = number of phosphate groups, and (HBD/MW) =number of hydrogen bond donors divided by molecular weight

No	Author	Journal information	Details
6	Eugster, P. J. Boccard, J. Debrus B, Bréant L, Wolfender J L, Martel S and Carrupt P A.	Phytochemistry, (2014) 108: 196–207 Doi: 10.1016/j.phytochem.2014.10.005	Descriptors including: (α^H, β^H, π^*, V : Abraham's solvation parameters), Molecular weight, TPSA (total polar surface area), S+logP and Rotatable bond
7	Jalali-Heravi M, Kyani A, Afsari-Mamaghani S and Ghadiri-Bidhendi A.	QSAR Combinatorial Science, (2008) 27:407 – 416 Doi: 10.1002/qsar.200630163	Descriptors including: RDF010v(Radial Distribution Function-1.0/ weighted by atomic van der Waals volumes), N-072 (number of functional groups of $\mathrm{R-CO-N}$<and>$\mathrm{N-C\equiv N}$) and Mor06p (atomic polarizabilities)
8	Ledesma E B and Wornat M J.	Analytical Chemistry, (2000) 72:5437-5443 DOI: 10.1021/ac000296r	Descriptors including: moment of inertia, total energy, polarizability, ionization potential, dipole moment, subpolarity
9	Schefzick S, Kibbey C. and Bradley M P.	Journal of Combinatorial Chemistry, (2004) 6: 916-927 DOI: 10.1021/cc049914y	Descriptors including: Acceptors, Donors, Rotatable bonds, CLogP, Molecular weight and Polar surface area
10	D'Archivio A A, Maggib M A and Ruggieri F	Journal of Pharmaceutical and Biomedical Analysis, (2014) 96: 224–230 DOI: 10.1016/j.jpba.2014.04.006	ATSCIe (Centred Broto-Moreau autocorrelation of lag 1 weighted by Sanderson electronegativity), DLS (Modified drug-like score), Eig07-AEA(dm) (Eigenvalue n. 7 from augmented edge adjacency matrix. weighted by dipole moment), H-052 (H attached to C (sp3) with 1X attached to next), HATS8u (Leverage-weighted autocorrelation of lag 8/unweighted) and MATS2e (Moran autocorrelation of lag 2 weighted by Sanderson electronegativity)

No	Author	Journal information	Details
11	Akbar J, Iqbal S, Batool F, Karim A and Chan K W	International Journal of Molecular Sciences, (2012), 13: 15387-15400 Doi:10.3390/ijms131115387	Mp (mean atomic polarizability), IDM (mean information content on the distance magnitude), DISPm (d COMMA2 value weighted by atomic masses), Mor22v (3D-MoRSE signal-22, weighted by atomic van der Waals volumes) annd Mor28e (3D-MoRSE signal-32, weighted by atomic Sanderson electronegativities)
12	Kaliszan R and Markuszewski M J	8. Quantitative Structure–Retention Relationships in Studies of Monolithic Materials, (2011):pp.159-172 Doi: 10.1002/9783527633241.ch8	CLogP
13	Babushok V I and Zenkevich I G	Chromatographia, (2010) 72:781-797 Doi: 10.1365/s10337-010-1721-80009-5893/10/11	sum of hydrophobicities
14	Bączek T, Wiczling P, Marszałł M, Heyden Y V and Kaliszan R.	Journal of Proteome Research, (2005) 4: 555–563 Doi: 10.1021/pr049780r	Experimental descriptor (log SumAA), logarithm of van der Waals volume, CLogP

CHAPTER 3

PURPOSE OF THE STUDY

From the literature review is evident that a wide-scope retention time prediction model is missing to support suspect and non-target LC-HRMS screening. Therefore, the main effort of this study was focused on the development of two widely applicable and acceptable models for negative and positive ionization mode in reversed phase liquid chromatography (RP-LC),meeting all validation criteria, to support the LC-HRMS suspect and non-target screening of environmental emerging contaminants.

Fulfilling this task, k-Nearest Neighborhood (k-NN) and Principle Component Analysis (PCA) were used for dividing the dataset into training and test set to prevent any biases (i.e. chemical structure diversity and retention time distribution) in selection of data points. This is also to remove any information lost or presence of individuals in components of models. The most relevant descriptors, regarding the observed retention times, were selected; for this purpose stepwise (SW) and genetic algorithm (GA) were used.

Multiple Linear Regression (MLR), Artificial Neural Networks (ANN), and Support Vector Machine (SVM) were used to correlate the selected molecular descriptors with the experimental retention times. The final models were evaluated internally and externally and the presence of possible outliers was studied carefully. Based on the statistical results, robust models were selected for the prediction of the retention time of suspect compounds in a LC-QTOFMS screening of a surface water sample (from Danube river as a part of a collaborative trial of the Joint Danube Survey), as external evaluation set. Extra protocols for the outlier detection and also the interpretation of the results were provided to result in accurate retention time prediction. A visualization software was developed (OTrAMS) to facilitate the detection of outliers and to understand the origin of failure. This was an important step in filtering of the screening results in order to reject the false positive results.

CHAPTER 4

LABORATORY EQUIPMENT, INSTRUMENTS AND REAGENTS

4.1. Chemicals

The reference standards of the pesticides were donated to the laboratory by Bruker Daltonics, at a concentration of 1 mg/L in methanol. The rest of the compounds included in the study were all purchased from Sigma–Aldrich (Germany) and are presented in Table S1 in electronic material. Individual stock solutions of these compounds were prepared in methanol at a concentration of 1 g/L and stored at -20 °C. Then, working solutions were prepared in methanol at a concentration of 1 mg/L. Methanol, LC-MS grade, was purchased from Merck (Germany), whereas 2-propanol of LC-MS grade was from Fisher Scientific (Geel, Belgium). Sodium hydroxide monohydrate (NaOH) for trace analysis ≥99.9995%, ammonium acetate, ammonium formate and formic acid, all LC-MS grade, were purchased from Fluka, Sigma–Aldrich (Germany). Distilled water used for LC–MS analysis was provided by a Milli-Q purification apparatus (Millipore Direct-Q UV, Bedford, MA, USA). Regenerated cellulose (RC) syringe filters (15 mm diameter, 0.22 μm pore size) were provided from Phenomenex (Torrance, CA, USA).

4.2. Chromatographic system

An ultrahigh-performance liquid chromatography (UHPLC) system with a LPG-3400 pump (DionexUltiMate 3000 RSLC, Thermo Fisher Scientific, Germany), interfaced to a QTOF mass spectrometer (Maxis Impact, Bruker Daltonics, Bremen, Germany) was used for the screening analysis.

The chromatographic separation was performed on an Acclaim RSLC C18 column (2.1 × 100 mm, 2.2 μm) from Thermo Fisher Scientific (Driesch, Germany) preceded by a guard column, ACQUITY UPLC BEH C18 1.7 μm, VanGuard Pre-Column, Waters (Ireland), thermostated at 30 ˚C. Mobile phase composition in positive ionization mode (PI) is (A) H_2O: MeOH (90:10) with 5 mM ammonium formate and 0.01% formic acid and (B) MeOH with 5 mM ammonium formate and 0.01% formic acid. For the negative ionization mode (NI), the mobile phase is (A) H_2O: MeOH (90:10) with 5 mM ammonium acetate and (B) MeOH with 5 mM ammonium acetate.

The gradient elution program was the same for the 2 ionization modes and the chromatogram lasts 15.5 min, with 5 min of re-equilibration of the column for the next injection. It starts with 1% B with a flow rate of 0.2 mL min^{-1} for 1 min. and it increases to 39 % in 2 min (flow rate 0.2 mL min^{-1}), and then to 99.9 % (flow rate 0.4 mLmin^{-1}) in the following 11 min. Then it keeps constant for 2 min (flow rate 0.48 mL min^{-1}) and then initial conditions were restored within 0.1 min and the flow rate decreased to 0.2 mL min^{-1}. The injection volume was set up to 5 µL.

The operating parameters of the electrospray ionization interface (ESI) are for PI mode: capillary voltage, 2500 V; end plate offset, 500 V; nebulizer, 2 bar; drying gas, 8 L min^{-1}; dry temperature, 200 °C; and for NI mode: capillary voltage, 3500 V; end plate offset, 500 V; nebulizer, 2 bar; drying gas, 8 L min^{-1}; dry temperature, 200 °C.

The QTOF MS system operates in broadband collision induced dissociation (bbCID) acquisition mode and records spectra over the range m/z 50−1000 with a scan rate of 2 Hz. The Bruker bbCID mode provides MS and MS/MS spectra at the same time, while it works at two different collision energies. At low collision energy (4 eV), MS spectra were acquired and at high collision energy (25 eV), fragmentation is taking place at the collision cell resulting in MS/MS spectra.

A QTOF external calibration was daily performed with a sodium formate solution, and a segment (0.1−0.25 min) in every chromatogram was used for internal calibration, using a calibrant injection at the beginning of each run. The sodium formate calibration mixture consists of 10 mM sodium formate in a mixture of water/isopropanol (1:1). The theoretical exact masses of calibration ions with formulas $Na(NaCOOH)_{1-14}$ in the range of 50−1000 Da were used for calibration. The instrument provided a typical resolving power of 36000−40000 during calibration (39274 at m/z 226.1593, 36923 at m/z 430.9137, and 36274 at m/z 702.8636). Mass spectra acquisition and data analysis was processed with Data Analysis 4.1 and Target Analysis 1.3 (Bruker Daltonics, Bremen, Germany).

4.3. Development of the dataset

A very important step is the selection of the analytes to build the database and then find a representative subset to provide for the modeling of the retention time prediction. The list includes pesticides from different classes and modes of actions, like organophosphorous,

carbamates, neonicotinoids, pyrethroids, ureas and many more, and some other emerging contaminants, like pharmaceuticals, illicit drugs, sweeteners, anti-corrosion agents, and perfluorinated compounds. To begin, the selection of the list was based initially on the diversity of the compounds, in order to cover the whole range of physicochemical properties of possible emerging contaminants. Moreover, the ionization efficiency of the compounds was examined; both positively and negatively ionizable compounds were selected. There is a higher number of compounds in positive ionization mode than in negative, as that is the case in the screening database, as well. Finally, different functional groups were selected through the compounds, since they play an important role in the retention time of a compound.

After the selection of the list of compounds that would be used to build the models, reference standard solutions of all the compounds at concentration 1 mg/L were injected at the chromatographic system in triplicate, in both polarities. Retention time of the compounds was recorded and was further evaluated for the models.

4.4. Sample preparation

The sample analyzed for this study was part of a collaborative trial organized by the NORMAN Association (www.normannetwork.net), where one of the main purposes was the comparison and harmonization of non-target screening methods [55].

The sample used in the collaborative trial was collected from location JDS57, downstream of Ruse/Giurgiu (RO/BG; rkm 488; coordinates N43.890150, E26.017067) on September 18, 2013 as a part of the Third Joint Danube Survey, organized by the International Commission for the Protection of the Danube River (ICPDR). The sample preparation included a large-volume solid-phase extraction (LVSPE) of 1000 litres of water. Briefly, the sampler cartridge was filled with 160 g of Macherey Nagel Chromabond® HR-X (neutral resin) and 100 g each of Chromabond® HR-XAW (anionic) and HR-XCW (cationic exchange resin). The retained compounds were extracted from the sorbents with 500 mL each of ethyl acetate and methanol (HR-X), 500 mL methanol with 2% 7 M ammonia in methanol (HR-XAW) or 500 mL methanol with 1% formic acid (HR-XCW). The extracts were then combined, neutralized, filtered (What man GF/F) and reduced to a final volume of 1 L using rotary evaporation. Aliquots of 1.5 mL, equivalent to 1.5 L of river water, were

transferred into vials and evaporated to dryness under nitrogen. These were sent to each participant along with a laboratory blank. The samples were reconstituted in MeOH:H_20 (50:50) in 1.5 mL and filtered through RC syringe filters prior to analysis [55].

4.5. QSRR methodology

All the chemical structures of the selected compounds were drawn in Hyperchem 7.03 [21] and then the initial geometry optimization calculations which employ energy minimization algorithms to locate the most stable structures were used. Here, all molecules were pre-optimized by using molecular mechanics force field (MM+), and then, final optimization was carried out by using semi-empirical (AM1) method with root mean square gradient of 0.01 kcal mol^{-1}. Dragon program was employed to calculate molecular descriptors for each optimized molecule [22]. The descriptors were grouped in 22 different types, including: constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelation, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, Radial Distribution Function (RDF) descriptors, 3D-MoRSE (3D Molecular Representation of Structure based on Electron diffraction) descriptors, WHIM (Weighted Holistic Invariant Molecular) descriptors, GETAWAY (geometry, topology and atoms-weighted Assembly) descriptors, functional group counts, atom-centred fragments, charge descriptors, molecular properties, 2D binary fingerprints and 2D frequency fingerprints [23-26]. Among the above descriptors, the constitutional descriptors are referring to atomic or molecular properties and are independent of the overall molecular connectivity. These types of descriptors encode the size of molecules and chemical properties. Geometrical descriptors are presenting features of the molecular geometry, e.g. distances between particular points on the molecular surface and distances between given chemical groups. Topological descriptors reflect the type and the connection of atoms in the 2D space [27]. In addition to the above descriptors, since the compounds contained ionizable functional groups in relevant pH, Log D, which encodes the lipophilicity of a molecule in aqueous phase with different pH, was calculated for each compound (at pH=3.6 for positive ionization compounds and pH=6.2 for negative ionization compounds) by using ChemAxon package [28]. The calculated descriptors for molecules in both

ionizations were pre-treated in order to remove the constant and near constant descriptors. Moreover, the remained variables were checked for existence of collinearity, so as to decrease the redundancy of the descriptor data matrix [e.g. among the detected collinear descriptors (r>0.9), the one showing the highest correlation with the activity/property was retained, and the others were removed from the data matrix]. To build predictive models for predicting the retention time behavior of suspect compounds, datasets (separately for positive and negative ionization) were split into training and test set using K-nearest neighborhood and principle component analysis. To derive the most relevant descriptors that are correlating to the retention time, and present the less inter-correlation values, stepwise and genetic algorithms were used. Genetic algorithm and stepwise methods as selection tools were written in MATLAB 6.5 program [29]. The results of each variable selection technique were then used as input for several modeling techniques such as Multiple Linear Regressions (MLR), Support Vector Machines (SVM) and Artificial Neural Networks (ANN).The derived models were compared and the most reliable models for identification purposes were finally selected.

4.6. OTrAMS

A novel display was developed to visualize the correlation between the activities, similarity, and standard residuals to fully understand the origin of residuals between the experimental and predicted retention time. In this technique the following steps are performed for obtaining the visualization:

- The dataset consists of three sub groups (train, test, suspect), expressed by their experimental retention time, standard residuals and normalized mean distances.
- In the next step, a3D-plot is produced by 4 boxes for the given data set (training and test set) in which each box corresponds to the range of standardized residuals. Box 1 is showing the range between ±1δ, box 2 denotes the range between ±1δ and ±2δ, box 3 indicates the range between ±2δ and ±3δ and box 4 indicates beyond ±3δ. Standardized residuals (δ) are raw residuals divided by their estimated standard deviation. The cut-off value for standardized residuals (±3δ) is set based on 99% confidence value of the modeling results.
- Next, the percent and number of available molecules in each box are calculated and saved in output file to show the distribution of data set in each box (the code written

in MATLAB to derive this plot is available in electronic material (OTrAMS.p)). The large presence of compounds in box 1 represents the compounds with the less error made by the model.

- Then, a visualization plot for box 3, which affects the model highly due to striking residuals, is demonstrated. In this plot, a compound in box 3 can be analyzed based on its distance from the mean value of the training set to understand the origin of the residual. The size of a bubble is proportional to leverage values and hence this plot can provide a quick analysis of outliers and their origins.

This step is crucial for accepting whether the retention times of the suspect compounds are correct or not. Based on the calculation of step 4 for the suspect list, any outliers located in box 4 and box 3 can be identified. Therefore, a high similarity distance from the mean value (training set) indicates that the suspect molecule cannot be studied by the model due to its dissimilarity and its unique structure. If the similarity distance for a suspect compound is low but the observed retention time shows higher distance from the mean of the training set, it indicates that this suspect compound is not correct and the response cannot be related to the provided structure. This is a major filtering step of reducing the false positive results of a screening HRMS procedure. The results of all above steps are saved finally for further analyses purposes.

CHAPTER 5

RESULTS AND DISCUSSION

5.1. Developed model for negative Electrospray Ionization Mode ((-)ESI)

5.1.1 PCA-SW-MLR

The selection of the test set based on PCA method is shown in figure 6. For selection of test set, the distribution of data points in score plot and also their retention time were considered.

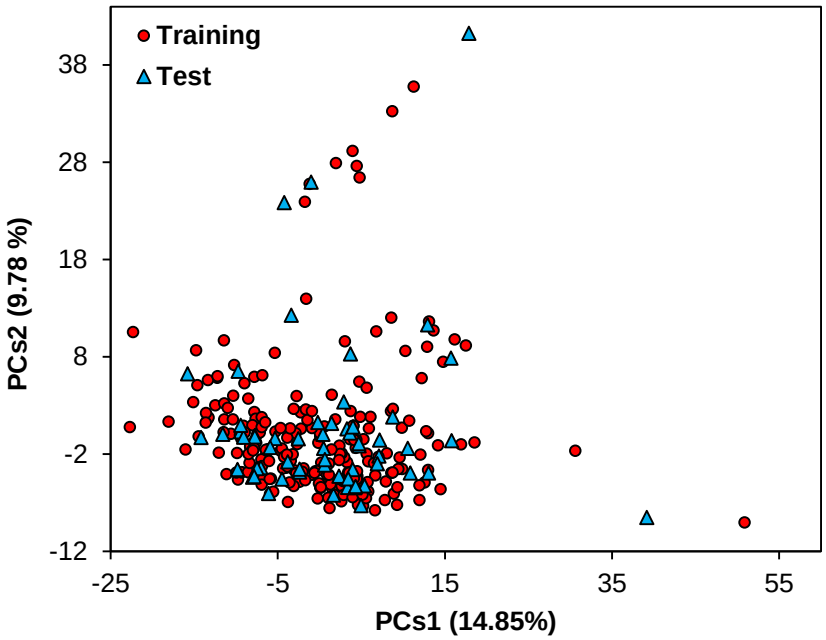


Figure 6: PCA analysis for negative ionization compounds (sample test set for SW-MLR)

After classification of data set by PCA method into training and test set, the stepwise method was used to select the most respective variables to understand the correlation of molecular structures with retention time. Based on the stepwise method as explained in section 1.2.4.1, the most seven relevant descriptors were selected and then the linear regression model was built. The linear model, based on the selection of test set on the biases of PCA, has obtained as follows:

Rt = -0.4879 (±0.6701) - 0.5351(±0.1141) nR06 + 0.9952(±0.2119) ICR + 0.8935(±0.2514) ATS3p - 0.6955(±0.1018) EEig13d + 0.9912(±0.1543) R3e + 0.5276(±0.0767) ALOGP + 0.7372(±0.06057) Log D(pH at 6.20) $(Eq. 20)$

N_{train}=241, R^2_{train}=0.854, $RMSE_{train}$=1.053, R^2_{adj}=0.850, F_{train}=195.523, Q^2_{LOO}=0.844, Q^2_{LGO}=0.777, Q^2_{BOOT}=0.842, N_{test}=59, R^2_{test}=0.782, $RMSE_{test}$=1.367, F_{test}=28.30, rm^2_{test} =0.724, CCC_{test}=0.8791, CCC_{train}=0.9216

Where N is the number of compounds, R^2 is the squared correlation coefficient, R^2_{adj} is the adjusted R^2, $Q^2_{LOO,}$ Q^2_{BOOT} and Q^2_{LGO} are the squared cross-validation coefficients for leave one out, bootstrapping and leave group out, respectively, RMSE is the root mean square error and F is the Fisher F statistic. As it can be seen, the obtained model shows the acceptable statistical parameters with higher square correlation coefficient (R^2), Fisher F statistic (F) and concordance correlation coefficient for both sets with lower RMSE values. The predicted retention time values for the whole range of the compounds in training and test sets using equation 20 have been plotted against the observed retention time values in figure 7, and listed in Table S1(electronic material). The corresponding VIF values and inter-correlation values of the selected seven descriptors are shown in Table 3.

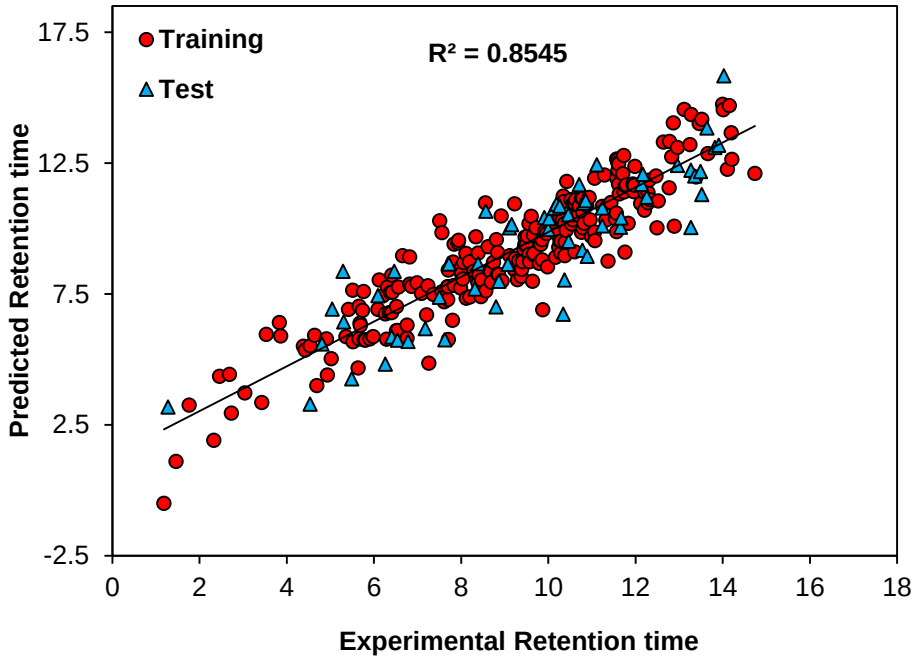


Figure 7: The plot of predicted retention time against the observed retention time values based on PCA-SW-MLR

Table 3: The correlation coefficient of selected descriptors and corresponding VIF values by PCA-SW-MLR

Variables	nR06	ICR	ATS3p	EEig13d	R3e	ALOGP	Log D(6.20)	VIF[a]
nR06	1	0	0	0	0	0	0	1.875
ICR	0.488	1	0	0	0	0	0	2.052
ATS3p	0.61	0.673	1	0	0	0	0	2.475
EEig13d	0.408	0.479	0.538	1	0	0	0	1.713
R3e	-0.051	0.303	0.277	0.381	1	0	0	1.492
ALOGP	0.356	0.45	0.488	0.315	0.263	1	0	3.287
Log D(6.20)	0.428	0.46	0.496	0.308	0.118	0.81	1	3.335

[a] Variation inflation factor

As can be seen from this Table, all variables have VIF values less than 5, indicating that the obtained model has appropriate selected variables. Also low R^2 and Q^2 values were obtained by Y-randomization test (Table 4).

Table 4: The Q^2_{Loo} and $R^2_{training}$ values after several Y-randomization tests for PCA-SW-MLR

No	Q^2	R^2
1	0.0047	0.0218
2	0.0002	0.0411
3	0.0071	0.0507
4	0.0082	0.0209
5	0.0004	0.038
6	0.0182	0.0163
7	0.0103	0.0201
8	0.0444	0.0126
9	0.0014	0.0251
10	0.0093	0.0579

The robustness of the proposed model and its predictive ability was guaranteed by the high Q^2_{BOOT} based on bootstrapping repeated 5000 times. Applicability domain was used and outliers were detected and removed; the final model was generated and showed two outliers that possessed residuals more than ±3δ (figure 8). These two compounds are belonged to the test set, and didn't include in model development; therefore, their omission just benefits the outcome of test set (R^2 from 0.782 to 0.818). Before interpreting the descriptors based on PCA-SW-MLR, genetic algorithms technique is also used to compare the two methods and their results.

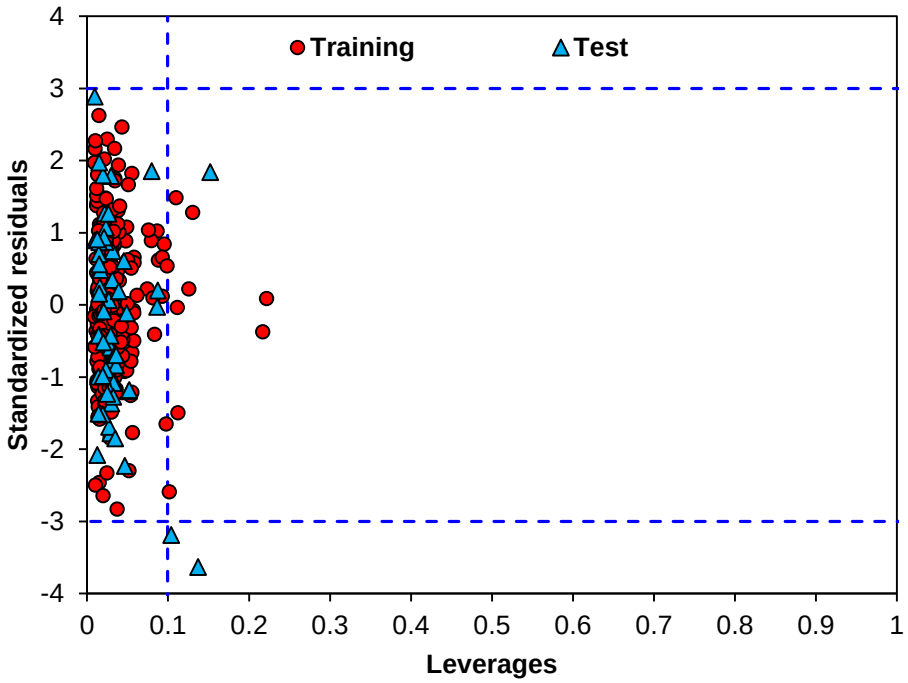


Figure 8: William plot of PCA-SW-MLR model (equation 9): h* warning leverage value is 0.09985.

5.1.2 PCA-GA-MLR

After classification of the data set by the same procedure done in PCA-SW-MLR, the genetic algorithm was used to select the most relevant descriptors. For selection of the best subset of descriptors, genetic algorithms technique was performed for different times and among the generated results, the best model which could present higher statistical

parameters where chosen. The results of combinations of different couples of descriptors selected by GAs were listed in Table 5.

Table 5: Comparison of statistical parameters for different selected descriptors by PCA-GA-MLR

Linear model equations	
Model 1:	Rt= 1.087(±1.116) +0.59215(±0.0670) Log D(6.20) -0.369 (±0.0854) BLTA96 +0.262(±0.0632) ALOGP -0.118 (±0.0630) nO +1.14073(±0.4086) BEHm4 +0.107 (±0.0531) RBN +0.4603(±0.1773) CIC1
Model 2:	Rt= 2.730(±0.526) +0.627 (±0.0616) Log D(6.20) -0.3601 (±0.0813) BLTA96 +0.285(±0.0598) ALOGP -0.158(±0.0983) O-058 +0.1085(±0.0471) RBN -0.00758(±0.00311) TPSA(Tot) +1.638 (±0.279) R2e
Model 3:	Rt= 2.472 (±0.532) +0.575 (±0.0702) Log D(6.20) -0.386(±0.0904) BLTA96 +0.336 (±0.0642) ALOGP -0.0737 (±0.0621) nO +0.804 (±0.493) BELm3 +0.262(±0.122) H3m +0.680 (±0.304) ICR
Model 4 :	Rt= 3.561 (±0.395) +0.545 (±0.0727) Log D(6.20) -0.540 (±0.0873) BLTA96 +0.310 (±0.0643) ALOGP -0.00015(±0.00249) TPSA(Tot) -0.258 (±0.2741) Mor23u -0.494 (±0.170) O-057 +1.481 (±0.282) B06[C-C]
Model 5:	Rt= 1.581 (±1.461) +0.657 (±0.0668) Log D(6.20) -0.378(±0.0951) BLTA96 +0.169 (±0.0691) ALOGP -0.0101 (±0.00365) TPSA(Tot) -0.0934(±0.0506) HGM +1.64 (±0.490) BEHm4 +0.358(±0.273) GATS1m

Statistical Resuls									
	r^2_{train}	**$RMSE_{train}$**	**F_{train}**	**r^2_{test}**	**$RMSE_{test}$**	**F_{test}**	**Q^2_{LOO}**	**Q^2_{Boot}**	**rm^2_{test}**
Model 1	0.799	1.24	132.81	0.784	1.363	27.05	0.768	0.767	0.742
Model 2	0.821	1.171	153.27	0.766	1.45	25.76	0.798	0.797	0.687
Model 3	0.792	1.263	126.91	0.79	1.353	28.96	0.763	0.761	0.731
Model 4	0.806	1.219	138.76	0.765	1.45	25.59	0.761	0.764	0.688
Model 5	0.789	1.27	125.1	0.788	1.36	27.49	0.755	0.754	0.743
Main Model	*0.812*	*1.201*	*143.94*	*0.786*	*1.379*	*28.08*	*0.789*	*0.788*	*0.721*

Rt = 1.789(±0.5342) + 0.6561(±0.06779) LogD(pH at 6.20) - 0.5386 (±0.08216) BLTA96 + 0.3083(±0.06447) ALOGP + 0.1038(±0.1205) nROH + 0.5174 (±0.376) HATS6m + 1.591(±0.29003) R2e - 0.2762 (±0.2209) Mor25e $(Eq. 21)$

N_{train}=242, R^2_{train}=0.812, $RMSE_{train}$=1.201, R^2_{adj}=0.806, F_{train}=143.94, Q^2_{LOO}=0.789, Q^2_{LGO}=0.700, Q^2_{BOOT}=0.788, N_{test}=59, R^2_{test}=0.786, $RMSE_{test}$=1.379, F_{test}=28.08, rm^2_{test} =0.721, CCC_{test}=0.8775, CCC_{train}=0.8960

The obtained statistical parameters (high squared correlation coefficient, CCC, Q^2_{BOOT} and Q^2_{LOO}) show that genetic algorithms technique is better than stepwise method for selecting of descriptors as model variables. To find out that the selected descriptors are statistically meaningful, the Y-randomization test was used(Table 6).

Table 6: The Q^2_{LOO} and $R^2_{training}$ values after several Y-randomization tests for PCA-GA-MLR

No	Q^2	R^2
1	0.0006	0.0248
2	0.0026	0.0417
3	0.0286	0.0131
4	0.0665	0.0034
5	0.0053	0.0166
6	3.30E-06	0.0282
7	0.003	0.0235
8	0.0027	0.0352
9	0.0004	0.0286
10	0.0388	0.0073

In this method, the properties for a group of compounds were shuffled, and then, a new model was built. The new QSPR models as outcome of this method should present low R^2 and Q^2_{LOO} values so as to be confident that the models are directly in relation with the selected variables. The predicted retention time values for all the compounds in training and test sets, using the equation 21, plotted against the observed retention time values are shown in figure 9, and listed in Table S1.

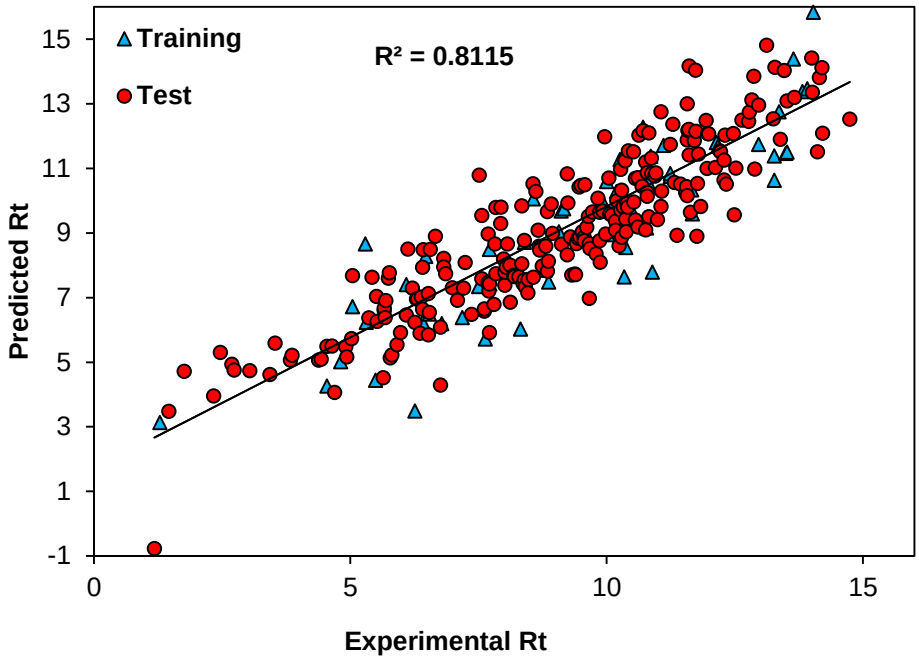


Figure 9: The plot of predicted retention time against the observed retention time values based on PCA-GA-MLR

The corresponding VIF values and inter-correlation values of the selected seven descriptors are listed in Table 7.

Table 7: The correlation coefficient of selected descriptors and corresponding VIF values by PCA-GA-MLR

Variables	Log D(6.20)	BLTA96	ALOGP	nROH	HATS6m	R2e	Mor25e	VIF[a]
Log D(6.20)	1	0	0	0	0	0	0	3.224
BLTA96	-0.638	1	0	0	0	0	0	2.464
ALOGP	0.712	-0.533	1	0	0	0	0	2.543
nROH	-0.243	0.0585	-0.0785	1	0	0	0	1.304
HATS6m	0.0774	-0.157	0.256	0.0136	1	0	0	1.300
R2e	0.228	-0.203	0.089	0.081	0.357	1	0	1.689
Mor25e	0.424	-0.59	0.378	0.2603	0.246	0.552	1	2.472

[a] Variation inflation factor

As can be seen from this Table, all variables have VIF values less than 5, indicating that the obtained model has excellent selected descriptors. Applicability domain was also obtained for the generated model and showed no outliers that possessed the residuals more than ±3δ (figure 10).

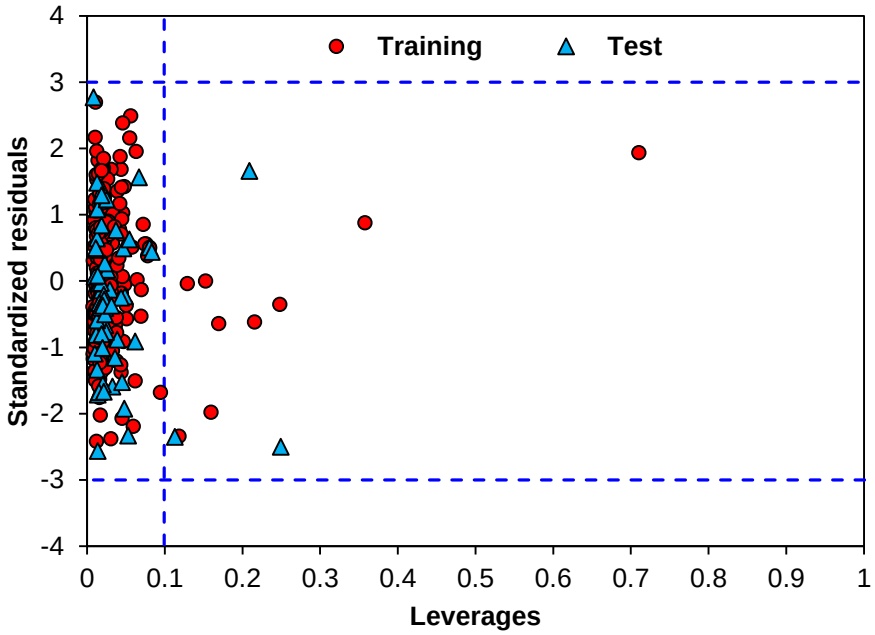


Figure 10: William plot of PCA-GA-MLR model (equation 10): h* warning leverage value is 0.09917

The PCA-GA-MLR model (eq 21) was obtained after the removal of compounds semduramicin and alitame, and the second built model did not show any outliers for the training set, so as to rebuild the model. Some other compounds were located outside the warning leverage value, however they did not show high (more than ±3δ) residuals, and therefore they did not treated as outliers. To understand the reason of these two outliers, the molecular descriptors which were selected by GAs can be used as input for Euclidean based applicability domain (figure 11), so as to explain the diversity of compounds based on the selected descriptors. As it can be seen, the origin of outliers are not derived from structural diversity, since they are within the capability of the model to be predicted, however the observed response did not match to the given structure. Therefore, the PCA-

GA-MLR model can be accepted as an initial model for predicting purposes. This workflow can help to understand if the screened unknown and suspect compounds to be studied further are within the capability of the models or not, before predicting their retention times values.

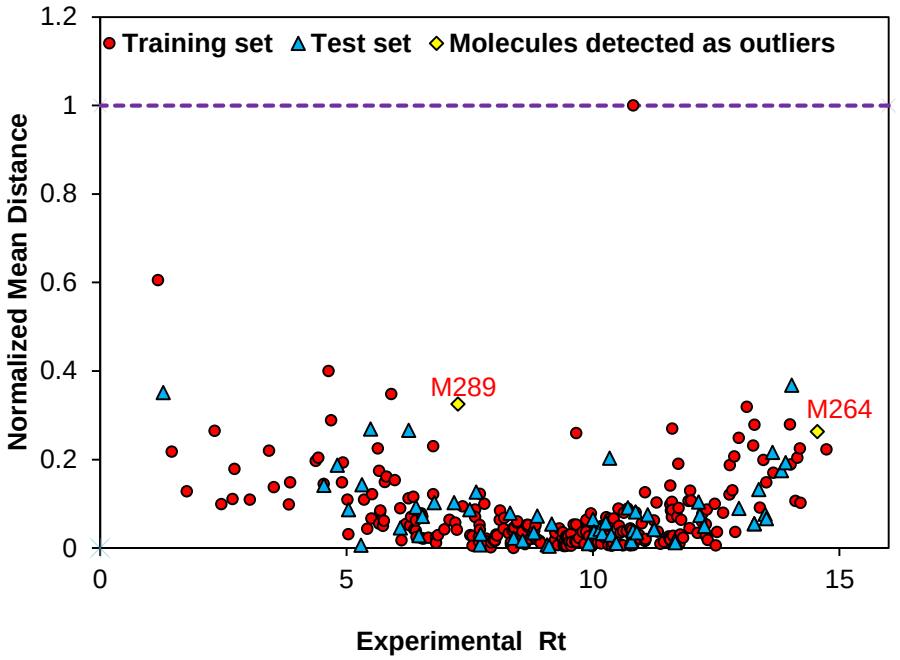


Figure 11: Euclidean based applicability domain of the compounds for PCA-GA-MLR

5.1.3kNN-SW-MLR

The same procedures were used for developing the linear and non-linear models; however the data set was spilt based on the results of a kNN dendrogram. The test compounds were marked in Table S1 and were shown in figure S1. Since the all interpretations of the results were explained above, therefore, here we are just presenting the obtained results for KNN-SW-MLR; the linear model was calculated as follows:

Rt = -0.5622 (±0.6977) + 0.9148 (±0.2304) ATS3p + 1.657(±0.3289) GATS2m - 1.006(±0.1392) EEig14r +1.601 (±0.2167) R3u +0.4980(±0.07977) ALOGP - 0.7737(±0.1948) B02[C-S] +0.7197(±0.0623) LogD(pH at 6.20) $(Eq. 22)$

N_{train}=241, R^2_{train}=0.842, $RMSE_{train}$=1.107, R^2_{adj}=0.837, F_{train}=176.95, Q^2_{LOO}=0.829, Q^2_{LGO}=0.752, Q^2_{BOOT}=0.827, N_{test}=60, R^2_{test}=0.822, $RMSE_{test}$=1.209, F_{test}=29.41, rm^2_{test} =0.770, CCC_{test}=0.8941, CCC_{train}=0.9140

The Y-randomization test was also used, and the results indicated that the developed model is acceptable (Table 8).

Table 8: The Q^2_{LOO} and $R^2_{training}$ values after several Y-randomization tests for kNN-SW-MLR

No	Q^2	R^2
1	0.0076	0.0193
2	0.0059	0.0224
3	0.075	0.0089
4	0.0015	0.0417
5	0.0028	0.0259
6	0.0335	0.0119
7	0.0053	0.0197
8	0.0005	0.0355
9	0.0053	0.0214
10	0.0023	0.0259

William plot was also calculated to detect the possible outliers, however non-outliers were seen for the training set (figure 12), and only one molecule which belonged to the test set was detected as outlier, in which its omission will not benefit the model, since it was not included in the model construction. The predicted retention time values using the equation 22 plotted against the observed retention time values are given in figure 13, and the results for the whole data set are listed in Table S1.

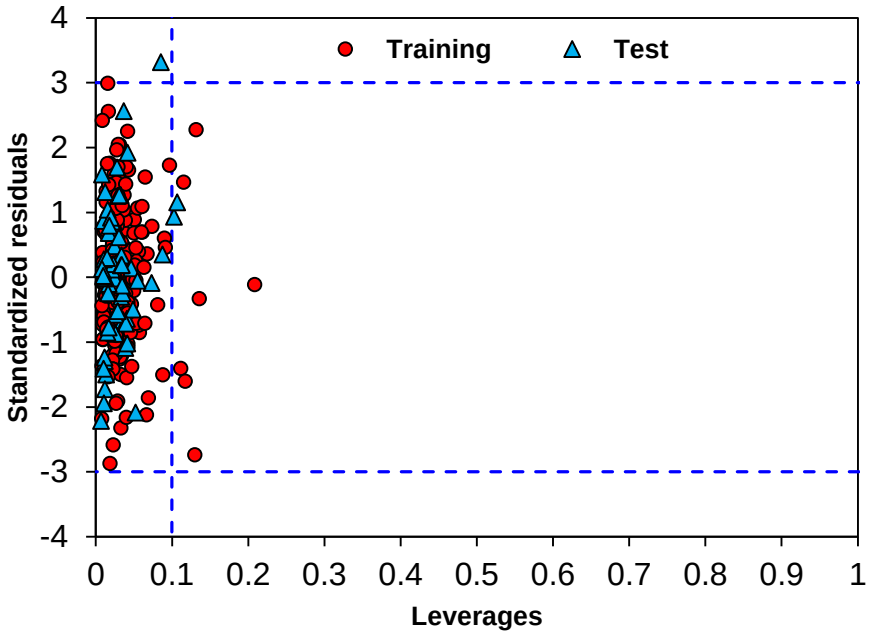


Figure 12: William plot of kNN-SW-MLR model (equation 11): h* warning leverage value is 0.099585

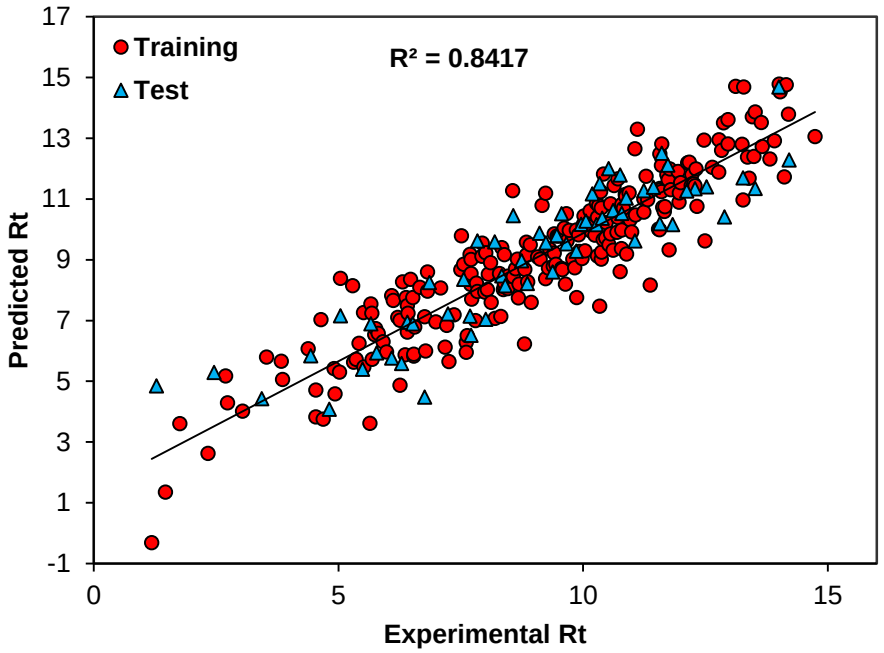


Figure 13: The plot of predicted retention time against the observed retention time values based on kNN-SW-MLR

5.1.4 kNN-GA-MLR

The obtained results for kNN-GA-MLR, as a general linear model, were calculated as follows:

Rt = -0.4297(±1.012) + 0.6242(±0.06824) LogD(pH at 6.20) + 0.4649(±0.1027) ALOGP - 0.08647(±0.09383) BLTA96 - 0.6998(±0.1527) EEig14r + 0.7589(±0.1320) CIC1 + 1.551(±0.3386) BEHm4 + 0.7907 (±0.3687) HATS6m $(Eq. 23)$

N_{train}=241, R^2_{train}=0.820, $RMSE_{train}$=1.169, R^2_{adj}=0.815, F_{train}=152.01 Q^2_{LOO}=0.806, Q^2_{LGO}=0.781, Q^2_{BOOT}=0.803, N_{test}=60, R^2_{test}=0.835, $RMSE_{test}$=1.228, F_{test}=27.74, rm^2_{test} =0.745, CCC_{test}=0.8935, CCC_{train}=0.9013

The equation 23 was obtained after removal a compound which was detected as outlier. The different selected compounds as test set and training set caused better prediction which was compared with other methods in Table 13. The different combinations of molecular descriptors based on training and test set selected by kNN were listed in Table 9.

Table 9: Statistical parameters comparison based on different selected descriptors by kNN-GA-MLR

Linear model equations	
Model 1:	Rt= 2.47 (±0.505) +0.643 (±0.0696) Log D(6.20) +0.418(±0.105) ALOGP -0.223 (±0.0938) BLTA96 -0.582 (±0.257) nPyridines +0.0822(±0.427) HATS6m +1.630 (±0.278) R2e +0.288(±0.1146) Cl-089
Model 2:	Rt= 4.27 (±0.449) +0.497(±0.0789) Log D(6.20) +0.451(±0.114) ALOGP -0.269 (±0.106) BLTA96 +0.357 (±0.0890) TI2 -0.473 (±0.176) O-057 +1.374(±0.560) R3p -0.0070(±0.0033) TPSA(Tot)
Model 3:	Rt= 3.361 (±0.464) +0.586(±0.0666) Log D(6.20) +0.460(±0.0986) ALOGP -0.0968(±0.1003) BLTA96 -1.147(±0.257) GATS1m +0.307(±0.148) CIC1 +1.88(±0.297) R2e -0.0047(±0.0028) TPSA(Tot)
Model 4 :	Rt= 2.341(±0.461) +0.598 (±0.0685) Log D(6.20) +0.398 (±0.106) ALOGP -0.231 (±0.1011) BLTA96 +0.157(±0.0542) F03[C-Cl] +1.140(±0.278) R2e +0.0792(±0.0665) S3K +0.458 (±0.171) CIC1
Model 5:	Rt= 4.28 (±0.320) +0.597 (±0.0788) Log D(6.20) +0.483 (±0.109) ALOGP -0.216 (±0.0984) BLTA96 +0.603(±0.256) F04[Cl-Cl] +0.614 (±0.206) CIC1 -0.308(±0.363) Mor28u +0.0470(±0.148) nROH

	Statistical Resuls								
	r^2_{train}	$RMSE_{train}$	F_{train}	r^2_{test}	$RMSE_{test}$	F_{test}	Q^2_{LOO}	Q^2_{Boot}	rm^2_{test}
Model 1	0.811	1.12	142.74	0.807	1.30	26.22	0.796	0.795	0.751
Model 2	0.796	1.245	130.01	0.790	1.355	25.58	0.775	0.773	0.753
Model 3	0.821	1.167	152.72	0.766	1.420	22.25	0.807	0.805	0.733
Model 4	0.812	1.198	143.30	0.820	1.265	28.61	0.796	0.795	0.761
Model 5	0.792	1.258	126.88	0.824	1.255	28.62	0.774	0.772	0.757
Main Model	*0.82*	*1.169*	*152.01*	*0.835*	*1.228*	*27.74*	*0.806*	*0.803*	*0.745*

The Y-randomization test was employed again, and the results were indicated that the developed model is acceptable (Table 10).

Table 10: The Q^2_{LOO} and $R^2_{training}$ values after several Y-randomization tests for KNN-GA-MLR

No	Q^2	R^2
1	5.12E-05	0.0273
2	0.03087	0.0156
3	0.00635	0.0493
4	0.001	0.0302
5	0.00551	0.021
6	0.00764	0.0203
7	0.13767	0.0026
8	0.01656	0.0188
9	0.01145	0.018
10	0.03597	0.0117

William plot was also calculated to detect the possible outliers for the final model; however non-outliers were observed (figure 14). The predicted retention time values using the equation 23 plotted versus the observed retention time values were shown in figure 15, and the results for the whole data set were listed in Table S1.

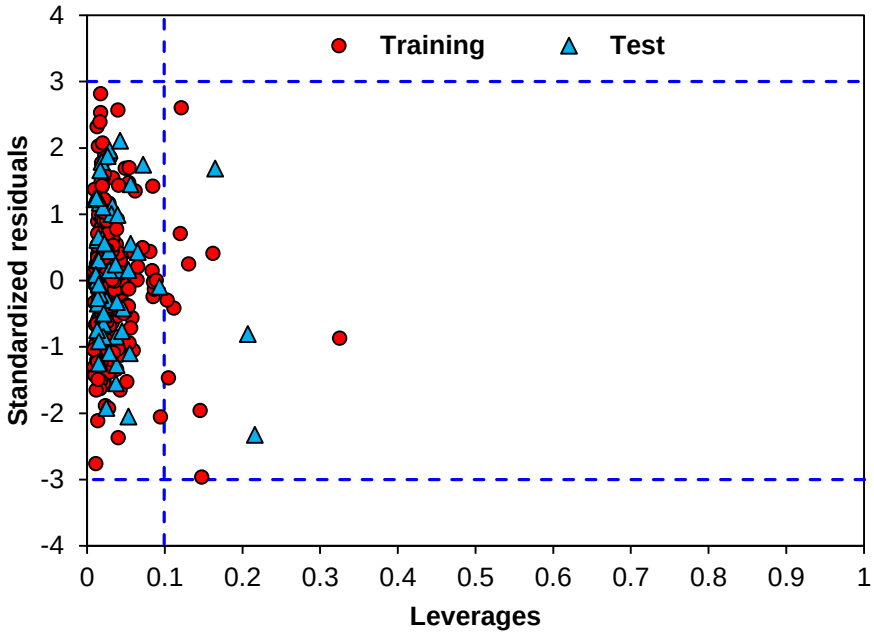


Figure 14: William plot of kNN-GA-MLR model (negative ionization): h* warning leverage value is 0.09914

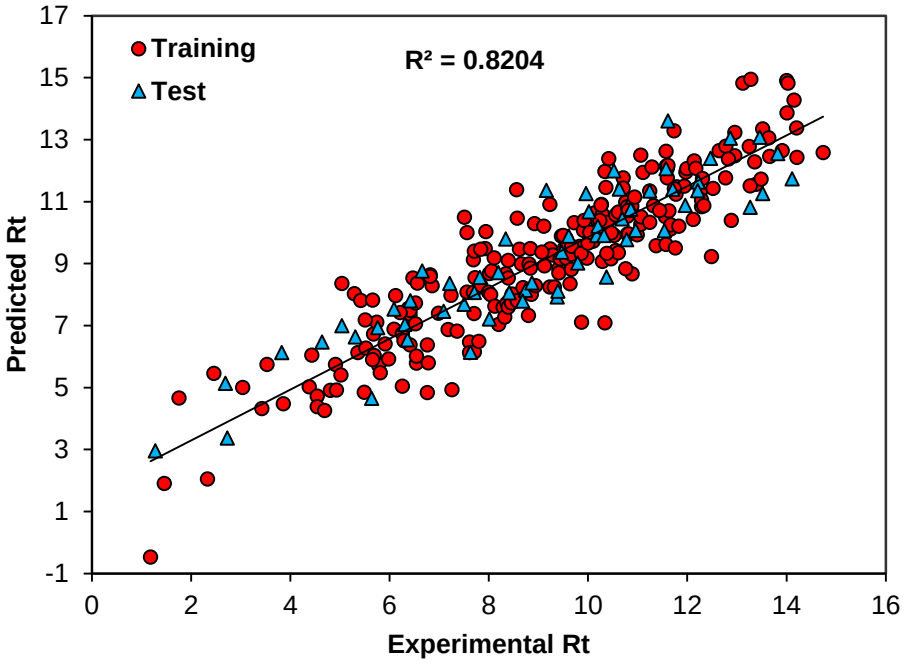


Figure 15: The plot of predicted retention time against the observed retention time values based on KNN-GA-MLR

5.1.5PCA-SW-SVM

After successful linear modeling based on both stepwise and genetic algorithms techniques, support vector machine method was used as non-linear modeling technique on the same subsets of descriptors used in linear modeling. As explained before, SVM regression depends on the combination of different factors such as kernel function type, capacity parameter *C*, ε of ε-insensitive loss function, and its corresponding parameters[47]. For generating the SVM model, firstly, the Kernel function type should be declared in which determines the sample distribution in space. As said above, in this work the radial basis function (RBF) was used due to its good general performance [48]. Considering equation 6, the γ parameter can be provided. γ is in close relation with SVM performance (its training time) where controls the generalization ability of SVM. Generally, to get the optimum value for γ, it is being measured from 0.1 to 5 with incremental steps of 0.1. To get better insight about the optimized values, the root mean square errors (RMSE) of cross-validation were obtained in each step. Figure 16 represents the plot of γ versus RMSE on the leave one out cross-validation. Here the optimal value of 2.7 has been obtained forγ.

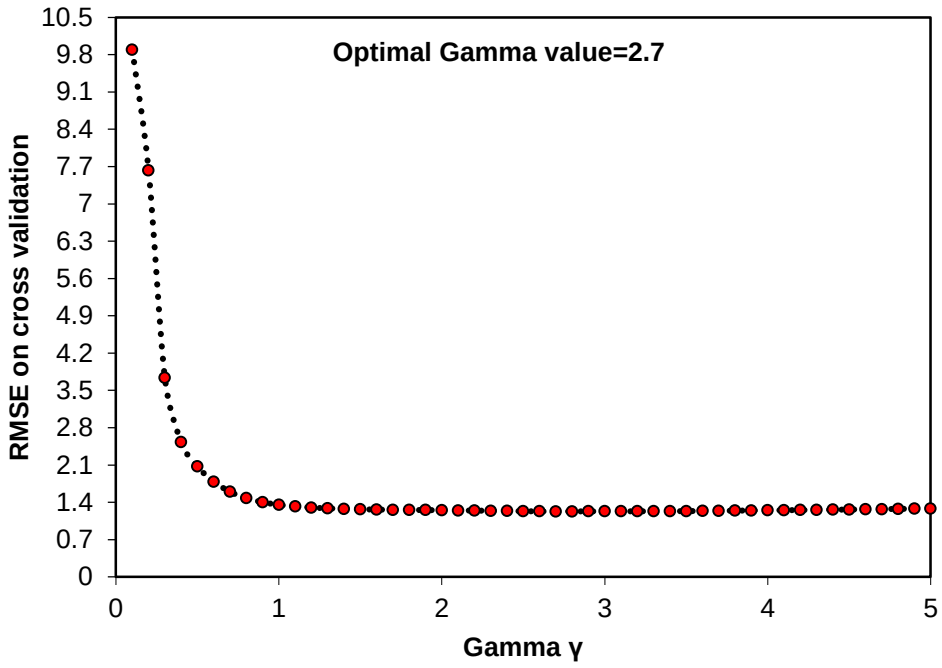


Figure16: The gamma(γ) vs. RMSE for the training set based on PCA-SW-SVM

Parameter ε-insensitive prevents the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulation's solution. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. ε-insensitive has an effect over smoothness of the response of SVM, and also influence the number of support vectors. An increase in ε-insensitive value reflects the reduction in requirements for the desired accuracy approximation. Therefore, if ε-insensitive is zero, there is an over-fitting issue, and if it presents larger values than the range of target values, the obtained results are not appropriate. The RMSEs of cross-validation for different ε values from 0.01 to 0.1 with incremental steps of 0.01 are shown by figure 17. The optimal value for ε-insensitive is 0.01.

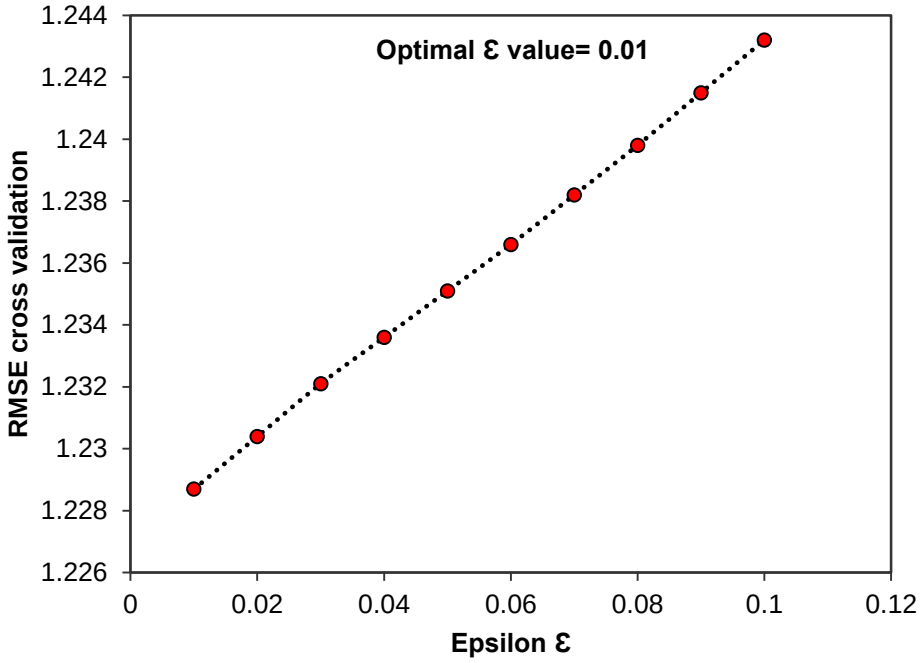


Figure 17: The epsilon (ε) vs. RMSE for the training set based on PCA-SW-SVM.

The final parameter which should be optimized was C where is a regularization parameter that controlled the tradeoff between maximizing the margin and minimizing the training error. The small values for C parameter would increase the number of training errors, and a large value would cause hard-margin SVM behavior. The capacity parameter C was checked from 1 to 50 with incremental steps of 1 and is shown in figure 18. The optimal value for capacity parameter is 50.

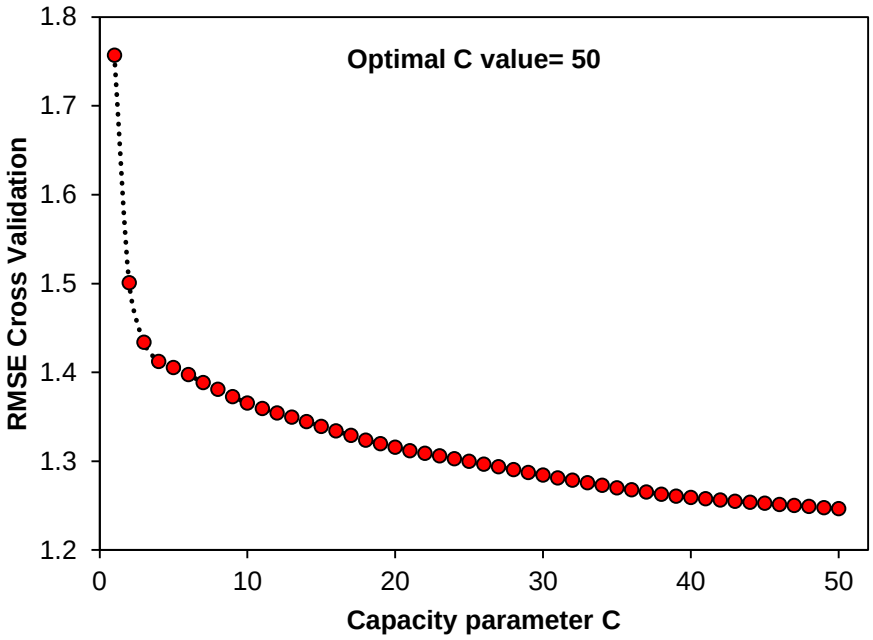


Figure 18: The capacity parameter(C) vs. RMSE for the training set based on PCA-SW-SVM

The parameters of SVM model were optimized as C=50, ε=0.01, γ=2.7. The predicted values for retention time by SVM method were given in Table S1.Also, the predicted versus experimental retention time values for both the training set and test set based on SVM model was implemented in figure 19.

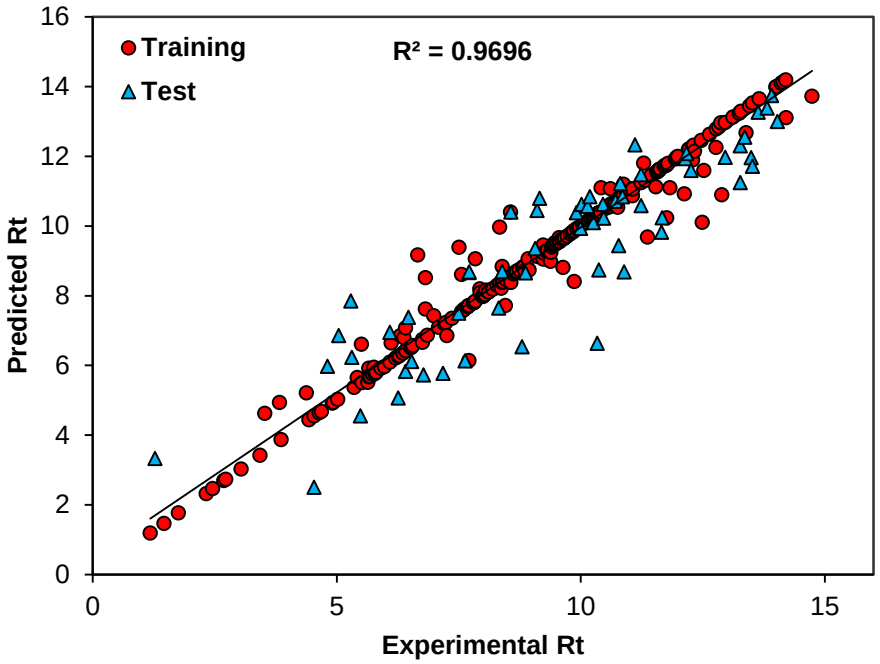


Figure 19: The plot of predicted retention time against the observed retention time values based on PCA-SW-SVM

The statistical parameters for PCA-SW-SVM model showed RMSE values with 0.486 for the training set, 1.25 for the test set, and the squared correlation coefficients (R^2) of 0.970 and 0.818 for training and test set, namely. Table 8 presents the statistical parameters of the results obtained from the studied models for the same set of compounds. For obtaining better results, the above workflow were performed for compounds of the test set and training set which were selected by K-nearest neighborhood clustering technique.

5.1.6PCA-GA-SVM

The same procedure employed in PCA-SW-SVM model was performed in this part; however the non-linear model was built based on the selected descriptors using genetic algorithms as a selection tool. The test set compounds were marked in Table S1 which were the same used in generation of PCA-GA-MLR model. The parameters of SVM model were optimized as C=50, ε=0.01, γ=1.9. The result of each optimization was shown in figures 20-22. The predicted values for the retention time by PCA-GA-SVM method were given in Table S1, and then plotted versus the observed retention time and shown in figure 23. The statistical results of PCA-GA-SVM were listed in Table 13.

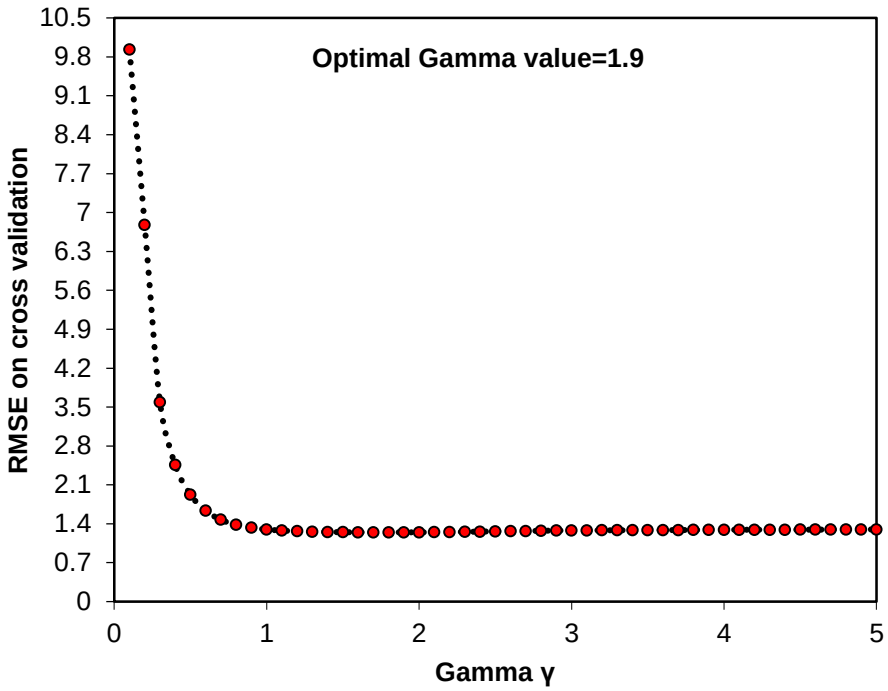


Figure20: PCA-GA-SVM optimized parameters for the gamma (γ) vs. RMSE

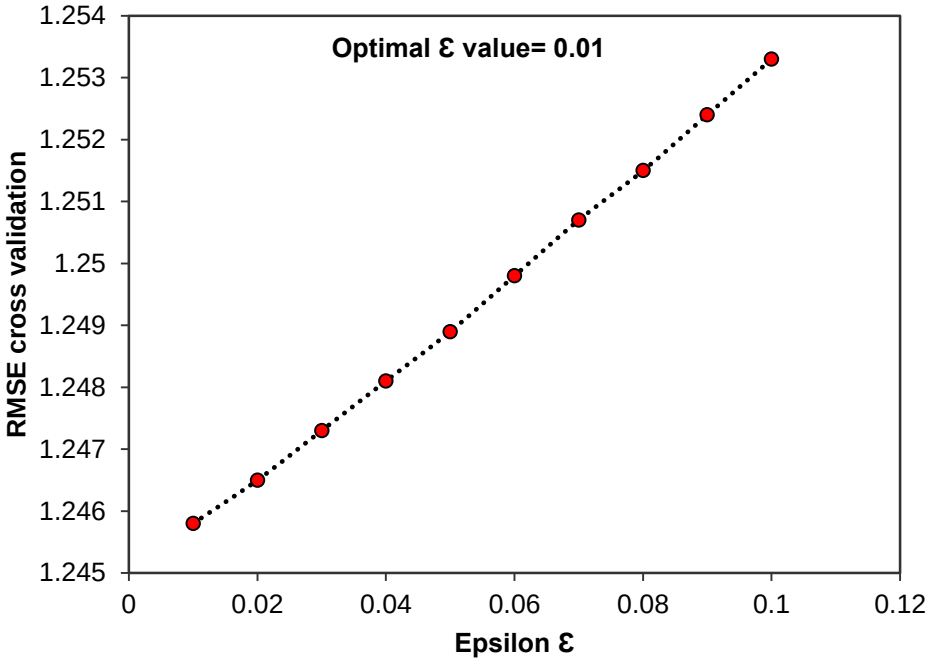


Figure 21: PCA-GA-SVM optimized parameters for the epsilon (ε) vs. RMSE

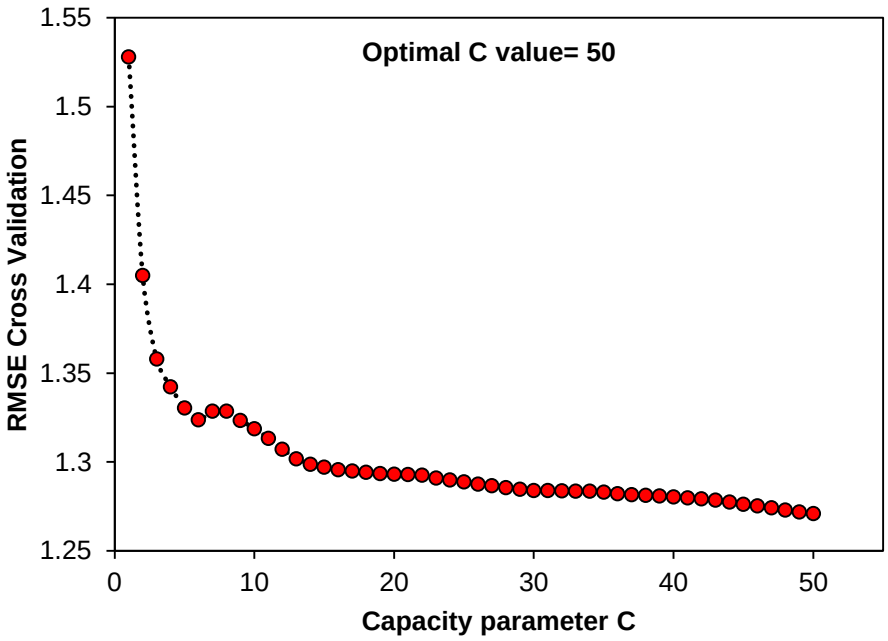


Figure 22: PCA-GA-SVM optimized parameters for the capacity (C) vs. RMSE

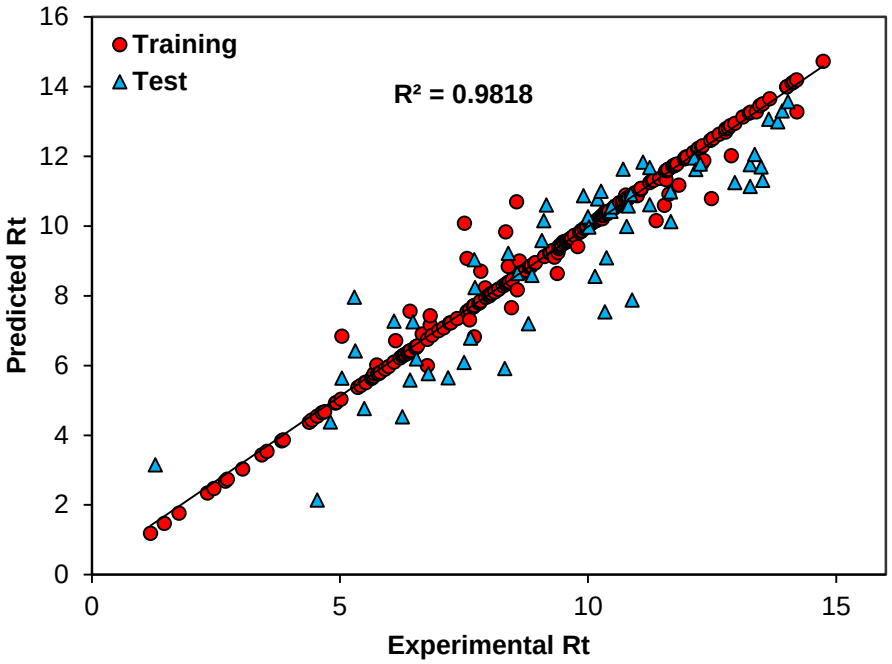


Figure 23: The plot of predicted retention time against the observed retention time values based on PCA-GA-SVM

5.1.7 kNN-SW-SVM

The non-linear model was built based on the same selected compounds as training set in KNN-SW-MLR, and the optimized parameters were calculated as follows (figures 24-26):

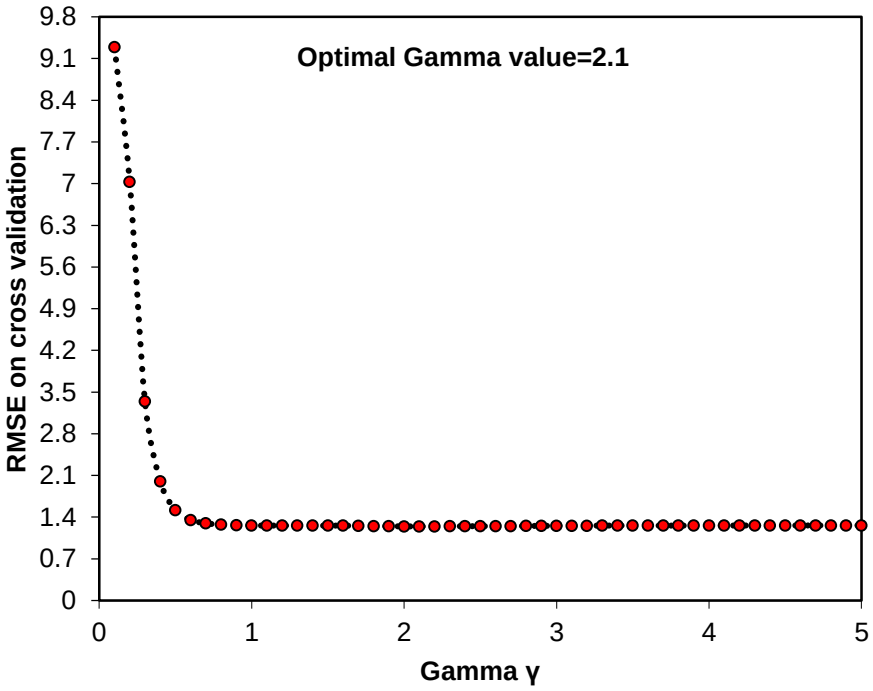


Figure 24: kNN-SW-SVM optimized parameters for the gamma (γ) vs. RMSE

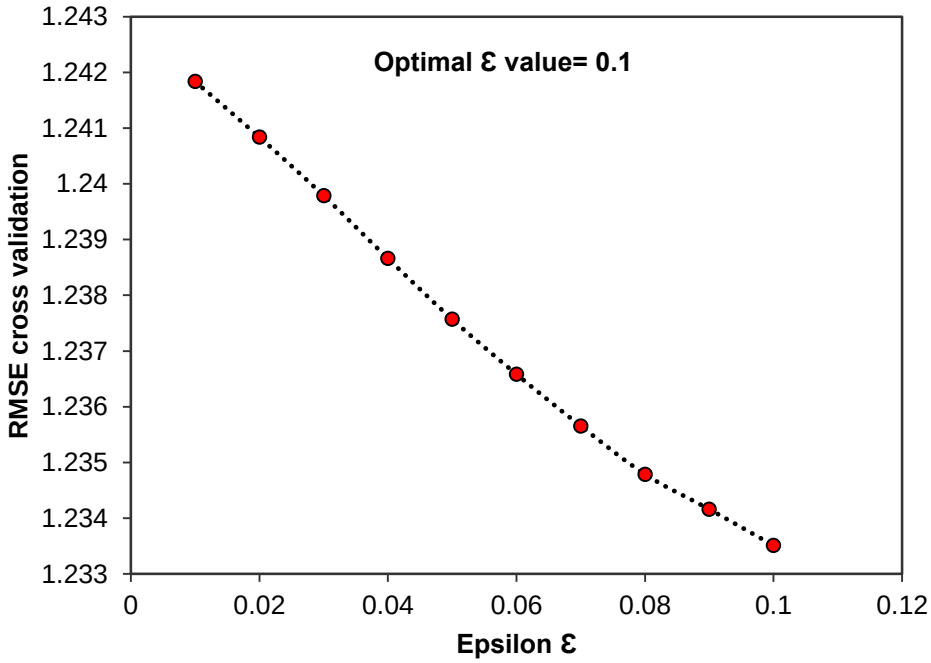


Figure 25: kNN-SW-SVM optimized parameters for the epsilon (ε) vs. RMSE

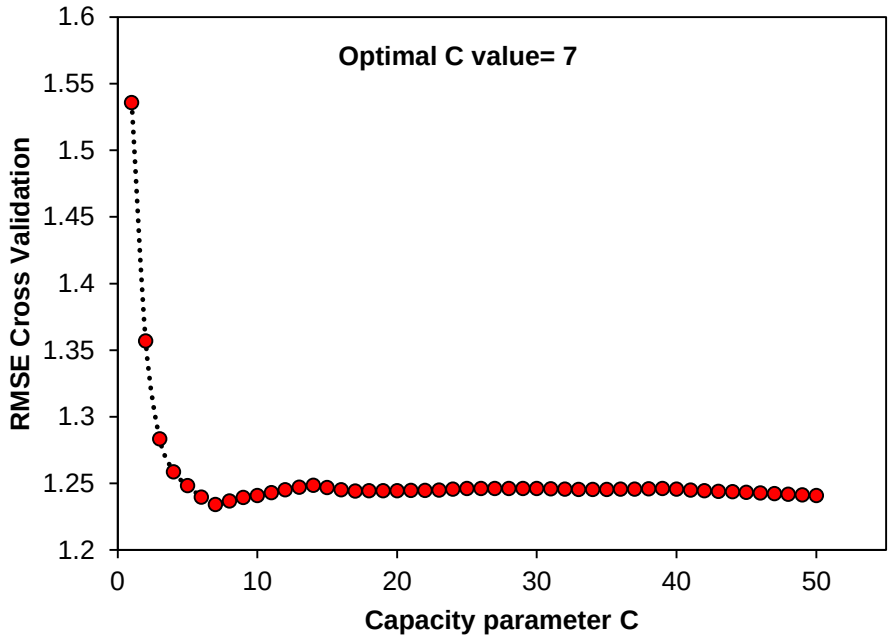


Figure 26: kNN-SW-SVM optimized parameters for the capacity (C) vs. RMSE

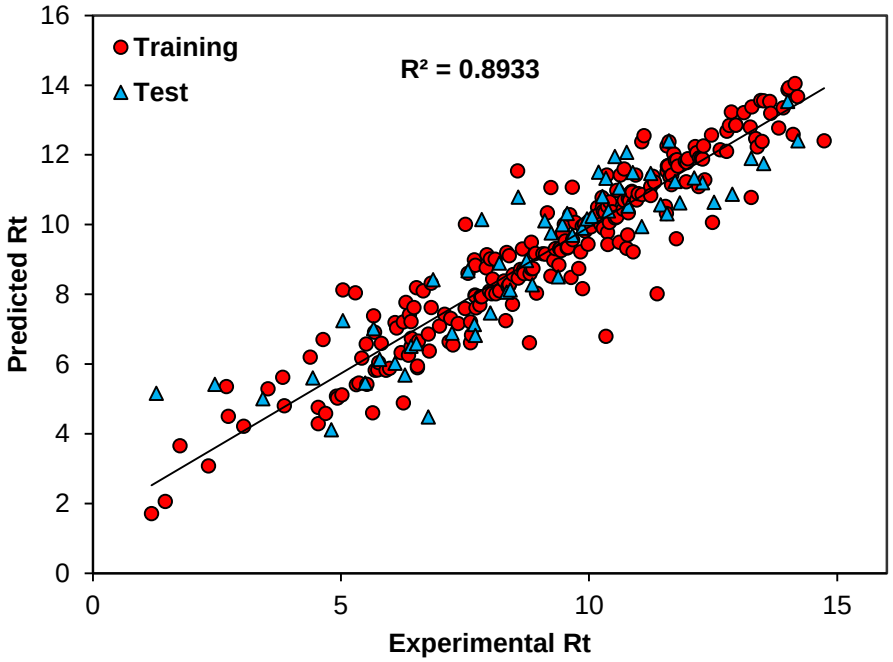


Figure 27: The plot of predicted retention time against the observed retention time values based on kNN-SW-SVM

5.1.8 kNN-GA-SVM

The non-linear model was built based on the same selected compounds as training set in kNN-GA-MLR, and the optimized parameters were calculated as C=45, ε=0.04, γ=2.0. The result of each optimization was shown in figures 28-30. The predicted values for retention time by kNN-GA-SVM method were given in Table S1, and then plotted versus the observed retention time and shown in figure 31.

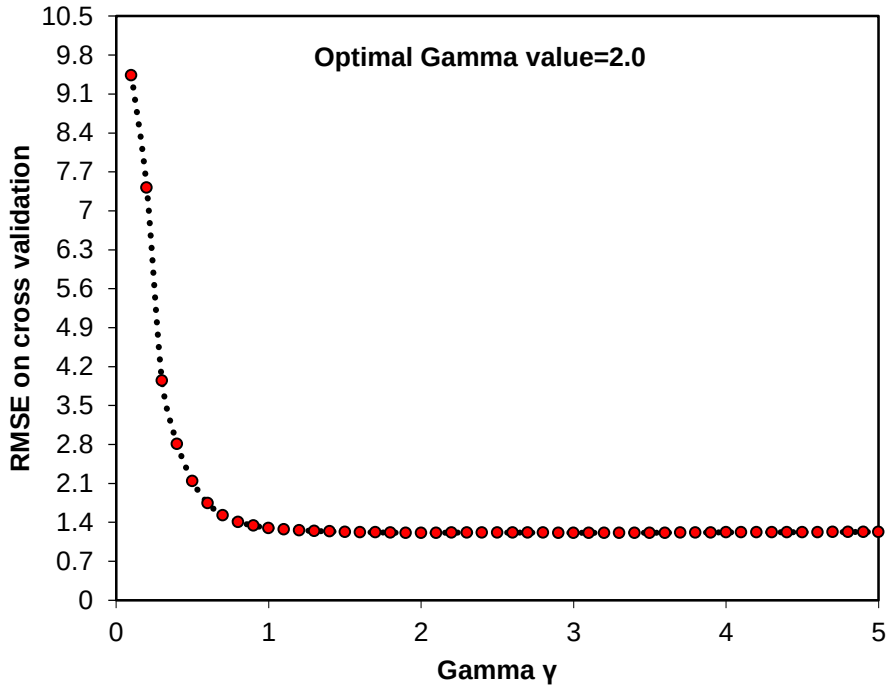


Figure 28: kNN-GA-SVM optimized parameters for the gamma (γ) vs. RMSE

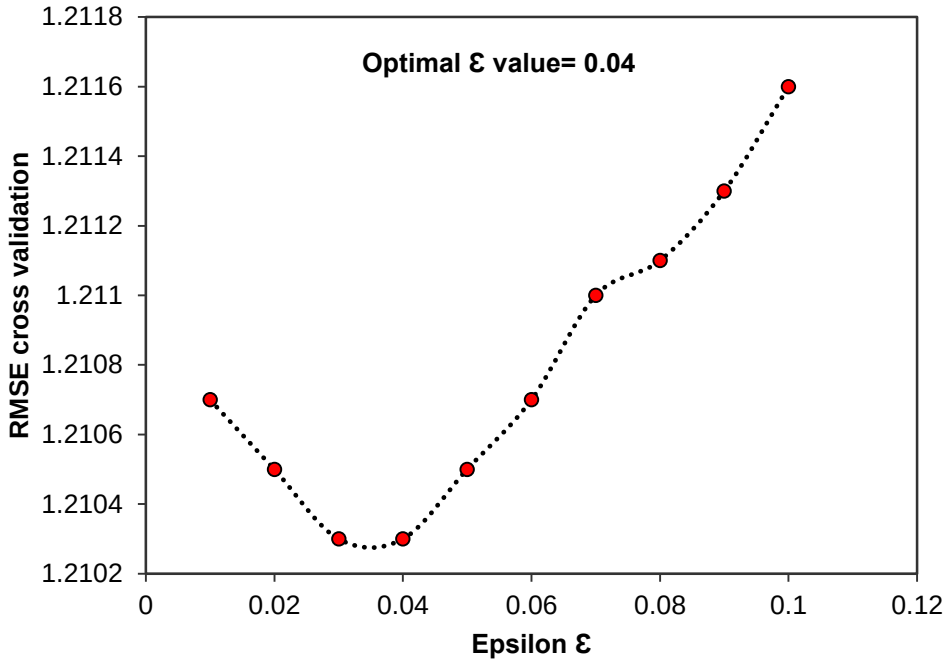


Figure 29: kNN-GA-SVM optimized parameters for the epsilon (ε) vs. RMSE

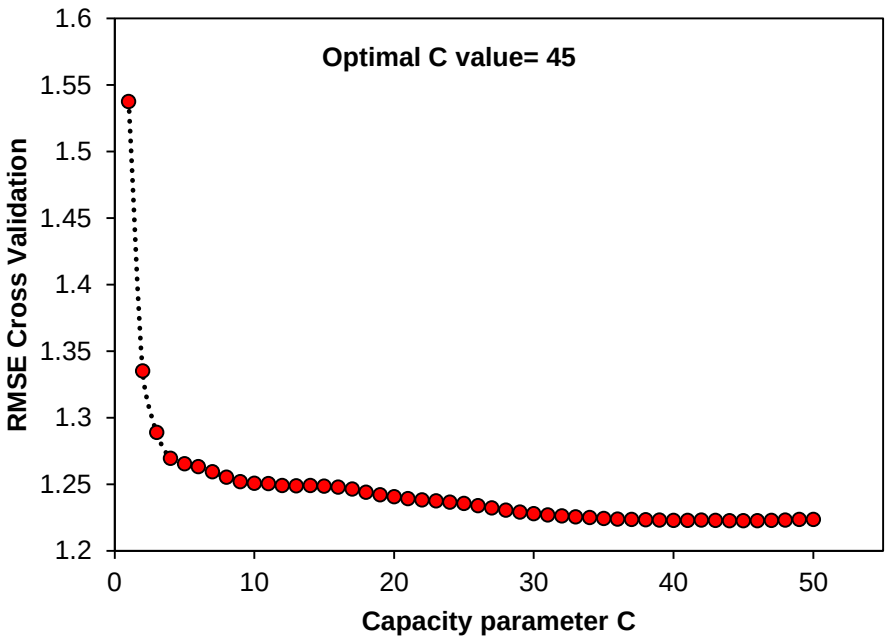


Figure 30: kNN-GA-SVM optimized parameters for the capacity (C) vs. RMSE

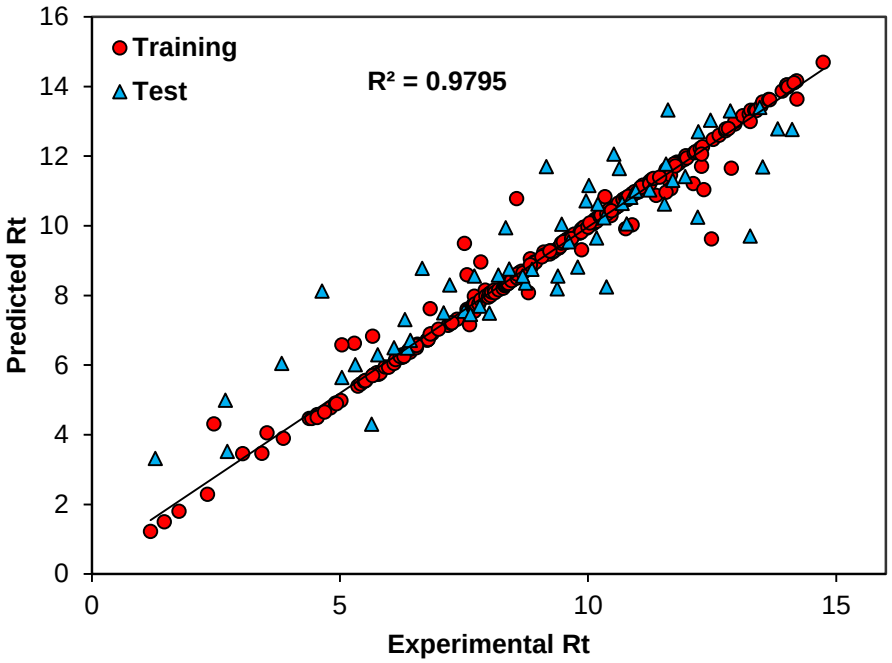


Figure 31: The plot of predicted retention time against the observed retention time values based on kNN-GA-SVM

The statistical results of this model were listed in Table 8. The comparison of built models is suggesting that kNN-GA-SVM is the most appropriate non-linear model for the prediction purposes, however PCA-GA-SVM can also be employed. From the linear models, both kNN-GA-MLR and kNN-SW-MLR can be used. The final validations for these models were carried out using Golbraikh and Tropsha acceptable model criteria's. The results are shown in Table 11.

Table 11:Golbraikh and Tropsha acceptable model criteria's for MLR and SVM

	kNN-SW-MLR	**kNN-GA-MLR**	**kNN-GA-SVM**
Condition I	0.822	0.806	0.833
Condition II	0.829	0.835	0.772
Condition III	K=0.9959 K′= 0.9877 $R^2 - {R_0}^2/R^2$ = 0.0048 ${R_0}^2 - {R_0}'^2/R^2$ = 0.0881	K=0.9966 K′= 0.9866 $R^2 - {R_0}^2/R^2$ = 0.0140 ${R_0}^2 - {R_0}'^2/R^2$ = 0.1202	K=0.9858 K′= 0.9979 $R^2 - {R_0}^2/R^2$ = 0.0077 ${R_0}^2 - {R_0}'^2/R^2$ = 0.0927
Condition IV	${R_0}^2 - {R_0}'^2$= 0.06845	${R_0}^2 - {R_0}'^2$= 0.08869	${R_0}^2 - {R_0}'^2$= 0.07082
Acceptance	Passed	Passed	Passed

5.1.9 kNN-GA-ANN

Since the models based on kNN and genetic algorithm showed appropriate internal and external results, the non-linear model based on ANN was developed based on kNN-genetic algorithm technique. It is accepted that for the generation of ANN models employing variable selection is not necessary, but it can be useful to get better results. Therefore, we used genetic algorithm for descriptors subset selection in ANN. The common problem with ANN, is to select the right node where in most literature the RMSE values are being considered for final model construction. Here, we reported and selected the ANN model based on the modified r^2 value, CCC value, and RMSE. Therefore, considering the over-fitting problem in higher nodes, the right nodes can be selected using their CCC values first that encodes the accuracy and precision, and then provided modified r^2 value for test set to select the nodes. Finally, for the couple of nodes with acceptable results for the test set, the one which shows also less RMSE value for the training set can be selected as the final node for subsequent analysis. The results of this procedure are shown in Table 12. From Table 12, it can be seen that the model built based on node=7 shows the highest CCC value for both the test and the training set, and the modified r^2 value for test set is the highest one among the other nodes. Considering the RMSE value between node 5 and 7, consequently the model based on 7 nodes is being selected. The MPD values for training set were calculated for the all nodes and given in Table 12 as follows:

$$MPD = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (Eq.24)$$

where y_i is the observed retention time, and $\hat{y}_i$ the calculated retention time and N denotes the number of data points. This formula measures the accuracy of the generated models based on each node and the lower value indicate the good fitted point. The predicted values based on kNN-GA-ANN are listed in Table S1 and their strength as prediction tool are compared in Table 13. The correlation plot of observed and predicted retention time is shown in figure 32.

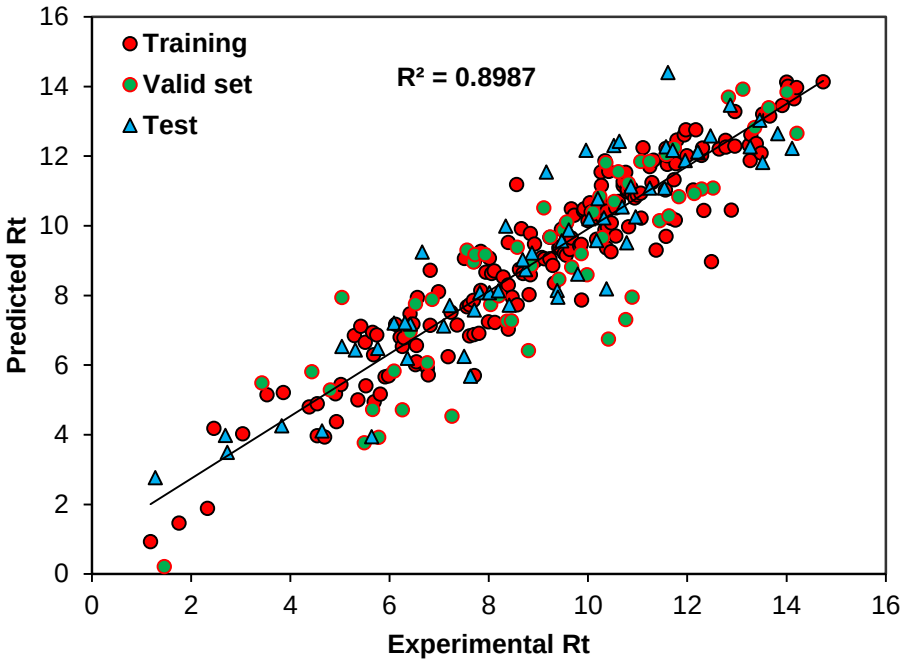


Figure 32: The plot of predicted retention time against the observed retention time values based on kNN-GA-ANN

Table 12:Comparisons of the developed models for negative ionization compounds based on different nodes in ANN

	Test					Valid				Train				
	R^2m	CCC	F	RMSE	R^2	CCC	F	RMSE	R^2	MPD	CCC	F	RMSE	R^2
Node1	0.790	0.907	34.83	1.157	0.848	0.9	35.52	1.197	0.814	12.066	0.905	117.06	1.15	0.827
Node2	0.624	0.900	31.62	1.216	0.83	0.907	38.61	1.158	0.826	11.503	0.912	128.25	1.11	0.839
Node3	0.782	0.888	30.55	1.333	0.795	0.879	29.25	1.321	0.774	11.623	9160	133.81	1.085	0.846
Node4	0.814	0.9	33.22	1.244	0.819	0.882	28.69	1.272	0.786	10.655	0.928	158.37	1.014	0.865
Node5	0.827	0.923	46.39	1.127	0.853	0.888	32.34	1.285	0.794	8.649	0.94	193.91	0.929	0.887
Node6	0.497	0.789	17.50	1.888	0.624	0.892	32.91	1.253	0.798	9.025	0.938	186.68	0.946	0.883
Node7	0.834	0.92	44.54	1.141	0.85	0.878	31.45	1.378	0.777	8.667	0.947	217.98	0.88	0.899
Node8	0.668	0.879	31.11	1.451	0.772	0.839	20.79	1.489	0.712	8.067	0.955	261.18	0.809	0.914
Node9	0.594	0.870	31.40	1.566	0.765	0.832	22.67	1.614	0.697	7.138	0.961	303.18	0.76	0.924
Node10	0.549	0.828	22.11	1.735	0.687	0.849	26.56	1.573	0.73	7.031	0.961	302.89	0.756	0.925
Node11	0.484	0.821	24.15	1.918	0.695	0.86	25.05	1.416	0.742	7.058	0.959	287.77	0.776	0.921
Node12	0.59	0.867	30.41	1.575	0.76	0.816	21.08	1.702	0.668	6.548	0.967	360.89	0.696	0.937
Node13	0.693	0.855	23.48	1.513	0.736	0.878	28.56	1.315	0.777	5.77	0.969	390.44	0.673	0.941
Node14	0.487	0.784	17.02	1.907	0.622	0.842	23.27	1.535	0.711	6.446	0.968	368.9	0.693	0.937
Node15	0.564	0.828	21.58	1.713	0.689	0.827	23.19	1.687	0.691	5.732	0.974	459.1	0.621	0.949
Node16	0.612	0.858	28.50	1.629	0.748	0.773	17.26	1.897	0.601	4.77	0.976	509.32	0.592	0.954
Node17	0.506	0.812	20.66	1.836	0.661	0.867	27.71	1.407	0.757	5.363	0.976	509.27	0.593	0.954
Node18	0.349	0.715	14.05	2.298	0.516	0.864	29.94	1.503	0.767	5.182	0.975	485.95	0.606	0.952
Node19	0.69	0.886	34.06	1.425	0.79	0.826	20.05	1.575	0.684	5.463	0.974	462.28	0.618	0.95
Node20	0.612	0.867	29.70	1.559	0.755	0.837	24.95	1.645	0.735	5.098	0.976	494.3	0.601	0.953

Table 13: Comparison of the developed models for negative ionization compounds

	Test					Train				
models	**R^2m**	**CCC[a]**	**RMSE**	**F**	**R^2**	**CCC[a]**	**Q^2_{LOO}**	**RMSE**	**F**	**R^2**
kNN-SW-MLR[b]	0.77	0.8941	1.209	29.406	0.822	0.914	0.829	1.107	176.95	0.842
kNN-GA-MLR[b]	0.745	0.8935	1.228	27.74	0.835	0.9013	0.806	1.169	152.01	0.820
kNN-SW-SVM	0.749	0.8876	1.234	27.045	0.818	0.9384	0.545	0.9234	238.75	0.893
kNN-GA- SVM[c]	0.767	0.8983	1.222	30.634	0.833	0.9891	0.772	0.4006	1468.84	0.980
PCA-SW-MLR	0.724	0.8791	1.367	28.301	0.782	0.9216	0.844	1.0527	195.523	0.854
PCA-GA-MLR	0.721	0.8775	1.379	28.079	0.786	0.896	0.789	1.2009	143.94	0.812
PCA-SW-SVM	0.786	0.8978	1.247	32.848	0.818	0.9839	0.653	0.4856	995.05	0.970
PCA-GA- SVM	0.763	0.8967	1.271	33.565	0.824	0.9907	0.538	0.3745	1755.2	0.982
	R^2m	CCC[a]	RMSE	F	R^2	CCC[a]	MDP	RMSE	F	R^2
kNN-GA-ANN	0.834	0.9203	1.141	44.544	0.850	0.9465	8.667	0.88	217.983	0.899

[a] Concordance correlation coefficient
[b] The best linear model
[c] The best non-linear model

5.1.10Interpretation of Molecular descriptors

The descriptors which were selected by the models are so important to be interrelated since each of them describes the molecular structure properties and its relationship with retention time. Therefore, by understanding their effect and definitions, the other compounds and their possible retention time can be provided. Here, since the model based on genetic algorithm-SVM showed appropriate results, the descriptors selected by genetic algorithms are being discussed. The relative importance of selected descriptors is shown in figure 37.

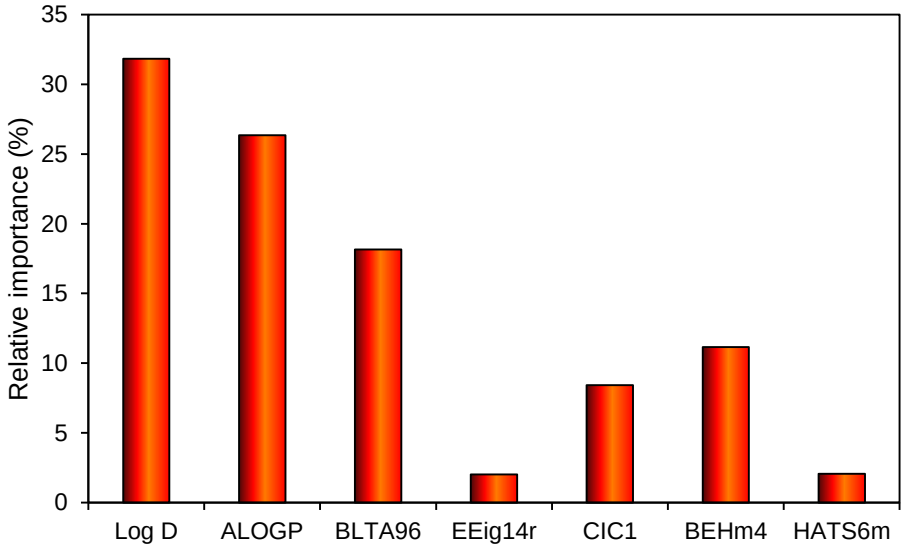


Figure 33: The relative importance of selected molecular descriptors

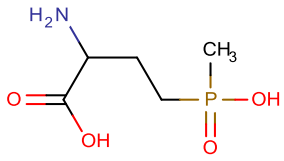
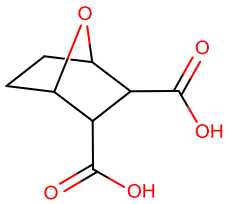
The first selected descriptor based on the genetic algorithms is LogD (pH at 6.20). By definition logP refers to neutral molecules. If a molecule contains basic or acidic groups, it can become ionized in the mobile phase and its distribution in octanol-water becomes pH-dependent. The pH-dependent distribution coefficient is defined as logD and it is calculated from the following equation (equation 25):

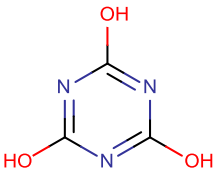
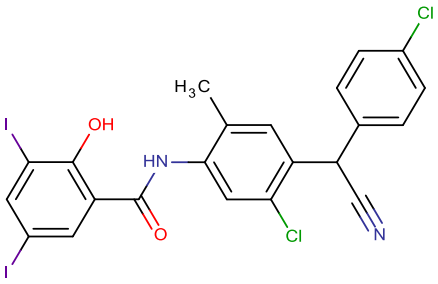
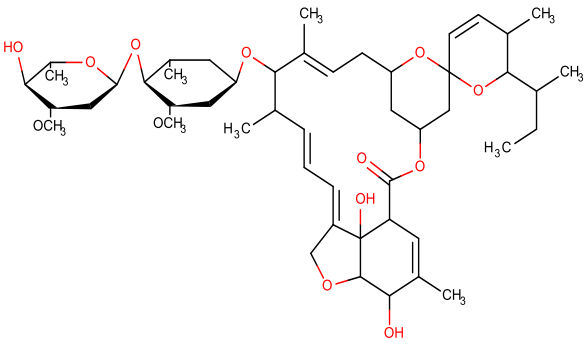
$$logD(pH) = logP - \log\left(1 + 10^{(pH-pKa)\Delta_i}\right) \qquad (Eq.25)$$

Where $\Delta_i = \{1, -1\}$is for acids and bases, respectively. The distribution coefficient, D, is a pH dependant measure of the propensity of a molecule to differentially dissolve in two immiscible phases, taking into account all ionized and unionized forms (microspecies). In our work, to obtain the logD values for each compound, ChemAxon package was used at

pH=6.2, so as to enable the model for better predictions by considering the ionized status of a molecule which contains basic or acidic groups. Both logD and logP are two main factors for risk assessment, drug design and toxicity of compounds since their values would help us to understand any properties of molecule in different conditions. As it can be seen from the linear equation 23, LogD is in direct relationship with retention time, where the lower value of logD would cause decrease in retention time value, too. To understand its effect clearly, we can use the definition of LogS (solubility), where at the certain pH, the compounds with high solubility should indicate the lower LogD. Therefore, based on the molecular structures, its solubility and LogD, the effect of LogD on retention time can be easily interpreted. Some compounds were selected from Table S1 (among the compounds of negative ionization) to investigate this effect (Table 14). LogS values were calculated using ChemAxon package [56].It can be seen that compounds with the lower LogD (-6.03) has the higher solubility (LogS (pH=6.2) =3.99), and hence, it results to the decrease of the retention time. Therefore, it is expected that compounds with lower retention time, have more solubility. Some more examples were added to the Table 14. If we consider molecule M261, it can be seen that the observed value presents the lower LogS, and therefore, it is being expected to have retention time at higher retention time.

Table 14: Effect of LogD and AlogP on retention time.

Mol.	Chemical Structure	Exp.Rt (min)	Log D	Log S	ALogP	BLTA96
M92		1.18	-6.03	3.99	-2.28	1.72
M64		1.28	-3.74	1.74	-0.381	-1.17

Mol.	Chemical Structure	Exp.Rt (min)	Log D	Log S	ALogP	BLTA96
M299		1.76	1.07	-1.15	0.706	-1.21
M261		13.12	7.61	-7.24	6.73	-6.83
M4		14.74	5.94	-6.68	5.04	-3.43

The second descriptor is ALogP (Ghose-Crippen octanol-water partition coefficient) which belongs to molecular properties descriptors, it is a measure of the lipophilicity of the molecule, and it is estimated using the Ghose–Crippen contribution method based on the hydrophobic atomic constants of atoms in the molecule [20, 57, 58]. Lipophilicity indicates the affinity of a molecule or a moiety for a lipophilic environment. The hydrophobicity represents the meaning of the association of non-polar groups or molecules in an aqueous environment which arises from the tendency of water to exclude non-polar molecules. In other words, the lipophilic character can affect the retention time significantly: the higher ALogP is, the higher retention is observed in C18 columns. As it can be seen from Table 14, LogD and ALogP have the same effect on the retention time, but since in ALogP the ionized effect of compounds is not being considered, the obtained values were less significant than LogD values. Therefore, compounds with high logP values have low

hydrophilicity, and since it is in direct relationship with the retention time: the higher AlogP would present higher retention time, as expected on a C18 column.

The next descriptor is BLTA96 (Verhaar Algae base-line toxicity from MLOGP (mmol/l)) in which is actually the toxicity index of given compounds against algae[20]. In the aquatic environment[59], there are at least 19 different models for the determination of toxicity. The DRAGON software has implemented enumerating indicators: toxicity in fish, daphnia and algae. In our model, it is a correlation between the retention times of the indicator of toxicity, expressed relative to the algae according to Verhaar Algae model [60]. In numerous biological studies, it has been proven that algae in the aquatic environment act as detoxification device (a kind of a "green liver").The paths of metabolism of xenobiotics in algae are close to the corresponding metabolic pathways in mammal body [60]. Apart from this reason, the used compounds (mostly pesticides) have also demonstrated the dominant toxicity where the selection of BLTA96 descriptor seems to be rational. This descriptor demonstrated the negative effect in linear equation and indicating that the increase of BLTA96 of compounds would results in lower retention time. In our dataset the value of this descriptor is ranged between -8.05 (most toxic compound) and 1.72 (least toxic compound) suggesting that compound with less BLTA96 is more lipophilic and thus resulting in decrease of polarity and increase of retention time. The BLTA96 values for some compounds were shown in Table 14, and, as it can be seen, M92 has the lowest retention time among the other compounds, and it presents the higher BLTA96 value.

The next selected descriptor is Eigenvalue 14 from edge adj. matrix weighted by resonance integrals (EEig14r) which belongs to the edge adjacency indices and encodes the connectivity between graph edges [20]. The edge adjacency matrix denoted as ${}^{e}A$ and shows the whole set of connections between pairs of atoms in which is calculating as follows:

$$[A]_{ij} = \begin{cases} 1 \; if (i,j) \in E(G) \\ 0 \; otherwise \end{cases} \qquad (Eq.26)$$

where1 is showing that the atoms in i and j were bounded, while otherwise is zero. Resonance effect is a kind of energy stabilizing due to the delocalization of electrons in a bond network available in a compound. It can cause the mesomeric effect (i.e

delocalization of π electrons in its π orbital) and secondary mesomeric effect which is the repulsion of the π electrons by non-bonded electrons on solvent or special substituent. As it can be seen, this descriptor represented negative effect in the model. This means that increase of the EEig14r value would reduce retention time.

The fifth descriptor is Complementary Information Content index (neighborhood symmetry of 1-order (CIC1). The information content index descriptors are calculated based on the pair wise equivalence atoms in a Hydrogen-filled molecule [20]. A pair of atoms are said to be equivalent at a particular level-r, if they are of the same element and their neighborhood is equivalent up to level-r. For the CIC_r, the *r*-th order measures the deviation of IC_r from its maximum value. It corresponds to the vertex partition into equivalence classes that are including one element each. CIC_r is calculating based on the equation 27.

$$CIC_r = \log_2 A - IC_r \qquad (Eq.27)$$

where A is the atom number. IC_r is defined below:

$$IC_r = -\sum_{g=1}^{G} \frac{A_g}{A} . \log_2 \frac{A_g}{A} = -\sum_{g=1}^{G} P_g . \log_2 P_g \qquad (Eq.28)$$

where g runs over the G equivalence classes, A_g is cardinality of the *g*th equivalence class, ***A*** is the total number of atoms, and p_g is the probability of randomly selecting a vertex of the *g* th class. It represents a measure of structural complexity per vertex. This descriptor showed a positive effect on the retention time: the increase of this descriptor (presence of two or more vertices that topologically equivalent with the same coordinates) would cause an increase to the retention time.

The sixth selected descriptor is BEHm4 (highest eigenvalue n.4 of Burden matrix / weighted by atomic masses). This descriptor is encoding the Burden eigenvalues descriptor and is calculated based on hydrogen included molecular graph weighted by atomic masses[61]. The positive sign of this descriptor (see equation 23) suggests that the retention time values are directly related to this descriptor positively. Increasing the atomic mass by adding more hydrogen atoms in the molecular structure would result to the increase of the retention time.

The last selected descriptor is leverage-weighted autocorrelation of lag 6/weighted by mass (HATS6m) and belongs to the GETAWAY H-indices descriptors family, and explains the

influence of atomic mass over probability interaction of leverage[62]. The GETAWAY (Geometry, Topology, and Atom-Weights Assembly) descriptors have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the Molecular Influence Matrix (MIM) [62]. Since the sign of this descriptor is positive in equation 23, the increase in its value by increasing the mass of compound would result in increase of retention time.

5.1.11 Applicability domain study of kNN-GA-MLR model for suspects

Some compounds (as suspect compounds) were used as evaluation set so as to predict their retention time based on the developed models. The results of prediction for these compounds along with their experimental determined retention time were listed in Table 15. Among the suspect compounds, some significant residuals were observed. The William plot and Euclidean based applicability domain were used to calculate the standard residuals and normalized mean distance values as inputs for generating the visualization of the outliers. This display would help to understand the origin of outliers more easily. Boxes based on the training and test set (figure 34) are presented, and then for the taken compounds as suspect list, the analysis was carried out. Results indicates that out of 63 as suspect compounds, 30 compounds were predicted very well and 33 compounds are belonged to box3 and box4. The results of the analyses were listed in Table 16. Considering these results and the visualization plot (figure 35), it can be concluded that out of 20 compounds in box4, six compounds (Oxadiazon, Carbuterol, Pivenfrine, Amoxecaine, Hexamidine, 4-Aminosalicylic acid) are within the applicability domain of models, but the suggested retention times are not matched with the structure and therefore, we can be sure that the suggested compound as suspect molecule cannot be correct. The compounds located in box 4 were shown in red color in figure 34.

Table 15: Retention time predicted values of of suspect compounds in negative ionization as evaluation set by kNN-GA-SVM

suspectlist	Exp. Rt	KNN-GA-SVM Predicted Rt

			The suggested structure can be accepted	The suggested structure is rejected
N1	metominostrobin	9.51	9.23	
N2	Carbofuran-3-hydroxy	6.22	7.38	
N3	Oxadiazon	1.38		13.08
N4	Carbaryl	7.21	9.05	
N5	Ancymidol	7.8	6.94	
N6	Citronellalhydrate	10.9	7.76	
N7	Linalylacetate	9.7	9.73	
N8	Crotethamide	4.8		9.1
N9	Diisopropyladipate	7.5		10.56
N10	Ethofumesate	5.4		8.83
N11	Carbuterol	12.6		4.67
N12	Etoxazene	5.3		9.72
N13	Furmecyclox	12.8	9.58	
N14	Pivenfrine	12.8		6.25
N15	Irone	13.3	11.15	
N16	Loxanast	13.2	12.46	
N17	Phenylacetylsalicylate	10.2	10.23	
N18	Menthylisovalerate	13.7	12.99	
N19	Mazindol	1.4		11.2
N20	Embelin	9.7	10.93	
N21	Amoxecaine	14.4		7.46
N22	Ipriflavone	13.5	10.72	
N23	Phenprocoumon	13.5	10.58	
N24	Dibenzylsuccinate	8.3	10.19	
N25	Metochalcone	8.3	10.44	

	suspectlist	Exp. Rt	KNN-GA-SVM Predicted Rt	
			The suggested structure can be accepted	The suggested structure is rejected
N26	γ-Linolenicacid	14	13.73	
N27	Hexyldodecanoate	14.9	14.08	
N28	Stearicacid	14.9	13.95	
N29	Dodecylgallate	14.2	13.35	
N30	Piperonylbutoxide	14.2		10.07
N31	Hexamidine	13.4		7.21
N32	Neraminol	13.4		7.84
N33	Dehydroabieticacid	13.9	12.9	
N34	Isotretinoin	13.9	12.98	
N35	Metandienone	13.9		10.76
N36	Nordinone	13.9		10.48
N37	Norgesterone	13.9		9.96
N38	Norvinisterone	13.9		10.26
N39	Tretinoin	13.9		7.72
N40	Algestone	11.2		7.72
N41	Corticosterone	11.2		7.5
N42	Cortodoxone	11.2		7.88
N43	Doxaprost	15.5		11.31
N44	Canrenone	14.8		9.32
N45	Hydroxymethyleneprogesterone	14.8		8.88
N46	Norethindroneacetate	14.8		10.47
N47	Cloprostenol	14.1		8.34
N48	Etretinate	13.5	13.47	
N49	Melengestrol	13.5		9.38

	suspectlist	Exp. Rt	KNN-GA-SVM Predicted Rt	
			The suggested structure can be accepted	The suggested structure is rejected
N50	Medrogestone	14.8		11.09
N51	Desmethylmoramide	5.8		9.85
N52	Doxapram	5.8		9.5
N53	Fenoctimine	15.1	13.1	
N54	Hydrocortamate	14.6		8.94
N55	Dotarizine	13.5	11.92	
N56	Picricacid	6.7	7.14	
N57	DNOC_ 2_4-Dinitro-o-kresol	6.4	7.7	
N58	4-Aminosalicylic acid	8.3		3.51
N59	Caprylicacid/ Octanoicacid	8	8.29	
N60	Benzylformate	6.1	6.75	
N61	8-Hydroxychinolin	6.8	6.59	
N62	Caffeicacid	4.9	4.69	
N63	4-Hydroxyphenyl-pyruvic acid	4.9	3.71	

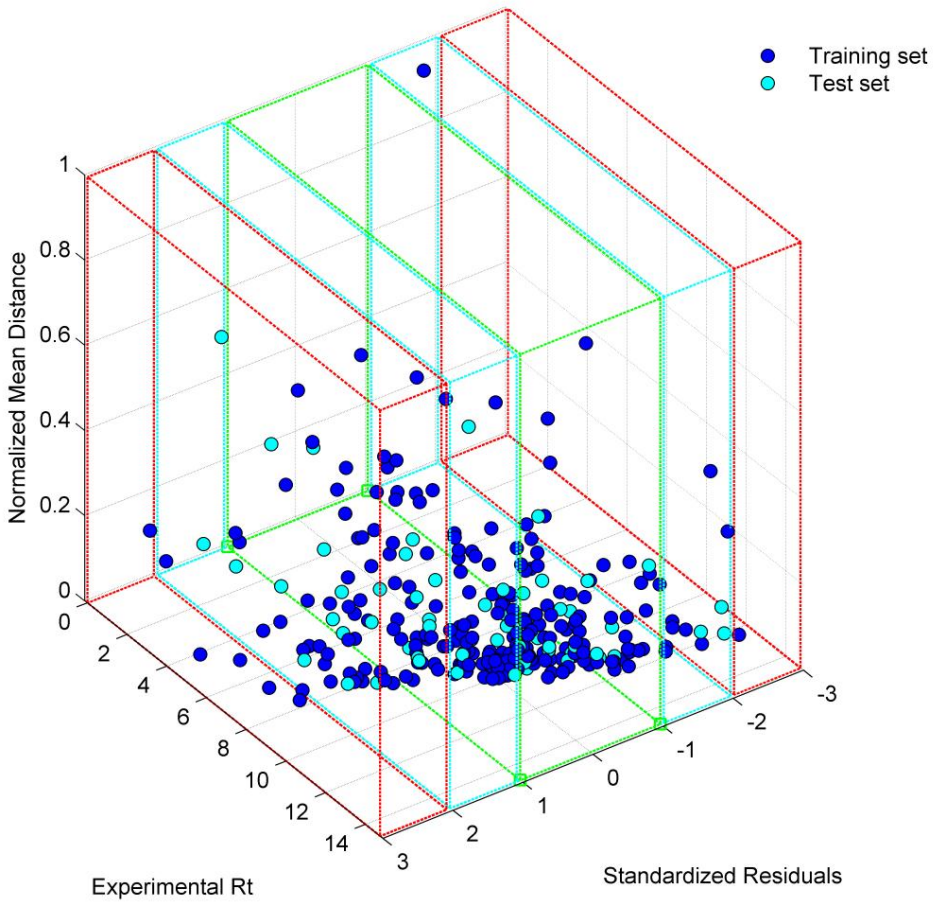


Figure 34: Visualization of data distribution

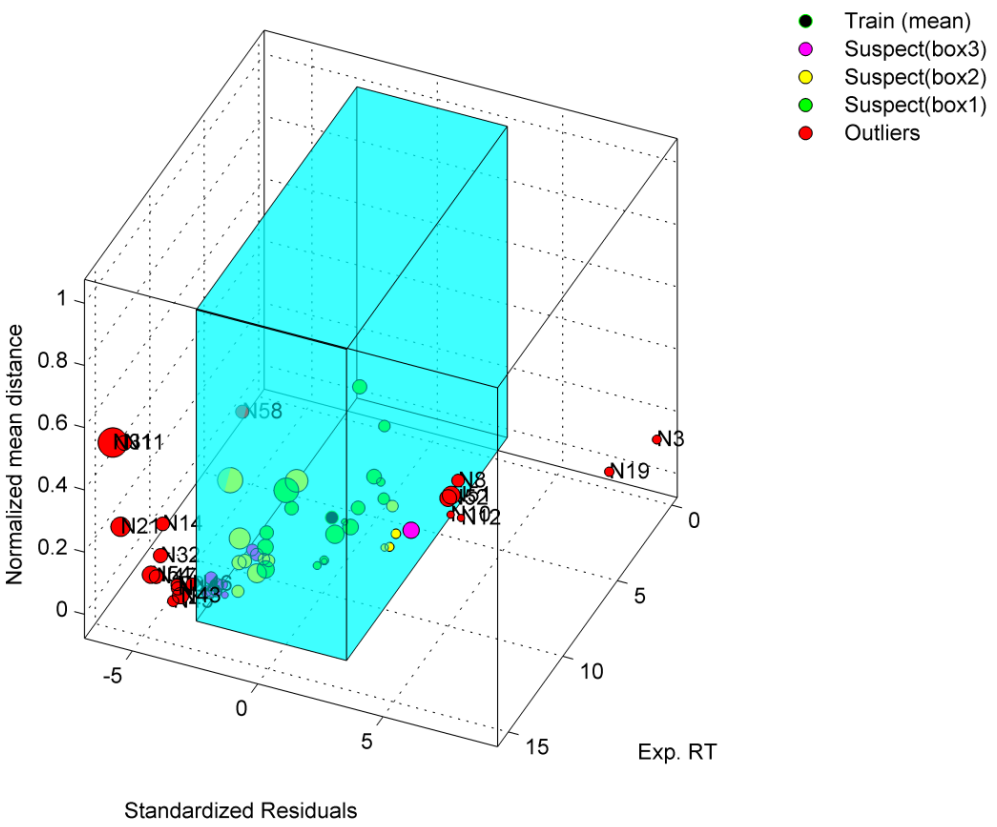


Figure 35: Origin of outliers for suspect compounds in negative ionization

Table 16: The analysis of visualization of outliers for linear model (kNN-GA-MLR)

Boxes	Origin of outliers	compounds
Box 3	The origin of residuals is mostly due to structural diversity. The model cannot predict their Rt	Citronellal hydrate, Diisopropyl adipate, Furmecyclox, Ipriflavone, Phenprocoumon, Metandienone, Nordinone, Norgesterone, Norvinisterone, Corticosterone, Melengestrol
	The origin of residuals is mostly due to Response. The suspect compounds are rejected.	Piperonyl butoxide, Medrogestone
Box 4	The origin of residuals is mostly due to structural diversity. The model cannot predict their Rt	Crotethamide, Ethofumesate, Etoxazene, Mazindol , Neraminol , Tretinoin, Doxaprost, Canrenone, Hydroxymethyleneprogesterone, Norethindrone acetate, Cloprostenol, Desmethylmoramide, Doxapram ,Hydrocortamate
	The origin of residuals is mostly due to Response. The suspect compounds are rejected.	Oxadiazon, Carbuterol , Pivenfrine, Amoxecaine, Hexamidine, 4-Aminosalicylic acid

5.2. Developed model for positive Electrospray Ionization Mode ((+)ESI)

5.2.1 PCA-SW-MLR

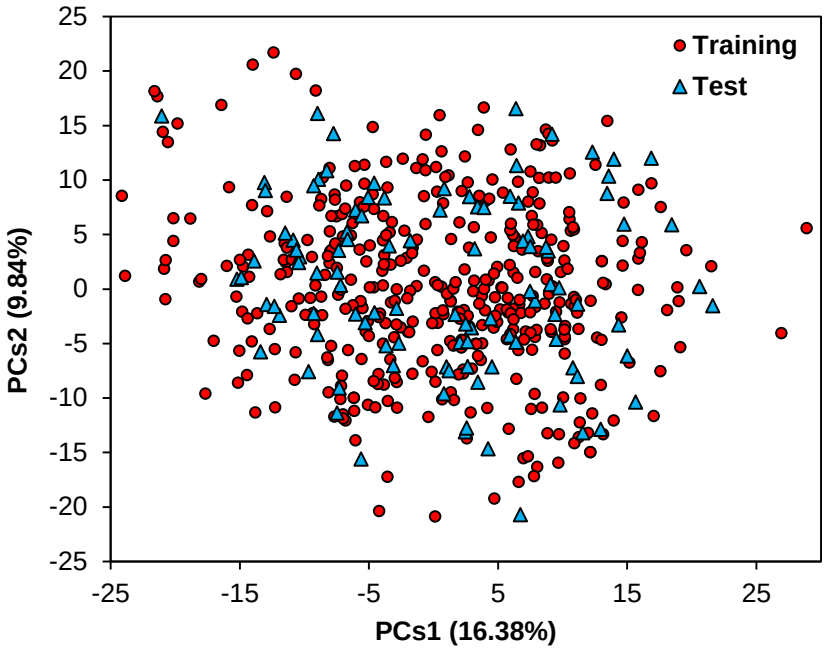


Figure 36: PCA analysis for the positive ionization compounds

Since the workflow was the same as that employed in negative ionization, here the results of each step were discussed in less details. The linear models based on stepwise variable selection tool and selected test set compounds by PCA and kNN are calculated initially before performing the genetic algorithms technique. The selected test set compounds based on each splitting techniques were shown in Table S1 (positive ionization). The model based on the PCA-SW-MLR is as follows:

Rt= 2.021 (±1.351) + 1.899(±1.897) Mv + 0.1021(±0.0291) RBN + 0.8486(±0.1384) CIC1 - 0.3978(±0.05838) C-025 + 0.0513(±0.01264) MLOGP2 + 1.685(±0.2639) B06[C-C] + 1.097(±0.05665) LogD (3.6) $(Eq. 29)$

N_{train}=422, R^2_{train}=0.846, $RMSE_{train}$=1.061, R^2_{adj}=0.843, F_{train}=324.49 Q^2_{LOO}=0.840, Q^2_{LGO}=0.478, Q^2_{BOOT}=0.838, N_{test}=105, R^2_{test}=0.843, $RMSE_{test}$=1.127, F_{test}=78.49, rm^2_{test} =0.765, CCC_{test}=0.9127, CCC_{train}=0.9165

The Y-randomization test was calculated, and the results were indicated that developed model is acceptable (Table 17).

Table 17:The Q^2_{Loo} and $R^2_{training}$ values after several Y-randomization tests for PCA-SW-MLR.

No	Q^2	R^2
1	0.0031	0.0286
2	0.0059	0.033
3	0.0153	0.0082
4	0.0104	0.0076
5	7.10E-06	0.0183
6	0.0083	0.0115
7	8.79E-05	0.0162
8	0.0027	0.0119
9	0.0047	0.0113
10	0.0304	0.0052

William plot detected 3 outliers (1 for training and 2 for test set) in model, and after removal of the detected outlier from the training set, the final predictive model obtained (figure 37).

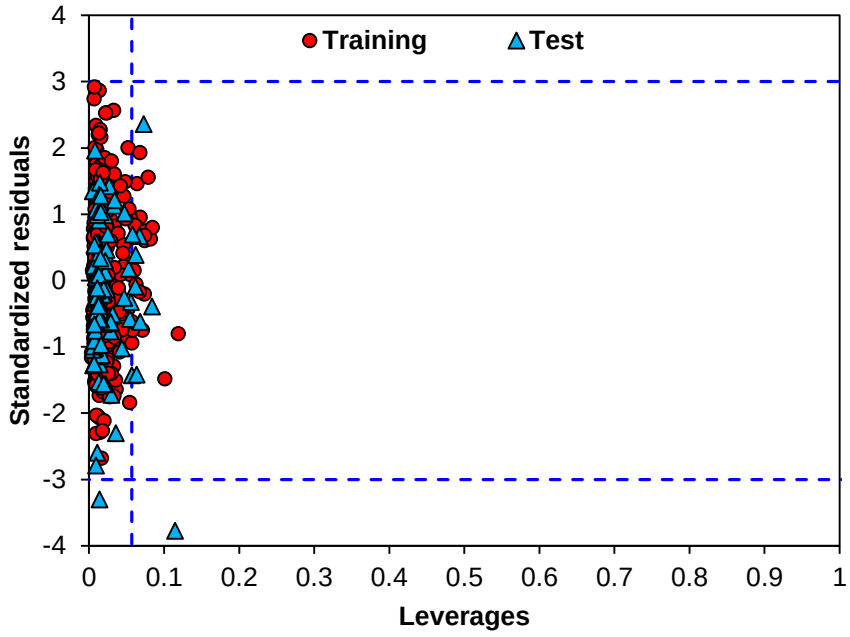


Figure 37: William plot of PCA-SW-MLR model (equation 29): h* warning leverage value is 0.056872.

VIF values for each selected descriptor along with correlation values between pair descriptors are listed in Table 18. The predicted retention time values using the equation 29 plotted versus the observed retention time values are shown in figure 38.

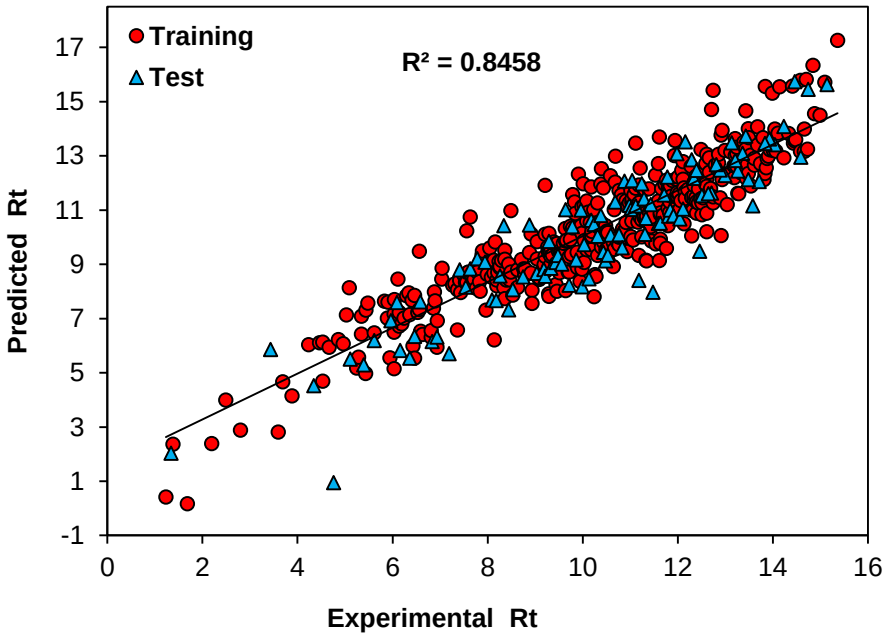


Figure 38: The plot of predicted retention time against the observed retention time values based on PCA-SW-MLR

Table 18: The correlation coefficient of selected descriptors and corresponding VIF values by PCA-SW-MLR

Variables	Mv	RBN	CIC1	C-025	MLOGP2	B06[C-C]	LogD (3.6)	VIF[a]
Mv	1	0	0	0	0	0	0	3.137
RBN	-0.35	1	0	0	0	0	0	1.775
CIC1	-0.57	0.536	1	0	0	0	0	2.547
C-025	0.223	-0.03	0.051	1	0	0	0	1.242
MLOGP2	0.363	0.151	0.168	0.417	1	0	0	2.22
B06[C-C]	0.013	0.236	0.214	0.138	0.197	1	0	1.204
LogD (3.6)	0.399	0.36	0.222	0.38	0.71	0.376	1	3.361

[a] Variation inflation factor

5.2.2 PCA-GA-MLR

The linear model based on genetic algorithms was also developed to compare the results. The model based on the PCA-GA-MLR is as follows:

Rt= 3.559(±0.3267) +0.9348(±0.06201) LogD(pH=3.6) -0.2956 (±0.0704) BLTA96 +0.1394 (±0.02849) RBN +0.00408(±0.00926) ALOGP2 -0.2621(±0.0686) nHDon +0.5871 (±0.1086) CIC1 +1.282(±0.2610) B06[C-C] $(Eq. 30)$

N_{train}=421, R^2_{train}=0.849, $RMSE_{train}$=1.048, R^2_{adj}=0.846, F_{train}=331.19 Q^2_{LOO}=0.842, Q^2_{LGO}=0.731, Q^2_{BOOT}=0.841, N_{test}=104, R^2_{test}=0.816, $RMSE_{test}$=1.154, F_{test}=59.00, rm^2_{test} =0.814, CCC_{test}=0.8966, CCC_{train}=0.9182

The Y-randomization test and VIF values were given in Table 19 and 20, respectively.

Table 19: The Q^2_{LOO} and $R^2_{training}$ values after several Y-randomization tests for PCA-GA-MLR

No	Q^2	R^2
1	0.0069	0.0341
2	0.0002	0.021

No	Q^2	R^2
3	6.79E-05	0.0179
4	0.0121	0.0085
5	0.0053	0.0123
6	0.0002	0.0155
7	0.0008	0.0147
8	0.0003	0.022
9	0.0027	0.0113
10	0.0027	0.0244

Table 20: The correlation coefficient of selected descriptors and corresponding VIF values by PCA-GA-MLR

Variables	LogD (3.6)	BLTA96	RBN	ALOGP2	nHDon	CIC1	B06[C-C]	VIF[a]
LogD (3.6)	1	0	0	0	0	0	0	4.499
BLTA96	-0.773	1	0	0	0	0	0	3.013
RBN	0.36	-0.131	1	0	0	0	0	1.762
ALOGP2	0.808	-0.722	0.307	1	0	0	0	3.470
nHDon	-0.37	0.297	-0.365	-0.222	1	0	0	1.276
CIC1	0.224	-0.201	0.549	0.221	-0.327	1	0	1.480
B06[C-C]	0.36	-0.343	0.230	0.204	-0.292	0.228	1	1.270

[a] Variation inflation factor

William plot detected 2 outliers (1 for training and 1 for test set) in model, and after removal of the detected outlier from the training set, the final predictive model obtained (figure 39). The predicted retention time values versus the observed retention time values for PCA-GA-MLR model are presented in figure 40.

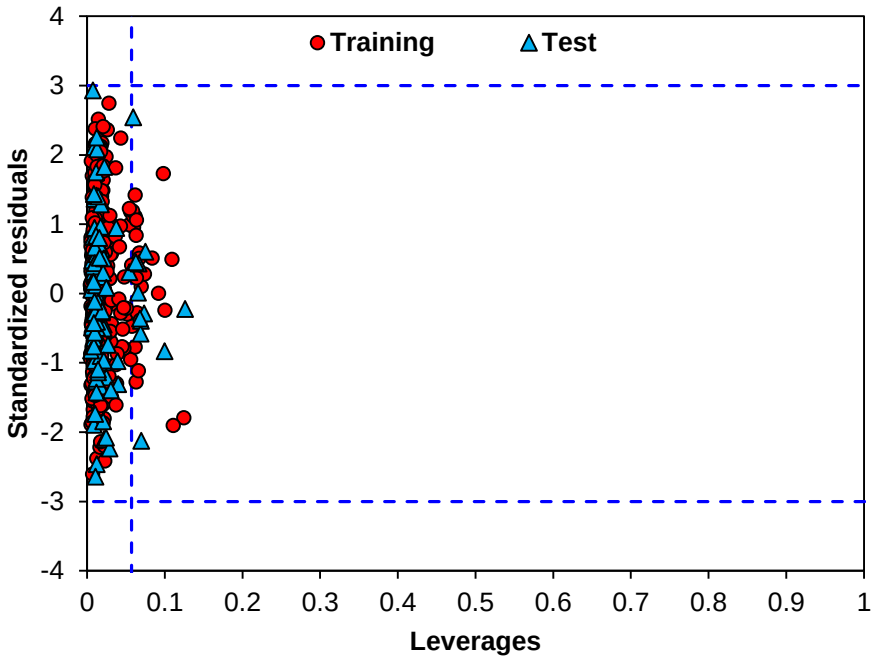


Figure 39: William plot of PCA-GA-MLR model (equation 30): h* warning leverage value is 0.057007.

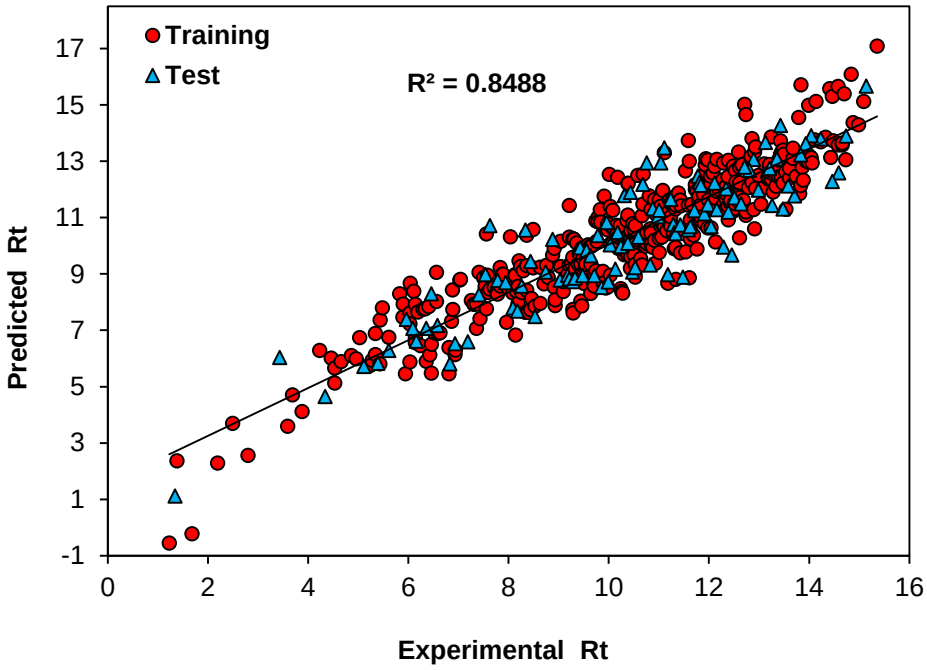


Figure 40: The plot of predicted retention time against the observed retention time values based on PCA-GA-MLR

5.2.3 kNN-SW-MLR

The same procedures were done for the data set split by kNN technique. The dendrogram for positive ionization can be found in the electronic supplementary material (figure S2).The results for kNN-SW-MLR were derived as follows:

Rt= 2.159(±1.367) +2.126(±1.9101) Mv +0.1228 (±0.0277) RBN +0.7832(±0.1381) CIC1 -0.409 (±0.0575) C-025 +0.0500(±0.0128) MLOGP2 +1.501(±0.23142) B06[C-C] +1.0854 (±0.0551) LogD(pH=3.6) $(Eq. 31)$

N_{train}=422, R^2_{train}=0.847, $RMSE_{train}$=1.051, R^2_{adj}=0.844, F_{train}=327.20 Q^2_{LOO}=0.841, Q^2_{LGO}=0.744, Q^2_{BOOT}=0.840, N_{test}=105, R^2_{test}=0.826, $RMSE_{test}$=1.162, F_{test}=67.03, rm^2_{test} =0.809, CCC_{test}=0.9050, CCC_{train}=0.9171

VIF values for each selected descriptor along with correlation values between pair descriptors are listed in Table 21. William plot detected two outliers for final kNN-SW-MLR model (2 compounds for test set) (figure 41). The predicted retention time versus the observed retention time values based on kNN-GA-MLR were shown in figure 42.

Table 21: The correlation coefficient of selected descriptors and corresponding VIF values by kNN-SW-MLR

Variables	Mv	RBN	CIC1	C-025	MLOGP2	B06[C-C]	LogD (3.6)	VIF[a]
Mv	1	0	0	0	0	0	0	3.137
RBN	-0.364	1	0	0	0	0	0	1.775
CIC1	-0.602	0.53	1	0	0	0	0	2.547
C-025	0.164	0.027	0.049	1	0	0	0	1.242
MLOGP2	0.366	0.142	0.144	0.425	1	0	0	2.220
B06[C-C]	0.007	0.244	0.175	0.169	0.225	1	0	1.204
LogD (3.6)	0.368	0.36	0.212	0.369	0.72	0.376	1	3.361

[a] Variation inflation factor

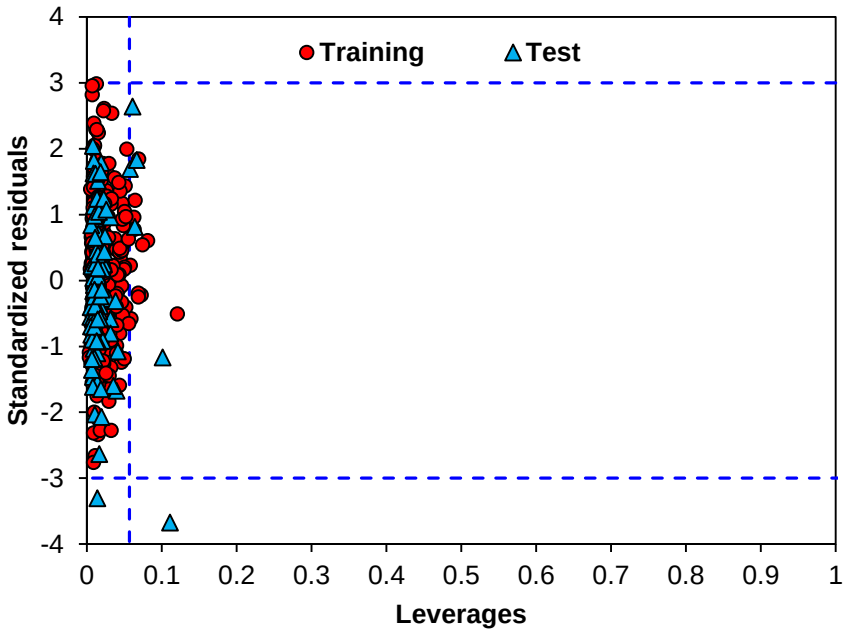


Figure 41: William plot of kNN-SW-MLR model: h* warning leverage value is 0.05687, namely.

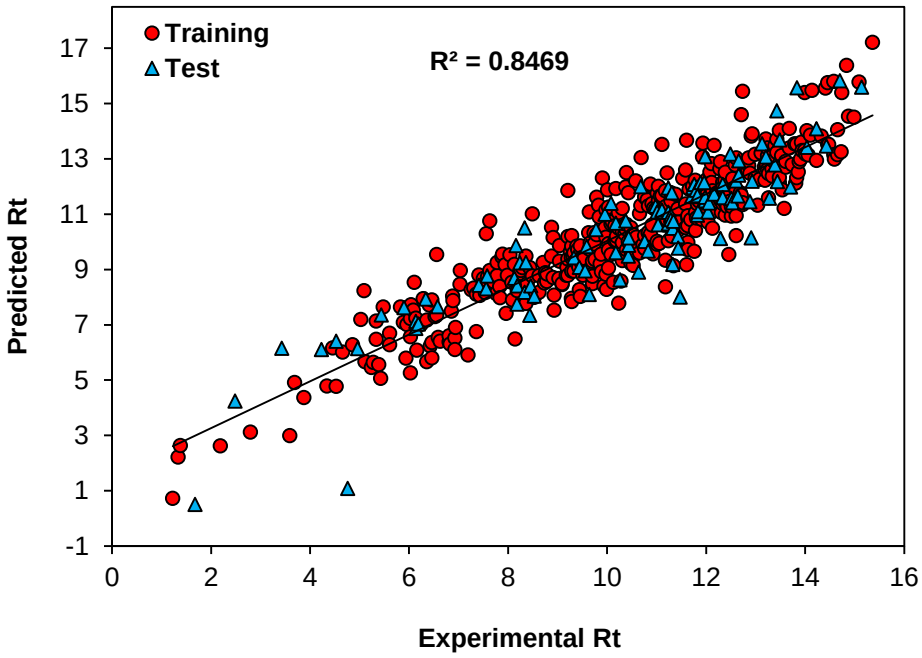


Figure 42: The plot of predicted retention time against the observed retention time values based on kNN-SW-MLR model

5.2.4 kNN-GA-MLR

The results for kNN-GA-MLR were obtained as below:

Rt= 3.442(±0.925) +0.8593(±0.0643) LogD(pH=3.6) -0.2826(±0.0715) BLTA96 +1.448(±0.248) B06[C-C] +0.3711 (±0.295) BEHp2 +0.0104(±0.0099) ALOGP2 +0.260 (±0.0250) RBN -0.0145(±0.00222) TPSA(NO) $(Eq. 32)$

N_{train}=422, R^2_{train}=0.840, $RMSE_{train}$=1.069, R^2_{adj}=0.838, F_{train}=310.17 Q^2_{LOO}=0.834, Q^2_{LGO}=0.798, Q^2_{BOOT}=0.832, N_{test}=105, R^2_{test}=0.846, $RMSE_{test}$=1.093, F_{test}=72.96, rm^2_{test} =0.838, CCC_{test}=0.9146, CCC_{train}=0.9133

For the linear generated model, it can be seen that, the external ability of the model is better than other linear models, and therefore, it can be employed as the best linear model to predict the retention time. Variation Inflation Factor (VIF) values of each chosen descriptor with its correlation values with other selected descriptors were listed in Table 22.

Table 22: The correlation coefficient of selected descriptors and corresponding VIF values by kNN-GA-MLR

Variables	LogD (3.6)	BLTA96	B06[C-C]	BEHp2	ALOGP2	RBN	TPSA(NO)	VIF[a]
LogD (3.6)	1	0	0	0	0	0	0	4.291
BLTA96	-0.771	1	0	0	0	0	0	3.054
B06[C-C]	0.377	-0.341	1	0	0	0	0	1.344
BEHp2	0.551	-0.453	0.419	1	0	0	0	1.803
ALOGP2	0.816	-0.73	0.229	0.452	1	0	0	3.315
RBN	0.363	-0.13	0.244	0.38	0.319	1	0	1.348
TPSA(NO)	-0.165	0.222	0.061	0.138	-0.13	0.192	1	1.211

[a] Variation inflation factor

The Y-randomization test was calculated for kNN-GA-MLR model and the results indicated that the developed model is acceptable (Table 23).

Table 23: The Q^2_{LOO} and $R^2_{training}$ values after several Y-randomization tests for kNN-SW-MLR and kNN-GA-MLR

	kNN-SW-MLR		kNN-GA-MLR	
No	**Q^2**	**R^2**	**Q^2**	**R^2**
1	0.0121	0.0432	0.0008	0.0269
2	0.0026	0.0132	0.0006	0.0241
3	0.0895	0.0022	0.0034	0.0135
4	1.50E-05	0.0192	0.0005	0.0159
5	0.0003	0.0164	0.0057	0.0105
6	0.0014	0.0154	0.0013	0.0136
7	0.0051	0.0313	0.0392	0.0042
8	0.0362	0.0057	0.0038	0.0313
9	0.0056	0.0099	1.17E-06	0.0200
10	0.0002	0.0184	0.0047	0.0104

Williams plot detected non-outliers for the final kNN-GA-MLR (figure 43). The predicted retention time versus the observed retention time values based on kNN-GA-MLR were shown in figure 44.

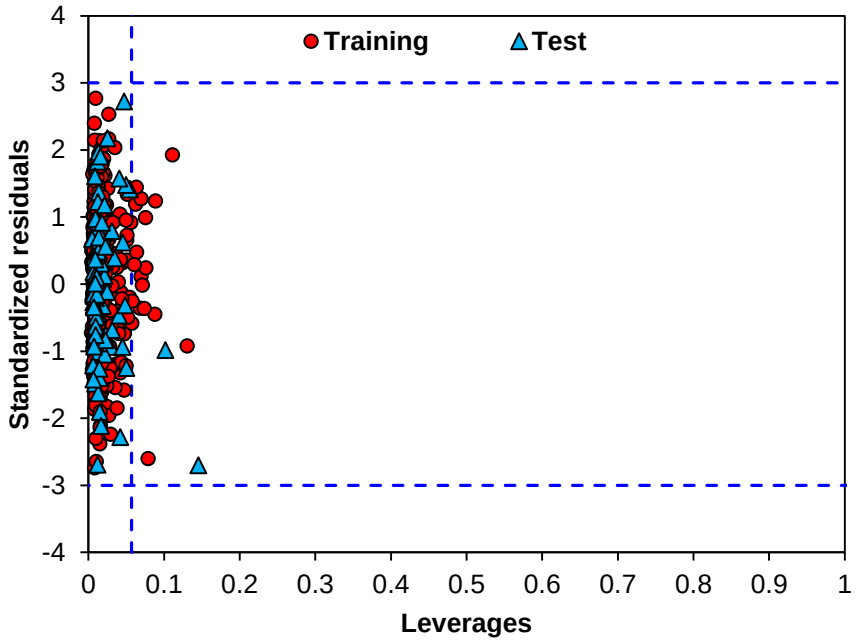


Figure 43: William plot of kNN-GA-MLR model (positive ionization): h* warning leverage value is 0.05714

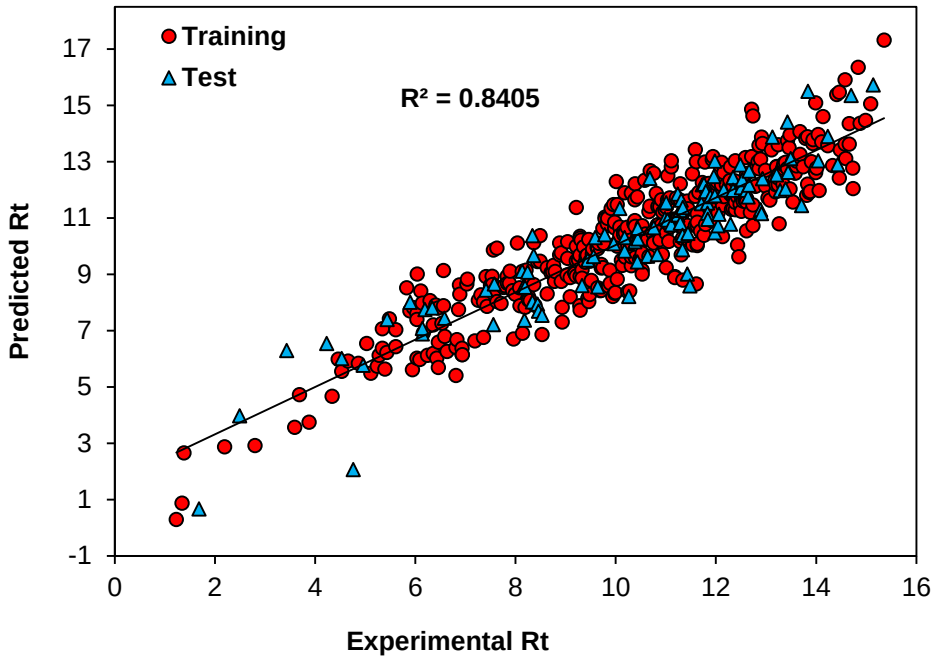


Figure 44: The plot of predicted retention time against the observed retention time values based on kNN-GA-MLR model

5.2.5 PCA-SW-SVM

The used methodology for developing support vector machine discussed in negative ionization was employed here. The results of the optimization of parameters for each SVM model based on stepwise and genetic algorithms with different splitting technique were listed in Table 24 and PCA-SW-SVM results were shown in figures 45-47.

Table 24: Optimized parameters values for SVM models

Models	Epsilon (ε)	Gamma (γ)	Capacity (C)
PCA-SW-SVM	0.1	5	22
PCA-GA-SVM	0.03	5	29
kNN-SW-SVM	0.1	5	31
kNN-GA-SVM	0.1	3.5	50

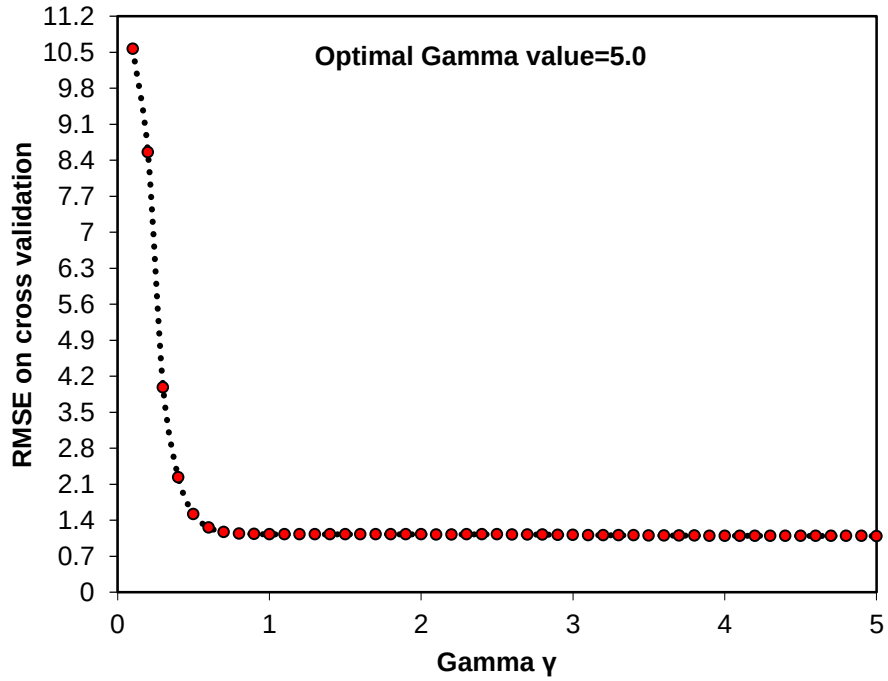


Figure 45: PCA-SW-SVM optimized parameters for the gamma (γ) vs. RMSE

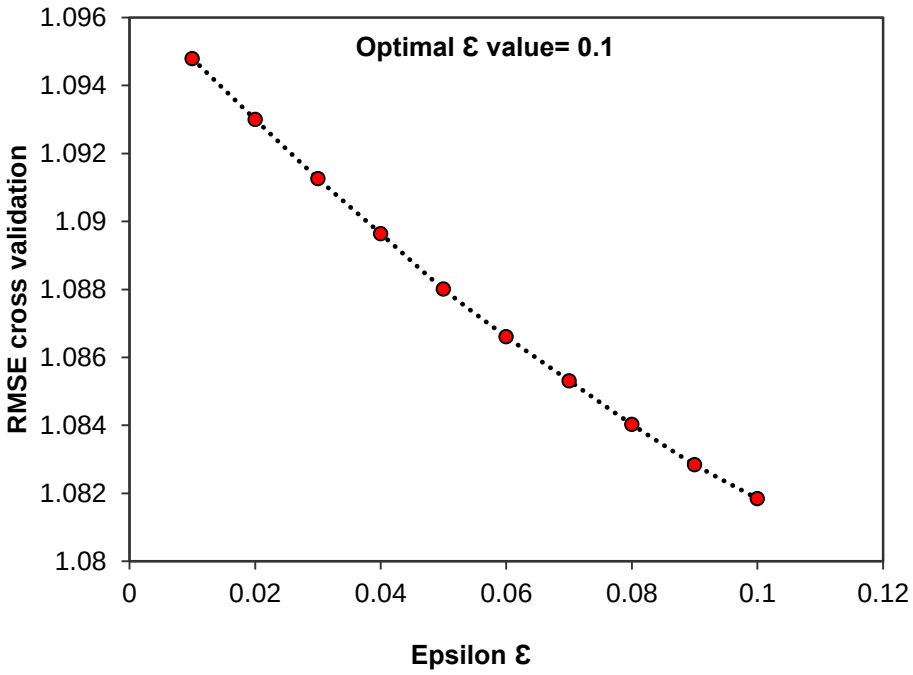


Figure 46: PCA-SW-SVM optimized parameters for the epsilon (ε) vs. RMSE

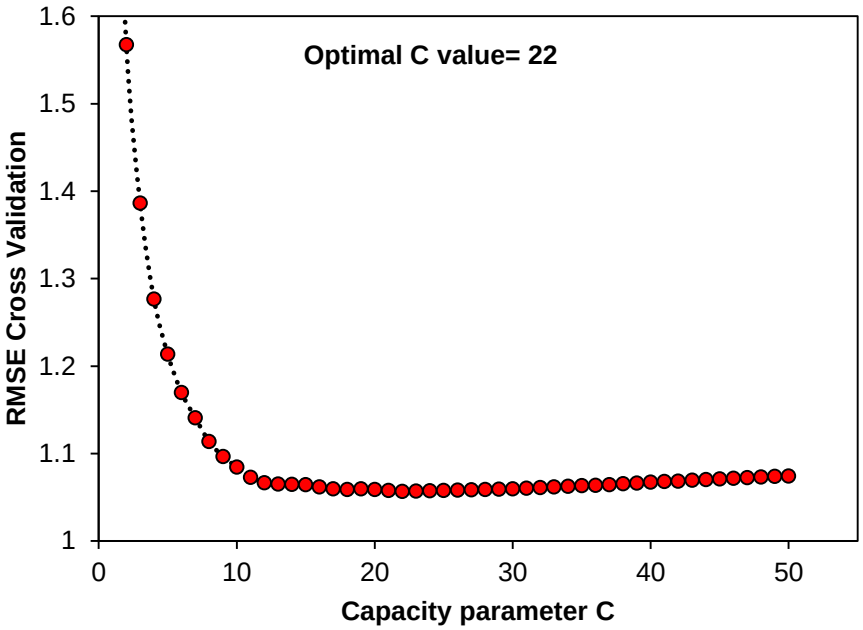


Figure 47: PCA-SW-SVM optimized parameters for the capacity (C) vs. RMSE

For each non-linear model, the data set was shown in Table S1 (positive ionization), and the results of each models were compared to the linear models and presented in Table 27. As it can be seen from Table 27, the best non-linear model was obtained based on the kNN-GA-SVM. The plot of predicted retention time against the observed retention time values based on PCA-SW-SVM method was shown in figures 48.

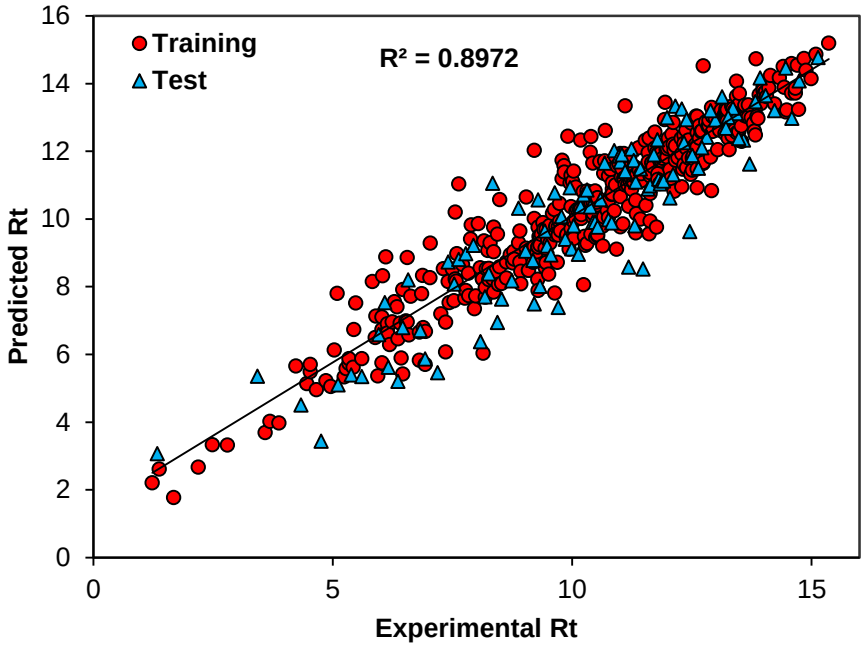


Figure 48: The plot of predicted retention time against the observed retention time values based on PCA-SW-SVM

5.2.6 PCA-GA-SVM

The optimum parameters of SVM for PCA-GA were derived as described above and shown in figures 49, 50 and 51.The plot of predicted retention time against the observed retention time values based on PCA-GA-SVM method was shown in figure52.

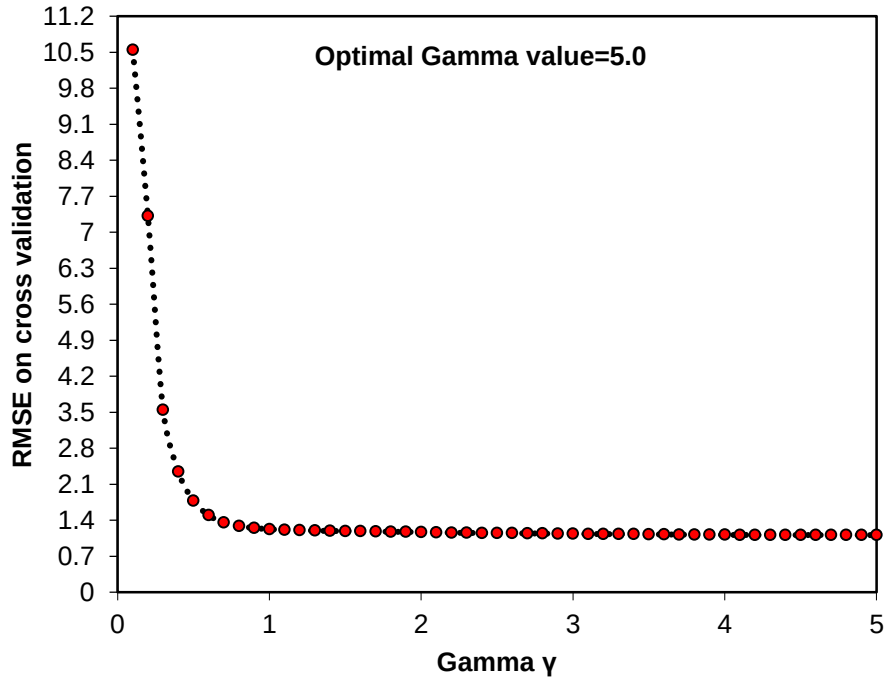


Figure 49: PCA-GA-SVM optimized parameters for the gamma (γ) vs. RMSE

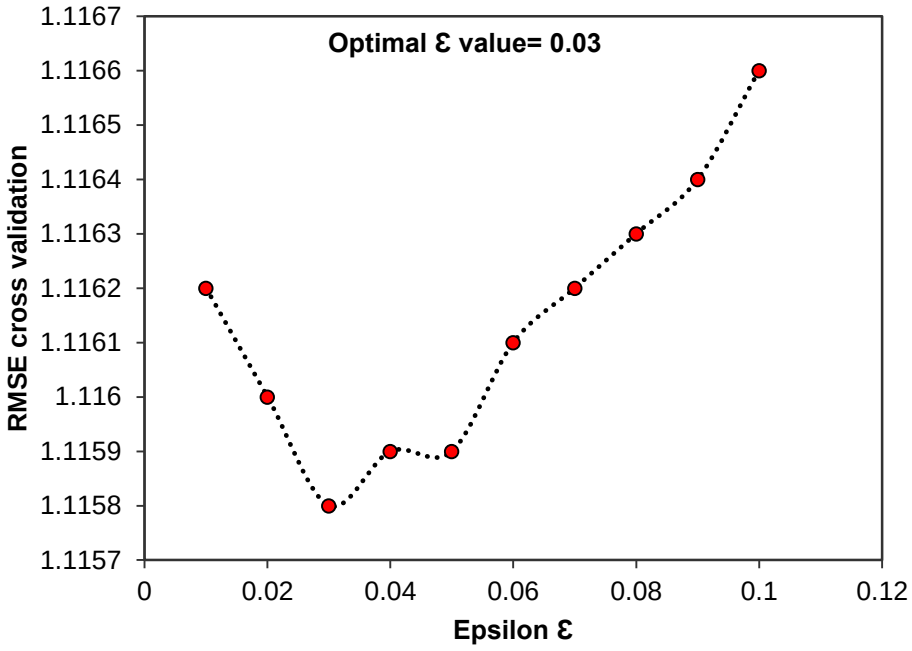


Figure 50: PCA-GA-SVM optimized parameters for the epsilon (ε) vs. RMSE

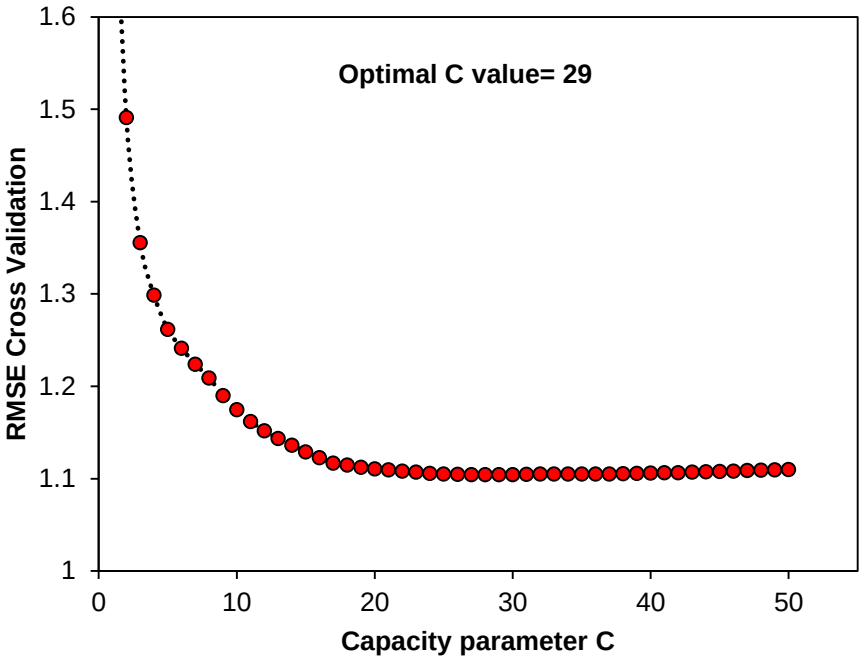


Figure 51: PCA-GA-SVM optimized parameters for the capacity (C) vs. RMSE

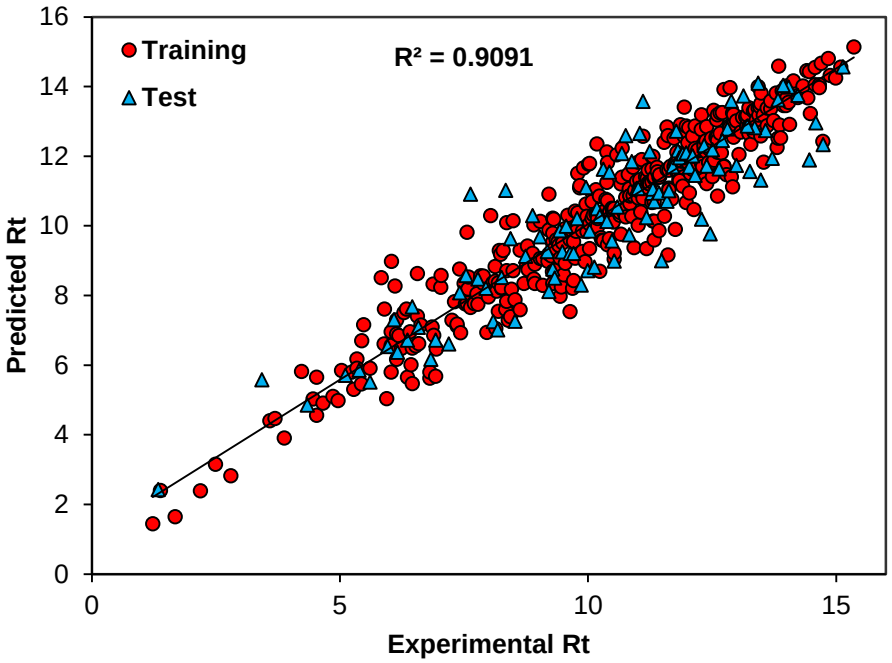


Figure 52: The plot of predicted retention time against the observed retention time values based on PCA-GA-SVM

5.2.7 kNN-SW-SVM

The optimum parameters of SVM for PCA-GA were derived as described above and shown in figures 53, 54 and 55.The plot of predicted retention time against the observed retention time values based on kNN-SW-SVM method was shown in figure56.

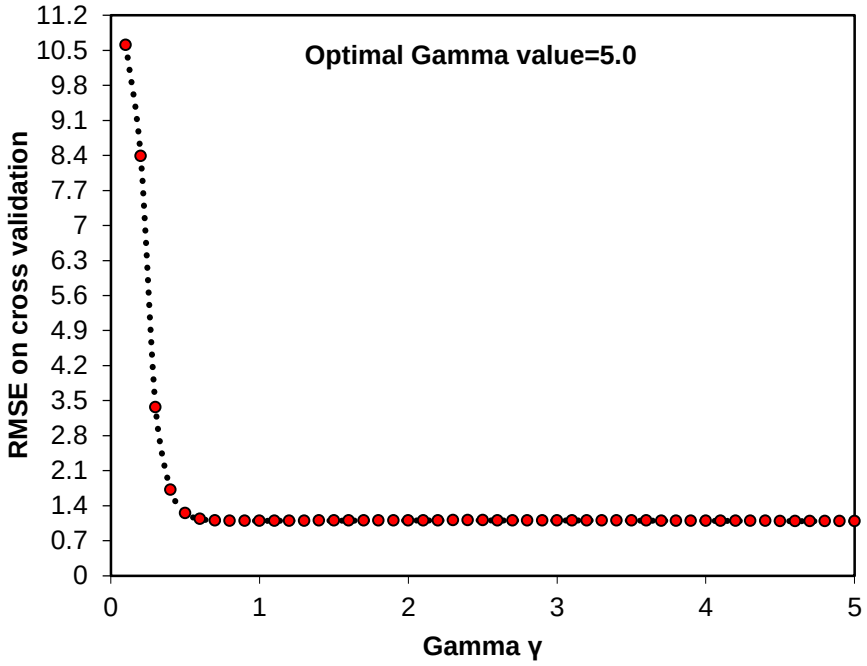


Figure 53: kNN-SW-SVM optimized parameters for the gamma (γ) vs. RMSE

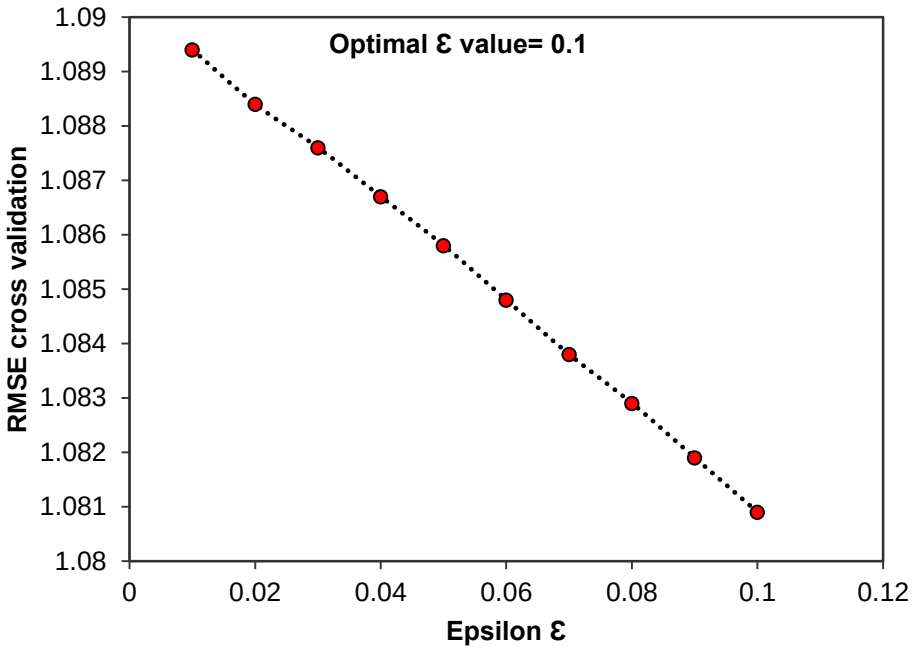


Figure 54: kNN-SW-SVM optimized parameters for the epsilon (ε) vs. RMSE

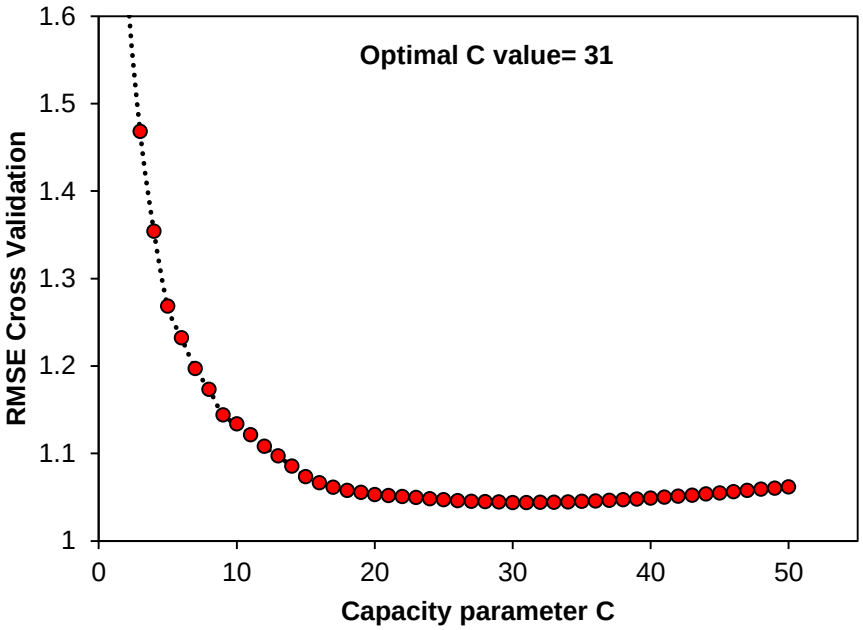


Figure 55: kNN-SW-SVM optimized parameters for the capacity (C) vs. RMSE

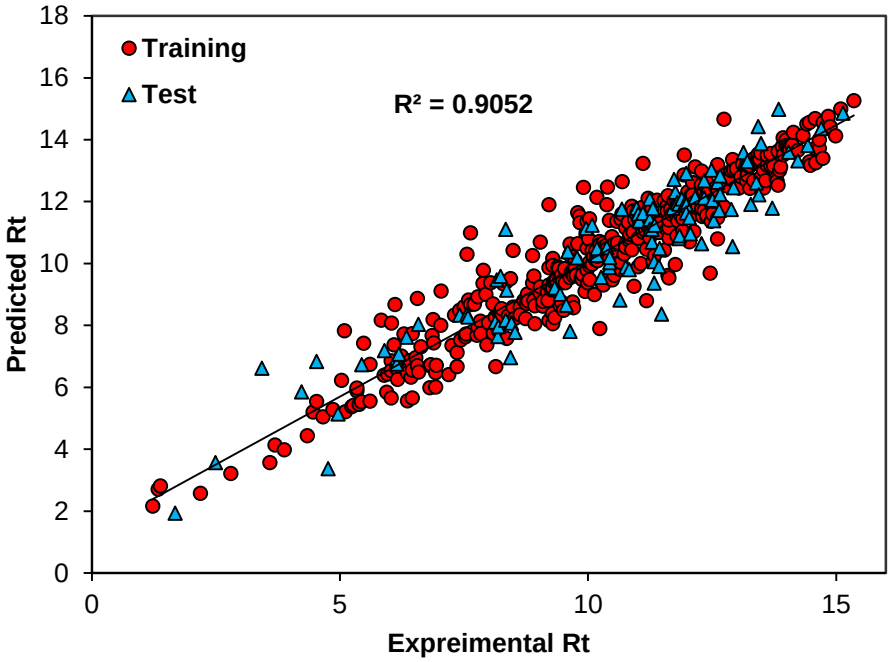


Figure 56: The plot of predicted retention time against the observed retention time values based on kNN-SW-SVM

5.2.8 kNN-GA-SVM

The selected kNN-GA-SVM non-linear model was built based on the same selected compounds as training set in kNN-GA-MLR, and the optimized parameters were calculated as C=50, ε=0.1, γ=3.5. The result of each optimization was shown in Figures 57-59. The predicted values for retention time by KNN-GA-SVM method for positive ionization compounds were given in Table S1, and plotted versus the observed retention time and shown in figure 60. The comparison of the built models (Table 27) suggests that kNN-GA-SVM is the most appropriate non-linear model for prediction purpose; however PCA-GA-SVM can also be employed. From the linear models, kNN-GA-MLR can be used due to the satisfactory external results. The final validations for these two selected models were carried out using Golbraikh and Tropsha acceptable model criteria's. The results are shown in Table 25.

Table 25: Golbraikh and Tropsha acceptable model criteria's for MLR and SVM

	kNN-GA-MLR	**kNN-GA-SVM**
Condition I	0.834	0.501
Condition II	0.846	0.887
Condition III	K=1.00824 K′= 0.98166 $R^2 - {R_0}^2/R^2$ = 0.00009 ${R_0}^2 - {R_0}'^2/R^2$ = 0.03518	K=1.00898 K′= 0.98359 $R^2 - {R_0}^2/R^2$ = 0.00092 ${R_0}^2 - {R_0}'^2/R^2$ = 0.02507
Condition IV	${R_0}^2 - {R_0}'^2$= 0.02968	${R_0}^2 - {R_0}'^2$= 0.02142
Acceptance	Passed	Passed

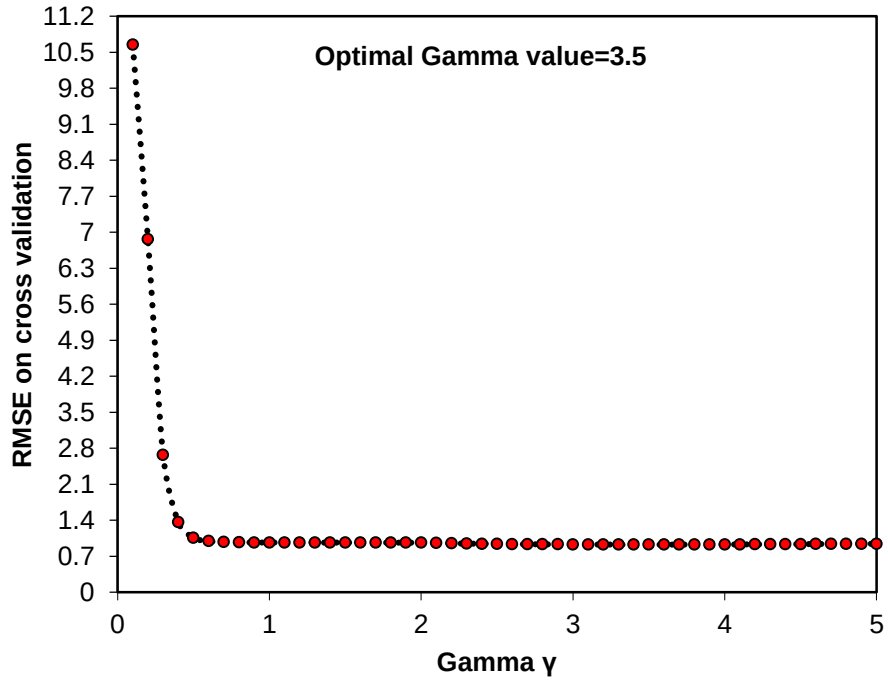


Figure 57: kNN-GA-SVM optimized parameters for the gamma (γ) vs. RMSE

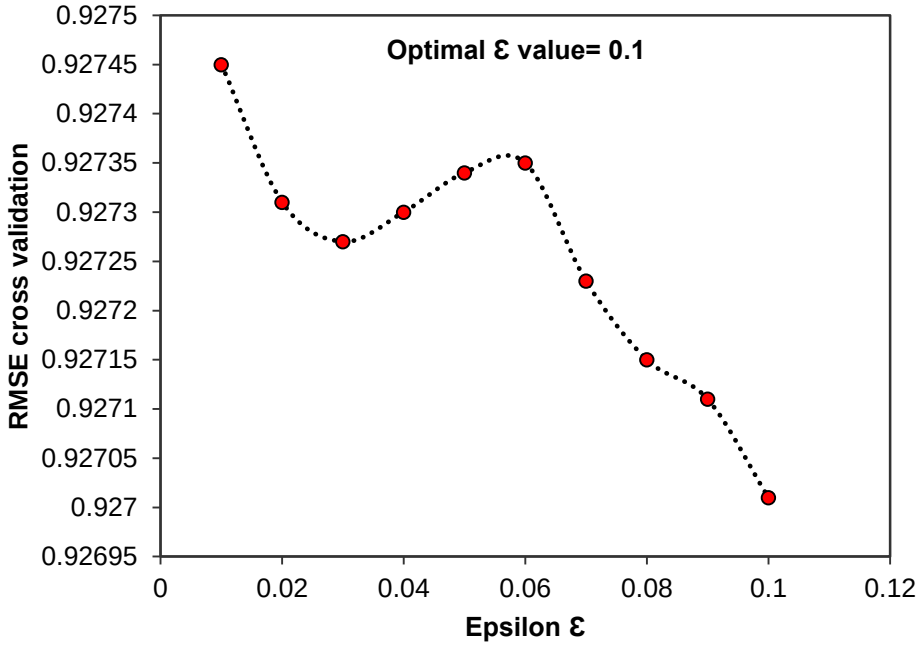


Figure 58: kNN-GA-SVM optimized parameters for the epsilon (ε) vs. RMSE

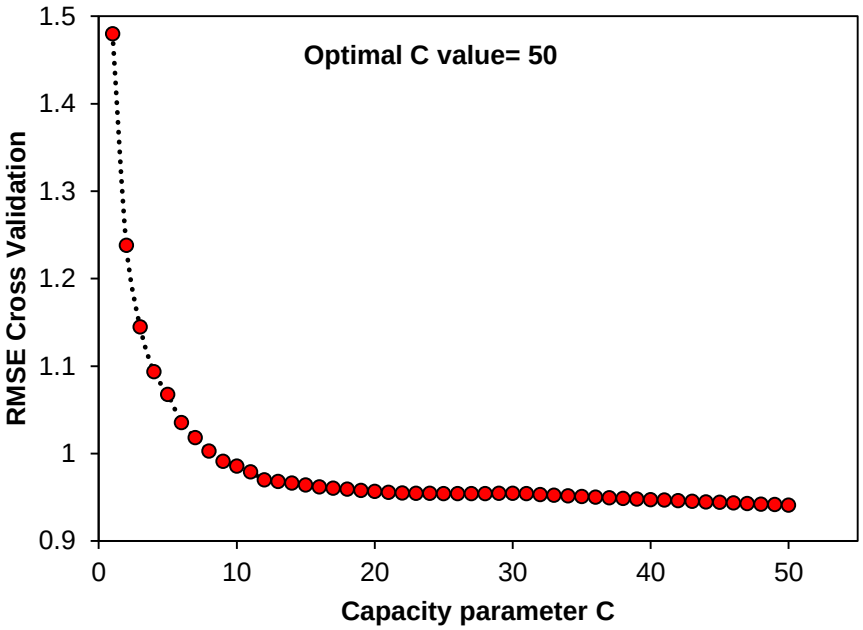


Figure 59: kNN-GA-SVM optimized parameters for the capacity (C) vs. RMSE

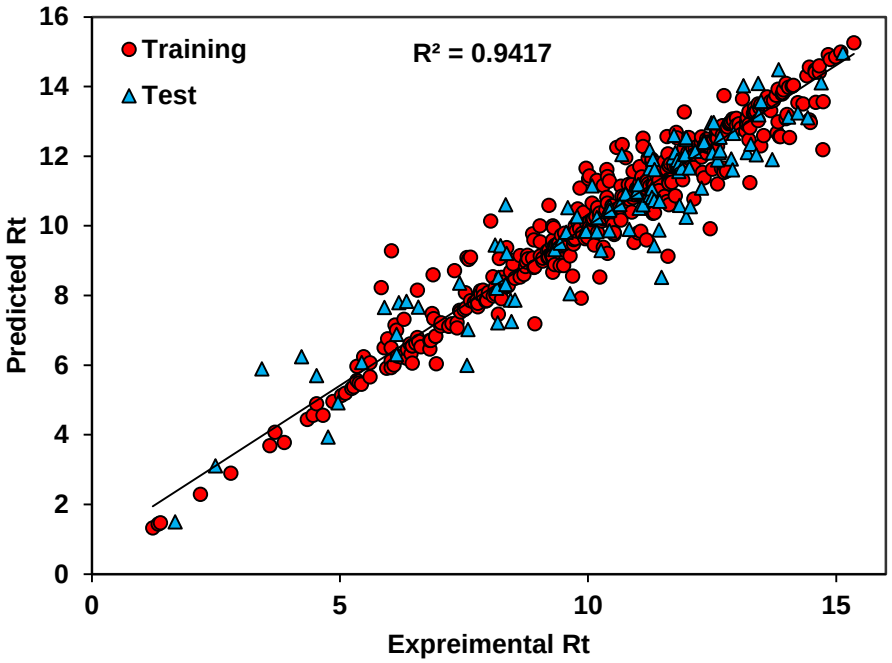


Figure 60: The plot of predicted retention time against the observed retention time values based on kNN-GA-SVM

5.2.9 kNN-GA-ANN

Since the models based on kNN and genetic algorithm showed appropriate internal and external results, for the generation of non-linear model based on ANN, kNN-GA-ANN technique was developed. The selected compounds as valid and test set were marked in Table S1 (positive ionization). The same newly introduced technique for choosing the nodes in negative ionization compounds were used here to develop accurate models without the over-fitting problem. The results of this methodology are shown in Table 26. From Table 26, it can be seen that the model built based on node=6 shows the highest CCC value for both the test and the training set. Moreover, the calculated modified r^2 value for test set is the highest one among the other nodes. Considering the RMSE value for node 3 and 6, consequently the model based on 6 nodes is being selected. The MPD values for training set with the different nodes were calculated using equation 13 and are given in Table 26. The obtained MPD value for node 3 and 6 indicates that the model based on 6 nodes represents an appropriate fitting. The predicted values based on kNN-GA-ANN are listed in Table S1 and its strength as prediction tool is compared in Table 27. The predicted values for the retention time by kNN-GA-ANN method for the positive ionization compounds versus the observed retention time are shown in figure 61.

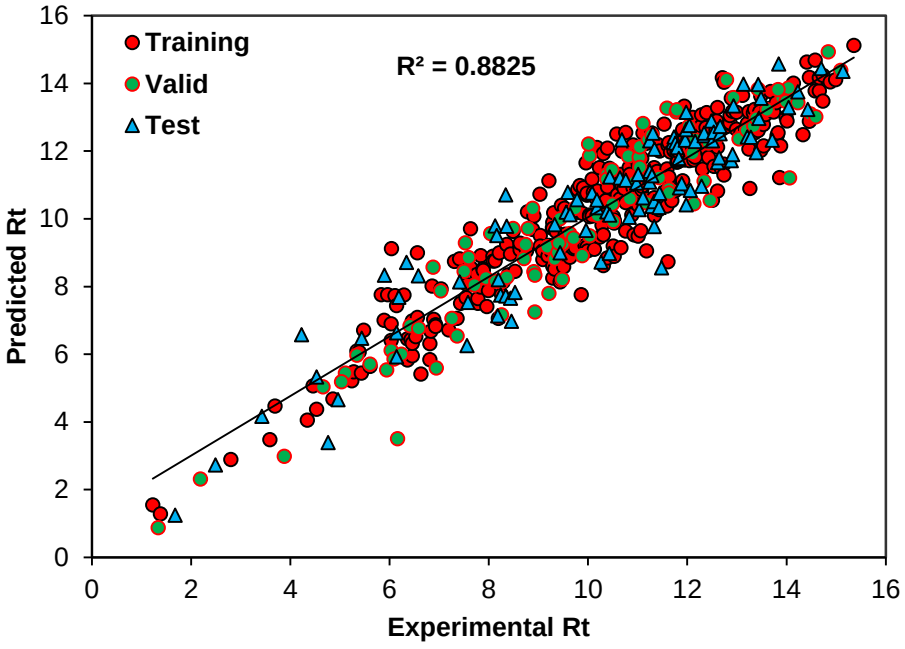


Figure 61: The plot of predicted retention time against the observed retention time values based on kNN-GA-ANN

Table 26:Comparison of the developed models for positive ionization compounds based on different nodes in ANN

	Test					Valid				Train				
	R^2m	CCC	F	RMSE	R^2	CCC	F	RMSE	R^2	MPD	CCC	F	RMSE	R^2
Node1	0.837	0.922	79.10	1.044	0.859	0.938	105.4	0.976	0.883	9.384	0.912	227.9	1.044	0.839
Node2	0.818	0.92	75.34	1.048	0.86	0.932	100.7	1.047	0.869	8.449	0.922	259.3	0.992	0.855
Node3	0.833	0.929	85.25	0.99	0.875	0.92	78.2	1.092	0.857	8.998	0.92	251.3	0.999	0.853
Node4	0.804	0.914	76.73	1.12	0.839	0.939	115.0	1.006	0.882	7.877	0.929	287.8	0.946	0.868
Node5	0.784	0.913	77.65	1.14	0.835	0.925	93.9	1.121	0.855	7.532	0.935	314.0	0.912	0.877
Node6	0.864	0.933	96.76	0.99	0.873	0.945	128.0	0.948	0.894	7.184	0.937	327.7	0.893	0.883
Node7	0.776	0.898	64.14	1.214	0.81	0.923	85.5	1.099	0.854	6.651	0.943	361.3	0.856	0.892
Node8	0.784	0.898	58.10	1.172	0.824	0.886	61.2	1.373	0.785	6.662	0.945	373.2	0.843	0.895
Node9	0.797	0.909	71.72	1.152	0.829	0.938	110.4	0.995	0.881	7.026	0.945	373.3	0.844	0.895
Node10	0.839	0.929	92.37	1.023	0.867	0.94	120.9	1.012	0.887	6.792	0.942	357.2	0.862	0.89
Node11	0.773	0.905	69.71	1.186	0.821	0.923	88.5	1.118	0.852	6.372	0.95	414.8	0.805	0.905
Node12	0.792	0.907	70.08	1.163	0.827	0.925	94.7	1.125	0.856	6.54	0.946	385.3	0.831	0.898
Node13	0.827	0.919	80.39	1.082	0.851	0.918	85.5	1.164	0.843	6.435	0.949	407.9	0.811	0.903
Node14	0.752	0.899	66.20	1.226	0.81	0.902	69.7	1.258	0.815	5.671	0.954	456.6	0.77	0.913
Node15	0.824	0.928	92.74	1.037	0.863	0.912	87.0	1.263	0.843	5.898	0.956	471.3	0.759	0.915
Node16	0.805	0.907	69.35	1.156	0.827	0.924	92.1	1.121	0.854	6.019	0.954	448.7	0.774	0.912
Node17	0.819	0.914	75.32	1.119	0.838	0.915	83.0	1.182	0.84	5.636	0.954	485.4	0.748	0.918
Node18	0.776	0.928	98.19	1.064	0.864	0.929	102.6	1.097	0.866	5.757	0.957	505.4	0.734	0.921
Node19	0.807	0.925	89.78	1.065	0.856	0.934	113.5	1.081	0.878	5.361	0.962	551.6	0.704	0.927
Node20	0.789	0.903	67.70	1.193	0.818	0.93	104.0	1.096	0.867	5.08	0.963	566.1	0.697	0.929

Table 27: Comparison of the developed models for positive ionization compounds

	Test					Train				
models	**R^2m**	**CCC[a]**	**RMSE**	**F**	**R^2**	**CCC[a]**	**Q^2_{LOO}**	**RMSE**	**F**	**R^2**
PCA-SW-MLR	0.765	0.913	1.127	78.487	0.843	0.916	0.84	1.061	324.492	0.846
PCA-GA-MLR	0.814	0.897	1.154	59.002	0.816	0.918	0.843	1.048	331.19	0.849
PCA-SW-SVM	0.806	0.922	1.057	85.618	0.854	0.943	0.497	0.873	473.063	0.897
PCA-GA- SVM	0.826	0.906	1.104	65.07	0.83	0.951	0.583	0.814	567.91	0.909
kNN-SW-MLR	0.809	0.905	1.163	67.029	0.826	0.917	0.841	1.051	327.203	0.847
kNN-GA-MLR[b]	0.838	0.915	1.093	72.961	0.846	0.913	0.834	1.069	310.173	0.841
kNN-SW-SVM	0.833	0.921	1.044	77.905	0.86	0.948	0.622	0.833	523.31	0.905
kNN-GA- SVM[c]	0.862	0.937	0.941	98.851	0.887	0.969	0.501	0.649	898.019	0.942
	R^2m	CCC[a]	RMSE	F	R^2	CCC[a]	MDP	RMSE	F	R^2
kNN-GA-ANN	0.784	0.913	1.14	77.645	0.835	0.935	7.532	0.912	313.988	0.877

[a] Concordance correlation coefficient

[b] The best linear model

[c] The best non-linear model

5.2.10 Interpretation of Molecular descriptors

The descriptors which were selected based on the genetic algorithms showed to have striking effects and appropriate correlations with observed retention time. In the derived prediction analyses for positive ionization, some descriptors were presented as they were chosen in negative ionization, and therefore, it reflects that these three descriptors (LogD(at certain pH), ALOGP, and BLTA96 are more responsible for the chemical behavior in regards of the retention time. For the compounds in positive ionization, these three descriptors have the same impact on the retention time (equation 32), as in negative ionization compounds, since the sign of the correlation coefficients for LogD and ALOGP2 is positive and the sign for BLTA96 is negative (equation 23). The relative importance of the selected descriptors is shown in figure 62.

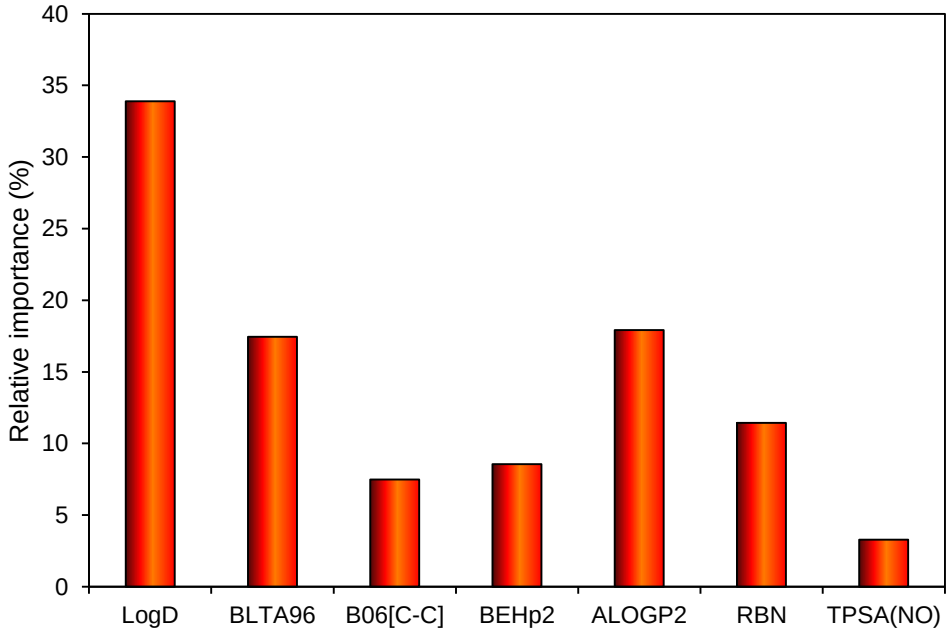


Figure 62: The relative importance of selected descriptors in positive ESI.

ALOGP2 is squared Ghose-Crippen octanol-water partition coefficient which is belonged to molecular properties descriptors and is a measure of the lipophilicity of the molecule, the same as ALOGP. As it was discussed previously, the Ghose-Crippen contribution method [20, 57, 58] is based on the hydrophobic atomic constants a_k that is measuring the lipophilic contribution of atoms in the molecule as follows:

$$\mathrm{Log}P = \sum_{\mathrm{k}} \mathrm{a_k}.\mathrm{N_k} \qquad (Eq.33)$$

Where N_k is the occurrence of the k th atom type, and the hydrophobic constant have been evaluated for hydrogen atoms, carbon atoms and heteroatoms.

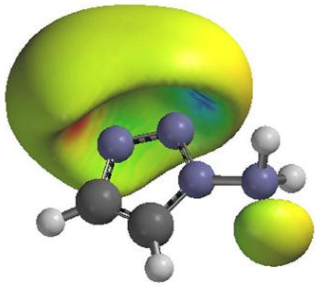
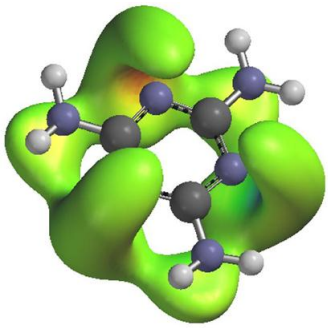
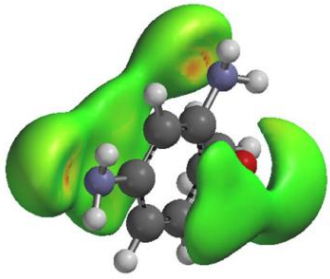
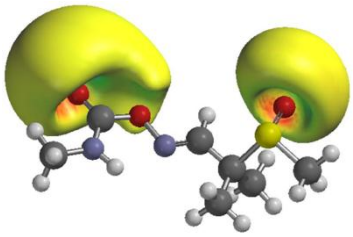
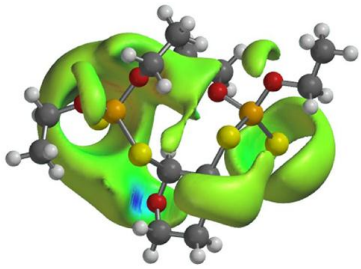
In addition to the above descriptors, the linear model based on kNN-GA-MLR showed another descriptor, B06[C-C]. This descriptor is a type of 2D binary atom pairs of order 6 descriptors and defines the presence/absence of C - C at topological distance 6. This kind of descriptor describes the pairs of atoms and bond types connecting them based on the topological representation of molecules. Two carbon atoms and inter atomic separation is defined as:

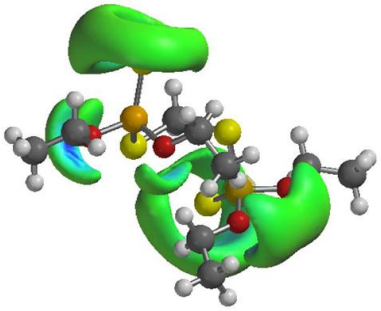
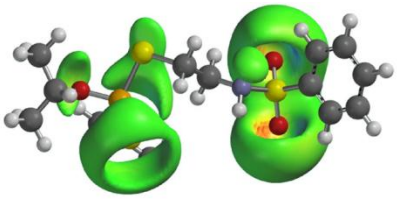
$$\mathrm{AP} = \{[i\text{th atom description}][\text{separation}][j\text{th atom description}]\} \qquad (Eq.34)$$

Therefore, the separation is the topological distance between these two carbon atoms. As it can be seen, this descriptor has positive sign in the linear equation, which encodes that the availability of such binary atom pairs in molecular structure would cause an increase to the retention time.

The next selected descriptor is BEHp2 (highest eigenvalue n.2 of Burden matrix / weighted by atomic polarizability). As it was discussed in negative ionization, the Burden eigenvalue descriptors [63] represent the chemical structural diversity or similarity of a molecule based on the Burden approach [62]. Another useful benefit is associated with the eigenvectors, which can be used to determine the attribution of each atom to substructures upon disconnection of the main structure into distinct fragments. In this context, atomic polarizibility that is relevant to intermolecular interactions are supported and since its correlation coefficient has positive sign in equation 32, increasing the atomic polarizability relevant to intermolecular interactions would increase the retention time. To understand BEHp2 effects and also its relationship with the polar atoms (O,N S and P), LogD values as well as charges potential and also BEHp2 values for some compounds were studied and listed in Table 28. Polar surface area (PSA) and charges potential were calculated based on DFT study on the basis of B3LYP/6-31*G method.

Table 28: Relationship between BEHp2, Retention time, LogD and charges potential

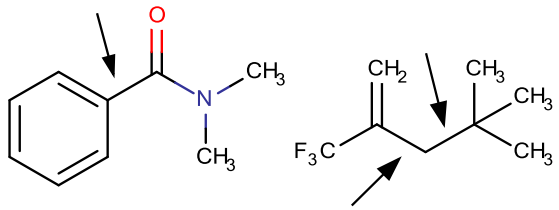
Name	LogD	Charges potential	BEHp2	Rt	PSA (Å^2)
m12	-1.07		2.319	1.38	56.095
m489	-2.56		2.633	1.34	97.207
m219	-1.81		2.911	2.19	55.698
m7	-0.59		3.134	4.66	56.169
m125	4.38		4.026	13.21	46.676

m145	3.93		3.994	13.53	31.169
m31	3.45		3.975	11.66	61.812

BEHp2 can be assumed to be measure of polarizibilities in substracture or fragments of a compound and thus lower value can represent the less number of fragments in molecular graph as well. For example, from Table 26, m12 (Amitrole) showed the lowest BEHp2 in contrast to whole data set suggesting that atomic prolazibility in substracture is so low. It is also a small substance which limits the fragmentation. Reported Rt as well as LogD are low. It seems that presence of Nitrogen in molecular graph decreases the BEHp2 values in contrast to presence of Oxygen, Sulfur and Phosphore. Comparing compound m12 with m219, it can be seen that addition of oxygen group inceased BEHp2 values but decreased LogD values suggesting that molecule is more polar, however as BEHp2 increasedin compound m125, the LogD is increased. Therefore, BEHp2 is not representing the atomic polarizibility of molecule but its fragments. There is also a masking effect of LogD which its effect is more dominant than BEHp2. Therefore, it can be concluded that if a molecule represents more fragmentations with less number of Nitrogen, the BEHp2 will be increased directly leading to increase of Rt while it is not representing the true effect of polarizibility of compound since its LogD might affect Rt inversely. It is also a good agreement between retention time and polar surface area calcualted by DFT study (B3LYP/6-31G*).

The next descriptor is RBN (number of rotatable bonds) in which belongs to constitutional descriptors and encodes the number of bonds that have free rotation around themselves. It is defined in a single bond type which is not in a ring. In addition, bonds between amide (C-

N) are excluded from this calculation due to the high rotational energy barrier. To understand this effect more deeply, some compounds were presented below:



RBN=1 RBN=2

Apparently, the RBN value is more in alkyl chains than in its ring form. It seems that compounds with more RBN and in other words alkyl chain have good interaction with stationary phase alkyl groups and thus postponing the elution. Since this descriptor indicated positive sign in equation, increasing number of rotatable bond would results in an increase of the retention time.

The last selected descriptor based on genetic algorithms is the topological polar surface area using N, O polar contributions (TPSA(NO)). This descriptor represents the influence of a particular functional group (especially based on the compounds with higher electronegativity atoms) [60]. Since this descriptor has a negative sign in the equation, the presence of N, O polar contributions in molecular structure would cause a decrease to the retention time. Therefore, compounds with higher TPSA(NO) value would show lower retention time, however in comparison to effects by other selected descriptors, the mean effect of this descriptor is lower.

5.2.11 Applicability domain study of kNN-GA-MLR model for suspects

Some compounds (as suspect compounds) were used to predict their retention time based on the developed models so as to figure it out whether the suggested compounds can be the correct candidates or not. The results of prediction for these compounds along with their experimental retention time were listed in Table 29.

Table 29: Retention time prediced values of suspect compounds in positive ionization by KNN-GA-SVM

	Suspectlist	Exp. Rt	KNN-GA-SVM Predicted Rt	
			The suggested structure can be accepted	The suggested structure is rejected
P1	(1-Hydroxy-iso-propyl)acetophenone	6.34		10.10
P2	1,3,3-Trimethyl-2-methyleneindoline	13.49		8.19
P3	1.2.3.6-Tetrahydrophthalimide (cis-)	5.71	4.87	
P4	17-alpha-Estradiol	12.98	11.22	
P5	17-beta-Estradiol	13.16	11.22	
P6	2-[2-[4-(1-1-3-3-tetramethylbutyl)phenoxy]ethoxy]ethanol / 4-Octylphenol di-ethoxylate	12.44	12.98	
P7	2-2-4-trimethylpentane-1-3-diol diisobutyrate	13.41	13.21	
P8	2-3-Dihydro-1-methyl-1H-indol (1-Methyl-2-oxindole)	6.53	6.44	
P9	2-6-Di-tert-butyl-4-hydroxy-4-methyl-2-5-cyclohexadien-1-one	11.71	10.28	
P10	2-6-Di-tert-butylphenol	12.26	11.97	
P11	2-Methyl-1-phenylpropan-2-ol	12.44	9.28	
P12	3-5-Di-tert-butyl-4-hydroxyacetophenone	10.16	11.52	
P13	4-acetyl-amino-antipyrine (AAA)	5.29	6.87	
P14	4-formyl-amino-antipyrine (FAA)	5.24	6.65	
P15	4-iso-Propenylacetophenone	7.34	9.67	
P16	4-tert-Butylphenol	12.44		8.36
P17	5 6-di-Me-Benzotriazole	5.69	7.10	
P18	Amitraz	8.79		11.28
P19	Ancymidol	7.78	8.26	
P20	Atrazine-desethyl	6.86	6.83	
P21	Benefin	8.88		12.69
P22	Benzylbutylphthalate	12.78	13.31	
P23	Bis-(2-ethylhexyl) phthalate	14.33	15.33	
P24	Bisoprolol	11.81	8.94	
P25	Butylbenzoate	7.34		11.43
P26	Butylmethoxydibenzoylmethane	14.86	13.04	
P27	Camphor	11.88		6.89
P28	Cyclohexylisocyanate	6.13	6.57	
P29	Desethylterbuthylazine	8.54	7.07	
P30	Diisononylphthalate	12.76	15.18	
P31	Di-iso-propylphenol	12.44	8.89	
P32	Di-n-butylphthalate	12.84	13.42	
P33	Dinex (2-Cyclohexyl-4.6-dinitrophenol)	8.63	9.69	

Suspectlist		Exp. Rt	KNN-GA-SVM Predicted Rt	
			The suggested structure can be accepted	The suggested structure is rejected
P34	Dinocap	10.09		14.17
P35	Di-n-octylphthalate	14.91	15.09	
P36	d-Limonene	12.44	10.63	
P37	Estriol	9.41	9.96	
P38	Estrone	11.23	11.39	
P39	Ethylenebrassylate	10.24	10.44	
P40	Ethylhexylmethoxycinnamate	14.63	14.02	
P41	g-Methylionone	12.26	11.11	
P42	Hexa(methoxymethyl)melamine	8.71	10.24	
P43	HydroxyIbuprofen	8.28	9.10	
P44	Icaridin	9.46	8.09	
P45	JWH-210	8.01		14.75
P46	Melamine	1.34	1.440	
P47	Merphos	6.61		14.38
P48	Methylneodecanamide	12.19	11.91	
P49	Methyl-iso-propylcyclohexenone-Carvone	12.44	9.74	
P50	Methylphenobarbital	7.81	7.29	
P51	Mutagen X	1.23		4.9
P52	N-Acetylmorpholine	6.21		3.57
P53	N-Methyl-2-pyrrolidone	3.71	3.51	
P54	N-Methylphenacetine	3.84		8.48
P55	N-N'-Diethyl-N-N'-diphenylurea	11.88	11.33	
P56	N-nitrosodiethylamine	1.23		5.82
P57	N-Nitrosopyrrolidine	1.23	3.47	
P58	N-phenyl-naphthylamine	12.71	12.00	
P59	Octocrylene	14.26	15.06	
P60	Oxadine A / 4-4-dimethyloxazolidine	5.18		2.15
P61	Phenytoine	7.26	9.00	
P62	Prometon	6.76	9.70	
P63	Pyrimidifen	8.39		13.48
P64	Sebuthylazine	10.43	10.39	
P65	Secbumeton	6.76		9.79
P66	Spectinomycin	8.19		3.94
P67	Styrene	12.44		7.83
P68	tert-Butylhydroquinone	8.56	8.17	
P69	Tributylphosphate (TBP)	12.53	13.13	
P70	Tributylacetylcitrate	13.53	12.95	
P71	Triethylcitrate	6.59	8.50	
P72	Trifluralin	8.88		12.81
P73	Triphenylphosphineoxide	9.96	12.04	
P74	Tris(1-chloro-2-propanyl) phosphate	10.53	11.45	

Suspectlist		Exp. Rt	KNN-GA-SVM Predicted Rt	
			The suggested structure can be accepted	The suggested structure is rejected
P75	Tris(2-methylpropyl) phosphate	12.66	12.47	
P76	Tris(methylphenyl) phosphate	13.34	14.10	
P77	Viridine	8.56	9.25	

The workflow introduced in section 4.6 was employed to visualize the results and detect the possible outliers. Based on the training and test sets (figure 63), boxes are presented. For the compounds in suspect list, the same analysis was carried out, and results illustrated that out of 77 compounds as suspect compound, 47 compounds are predicted very well and 30 compounds are belonged to box3 and box4. The results are listed in Table 30. According to Table 30and visualization plot (figure 64), it can be concluded that out of 21 compounds in box4, twelve compounds ((1-Hydroxy-iso-propyl) acetophenone, 1,3,3-Trimethyl-2-methyleneindoline, 4-tert-Butylphenol, Benefin, Bis-(2-ethylhexyl) phthalate, Butyl benzoate, Camphor, N-Methylphenacetine, N-nitrosodiethylamine, N-Nitrosopyrrolidine, Pyrimidifen, Trifluralin) are within the applicability domain of models, but the suggested retention times are not matched with the structure and therefore, we can be confident that the suggested compounds cannot be correct. The rest of the compounds showed high residuals due to the chemical structural diversity which is beyond the applicability domain of the generated models. Molecules belonged to box 4 (outliers) were shown in red color in figure 64.

Table 30: The analysis of visualization of outliers for linear model (kNN-GA-MLR)

Boxes	Origin of outliers	compounds
Box 3	The origin of residuals is mostly due to structural diversity. The model cannot predict their Rt.	17-alpha-Estradiol
	The origin of residuals is mostly due to Response. The suspect compounds are rejected.	2-Methyl-1-phenylpropan-2-ol, 4-iso-Propenylacetophenone, Di-iso-propylphenol, Methyl-iso-propylcyclohexenone- Carvone, N-Acetylmorpholine, Oxadine A / 4-4-dimethyloxazolidine, Triphenylphosphine oxide

	The origin of residuals is mostly due to structural diversity. The model cannot predict their Rt.	Bisoprolol, Diisononyl phthalate, Dinocap, Di-n-octyl phthalate, JWH-210, Merphos, Mutagen X, Spectinomycin, Styrene
Box 4	The origin of residuals is mostly due to Response. The suspect compounds are rejected.	(1-Hydroxy-iso-propyl)acetophenone, 1,3,3-Trimethyl-2-methyleneindoline, 4-tert-Butylphenol, Benefin, Bis-(2-ethylhexyl) phthalate, Butyl benzoate, Camphor, N-Methylphenacetine, N-nitrosodiethylamine, N-Nitrosopyrrolidine, Pyrimidifen, Trifluralin

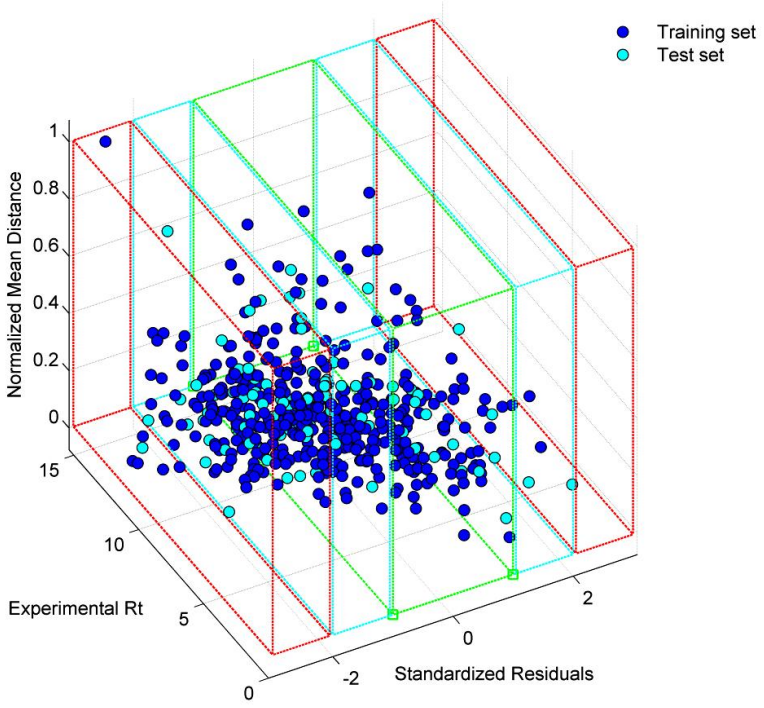


Figure 63: Visualization of the data distribution for (+) ESI compounds

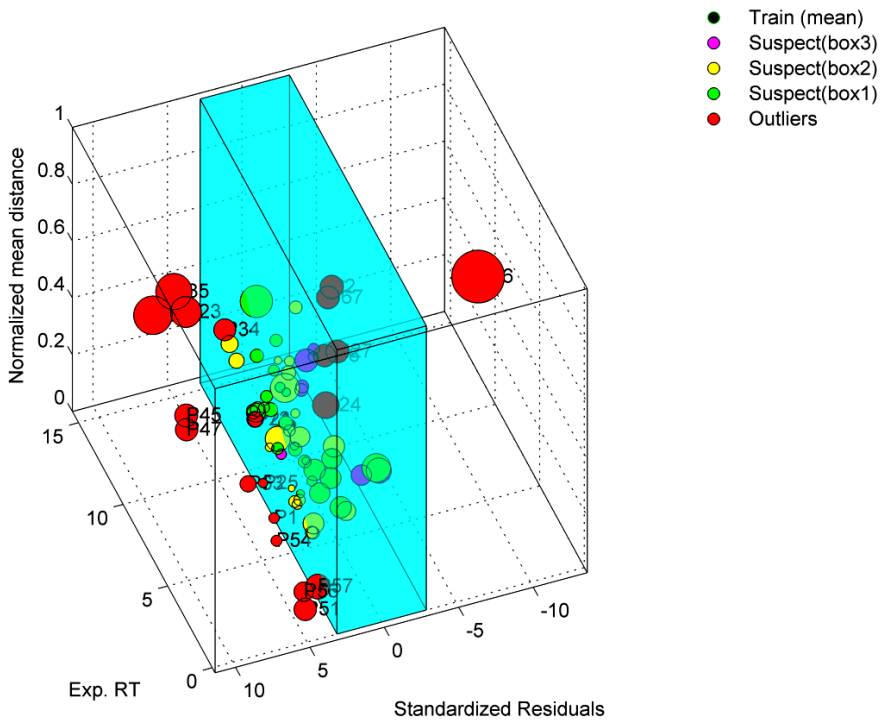


Figure 64: Origin of the outliers for the suspect compounds in positive ionization

CHAPTER 6

CONCLUSIONS

The obtained QSRR models for a large data set measured in two ionization modes by liquid chromatography–quadrupole – time of flight mass spectroscopy (LC-QTOF MS) supported the safe identification of suspect compounds. In this work, some novel methodologies were presented to understand the origin of miscalculation for suspect list compounds where it enabled the researchers for better understanding of rejection of a compound as suspect compound. Different models for both datasets in two ionization modes were provided and compared. The results indicated that the generated models using kNN for the appropriate division of data set, and employing of genetic algorithms as variable selection technique, would lead to strong models for the prediction purposes. MLR based on kNN-GA as a simple linear model for both datasets was validated by employing different validation techniques, and it proved to be applicable for prediction purposes, however among the non-linear models, SVM with same selection technique showed the highest statistical confidence and accuracy for the prediction of the retention time. The newly suggested remarks for ANN models generation reported in supplementary material file could also lead to better prediction in comparison to the MLR models, so that it could present better statistical results with acceptable criteria. However, its results were weak in comparison to SVM model. The provided large dataset could also make the application of models possible in different disciplines (environmental chemistry, pharmaceutical analysis, metabolomics, forensics and doping analysis). Consequently, the derived workflows for model generations and validations besides the visualization of outliers technique, showed great potential for the identification of suspect compounds.

APPENDIX A

```
OTrAMS(ESI,save,PD)
%%%%%%%%%%%About%%%%%%%%%%%%%%%%%
%This code is written by Reza Aalizadeh ©
%Laboratory of Analytical Chemistry,
%Department of Chemistry,
%National and Kapodistrian University of Athens,
%Panepistimiopolis Zographou, 15771 Athens, Greece
%Usage:
%ESI is electrospray ionization mode: 'Negative' or 'Positive'
%save is to export the outlier analysis to excel format: save=1, not save=0
%PD is plotting options: for separated figures and 2D dimension >PD=1, for 3D
%dimension and it's criteria >PD=2
% For Negative ESI, type the following command in MATLAB: OTrAMS('Negative',1,2)
% For Positive ESI, type the following command in MATLAB: OTrAMS('Positive',1,2)
```

REFERENCES

1. M. Krauss, H. Singer, J. Hollender, "*LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns*", Analytical and Bioanalytical Chemistry, 397 (2010) 943-951.
2. R.-J. Hu, H.-X. Liu, R.-S. Zhang, C.-X. Xue, X.-J. Yao, M.-C. Liu, Z.-D. Hu, B.-T. Fan, "*QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors*", Talanta, 68 (2005) 31-39.
3. Z. Dashtbozorgi, H. Golmohammadi, E. Konoz, "*Support vector regression based QSPR for the prediction of retention time of pesticide residues in gas chromatography–mass spectroscopy*", Microchemical Journal, 106 (2013) 51-60.
4. Y. Du, Y. Liang, "*Data mining for seeking accurate quantitative relationship between molecular structure and GC retention indices of alkanes by projection pursuit*", Computational Biology and Chemistry, 27 (2003) 339-353.
5. M. Turowski, T. Morimoto, K. Kimata, H. Monde, T. Ikegami, K. Hosoya, N. Tanaka, "*Selectivity of stationary phases in reversed-phase liquid chromatography based on the dispersion interactions*", Journal of Chromatography A, 911 (2001) 177-190.
6. R. Kaliszan, "*QSRR: Quantitative Structure-(Chromatographic) Retention Relationships*", Chemical Reviews, 107 (2007) 3212-3246.
7. L. Wu, Y. Wu, H. Shen, P. Gong, L. Cao, G. Wang, H. Hao, "*Quantitative structure-ion intensity relationship strategy to the prediction of absolute levels without authentic standards*", Analytica Chimica Acta, 794 (2013) 67-75.
8. B. Zonja, A. Delgado, S. Pérez, D. Barceló, "*LC-HRMS Suspect Screening for Detection-Based Prioritization of Iodinated Contrast Media Photodegradates in Surface Waters*", Environmental Science & Technology, 49 (2015) 3464-3472.
9. M. Molíková, M.J. Markuszewski, R. Kaliszan, P. Jandera, "*Chromatographic behaviour of ionic liquid cations in view of quantitative structure-retention relationship*", Journal of Chromatography A, 1217 (2010) 1305-1312.
10. H.P. Varbanov, M.A. Jakupec, A. Roller, F. Jensen, M. Galanski, B.K. Keppler, "*Theoretical Investigations and Density Functional Theory Based Quantitative Structure–Activity Relationships Model for Novel Cytotoxic Platinum(IV) Complexes*", Journal of Medicinal Chemistry, 56 (2013) 330-344.
11. R. Leardi, A. Lupiáñez González, "*Genetic algorithms applied to feature selection in PLS regression: how and when to use them*", Chemometrics and Intelligent Laboratory Systems, 41 (1998) 195-207.
12. R. Leardi, "*Genetic algorithms in chemistry*", Journal of Chromatography A, 1158 (2007) 226-233.
13. K. Roy, S. Kar, P. Ambure, "*On a simple approach for determining applicability domain of QSAR models*", Chemometrics and Intelligent Laboratory Systems, 145 (2015) 22-29.
14. N.L. Allinger, "*Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms*", Journal of the American Chemical Society, 99 (1977) 8127-8134.
15. J.-H. Lii, S. Gallion, C. Bender, H. Wikström, N.L. Allinger, K.M. Flurchick, M.M. Teeter, "*Molecular mechanics (MM2) calculations on peptides and on the protein*

Crambin using the CYBER 205", Journal of Computational Chemistry, 10 (1989) 503-513.
16. R. Todeschini, M. Lasagni, E. Marengo, "*New molecular descriptors for 2D and 3D structures. Theory*", Journal of Chemometrics, 8 (1994) 263-272.
17. R. Todeschini, G. Moro, R. Boggia, L. Bonati, U. Cosentino, M. Lasagni, D. Pitea, "*Modeling and prediction of molecular properties. Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptors*", Chemometrics and Intelligent Laboratory Systems, 36 (1997) 65-73.
18. R. Todeschini, R. Cazar, E. Collina, "*The chemical meaning of topological indices*", Chemometrics and Intelligent Laboratory Systems, 15 (1992) 51-59.
19. R. Todeschini, P. Gramatica, R. Provenzani, E. Marengo, "*Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons*", Chemometrics and Intelligent Laboratory Systems, 27 (1995) 221-229.
20. R. Todeschini, V. Consonni, "*Handbook of molecular descriptors*", Wiley-VCH, Weinheim, 2000.
21. J. Galvez, R. Garcia, M.T. Salabert, R. Soler, "*Charge Indexes. New Topological Descriptors*", Journal of Chemical Information and Computer Sciences, 34 (1994) 520-525.
22. V.C. Roberto Todeschini, "*Molecular Descriptors for Chemoinformatics*", Wiley-VCH, New York, USA, 2009.
23. L.B. Kier, L.H. Hall, "*Intermolecular Accessibility: The Meaning of Molecular Connectivity*", Journal of Chemical Information and Computer Sciences, 40 (2000) 792-795.
24. M. Randic, "*Characterization of molecular branching*", Journal of the American Chemical Society, 97 (1975) 6609-6615.
25. G. Ruecker, C. Ruecker, "*Counts of all walks as atomic and molecular descriptors*", Journal of Chemical Information and Computer Sciences, 33 (1993) 683-695.
26. V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, "*Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies*", Journal of Chemical Information and Computer Sciences, 42 (2002) 693-705.
27. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, "*Chemical Information in 3D Space*", Journal of Chemical Information and Computer Sciences, 36 (1996) 1030-1037.
28. J.H. Schuur, P. Selzer, J. Gasteiger, "*The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity*", Journal of Chemical Information and Computer Sciences, 36 (1996) 334-344.
29. O. Devinyak, D. Havrylyuk, R. Lesyk, "*3D-MoRSE descriptors explained*", Journal of Molecular Graphics and Modelling, 54 (2014) 194-203.
30. I.T. Jolliffe, "*Principal Component Analysis*", 2nd ed.2002.
31. J.A. Hartigan, "*Clustering Algorithms*", John Wiley & Sons, Inc.1975, pp. 351.
32. A.K. Jain, "*Data clustering: 50 years beyond K-means*", Pattern Recognition Letters, 31 (2010) 651-666.

33. R. Dubes, A.K. Jain, "*Clustering techniques: The user's dilemma*", Pattern Recognition, 8 (1976) 247-260.
34. A. Toubaei, H. Golmohammadi, Z. Dashtbozorgi, W.E. Acree Jr, "*QSPR studies for predicting gas to acetone and gas to acetonitrile solvation enthalpies using support vector machine*", Journal of Molecular Liquids, 175 (2012) 24-32.
35. K. Bodzioch, A. Durand, R. Kaliszan, T. Bączek, Y. Vander Heyden, "*Advanced QSRR modeling of peptides behavior in RPLC*", Talanta, 81 (2010) 1711-1718.
36. R.R. Hocking, "*A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression*", Biometrics, 32 (1976) 1-49.
37. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S.D. Jong, P.J. Lewi, J. Smeyers-Verbeke, "*Handbook of Chemometrics and Qualimetrics, Part A*", Elsevier, Amsterdam, 1997.
38. N.R. Draper, H. Smith, "*Applied Regression Analysis*", 2d Edition ed., John Wiley & Sons, Inc, New York, 1981.
39. J.M. Wagner, D.G. Shimshak, "*Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives*", European Journal of Operational Research, 180 (2007) 57-67.
40. R. Leardi, R. Boggia, M. Terrile, "*Genetic algorithms as a strategy for feature selection*", Journal of Chemometrics, 6 (1992) 267-281.
41. A. Niazi, R. Leardi, "*Genetic algorithms in chemometrics*", Journal of Chemometrics, 26 (2012) 345-351.
42. R. Liu, S.-S. So, "*Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility*", Journal of Chemical Information and Computer Sciences, 41 (2001) 1633-1639.
43. F. Sun, Q. Yu, J. Zhu, L. Lei, Z. Li, X. Zhang, "*Measurement and ANN prediction of pH-dependent solubility of nitrogen-heterocyclic compounds*", Chemosphere, 134 (2015) 402-407.
44. S. Riahi, M.F. Mousavi, M. Shamsipur, "*Prediction of selectivity coefficients of a theophylline-selective electrode using MLR and ANN*", Talanta, 69 (2006) 736-740.
45. S. Thamarai Selvi, S. Arumugam, L. Ganesan, "*BIONET: an artificial neural network model for diagnosis of diseases*", Pattern Recognition Letters, 21 (2000) 721-740.
46. A. Habibi-Yangjeh, E. Pourbasheer, M. Danandeh-Jenagharad, "*Prediction of basicity constants of various pyridines in aqueous solution using a principal component-genetic algorithm-artificial neural network*", Monatshefte für Chemie - Chemical Monthly, 139 (2008) 1423-1431.
47. V. Vapnik, "*Statistical learning theory*", Wiley, New York, 1998.
48. S. Riahi, E. Pourbasheer, M.R. Ganjali, P. Norouzi, "*Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine*", Journal of Hazardous Materials, 166 (2009) 853-859.
49. L. Lin, "*A concordance correlation coefficient to evaluate reproducibility.*", Biometrics, 45 (1989) 255-268.
50. A. Tropsha, "*Best Practices for QSAR Model Development, Validation, and Exploitation*", Molecular Informatics, 29 (2010) 476-488.
51. A. Golbraikh, A. Tropsha, "*Beware of q2!*", Journal of Molecular Graphics and Modelling, 20 (2002) 269-276.

52. F. Ruggiu, P. Gizzi, J.-L. Galzi, M. Hibert, J. Haiech, I. Baskin, D. Horvath, G. Marcou, A. Varnek, "*Quantitative Structure–Property Relationship Modeling: A Valuable Support in High-Throughput Screening Quality Control*", Analytical Chemistry, 86 (2014) 2510-2520.
53. T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang, "*Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships*", Alternatives to Laboratory Animals, 33 (2005) 1-19.
54. H. Golmohammadi, Z. Dashtbozorgi, W.E. Acree Jr, "*Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine*", European Journal of Pharmaceutical Sciences, 47 (2012) 421-429.
55. E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibáñez, T. Portolés, R.d. Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipaničev, P. Rostkowski, J. Hollender, "*Non-target screening with high resolution mass spectrometry: Critical review using a collaborative trial on water analysis*", Anal. Bioanal. Chem., (2015).
56. Partitioning(logD), "*Marvin 6.3.1*", ChemAxon (http://www.chemaxon.com)" 2014.
57. A.K. Ghose, G.M. Crippen, "*Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity*", Journal of Computational Chemistry, 7 (1986) 565-577.
58. A.K. Ghose, A. Pritchett, G.M. Crippen, "*Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions*", Journal of Computational Chemistry, 9 (1988) 80-90.
59. M.A. Torres, M.P. Barros, S.C.G. Campos, E. Pinto, S. Rajamani, R.T. Sayre, P. Colepicolo, "*Biochemical biomarkers in algae and marine pollution: A review*", Ecotoxicology and Environmental Safety, 71 (2008) 1-15.
60. K. Goryński, B. Bojko, A. Nowaczyk, A. Buciński, J. Pawliszyn, R. Kaliszan, "*Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds*", Analytica Chimica Acta, 797 (2013) 13-19.
61. S. He, Y. Shao, L. Fan, Z. Che, H. Xu, X. Zhi, J. Wang, X. Yao, H. Qu, "*Synthesis and Quantitative Structure–Activity Relationship (QSAR) Study of Novel 4-Acyloxypodophyllotoxin Derivatives Modified in the A and C Rings as Insecticidal Agents*", Journal of Agricultural and Food Chemistry, 61 (2013) 618-625.
62. V. Consonni, R. Todeschini, M. Pavan, "*Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors*", Journal of Chemical Information and Computer Sciences, 42 (2002) 682-692.
63. F.R. Burden, "*Molecular identification number for substructure searches*", Journal of Chemical Information and Computer Sciences, 29 (1989) 225-227.