Chemical Methodologies

journal homepage: http://chemmethod.com

Original Research article

# Advanced QSRR Modeling of Organic Pollutants in Natural Water and Wastewater in Gas Chromatography Time-of-Flight Mass Spectrometry

Mehrdad Shahpar[a] * and Sharmin Esmaeilpoor[b]

a Director of Ilam Petrochemical Company, Ilam, Iran

b Department of Chemistry, Payame Noor University, P.O. BOX 19395-4697, Tehran, Iran

**ARTICLE INFORMATION**

Received: 07 October 2017
Received in revised: 17 November 2017
Accepted: 10 December 2017
Available online: 30 December 2017

DOI:
**10.22631/chemm.2017.100307.1012**

**KEYWORDS**

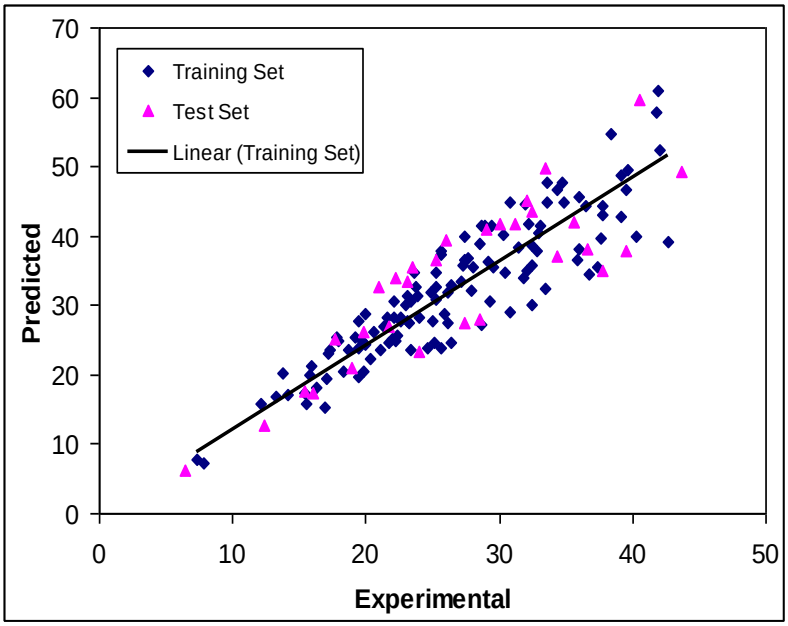
Water pollution
Hazardous chemicals
Organic pollutants
Gas chromatography
Time-of-flight mass spectrometry
Chemometrics
Levenberg-Marquardt artificial neural network

**ABSTRACT**

Water pollution is a major global problem that requires ongoing evaluation and revision of water resource policy . Water pollution is the major cause of death and diseases, resulting in deaths of more than 14,000 people daily. Genetic algorithm-partial least square (GA-PLS), Kernel partial least square (GA-KPLS) and Levenberg-Marquardt artificial neural network (L-M ANN) techniques were used to investigate the correlation between the retention time (RT) and descriptors for 150 organic contaminants in natural water and wastewater which obtained by gas chromatography coupled to high-resolution time-of-flight mass spectrometry (GC-TOF MS). The L-M ANN model showed a better performance in comparison with other models, indicating that L-M ANN model can be used as an alternative modeling tool for quantitative structure–retention relationship (QSRR) studies.

** Corresponding author: E-mail address: Shahpar2012@gmail.com*
*Director of Ilam Petrochemical Company, Tel.: +98 9143100801*

**Graphical Abstract**




**Introduction**

Water pollution is the contamination of the water bodies including lakes, rivers, oceans, and groundwater. Water pollution occurs when pollutants are discharged directly or indirectly into the water bodies without adequate treatment to remove the harmful compounds. Water pollution affects plants and organisms living in the water bodies. In almost all cases, not only it has a negative effect on the individual species and populations, but also it damages the natural biological communities [1,2].

An estimated 700 million Indians have no access to a proper toilet, and 1,000 Indian children die because of diarrheal sickness every day. 90% of cities in China suffers from water pollution, and around 500 million people do not have access to a safe drinking water [2,3]. In addition to the acute problems of water pollution in developing countries, the developed countries struggle with the pollution problems as well. In the most recent national report on water quality in the United States, 45 % of the assessed stream miles, 47 % of the assessed lake acres, and 32 % of the assessed bay and estuarine square miles were classified as pollution [3].

atural phenomena such as volcanoes, algae blooms, storms, and earthquakes also cause major changes in water quality and the ecological status of water Surface water and groundwater have often been studied and managed as separate resources, although they are interrelated. Surface water seeps through the soil and becomes groundwater. Conversely, groundwater can also feed

surface water sources. Sources of surface water pollution are generally grouped into two categories based on their origin [4].

Contaminants in water may include organic and inorganic substances. Some organic water pollutants are:

Insecticides and herbicides, a huge range of organohalide and other chemicals Bacteria, often is from sewage or livestock operations Food processing waste, including pathogens Tree and brush debris from logging operations VOCs (Volatile Organic Compounds, industrial solvents) from improper storage [5, 6]. Some inorganic water pollutants include: Heavy metals including acid mine drainage Acidity caused by industrial discharges (especially sulfur dioxide from power plants) Chemical waste as industrial by products Fertilizers, in runoff from agriculture including nitrates and phosphates Silt in surface runoff from construction sites, logging, slash and burn practices or land clearing sites [7]. Organic pollution occurs when an excess of organic matter, such as manure or sewage enters the water. When organic pollution increases in a pond, the number of decomposers will increase. As the aquatic organisms die, they are broken down by decomposers, leading to further depletion of the oxygen. A type of organic pollution can occur when inorganic pollutants such as nitrogen and phosphates accumulate in aquatic ecosystems. High level of these nutrients cause an overgrowth of plants and algae. As the plants and algae die, they become organic material in the water. The enormous decay of this plant matter, in turn, lowers the oxygen level. The process of rapid plant growth followed by increased activity by decomposers and a depletion of the oxygen level is called *eutrophication* [5, 6].

There are several important reasons why social scientists should examine the causes of organic water pollution. First, it is largely the result of human activities. The industrial activities that contribute to organic water pollution include manufacturing of glass, pesticides, medicines, plastics, ceramics, textiles, metals, and paper [8]. Some other activities that contribute to water pollution include food processing facilities with inadequate disposal facilities and the dispersing of water used to cool coke during steel production. The chemicals and byproducts of these manufacturing and industrial processes often end up as waste and are disposed of by being dumped into rivers, lakes, and streams [9].

Second, water pollution has been associated with many other environmental problems. For instance, many chemicals that dumped into waterways are not only highly toxic but also take a long time to decompose. Consequently, there is a shift in the pH of water. The pH shift causes certain plants and animals to die off while allowing others to reproduce unchecked, thereby reducing biodiversity. Some water pollutants also stimulate oxygen consumption by plants, algae, and

bacteria. This process reduces levels of dissolved oxygen creating a situation of chronic “stress” that lowers the body weight of aquatic animals and makes them less able to compete for food and habitat. It also creates a situation that is toxic to some fish and aquatic invertebrates, which die due to lack of oxygen [10].

Third, water pollution from industrial and manufacturing activity has serious health effects in humans [11, 12]. The toxic chemicals found in water supplies affect people through the process of “bioaccumulation” or the building up of toxins in the fatty tissue of mammals. The long-term effects of bioaccumulation in adults include cancer, blood disorders, immunity suppression, and spontaneous abortions. The buildup of these pollutants has been linked to birth defects.

The United States Environmental Protection Agency (EPA) monitors and analyzes organic pollutants in water. The EPA has established a list of a "dirty dozen" particularly widespread and persistent organic pollutants (POPs). Part of the EPA's mandate is to identify where these pollutants occur in water resources and to contain or mitigate POPs.

The POPs include intentionally produced chemicals such as pesticides as well as industry or combustion by-products. The dirty dozen are aldrin, chlordane, DDT, dieldrin, endrin, heptachlor, hexachlorobenzene, mirex, toxaphene, PCBs, dioxins and furans.

EPA laboratories in Cincinnati, Ohio and Athens, Georgia investigated analytical methods to analyze organic pollutants in water based on gas chromatography separation of the pollutants and mass spectrometer identification and quantification. The research results were published as the EPA's test methods 624 and 625 for the standard analysis of organic pollutants in municipal and industrial effluent [13].

Gas chromatography separates organic pollutants for further analysis. The researcher injects a sample into the gas chromatography instrument. The instrument heats the sample to a gas and injects it into the gas chromatography tube or column. As the sample travels the length of the column, the different organic molecules condense and liquefy and then vaporize as a gas again. As a liquid, the molecules stick to the column, but as a gas, they travel through the column quickly. Different pollutants have different ratios of gas to liquid, so they each travel through the column at different speeds [14]. The separated pollutants are then analyzed by mass spectrometry. A mass spectrometer ionizes a sample and shoots it through an electric field. The electric field bends the path (trajectory) of lighter molecules more than that of heavy molecules. The sample strikes a detector at a certain position based on its mass. This method identifies and quantifies organic pollutants in water after they have been separated by gas chromatography. The combination of gas

chromatography and mass spectrometry give researchers complete information on the type of organic pollutants in a sample and the concentration of each pollutant in the sample.

Most of these methods are focused on target analysis with quantitative purposes and their scope rarely exceeds several tens of analytes, being quite unusual to find analytical methods for the determination of more than 100 organic pollutants. In the last decade there has been a notable increase in the use of full spectrum acquisition techniques, such as time-of-flight mass spectrometry (TOF MS), which allows acquiring huge amount of chemical information on the sample in a single analysis [15, 16]. This facilitates widening the number of analytes that can be searched in a single experiment, with the additional advantage that data can be re-examined at any time to search for other compounds not included in the first screening, without the need of additional analysis. TOFMS and hybrid quadrupole-TOFMS have been successfully applied for screening purposes in combination with gas chromatography (GC) or liquid-chromatography (LC) in different applied fields, like environmental analysis, food safety or toxicology. This analyzer provides the selectivity and sensitivity required for wide-scope screening, as it combines high full-spectral sensitivity with high mass resolution. Accurate mass data obtained can be processed in both "post-target" and/or non-target way, which gives high versatility to the instrument which allows the user to tackle an analytical problem in different ways, depending on the aim of the analysis [15-17].

Prediction of physico–chemical properties of materials based on their molecular structure has been one of the wishes of scientists and engineers for a long time. One of the best methods which have been applied for this purpose is quantitative structure–property relationships (QSRR). QSRR analysis is now a well established and highly respected technique to correlate chromategraphic retention time of a compound with its molecular structure, through a variety of descriptors. The basic strategy of QSRR analysis is to find optimum quantitative relationships, which can then be used for the prediction of the retention from molecular structures [18, 19]. Once a reliable relation has been obtained, it is possible to use it to predict that retention for other structures not yet measured or even not yet prepared. QSRR on the retention time have been reported for different types of organic compounds [20-22].

The application of this technique usually requires variable selection for building well-fitted models. Nowadays, the genetic algorithm method (GA) is well known as an interesting and more widely used variable selection method. GA is a stochastic method that solves the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation [23, 24].

In this work, we aim to construct a QSRR model of the retention time of organic contaminants in natural water and wastewater and their theoretically derived descriptors. After the variables were selected, the linear multivariate regressions (e.g. the partial least squares (PLS)) as well as the non-linear regressions (e.g. the kernel PLS (KPLS), Levenberg- Marquardt artificial neural network (L-M ANN)) were utilized to construct the linear and non-linear QSRR models. The sets of variables, which provide the best-fitted models for PLS and KPLS methods, were selected with the help of the genetic algorithm.

## Materials and methods

### Equipment

A Pentium IV personal computer (CPU at 3.06 GHz) with the Windows XP operating system was used. The geometry optimization was performed with HyperChem (Version 7.0 Hypercube, Inc). For the calculation of the molecular descriptors, the Dragon 2.1 software was used. The GA-PLS, GA-KPLS, L-M ANN, cross validation and the other calculations were performed in the MATLAB (Version 7.0, Math works, Inc).

### Data set and descriptor generation

The data set used in this study, is the retention time (RT) of organic contaminants in natural water and wastewater (a total number of 150 molecules), which obtained by gas chromatography time-of-flight mass spectrometry (GC-TOF) were taken from the literature [25] is shown in Table 1 and Table 2. The constituents of organic pollutants in natural water and wastewater includes: PAHs, octyl/nonyl phenols, PCBs, PBDEs and a notable number of pesticides, such as insecticides (organochlorines, organophosphorus, carbamates and pyrethroids), herbicides (triazines and chloroacetanilides), fungicides and several relevant metabolites. Water samples of different types and origin were collected from different sites of the Castellón province (Spain). Concretely, two surface water (SW) (Villarreal and Burriana), two ground water (GW) (Almassora and Castellón), and two effluent water samples (EWW) from a wastewater treatment plant (WWTP) of Castellón were collected. The chemical structure of the 150 studied molecules were drawn with the Hyperchem software and saved with the HIN extension. To optimize the geometry of the studied molecules, the AM1 geometrical optimization was applied. The DRAGON software was used to calculate the descriptors in this research and a total of 1497 molecular descriptors, belonging to 18 different types of the theoretical descriptors, were calculated for each molecule.

### Instrumentation

GC instrumentation consisted of an Agilent 6890N GC system (Paloalto, CA, USA), equipped with an Agilent 7683 autosampler, coupled to a time-of-flight mass spectrometer, GCT (Waters Corporation, Manchester, UK), operating in electron ionization (EI) mode. The GC separation was performed using a fused silica HP-5MS capillary column of 30m×0.25mm i.d. and a film thickness of 0.25 $\mu$ m (J&W Scientific, Folson, CA, USA). The oven temperature was programmed as follows: 90 °C (1min); 5 °C/min to 300 °C (2min). Splitless injections of 1 $\mu$ L sample were carried out. Helium was used as carrier gas at 1mL/min. The interface and source temperatures were both set to 250 °C and a solvent delay of 3min was selected. TOF MS was operated at 1 spectrum/s acquiring the mass range m/z 50–650 and using a multi-channel plate voltage of 2800V. TOF-MS resolution was about 8500 (FWHM) atm/z 614. Heptacosa, used for the daily mass calibration as well as lockmass, was injected via syringe in the reference reservoir at 30 °C. The m/z ion monitored was 218.9856. The application manager TargetLynx, a module of MassLynx software, was used to process data obtained from standards and samples for target compounds. The application manager ChromaLynx, also a module of MassLynx software, was used to investigate the presence of non-target compounds in samples. Library searching was performed using the commercial NIST library.

**Data pretreatment**

To decrease the redundancy existing in the descriptor data matrix, those descriptors which contribute either no information or whose information content is redundant with other descriptors present in the pool. Then, the remaining descriptors were collected in an n × m data matrix (D), where n = 150 and m=1019 are the number of the compounds and the descriptors, respectively. These descriptors were employed to generate the models with the GA-PLS and GA-KPLS program.

**Genetic algorithm**

Genetic algorithm is a problem-solving method that uses generic rules such as reproduction, crossover and mutation to build pseudo organisms that are then selected based on a fitness criterion to survive and pass information on to the next generation [26]. GA uses a binary bit string representation as the coding technique for a given problem; the presence or absence of a descriptor in a chromosome is coded by 1 or 0. A string is composed of several genes that represent a specific characteristic to be studied. In the present case, a string is composed of 561 genes representing the presence or absence of a descriptor. By encoding various descriptors with bit strings, called chromosomes, the initial population was created randomly. The population size was varied between 50 and 300 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations had taken the same fitness [27, 28]. In this paper, size of the

population is 30 chromosomes, the probability of initial variable selection is 5:V (V is the number of independent variables), crossover is multi Point, the probability of crossover is 0.5, mutation is multi Point, the probability of mutation is 0.01 and the number of evolution generations is 1000. For each set of data, 5000 runs were performed.

**Nonlinear model**

**Artificial neural network**

A three-layer back propagation artificial neural network ANN with a sigmoid transfer function was used in the investigation of feature sets. The descriptors from the training set were used for the model generation whereas the descriptors from the validation set were used to stop the overtraining of network. And the descriptors from the validation set were used to verify the predictivity of the model. Before training the networks, the input and output values were normalized with auto-scaling of all data [29, 30]. To compare the results, the same number of hidden layer nodes was used for the ANN models from all other feature sets of each database. The goal of training the network is to minimize the output errors by changing the weights between the layers.

$$\Delta W_{ij,n} = F_n + \alpha \Delta W_{ij,n-1} \quad (1)$$

In this, $\Delta W_{ij}$ is the change in the weight factor for each network node, α is the momentum factor, and F is a weight update function, which indicates how weights are changed during the learning process. The weights of hidden layer were optimized using the Levenberg-Marquardt algorithm, a second derivative optimization method [31].

**Levenberg-Marquardt Algorithm**

In Levenberg-Marquardt algorithm, the update function, $F_n$, was calculated using equations (2-4).

$$F_0 = -g_0 \quad (2)$$

$$g = J^T e \quad (3)$$

$$F_n = -[J^T \times J + \mu I]^{-1} \times J^T \times e \quad (4)$$

Where g is gradient, and J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights, and e is a vector of network errors. The parameter μ is multiplied by some factor (λ) whenever a step would result in an increased e and when a step reduces e, μ is divided by λ [32, 33].

## Results and discussion

### Linear model

### Results of the GA-PLS model

The best model is selected based on the highest square correlation coefficient leave-group-out cross validation ($R^2$), the least root mean squares error (RMSE) and relative error (RE) of prediction. These parameters are probably the most popular measure of how well a model fits the data. The best GA-PLS model contains sixteen selected descriptors in seven latent variables space. These descriptors were obtained constitutional descriptors (mean electrotopological state (Ms)), 2D autocorrelations (Broto-Moreau autocorrelation of a topological structure - lag 5/weighted by atomic masses (ATS5m), Broto-Moreau autocorrelation of a topological structure - lag 5 / weighted by atomic van der Waals volumes (ATS5v), Broto-Moreau autocorrelation of a topological structure - lag 5 / weighted by atomic Sanderson electronegativities (ATS5e), Moran autocorrelation - lag 3/weighted by atomic polarizabilities (MATS3p), Geary autocorrelation - lag 5/weighted by atomic Sanderson electronegativities (GATS5e) and Geary autocorrelation - lag 5 / weighted by atomic polarizabilities (GATS5p)), geometrical descriptors (spherosity (SPH)), absolute eigenvalue sum on geometry matrix (SEig)), 3D-MoRSE descriptors (3D-MoRSE - signal 11/weighted by atomic masses (Mor11m), 3D-MoRSE - signal 29 / weighted by atomic masses (Mor29m) and 3D-MoRSE - signal 12/weighted by atomic van der Waals volumes (Mor12v)), GETAWAY descriptors (leverage-weighted autocorrelation of lag 5 / unweighted (HATS5u)), atom-centred fragments (CR3X (C-011) and R--CX..X (C-035)) and charge descriptors (total negative charge (Qneg)). The $R^2$ and RMSE for training and validation sets were (0.809, 0.740) and (0.599, 1.055), respectively. The predicted values of RT are plotted against the experimental values for training and test sets in Figure 1. For this in general, the number of components (latent variables) is less than the number of independent variables in PLS analysis. The PLS model uses higher number of descriptors that allow the model to extract better structural information from descriptors to result in a lower prediction error.

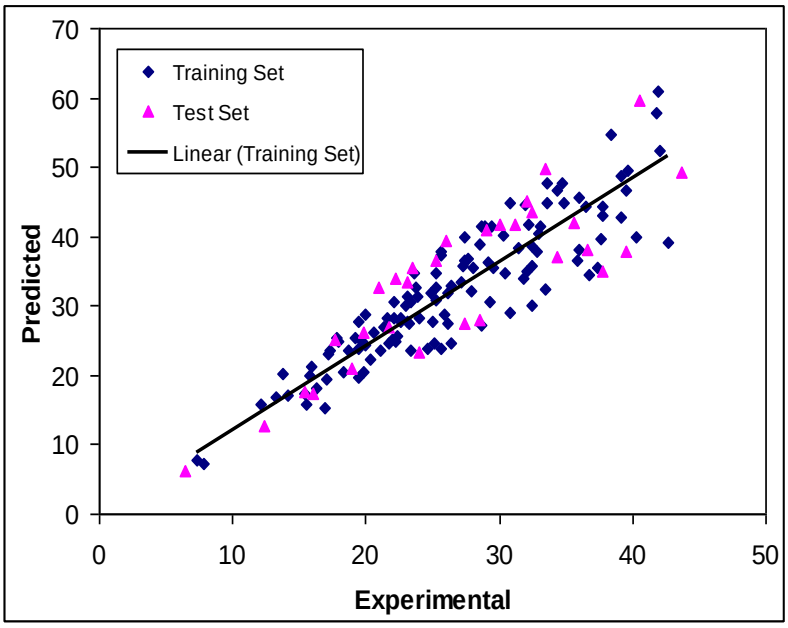



**Figure 1.** Plots of predicted retention time against the experimental values by GA-PLS model

**Nonlinear model**

**Results of the GA-KPLS model**

PLS is useful in situations where the number of explanatory variables exceeds the number of observations and/or a high level of multicollinearity among those variables is assumed. Motivated by this fact we will provide a kernel PLS algorithm for construction of nonlinear regression models in possibly high-dimensional feature spaces. PLS has proven to be useful in situations when the number of observed variables (N) is significantly greater than the number of observations (n) and high multicollinearity among the variables exists. This situation when N ≥ n is common in chemometrics and gave rise to the modification of classical principal component analysis (PCA) and linear PLS methods to their kernel variants. However, rather than assuming a nonlinear transformation into a feature space of arbitrary dimensionality the authors attempted to reduce computational complexity in the input space. Motivated by these works we propose a more general nonlinear kernel PLS algorithm.

In this paper a radial basis kernel function, $k(x,y)= \exp(||x-y||^2/c)$, was selected as the kernel function with ($c = rm\sigma^2$) where r is a constant that can be determined by considering the process to be predicted (here r was set to be 1), m is the dimension of the input space and $\sigma^2$ is the variance of the data [34]. It means that the value of c depends on the system under the study. The 13 descriptors in 5 latent variables space chosen by GA-KPLS feature selection methods were

contained. These descriptors were obtained topological descriptors (Schultz Molecular Topological Index (MTI) (SMTI), Harary H index (Har), average eccentricity (AECC) and eccentric connectivity index (CSI)), 2D autocorrelations (Broto-Moreau autocorrelation of a topological structure - lag 5/weighted by atomic van der Waals volumes (ATS5v) and Moran autocorrelation - lag 3/weighted by atomic polarizabilities (MATS3p)), Burden eigenvalues (lowest eigenvalue n. 1 of Burden matrix/weighted by atomic Sanderson electronegativities (BELe1), geometrical descriptors (average span R (SPAM)), 3D-MoRSE descriptors (3D-MoRSE - signal 03 / weighted by atomic masses (Mor03m), 3D-MoRSE - signal 19 / weighted by atomic masses (Mor19m), 3D-MoRSE - signal 23 / weighted by atomic masses (Mor23m), 3D-MoRSE - signal 17 / weighted by atomic van der Waals volumes (Mor17v)), molecular properties (Squared Moriguchi octanol-water partition coeff. (logP^2) (MLOGP2)). The $R^2$ and RMSE for training and test sets were (0.781, 0.716) and (0.649, 1.293), respectively. Figure 2 shows the plot of the GA-KPLS predicted versus experimental values for RT of all of the molecules in the data set. It can be seen from these results that statistical results for GA-PLS model are superior to GA-KPLS method.

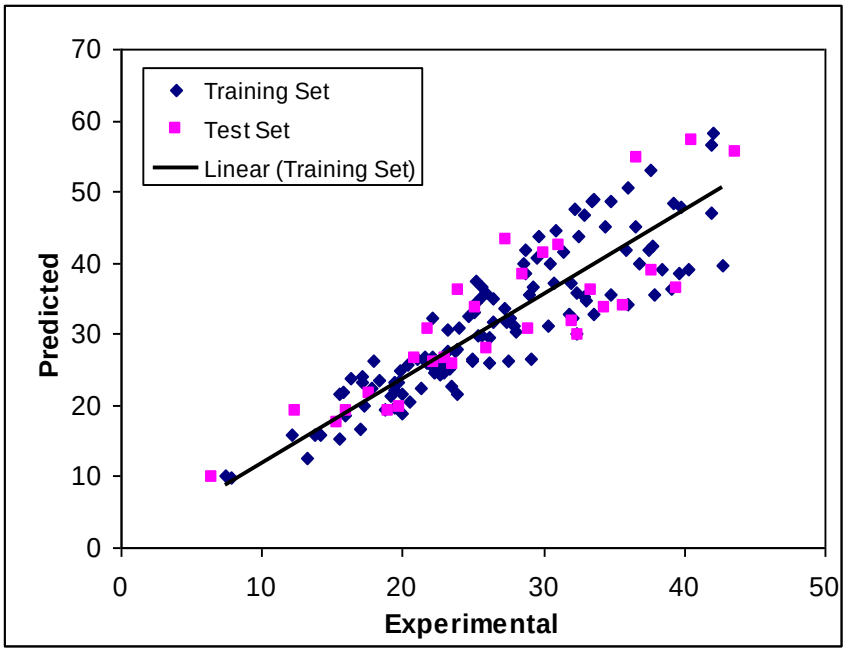
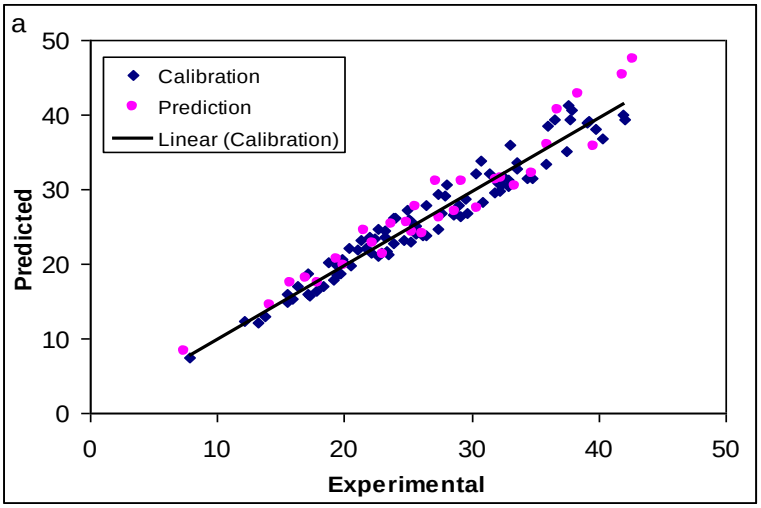



**Figure 2.** Plot of predicted RT obtained by GA-KPLS against the experimental values

**Results of the L-M ANN model**

The networks were generated using descriptors appearing in the GA-PLS model as inputs. For ANN generation, dataset was separated into three groups: calibration, prediction and test sets. Before training, the input and output values were normalized between 0 and 1. Number of neurons in the

hidden layer, learning rate and momentum were optimized. A feed-forward neural network with back-propagation algorithm was constructed to model the retention relationship [35]. This method is an iterative algorithm that allows training of multilayer networks. The algorithm looks for the minimum of the error function. In this way, the training process tries to diminish the difference between the outputs of the network and the expected values. Of course, there are some other approaches such as Levenberg Marquardt algorithm, gradient descent with variable learning rate back-propagation and resilient back-propagation. These networks are different in weight update functions and can converge faster than steepest decent method [36]. But this paper has not focused on investigating the role of weight update functions or calculation time in artificial neural networks. Our network has nine input layer, four hidden layer and one output layer. A bias unit with a constant activation of unity is connected to each unit in the hidden and output layers. Once the best topology of the network is obtained and the convergence criterion is reached, a leave-4- out cross-validation procedure is also employed to more validate the performances of the resulted networks. To evaluate the performance of the ANN, RMSE of the calibration was used. The number of neurons in the hidden layer with the minimum value of RMSE was selected as the optimum number. Learning rate and momentum were optimized in a similar way. It was realized that the RMSE for the training and test sets are minimum when four neurons were selected in the hidden layer. The $R^2$ and RMSE for calibration, prediction and test sets were (0.945, 0.929, 0.861) and (0.165, 0.353, 0.522), respectively. Inspection of the results reveals a higher $R^2$ and lowers other values parameter for the test set compared with their counterparts for other models. Plots of predicted RT versus experimental RT values by L-M ANN for calibration, prediction and test sets are shown in Figure 3a, 3b, respectively.



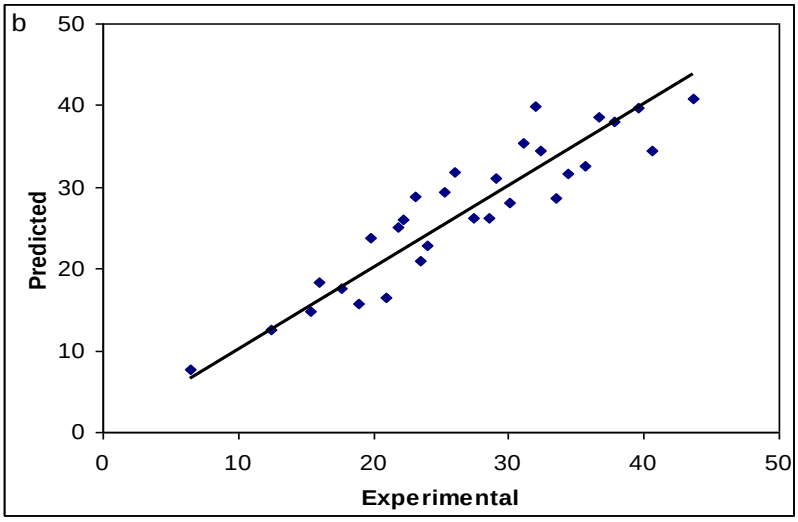
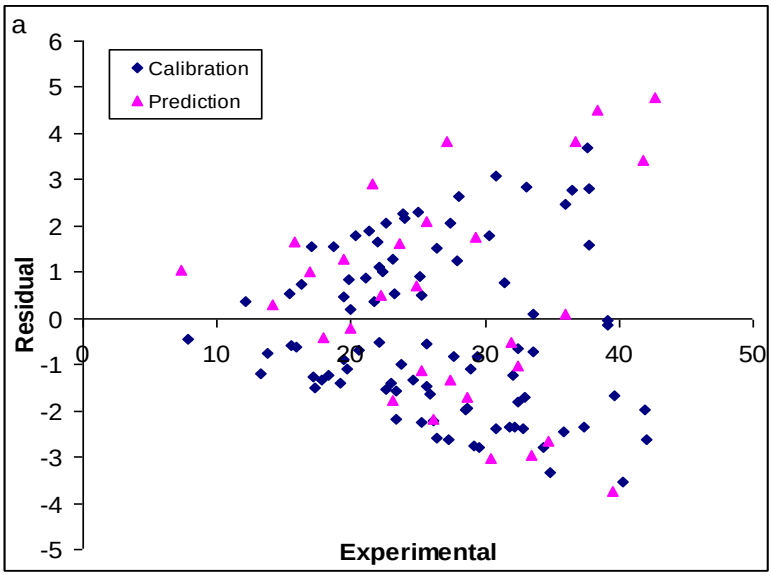



**Figure 3.** Plot of predicted RT obtained by L-M ANN against the experimental values (a) calibration and prediction sets of molecules and (b) for validation set

The residuals (predicted RT– experimental RT) obtained by the L-M ANN modeling versus the experimental RT values are shown in Figure 4a, 4b. As the calculated residuals are distributed on both sides of the zero line, one may conclude that there is no systematic error in the development of the neural network.



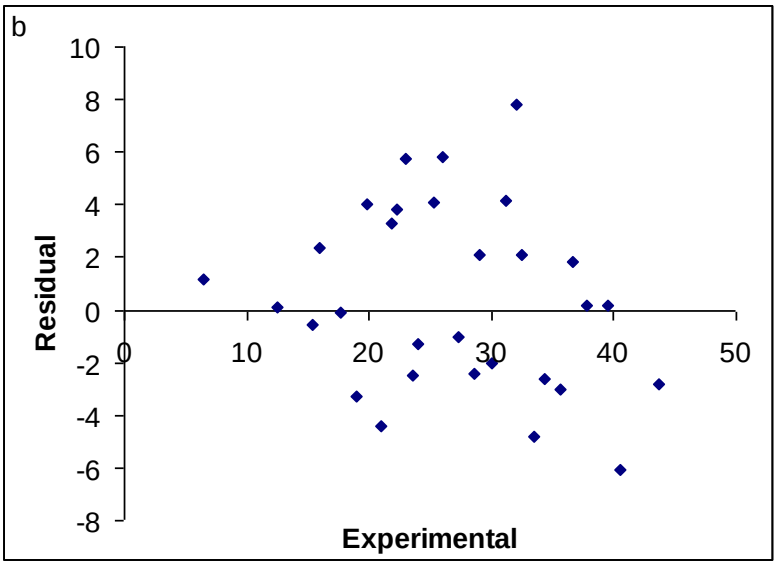



**Figure 4.** Plot of residuals obtained by L-M ANN against the experimental RT values (a) training set of molecules and (b) for test set

The values of experimental, calculated and RMSE are shown in Table 1 and Table 2 for training and test sets which obtained by L-M ANN model. The $Q^2$ of training and test sets for the GA-PLS and GA-KPLS models are (0.802, 0.734) and (0.775, 0.712) respectively which would be compared with the values of (0.943, 0.924, 0.853), respectively, for L-M ANN model. Comparison between these values and other statistical parameters reveals the superiority of the L-M ANN model over other models. The key strength of neural networks, unlike regression analysis, is their ability to flexible mapping of the selected features by manipulating their functional dependence implicitly. The statistical parameters reveal the high predictive ability of L-M ANN model.

**Table 1.** The data set, structure, the corresponding observed, calculate and root mean square error values retention time of training set for L-M ANN

| No | Name | Molecular formula | $RT_{Exp}$ | $RT_{Cal}$ | RMSE |
|---|---|---|---|---|---|
| | **Calibration Set** | | | | |
| 1 | Dichlorvos | $C_4H_7Cl_2O_4P$ | 7.85 | 7.41 | 0.047 |
| 2 | Mevinphos | $C_7H_{13}O_6P$ | 12.08 | 12.45 | 0.039 |
| 3 | Acenaphthene | $C_{12}H_{10}$ | 13.25 | 12.06 | 0.125 |
| 4 | Methacrifos | $C_7H_{13}O_5PS$ | 13.8 | 13.05 | 0.079 |
| 5 | Heptenophos | $C_9H_{12}ClO_4P$ | 15.45 | 15.97 | 0.055 |
| 6 | Fluorene | $C_{13}H_{10}$ | 15.47 | 14.89 | 0.062 |
| 7 | Tecnazene | $C_6HCl_4NO_2$ | 15.95 | 15.33 | 0.065 |
| 8 | Diphenylamine | $C_{12}H_{11}N$ | 16.33 | 17.05 | 0.076 |
| 9 | Chlorpropham | $C_{10}H_{12}ClNO_2$ | 17.08 | 18.63 | 0.164 |
| 10 | Terbumeton desethyl | $C_8H_{15}N_5O$ | 17.18 | 15.91 | 0.134 |

| | | | | | |
|---|---|---|---|---|---|
| 11 | Atrazine desethyl | $C_6H_{10}ClN_5$ | 17.28 | 15.79 | 0.157 |
| 12 | Trifluraline | $C_{13}H_{16}F_3N_3O_4$ | 17.79 | 16.46 | 0.141 |
| 13 | Hexachlorobenzene | $C_6Cl_6$ | 18.3 | 17.08 | 0.129 |
| 14 | Dimethoate | $C_5H_{12}NO_3PS_2$ | 18.68 | 20.22 | 0.162 |
| 15 | Atrazine | $C_8H_{14}ClN_5$ | 19.2 | 17.80 | 0.147 |
| 16 | Lindane | $C_6H_6Cl_6$ | 19.39 | 19.84 | 0.048 |
| 17 | Terbumeton | $C_{10}H_{19}N_5O$ | 19.47 | 18.57 | 0.095 |
| 18 | Phenanthrene | $C_{14}H_{10}$ | 19.72 | 18.62 | 0.116 |
| 19 | Fonofos | $C_{10}H_{15}OPS_2$ | 19.8 | 20.64 | 0.089 |
| 20 | Propyzamide | $C_{12}H_{11}Cl_2NO$ | 19.92 | 20.12 | 0.021 |
| 21 | Diazinon | $C_{12}H_{21}N_2O_3PS$ | 20.37 | 22.17 | 0.190 |
| 22 | Terbacil | $C_9H_{13}ClN_2O_2$ | 20.54 | 19.84 | 0.074 |
| 23 | Endosulfan ether | $C_9H_6Cl_6O$ | 21.04 | 21.90 | 0.091 |
| 24 | Pirimicarb | $C_{11}H_{18}N_4O_2$ | 21.35 | 23.24 | 0.200 |
| 25 | PCB 28 | $C_{12}H_7Cl_3$ | 21.69 | 22.04 | 0.037 |
| 26 | Chlorpyrifos methyl | $C_7H_7Cl_3NO_3PS$ | 21.95 | 23.62 | 0.176 |
| 27 | Parathion methyl | $C_8H_{10}NO_5PS$ | 22.05 | 23.15 | 0.116 |
| 28 | Chlozolinate | $C_{13}H_{11}Cl_2NO_5$ | 22.08 | 21.55 | 0.056 |
| 29 | Alachlor | $C_{14}H_{20}ClNO_2$ | 22.39 | 23.40 | 0.107 |
| 30 | Fenchlorphos | $C_8H_8Cl_3O_3PS$ | 22.62 | 21.09 | 0.161 |
| 31 | Metalaxyl | $C_{15}H_{21}NO_4$ | 22.63 | 24.69 | 0.218 |
| 32 | Methiocarb sulfone | $C_{11}H_{15}NO_4S$ | 22.92 | 21.53 | 0.147 |
| 33 | Methiocarb | $C_{11}H_{15}NO_2S$ | 23.14 | 24.42 | 0.135 |
| 34 | Fenitrothion | $C_9H_{12}NO_5PS$ | 23.17 | 23.70 | 0.055 |
| 35 | Pirimiphos methyl | $C_{11}H_{20}N_3O_3PS$ | 23.32 | 21.74 | 0.166 |
| 36 | Dichlofluanide | $C_9H_{11}Cl_2FN_2O_2S_2$ | 23.42 | 21.22 | 0.232 |
| 37 | Metolachlor | $C_{15}H_{22}ClNO_2$ | 23.79 | 22.78 | 0.106 |
| 38 | Fenthion | $C_{10}H_{15}O_3PS_2$ | 23.92 | 26.17 | 0.238 |
| 39 | Chlorpyrifos | $C_9H_{11}Cl_3NO_3PS$ | 24 | 26.16 | 0.228 |
| 40 | Isodrin | $C_{12}H_8Cl_6$ | 24.62 | 23.28 | 0.141 |
| 41 | Cyprodinil | $C_{14}H_{15}N_3$ | 24.95 | 27.24 | 0.241 |
| 42 | Heptachlor epoxide B | $C_{10}H_5Cl_7O$ | 25.09 | 25.99 | 0.095 |
| 43 | Fluoranthene | $C_{16}H_{10}$ | 25.2 | 22.96 | 0.236 |
| 44 | Heptachlor epoxide A | $C_{10}H_5Cl_7O$ | 25.25 | 25.75 | 0.053 |
| 45 | Chlorfenvinphos | $C_{12}H_{14}Cl_3O_4P$ | 25.57 | 25.02 | 0.058 |
| 46 | Isofenphos | $C_{15}H_{24}NO_4PS$ | 25.6 | 24.14 | 0.154 |
| 47 | Procymidone | $C_{13}H_{11}Cl_2NO_2$ | 25.85 | 24.21 | 0.173 |
| 48 | Methidathion | $C_6H_{11}N_2O_4PS_3$ | 26.12 | 23.91 | 0.233 |
| 49 | Fenoxycarb | $C_{17}H_{19}NO_4$ | 26.37 | 23.77 | 0.274 |
| 50 | $\alpha$-Endosulfan | $C_9H_6Cl_6O_3S$ | 26.42 | 27.95 | 0.161 |
| 51 | PCB 77 | $C_{12}H_6Cl_4$ | 27.32 | 24.69 | 0.278 |
| 52 | Dieldrin | $C_{12}H_8Cl_6O$ | 27.39 | 29.46 | 0.218 |
| 53 | PCB 81 | $C_{12}H_6Cl_4$ | 27.69 | 26.87 | 0.086 |
| 54 | Buprofezin | $C_{16}H_{23}N_3OS$ | 27.87 | 29.11 | 0.131 |
| 55 | Bupimirate | $C_{13}H_{24}N_4O_3S$ | 28.07 | 30.70 | 0.277 |
| 56 | $\beta$-Endosulfan | $C_9H_6Cl_6O_3S$ | 28.52 | 26.56 | 0.207 |

| | | | | | |
|---|---|---|---|---|---|
| 57 | BDE 28 | $C_{12}H_7OBr_3$ | 28.68 | 26.75 | 0.204 |
| 58 | p, $p'$-DDD | $C_{14}H_{10}Cl_4$ | 28.97 | 27.89 | 0.114 |
| 59 | Oxadixyl | $C_{14}H_{18}N_2O_4$ | 29.15 | 26.41 | 0.289 |
| 60 | PCB 153 | $C_{12}H_4Cl_6$ | 29.47 | 28.63 | 0.089 |
| 61 | PCB 123 | $C_{12}H_5Cl_5$ | 29.59 | 26.79 | 0.295 |
| 62 | p, $p'$-DDT | $C_{14}H_9Cl_5$ | 30.3 | 32.08 | 0.188 |
| 63 | PCB 126 | $C_{12}H_5Cl_5$ | 30.75 | 33.82 | 0.324 |
| 64 | Tebuconazole | $C_{16}H_{22}ClN_3O$ | 30.8 | 28.40 | 0.253 |
| 65 | PCB 156 | $C_{12}H_4Cl_6$ | 31.45 | 32.22 | 0.081 |
| 66 | Benzo(a)anthracene | $C_{18}H_{12}$ | 31.84 | 29.50 | 0.247 |
| 67 | Phosmet | $C_{11}H_{12}NO_4PS_2$ | 32.08 | 30.86 | 0.129 |
| 68 | PCB 157 | $C_{12}H_4Cl_6$ | 32.24 | 29.89 | 0.248 |
| 69 | Bifenthrin | $C_{23}H_{22}ClF_3O_2$ | 32.39 | 30.60 | 0.189 |
| 70 | PCB 167 | $C_{12}H_4Cl_6$ | 32.44 | 31.77 | 0.071 |
| 71 | PCB 180 | $C_{12}H_3Cl_7$ | 32.84 | 30.44 | 0.253 |
| 72 | BDE 47 | $C_{12}H_6OBr_4$ | 32.92 | 31.22 | 0.180 |
| 73 | Tetradifon | $C_{12}H_6Cl_4O_2S$ | 33.07 | 35.90 | 0.298 |
| 74 | PCB 169 | $C_{12}H_4Cl_6$ | 33.55 | 32.82 | 0.077 |
| 75 | Mirex | $C_{10}Cl_{12}$ | 33.62 | 33.70 | 0.008 |
| 76 | Fenarimol | $C_{17}H_{12}Cl_2N_2O$ | 34.39 | 31.59 | 0.295 |
| 77 | PCB 189 | $C_{12}H_3Cl_7$ | 34.82 | 31.49 | 0.351 |
| 78 | Permethrin II | $C_{21}H_{20}Cl_2O_3$ | 35.9 | 33.46 | 0.258 |
| 79 | Coumaphos | $C_{14}H_{16}ClO_5PS$ | 36.02 | 38.50 | 0.262 |
| 80 | Benzo(b)fluoranthene | $C_{20}H_{12}$ | 36.55 | 39.32 | 0.292 |
| 81 | Cypermethrin I | C22H19Cl2NO3 | 37.42 | 35.08 | 0.247 |
| 82 | Cypermethrin II | $C_{22}H_{19}Cl_2NO_3$ | 37.62 | 41.32 | 0.390 |
| 83 | Cypermethrin IV | $C_{22}H_{19}Cl_2NO_3$ | 37.79 | 39.39 | 0.169 |
| 84 | Benzo(a)pyrene | $C_{20}H_{12}$ | 37.81 | 40.62 | 0.296 |
| 85 | Fenvalerate I | $C_{25}H_{22}ClNO_3$ | 39.15 | 38.99 | 0.017 |
| 86 | BDE 154 | $C_{12}H_4OBr_6$ | 39.17 | 39.14 | 0.003 |
| 87 | $\tau$-Fluvalinate II | $C_{26}H_{22}ClF_3N_2O_3$ | 39.7 | 38.01 | 0.178 |
| 88 | BDE 153 | $C_{12}H_4OBr_6$ | 40.3 | 36.76 | 0.373 |
| 89 | Indeno(1,2,3,cd)pyrene | $C_{22}H_{12}$ | 41.89 | 39.90 | 0.210 |
| 90 | Dibenzo(a,h)anthracene | $C_{22}H_{14}$ | 42.07 | 39.43 | 0.278 |
| | Prediction Set | | | | |
| 91 | Methamidophos | $C_2H_8NO_2PS$ | 7.35 | 8.40 | 0.191 |
| 92 | Pentachlorobenzene | $C_6HCl_5$ | 14.09 | 14.40 | 0.056 |
| 93 | Omethoate | $C_5H_{12}NO_4PS$ | 15.72 | 17.36 | 0.300 |
| 94 | Atrazine desisopropyl | $C_5H_8ClN_5$ | 16.98 | 18.01 | 0.187 |
| 95 | Phorate | $C_7H_{17}O_2PS_3$ | 17.97 | 17.55 | 0.077 |
| 96 | 4-n-Octylphenol | $C_{14}H_{22}O$ | 19.44 | 20.71 | 0.232 |
| 97 | Anthracene | $C_{14}H_{10}$ | 19.92 | 19.71 | 0.038 |
| 98 | 4-n-Nonylphenol | $C_{15}H_{24}O$ | 21.57 | 24.47 | 0.530 |
| 99 | Carbaryl | $C_{12}H_{11}NO_2$ | 22.22 | 22.71 | 0.089 |
| 100 | Terbutryn | $C_{10}H_{19}N_5S$ | 23.09 | 21.31 | 0.326 |
| 101 | Malathion | $C_{10}H_{19}O_6PS_2$ | 23.67 | 25.29 | 0.296 |

| | | | | | |
|---|---|---|---|---|---|
| 102 | Pirimiphos ethyl | $C_{13}H_{24}N_3O_3PS$ | 24.9 | 25.62 | 0.131 |
| 103 | Thiabendazole | $C_{10}H_7N_3S$ | 25.3 | 24.17 | 0.206 |
| 104 | Quinalphos | $C_{12}H_{15}N_2O_3PS$ | 25.65 | 27.73 | 0.381 |
| 105 | Pyrene | $C_{16}H_{10}$ | 26.15 | 23.98 | 0.396 |
| 106 | Imazalil | $C_{14}H_{14}Cl_2N_2O$ | 27.2 | 31.04 | 0.700 |
| 107 | p, $p'$-DDE | $C_{14}H_8Cl_4$ | 27.45 | 26.13 | 0.241 |
| 108 | PCB 118 | $C_{12}H_5Cl_5$ | 28.64 | 26.93 | 0.312 |
| 109 | Ethion | $C_9H_{22}O_4P_2S_4$ | 29.24 | 30.98 | 0.318 |
| 110 | PCB 138 | $C_{12}H_4Cl_6$ | 30.45 | 27.41 | 0.556 |
| 111 | Iprodione | $C_{13}H_{13}Cl_2N_3O_3$ | 31.89 | 31.38 | 0.094 |
| 112 | BDE 71 | $C_{12}H_6OBr_4$ | 32.4 | 31.39 | 0.185 |
| 113 | BDE 66 | $C_{12}H_6OBr_4$ | 33.47 | 30.50 | 0.543 |
| 114 | Pyrazophos | $C_{14}H_{20}N_3O_5PS$ | 34.74 | 32.08 | 0.485 |
| 115 | BDE 100 | $C_{12}H_5OBr_5$ | 35.95 | 36.04 | 0.016 |
| 116 | BDE 99 | $C_{12}H_5OBr_5$ | 36.8 | 40.64 | 0.700 |
| 117 | BDE 85 | $C_{12}H_5OBr_5$ | 38.35 | 42.86 | 0.824 |
| 118 | $\tau$ -Fluvalinate I | $C_{26}H_{22}ClF_3N_2O_3$ | 39.57 | 35.84 | 0.682 |
| 119 | BDE 138 | $C_{12}H_4OBr_6$ | 41.85 | 45.27 | 0.625 |
| 120 | Benzo(g,h,l)perylene | $C_{22}H_{12}$ | 42.69 | 47.47 | 0.873 |

**Table 2 .** The data set, structure, observed, calculate and RMSE values RT for test set by L-M ANN

| No | Name | Molecular formula | $RT_{Exp}$ | $RT_{Cal}$ | RMSE |
|---|---|---|---|---|---|
| 1 | Naphthalene | $C_{10}H_8$ | 6.5 | 7.66 | 0.212 |
| 2 | Acenaphthylene | $C_{12}H_8$ | 12.43 | 12.52 | 0.016 |
| 3 | Molinate | $C_9H_{17}NOS$ | 15.38 | 14.83 | 0.101 |
| 4 | 4-t-Octylphenol | $C_{14}H_{22}O$ | 15.99 | 18.37 | 0.434 |
| 5 | Terbuthylazine desethyl | $C_7H_{12}ClN_5$ | 17.68 | 17.59 | 0.016 |
| 6 | Simazine | $C_7H_{12}ClN_5$ | 18.95 | 15.64 | 0.604 |
| 7 | Terbuthylazine | $C_9H_{16}ClN_5$ | 19.77 | 23.80 | 0.735 |
| 8 | Etrimfos | $C_{10}H_{17}N_2O_4PS$ | 20.92 | 16.50 | 0.807 |
| 9 | Fosfamidon | $C_{10}H_{19}ClNO_5P$ | 21.78 | 25.09 | 0.604 |
| 10 | Heptachlor | $C_{10}H_5Cl_7$ | 22.2 | 26.02 | 0.697 |
| 11 | PCB 52 | $C_{12}H_6Cl_4$ | 23.05 | 28.83 | 1.056 |
| 12 | Aldrin | $C_{12}H_8Cl_6$ | 23.52 | 21.05 | 0.451 |
| 13 | Parathion ethyl | $C_{10}H_{14}NO_5PS$ | 24.02 | 22.75 | 0.231 |
| 14 | Penconazole | $C_{13}H_{15}Cl_2N_3$ | 25.25 | 29.31 | 0.742 |
| 15 | Hexythiazox | $C_{17}H_{21}ClN_2O_2S$ | 26.04 | 31.87 | 1.064 |
| 16 | Profenofos | $C_{11}H_{15}BrClO_3PS$ | 27.35 | 26.30 | 0.193 |
| 17 | PCB 105 | $C_{12}H_5Cl_5$ | 28.55 | 26.14 | 0.441 |
| 18 | PCB 114 | $C_{12}H_5Cl_5$ | 29.04 | 31.15 | 0.384 |
| 19 | Endosulfan sulfate | $C_9H_6Cl_6O_4S$ | 30.09 | 28.10 | 0.364 |
| 20 | Diflufenican | $C_{19}H_{11}F_5N_2O_2$ | 31.14 | 35.31 | 0.762 |
| 21 | Chrysene | $C_{18}H_{12}$ | 32.02 | 39.80 | 1.420 |
| 22 | Metoxychlor | $C_{16}H_{15}Cl_3O_2$ | 32.42 | 34.52 | 0.383 |
| 23 | Phosalone | $C_{12}H_{15}ClNO_4PS_2$ | 33.44 | 28.64 | 0.876 |
| 24 | $\lambda$ -Cyhalothrin | $C_{23}H_{19}ClF_3NO_3$ | 34.34 | 31.73 | 0.477 |

| | | | | | |
|---|---|---|---|---|---|
| 25 | Permethrin I | $C_{21}H_{20}Cl_2O_3$ | 35.65 | 32.60 | 0.557 |
| 26 | Benzo(k)fluoranthene | $C_{20}H_{12}$ | 36.65 | 38.49 | 0.337 |
| 27 | Cypermethrin III | $C_{22}H_{19}Cl_2NO_3$ | 37.79 | 37.99 | 0.036 |
| 28 | Fenvalerate II | $C_{25}H_{22}ClNO_3$ | 39.55 | 39.72 | 0.031 |
| 29 | Deltamethrin | $C_{22}H_{19}Br_2NO_3$ | 40.55 | 34.50 | 1.105 |
| 30 | BDE 183 | $C_{12}H_3OBr_7$ | 43.65 | 40.81 | 0.519 |

The whole of these data clearly displays a significant improvement of the QSRR model consequent to nonlinear statistical treatment. Obviously, there is a close agreement between the experimental and predicted RT and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. As can be seen in this section, the L-M ANN is more reproducible than GA-PLS and GA-KPLS for modeling the retention time of organic contaminants in natural water and wastewater.

**Model validation and statistical parameters**

**Model validation**

Validation is a crucial aspect of any QSPR/QSRR modeling. The accuracy of proposed models was illustrated using the evaluation techniques such as leave-group-out cross validation (LGO-CV) procedure and validation through an external test set.

3.3.2 Cross validation technique Cross validation is a popular technique used to explore the reliability of statistical models. Based on this technique, many modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input–output model is developed, based on the utilized modeling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones or group data that have not been utilized in the development of the model) [37]. The LGO procedure was utilized in this study. A QSRR model was then constructedbased on this reduced data set and subsequently used to predict the removed data. This procedure was repeated until a complete set of predicted was obtained. The statistical significance of the screened model was judged by the correlation coefficient ($Q^2$).

The accuracy of cross validation results is extensively accepted in the literature considering the $Q^2$ value. In this sense, a high value of the statistical characteristic ($Q^2 > 0.5$) is considered as proof of the high predictive ability of the model. However, this assumption is in many cases incorrect and can be that exist the lack of the correlation between the high LGO $Q^2$ and the high predictive ability of QSRR models has been established and corroborated recently [38]. Thus, the high value of LGO-CV $Q^2$ appears to be necessary but not sufficient condition for the models to have a high predictive

power. These authors stated that an external set is necessary. As a next step, further analysis was also followed for chemical property of the new set of compounds using the developed QSRR model.

**Validation through the external validation set**

Validating QSRR with external data (i.e. data not used in the model development) is the best method of validation. However, the availability of an independent external validation set of several compounds is rare in QSRR. Thus, the predictive ability of a QSRR model with the selected descriptors was further explored by dividing the full data set. The predictive power of the models developed on the selected training set is estimated on the predicted values of test set chemicals. The data set was randomly divided into training (calibration and prediction sets) and test sets after sorting based on the RT values. The data set was randomly divided into three groups including calibration and prediction sets (training set) and test set. The calibration set was used for model generation. The prediction set was applied deal with overfitting of the network, whereas test set which its molecules have no role in model building was used for the evaluation of the predictive ability of the models for external set. The calibration set consisted of 90 molecules; prediction set consisted of 30 molecules and the test set, consisted of 30 molecules. The whole of these data clearly displays a significant improvement of the QSRR model consequent to non-linear statistical treatment and a substantial independence of model prediction from the structure of the test molecule. In the above analysis, the descriptive power of a given model has been measured by its ability to predict retention of unknown molecules. For instance, as to prediction ability, it can be observed in Figure 3 that scattering of data points from the ideal trend in test set is poor.

**Statistical parameters**

For the constructed models, some general statistical parameters were selected to evaluate the predictive ability of the models for RT values. In this case, the predicted RT of each sample in prediction step was compared with the experimental acidity constant.

Root mean square error (RMSE) is a measurement of the average difference between predicted and experimental values, at the prediction step. RMSE can be interpreted as the average prediction error, expressed in the same units as the original response values. Its small value indicates that the model predicts better than chance and can be considered statistically significant. The RMSE was obtained by the following formula:

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{\wedge})^2]^{\frac{1}{2}} \quad (5)$$

The other statistical parameter was relative error (RE) that shows the predictive ability of each component, and is calculated as:

$$RE(\%) = 100 \times \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i^{\wedge} - y_i)}{y_i} \right] \quad (6)$$

The predictive ability was evaluated by the cross validation coefficient ($Q^2$ or $R^2_{cv}$) which is based on the prediction error sum of squares (PRESS) and was calculated by following equation:

$$R_{cv}^2 \equiv Q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^{\wedge})^2}{\sum_{i=1}^{n} (y_i - y^{-})^2} \quad (7)$$

Where $y_i$ is the experimental RT in the sample i, $\hat{y}_i$ represented the predicted RT in the sample i, $\bar{y}$ is the mean of experimental RT in the prediction set and n is the total number of samples used in the test set [39, 40].

**Conclusion**

Organic pollutants in water can harm the environment and pose health risks for humans. Organic pollutants pose special risks because they are often not naturally broken down and can remain in water sources for decades or longer. The analysis of organic pollutants in water allows managers to assess the quality and safety of water sources. The GA-PLS, GA-KPLS and L-M ANN modeling was applied for the prediction of the retention time of 150 organic contaminants in natural water and wastewater. High correlation coefficients and low prediction errors confirmed the good predictability of models. Application of the developed model to a validation set of 30 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. Three methods seemed to be useful, although a comparison between these methods revealed the slight superiority of the L-M ANN over the other models.

**References**

[1] Jury W.A., Vaux J. *Adv. Agron*., 2007**, 95**:1

[2] Budka M., Gabrys B., Ravagnan E. *Water Res*., 2010, **44**:3294

[3] Valsson T., Ulfarsson G.F. *Futures*., 2012, **44**:91

[4] Baresel Ch., Destouni G., Gren I. *J. Environ. Manage*., 2006, **78**:138

[5] Matilainen A., Vepsäläinen M., Sillanpää M. *Adv. Colloid Interface Sci*., 2010, **159**:189

[6] Matilainen A., Sillanpä M. *Chemosphere.* 2010, **80**:351

[7] Yu X., Wei Ch., Ke L., Wu H., Chai X., Hu Y. *J. Colloid Interface Sci*., 2011, **6**:203

[8] Hettige H., Mani M., Wheeler D. *J. Dev. Econ*., 2000, **62**:445

[9] Miller G.T. Living in the Environment: Principles, Connections, and Solutions. Pacific Grove: Brooks and Cole, 2002

[10] World Bank, World Development Indicators. Washington: Compact Disk, 2003

[11] Jorgenson A.K., Burns T.J. *Humboldt J. Soc. Relat*., 2004, **28**:7

[12] Orgenson A.K. *Soc. Sci. Quart*., 2006, **87**:711

[13] Clarke B.O., Smith S.R. *Environ. Int*., 2011, **37**:226

[14] Sánchez-Avila J., Quintana J., Ventura F., Mar R. *Poll. Bull*., 2010, **60**:103

[15] Ibáñez M., Sancho J.V., Hernández F., McMillan D., Rao R. *TrAC Trends Anal. Chem.,* 2008, **27**:481

[16] Esteve-Turrillas F.A., Yusà V., Pastor A. *Talanta.* 2008, **74**:443

[17] Serrano R., Nácher-Mestre J., Portolés T., Amat F., Hernández F. *Talanta*. 2011, **85**:877

[18] Gupta V., Khani H., Ahmadi-Roudi B., Mirakhorli Sh., Fereyduni E., Agarwal S. *Talanta.* 2011, **83**:1014

[19] Matteis C.I.D., Simpson D.A., Doughty S.W., Euerby M.R., Shaw M.P., Barrett D.A. *J. Chromatogr. A*., 2010, **1217**:6987

[20] Liu T., Nicholls I.A., Öberg T.. *Anal. Chim. Acta.*, 2011, **702**:37

[21] Riahi S., Pourbasheer E., Ganjali M.R., Norouzi P. *J. Hazard. Mater.,* 2009, **166**:853

[22] Bodzioch K., Durand A., Kaliszan R., Bączek T., Vander Heyden Y. *Talanta.* 2010, **81**:1711

[23] Leardi R. *Comper. Chemom*. 2009, **20**:631

[24] Ferrand M., Huquet B., Barbey S., Barillet F., Faucon F., Larroque H., Leray O., Trommenschlager J.M., Brochard M. *Chemom. Intell. Lab. Syst*., 2011, **106**:183

[25] Portolés T., Pitarch E., López F.J., Hernández F. *J. Chromatogr. A*., 2011, **1218**:303

[26] Devos O., Duponchel L. *Chemom. Intell. Lab. Syst*., 2011, **107**:50

[27] Hemmateenejad B., Shamsipur M., Zare-Shahabadi V., Akhond M.N. *Anal. Chim. Acta.,* 2011, **8**:13

[28] Pourbasheer E., Riahi S., Ganjali M.R., Norouzi P.. *Eur. J. Med. Chem*., 2009, **44**:5023

[29] Singh K.P., Ojha P., Malik A., Jain G.. *Chemom. Intell. Lab. Syst*., 2009, **99**:150

[30] Jančić-Stojanović B., Ivanović D., Malenović A., Medenica M. *Talanta.* 2009, **78**:107

[31] Jalali-Heravi M., Asadollahi-Baboli M., Shahbazikhah P., *Eur. J. Med. Chem*., 2008, **43**:548

[32] Xuefeng Y. *Chemom. Intell. Lab. Syst*., 2010, **103**:152

[33] Singh K.P., Basant N., Malik A., G. Jain. *Anal. Chim. Acta*., 2010, **658**:1.

[34] Noorizadeh H., Farmany A. *Drug Test Anal.*, 2013, **5**:320

[35] D'Archivio A.A., Maggi M.A., Mazzeo P., Ruggieri F. *Anal. Chim. Acta*. 2008, **628**:162

[36] Jančić B., Medenica M., Ivanović D., Janković S., Malenović A. *J. Chromatogr. A*., 2008, **1189**:366

[37] Deeb O. *Chemom. Intell. Lab. Syst*., 2010, **104**:181

[38] Kishore D.P., Balakumar C., Raghuram Rao A., Pratim Roy P., Roy K. *Bioorg. Med. Chem. Lett.,* 2011, **21**:818

[39] Hemmateenejad B., Javadnia K., Elyasi M. *Anal. Chim. Acta*., 2007, 592:72

[40] Noorizadeh H., Farmany A. *Environ Sci Pollut Res.,* 2012, **19**:1252

**How to cite this manuscript:** Mehrdad Shahpar*, Sharmin Esmaeilpoor . Advanced QSRR Modeling of Organic Pollutants in Natural Water and Wastewater in Gas Chromatography Time-of-Flight Mass Spectrometry. Chemical Methodologies 2(1), 2018, 1-22. DOI: 10.22631/chemm.2017.100307.1012.