*Asian Journal of Chemistry* Vol. 21, No. 4 (2009), 2521-2532

# Prediction of Solute Descriptors in LSER Equation Using Quantitative Structure-Property Relationship Methodology

M.H. FATEMI* and M.A. GHASEMI
*Department of Chemistry, Mazandaran University, Babolsar, Iran*
*Tel/Fax: (98)(112)5242002; E-mail: mhfatemi@umz.ac.ir*

In this study, a quantitative structure-property relationship method based on multiple linear regressions (MLR) and artificial neural network (ANN) techniques were applied for the calculation/prediction of $\Sigma\beta_2^H$ and $\pi_2^H$ parameters of the linear solvation energy relationship (LSER). The selected descriptors that appear in multiple linear regression models for $\Sigma\beta_2^H$ are: maximal electrotopological positive variation, average connectivity index chi-5, Geary autocorrelation-lag1/weighted by atomic polarizabilities, radial distribution function-2/unweighted and leverage-weighted autocorrelation-lag 4/unweighted. Also descriptors that appear in MLR model for $\pi_2^H$ are: Geary autocorrelation-lag2/weighted by atomic Sanderson electronegativites, 2nd component accessibility directional WHIM index/weighted by atomic vander Waals volumes, d COMMA-2 value/weighted by atomic Sanderson electronegativites, number of H attached to $C_1(sp^3)/C_0(sp^2)$ and dipole moment. These descriptors were used as inputs for two ANNs. After training and optimization of these ANNs, they were used to prediction of $\pi_2^H$ and $\Sigma\beta_2^H$ values of the test set compounds, separately. Analysis of the results obtained indicates that the models we proposed can correctly represent the relationship between these LSER solute parameters and theoretically calculated molecular descriptors. Also results showed the superiority of neural networks over regression models.

**Key Words: Neural network, Quantitative structure-property relationship, Linear solvation energy relationship, Multiple linear regression, Molecular descriptors.**

## INTRODUCTION

Important progress has been made over the last years in understanding the relationships between various properties of organic compounds and their chemical structures. Numerous predictive models were developed that aim to predict mixture thermodynamic properties from parameters that quantify the structure of the molecules. Such investigations are called quantitative structure-property relationship (QSPR) studies. The key to the QSPR methodology is the accurate characterization of structural features (molecular descriptors) that related to the observed property. The advantage of QSPR study as compared with other methods is that no experimental parameters are required. In view of the fact that the most chemicals have very little

testing data, it would be desirable if one could develop QSPR model only from molecular descriptors that can be calculated directly from the chemical structure. Among one of the most significant achievements of QSPRs is the linear solvation energy relationship (LSER) of Kamlet *et al.*[1]. The LSER model was first developed by Kamlet and Taft to describe solvation effects on physico-chemical processes[2-5]. The descriptors in this model were later adapted to describe solute characteristic in order to investigate the solubility properties in various media[6,7]. The LSER approach has been applied extensively to the study of retention in gas chromatography[8-13], reverse phase liquid chromatography (RPLC)[14-20] and to some extent in normal-phase liquid chromatography[21-24]. Based on this model, a free energy related term in a phase transfer process could be correlated with various fundamental molecular solute descriptors properties. The LSER model proposed by Abraham and coworker[25-27] to express a solute property (SP) as follows:

$$\log SP = c + rR_2 + mV_2 + s\pi_2^H + a\Sigma\alpha_2^H + b\Sigma\beta_2^H \quad (1)$$

In this equation, the subscript 2 denotes the solute descriptors, which include excess molar refraction ($R_2$), McGawans molecular volume ($V_2$), dipolarity/polarizability ($\pi_2^H$), overall hydrogen bond acidity ($\Sigma\alpha_2^H$) and overall hydrogen bond basicity ($\Sigma\beta_2^H$). These descriptors represent the ability of solute to participate in various solute-phase interactions. Each solute descriptor is multiplied by a coefficient (c, r, s, a, b, l) that represents the system response to these interactions. Detailed descriptions of the parameters in eqn. 1 have been extensively presented in the literature[25-28]. The descriptors $R_2$ and $V_2$ in eqn. 1 are easily calculated from structure, but traditionally the polarity and hydrogen bonding descriptors had to be determined experimentally. This could be either directly from complexation measurements or indirectly *via* back calculations from partition measurements, which can be difficult and time consuming. For example, some of solute descriptors obtained from McReynolds gas chromatographic retention data by Ballantine and Callihan[29]. The values of $\pi_2^H$, $\Sigma\beta_2^H$ and $\Sigma\alpha_2^H$ parameters obtained from GC measurements by Abraham *et al.*[30]. Also, Roglaski and Mutelet[31] have used temperature gradient gas chromatography to determine/predict LSER parameters of highly boiling organic compounds. Whereas the experimental determinations of these parameters are expensive, time consuming and needs to pure organic compound, therefore development of a theoretical based method for calculations/predictions of these parameters are necessary. Recently, a theoretical method has been developed for the estimation of these parameters ($\pi_2^H$, $\Sigma\beta_2^H$ and $\Sigma\alpha_2^H$) based on fragmental contributions[32]. The generality of this method is limited by the lack of experimental data for important fragments. The fragmental contributions to descriptors were taken from an experimental data base of descriptors and clearly, if a given fragment is not present in the data base then no values can be assigned. Also the calculations of the $\pi_2^H$, $\Sigma\beta_2^H$ and $\Sigma\alpha_2^H$ parameters separately obtained from other types of fragmentation method for some solute by Platts *et al.*[33-35]. But these manual fragmentation approaches is slow, time consuming and limiting the use of eqn. 1 for large datasets. Svozil *et al.*[36] reported

an artificial neural network (ANN) approach to estimating $\pi_2^H$. They took a number of topological and quantum mechanical properties as input, combining them non-linearly *via* a feed-forward ANN. In order to get an acceptable model, they restricted the diversity of compounds to benzene and phenol types. The calculations performed using the ANN method resulted in $R^2 = 0.979$ for the training set (62 compounds) and $R^2 = 0.932$ for the test set (after removal of statistical outliers) for a model with 16 descriptors. They reported that if they use smaller ANN model with 7 descriptors as input, their correlation coefficients of obtained model were 0.908 and 0.537 for training and testing set, respectively. In this paper, the calculation/prediction of $\pi_2^H$ and $\Sigma\beta_2^H$ parameters from the theoretical calculated parameters based on QSPR methods and using artificial neural network is discussed.

## EXPERIMENTAL

**Data set:** The data set of solute descriptors in LSER was taken from the values reported by Abraham *et al.*[30]. The molecules in the data set including aliphatic and aromatic compounds are shown in Table-1. The $\pi_2^H$ values fall in the range of 0.21 to 1.8 for *n*-butyl ether and *n*-benzyl formamide, respectively and also for the $\Sigma\beta_2^H$ fall in the range of 0.02 to 0.80 for chloroform and N,N-dibutyl formamide, respectively. Data set were split into 3 separation section; the training; test and external validation sets, consist of 60, 13 and 13 members, respectively. The training set was used to adjust the parameters of model, the test set was used to prevent the network from over-fitting and external validation set was used to evaluate the prediction ability of constructed models.

**Descriptors:** The parameters of $\pi_2^H$ represent the ability of solute molecules to interact with solvent by electronic interactions. The value of these parameters depends on the polarizability, dipolarity, size of molecule and the strength of interaction between the solute and solvent molecules. Also the value of $\Sigma\beta_2^H$ depends on the ability for the formation of hydrogen bonds, polarity and the strength of interaction between the solute and solvent molecules. Some of the molecular descriptors were used to search for the best models of the solute descriptors ($\pi_2^H$ and $\Sigma\beta_2^H$) were calculated by the Dragon package[37] on the basis of the minimum energy molecular geometries optimized by the Hyperchem 5.02 (Hypercube, 1997 ). Dragon is new, freely available software (by Milano Chemometrics and the QSAR Research Group) for the calculation of more than 800 molecular descriptors. Also some electronic descriptors were calculated using the MOPAC program (version 6)[38]. Subsequently, the method of stepwise multiple linear regression (MLR) was used to select the most important descriptors and to calculate the coefficients relating the descriptors to solute parameters ($\pi_2^H$ and $\Sigma\beta_2^H$). The selected descriptors that appear in MLR model for $\pi_2^H$ are shown in Table-2. These descriptors are: Geary autocorrelation -lag2/weighted by atomic Sanderson electronegativites (GATS2E), 2nd component accessibility directional WHIM (weighted holistic invariant molecular descriptors) index/weighted by atomic vander Waals volumes (E2V), d COMMA (comparative

TABLE-1
DATA SET AND CORRESPONDING OBSERVED AND PREDICTED VALUES OF THE $\pi_2^H$ AND $\Sigma\beta_2^H$ FOR TRAINING, TEST AND VALIDATION SET[a]

| No. | Name | $\Sigma\beta_2^H$ | | | $\pi_2^H$ | | |
|---|---|---|---|---|---|---|---|
| | | MLR | ANN | EXP | MLR | ANN | EXP |
| **Training set** | | | | | | | |
| 1 | 1-Butanol | 0.54 | 0.48 | 0.48 | 0.49 | 0.43 | 0.42 |
| 2 | 2-Propanol | 0.53 | 0.55 | 0.56 | 0.39 | 0.34 | 0.36 |
| 3 | Cyclohexanol | 0.56 | 0.58 | 0.57 | 0.55 | 0.53 | 0.54 |
| 4 | 1-Butanal | 0.50 | 0.41 | 0.45 | 0.51 | 0.64 | 0.65 |
| 5 | 1-Hexanal | 0.39 | 0.37 | 0.45 | 0.55 | 0.64 | 0.63 |
| 6 | N,N-Dimethyl formamide | 0.76 | 0.74 | 0.74 | 1.13 | 1.30 | 1.31 |
| 7 | N,N-Dibutyl formamide | 0.71 | 0.80 | 0.80 | 1.16 | 1.19 | 1.19 |
| 8 | N,N-Dimethyl acetamide | 0.61 | 0.78 | 0.78 | 1.31 | 1.34 | 1.33 |
| 9 | *n*-Propyl formate | 0.46 | 0.37 | 0.38 | 0.67 | 0.62 | 0.63 |
| 10 | *n*-Butyl acetate | 0.47 | 0.41 | 0.45 | 0.59 | 0.62 | 0.60 |
| 11 | Ethyl propionate | 0.41 | 0.46 | 0.45 | 0.67 | 0.60 | 0.58 |
| 12 | Ethyl butyrate | 0.50 | 0.47 | 0.45 | 0.73 | 0.57 | 0.58 |
| 13 | *n*-Butyl ether | 0.49 | 0.45 | 0.45 | 0.44 | 0.21 | 0.21 |
| 14 | Acetone | 0.46 | 0.50 | 0.49 | 0.64 | 0.70 | 0.70 |
| 15 | 2-Butanone | 0.49 | 0.49 | 0.51 | 0.81 | 0.70 | 0.70 |
| 16 | 2-Nonanone | 0.43 | 0.52 | 0.51 | 0.68 | 0.60 | 0.62 |
| 17 | Cyclopentanone | 0.38 | 0.52 | 0.52 | 0.71 | 0.87 | 0.86 |
| 18 | *n*-Propionitrile | 0.32 | 0.35 | 0.36 | 0.72 | 0.91 | 0.90 |
| 19 | *n*-Hexyl cyanide | 0.31 | 0.37 | 0.36 | 0.75 | 0.86 | 0.86 |
| 20 | *n*-Nitropropane | 0.50 | 0.36 | 0.31 | 0.99 | 0.92 | 0.95 |
| 21 | *n*-Nitrobutane | 0.32 | 0.30 | 0.31 | 0.94 | 0.96 | 0.93 |
| 22 | *n*-Nitropentane | 0.43 | 0.41 | 0.31 | 0.95 | 0.93 | 0.91 |
| 23 | Methylene chloride | 0.04 | 0.04 | 0.05 | 0.49 | 0.56 | 0.57 |
| 24 | Chloroform | 0.02 | 0.04 | 0.02 | 0.46 | 0.49 | 0.49 |
| 25 | Dibromomethane | 0.11 | 0.12 | 0.10 | 0.62 | 0.68 | 0.67 |
| 26 | 3-Phenyl propanol | 0.62 | 0.67 | 0.67 | 0.91 | 0.90 | 0.90 |
| 27 | Benzaldehyde | 0.44 | 0.41 | 0.39 | 0.94 | 0.99 | 1.00 |
| 28 | N-Benzyl formamide | 0.56 | 0.62 | 0.63 | 1.28 | 1.74 | 1.80 |
| 29 | Methyl benzoate | 0.49 | 0.47 | 0.46 | 1.06 | 0.85 | 0.85 |
| 30 | Ethyl benzoate | 0.50 | 0.45 | 0.46 | 1.11 | 0.86 | 0.85 |
| 31 | Anisole | 0.30 | 0.28 | 0.29 | 0.68 | 0.76 | 0.75 |
| 32 | Acetophenone | 0.34 | 0.48 | 0.48 | 1.13 | 1.01 | 1.01 |
| 33 | Propiophenone | 0.42 | 0.51 | 0.51 | 1.15 | 0.95 | 0.95 |
| 34 | Benzophenone | 0.53 | 0.50 | 0.50 | 1.56 | 1.50 | 1.50 |
| 35 | *m*-Toluenitrile | 0.29 | 0.32 | 0.34 | 1.20 | 1.09 | 1.10 |
| 36 | Nitrobenzene | 0.25 | 0.28 | 0.28 | 1.29 | 1.11 | 1.11 |
| 37 | *o*-Nitrotoluene | 0.34 | 0.29 | 0.27 | 1.32 | 1.09 | 1.11 |
| 38 | *p*-Nitrobenzyl bromide | 0.35 | 0.41 | 0.40 | 1.31 | 1.48 | 1.50 |
| 39 | *p*-Nitrobenzyl chloride | 0.34 | 0.38 | 0.40 | 1.33 | 1.38 | 1.34 |
| 40 | Fluoro benzene | 0.35 | 0.10 | 0.10 | 0.77 | 0.57 | 0.57 |
| 41 | Chloro benzene | 0.11 | 0.03 | 0.07 | 0.76 | 0.67 | 0.65 |
| 42 | Bromo benzene | 0.09 | 0.11 | 0.09 | 0.83 | 0.70 | 0.73 |
| 43 | Iodo benzene | 0.05 | 0.04 | 0.12 | 0.71 | 0.82 | 0.82 |

| No. | Name | $\Sigma\beta_2^H$ | | | $\pi_2^H$ | | |
|---|---|---|---|---|---|---|---|
| | | MLR | ANN | EXP | MLR | ANN | EXP |
| 44 | Benzyl bromide | 0.16 | 0.15 | 0.20 | 0.84 | 0.98 | 0.98 |
| 45 | *p*-Chloro toluene | 0.19 | 0.09 | 0.07 | 0.71 | 0.65 | 0.67 |
| 46 | *p*-Bromo toluene | 0.16 | 0.15 | 0.09 | 0.81 | 0.75 | 0.74 |
| 47 | *p*-Dichloro benzene | 0.11 | 0.04 | 0.02 | 0.61 | 0.74 | 0.75 |
| 48 | Benzene | 0.07 | 0.16 | 0.14 | 0.57 | 0.54 | 0.52 |
| 49 | Toluene | 0.16 | 0.13 | 0.14 | 0.59 | 0.51 | 0.52 |
| 50 | Ethyl benzene | 0.27 | 0.14 | 0.15 | 0.59 | 0.47 | 0.51 |
| 51 | *tert*-Butyl benzene | 0.17 | 0.18 | 0.16 | 0.69 | 0.52 | 0.49 |
| 52 | *p*-Xylene | 0.14 | 0.14 | 0.16 | 0.53 | 0.55 | 0.52 |
| 53 | Mesitylene | 0.20 | 0.21 | 0.19 | 0.59 | 0.51 | 0.52 |
| 54 | Biphenyl | 0.26 | 0.23 | 0.22 | 0.99 | 0.98 | 0.99 |
| 55 | Naphthalene | 0.10 | 0.17 | 0.20 | 0.84 | 0.91 | 0.92 |
| 56 | Anthracene | 0.14 | 0.24 | 0.26 | 1.06 | 1.35 | 1.34 |
| 57 | Phenol | 0.28 | 0.29 | 0.30 | 0.81 | 0.86 | 0.89 |
| 58 | *o*-Cresol | 0.37 | 0.33 | 0.30 | 0.73 | 0.88 | 0.86 |
| 59 | *p*-Ethyl phenol | 0.38 | 0.31 | 0.36 | 0.75 | 0.92 | 0.90 |
| 60 | *p*-Chloro phenol | 0.26 | 0.24 | 0.20 | 0.82 | 1.08 | 1.08 |
| **Test set** | | | | | | | |
| 1 | 1-Octanol | 0.49 | 0.44 | 0.48 | 0.48 | 0.46 | 0.34 |
| 2 | 1-Heptanal | 0.41 | 0.39 | 0.45 | 0.55 | 0.63 | 0.61 |
| 3 | N,N-Diethyl formamide | 0.68 | 0.73 | 0.76 | 1.08 | 1.22 | 1.25 |
| 4 | *n*-Hexyl acetate | 0.49 | 0.44 | 0.45 | 0.56 | 0.53 | 0.56 |
| 5 | *n*-Propylether | 0.45 | 0.40 | 0.45 | 0.37 | 0.16 | 0.23 |
| 6 | 2-Heptanone | 0.41 | 0.52 | 0.51 | 0.72 | 0.65 | 0.66 |
| 7 | *n*-Valeronitrile | 0.22 | 0.34 | 0.36 | 0.77 | 0.95 | 0.90 |
| 8 | *n*-Heptyl cyanide | 0.33 | 0.35 | 0.36 | 0.73 | 0.81 | 0.84 |
| 9 | Benzyl alcohol | 0.46 | 0.43 | 0.56 | 0.86 | 0.98 | 0.87 |
| 10 | Benzonitrile | 0.23 | 0.23 | 0.33 | 1.10 | 1.05 | 1.11 |
| 11 | *m*-Nitro toluene | 0.32 | 0.27 | 0.25 | 1.30 | 1.09 | 1.10 |
| 12 | *n*-Propyl benzene | 0.31 | 0.08 | 0.15 | 0.54 | 0.46 | 0.50 |
| 13 | *p*-Cresol | 0.36 | 0.34 | 0.31 | 0.73 | 0.83 | 0.87 |
| **Validation set** | | | | | | | |
| 1 | 1-Hexanol | 0.44 | 0.35 | 0.48 | 0.49 | 0.48 | 0.38 |
| 2 | 1-Octanal | 0.44 | 0.41 | 0.45 | 0.56 | 0.62 | 0.59 |
| 3 | N,N-Diethyl acetamide | 0.65 | 0.74 | 0.78 | 1.21 | 1.47 | 1.30 |
| 4 | *n*-Amyl acetate | 0.48 | 0.42 | 0.45 | 0.56 | 0.59 | 0.58 |
| 5 | Ethyl ether | 0.53 | 0.50 | 0.45 | 0.37 | 0.27 | 0.25 |
| 6 | 2-Hexanone | 0.35 | 0.47 | 0.51 | 0.73 | 0.74 | 0.68 |
| 7 | *n*-Hexanitrile | 0.28 | 0.33 | 0.36 | 0.74 | 0.88 | 0.88 |
| 8 | *n*-Octyl cyanide | 0.41 | 0.34 | 0.36 | 0.73 | 0.71 | 0.82 |
| 9 | 2-Phenyl ethanol | 0.51 | 0.58 | 0.64 | 0.87 | 0.98 | 0.91 |
| 10 | Benzyl cyanide | 0.30 | 0.31 | 0.45 | 1.16 | 1.11 | 1.15 |
| 11 | *p*-Nitro toluene | 0.34 | 0.29 | 0.28 | 1.25 | 1.14 | 1.11 |
| 12 | *n*-Butyl benzene | 0.34 | 0.09 | 0.15 | 0.53 | 0.51 | 0.51 |
| 13 | *m*-Cresol | 0.36 | 0.32 | 0.34 | 0.79 | 0.97 | 0.88 |

[a]Exp refers to experimental; ANN refers to artificial neural network; MLR refers to multiple linear regression determined value of $\pi_2^H$ and $\Sigma\beta_2^H$.

TABLE-2
SPECIFICATION OF MULTIPLE LINEAR REGRESSIONS FOR THE MODELING OF $\pi_2^H$

| Descriptor | Notation | Coefficient | Main effect |
|---|---|---|---|
| Geary autocorrelation –lag2/weighted by atomic Sanderson electronegativites | GATS2E | -0.489 (±0.104) | -0.370 |
| 2nd Component accessibility directional WHIM index/weighted by atomic van der Waals volumes | E2V | 0.961 (±0.230) | 0.290 |
| d COMMA 2 value/weighted by atomic Sanderson electronegativites | DISPE | -0.571(±0.188) | -0.142 |
| Number of H attached to $C_1(sp^3)/C_0(sp^2)$ | HAC | 0.089 (±0.009) | 0.352 |
| Dipole moment | DP | 0.381 (±0.062) | 0.325 |
| Constant | | 0.470 (±0.084) | |

molecular moment analysis) value/weighted by atomic Sanderson electronegativites (DISPE), number of H attached to $C_1(sp3)/C_0(sp2)$ (HAC) and dipol moment (DP). Also the names of the descriptors that appear in the best MLR equation for modeling of $\Sigma\beta_2^H$ parameter are shown in Table-3. These descriptors are: maximal electrotopological positive variation (MAXDP), average connectivity index chi-5 (X5A), Geary autocorrelation-lag1/weighted by atomic polarizabilities (GATS1P), Radial distribution function-2/unweighted (RDF020U) and leverage-weighted autocorrelation-lag 4/unweighted (HATS4U). These descriptors were used as inputs for generated ANNs. A detailed description of these descriptors has been adequately described elsewhere[39-50].

TABLE-3
SPECIFICATION OF MULTIPLE LINEAR REGRESSIONS FOR THE MODELING OF $\Sigma\beta_2^H$

| Descriptor | Notation | Coefficient | Main effect |
|---|---|---|---|
| Maximal electro topological positive variation | MAXDP | 0.073 (±0.009) | 0.155 |
| Average connectivity index chi-5 | X5A | -0.723 (±0.198) | -0.069 |
| Geary autocorrelation-lag1/weighted by atomic polarizabilities | GATS1P | 0.129 (±0.031) | 0.149 |
| Leverage-weighted autocorrelation of 4/unweighted | HATS4U | 0.089 (±0.030) | 0.068 |
| Radial distribution function-2/unweighted | RDF020U | 0.069 (±0.010) | 0.131 |
| Constant | | -0.092 (±0.037) | |

**Artificial neural network:** Artificial neural networks have been applied in QSPR analysis since the late of 1980s due to its flexibility in modeling of non-linear problems. They are parallel computational devices consisting of groups of highly interconnected processing elements called neurons. They are characterized by topology, computational characteristics of their elements and training rules. Traditional neural

networks have neurons arranged in a series of layers. The first layer is termed the input layer, each of its neurons receives information from the exterior, corresponding to one of the independent variables used as inputs. The last layer is the output layer and its neurons handle the output from the network. The layers of neurons between the input and output layers are called hidden layers. Each layer may make its independent computations and may pass the results to another layer. In feed forward neural networks the connections among neurons are directed upwards. ANN offers attractive possibilities for non-linear modeling and optimization when underlying mechanisms are complex. They have been widely used to predict many physico-chemical properties. The theories behind of artificial neural networks have been adequately described[51-53].

The program for the feed-forward neural networks that were trained by back-propagation algorithm was written by VISUAL FORTRAN in the laboratory. These networks have 5 nodes in the input layer and 1 node in the output layer. Descriptors that appeared in the selected MLR models were used as inputs for the generated ANNs and their outputs are the values of $\pi_2^H$ or $\Sigma\beta_2^H$ for the molecules of interest. The number of nodes in the hidden layer would be optimized. The initial weights were randomly selected from a uniform distribution that ranged between -0.3 and +0.3. The initial bias values were set to be one. These values were optimized during the network training. The value of each input was divided into its mean value to bring them into the dynamic range of the sigmoid transfer function of the ANN. Before training, the network was optimized for the number of nodes in the hidden layer, learning rate and momentum and then the network was trained using the training and test sets to optimize the values of weights and biases. The test set was used to prevent the network from overfitting. In order to evaluate the prediction power of the ANN, trained network was employed to calculate the solute parameters for the validation set.

## RESULTS AND DISCUSSION

The data set and corresponding observed and ANN predicted values of $\pi_2^H$ and $\Sigma\beta_2^H$ for all molecules studied in this work are shown in Table-1. The selected descriptors that appear in MLR model for $\pi_2^H$ are shown in Table-2. In this table, the parameter of GATS2E has maximum main effect. This parameter is an auto-correlation descriptor that weighted by electronegativity and has a negative effect on the value of $\pi_2^H$. Increasing of GATS2E causes a decreasing in $\pi_2^H$ value due to the decreasing in the electronegativity effect. Next descriptor that has high main effect is HAC. This descriptor shows the sum of the number of hydrogen atoms connected to $SP^2$ carbon atom and the number of hydrogen atoms attached to carbon atoms, which this carbon connected to an electronegative atom. The value of this descriptor represents the polarity characteristic of a molecule and inclusion of this parameter in $\pi_2^H$ model reveals the role of dipolarity/polarizability interaction in solute properties. Other parameter in this model is dipole moment (DP) that has a

direct effect on polarity parameter. The descriptor of E2V shows the electron density around the Y-axis in a molecule and has a positive main effect. This parameter has direct relation with the $\pi_2^H$ value because increasing in electron density around the Y-axis causes that this molecule more stretched and the value of $\pi_2^H$ would be increased. The final descriptor that appears in $\pi_2^H$ model with the smallest main effect is DISPE, which is a component of COMMA type's descriptors. This descriptor indicates displacement between the geometric center and the center of the considered property field with respect to the molecular principal axes. The considered property field for the DISPE descriptor is electronegativites. Increasing in the displacement between the geometric center and the center of electronegative field causes a decrease in the electronegative effectiveness and decreasing in the value of $\pi_2^H$.

The name of descriptors and their main effects that appear in MLR model for $\Sigma\beta_2^H$ are shown in Table-3. The parameter of MAXDP has a maximum positive main effect and describes compound electrophilicity, which relate to the types of polar groups in the molecules. Next descriptor is GATS1P with positive main effect, which is an autocorrelation descriptor that weighted by polarizability. This descriptor describes how considered property (polarizability) was distributed along a topological structure of a molecule. Increasing in GATS1P causes an increasing in $\Sigma\beta_2^H$ value due to the increasing in charge distribution in the molecule. Other descriptor with a positive main effect in $\Sigma\beta_2^H$ model was radial distribution function (RDF020U). This descriptor provides information about atom types, ring types and planar or non-planar systems of a molecule. As shown in Table-3, another descriptor is HATS4U, which is a one type of the GETAWAY descriptors and encoding information on the effective position of substituent, fragments in the molecular space and the molecular size for specific atomic properties and has a positive effect on the value of $\Sigma\beta_2^H$. The final descriptor in $\Sigma\beta_2^H$ model with the smallest main effect was X5A. The X5A parameter is one of the topological descriptors that related to the valence layer electrons and the number of atoms in molecule, this descriptor shows the effect of these two factors simultaneously and has a negative effect on $\Sigma\beta_2^H$. The inclusion of these descriptors in $\pi_2^H$ and $\Sigma\beta_2^H$ models reveals the important roles of electronic and steric interactions in solute characteristics.

In the next step, two separately ANN developed to calculate of the $\pi_2^H$ and $\Sigma\beta_2^H$. Before training the networks, the parameters of the number of nodes in the hidden layer, weights and biases learning rates and momentum values were optimized. The procedure for the optimization of these parameters is reported in previous papers[54-56]. Table-4 shows the architecture and specifications of the optimized ANNs parameters for modeling the values of $\pi_2^H$ and $\Sigma\beta_2^H$. These networks were then trained by using the training set for the optimization of the weights and biases values by back propagation algorithm. It is known that neural network can become over-trained. An over-trained network has usually learned perfectly the stimulus pattern it has seen but can not give accurate prediction for unseen stimuli and it no

TABLE-4
ARCHITECTURE AND SPECIFICATION OF THE GENERATED ANNs FOR $\pi_2^H$ and $\Sigma\beta_2^H$ MODELING

| Parameter | $\pi_2^H$ ANN | $\Sigma\beta_2^H$ ANN |
|---|---|---|
| No. of nodes in the input layer | 5 | 5 |
| No. of nodes in the hidden layer | 8 | 9 |
| No. of nodes in the output layer | 1 | 1 |
| Weights learning rate | 0.210 | 0.250 |
| Biases learning rate | 0.320 | 0.850 |
| Momentum | 0.391 | 0.480 |
| Transfer function | Sigmoid | Sigmoid |

longer able to generalize. There are several methods for overcoming this problem. One method is to use a test set to evaluate the prediction power of the network during its training. In this method, after each 100 training iteration the network was used to calculate solute parameters ($\pi_2^H$ or $\Sigma\beta_2^H$) of molecules included in the test set. To maintain the predictive power of the network at a desirable level, training was stopped when the value of errors for the test set started to increase. Since this error is not a good estimate of the generalization error, prediction potential of the model was evaluated on a third set of data, named validation set. The compounds in the validation set were not used during the training process and were reserved to evaluate the predictive power of the generated ANN.

Table-1 represent the experimental and ANNs predicted values of $\pi_2^H$ and $\Sigma\beta_2^H$ for the training, test and validation sets. Also in Table-5, the statistical parameters of ANNs and MLRs models for the $\pi_2^H$ and $\Sigma\beta_2^H$ parameters are shown. For the $\Sigma\beta_2^H$ model, the standard errors of training, test and validation sets for the MLR model are 1.439, 1.22 and 0.928, respectively which would be compared with the values of 0.177, 0.638 and 0.670, respectively, for the ANN model. Also, for the $\Sigma\beta_2^H$ model, the standard errors of training, test and validation sets for the MLR model are 0.716, 0.728 and 1.086, respectively which would be compared with the values of 0.293, 0.475 and 0.512, respectively, for the ANN model. Comparison between these values and other statistical values in Table-5 shows the superiority of ANNs over MLR models.

TABLE-5
STATISTICAL PARAMETERS OBTAINED USING THE ANN AND MLR MODELS[a]

| Parameter | Model | $(SE)_C$ | $(SE)_p$ | $(SE)_v$ | $R_t$ | $R_p$ | $R_v$ | $F_t$ | $F_p$ | $F_v$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\pi_2^H$ | ANN | 0.177 | 0.638 | 0.670 | 0.998 | 0.980 | 0.978 | 18013 | 266 | 244 |
| | MLR | 1.439 | 1.221 | 0. 928 | 0.887 | 0.924 | 0.958 | 214 | 65 | 123 |
| $\Sigma\beta_2^H$ | ANN | 0.293 | 0.475 | 0.512 | 0.989 | 0.955 | 0.950 | 2505 | 113 | 101 |
| | MLR | 0.716 | 0.728 | 1.086 | 0.916 | 0.828 | 0.748 | 303 | 24 | 14 |

[a]t refers to the training set; p refer to the prediction set; v refers to the validation set; $(SE)_C$ is the standard error of training; set; $(SE)_p$ is the standard error of test set; $(SE)_v$ is the standard error of validation set; R is the correlation coefficient; F is the statistical F value.

Figs. 1 and 2 shows the plot of the ANN predicted *versus* the experimental values of $\pi_2^H$ and $\Sigma\beta_2^H$ for the training, test and validation sets, respectively. The propagation of results in both sides of regression line indicates that no systematic error exists in the generated ANNs models.

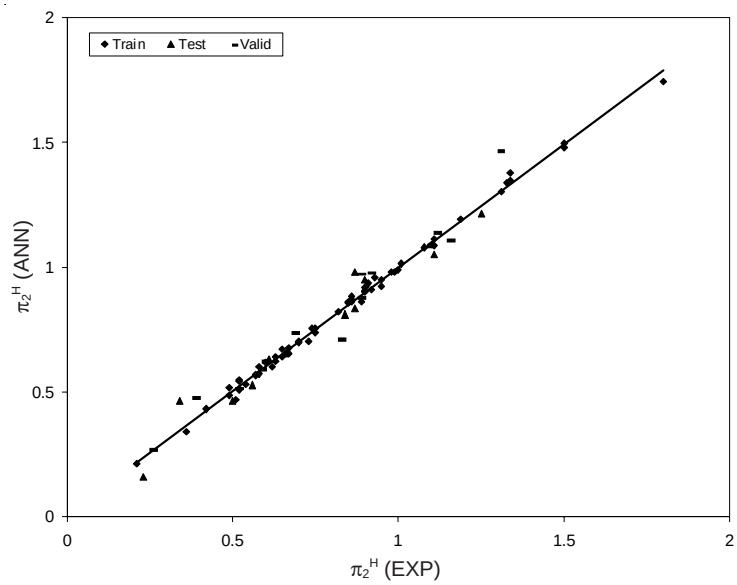


Fig. 1. Plot of the calculated $\pi_2^H$ against the experimental values

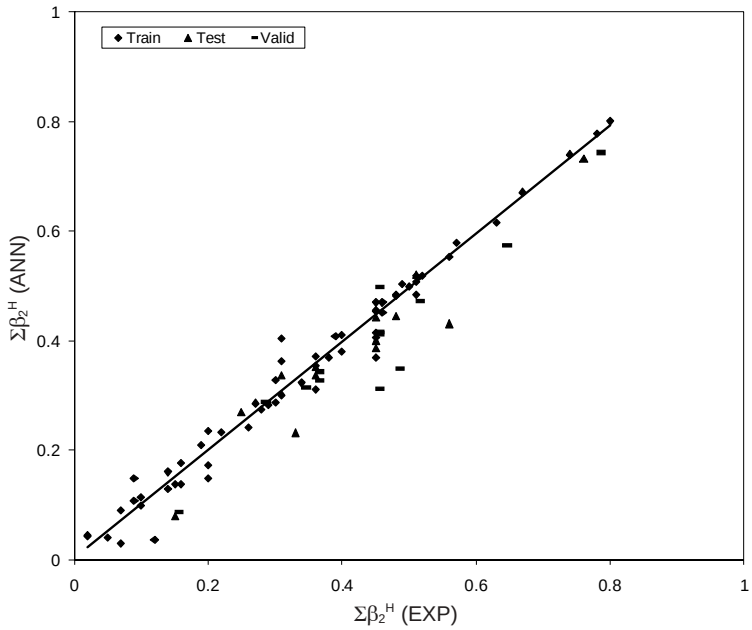


Fig. 2. Plot of the calculated $\Sigma\beta_2^H$ against the experimental values

## Conclusion

The QSPR model presented here is better than those reported by Svozil and co-worker[36] due to higher correlation coefficient; lower the number of used descriptors and diversity of compound, studied in this work. The results of this study demonstrate that the QSPR method using the ANN techniques can generate suitable models for the prediction of the $\pi_2^H$ and $\Sigma\beta_2^H$ values for some aliphatic and aromatic compounds. The key strength of the neural networks is their ability to allow for flexible mapping of the selected features by manipulating their functional dependence implicitly, unlike regression analysis. Neural network handles both linear and non-linear relationships without adding complexity to model. This capability offset the large computing time required and complexity of the ANN method with respect to MLR. Also the analysis of the results obtained indicates that the models one can proposed correctly represent the relationship between these LSER solute parameters and theoretically calculated molecular descriptors.

## REFERENCES


1. M.J. Kamlet, R.W. Taft and J.M. Abboud, *J. Am. Chem. Soc.*, **25**, 91 (1977).
2. M.J. Kamlet and R.W. Taft, *J. Am. Chem. Soc.*, **98**, 377 (1976).
3. R.W. Taft and M.J. Kamlet, *J. Am. Chem. Soc.*, **98**, 2886 (1976).
4. T. Yokoyama, R.W. Taft and M.J. Kamlet, *J. Am. Chem. Soc.*, **98**, 3323 (1976).
5. M.J. Kamlet, J.M. Abboud and R.W. Taft, *J. Am. Chem. Soc.*, **98**, 6027 (1976).
6. M.J. Kamlet, M.H. Abraham, R.M. Doherty and R.W. Taft, *J. Am. Chem. Soc.*, **106**, 464 (1984).
7. M.J. Kamlet, J.M. Abboud, R.M. Doherty, M.H. Abraham and R.W. Taft, *Chemtech*, **16**, 556 (1986).
8. M.H. Abraham, G.S. Whiting, R.M. Doherty and W.J. Shuely, *J. Chromatogr.*, **587**, 229 (1991).
9. M.H. Abraham, G.S. Whiting, J. Andonian-Haftven, J.W. Steed and J.W. Grate, *J. Chromatogr.*, **588**, 361 (1991).
10. M.H. Abraham, G.S. Whiting, R.M. Doherty, W.J. Shuely and Posakellariou, *Polymer*, **33**, 2162 (1992).
11. M.H. Abraham, I. Hamerton, J.B. Rose and J.W. Grate, *J. Chem. Soc. Perkin. Trans. II*, 1417 (1991).
12. M.H. Abraham and G.S. Whiting, *J. Am. Oil Chem. Soc.*, **69**, 1236 (1992).
13. M.H. Abraham andonian-Haftven, J.P. Osei-Owusu, P. Sakellarious, J.S. Urieta, M.C. Lopez and R. Fuchs, *J. Chem. Soc. Perkin Trans. II*, 299 (1993).
14. N. Chen, Y. Zhang and Polu, *J. Chromatogr.*, **606**, 1 (1992).
15. F. Gritti, G. Felix, W.F. Achard and F. Hardouin, *J. Chromatogr. A*, **922**, 51 (2001).
16. A. Wang, L.C. Tan and P.W. Carr, *J. Chromatogr. A*, **848**, 21 (1999).
17. J.H. Park, M.D. Jang and S.T. Kim, *Bull. Korean Chem. Soc.*, **2**, 297 (1990).
18. L.C. Tan, P.W. Carr and M.H. Abraham, *J. Chromatogr.*, **752**, 1 (1996).
19. D.E. Leahy, P.W. Carr, R.S. Pearlman, R.W. Taft and M.J. Kamlet, *Chromatographia*, **21**, 2674 (1986).
20. L.C. Tan and P.W. Carr, *J. Chromatogr.*, **799**, 1 (1998).
21. J.H. Park and P.W. Carr, *J. Chromatogr.*, **465**, 123 (1989).
22. J. Li and D.A. Whitman, *Anal. Chim. Acta*, **368**, 141 (1998).
23. J.H. Park, M.H. Youn, Y.K. Ryu, B.E. Kim, J.W. Ryu and M.D. Jang, *J. Chromatogr. A*, **769**, 249 (1998).
24. F.Z. Oumada, M. Roses, E. Bosch and M.H. Abraham, *Anal. Chim. Acta*, **382**, 301 (1999).
25. M.H. Abraham, P.L. Grellier and R.A. Megill, *J. Chem. Soc. Perkin Trans. II*, 797 (1987).
26. M.H. Abraham, *Chem. Soc. Rev.*, **110**, 73 (1993).
27. M.H. Abraham, G.S. Whiting, R.M. Doherty and W.J. Shuely, *J. Chromatogr.*, **587**, 213 (1991).
28. M.H. Abraham, A. Ibrahim and A.M. Zissimos, *J. Chromatogr.*, **1037**, 29 (2004).

29. D.S. Ballantine and B.K. Callihan, *J. Chromatogr. A*, **915**, 177 (2001).
30. M.H. Abraham, G.S. Whiting, R.M. Doherty and W.J. Shuely, *J. Chromatogr.*, **587**, 213 (1991).
31. M. Rogalski and F. Mutelet, *J. Chromatogr. A*, **988**, 117 (2003).
32. J.A. Platts, M.H. Abraham and A. Hersey, *J. Chem. Inf. Comput. Sci.*, **39**, 835 (1999).
33. J.A. Platts, *Phys. Chem. Chem. Phys.*, **2**, 973 (2000).
34. J.A. Platts, *Phys. Chem. Chem. Phys.*, **2**, 3115 (2000).
35. O. Lamache, J.A. Platts and A. Hersey, *Phys. Chem. Chem. Phys.*, **3**, 2747 (2001).
36. D. Svozil, V. Kvasnicka and J.G. Sevcik, *J. Chem. Inf. Comput. Sci.*, **37**, 338 (1997).
37. http: www.disat.unimib.it chm (2001)
38. J.J.P. Stewart, MOPAC, Semi Empirical Molecular Orbital Program, QCPE, Vol. 455, Research Laboratory, United States Air Force Academy, Version 6 (1990).
39. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim (2000).
40. V.N. Viswanadhan, A.K. Ghose, G.R. Revankar and R.K. Robins, *J. Chem. Inf. Comput. Sci.*, **29**, 163 (1989).
41. L.B. Kier and L.H. Hall, Molecular Connectivity in Structure- Activity Analysis, Research Studies Press-Wiley, Chichester (UK) (1986).
42. L.B. Kier and L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York (1976).
43. M. Randic, *J. Mol. Graph. Model.*, **20**, 19 (2001).
44. R.C. Geary, *Incorp. Statist.*, **5**, 115 (1954).
45. M.C. Hemmer, V. Steinhauer and J. Gasteiger, *Vibrat. Spect.*, **19**, 151 (1999).
46. R. Todeschini, M. Lasagni and E. Marengo, *J. Chemom.*, **8**, 263 (1994).
47. B.D. Silverman, *J. Chem. Inf. Comput. Sci.*, **40**, 1470 (2000).
48. V. Consonni, R. Todeschini and M. Pavan, *J. Chem. Inf. Comput. Sci.*, **42**, 682 (2002).
49. V. Consonni, R. Todeschini and M. Pavan, *J. Chem. Inf. Comput. Sci.*, **42**, 693 (2002).
50. P. Gramatica, M. Corradi and V. Consonni, *Chemosphere*, **41**, 763 (2000).
51. J. Zupan and J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim (1999).
52. S. Haykin, Neural Networks, Prentice-Hall, Englewood Cliffs, N.J. (1994).
53. N.K. Bose and P. Liang, Neural Networks, Fundamentals, Mc Graw-Hill, New York (1996).
54. M. Jalali-Heravi and M.H. Fatemi, *Anal. Chim. Acta*, **415**, 95 (2000).
55. M. Jalali-Heravi and M.H. Fatemi, *J. Chromatogr. A*, **915**, 177 (2001).
56. M.H. Fatemi, *J. Chromatogr. A*, **1038**, 231 (2004).

(*Received*: 24 April 2007; *Accepted*: 1 January 2009) AJC-7087