Polymer 51 (2010) 3568–3574

Contents lists available at ScienceDirect

Polymer

journal homepage: www.elsevier.com/locate/polymer

# Prediction of inherent viscosity for polymers containing natural amino acids from the theoretical derived molecular descriptors

Shadpour Mallakpour [a,*], Mehdi Hatami [a], Hassan Golmohammadi [b]

[a] Organic Polymer Chemistry Research Laboratory, Department of Chemistry, Isfahan University of Technology, Isfahan 84156-83111, Islamic Republic of Iran
[b] Department of Chemistry, University of Mazandaran, Babolsar, Islamic Republic of Iran

ARTICLE INFO

*Article history:*
Received 19 February 2010
Received in revised form
17 May 2010
Accepted 17 May 2010
Available online 1 June 2010

*Keywords:*
Quantitative structure-property relationship
Inherent viscosity
Artificial neural network

ABSTRACT

The main aim of the present work was development of a quantitative structure-property relationship (QSPR) method using an artificial neural network (ANN) for the prediction of inherent viscosity ($\eta_{inh}$) of a data set of 75 optically active polymers containing natural amino acids. The total of 540 descriptors was calculated for all molecules in the data set. In the next step an ANN was constructed and trained for the prediction of $\eta_{inh}$ of polymers. The inputs of this neural network are theoretically derived descriptors that were chosen by genetic algorithm (GA) and multiple linear regression (MLR) feature selection techniques. The values of standard errors for the neural network calculated $\eta_{inh}$ of training, test and validation sets are 0.023, 0.030 and 0.031, respectively. Comparison between these values and other statistical values reveal the superiority of the ANN model over the MLR one.



## 1. Introduction

Synthesis of optically active polymers is an important field in macromolecular science as they find a wide variety of potential applications based on the chiral structure [1–3]. One of the most practical and widely accepted applications of chiral polymers is the use as chiral stationary phase (CSP) for high-performance liquid chromatography (HPLC) for the separation of racemic compounds (resolution). Optically active polymers are newly considerable topics which have been paid attention lately. Because polymers with chiral structures are biologically very important, most of the natural polymers are optically active and have special chemical activities such as catalytic properties that exist in genes, proteins and enzymes. Some other applications could be listed as: (1) constructing chiral media for asymmetric synthesis, (2) chiral stationary phases for resolution of enantiomers in chromatographic techniques, (3) chiral liquid crystal in ferro-electric and non-linear optical devices [4–6].

Inherent viscosity of an optically active polymer is the ratio of natural logarithm of the relative viscosity, $\eta_r$ to the mass concentration of the polymer, C, i.e.:

$$\eta_{inh} = \frac{(\ln \eta_r)}{C} \quad (1)$$

In view of fact that experimental determination of the inherent viscosity of a large set of polymers is expensive, time-consuming and required high-purity samples and skilled operators, so the development of an alternative method such as quantitative structure-property relationship (QSPR) would be useful for the theoretical calculation of $\eta_{inh}$ values. QSPR is a mathematical method that relates the properties of interested molecule to its structural features. A current trend in quantitative structure-property relationship studies is the use of theoretical molecular descriptors, which can be calculated directly from molecular structure. Obtained QSPR model, can be used to estimate the properties of other polymers even when their structure is only sketchy. Numerous researchers have attempted to predict some physical properties values for polymers on the basis of quantitative structure-property relationships (QSPRs). Yu et al. [7] developed a QSPR model for prediction of glass transition temperature ($T_g$) of 107 polystyrenes using multiple linear stepwise regression and four descriptors, RSC, SMC, DHB and MPE. Afantitis et al. [8] predicted intrinsic viscosity in polymer–solvent combinations using a novel QSPR model based on multiple linear regression (MLR) technique. Liu et al. [9] constructed a model to correlate molar volumes (V), refractive index (n), and glass transition temperature ($T_g$) to the structural units of 35 polymethacrylates by stepwise regression and artificial neural network (ANN)

* Corresponding author. Tel.: +98 311 391 3267; fax: +98 311 391 2350.
*E-mail addresses:* mallak@cc.iut.ac.ir, mallakpour84@alumni.ufl.edu, mallak777@yahoo.com (S. Mallakpour).

0032-3861/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2010.05.033

methods. Yu et al. [10] estimated the dielectric dissipation factor (tanδ) of 92 polymers with an ANN model based on the DFT calculation.

Recently artificial neural networks (ANNs) have been used for a wide variety of chemical problems such as mass spectral search [11], prediction of enthalpy of alkanes [12] classification of ion mobility spectra [13], and prediction of dielectric constants [14]. ANNs were also used in quantitative structure property relationship studies [15–19]. In this investigation, the calculated descriptors from structures were used lonely to predict the inherent viscosity of 75 optically active polymers using the ANN and QSPR methods.

## 2. Methods

### 2.1. Dataset

The inherent viscosity ($\eta_{inh}$) of 75 optically active polymers were taken from Refs. [20–29], which was used as data set. The molecules in the data set are shown in Table 1. Data set was randomly divided into three separate sections, the training, test, and external validation sets, consisting of 51, 12, and 12 members, respectively. The training set was used to adjust the parameters of models; the test set was used for monitoring the extent of overtraining and external validation set was used for evaluation of the prediction power of obtained model.

### 2.2. Molecular descriptors generation

Molecular descriptors are mathematical values that describe the structure or shape of molecules, helping predict the activity and properties of molecules in complex experiments [30]. It is impossible to calculate descriptors directly for an entire molecule because all polymers have wide distribution of molecular weight and possess high molecular weight. It is understandable that, if the molecular weight is high enough, the terminal groups hold only a very small proportion in a polymer and its effect on the properties can be ignored. Molecular descriptors calculated directly from the structure of the repeating units (RU) can be used for the study of QSPRs for polymers, since all the properties depend on the chemical structure of the polymer molecules, and all these structures were conditioned by the RU structure. Therefore, we adopt this method and focus on the following model to calculate molecular descriptors. The structures for polymers were endcapped with last group of opposing side. In the next step, the molecular structure of monomer compounds used in the polymerization, were used to determine the molecular descriptors of obtained polymers. After providing the data set, all molecules (RU) were drawn into Hyperchem software [31] and optimized using the PM3 semiempirical method. In a next step, the Hyperchem output files were used by the dragon package to calculate molecular descriptors. Dragon is new, freely available software (by Milano Chemometrics and the QSAR Research Group) for the calculation of more than 800 molecular descriptors. Some of generated descriptors for each compound encoded similar information about the molecule of interest, therefore it was desirable to test each descriptor and eliminate those that show high correlation ($R > 0.9$) with each other. A total of 124 out of 540 descriptors showed high correlation and were removed from the next consideration. Subsequently, the method of genetic algorithm (GA) and stepwise multiple linear regression (MLR) were used to select the most important descriptors and to calculate coefficients relating the descriptors to inherent viscosity. The descriptors that appear in the best MLR equation are shown in Table 2.

**Table 1**
Data set and corresponding observed and predicted values of inherent viscosity of polymers

| Number | Name of monomers of polymers | ηinh (EXP) | ηinh (MLR) | ηinh (ANN) |
|---|---|---|---|---|
| | Training set | | | |
| 1 | [a] asnd *p*-phenylenediamine | 0.27 | 0.29 | 0.25 |
| 2 | [a] and 2,4-diaminotoluene | 0.30 | 0.24 | 0.31 |
| 3 | [a] and *m*-phenylene diamine | 0.62 | 0.59 | 0.62 |
| 4 | [a] and enzidine | 0.55 | 0.63 | 0.58 |
| 5 | [a] and 4,4'-diaminodiphenylether | 0.78 | 0.65 | 0.74 |
| 6 | [a] and 4,4'-diaminodiphenyl methane | 0.49 | 0.45 | 0.47 |
| 7 | [b] and bisphenol A | 0.49 | 0.51 | 0.53 |
| 8 | [b] and 4,4'-dihydroxydiphenyl sulphide | 0.54 | 0.49 | 0.52 |
| 9 | [b] and bisphenyl-2,2'-diol | 0.35 | 0.30 | 0.38 |
| 10 | [b] and 4,4'-dihydroxydiphenyl sulphone | 0.45 | 0.45 | 0.45 |
| 11 | [c] and phenol phthalein | 1.11 | 0.90 | 1.09 |
| 12 | [c] and bisphenol-A | 0.58 | 0.51 | 0.59 |
| 13 | [c] and 1,4-dihydroxyanthraquinone | 0.42 | 0.41 | 0.41 |
| 14 | [c] and 1,5-dihydroxy naphthalene | 0.40 | 0.35 | 0.40 |
| 15 | [c] and dihydroxy biphenyl | 0.47 | 0.41 | 0.48 |
| 16 | [c] and 2,4-dihydroxyacetophenone | 0.52 | 0.47 | 0.53 |
| 17 | [d] and benzidine | 0.22 | 0.29 | 0.24 |
| 18 | [d] and 1,5-diaminoanthraquinone | 0.09 | 0.07 | 0.10 |
| 19 | [d] and 4,4'-sulfonyldianiline | 0.09 | 0.08 | 0.11 |
| 20 | [d] and 3,3'-diaminobenzophenone | 0.22 | 0.25 | 0.23 |
| 21 | [d] and 2,6-diaminopyridine | 0.12 | 0.13 | 0.11 |
| 22 | [e] and 4,4'-diaminodiphenylmethane | 0.28 | 0.25 | 0.26 |
| 23 | [e] and 1,4-phenylenediamine | 0.32 | 0.27 | 0.31 |
| 24 | [e] and 1,3-phenylenediamine | 0.28 | 0.20 | 0.25 |
| 25 | [e] and 4,4'-diaminobiphenyl | 0.31 | 0.28 | 0.34 |
| 26 | [f] and bisphenol-A | 0.56 | 0.76 | 0.54 |
| 27 | [f] and phenolphthalein | 0.63 | 0.58 | 0.65 |
| 28 | [f] and 4,6-dihydroxypyrimidine | 0.49 | 0.54 | 0.48 |
| 29 | [f] and 2,4-dihydroxyacetophenone | 0.67 | 0.49 | 0.69 |
| 30 | [g] and 4,4'-sulphonyldianiline | 0.34 | 0.26 | 0.36 |
| 31 | [g] and 4,4'-diaminodiphenylether | 0.26 | 0.30 | 0.29 |
| 32 | [g] and *p*-phenylenediamine | 0.33 | 0.29 | 0.31 |
| 33 | [g] and 4,4'-diaminobiphenyl | 0.39 | 0.25 | 0.35 |
| 34 | [h] and phenolphthalein | 0.70 | 0.63 | 0.67 |
| 35 | [h] and 1,4-dihydroxybenzene | 0.67 | 0.58 | 0.69 |
| 36 | [h] and 4,4'-dihydroxydiphenyl sulphide | 0.55 | 0.60 | 0.57 |
| 37 | [h] and 4,4'-dihydroxydiphenyl sulphone | 0.57 | 0.53 | 0.58 |
| 38 | [h] and 2,6 dihydroxytoluene | 0.86 | 0.61 | 0.84 |
| 39 | [i] and 4,4'-sulphonyldianiline | 0.42 | 0.36 | 0.41 |
| 40 | [i] and 4,4'-diaminobiphenyl | 0.37 | 0.42 | 0.34 |
| 41 | [j] and *p*-phenylenediamine | 0.28 | 0.36 | 0.29 |
| 42 | [j] and *m*-phenylenediamine | 0.33 | 0.24 | 0.37 |
| 43 | [k] and bisphenol-A | 0.20 | 0.28 | 0.21 |
| 44 | [k] and 4,4'-hydroquinone | 0.20 | 0.25 | 0.22 |
| 45 | [k] and 1,8-dihydroxyanthraquinone | 0.17 | 0.22 | 0.19 |
| 46 | [k] and 4,4-dihydroxy biphenyl | 0.27 | 0.21 | 0.24 |
| 47 | [k] and 2,4-dihydroxyacetophenone | 0.10 | 0.14 | 0.11 |
| 48 | [l] and 4,4'-diaminodiphenyl methane | 0.52 | 0.46 | 0.56 |
| 49 | [l] and 2,4-diaminotoluene | 0.41 | 0.34 | 0.38 |
| 50 | [l] and 4,4'-sulfonyldianiline | 0.57 | 0.47 | 0.54 |
| 51 | [l] and *m*-phenylenediamine | 0.37 | 0.44 | 0.40 |
| | Test set | | | |
| 52 | [a] and 4,4'-diaminodiphenylsulfone | 0.34 | 0.32 | 0.36 |
| 53 | [b] and 2,6- dihydroxy toluene | 0.41 | 0.36 | 0.44 |
| 54 | [c] and 4,4'-hydroquinone | 0.72 | 0.64 | 0.72 |

| Number | Name of monomers of polymers | ηinh (EXP) | ηinh (MLR) | ηinh (ANN) |
|---|---|---|---|---|
| 55 | [d] and 4,4'-diaminodiphenylmethane | 0.29 | 0.23 | 0.26 |
| 56 | [e] and 4,4'-diaminodiphenylether | 0.25 | 0.20 | 0.22 |
| 57 | [f] and 4,4'-dihydroxydiphenyl sulfone | 0.48 | 0.53 | 0.52 |
| 58 | [g] and *m*-phenylenediamine | 0.29 | 0.42 | 0.27 |
| 59 | [h] and 2,4-dihydroxyacetophenone | 0.80 | 0.64 | 0.76 |
| 60 | i and 1,3-phenylenediamine | 0.33 | 0.29 | 0.36 |
| 61 | [j] and 4,4'-diaminobiphenyl | 0.43 | 0.47 | 0.40 |
| 62 | [k] and phenol phthalein | 0.15 | 0.11 | 0.17 |

(*continued on next page*)

**Table 1** (*continued*)

| Number | Name of monomers of polymers | ηinh (EXP) | ηinh (MLR) | ηinh (ANN) |
|---|---|---|---|---|
| 63 | [l] and 4,4'-diaminodiphenylether Validation set | 0.56 | 0.48 | 0.58 |
| 64 | [a] and 2,6-diaminopyridine | 0.38 | 0.44 | 0.42 |
| 65 | [b] and 1,4-dihydroxybenzene | 0.45 | 0.53 | 0.41 |
| 66 | [c] and 1,8-dihydroxyanthraquinone | 0.53 | 0.45 | 0.50 |
| 67 | [d] and *p*-phenylenediamine | 0.15 | 0.19 | 0.13 |
| 68 | [e] and 4,4'-sulphonyldianiline | 0.35 | 0.28 | 0.31 |
| 69 | [f] and 1,4-dihydroxybenzene | 0.55 | 0.41 | 0.54 |
| 70 | [g] and 4,4'-diaminodiphenyl methane | 0.42 | 0.30 | 0.46 |
| 71 | h and bisphenol-A | 1.00 | 0.81 | 1.04 |
| 72 | i and 4,4'-diaminodiphenylether | 0.35 | 0.29 | 0.33 |
| 73 | [j] and 4,4'-diaminodiphenylether | 0.36 | 0.32 | 0.40 |
| 74 | [k] and 1,5-dihydroxy naphthalene | 0.17 | 0.23 | 0.15 |
| 75 | [l] and *p*-phenylenedi-amine | 0.44 | 0.48 | 0.41 |

[a] is 4,4'-(hexafluoroisopropylidene)-*N,N'*-bis-(phthaloyl-L-leucine-p-amidobenzoic acid)
[b] is *N,N'*-(4,4'-hexafluoroisopropylidendiphthaloyl)-bis-*L*-isoleucine
[c] is 4,4'-(hexafluoroisopropylidene)-*N,N'*-bis-(phthaloyl-L-leucine) diacid chloride
[d] is 4,4'-(hexafluoroisopropylidene) bis-(phthaloyl-L-leucine)
[e] is 4,4'–(hexafluoroisopropylidene)-*N,N'*-bis-(phthaloyl-L-methionine) diacid chloride
[f] is *N,N'*-(4,4-hexafluoroisopropylidenediphthaloyl)-bis-L-methionine
[g] is *N,N'*-(4,4'-oxydiphthaloyl)-bis-*L*-isoleucine diacid chloride
[h] is *N,N'*-(4,4'-oxydiphthaloyl)-bis-*L*-leucine
[i] is *N,N'*-(4,4'-oxydiphthaloyl)-bis-L-methionine diacid chloride
[j] is *N,N'*-(4,4'-oxydiphthaloyl)-bis-(s)-(+)-valine diacid chloride
[k] is *N,N'*-(pyromellitoyl)-bis-L-leucine diacid chloride
[l] is *N,N'*-(4,4'-carbonyldiphthaloyl)-bis-L-leucine diacid chloride

### 2.3. Variable selection using genetic algorithm

Genetic algorithms are adaptive heuristic search algorithms that can be applied when the dimension of the data space is too large for an exhaustive search. They have been proved to be an efficient method in the feature selection problems [32,33]. GAs have several advantages in comparison with other optimization algorithms. They have the ability to move from local optima present on the response surface. They require no knowledge or gradient information about the response surface and can be employed for a wide variety of optimization problem [34]. The major drawbacks to GA are that, there can be difficulties in finding the exact global optimum, which requires a large number of response (fitness) function evaluations and configuring the problem is not straightforward [35]. There are some basic steps in genetic algorithms as follow: (1) a chromosome is represented by a binary bit string and an initial population of chromosomes is created in a random way; (2) a value for the fitness function of each chromosome is evaluated; (3) according to the values of fitness function, the chromosomes of the next generation are reproduced by selection, crossover and mutation operations. In this paper, GA program was written with MATLAB 7.0 [36] and based on Leardi's method [37] with a few minor modifications in our laboratory. The size of population is 30, the probability of cross over is 0.5, the probability of mutation is 0.01 and the number of evaluation is 200. For each set of data 100 runs were performed. Here, we try to use varieties of fitness functions which are proportional to the residual error of the training set, test set and the number of selected variables according to the following equation:

$$\text{fitness} = \frac{1}{\text{SEC} + \text{SEP} + (m)^w} \tag{2}$$

In this equation, SEC and SEP are standard error of calibration (training) and test set, respectively; $m$ is the number of variables in the represented model and $w$ is a numerical value that implies the weights of $m$ in the value of fitness. In fact, the value of $w$ determine the number of variables consist in selected chromosome. Some experiments were done using different value for $w$. Acquired results showed that for small value of w, the number of variables in the fittest individual was high and on the other hand if the value of $w$ was to be high, the number of variables in the best chromosome was small. Hence, after some experiments the value of $w$ was set to be 0.3. It is worth noting that the parameter of $w$ was determined in a preliminary study, before the overall genetic algorithm optimization has been carried out.

**Table 2**
Specification of multiple linear regression model.

| Descriptor | Notation | Coefficient | Mean effect | VIF |
|---|---|---|---|---|
| Molecular weight of repeating unit | MW | 0.001 | 0.781 | 6.605 |
| Connectivity index chi-3 | $^3$X | −0.052 | −1.036 | 8.785 |
| 2nd component accessibility directional WHIM index/ weighted by atomic van der Waals volumes | E2V | 1.656 | 0.572 | 1.223 |
| distance/detour ring index of order 6 | DDR06 | −0.093 | −0.110 | 3.238 |
| Balaban index | J | −10.497 | −0.150 | 1.172 |
| Constant | | 0.339 | | |

### 2.4. Multiple linear regression (MLR)

Multiple linear regression is common method used in QSPR study. Equation linking the structural features to the $\eta_{inh}$ is developed with the form:

$$\eta_{inh} = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n \tag{3}$$

where $a_0$ is the intercept and $a_1$, $a_2$, …, $a_n$ are the regression coefficients of the descriptors. The descriptors ($x_1$, $x_2$, …, $x_n$) included in the equation are used to describe chemical structure of compounds and $n$ is the number of the descriptors to find the best regression model. The main goal of the generation of the MLR model was to choose a set of suitable descriptors that can be used as inputs for generation of the ANN model. A stepwise procedure was used for selection of descriptors. This method combines the forward and backward procedures. Due to the complexity of inter-correlations, the variance explained by certain variables will change when new variables enter the equation. Sometimes a variable that is qualified to enter loses some of its predictive validity when other variables enter. If this takes place, the stepwise method will remove the weakened variable. A final set of selected equations was then tested for stability and validity through a variety of statistical methods. The choice of equation suitable for further consideration was made by using four criteria, namely, multiple correlation coefficients ($R$), standard error (S.E.), $F$-statistic and the number of descriptors in the model. The orthogonality of the descriptors in the model was established through variance inflation factor (VIF) [38,39]. The VIF is defined as $1/(1-R_i^2)$ where $R_i$ is the multiple coefficient of determination in a regression of the $i$th predictor on all other predictors. A VIF value larger than 10 indicates that the information of the descriptors may be hidden by the correlation of the other descriptors. The best multiple linear regression (MLR) model is one that has high $R$ and $F$-values, low standard error, least number of descriptors and high ability for prediction. The statistical characteristics of the best MLR model are shown in Table 2. The orthogonality of the descriptors (VIF) in the MLR model is in agreement with the limit.

### 2.5. Artificial neural network

Artificial neural networks (ANNs) are basically a data-driven black-box model capable of solving highly non-linear complex

problems. They have the ability to capture the relationship between input and output variables from given patterns (historical data or measured data on input and output variables of the system of the concern) and this enables them to solve large-scale complex problems. The network learns basically by finding the optimal network-connection-weights that would generate an output vector as close as possible to the target values of the output vector, with the selected accuracy. A detailed description of the theory behind a neural network has been adequately described elsewhere [37,40,41]. The program for the feedforward neural network that was trained by backpropagation algorithm was written with MATLAB 7 in our laboratory. This network has five nodes in the input layer and one node in the output layer. Descriptors that appeared in the selected MLR model were used as inputs for the generated ANN and its output was the inherent viscosity for the molecules of interest. The number of nodes in the hidden layer would be optimized. The initial weights were randomly selected 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 rates and momentum, then the network was trained using the training set to optimize the values of weights and biases. Finally in order to evaluate the prediction power of the ANN, trained network was employed to calculate the inherent viscosity for the external validation set.

### 2.6. Estimation of the predictive ability of a QSPR model

For the optimized QSPR model several parameters were selected to test prediction ability of the model. A real QSPR model may have a high predictive ability, if it is close to ideal one. This may imply that the correlation coefficient R between the experimental (actual) y and predicted $\tilde{y}$ properties must be close to 1 and regression of y against $\tilde{y}$ or $\tilde{y}$ against y through the origin, i.e. $y^{r0} = k\tilde{y}$ and $\tilde{y}^{r0} = k'y$, respectively, should be characterized by at least either k or $k'$ close to 1 [42]. Slopes k and $k'$ are calculated as follows:

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \tag{4}$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \tag{5}$$

The criteria formulated above may not be sufficient for a QSPR model to be truly predictive. Regression lines through the origin defined by $y^{r0} = k\tilde{y}$ and $\tilde{y}^{r0} = k'y$(with the intercept set to one) should be close to optimum regression lines $y^r = a\tilde{y} + b$ and $\tilde{y}^r = a'y + b'$ (b and b′ are intercepts). Correlation coefficients for these lines $R_0^2$ and $R_0'^2$ are calculated as follows:

$$R_0^2 = 1 - \frac{\sum \left(\tilde{y}_i - y_i^{r0}\right)^2}{\sum \left(\tilde{y}_i - \bar{\tilde{y}}\right)^2} \tag{6}$$

$$R_0'^2 = 1 - \frac{\sum \left(y_i - \tilde{y}_i^{r0}\right)^2}{\sum (y_i - \bar{y})^2} \tag{7}$$

where $\bar{y}$ and $\bar{\tilde{y}}$ are the average values of the observed and predicted properties, respectively and the summations are over all n compounds in the validation set.

A difference between $R_2$ and $R_0^2$ values ($R_m^2$) needs to be studied to explore the prediction potential of a model [43]. This term was defined in the following manner:

$$R_m^2 = R^2\left(1 - \left|\sqrt{R^2 - R_0^2}\right|\right) \tag{8}$$

Finally, the following criteria for evaluation of the predictive ability of QSPR models should be considered:

1. High value of cross-validated $R^2$ ($q^2 > 0.5$).
2. Correlation coefficient R between the predicted and actual properties from an external test set close to 1. $R_0^2$or $R_0'^2$ should be close to $R^2$.
3. At least one slope of regression lines (k or k′) through the origin should be close to 1.
4. $R_m^2$should be greater than 0.5.

## 3. Result and discussion

Table 1 shows the data set and corresponding observed MLR and ANN predicted values of inherent viscosity of all polymers studied in this work. It can be seen from Table 2 that five descriptors appeared in the MLR model. These descriptors are: molecular weight (MW), Randic index order 3 ($^3\chi$), 2nd component accessibility directional WHIM index/weighted by atomic van der Waals volumes (E2V), distance/detour ring index of order 6 (DDR06) and Balaban index (J). The numerical values of these descriptors are shown in Table 3. Table 4 represents the correlation matrix for these descriptors. By interpreting the descriptors in the models, it is possible to gain some insight into factors that are likely related to inherent viscosity of polymers.

The first descriptor described here is MW, which is a constitutional descriptor. This simple descriptor reflects only the molecular composition of the compound without using the geometry or electronic structure of the molecule. The second descriptor is “2nd component accessibility directional WHIM index/weighted by atomic van der Waals volumes (E2V)”. WHIM descriptors are the molecular descriptors based on statistical indices calculated on the projections of the atoms along principal axes [44]. They are built in such a way to capture relevant molecular 3-dimensional information regarding to the molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. These indices are calculated from (x, y, z)-coordinates of a molecule within different weighting schemes in a straightforward manner and represent a very general approach to describe molecules in a unitary conceptual framework. A detailed description of their chemical meaning and of the WHIM theory is reported elsewhere [45]. As it shown in Table 2 these two descriptors have positive signs for their effects, which reveals that by increasing the values of these descriptors, the values of $\eta_{inh}$ increase. The next descriptor is Randic index order 3 ($^3\chi$), a topological descriptor. This descriptor is defined by the following formula:

$$^3\chi = \sum_{path} (D_i D_j \ldots D_k)^{-1/2} \tag{9}$$

where $D_i$ and $D_j$ are the edge degrees (atom connectivities) of the molecular graph. Topological descriptors (also called topological indices) describe the atomic connectivity in the molecule [46–48]. The forth descriptor is Balaban index ($J$). This descriptor is defined by the following formula:

$$J = \left(\frac{q}{\mu + 1}\right) \sum_{i,j}^{q} (S_i S_j)^{-\frac{1}{2}} \tag{10}$$

**Table 3**
The values of the descriptors that were used in this work.

| Number[a] | MW | $^3$X | E2V | DDR06 | J |
|---|---|---|---|---|---|
| 1 | 981.01 | 24.529 | 0.298 | 1.048 | 0.016 |
| 2 | 1190.99 | 2.860 | 0.329 | 2.346 | 0.038 |
| 3 | 996.96 | 2.199 | 0.281 | 2.231 | 0.000 |
| 4 | 1057.11 | 23.012 | 0.507 | 0.912 | 0.033 |
| 5 | 1073.11 | 23.162 | 0.523 | 0.902 | 0.034 |
| 6 | 1071.14 | 27.162 | 0.529 | 0.902 | 0.034 |
| 7 | 862.9 | 21.400 | 0.382 | 1.198 | 0.011 |
| 8 | 852.88 | 21.201 | 0.380 | 1.192 | 0.013 |
| 9 | 820.81 | 21.098 | 0.286 | 1.264 | 0.013 |
| 10 | 884.88 | 22.400 | 0.374 | 1.198 | 0.013 |
| 11 | 952.93 | 17.877 | 0.434 | 0.985 | 0.010 |
| 12 | 862.9 | 21.545 | 0.412 | 1.191 | 0.015 |
| 13 | 874.81 | 22.726 | 0.352 | 1.137 | 0.011 |
| 14 | 794.77 | 19.743 | 0.340 | 1.259 | 0.021 |
| 15 | 820.81 | 20.197 | 0.350 | 1.202 | 0.018 |
| 16 | 786.75 | 18.779 | 0.337 | 1.409 | 0.012 |
| 17 | 750.83 | 18.886 | 0.349 | 1.616 | 0.025 |
| 18 | 832.88 | 22.347 | 0.174 | 1.185 | 0.013 |
| 19 | 872.85 | 24.612 | 0.259 | 1.132 | 0.018 |
| 20 | 882.92 | 19.545 | 0.168 | 1.191 | 0.013 |
| 21 | 742.75 | 20.714 | 0.257 | 1.425 | 0.018 |
| 22 | 832.88 | 19.201 | 0.187 | 1.192 | 0.013 |
| 23 | 742.75 | 20.568 | 0.333 | 1.436 | 0.017 |
| 24 | 818.85 | 23.052 | 0.306 | 1.210 | 0.016 |
| 25 | 742.75 | 20.478 | 0.326 | 1.440 | 0.015 |
| 26 | 756.78 | 17.140 | 0.451 | 1.134 | 0.010 |
| 27 | 780.79 | 20.009 | 0.306 | 1.414 | 0.012 |
| 28 | 888.96 | 22.642 | 0.402 | 1.178 | 0.012 |
| 29 | 920.96 | 21.840 | 0.351 | 1.185 | 0.012 |
| 30 | 748.89 | 21.514 | 0.297 | 1.040 | 0.012 |
| 31 | 700.82 | 20.315 | 0.319 | 1.026 | 0.013 |
| 32 | 608.72 | 17.682 | 0.303 | 1.209 | 0.014 |
| 33 | 684.82 | 20.165 | 0.294 | 1.041 | 0.013 |
| 34 | 818.9 | 19.991 | 0.416 | 0.881 | 0.011 |
| 35 | 610.68 | 16.828 | 0.435 | 1.199 | 0.012 |
| 36 | 718.85 | 19.461 | 0.455 | 1.019 | 0.012 |
| 37 | 750.85 | 20.659 | 0.432 | 1.033 | 0.012 |
| 38 | 624.71 | 17.537 | 0.455 | 1.211 | 0.010 |
| 39 | 784.97 | 18.954 | 0.264 | 1.027 | 0.013 |
| 40 | 720.83 | 20.451 | 0.385 | 1.023 | 0.013 |
| 41 | 580.66 | 16.620 | 0.286 | 1.195 | 0.014 |
| 42 | 580.66 | 16.529 | 0.280 | 1.200 | 0.013 |
| 43 | 690.71 | 20.685 | 0.302 | 0.995 | 0.014 |
| 44 | 728.81 | 21.750 | 0.314 | 0.996 | 0.015 |
| 45 | 820.81 | 23.584 | 0.267 | 0.873 | 0.014 |
| 46 | 820.81 | 22.584 | 0.264 | 0.871 | 0.014 |
| 47 | 740.77 | 22.715 | 0.278 | 0.925 | 0.014 |
| 48 | 710.86 | 20.145 | 0.390 | 1.042 | 0.010 |
| 49 | 634.76 | 18.083 | 0.383 | 1.233 | 0.016 |
| 50 | 760.9 | 21.344 | 0.423 | 1.054 | 0.013 |
| 51 | 620.73 | 17.422 | 0.367 | 1.234 | 0.016 |
| 52 | 1121.18 | 30.360 | 0.495 | 0.901 | 0.030 |
| 53 | 758.74 | 21.140 | 0.351 | 1.434 | 0.010 |
| 54 | 744.71 | 17.714 | 0.433 | 1.425 | 0.012 |
| 55 | 818.85 | 21.197 | 0.286 | 1.202 | 0.020 |
| 56 | 834.85 | 23.201 | 0.294 | 1.192 | 0.015 |
| 57 | 782.77 | 19.918 | 0.411 | 1.419 | 0.012 |
| 58 | 608.72 | 17.592 | 0.371 | 1.214 | 0.013 |
| 59 | 612.66 | 16.737 | 0.382 | 1.203 | 0.012 |
| 60 | 1289.58 | 31.490 | 0.266 | 0.710 | 0.009 |
| 61 | 656.76 | 19.103 | 0.403 | 1.025 | 0.012 |
| 62 | 898.93 | 27.849 | 0.323 | 0.786 | 0.015 |
| 63 | 712.83 | 20.145 | 0.431 | 1.042 | 0.015 |
| 64 | 90.16 | 2.199 | 0.264 | 2.231 | 0.012 |
| 65 | 744.71 | 18.568 | 0.399 | 1.436 | 0.013 |
| 66 | 874.81 | 19.612 | 0.289 | 1.135 | 0.013 |
| 67 | 846.86 | 23.017 | 0.272 | 1.188 | 0.015 |
| 68 | 743.74 | 19.623 | 0.311 | 1.429 | 0.017 |
| 69 | 989.01 | 21.172 | 0.294 | 0.980 | 0.010 |
| 70 | 698.85 | 20.315 | 0.351 | 1.026 | 0.012 |
| 71 | 728.87 | 16.659 | 0.479 | 1.033 | 0.010 |
| 72 | 1289.58 | 31.490 | 0.266 | 0.710 | 0.009 |
| 73 | 672.76 | 19.253 | 0.385 | 1.010 | 0.013 |
| 74 | 808.9 | 24.517 | 0.349 | 0.891 | 0.015 |
| 75 | 620.73 | 17.512 | 0.430 | 1.230 | 0.018 |

The definitions of the descriptors are given in Table 2.
[a] The numbers refer to the numbers of the molecules given in Table 1.

where $q$ is the number of edges in the molecular graph, $\mu$ is the cyclometric number and $S_i$ and $S_j$ are the distance sums (or distance degrees), obtained by summation the row $i$ and column $i$ (or row $j$ and column $j$, respectively) of the distance matrix between atoms in the molecule. The last descriptor that is presented here is distance/detour ring index of order 6 (DDR06). This descriptor is also a topological descriptor. Topological descriptors are derived entirely from 2D structural formulas and. Therefore, missing parameters, conformational flexibility, or molecular alignment do not have to be taken into account. Topological descriptors can be easily calculated from molecular graphs in which the atoms are represented by vertices and the bonds by edges. The connections between the atoms can be described by various types of topological matrices, which can be mathematically manipulated so as to derive a single number, usually known as graph invariant, graph-theoretical index or topological Index.

These three descriptors have negative signs for their effects, which reveal that by increasing the values of these descriptors, the values of $\eta_{inh}$ decrease.

From the above discussion, it can be seen that all descriptors involved in the QSPR model have physically meaning, and these descriptors can account for structural features that affect the inherent viscosity of the interested polymers.

The next step was the construction of an artificial neural network. During the training of the ANN, the parameters of network including the number of nodes in the hidden layer, weights and biases learning rates and momentum values were optimized. Table 5 shows the architecture and specification of the optimized network. After optimization of the network parameters, the network was trained by using training set for adjustment of the weights and biases values by backpropagation 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. 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 1000 training iterations, the network was used to calculate $\eta_{inh}$ 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. Results obtained showed overtraining began after 26000 iterations.

The predictive power of the ANN models developed on the selected training sets are estimated on the predictions of validation set chemicals, by calculating the $q^2$ that is defined as follow:

**Table 4**
Correlation matrix between selected descriptors.

| | MW | $^3$X | E2V | DDR06 | J |
|---|---|---|---|---|---|
| MW | 1 | 0.894 | 0.239 | −0.729 | 0.184 |
| $^3$X | | 1 | 0.110 | −0.803 | 0.097 |
| E2V | | | 1 | −0.219 | 0.308 |
| DDR06 | | | | 1 | 0.072 |
| J | | | | | 1 |

**Table 5**
Architecture and specifications of optimized ANN model.

| | |
|---|---|
| Number of nodes in the input layer | 5 |
| Number of nodes in the hidden layer | 5 |
| Number of nodes in the output layer | 1 |
| Weights learning rate | 0.2 |
| Biases learning rate | 0.1 |
| Momentum | 0.5 |
| Transfer function | Sigmoid |

**Table 6**
Statistical parameters obtained using the ANN and MLR models.[a]

| Model | $SE_c$ | $SE_t$ | $SE_v$ | $R_c$ | $R_t$ | $R_v$ | $F_c$ | $F_t$ | $F_v$ |
|---|---|---|---|---|---|---|---|---|---|
| ANN | 0.023 | 0.030 | 0.031 | 0.994 | 0.989 | 0.991 | 3972 | 429 | 522 |
| MLR | 0.079 | 0.077 | 0.088 | 0.924 | 0.923 | 0.922 | 286 | 58 | 56 |

[a] c refers to the calibration (training) set; t refers to test set; v refers to validation set; R is the correlation coefficient; SE is standard error and F is the statistical F value.

$$q^2 = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum(y_i - \overline{y})^2} \quad (11)$$

where $y_i$ and $\hat{y}_i$, respectively are the measured and predicted values of the dependent variable (inherent viscosity), $\overline{y}$ is the averaged value of dependent variable of the training set and the summations cover all the compounds. The calculated value of $q^2$ was 0.978.

Table 1 represents the experimental, MLR and ANN calculated values of inherent viscosity for the training, test and validation sets. The statistical parameters obtained by ANN and MLR models for these sets are shown in Table 6. The standard errors of training, test and validation sets for the MLR model are 0.079, 0.077, and 0.088, respectively which would be compared with the values of 0.023, 0.030, and 0.031, respectively, for the ANN model. Comparison between these values and other statistical parameters in Table 6 reveals the superiority of the ANN model over MLR one. The key strength of neural networks, unlike MLR analysis, is their ability to flexible mapping of the selected features by manipulating their functional dependence implicitly.

The statistical values of validation set for the ANN model was characterized by $q^2 = 0.978$, $R^2 = 0.982$ ($R = 0.991$), $R_0^2 = 0.978$, $R_m^2 = 0.921$ and $k = 0.993$. These values and other statistical parameters which are shown in Table 6 reveal the high predictive ability of the model. Fig. 1 shows the plot of the ANN predicted *versus* experimental values for inherent viscosity of all of the molecules in data set. The residuals of the ANN calculated values of the inherent viscosity are plotted against the experimental values in Fig. 2. The propagation of the residuals in both sides of zero line indicates that no systematic error exists in the constructed QSPR model.

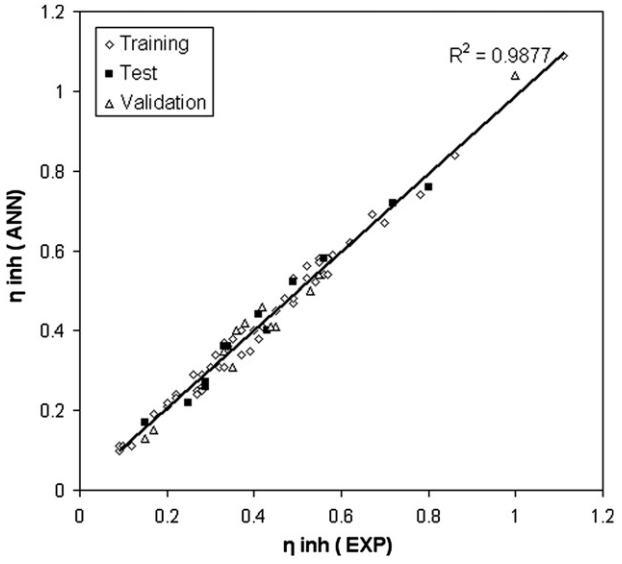


**Fig. 1.** Plot of ANN calculated inherent viscosity against experimental values.

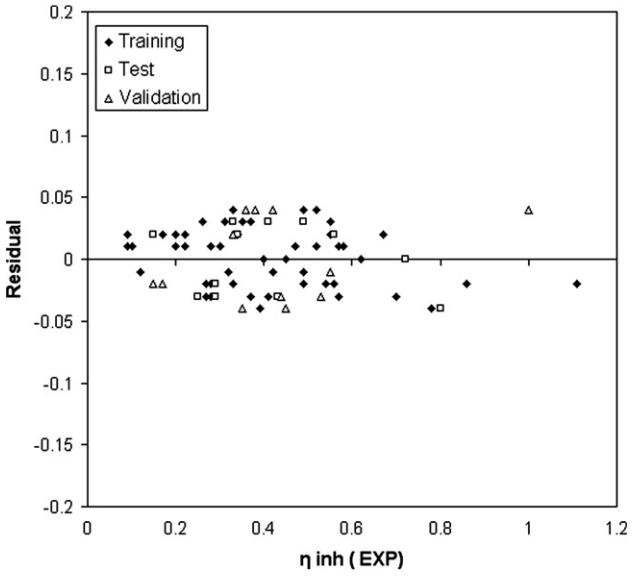


**Fig. 2.** Plot of residual *versus* experimental values of inherent viscosity.

## 4. Conclusions

In the present work GA as a feature selection tool and MLR and ANN as feature mapping techniques were used for prediction of the inherent viscosity of 75 optically active polymers. The optimized 5-5-1 ANN model showed a remarkable improvement over the linear model. The GA-based MLR approach is especially useful for modeling a large variable data set. The physical meaning of the selected subset of descriptors, which are the most predictive and informative, from the GA method, is determined. The inherent viscosities of investigated polymers were interpreted rationally with these five descriptors. Result obtained indicate that while the GA and MLR method could be more powerful in precise selecting of important parameters and assume the significance of each of descriptors, introduction of neural network gives a significant improvement of prediction quality.

## Acknowledgements

We wish to express our gratitude to the Research Affairs Division Isfahan University of Technology (IUT), Isfahan, for partial financial support. Further financial support from National Elite Foundation (NEF) and Center of Excellency in Sensors and Green Chemistry Research (IUT) is gratefully acknowledged. The author also acknowledges Professor T. Khayamian for his helpful discussion.

## References

[1] Mallakpour S, Taghavi M. Polymer 2008;49:3239–49.
[2] Mallakpour S, Rafiee Z. Polymer 2008;49:3007–13.
[3] Mallakpour S, Rafiee Z. Polymer 2007;48:5530–40.
[4] Farina M. Top Stereochem 1987;17:1–111.

[5] Mallakpour SE, Hajipour AR, Vahabi R. J Appl Polym Sci 2002;84:35–43.
[6] Chen J, Nan Q, Guo L, Zhou Y. J Appl Polym Sci 2010;115(4):2190–6.
[7] Yu X, Wang X, Li X, Gao J, Wang H. Macromol Theory Simul 2006;15:94–9.
[8] Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O. Polymer 2006;47:3240–8.
[9] Liu W, Yi P, Tang Z. QSAR Comb Sci 2006;25(10):936–43.
[10] Yu X, Yi B, Liu F, Wang X. React Funct Polym 2008;68:1557–62.
[11] Jalali-Heravi M, Fatemi MH. Anal Chem Act 2000;415:95–103.
[12] Yao X, Zhang X, Zhang R, Liu M, Hu Z, Fan B. Comput Chem 2001;25:475–82.
[13] Schweitzer RC, Morris JB. Anal Chim Acta 1999;384:285–303.
[14] Fatemi MH. J Chromatogr A 2002;955:273–80.
[15] Golmohammadi H, Fatemi MH. Electrophoresis 2005;26:3438–44.
[16] Baher E, Fatemi MH, Konoz E, Golmohammadi H. Microchim Acta 2007;158:117–22.
[17] Konoz E, Golmohammadi H. Anal Chim Acta 2008;619:157–64.
[18] Golmohammadi HJ. Comput Chem 2009;30:2455–65.
[19] Golmohammadi H, Konoz E, Dashtbozorgi Z. Anal Sci 2009;25:1137–42.
[20] Mallakpour SE, Hajipour A, Khoee S. Eur Polym J 2002;38:2011–6.
[21] Mallakpour S, Moghaddam E. Iran Polym J 2006;15(7):547–54.
[22] Mallakpour SE, Hajipour A, Khoee S. J Appl Polym Sci 2000;77:3003–9.
[23] Mallakpour SE, Hajipour A, Khoee S. Polym Int 1999;48:1133–40.
[24] Mallakpour S, Kowsari E. Polym Bull 2006;57:169–78.
[25] Mallakpour S, Kowsari E. Polym Adv Technol 2006;17:174–9.
[26] Mallakpour S, Kowsari E. Iran Polym J 2005;14(9):799–806.
[27] Mallakpour S, Kowsari E. Iran Polym J 2006;15(6):457–65.
[28] Mallakpour S, Habibi S. Eur Polym J 2003;39:1823–9.
[29] Mallakpour SE, Hajipour A, Zamanlou MR. J Polym Sci 2001;39:177–86.
[30] Ohlenbusch G, Frimmel FH. Chemosphere 2001;45:323–7.
[31] Hyperchem, re. 4. for Windows, Autodesk, Sansalito, CA; 1995.
[32] Leardi R, Boggia R, Terrile M. J Chemom 1992;6:267–81.
[33] Leardi R, Gonzalez AL. Chemom Intell Lab Syst 1998;41:195–207.
[34] Chambers L. Practical handbook of genetic algorithms. Lewis Publishing; 1995.
[35] Hibbert DB. Chemom Intell Lab Syst 1993;19:277–93.
[36] MATLAB 7.0, http://www.mathworks.com.
[37] Blank TB, Brown ST. Anal Chem 1993;65:3081–9.
[38] Chatterjee S, Hadi A, Price B. Regression analysis by examples. 3rd ed. New York: Wiley-VCH; 2000.
[39] Shapiro S, Guggenheim B. Quant Struct Act Relat 1998;17:327–37.
[40] Beal TM, Hagan HB, Demuth M. Neural network design. Boston: PWS; 1996.
[41] Zupan J, Gasteiger J. Neural networks for chemists: an introduction. Weinheim: VCH; 1993.
[42] Golbraikh A, Tropsha A. J Mol Graphics Model 2002;20:269–76.
[43] Roy PP, Roy K. QSAR Comb Sci 2008;27:302–13.
[44] Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim, Germany: Wiley-VCH; 2000.
[45] Todeschini R, Gramatica P. Quant Struct Act Rel 1997;16:113–9.
[46] Kowalski BR, editor. Chemometrics. Dordrecht: Reidel; 1984.
[47] Stankevich MI, Stankevich IV, Zefirov NS. Russ Chem Rev 1988;57:191–208.
[48] El-Basil S, Randic M. Adv Quant Chem 1992;24:239–90.