Asian Journal of Nanoscience and Materials, 2018, 1(1), 25-35.

FULL PAPER

# Theoretical analysis of the retention behavior of pesticides and active pharmaceutical compounds in wastewater and river waters in liquid chromatography–quadrupole-time-of-flight mass spectrometry

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

[a]*Director of Ilam Petrochemical Company*

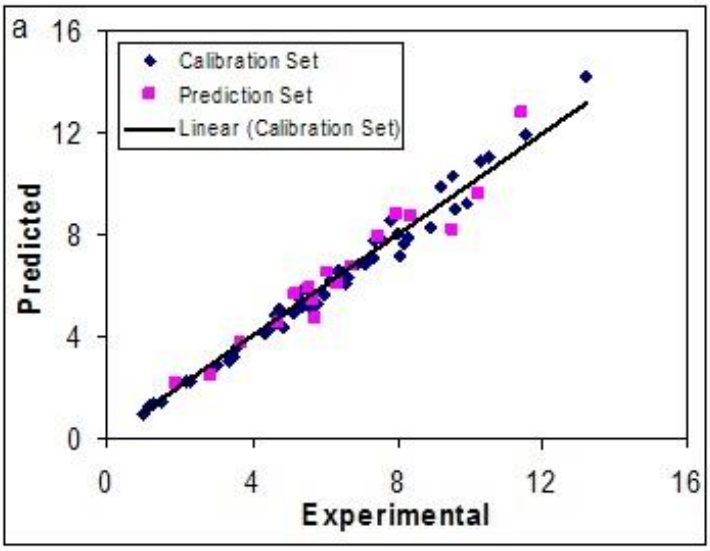
[b]*Department of Chemistry, Payame Noor University, P.O. BOX 19395-4697, Tehran, Iran.*

**Received: 29 January 2018, Revised: 07 March 2018 and Accepted: 09 March 2018.**

**ABSTRACT:** Silica supported-boron sulfonic acid (SBSA) was used as a cheap and mild bronsted acidic in the reaction of indole with aldehydes to afford the corresponding bis(indolyl)methanesin in solvent free grinding and room temperature. The catalyst is also effective in the reaction in good yields. This methodology offers several advantages, such as good yields, reusability of catalyst, short reaction times, simple procedure, and mild conditions. The catalyst can be recovered and reused without loss of activity. The work-up of the reaction consists of a simple filtration, followed by concentration of the crude product and purification.

**KEYWORDS:** Indole, Bis(indolyl)methane, Boron Sulfonic Acid, Aldehyde, Synthesis, Solvent-free, Green Chemistry.

**GRAPPHICAL ABSTACT:**




## 1 - Introduction

Water pollution is a major global problem which requires ongoing evaluation and revision of water resource policy at all levels (international down to individual aquifers and wells). It has been suggested that it is the leading worldwide cause of deaths and

*Corresponding author: Sharmin Esmaeilpoor, Email: Shahpar2012@gmail.com Tel.: +989143100801

**Asian Journal of Nanoscience and Materials**

diseases Also, it accounts for the deaths of more than 14,000 people daily. An estimated 700 million Indians have no access to a proper toilet, and 1,000 Indian children die of diarrheal sickness every day. Some 90% of China's cities suffer from some degree of water pollution, and nearly 500 million people lack access to safe drinking water [1]. The water pollution has damaged the food chain which is very important for the food preparation of plants through photosynthesis. When Filth is thrown in water the toxins travel from the water and when the animals drink that water they get contaminated and when humans tend to eat the meat of the animals is infected by toxins it causes further damage to the humans. Infectious diseases such as cholera and typhoid can be contracted from drinking contaminated water [2]. These may include organic and inorganic substances. Organic pollution (pesticides and active pharmaceutical compounds) 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. These decomposers grow rapidly and use a great deal of oxygen during their growth, leading to depletion of the oxygen as the decomposition process occurs. Lack of oxygen can kill the aquatic organisms [3]. There is currently a large amount of information from the studies conducted around the world regarding the concentrations of a wide range of organic pollutants in natural waters, wastewater, and drinking water. The water utilities should consider pro-active monitoring for any new chemicals that may be discharged into water in some form. This monitoring should, wherever possible, include metabolites and degradation by-products of these pollutants, which in some cases may have more significant consequences than the parent compounds. 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. Over 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 a huge amount of chemical information on the sample in a single analysis. TOFMS and specially hybrid quadrupole-TOFMS (QTOFMS) have been successfully applied for screening purposes in combination with liquid-chromatography (LC) in different applied fields, like environmental analysis, food safety or toxicology [4]. 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 (QSPR). Quantitative structure–retention relationships (QSRR) represent statistical models that quantify the relation between the structure of molecules and their chromatographic retention time, allowing the prediction of the RT of novel compounds [5]. QSRR on the RT have been reported for different types of compounds [6,7]. The aim of the present study is estimation of ability optimal descriptors calculated with linear regression (the partial least squares (PLS)) and non-linear regressions (the kernel partial

**Asian Journal of Nanoscience and Materials**

least squares (KPLS) and Levenberg-Marquardt artificial neural network (L-M ANN)) in QSRR analysis of retention time (RT) of pesticides and active pharmaceutical compounds in water samples of different types (surface and wastewaters). The stability and predictive power of these models were validated using Leave-Group-Out Cross-Validation (LGO CV) and external test set, techniques.

## 2 - Computational

### 2.1 - Computer hardware and software

A Pentium IV personal computer (CPU at 3.06 GHz) with the Windows XP operating system was used. The structures of the organic contaminants were drawn with HyperChem version 7.0. The all molecules were preoptimized using molecular mechanics AM1 method in the HyperChem program. Some quantum descriptor such as polarizability and orbital energy of LUMO were calculated by using the HyperChem software. The output files exported from Dragon for generating descriptors which was developed by Todeschini et al [8]. The GA-PLS, GA-KPLS, L-M ANN, cross validation, and other calculations were performed in MATLAB (Version 7.0, Math works, Inc).

### 2.2 - Data set

The group of pesticides and pharmaceutically active compounds and some of their more relevant metabolites comprise 87 organic pollutants belonging to different therapeutical groupsAll the materials purchased from Sigma–Aldrich (Steinheim, Germany), Merck (Mollet del Vallés, Spain), and LGC Promechem (Barcelona, Spain) at analytical grade (purity >95%). These data with their RT values extracted from the literature [9]. These data were obtained using liquid chromatography–electrospray quadrupole-time-of-flight mass spectrometry (LC–QTOFMS). The name of the studied compounds and their experimental RT values for data set are shown in Table 1.

### 2.3 - Determination of molecular descriptors

Molecular descriptors are defined as numerical characteristics associated with chemical structures. The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number applied to correlate physical properties. The Dragon software was used to calculate the descriptors in this research and a total of molecular descriptors, from 18 different types of theoretical descriptors, were calculated for each molecule. Since the values of many descriptors are related to the bonds length and bonds angles, the chemical structure ofthe every molecule must be optimized prior tocalculating its molecular descriptors. For this reason, the chemical structure of the 87 studied molecules were drawn with Hyperchem software and saved with the HIN extension. To optimize the geometry of these molecules, the AM1 geometrical optimization was applied. After optimizing the chemical structures of the all compounds, the molecular descriptors were calculated using Dragon. A wide variety of descriptors have been reported in the literature, having been used in QSRR analysis.

**Asian Journal of Nanoscience and Materials**

### 2.4 - Data pre-processing

Each set of the calculated descriptors was collected in a separate data matrix Di with a dimension of (m×n), where m and n are being the number of molecules and the number of descriptors, respectively. Grouping of descriptors was based on the classification achieved by Dragon software. In each group, the calculated descriptors were searched for constant or near constant values for all molecules and those detected were removed. Scaling and centering is one of the pre-processing methods we need before performing the regression methods combined with FE. The results of the projection methods depend on the normalization of the data. Descriptors with small absolute values have a small contribution to overall variances; this biases towards other descriptors with higher values. With appropriate scaling, equal weights are assigned to each descriptor, so that the important variables in the model can be focused.

**Table 1.** The compounds, retention time (min), calculated and RMSE values by GA-PLS, GA KPLS and L-M ANN models.

| No | Compounds | RT | Cal $_{GA-PLS}$ | RMSE | Cal $_{GA-KPLS}$ | RMSE | Cal $_{L-M ANN}$ | RMSE |
|---|---|---|---|---|---|---|---|---|
| | Training Set | | | | | | | |
| 1 | Trans3hy.cotinine | 1.02 | 1.05 | 0.004 | 1.02 | 0.987 | 0.97 | 0.007 |
| 2 | Nicotine | 1.19 | 1.17 | 0.002 | 1.19 | 0.993 | 1.25 | 0.009 |
| 3 | Salicylic acid | 1.29 | 1.16 | 0.016 | 1.29 | 1.103 | 1.33 | 0.006 |
| 4 | Amidotrizoate | 1.50 | 2.01 | 0.062 | 1.50 | 1.873 | 1.41 | 0.012 |
| 5p | Theobromine | 1.91 | 1.90 | 0.002 | 1.91 | 1.783 | 2.13 | 0.053 |
| 6 | Ranitidine | 2.18 | 2.70 | 0.064 | 2.18 | 3.032 | 2.22 | 0.006 |
| 7 | 4-aminoantipiryne | 2.23 | 2.28 | 0.012 | 2.18 | 2.009 | 2.23 | 0.006 |
| 8 | Amoxicillin | 2.31 | 2.40 | 0.011 | 2.31 | 2.124 | 2.24 | 0.010 |
| 9p | Theophylline | 2.84 | 3.13 | 0.036 | 2.84 | 4.249 | 2.49 | 0.085 |
| 10 | Atenolol | 3.03 | 2.73 | 0.036 | 3.03 | 3.357 | 2.84 | 0.027 |
| 11 | Lincomycin | 3.32 | 3.70 | 0.046 | 3.32 | 3.446 | 3.01 | 0.043 |
| 12 | Codeine | 3.43 | 4.06 | 0.077 | 3.43 | 3.498 | 3.20 | 0.032 |
| 13 | Paraxanthine | 3.55 | 2.87 | 0.082 | 3.55 | 3.747 | 3.58 | 0.004 |
| 14p | Salbutamol | 3.69 | 4.80 | 0.135 | 3.69 | 3.968 | 3.69 | 0.000 |
| 15 | Hydrochlorothiazide | 4.34 | 5.56 | 0.149 | 4.34 | 5.269 | 4.11 | 0.032 |
| 16 | Acetaminophen | 4.49 | 5.47 | 0.119 | 4.49 | 3.968 | 4.32 | 0.024 |
| 17 | Famotidine | 4.63 | 4.99 | 0.043 | 4.63 | 3.828 | 4.79 | 0.023 |
| 18p | Caffeine | 4.73 | 4.02 | 0.086 | 4.73 | 5.302 | 4.54 | 0.046 |
| 19 | 4-dimethylaminoantipiryne | 4.75 | 6.14 | 0.168 | 4.75 | 6.692 | 5.11 | 0.050 |
| 20 | Mepivacaine | 4.86 | 4.45 | 0.050 | 4.86 | 7.066 | 4.39 | 0.066 |
| 21 | Sulfadiazine | 5.13 | 5.47 | 0.042 | 5.13 | 5.054 | 4.94 | 0.027 |
| 22p | mecoprop | 5.16 | 4.17 | 0.120 | 5.16 | 4.707 | 5.65 | 0.118 |
| 23 | Ofloxacin | 5.22 | 5.14 | 0.009 | 5.22 | 7.365 | 5.51 | 0.041 |
| 24 | Antipyrine | 5.34 | 7.08 | 0.211 | 5.34 | 5.161 | 5.13 | 0.029 |

**Asian Journal of Nanoscience and Materials**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 25 | Tetracycline | 5.41 | 7.04 | 0.198 | 5.41 | 6.774 | 5.87 | 0.065 |
| 26p | Acetanilide | 5.57 | 5.57 | 0.000 | 5.57 | 8.226 | 5.86 | 0.070 |
| 27 | Norfloxacin | 5.59 | 7.43 | 0.223 | 5.59 | 7.074 | 5.79 | 0.028 |
| 28 | N-acetyl-4-aminoantipiryne | 5.60 | 6.09 | 0.060 | 5.60 | 7.892 | 5.12 | 0.067 |
| 29p | N-formyl-4-aminoantipiryne | 5.69 | 6.47 | 0.094 | 5.69 | 4.592 | 5.41 | 0.068 |
| 30 | Metoprolol | 5.73 | 6.70 | 0.118 | 5.73 | 5.134 | 5.24 | 0.069 |
| 31p | Sulfathiazole | 5.76 | 6.90 | 0.138 | 5.76 | 7.909 | 4.64 | 0.271 |
| 32 | Ciprofloxacin | 5.79 | 6.77 | 0.119 | 5.79 | 6.376 | 5.22 | 0.079 |
| 33 | Cefotaxime | 5.98 | 7.24 | 0.153 | 5.98 | 8.511 | 5.65 | 0.046 |
| 34p | Pentoxifylline | 6.08 | 6.73 | 0.078 | 6.08 | 7.992 | 6.51 | 0.105 |
| 35 | Omeprazole | 6.13 | 7.53 | 0.170 | 6.13 | 7.853 | 6.15 | 0.003 |
| 36 | Sulfamethazine | 6.17 | 6.84 | 0.082 | 6.17 | 6.431 | 6.12 | 0.007 |
| 37 | Venlafaxine | 6.34 | 7.38 | 0.126 | 6.34 | 5.809 | 6.60 | 0.036 |
| 38p | Propanolol | 6.36 | 8.78 | 0.293 | 6.36 | 5.781 | 6.02 | 0.082 |
| 39 | Ifosfamide | 6.38 | 8.25 | 0.227 | 6.38 | 5.958 | 6.49 | 0.015 |
| 40 | Clofibric acid | 6.47 | 6.41 | 0.007 | 6.47 | 7.089 | 6.48 | 0.002 |
| 41 | Cyclophosphamide | 6.57 | 7.57 | 0.122 | 6.57 | 9.193 | 6.01 | 0.078 |
| 42 | Pravastatin | 6.65 | 5.84 | 0.098 | 6.65 | 7.602 | 6.24 | 0.057 |
| 43p | Citalopram Hydrobr | 6.74 | 6.42 | 0.039 | 6.74 | 9.242 | 6.76 | 0.005 |
| 44 | Erythromycin | 6.94 | 8.90 | 0.237 | 6.94 | 7.790 | 6.77 | 0.023 |
| 45 | Lansoprazole | 7.15 | 8.59 | 0.174 | 7.15 | 8.009 | 6.83 | 0.045 |
| 46 | Amitriptyline | 7.34 | 5.93 | 0.171 | 7.34 | 8.188 | 7.74 | 0.056 |
| 47 | Sulfamethoxazole | 7.36 | 8.37 | 0.122 | 7.36 | 8.012 | 7.09 | 0.038 |
| 48p | Carbamazepine | 7.48 | 8.02 | 0.066 | 7.48 | 10.705 | 7.94 | 0.111 |
| 49 | Bezafibrate | 7.58 | 10.18 | 0.315 | 7.58 | 6.094 | 7.79 | 0.029 |
| 50 | Fluoxetine | 7.78 | 7.54 | 0.029 | 7.78 | 7.844 | 8.54 | 0.106 |
| 51p | Clotrimazole | 7.99 | 6.70 | 0.156 | 7.99 | 6.667 | 8.82 | 0.202 |
| 52 | Methylprednisolone | 8.02 | 8.71 | 0.084 | 8.02 | 10.308 | 8.02 | 0.001 |
| 53 | Propyphenazone | 8.06 | 11.21 | 0.381 | 8.06 | 10.116 | 7.16 | 0.126 |
| 54 | Loratadine | 8.21 | 11.11 | 0.352 | 8.21 | 12.217 | 7.64 | 0.080 |
| 55 | Ketorolac | 8.32 | 9.46 | 0.138 | 8.32 | 9.875 | 7.84 | 0.067 |
| 56p | Clomipramine | 8.37 | 9.97 | 0.194 | 8.37 | 9.040 | 8.74 | 0.089 |
| 57 | Tamoxifen | 8.94 | 11.30 | 0.286 | 8.94 | 8.396 | 8.25 | 0.097 |
| 58 | Fenoprofen | 9.22 | 7.96 | 0.152 | 9.22 | 11.411 | 9.80 | 0.081 |
| 59 | Indomethacin | 9.51 | 7.61 | 0.230 | 9.51 | 12.490 | 10.29 | 0.109 |
| 60p | Diazepan | 9.52 | 10.71 | 0.144 | 9.52 | 10.121 | 8.17 | 0.328 |
| 61 | Ketoprofen | 9.60 | 7.72 | 0.228 | 9.6 | 7.746 | 8.94 | 0.092 |

**Asian Journal of Nanoscience and Materials**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 62 | Ibuprofen | 9.910 | 10.61 | 0.085 | 9.91 | 10.841 | 9.16 | 0.105 |
| 63 | Fenofibric acid | 10.30 | 9.39 | 0.103 | 10.24 | 13.594 | 10.84 | 0.076 |
| 64p | Chlorophene | 10.24 | 13.10 | 0.340 | 10.3 | 13.964 | 9.56 | 0.164 |
| 65 | Gemfibrozil | 10.52 | 12.03 | 0.183 | 10.52 | 15.191 | 10.99 | 0.065 |
| 66p | Mevastatin | 11.43 | 9.25 | 0.264 | 11.43 | 15.903 | 12.77 | 0.326 |
| 67 | Mefenamic Ac | 11.52 | 14.68 | 0.383 | 11.52 | 12.182 | 11.87 | 0.049 |
| 68 | Fenofibrate | 13.21 | 15.09 | 0.228 | 13.21 | 16.309 | 14.16 | 0.133 |
| | Test Set | | | | | | | |
| 69 | Cotinina | 1.19 | 1.56 | 0.084 | 1.31 | 0.026 | 1.39 | 0.046 |
| 70 | 4-methylaminoantipyrine | 1.95 | 2.74 | 0.182 | 1.81 | 0.031 | 2.61 | 0.151 |
| 71 | Terbutaline | 2.77 | 2.95 | 0.042 | 2.07 | 0.160 | 1.98 | 0.181 |
| 72 | Metronidazole | 3.45 | 4.62 | 0.269 | 3.14 | 0.071 | 2.01 | 0.330 |
| 73 | Sotalol | 4.53 | 3.66 | 0.200 | 3.85 | 0.157 | 3.66 | 0.200 |
| 74 | Trimethoprim | 4.97 | 3.74 | 0.283 | 7.13 | 0.495 | 6.24 | 0.291 |
| 75 | Nadolol | 5.35 | 5.35 | 0.000 | 4.27 | 0.247 | 5.39 | 0.009 |
| 76 | Sulfapyridine | 5.60 | 5.59 | 0.002 | 3.98 | 0.372 | 6.52 | 0.211 |
| 77 | Furosemide | 5.95 | 6.47 | 0.119 | 5.92 | 0.006 | 5.03 | 0.211 |
| 78 | Azithromycin | 6.14 | 3.59 | 0.585 | 7.73 | 0.365 | 5.72 | 0.096 |
| 79 | Primidone | 6.57 | 6.40 | 0.040 | 9.27 | 0.620 | 6.19 | 0.087 |
| 80 | Paroxetine | 7.01 | 6.32 | 0.157 | 8.24 | 0.282 | 7.66 | 0.149 |
| 81 | Clarithromycin | 7.38 | 8.62 | 0.284 | 6.09 | 0.296 | 8.42 | 0.239 |
| 82 | Carbamazepine | 7.88 | 8.82 | 0.215 | 7.51 | 0.084 | 7.34 | 0.124 |
| 83 | Ketoprofen | 8.07 | 6.70 | 0.313 | 7.16 | 0.210 | 6.34 | 0.397 |
| 84 | Naproxen | 9.20 | 10.02 | 0.188 | 6.50 | 0.619 | 9.37 | 0.039 |
| 85 | Diclofenac | 9.55 | 10.47 | 0.212 | 9.72 | 0.038 | 11.42 | 0.429 |
| 86 | Indomethacin | 10.36 | 13.16 | 0.642 | 12.64 | 0.523 | 11.98 | 0.372 |
| 87 | Simvastatin | 12.51 | 10.64 | 0.429 | 9.61 | 0.666 | 15.21 | 0.619 |

**P**: Prediction Set

## 2.5 - Nonlinear model

### 2.5.1 - Artificial neural network

A three-layer back propagation artificial neural network (ANN) with a sigmoid transfer function was used to investigatethe feature sets. The descriptors from the calibration set were used for the model generation whereas the descriptors from the prediction set were used to stop the overtraining of network. The descriptors from the test set were used to verify the

**Asian Journal of Nanoscience and Materials**

predictivity of the model. Before training the networks, the input and output values were normalized with auto-scaling of the all data. 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 [10].

## 3 - Results and discussion

### 3.1 - Linear model

#### 3.1.1 - Results of the GA-PLS model

The best model is selected on the basis of the highest square correlation coefficient leave-group-out cross validation (R2), the least root mean squares error (RMSE) and relative error (RE). These parameters are probably the most popular measure of how well a model fits the data. The best GA-PLS model contains eleven selected descriptors in six latent variables space. The R2, mean RE, and RMSE for training and test sets were (0.855, 0.807), (16.76, 17.11) and (0.13, 0.22), respectively. The predicted values of RT are plotted against the experimental values for training and test sets in Fig 1a. The residuals (predicted RT− experimental RT) obtained by the GA-PLS modeling versus the experimental RT values are shown in Fig. 1b.

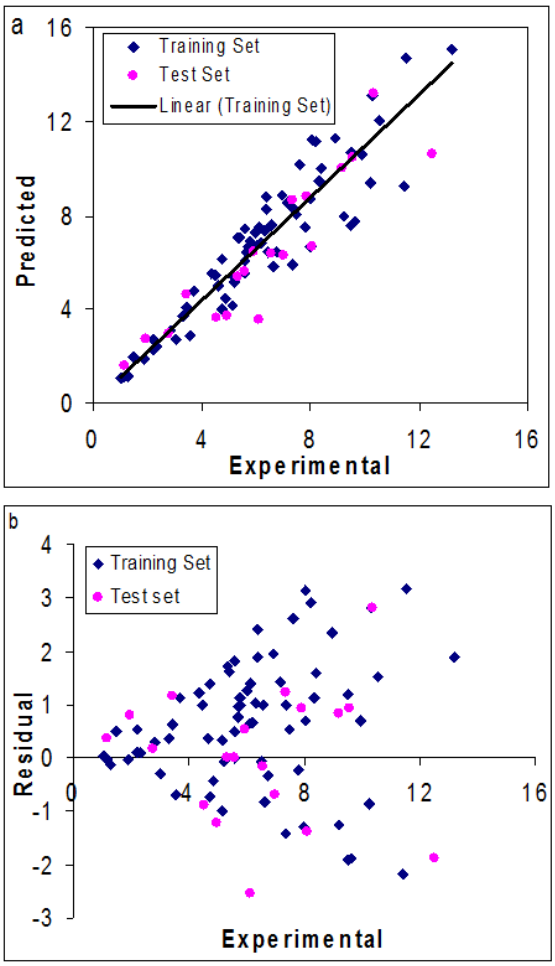



**Fig 1.** (a) Plots of predicted retention time against the experimental values and (b) the residual vs. the experimental RT by GA-PLS model

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, resultingin a lower prediction error. The values of the experimentally measured, calculated, and the RMSE are demonstratedin Table 1.

### 3.2. Nonlinear model

**Asian Journal of Nanoscience and Materials**

### 3.2.1 Results of the GA-KPLS model

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 [11]. It means that the value of c depends on the system under the study. The 9 descriptors in 5 latent variables space chosen by GA-KPLS feature selection methods were contained. The R2, mean RE, and RMSE for training and test sets were (0.848, 0.774), (20.34, 18.58) and (0.16, 0.28), respectively. the results show GA-PLS model are superior to GA-KPLS method. Fig. 2a illustrates the plot of the GA-KPLS predicted versus experimental values for RT of all of the molecules in the data set. The plots of the residuals versus the experimental RT values obtained by the GA-KPLS modeling, is demonstrated in Fig. 2b.

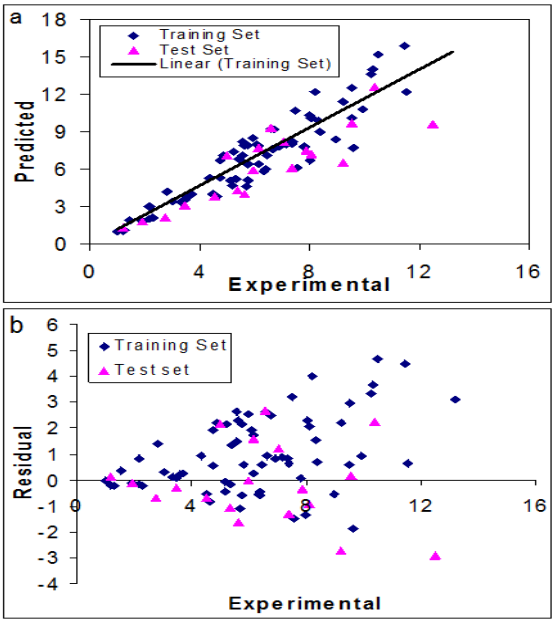



**Fig 1.** (a) Plots of predicted RT versus the experimental values and (b) the residual against the experimental RT by GA-KPLS model

### 3.2.2. Results of the L-M ANN model

To improve the predictive performance of the nonlinear QSRR model, the L-M ANN modeling was performed. For ANN generation, data set was separated into three groups: calibration and prediction (training) and test sets. All molecules were randomly placed in these sets. A three-layer network with a sigmoid transfer function was designed for each ANN. Before training the networks the input and output values were normalized between -1 and 1. The network was then trained using the training set by the back propagation strategy for optimization of the weights and bias values. The proper number of nodes in the hidden layer was determined by training the network with different number of nodes in the hidden layer. The root-mean-square error (RMSE) value measures how good the outputs are in comparison with the target values. It should be noted that for evaluating the overfitting, the training of the network for the prediction of RT must stop when the RMSE of the prediction set begins to increase while RMSE of calibration set continues to decrease. Therefore, training of the network was stopped when overtraining began. All of the above mentioned steps were carried out using basic back propagation, conjugate gradient and Levenberge-Marquardt weight update functions. It was realized that the RMSE for the training and test sets are minimum when three neurons were selected in the hidden layer. Finally, the number of the iterations was optimized with the optimum values for the variables. It was realized that after 18 iterations, the RMSE for prediction set were minimum. The RMSE, mean relative error and R2 for calibration, prediction and test sets were (0.05, 5.23, 0.978), (0.12, 7.81, 0.938) and (0.15, 12.90, 0.906), respectively. The residuals of L-M ANN predicted values of RT against the experimental values for

**Asian Journal of Nanoscience and Materials**

training and test sets are plotted in Fig.3a and Fig.3b.

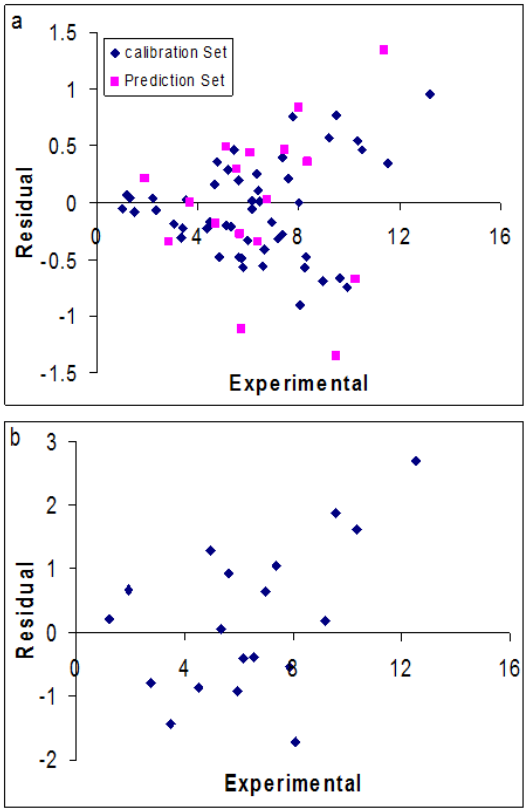



**Fig 2.** Plot of residuals obtained by L-M ANN against the experimental RT values (a) training set of molecules and (b) for test set

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. In Table 1, the predicted and RMSE values of RT obtained by the three models are presented. Comparison between these values and other statistical parameter reveals the superiority of the L-M ANN model over other model. 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. The plot of the predicted RT versus the experimental RT values by L-M ANN for training and test sets are present in Fig.4a and Fig 4b. 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 other models for modeling the RT of organic contaminants in waters.

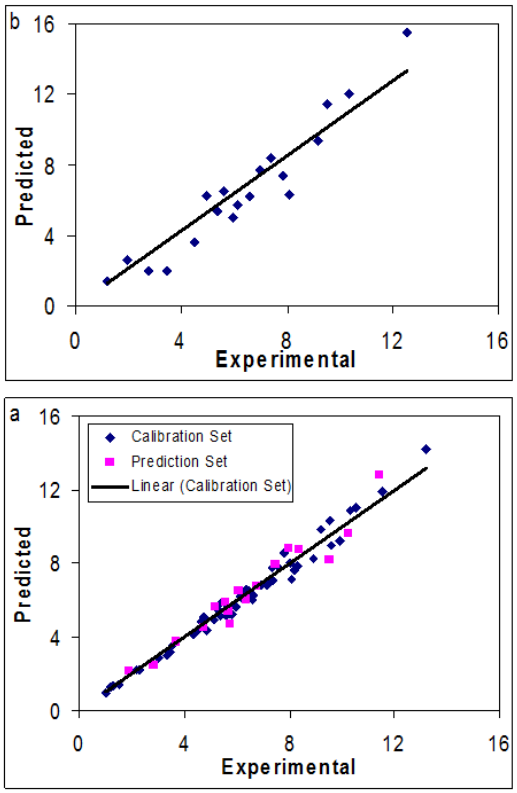



**Fig 4.** Plot of predicted RT obtained by L-M ANN against the experimental values (a) for training set and (b) test set.

### 3.3. Model validation

**Asian Journal of Nanoscience and Materials**

The accuracy of the proposed models was illustrated using the evaluation techniques such as leave group out cross-validation (LGO-CV) procedure, validation through an external test set. In addition, chance correlation procedure is a useful method to assess the accuracy of the resulted model, by which one can make sure if the results were obtained by chance or not. Cross validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of 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). In particular, the LGO-CV procedure was utilized in this study. A QSRR model was then constructed on the basis of 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 (Q2). The predictive ability was evaluated by using the cross validation coefficient (Q2 or R2cv). The accuracy of cross validation results is extensively accepted in the literature considering the Q2 value. In this sense, a high value of the statistical characteristic (Q2 > 0.5) is considered as proof of the high predictive ability of the model.The data set should be divided into three new sub-data sets, one for calibration and prediction (training), and the other one for testing. The calibration set was used for generating the model. The prediction set was applied dealing 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. In this work, in each running program, from all 87 components, 51 components are in calibration set, 17 components are in prediction set and 19 components are in test set. The result clearly displayed 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 RT of unknown drug molecules [4-7].

## 4. Conclusion

The GA-PLS, GA-KPLS, and L-M ANN models wereused to predict the RT values of large number pesticides and active pharmaceutical compounds in wastewater effluent and river water samples. High correlation coefficients and low prediction errors confirmed the good predictability of models. All methods seemed to be useful, although a comparison between these methods revealed the slight superiority of the L-M ANN over other models. Application of the developed model to a testing set of 19 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. The QSRR procedure allowed us to achieve a precise and relatively fast method for determination of RT of different series of organic micro-contaminants in wastewater and river to

**Asian Journal of Nanoscience and Materials**

predict with sufficient accuracy the RT of new substituted compounds.

**Acknowledgement**

The authors would like toacknowledge the support of the Payame Noor University (PNU) for its financial support.

**References**

1. Ayandiran TA., Fawole OO, Dahunsi SO (**2018**) Res. Ind 19:13-24
2. Wijesiri B, Deilami K, Goonetilleke A (**2018**) Environ. Pollut 234:480-486
3. You D, Min X, Liu L, Ren Zh, Xiao X, Luo J, Luo X (**2018**) J. Hazard. Mater 346:218-225
4. Zhu J, Yi X, Zhang J, Chen Sh, Wu Y (**2018**) J. Pharm. Biomed. Anal 151: 42-48
5. Noorizadeh H, EsmaeilpoorSh, Moghadam Z, NosratolahySh (**2014**) ICC 4: 283-299
6. EsmaeilpoorSh, Shirzadi Z, Noorizadeh H (**2014**) ICC 2: 56-71.
7. Shahpar M, EsmaeilpoorSh (**2017**) Asian J. Green Chem 2: 116-129
8. R. Todeschini, V. Consonni, A. Mauri, M. Pavan M, DRAGON-Software for the calculation of molecular descriptors; Version 3.0 for Windows, **2003**.
9. Gómez MJ, Gómez-Ramos MM, Malato O, Mezcua M, Férnandez-Alba AR (**2010**) J. Chromat. A 1217:7038–7054.
10. Shahpar M, Esmaeilpoor Sh (**2018**) Asian J. Nano. Mat. 1: 1-11
11. Shahpar M, EsmaeilpoorSh Chem. Method. Articles in Press

How to cite this manuscript: Mehrdad Shahpar and Sharmin Esmaeilpoor*. Theoretical analysis of the retention behavior of pesticides and active pharmaceutical compounds in wastewater and river waters in liquid chromatography–quadrupole-time-of-flight mass spectrometry. Asian Journal of Nanoscience and Materials, 2018, 1, 25-35.