Contents lists available at SciVerse ScienceDirect

Journal of the Taiwan Institute of Chemical Engineers

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

A molecular-based model for prediction of liquid viscosity of pure organic compounds: A quantitative structure property relationship (QSPR) approach

Farhad Gharagheizi [a,*], Seyyed Alireza Mirkhani [a], Mohammad Hossein Keshavarz [b], Nasrin Farahani [c], Kaniki Tumba [d]

[a] *Department of Chemical Engineering, Islamic Azad University, Buinzahra Branch, Buinzahra, Iran*
[b] *Department of Chemistry, Malek-ashtar University of Technology, Shahin-shahr P.O. Box 83145/115, Isfahan, Iran*
[c] *Department of Chemistry, Islamic Azad University, Buinzahra Branch, Buinzahra, Iran*
[d] *Department of Chemical Engineering, Mangosuthu University of Technology, Durban, South Africa*

ARTICLE INFO

Article history:
Received 6 July 2012
Received in revised form 14 December 2012
Accepted 19 December 2012
Available online 14 February 2013

Keywords:
Genetic algorithms
Subset variable selection
Feed forward neural networks
Quantitative structure–viscosity relationship

ABSTRACT

In this study, a new quantitative structure–property relationship (QSPR) is presented for prediction of the liquid viscosity of pure organic compounds. The model implements eight molecular descriptors selected using the genetic algorithm-based multivariate linear regression (GA-MLR) from more than various 3000 molecular descriptors to predict the liquid viscosity. To propose a comprehensive and predictive model, 2748 pure organic compounds are investigated. Furthermore, several statistical methods are applied to evaluate the predictive capability of the model. The root mean square of error and the average absolute percent error of the model are equal to 0.34 and 7%, respectively.

1. Introduction

Viscosity is one of the most important physical properties of liquids which has a broad spectrum of applications in science and engineering. From molecular point of view, it provides fundamental information about intermolecular forces in liquids. Although, a large number of experimental data have been reported for viscosity of diverse compounds in the literature, there is no available experimental data for many desired systems and conditions. Furthermore, experimentally measurement of viscosity is laborious, time consuming, and needs especial instruments. As a result, it is important to develop an accurate model to predict viscosity.

Many models have been reported in the literature for viscosity prediction. A good review of theoretical, empirical and semi empirical methods could be find elsewhere [1].

In recent years, new predictive approaches based on molecular structure such as quantitative structure–property relationships (QSPR) and group contribution methods (GC) grasp the attention of many researchers.

Cocchi *et al.* [2] studied a series of organic compounds of structural formula X-CH_2CH_2-Y (X, and Y: alkyls, aromatics, halogens, electron acceptors and donors, and hydrogen-bonding groups) in order to develop a QSPR model for prediction of their liquid viscosity at 25 °C. For this purpose, they investigated the log-based data for 46 pure compounds (37 compounds for training set and 9 compounds for test set). They developed several models by applying several methods existed in CODESSA software. Finally, they concluded that their partial least square-based model is the best among all the other applied methods. Their model contains 16 descriptors. The major drawback of their model is to apply a large number of parameters (16 descriptors) to predict the viscosity of a small number of compounds (totally 46 compounds). Despite of a considerable number of descriptors, their model do not show an excellent squared correlation coefficient ($R^2 = 0.9497$ over log-based data). Besides, their model cannot be generalized because it has obtained based on a small number of compounds.

In another survey, with CODESSA software, Ivanciuc *et al.* [3] developed a QSPR model for viscosity of liquids at 20 °C. In their study, 337 compounds from various classes of organic compounds were implemented. Their linear model contained five parameters and had a root mean square error and squared correlation coefficient (R^2) of 0.37 and 0.85 ($R = 0.920$), respectively, based on log-based data. The main limitation of their model is its

[*] Corresponding author. Fax: +98 21 77926580.
E-mail addresses: fghara@ut.ac.ir, fghara@gmail.com (F. Gharagheizi).

1876-1070/$ – see front matter © 2012 Taiwan Institute of Chemical Engineers. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jtice.2012.12.015

restriction to predict viscosity of liquids with more than one polar group.

Utilizing a data set of liquid viscosity of 361 structurally diverse organic compounds at 20 °C, Katritzky *et al.* [4] attempted to develop a QSPR model, again applying CODESSA software. Their final five-parameter model was obtained with a squared correlation coefficient (R^2) of 0.854 and a standard error (*SE*) of 0.22 log units. In addition, they tried to enhance the precision of their models by using Artificial Neural Networks (ANN). However, their effort did not lead to any better model.

Kauffman *et al.* [5] proposed an eight-descriptor QSPR model to estimate liquid viscosity. In their study, 212 compounds from various organic groups with the viscosity range of 0.215–944.06 mPa. s were included. Approximately 90% of the compounds were assigned to train the model. In order to subset variable selection, they applied simulated annealing method coupled with multiple linear regression (MLR) [6]. They considered models composed of 4–12 descriptors. Their final models from (multiple linear regression) MLR analysis were then submitted to three layers, fully connected, feed-forward (computational neural networks) CNNs. Finally, they concluded that an eight-parameter random forest tree can well describe the data. One of the drawbacks of the proposed model is that the viscosity predictions for compounds with two or three hydroxyl groups have a moderate accuracy.

Providing a liquid viscosity experimental data set for 403 compounds, Rajappan *et al.* [7] developed an eight-parameter model by applying the robust random forest regression algorithm (RF). Their model showed a high value for squared correlation coefficient ($R^2 = 0.98$). According to Liaw and Wiener [8] the maximum number of trees in RF method should be no more than one third of the data. Therefore, the number of trees should be no more than 135 (403/3), whereas Rajappan *et al.* [7] applied 500 trees in their model. It seems that some sort of over-fitting has happened in the study.

The group contribution type methods (GC) are another class of approaches applied to predict liquid viscosity [9–16]. In this category, Morejon [11] developed a method based on the molecular weight and group interactions contributions to predict liquid viscosity at 20 °C. He selected 35 simple groups to generate a consistent set of group interactions to treat a wide class of organics including 230 compounds. The model showed squared correlation coefficient (R^2) and standard deviation error of 0.957 and 2.474 (log-based data), respectively. The main drawback of the method is that the prediction of viscosity values for those compounds not available in the training set is rather high (12.7%). This demonstrates lack of predictive power for the model.

The main drawback of the proposed methods based on group interaction or group contribution is that they are incapable for prediction of viscosity of compounds containing new chemical groups.

In this study, a new comprehensive QSPR model is presented for prediction of liquid viscosity of pure organic compounds at 25 °C. To develop the model, a comprehensive experimental data set including more than 2700 compounds is investigated.

2. Development of model

2.1. Data set preparation

The accuracy and reliability of models for prediction of physical properties, especially the molecular-based ones, directly depends on the quality and comprehensiveness of the applied data set for their development. The aforementioned characteristics of such a model include both diversity in the investigated chemical families and the number of pure compounds available in the data set. In this study, the database prepared by Yaws [17] was implemented, which is one of the most comprehensive sources of physical property data for chemical species *e.g.* liquid viscosity. The liquid viscosity of 2748 compounds found in the database and used in this study. The chemical structures of the treated chemical compounds are presented as supplementary materials.

2.2. Determination of molecular descriptors

In this step, the molecular structures of 2748 compounds were drawn in to Hyperchem software [18] and optimized using MM+ molecular mechanics force field. Thereafter, the dragon software [19] was employed to calculate molecular descriptors from those optimized structures. Dragon software is capable of computing more than 3000 molecular descriptors for each molecule. The detailed information about the types of molecular descriptors and the procedure of their calculation in Dragon could be found elsewhere in the user's guide of the software. [19]

2.3. Model parameter selection

Advancement of QSPR theory in recent years makes a lot of molecular descriptors available to develop predictive models. Therefore, there are much more descriptors than required to develop the model. In such circumstances, it is required to select an optimal subset of descriptors that can describe the property under consideration, well. There are several methods applied for this purpose. A generally accepted method is the genetic algorithm-based multivariate linear regression (GA-MLR). In the method, at first, a linear correlation between the descriptors and the property is deemed. As a powerful search tool, the genetic algorithm is implemented to select the optimal subset of molecular descriptors subjected to an objective containing several mathematical conditions to avoid chance correlations or some other statistical problems. This method was presented by Leardi *et al.* [20] for the first time.

Many standard fitness functions such as R^2, adjusted R^2, Q^2, Akaike information content, LOF function and so on, could be used as objective function in GA-MLR technique [21].

RQK proposed as a fitness function for model searching to avoid unwanted model properties, such as chance correlation, presence of noisy variables in the models, and other model's defects, which have a negative impact on the prediction capability of the model. [21]

RQK is a constrained fitness function based on Q^2_{LOO} statistics (leave-one-out cross validated variance) and other four tests that must be fulfilled contemporarily. Q^2_{LOO} is defined as:

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{ic})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{1}$$

Where y_i is $\ln(\eta_L)$ for *i*th component, $\bar{y}$ is the mean value of $\ln(\eta_L)$ for all components, and $\hat{y}_{ic}$ is response of *i*th object estimated by a model obtained without using the *i*th object.

The above mentioned four constrains were presented by Todeschini *et al.* [21] as below:

$$\Delta K = K_{\mathrm{XY}} - K_{\mathrm{X}} > 0 \quad (\text{quick rule}) \tag{2}$$

$$\Delta Q = Q^2_{\mathrm{LOO}} - Q^2_{\mathrm{ASYM}} > 0 \quad (\text{asymptotic}\, Q^2 \,\text{rule}) \tag{3}$$

$$R^P > 0 \quad (\text{redundancy}\, RP \,\text{rule}) \tag{4}$$

$$R^N > 0 \quad (\text{Over fitting}\, RN \,\text{rule}) \tag{5}$$

since several conditions are evaluated, the validity of the model and its predictive power are assured. In this study, GA-MLR with

RQK fitness function is applied based on satisfactory results in our previous works.

Prior to applying GA-MLR, the data set should be divided into two new sub sets, one for training, and the other one for testing. By means of training set, the best model is found and then, its prediction power is checked by test set. In this work, 80% of the database was assigned to training purposes, and the remaining 20% was assigned to test purposes. The optimal allocation percent of the training set and the test set from the main data set studied elsewhere [22]. The process of the splitting was randomly performed (in each running program, from all 2748 components, 2200 components are in the training set and 548 components are in the test set).

2.4. Artificial neural network algorithm [23–27]

Artificial Neural Networks are extensively used in various scientific and engineering problems *e.g.* estimations of physical and chemical properties of different pure compounds These capable mathematical tools are generally applied to study the complicated systems. The theoretical explanations about Artificial Neural Networks can be found elsewhere [23,28–34].

In this step, in order to investigate the nonlinear relation between the molecular descriptors selected by GA-MLR, a three-layer feed forward artificial neural network (FFANN) is generated. To do so, during developing the model, all the parameters as well as the $\ln(\eta_L)$ values are normalized between -1 and +1 to decrease computational errors. This can be performed using maximum and minimum values of each parameter for input data and using maximum and minimum values of $\ln(\eta_L)$ for output parameter. Later, the main data set is divided into three new sub-data sets including the “Training” set, the “Validation (optimization)” set, and the “Test (prediction)” set. In this work, the “Training” set is used to generate the ANN structure, the “Validation (optimization)” set is applied for optimization of the model, and the “Test (prediction)” set is used to investigate the prediction capability and validity of the obtained model. The process of division of the main data set into three sub-data sets is performed, randomly. For this purpose, about 80%, 10%, and 10% of the main data set are randomly selected for the “Training” set (2200 compounds), the “Validation” set (274 compounds), and the “Test” set (274 compounds). The effect of the allocation percent of the three sub-datasets from the data of main data set on the accuracy of the ANN model has been studied elsewhere [22,23,33,34].

As a matter of fact, generating an ANN model is the determination of the weight matrices and bias vectors. [28] As shown in Fig. 1, there are two weight matrices and two bias vectors in a three layer FFANN: *W* and *b* for hidden layer, and *W*, *b* for the output layer. These parameters should be obtained by minimization of an objective function. The objective function used in this study is the sum of squares of errors between the outputs of the ANN (predicted $\ln(\eta_L)$) and the target values (experimental $\ln(\eta_L)$). This minimization is performed by Levenberg–Marquardt (LM) [28] optimization strategy. There are also more accurate optimization methods other than this algorithm; however, they need much more convergence time. In other words, the more accurate optimization, the more time is needed for the algorithm to be converged. The LM [28] is the most-widely used algorithm for training FFANNs due to being robust and accurate enough to deal with the considered system. In most cases, the number of neurons in the hidden layer (n) is fixed. Therefore, the main goal is to produce a FFANN model, which is able to predict the target values as accurately as expected. This step is repeated till the best FFANN is obtained. Generally and especially in three-layer FFANNs, it is more efficient that the number of neurons in the hidden layer is optimized according to the accuracy of the obtained FFANN.

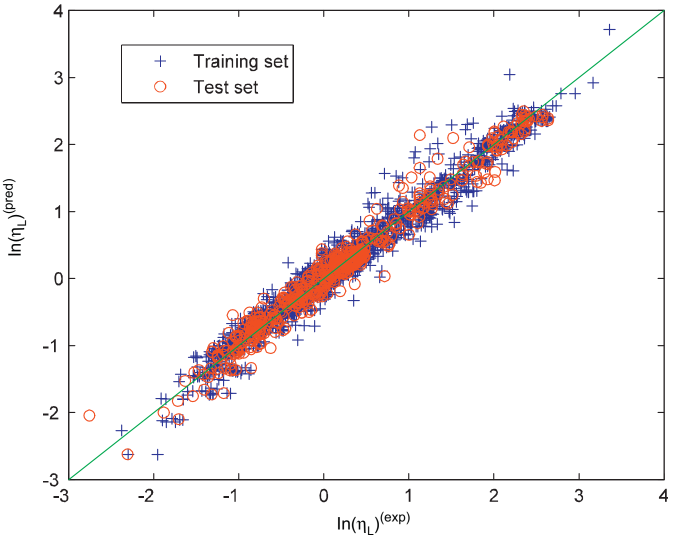

Fig. 1. Comparison between the predicted values by Eq. (6) and their corresponding experimental values of liquid viscosity (log based).

3. Results and discussion

The most accurate multivariate linear equation is obtained following the presented procedure. For obtaining this equation, the best one-molecular descriptor model is obtained at the first place. Later, the best two-molecular descriptor model is determined. This procedure is repeated to achieve the most accurate three-, four-, five- etc. molecular descriptor models. It is found that the most accurate multivariate linear model has eight parameters because further increase in the number of molecular descriptors does not lead to any considerable effects on the accuracy of the model. The final equation and its statistical parameters are presented as follows:

$$\begin{aligned}\ln(\eta_L) = {} & -4.08337(\pm 0.03977) + 0.62496(\pm 0.02312)n\mathrm{TB} \\ & + 0.30546(\pm 0.01146)n\mathrm{R06} \\ & + 1.81395(\pm 0.01965)\mathrm{ATS1m} \\ & - 0.06576(\pm 0.0033)\mathrm{BEHm1} \\ & + 0.06385(\pm 0.00208)\mathrm{RDF040}v \\ & + 1.87398(\pm 0.02224)n\mathrm{ROH} \\ & + 0.40289(\pm 0.02366)\mathrm{H_y} \\ & + 0.19265(\pm 0.01428)\mathrm{B02[C{-}O]} \qquad (6)\end{aligned}$$

the statistical parameters of model are:

$$n_{\mathrm{training}} = 2200;\quad n_{\mathrm{test}} = 548;\quad R^2_{\mathrm{training}} = 0.9675,\quad R^2_{\mathrm{test}} = 0.9697$$

$$Q^2_{\mathrm{LOO}} = 0.9671;\quad Q^2_{\mathrm{EXT}} = 0.9694;\quad Q^2_{\mathrm{LTO}} = 0.9637$$

$$s = 0.178;\quad Q^2_{\mathrm{BOOT}} = 0.9669;\quad a = -0.012;\quad F = 8162.57$$

with RQK function parameters:

$$(\Delta K = 0.069;\quad \Delta Q = 0.000;\quad R^P = 0.031;\quad R^N = 0.000)$$

where η_L is in cP unit. In the preceding equation [35]:

- “nTB” is the number of triple bonds.
- “nR06” is the number of 6-membered rings.
- “ATS1m” is the Broto–Moreau autocorrelation of a topological structure – lag 1/weighted by atomic masses.

- "BEHm1" is the highest eigenvalue *n*. 1 of Burden matrix/weighted by atomic masses.
- "RDF040*v*" is the Radial Distribution Function – 4.0/weighted by atomic van der Waals volumes.
- "*n*ROH" is the number of hydroxyl groups in a molecule. This descriptor is a functional group count.
- "H_y" is called "hydrophilicity" descriptor. It belongs to molecular properties. It is defined as:

$$H_y = \frac{(1+N_{H_y})\times\log_2(1+N_{H_y}) + nC\times((1/nSK)\times\log_2(1/nSK)) + \sqrt{N_{H_y}/nSK^2}}{\log_2(1+nSK)} \qquad (7)$$

where N_{H_y} is the number of hydrophilic groups (OH, SH, and NH groups), *n*C the number of carbon atoms, and *n*SK the number of hydrogen excluded atoms.

- "MLOGP" denotes the Moriguchi descriptors that belong to molecular properties. It is calculated as follows:

$$\begin{aligned}\mathrm{MLOGP} = {} & -1.014 + 1.244(F_{CX})^{0.6} - 1.017(N_O + N_N)^{0.9} \\ & + 0.406F_{PRX} - 0.145{N_{UNS}}^{0.8} + 0.511I_{HB} \\ & + 0.268N_{POL} - 2.215F_{AMP} + 0.912I_{ALK} - 0.392I_{RNG} \\ & - 3.6847F_{QN} + 0.474N_{NO_2} + 1.582F_{NCS} + 0.773I_{\beta L} \qquad (8)\end{aligned}$$

- F_{CX} is the summation of number of carbon and halogen atoms weighted by C = 1.0; F = 0.5; Cl = 0; Br = 1.5; I = 2.0.
- $N_O + N_N$ is the total number of nitrogen and oxygen atoms.
- F_{PRX} is the proximity effect of N/O: 2 for X–Y and 1 for X–A–Y (X, Y: N and/or O; A: C, S, or P; –: saturated or unsaturated bond) with a correction (−1) for −CON< and −SO_2N<.
- N_{UNS} is the total number of unsaturated bonds (not those in NO_2).
- I_{HB} is the dummy variable for the presence of intramolecular hydrogen bond as *ortho* -OH and -CO–R, -OH and -NH_2, -NH_2 and -COOH, or 8-OH/NH_2 in quinolines, 5 or 8–OH/NH_2 in quinoxalines, *etc*.
- N_{POL} is the number of aromatic polar substituents (aromatic substituents excluding Ar–C(X)(Y)– and Ar–C(X)=C; X, Y: C and/or H). Upper limit = 4.
- F_{AMP} is the amphoteric property; α-aminoacid = 1, aminobenzoic acid = 0.5, pyridinecarboxylic acid = 0.5.
- I_{ALK} is the dummy variable for alkane, alkene, cycloalkane, cycloalkene (hydrocarbons with 0 or 1 double bond) or hydrocarbon chain with at least 7 carbon atoms.
- I_{RNG} is the dummy variable for the presence of ring structures except benzene and its condensed rings (aromatic, heteroaromatic, and hydrocarbon rings).
- F_{QN} is the Quaternary nitrogen >N + <: 1; *N*-oxide: 0.5.
- N_{NO_2} is the number of nitro groups.
- F_{NCS} is a parameter that its value for Isothiocyanate (–N=C=S) is 1.0 and for thiocyanate (–S–C#N) is 0.5 (# means triple bond).
- $I_{\beta L}$ is the dummy variable for the presence of β-lactam.

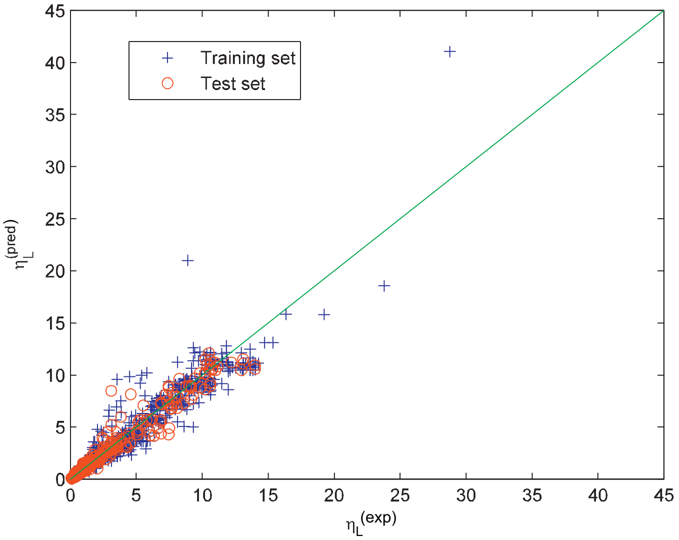

Fig. 2. Comparison between the predicted values by Eq. (6) and their corresponding experimental values of liquid viscosity.

- "B02[C–O]" is the presence/absence (1 or 0) of C–O pair at topological distance 2.

Two out of eight descriptors include "*n*ROH" and "H_y" are directly related to hydrogen bonding interactions which plays key role in the liquid viscosity.

Bootstrap technique, y-scrambling, and external validation techniques are used [35] to more evaluate the validity of the model. The bootstrapping is repeated 5000 times. Also, y-scrambling is repeated 300 times. As can be seen, the difference between Q^2_{LOO}, Q^2_{BOOT}, Q^2_{EXT}, and R^2 shows that the obtained model is a reliable correlation and has high predictive capability [35]. Furthermore, the intercept value of the y-Scrambling technique has low value (a = 0.010) that reveals the validity of the model (The y-scrambling, bootstrapping, and external validation techniques have been extensively presented by Todeschini *et al.* [35]).

The predicted liquid viscosity values by Eq. (6) in comparison with the experimental values [17] are shown in Fig. 1 (log based) and Fig. 2 as well. The predicted liquid viscosity values for the investigated chemical compounds, the calculated descriptors, and the status of all compounds ("Training" or "Test" sets) are presented as supplementary information.

Using the selected molecular descriptors by GA-MLR (in Eq. (6)), a three-layer feed forward neural network was obtained for prediction of the ln(η_L) of pure compounds. To determine the number of neurons in hidden layer of the neural network, numbers 1–50 were tested, so the number 20 showed the best results. Therefore, the best FFANN has a structure of 8-20-1. The mat file (MATLAB file format) of the obtained neural network containing all parameters of the obtained model is freely accessible from the authors by the schematic structure of the best FFANN obtained is shown in Fig. 3.

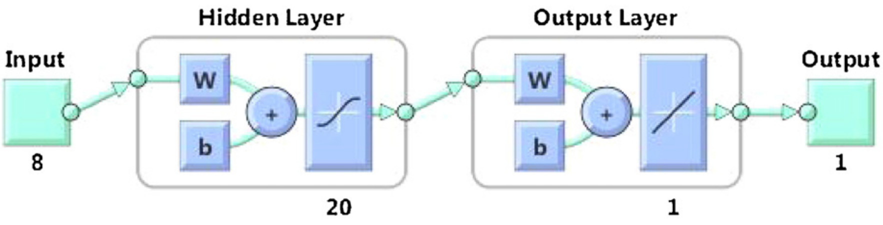

Fig. 3. The schematic structure of the FFANN obtained in this study.

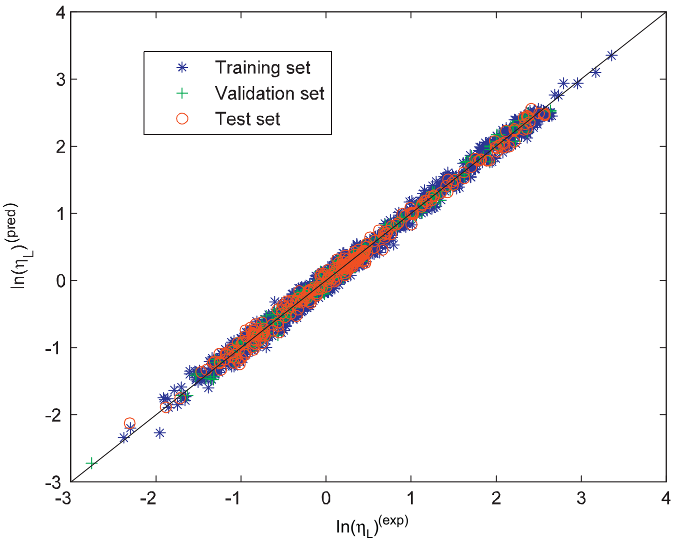

Fig. 4. Comparison between the predicted values by the FFANN and their corresponding experimental values of liquid viscosity (log based).

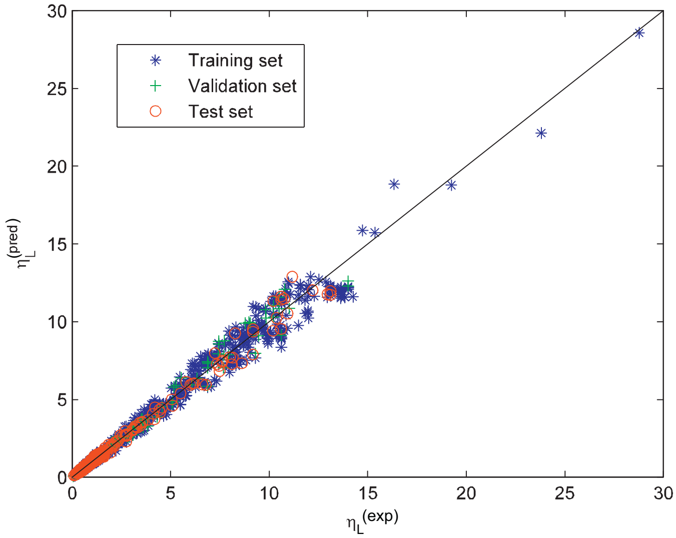

Fig. 5. Comparison between the predicted values by the FFANN and their corresponding experimental values of liquid viscosity.

Table 1
Statistical parameters of the model.

Statistical parameter	Value
Training set	
R^2	0.986
Average percent error	6.88%
Standard deviation error	2.913
Root mean square error	0.345
n	2200
Validation set	
R^2	0.987
Average percent error	6.39%
Standard deviation error	2.708
Root mean square error	0.316
n	274
Test set	
R^2	0.987
Average percent error	6.71%
Standard deviation error	2.751
Root mean square error	0.291
n	274
Training + validation + test set	
R^2	0.987
Average percent error	6.81%
Standard deviation error	2.876
Root mean square error	0.337
n	2748

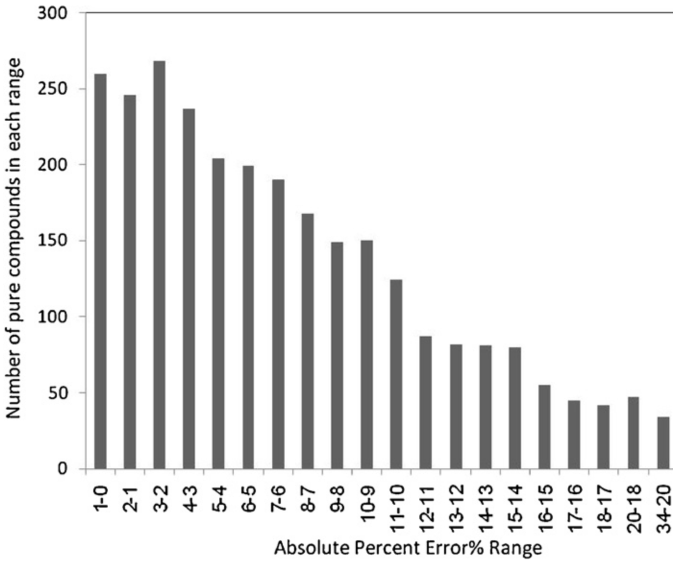

Fig. 6. Distribution of studied compounds in average percent error range.

The predicted liquid viscosity values by the optimized FFNN in comparison with the experimental values [17] are shown in Fig. 4 (log based) and Fig. 5. The values of the predicted liquid viscosity as well as the status of each pure compound in the model (belonging to the training set, the validation set, or the test set) are presented as supplementary information.

The statistical parameters of the best FFNN (based on η_L [cP]) are presented in Table 1. These results show that the squared correlation coefficient, average absolute percent error, and root mean square error of the model over the training set, the validation set, the test set and the main data set are respectively 0.986, 0.987, 0.987, 0.987, 6.88%, 6.39%, 6.71%, 6.81%, 0.345, 0.316, 0.291, and 0.337.

Also, the number of compounds in each integer the absolute percent error range is depicted in Fig. 6. The figure shows that the maximum absolute percent error obtained using the model is 34%. As can be seen, the absolute percent error for more than 1800 compounds is lower than 10%.

4. Conclusion

In this study, a molecular-based model was presented for estimation of the liquid viscosity at 25 °C A comprehensive data set composed of experimental data for 2748 pure compounds from diverse chemical families was utilized to develop a comprehensive model. This eight-parameter model is a three-layer Feed Forward Artificial Neural Networks (FFANN). The parameters were selected from more than 3000 molecular-based parameters by applying the genetic algorithm-based multivariate linear regression methods (GA-MLR). The predictive power of the obtained model was evaluated using several statistical methods.

Another element to consider is that, the presented model may be used as a technique to test the reliability of the experimental data reported in the literature.

Finally, the average percent error of the model results from experimental values demonstrates the accuracy of the presented model.

Supporting information

The supplementary tables contain the names and molecular structure of all the 2748 compounds, their molecular descriptors, and their predicted values of liquid viscosities.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jtice.2012.12.015.

References

[1] Viswanath DS, Ghosh TK, PDH L, Dutt NK. Viscosity of Liquids: theory, estimation, experiment and data. Dordrecht: Springer; 2007 .

[2] Cocchi M, De Benedetti PG, Seeber R, Tassi L, Ulrici A. Development of quantitative structure–property relationships using calculated descriptors for the prediction of the physicochemical properties (nD, ρ, bp, ε, η) of a series of organic solvents. J Chem Inf Model 1999;39:1190–203.

[3] Ivanciuc O, Ivanciuc T, Filip PA, Cabrol-Bass D. Estimation of the liquid viscosity of organic compounds with a quantitative structure–property model. J Chem Inf Model 1999;39:515–24.

[4] Katritzky AR, Chen K, Wang Y, Karelson M, Lucic B, Trinajstic N, et al. Prediction of liquid viscosity for organic compounds by a quantitative structure–property relationship. J Phys Org Chem 2000;13:80–6.

[5] Kauffman GW, Jurs PC. Prediction of surface tension, viscosity, and thermal conductivity for common organic solvents using quantitative structure–property relationships. J Chem Inf Model 2001;41:408–18.

[6] Sutter JM, Dixon SL, Jurs PC. Automated descriptor selection for quantitative structure–activity relationships using generalized simulated annealing. J Chem Inf Model 1995;35:77–84.

[7] Rajappan R, Shingade PD, Natarajan R, Jayaraman VK. Quantitative structure–property relationship (QSPR) prediction of liquid viscosities of pure organic compounds employing random forest regression. Ind Eng Chem Res 2009;48: 9708–12.

[8] Liaw A, Wiener M. Resampling methods in R: the boot package. R News 2002;2:1–39.

[9] Hsu H. Viscosity estimation at low temperatures ($T_r < 0.75$) for organic liquids from group contributions. Chem Eng J 2002;88:27–35.

[10] Luckas M, Lucas K. Viscosity of liquids: an equation with parameters correlating with structural groups. AlChE J 1986;32:139–41.

[11] Marrero-Morejón J. Estimation of liquid viscosity at ambient temperature of pure organic compounds by using group-interaction contributions. Chem Eng J 2000;79:69–72.

[12] Murata A, Tochigi K, Yamamoto H. Prediction of the liquid viscosities of pure components and mixtures using neural network and asog group contribution methods. Mol Simul 2004;30:451–7.

[13] Nannoolal Y, Rarey J, Ramjugernath D. Estimation of pure component properties. Part 4: Estimation of the saturated liquid viscosity of non-electrolyte organic compounds via group contributions and group interactions. Fluid Phase Equilib 2009;281:97–119.

[14] Stefanis E, Constantinou L, Tsivintzelis I, Panayiotou C. New group-contribution method for predicting temperature-dependent properties of pure organic compounds. Int J Thermophys 2005;26:1369–88.

[15] Yinghua L. Estimation of liquid viscosity of pure compounds at different temperatures by a corresponding-states group-contribution method. Fluid Phase Equilib 2002;198:123–30.

[16] Sastri SRS, Rao KK. A new group contribution method for predicting viscosity of organic liquids. Chem Eng J 1992;50:9–25.

[17] Yaws CL. Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds; 2003.

[18] Sanghvi R, Yalkowsky SH. Estimation of the normal boiling point of organic compounds. Ind Eng Chem Res 2006;45:2856–61.

[19] Talete srl. Dragon 5. 4 for windows (Software for molecular Descriptor Calculations); 2006.

[20] Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemometr 1992;6:267–81.

[21] Todeschini R, Consonni, Mauri R, Pavan M. Detecting bad regression models: multicriteia fitness functions in regression analysis. Anal Chim Acta 2004;515: 199–208.

[22] Gharagheizi F. QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN. Comput Mater Sci 2007;40:159–67.

[23] Gharagheizi F, Sattari M, Tirandazi B. Prediction of crystal lattice energy using enthalpy of sublimation: a group contribution-based model. Ind Eng Chem Res 2011;50:2482–6.

[24] Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. QSPR approach for determination of parachor of non-electrolyte organic compounds. Chem Eng Sci 2011;66:2959–67.

[25] Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Determination of critical properties and acentric factors of pure compounds using the artificial neural network group contribution algorithm. J Chem Eng Data 2011;56:2460–76.

[26] Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Use of artificial neural network – group contribution method to determine surface tension of pure compounds. J Chem Eng Data 2011;56:2587–601.

[27] Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Representation/prediction of solubilities of pure compounds in water using artificial neural network – group contribution method. J Chem Eng Data 2011;56:720–6.

[28] Hagan MT, Demuth HB, Beale M. Neural Network Design. Andover, MA: International Thomson; 2002.

[29] Zendehboudi S, Ahmadi MA, James L, Chatzis I. Prediction of condensate-to-gas ratio for retrograde gas condensate reservoirs using artificial neural network with particle swarm optimization. Energy Fuels 2012;26:3432–47.

[30] Yousefi F, Karimi H. Application of equation of state and artificial neural network to prediction of volumetric properties of polymer melts. J Ind Eng Chem 2012.

[31] Pazuki GR, Nikookar M, Dehnavi M, Al-Anazi B. The prediction of permeability using an artificial neural network system. Pet Sci Technol 2012;30:2108–13.

[32] Marinović S, Bolanča T, Ukić. Š, Rukavina V, Jukić A. Prediction of diesel fuel cold properties using artificial neural networks. Chem Technol Fuels Oils 2012;48:67–74.

[33] Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Artificial neural network modeling of solubilities of 21 commonly used industrial solid compounds in supercritical carbon dioxide. Ind Eng Chem Res 2011;50:221–6.

[34] Eslamimanesh A, Gharagheizi F, Mohammadi AH, Richon D. Artificial neural network modeling of solubility of supercritical carbon dioxide in 24 commonly used Ionic liquids. Chem Eng Sci 2011;66:3039–44.

[35] Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. 2nd rev. and enl. ed. Weinheim/Chichester: Wiley-VCH/John Wiley; 2009.