ICC
Iranian Chemical Communication
Original Research Article
Payame Noor University of Ilam http://icc.journals.pnu.ac.ir

An improved structure models to explain retention behavior of atmospheric nanoparticles

Sharmin Esmaeilpoor[a], Zahra Shirzadi[b], Hadi Noorizadeh[a,*]

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

[b]*Department of chemistry, Islamic Azad University, Shahreza Branch, Isfahan, Iran*

Received: 17 November 2013 , Accepted: 29 December 2013, Published: 1 February 2014

Abstract

The quantitative structure-retention relationship (QSRR) of nanoparticles in roadside atmosphere against the comprehensive two-dimensional gas chromatography which was coupled to high-resolution time-of-flight mass spectrometry was studied. The genetic algorithm (GA) was employed to select the variables that resulted in the best-fitted models. After the variables were selected, the linear multivariate regressions [e.g. the partial least squares (PLS)] as well as the nonlinear regressions [e.g. the kernel PLS (KPLS) and Levenberg- Marquardt artificial neural network (L-M ANN)] were utilized to construct the linear and nonlinear QSRR models. The correlation coefficient cross validation (Q^2) and relative error for test set L-M ANN model are 0.939 and 4.89, respectively. The resulting data indicated that L-M ANN could be used as a powerful modeling tool for the QSPR studies.

Keywords: Atmospheric nanoparticles, QSRR, GA-KPLS, Levenberg -Marquardt artificial neural network.

Introduction

Atmospheric nanoparticles (diameter of particle: Dp <50 nm) and ultrafine particles (Dp < 100 nm) have received special attention due to their potential affect to human health [1,2]. These are ubiquitous in the troposphere and exert an important influence on the global

*Corresponding author: Hadi Noorizadeh
Fax number: +98 (841) 3382681,Tel number: + 98 9183452507
E-mail: Hadinoorizadeh@yahoo.com

climate and environment. Increasing our understanding of the physical and chemical properties of aerosols is essential in order to properly assess their effects on various issues such as human health, air quality and global climate and ultimately establishing effective control strategies. The effects of atmospheric aerosol particles on the environment and also on human health are strongly dependent on their particle size and chemical composition [3]. Carbonaceous aerosol, including elemental carbon (EC, a chemical structure similar to impure graphite) and organic carbon (OC, a large variety of organic compounds), are important components of the Atmospheric nanoparticles [4].

For a time-series study, which was done to study the influence of organic aerosol compounds, it is necessary to have data of several compounds or groups of compounds at least with a daily resolution. Because most of the organic compounds occur in low concentrations in ambient aerosol, time-consuming analytical methods are required in our analysis. Thermal desorption (TD) has been used for extracting organic compounds from atmospheric particles. Generally, Thermal desorption has been employed for extracting volatile and semi-volatile organic species from adsorbing matrices such as solid sorbent tubes [5].

Comprehensive two-dimensional gas chromatography (GC×GC) is a novel technique, whereby a sample is separated (in two dimensions) with two comprehensively coupled gas chromatographic columns. Two different chromatographic mechanisms (i.e. volatility and polarity) are used to separate the compounds in the two columns. A promising technique for analyzing the air pollution research is GC×GC coupled to fast time-of-flight mass spectrometry (TOFMS) [6-8]. Due to the increased separation of GC×GC and also our one-dimensional GC, the mass spectra are of considerably increased quality (lower background level).

The problem of skewing of mass spectra in GC–MS experiments with scanning mass analysers is also not present in time-of-flight mass spectrometry. Thus, TOFMSs provide identical mass spectral patterns over a complete chromatographic peak for the same component. The TOFMS systems can readily achieve the required spectral acquisition rates for reliable GC×GC peak assignment and quantification [9,10].

The combination of TD, GC×GC and TOF-MS allowed detection of more than 10,000 individual organic compounds in aerosol samples [11]. For proper quantification, a more limited mass range

should be selected. Exact mass measurement (mass measurement with uncertainties of a few mDa) with fast acquisition (up to 25 Hz) have become available through recent progress in the GC–TOF-MS technology. In 2005, the coupling of GC×GC to a high resolution (HR) TOF-MS with an acquisition speed of 25 Hz was reported. The HRTOF-MS can be considered a candidate MS which provides high-resolution mass information for qualitative analysis in GC×GC [12].

In quantitative structure–retention relationships (QSRR), the retention of given chromatographic system was modeled as a solute (molecular) descriptors. Computationally determined retention parameters have become crucial in identifying potential nanoparticles candidates, and this technique is used in lead and clinical candidate optimization as well as in the selection of new compounds for screening. Only one report, dealing with QSRR nanoparticles calculation has been published in the literature [13].

The QSRR models which apply to partial least squares (PLS) method would often combine with genetic algorithms (GA) for feature selection [14, 15]. Because of the complexity of relationships between the property of molecules and structures, nonlinear models are also used to model the structure–property relationships. Levenberg -Marquardt artificial neural network (L-M ANN) is nonparametric nonlinear modeling technique that has attracted increasing interest. In the recent years, nonlinear kernel-based algorithms as kernel partial least squares (KPLS) have been proposed [16,17]. The basic idea of KPLS is first to map each point in an original data space into a feature space via nonlinear mapping and then to develop a linear PLS model in the mapped space. According to Cover's theorem, nonlinear data structure in the original space is most likely to be linear after high-dimensional nonlinear mapping [18]. Therefore, KPLS can efficiently compute latent variables in the feature space by means of integral operators and nonlinear kernel functions. Compared to other nonlinear methods, the main advantage of the kernel based algorithm is that it does not involve nonlinear optimization. It essentially requires only linear algebra, making it as simple as the conventional linear PLS. In this research, GA-PLS, GA-KPLS and L-M ANN were employed to generate QSRR models that correlate the structure of nanoparticles in roadside atmosphere.

Computational

Data set

Retention time of 40 nanoparticle compounds which were taken from the literature [12] is presented in Table 1. Thermal desorption- comprehensive two-dimensional gas chromatography-high resolution time-of-flight mass spectrometry (TD–GC×GC–HRTOF-MS) is applied to the analysis of 50 nanoparticles fraction with a diameter of 29–58 nm in roadside atmosphere. Sampling of size-resolved particles was performed with a low-pressure impactor. The separation in GC×GC was performed. The data acquisition speed was 25 Hz. In the current research, retention data were collected by second column used for QSRR models. The RT of Atmospheric nanoparticles was decreased in the range of 5.08 and 1.02 for both benzo[ghi]perylene and toluene, respectively.

Descriptor calculation

All structures of compounds were drawn with the HyperChem 6.0 program. The pre optimization of all molecules were performed using MM+ molecular mechanics force field. A more precise optimization was done with the semiempirical AM1 method in HyperChem. The molecular structures were optimized using the Fletcher-Reeves algorithm until the root mean square gradient was 0.01. Moreover, the calculated values of the quantum chemical features of molecules will be influenced by the related conformation. In this study, an attempt was made to use the most stable conformations. Some quantum chemical descriptors such as dipole moment and orbital energies of LUMO and HOMO were calculated using the HyperChem program. The output files were transferred into the DRAGON 3.0 program to calculate 1497 molecular descriptors [19].

Table 1. The data set and the corresponding observed and predicted RT values by L-M ANN for the calibration, prediction and test sets.

No	Name	RT_{Exp}	RT_{ANN}	RE (%)
	Calibration Set			
1	Ethyl benzene	1.02	0.99	2.94
2	Styrene	1.17	1.21	3.42
3	Benzofuran	1.38	1.45	5.07
4	Furfural	1.54	1.60	3.90
5	Benzaldehyde	1.58	1.61	1.90
6	Nicotine	1.66	1.67	0.60
7	Naphtho[2,1-b]furan	1.83	1.78	2.73

8	Quinoline	1.90	2.02	6.32
9	Nicotyrine	1.95	1.95	0.00
10	Benzophenone	1.99	2.16	8.54
11	Phthalic anhydride	2.03	2.08	2.46
12	Anthrone	2.07	2.05	0.97
13	Phenanthrened	2.11	2.18	3.32
14	9H-Fluorene-9-one	2.15	2.12	1.40
15	Fluoranthened	2.35	2.22	5.53
16	Pyrene, 2-methyl	2.40	2.24	6.67
17	9,10-Anthracenedione	2.48	2.51	1.21
18	Benzo[a]anthracened	2.68	2.75	2.61
19	Naphtho[1,2-c]furan-1,3-dione	2.72	2.85	4.78
20	Cyclopenta[cd]pyrene	2.80	2.83	1.07
21	7H-Benzo[de]anthracen-7-one	2.96	2.94	0.68
22	Perylened	3.25	3.18	2.15
23	Indeno[1,2,3-cd]pyrened	4.43	4.13	6.77
24	Benzo[ghi]perylened	5.08	4.77	6.10
	Prediction Set			
25	Benzofuran, 2-methyl	1.50	1.48	1.33
26	Benzonitrile, 2-methyl	1.87	1.97	5.35
27	Benzothiazole	1.91	1.93	1.05
28	Indandione	1.98	2.06	4.04
29	Anthracened	2.19	2.14	2.28
30	Pyrene, 1-methyl	2.44	2.60	6.56
31	Chrysene, 1-methyl	2.64	2.63	0.38
32	Benzo[a]pyrened	3.13	3.01	3.83
	Test Set			
33	Toluene	1.02	1.01	0.98
34	Phenol, 4-methyl	1.54	1.63	5.84
35	2(5H)Furanone, 3-methyl	1.83	1.87	2.19
36	Indanone	1.91	2.05	7.33
37	2,5-Furandicarboxaldehyde	1.99	1.94	2.51
38	Isoquinoline	2.11	2.25	6.64
39	Cyclopenta[def]phenanthrenone	2.48	2.27	8.47
40	Chrysened	2.68	2.82	5.22

Genetic algorithm

A detailed description of the genetic algorithm (GA) can be found in the literature [20-22]. Genetic algorithm would be our simulated methods based on ideas from Darwin's theory of natural selection and evolution (the struggle for life). In GA, a chromosome (or an individual) which can be defined as an enciphered entity of a candidate solution, is expressed in a set of variables. GA consist of the following basic steps: (1) A chromosome is represented by a binary bit string and then an initial population of chromosomes is created in a random way; (2) A value for the fitness function of each chromosome is evaluated; (3) Based on the values of the fitness functions, the chromosomes of the next generation are produced by selection, crossover and mutation operations. The fitness function was proposed by Depczynski *et al.* [23].

Linear model

Partial least squares

PLS is a linear multivariate method for relating the process variables X with responses Y. PLS can analyze data with strongly collinear, noisy, and numerous variables in both X and Y [24]. PLS reduces the dimension of the predictor variables by extracting factors or latent variables that are correlated with Y while capturing a large amount of the variations in X. This means that PLS maximizes the covariance between matrices X and Y. In PLS, the scaled matrices X and Y are decomposed into score vectors (t and u), loading vectors (p and q), and residual error matrices (E and F):

$$X = \sum_{i=1}^{a} t_i p_i^T + E$$
$$Y = \sum_{i=1}^{a} u_i q_i^T + F \qquad (1)$$

Where a is the number of latent variables, in an inner relation, the score vector t is linearly regressed against the score vector u.

$$U_i = b_i t_i + h_i \qquad (2)$$

Where b is regression coefficient, that is determined by minimizing the residual h, It is crucial to determine the optimal number of latent variables and cross validation which is a practical and reliable way to test the predictive significance of each PLS component. There are several algorithms to calculate the PLS model parameters. In this work, the NIPALS algorithm was used with the exchange of scores [25].

Nonlinear models

Kernel partial least squares

The KPLS method is based on the mapping of the original input data into a high dimensional feature space $\Im$ where a linear PLS model is created. By nonlinear mapping $\Phi : x \in \Re^n \rightarrow \Phi(x) \in \Im$, a KPLS algorithm can be derived from a sequence of NIPALS steps and has the following formulation [26]:

1. Initialize score vector w as equal to any column of Y.
2. Calculate scores $u = \Phi\Phi^T w$ and normalize u to ||u|| = 1, where Φ is a matrix of regressors.
3. Regress columns of Y on u: $c = Y^T u$, where c is a weight vector.
4. Calculate a new score vector w for Y: w = Yc and then normalize w to ||w||=1.
5. Repeat steps 2–4 until convergence of w.
6. Deflate $\Phi\Phi^T$ and Y matrices:

$$\Phi\Phi^T = (\Phi - uu^T\Phi)(\Phi - uu^T\Phi)^T \qquad (3)$$

$$Y = Y - uu^T Y \qquad (4)$$

7. Go to step 1 to calculate the next latent variable.

Without explicitly mapping into the high-dimensional feature space, a kernel function can be used to compute the dot products as follows:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \qquad (5)$$

$\Phi\Phi^T$ represents the (n×n) kernel Gram matrix K of the cross dot products between all mapped input data points $\Phi(x_i), i = 1,...,n$. The deflation of the $\Phi\Phi^T = K$ matrix after extraction of the u components is given by:

$$K = (I - uu^T)K(I - uu^T) \qquad (6)$$

Where "I" is an m-dimensional identity matrix, taking into account the normalized scores "u" of the prediction of KPLS model on training data, $\hat{Y}$ is defined as:

$$\hat{Y} = KW(U^T KW)^{-1} U^T Y = UU^T Y \qquad (7)$$

For predictions on new observation data $\hat{Y}_t$, the regression can be written as:

$$\hat{Y}_t = K_t W(U^T KW)^{-1} U^T Y \qquad (8)$$

Where K_t is the test matrix whose elements are $K_{ij} = K(x_i, x_j)$, x_i and x_j present the test and training data points, respectively.

Artificial neural network

An artificial neural network (ANN) with a layered structure is a mathematical system that stimulates the biological neural network which consists of computing units named neurons and connections between neurons named

synapses [27,29]. Input or independent variables are considered as neurons of input layer, while dependent or output variables are considered as output neurons. Synapses connect input neurons to hidden neurons and hidden neurons to output neurons. The strength of the synapse from neuron i to neuron j is determined by the means of a weight, W_{ij}. In addition, each neuron j from the hidden layer, and eventually the output neuron, are associated with a real value b_j, named the neuron's bias and with a nonlinear function, named the transfer or activation function. Because the artificial neural networks (ANNs) are not restricted to linear correlations, they can be used for nonlinear phenomena or curved manifolds [27]. Back propagation neural networks (BNNs) are most often used in analytical applications [28]. The back propagation network receives a set of inputs, which is multiplied by each node and then a nonlinear transfer function is applied. The goal of training the network is to change the weight between the layers in a direction to minimize the output errors. The changes in values of weights can be obtained using Eq. (9):

$$\Delta W_{ij,n} = F_n + \alpha \Delta W_{ij,n-1} \qquad (9)$$

Where ΔW_{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. There is no single best weight update function which can be applied to all nonlinear optimizations. One needs to choose a weight update function based on the characteristics of the problem and the data set of interest. Various types of algorithms have been found to be effective for most practical purposes such as Levenberg-Marquardt (L-M) algorithm.

Levenberg -Marquardt algorithm

While basic back propagation is the steepest descent algorithm, the Levenberg-Marquardt algorithm [30] is an alternative to the conjugate methods for second derivative optimization. In this algorithm, the update function, F_n, can be calculated using Eqs. (10) and (11):

$$F_0 = -g_0 \qquad (10)$$

$$F_n = -[J^T \times J + \mu I]^{-1} \times J^T \times e \qquad (11)$$

Where J is the Jacobian matrix, μ is a constant, I is an identity matrix, and e is an error function [31].

Software and programs

A Pentium IV personal computer (CPU at 3.06 GHz) with windows XP operational system was used. Geometry optimization was

performed by HyperChem (Version 7.0 Hypercube, Inc.); Dragon software was used to calculate the descriptors. MINITAB software (version 14, MINITAB) was used for the simple PLS analysis. Cross validation, GA-PLS, GA-KPLS, L-M ANN and other calculation were performed in the MATLAB (Version 7, Mathworks, Inc.) environment.

Results and discussion

Linear model

Results of the GA-PLS model

To reduce the original pool of descriptors to an appropriate size, the objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with other descriptors present in the pool. After this process, 1091 descriptors were remained. These descriptors were employed to generate the models with the GA-PLS and GA-KPLS program. The best model is selected on the basis of the highest multiple correlation coefficient leave-group-out cross validation (LGO-CV) (Q^2), the least root mean squares error (RMSE) and relative error (RE) of prediction and simplicity of the model. These parameters are probably the most popular measures of how well a model fits the data. The best GA-PLS model contains 18 selected descriptors in 11 latent variables space. For this, in general, the number of components (latent variables) is less than the number of independent variables in PLS analysis. The Q^2, mean RE and RMSE for training and test sets were (0.85, 0.69), (6.73, 14.07) and (0.27, 0.44), respectively. 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.

Nonlinear models

Results of the GA-KPLS model

The leave-group-out cross validation (LGO-CV) has been performed. 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 set to be 1), m is the dimension of the input space and σ^2 is the variance of the data [32]. 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 Q^2, mean RE and RMSE for training and test sets were (0.88, 0.79), (3.89, 8.64) and (0.19, 0.31), respectively.

The RMSE values of the GA-KPLS model for the training and test sets were much lower than GA-PLS model. From these results, it can be noticed that the GA-KPLS model gives the highest Q^2 values, so this model provides the most satisfactory results, compared with the results obtained from the GA-PLS model. The GA-PLS linear model has good statistical quality with low prediction error, while the corresponding errors obtained by the GA-KPLS model are lower. Consequently, the GA-KPLS approach currently constitutes the most accurate method for predicting the retention of these components than that of the GA-PLS method. This suggests that GA-KPLS hold promise for applications in choosing variables for L-M ANN systems. This result indicates that the RT of nanoparticle molecules possesses some nonlinear characteristics.

Results of the L-M ANN model

With the aim of improving the predictive performance of nonlinear QSRR model, L-M ANN modeling was performed. Descriptors of GA-KPLS model were selected as inputs in L-M ANN model. The network architecture consisted of nine neurons in the input layer corresponding to the five mentioned descriptors. The output layer had one neuron that predicts the RT. The number of neurons in the hidden layer is unknown and needs to be optimized. In addition to the number of neurons in the hidden layer, the learning rate, the momentum and the number of iterations also should be optimized. In this work, the number of neurons in the hidden layer and other parameters except the number of iterations were simultaneously optimized. A MATLAB program was written to change the number of neurons in the hidden layer from 2 to 7, the learning rate from 0.001 to 0.1 with a step of 0.001 and the momentum from 0.1 to 0.99 with a step of 0.01. The root mean square errors for training set was calculated for all of the possible combination of values for the mentioned variables in leave-group-out cross validation (LGO-CV). It was realized that the RMSE for the training set is minimum when two neurons were selected in the hidden layer and the learning rate and the momentum values were 0.6 and 0.4, respectively. Finally, the number of 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 values of experimental, calculated and percent relative error are shown in Table 1. The Q^2, RE and RMSE for calibration, prediction and test sets were (0.98, 0.97, 0.93), (3.24, 3.13, 4.89) and (0.10, 0.89, 0.12), respectively. For the constructed model, three general statistical

parameters were selected to evaluate the prediction ability of the model for the RT. The statistical parameters Q^2, RE and RMSE were obtained for proposed models. Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSRR model. Inspection of the results reveals a higher Q^2 and other parameter values for the training and test sets compared with their counterparts for GA-KPLS and GA-PLS. Plots of predicted RT versus experimental RT values by L-M ANN are shown in Fig. 1a, 1b. Obviously, there is a close agreement between the experimental and predicted RT, moreover, the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. This clearly shows the strength of L-M ANN as a nonlinear feature selection method. The key strength of L-M ANN is their ability to allow the flexible mapping of the selected features by manipulating their functional dependence implicitly. Neural network handles both linear and nonlinear relationship without adding complexity to the model. This capacity offsets the large computing time required and complexity of L-M ANN model with respect to other models.

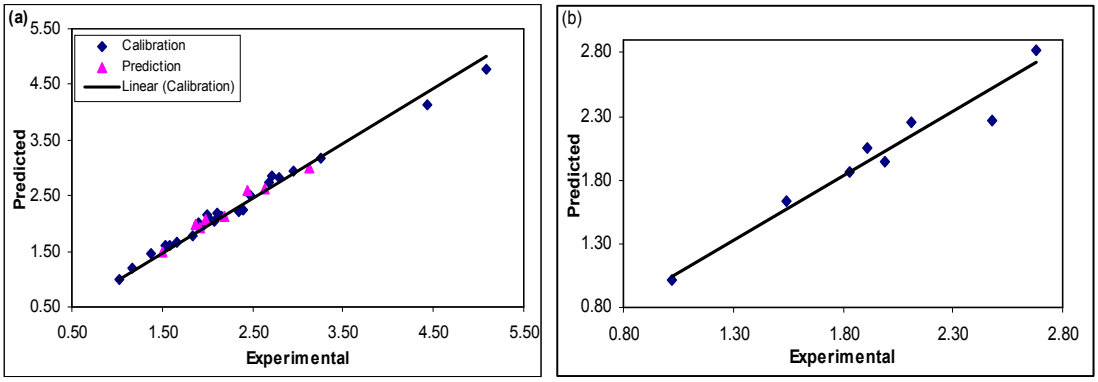

Figure 1. Plot of predicted log K_s obtained by L-M ANN against the experimental values (a) training set of molecules and (b) for test set

Model validation

Validation is a crucial aspect of any QSPR/QSRR modeling [33]. 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.

Cross validation technique

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) [34]. In particular, the LGO 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 prediction was obtained. The statistical significance of the screened model was judged by the correlation coefficient (Q^2). 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:

$$R^2_{cv} \equiv Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (12)$$

Where y_i, $\hat{y_i}$ and $\bar{y}$ were respectively the experimental, predicted, and mean RT values of the samples,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 [35]. Although this assumption is in many cases incorrect, it is worth mentioning that the lack of the correlation between the high Q^2 and the high predictive ability of QSPR/QSRR models has been established and corroborated recently [33]. Thus, the high value of 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. In our 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 test 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 test 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 which was developed on the selected training set is estimated on the predicted values of test set chemicals. The data set was randomly divided into three groups including calibration and prediction sets (training set) and test set, which consists of 24, 8 and 8 molecules, respectively. The calibration set was used for model generation. The applied prediction set deals with overfitting of the network, whereas test set in 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 result 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 partition of unknown drugs. For instance, as it was done for the prediction ability, it can be observed in Fig. 1 that scattering of data which points from the ideal trend in test set is poor.

Conclusion

In the present study, a linear method (GA-PLS) and two nonlinear methods (GA-KPLS and L-M ANN) were used to construct a quantitative relation between the retention of nanoparticles in roadside atmosphere and their calculated descriptors. The most important selected molecular descriptors represent the molecular properties, constitutional and quantum chemical descriptors that are known to be important in the retention mechanism of atmospheric molecules. The results obtained by L-M ANN were compared with the results obtained by other models. The results demonstrated that L-M ANN was more powerful in the retention prediction of these nanoparticle compounds than GA-PLS and GA-KPLS. A suitable model with high statistical quality and low prediction errors was eventually derived. It was easy to notice that there was a good prospect for the L-M ANN application in the QSRR modeling.

References

[1] D.M. Brown, M.R.Wilson, W. MacNee, V. Stone, K. Donaldson, *Toxicol. Appl. Pharm*., **2001**, *175*, 191-199.

[2] K. Inoue, H. Takano, R. Yanagisawa, M. Sakurai, T. Ichinose, K. Sadakane, T. Yoshikawa, *Respir. Res*., **2005**, *6*, 106-113.

[3] G. Buzorius, A. Zelenyuk, F. Brechtel, D. Imre, *Geophys Res Let.*, **2002**, *29*, 1974–1978.

[4] Zh. Chong-Shu, Ch-Ch. Chen, J-J. Cao, Ch-J. Tsai, Ch.C.-K. Chou, Sh-Ch. Liu, *Atmos. Environ*, **2010**, *44*, 2668-2673.

[5] S.S. Hang Ho, J. Zh. Yu, *J. Chromatogr. A*, **2004**, *1059*, 121–129.

[6] J. Dallüge, M. van Rijn, J. Beens, R.J.J. Vreuls, U.A.Th.Brinkman, *J. Chromatogr. A*, **2002**, *965*, 207-217.

[7] F. Adam, F. Bertoncini, N. Brodusch, E. Durand, D. Thiebaut, D. Espinat, M-C. Hennion, *J. Chromatogr. A*, **2007**, *1148*, 55–64.

[8] T. Hyölyläinen, M. Kallio, M. Shimmo, K. Saamio, K. Hartonen, M.L. Riekkola, *in: Presentation at the First International Symposium on Two-Dimensional Gas Chromatography*, Volendam, The Netherlands, **2003**.

[9] C. Muhlen, C. Alcaraz Zini, E.B. Caramao, P.J. Marriott, *J. Chromatogr. A*, **2008**, *1200*, 34–42.

[10] M. Adahchour, M. Brandt, H.U. Baier, R.J.J. Vreuls, A.M. Batenburg, U.A.Th. Brinkman, *J. Chromatogr. A*, **2004**, *1054*, 57-65.

[11] J.F. Hamilton, P.J. Webb, A.C. Lewis, J.R. Hopkins, S. Smith, P. Davy, *Atmos. Chem. Phys*, **2004,** *4*, 1279-1290.

[12] N. Ochiai, T. Ieda, K. Sasamoto, A. Fushimi, Sh. Hasegawa, K. Tanabe, Sh. Kobayashi, *J. Chromatogr. A*, **2007**, *1150*, 13–20.

[13] J. Leban, M. Baierl, J. Mies, V. Trentinaglia, S. Rath, K. Kronthaler, K. Wolf, A. Gotschlich, M.H.J. Seifert, *J. Bioorg. Med. Chem. Lett*., **2007**, *17*, 5858–5862.

[14] K. Bodzioch, A. Durand, R. Kaliszan, T. Bączek, Y. Vander Heyden, Talanta, **2010**, *81*, 1711-1718

[15] S. Riahi, E. Pourbasheer, M.R. Ganjali, P. Norouzi, *J.Hazard.Mater.,* **2009**, *166*, 853-859.

[16] S.H. Woo, J. ChO, Y.S. Yun, H. Choi, Ch.S. Lee, D.S. Lee, *J. Hazard. Mater*, **2009**, *161*, 538–544

[17] N. Krämer, A.L. Boulesteix, G. Tutz, *Chemom. Intell. Lab. Syst.,* **2008**, *94*, 60–69.

[18] S. Haykin, *Neural Networks*, Prentice-Hall, New Jersey, **1999**.

[19] R. Todeschini, V.Consonni, A.Mauri, M.Pavan, *DRAGON-Software for the calculation of molecular descriptors*. Version 3.0 for Windows, **2003**.

[20] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley–Longman, Reading, MA, USA, **2000**.

[21] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali, P. Norouzi, Exploring QSARs for antiviral activity of 4-alkylamino-6-(2-hydroxyethyl)-2- methylthiopyrimidines by support vector machine, *Chem. Biol. Drug Des*, **2008**, *72*, 205-216.

[22] J.A.d. Sousa, M.C. Hemmer, J. Casteiger, Prediction of H-1 NMR chemical shifts using neural networks, *Anal. Chem*, **2002**, *74*, 80-93.

[23] U. Depczynski, V.J. Frost, K. Molt, Genetic algorithms applied to the selection of factors in principal component regression, *Anal. Chim. Acta*, **2000**, *420*, 217-227.

[24] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst*, **2001**, *58*, 109-130.

[25] B.M. Nicolai, K.I. Theron, J. Lammertyn, *Chemom. Intell. Lab. Syst*., **2007**, *85*, 243–252

[26] R. Rosipal, L.J. Trejo, *J. Mach. Learning Res*, **2001**, *2*, 97-123.

[27] J. Zupan, J. Gasteiger, *Neural Network in Chemistry and Drug Design*, Wiley–VCH, Weinheim, **1999**.

[28] S. Kara, A.S. Güven, M. Okandan, F. Dirgenali, *Comput. Biol. Med*, **2006**, *36*, 473–483

[29] A. Yasri, D. Hartsough, *J. Chem. Inf. Comput. Sci*, **2001**, *41*, 1218-1227

[30] S. Kara, M. Okandan, *Pattern Recognit*, **2007,** *40*, 2967 – 2973

[31] M. Salvi, D. Dazzi, I. Pelistri, F. Neri, J.R. Wall, *Ophthalmology*, **2002**, *109*, 1703-1708.

[32] K.Kim, J.M.Lee, I.B. Lee, *Chemom. Intell. Lab. Syst*, **2005**, *79*, 22-30.

[33] J. Acevedo-Martınez, J.C. Escalona-Arranz, A. Villar-Rojas, F. Tellez-Palmero, R. Perez-Roses, L. Gonzalez, R. Carrasco-Velar, *J. Chromatogr. A*, **2006**, *1102*, 238-244.

[34] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-

Markopoulou, *Bioorg. Med. Chem*, **2006**, *14*, 6686-6694.

[35] A. Golbraikh, A. Tropsha, *J. Mol. Graphics Modell*, **2002**, *20*, 269-276.

[36] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany, **2000**.