JOURNAL OF CHEMICAL & ENGINEERING DATA

CORRELATION

pubs.acs.org/jced

Representation and Prediction of Molecular Diffusivity of Nonelectrolyte Organic Compounds in Water at Infinite Dilution Using the Artificial Neural Network-Group Contribution Method

Farhad Gharagheizi,[†] Ali Eslamimanesh,[‡] Amir H. Mohammadi,[*,‡,§] and Dominique Richon[‡]

[†]Saman Energy Giti Co., Postal Code: 3331619636 Tehran, Iran

[‡]MINES ParisTech, CEP/TEP - Centre Énergétique et Procédés, 35 Rue Saint Honoré, 77305 Fontainebleau, France

[§]Thermodynamics Research Unit, School of Chemical Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa

S *Supporting Information*

ABSTRACT: The determination of diffusion coefficients of pure compounds in water at infinite dilution is of utmost interest in chemical and environmental engineering, especially wastewater treatment processes. In this work, the artificial neural network-group contribution (ANN-GC) method is applied to represent and predict the molecular diffusivity of nonelectrolyte organic compounds in water at infinite dilution and 298.15 K. A total of 4852 pure compounds from various chemical families has been investigated to propose a predictive model. The obtained results show the squared correlation coefficient of 0.996, root-mean-square error of about 0.02, and average absolute deviation lower than 1.5 % for the calculated or predicted property from existing experimental values.

I. INTRODUCTION

During the past few decades, global environmental concerns have generated great interest in different industries.[1] Among these concepts, wastewater treatment is drastic due to the fact that water is one of the most vital and imperative substances in human life. Numerous types of water motion transport exist within a natural water sample, but they can be divided into two general categories: “advection” and “diffusion”. Generally, a combination of both groups is involved. However, the most important one is diffusion.[2] In 1855, Adolf Fick proposed the following equation to describe the diffusion mechanism:[2,3]

$$J_{Ax} = -D_{AB}\frac{\partial C_A}{\partial x} \qquad (1)$$

where J_{Ax} is mass flux of substance A in the x direction, D_{AB} is the diffusion coefficient of A in B, and C_A is the concentration of the substance A.

The most significant factor to consider for the determination of the diffusion coefficient is that the experimental values of this property are not always available especially for new chemical species applied in modern industries. On the other hand, experimental measurements of such properties may be expensive and time-consuming. Hence, general and reliable models are required for development with the aim of reducing significantly the required experimental work which is expensive and time-consuming.

The presented techniques so far for evaluating the binary liquid diffusion coefficients are based on the calculation of the diffusion coefficient of solute A in solvent B, which is diffusing at infinite dilution (D°_{AB}).[3] This parameter implies that each A molecule is in an environment of essentially pure B. For engineering purposes, D°_{AB} is assumed to be a representative diffusion coefficient even for concentrations of A of 5 to 10 mole fraction.[3]

Wilke and Chang[4] were the first to correlate the diffusion coefficient of solute A in solvent B at infinite dilution. They modified the Stokes–Einstein relation using the molecular weight of the solvent, temperature, viscosity of the solvent, molar volume of the solute at its normal boiling point temperature, and dimensionless association factor of the solvent as the parameters of the correlation. They compared the results of the presented correlation with experimental values of the diffusion coefficients of 250 binary mixtures containing water, methanol, and ethanol as solvents at different temperatures and obtained the average absolute deviations of around 10 %. Good reviews of proposed modifications of this correlation[5−9] especially for the case that the solvent is an organic liquid can be found elsewhere.[3] It has been shown that, although these correlations have brought about increase of the accuracy of the original equation results, none of them have been widely accepted among the researchers.[3]

In 1975, Tyn and Calus[10] related the diffusion coefficient of solute A into solvent B to the molar volume of the solvent at normal boiling point temperature, parachors of the solute and solvent, temperature, and viscosity of the solvent. They used the relation between the parachor and the surface tension to evaluate the required value of the parachor parameter. However, their proposed correlation has some limitations[3]; for example, it is not applicable to viscous solvents. Calculations of diffusion coefficients at infinite dilution for a number of systems show an absolute average deviation of 9 % for this correlation.[3] However,

Received: November 5, 2010
Accepted: March 20, 2011
Published: April 12, 2011

dx.doi.org/10.1021/je101190p | J. Chem. Eng. Data 2011, 56, 1741–1750

this correlation needs parachor values as one of the parameters. Unfortunately, experimental values of parachor may not be available for most of compounds of interest. Furthermore, the group contribution methods available in the literature for their evaluation do not cover many of chemical species. A similar approach has been pursued by Hayduk and Minhas,[11] who used more experimental data to develop such a correlation. They reported an absolute average deviation of 10 % for the same systems investigated by Tyn and Calus.[10] This correlation is generally applied by the researcher for the evaluation of diffusion coefficients of ordinary pure compounds in aqueous solutions.

Another attempt has been made by Nakanishi[12] in 1978. He correlated the values of the diffusion coefficients of chemical substances with molar volumes of solute and solvent, temperature, and viscosity of solvent. This correlation contains also four other factors, which are defined for investigated chemical families including alcohols, glycols, organic acids, highly polar materials, and paraffins in the original article. The values calculated by this correlation for a number of solute−solvent systems show an average absolute deviation of about 13 %. Recently, Gharagheizi and Sattari[2] have proposed a QSPR (quantitative structure property relationship) model, in which the diffusivity coefficients of 320 pure compounds in water have been calculated. They reported the squared correlation coefficient of 0.98 for the obtained results.

Although most of the aforementioned correlations have the advantage of possible application for the calculation of diffusion coefficients of chemical compounds in two or three different solvents including water, methanol, and ethanol at infinite dilution, they have several drawbacks:

1. They need the knowledge of the values of several quantities, for which the experimental values are not always available. In the case that additional estimation techniques are used, they may induce more errors in final calculation results.
2. They generally do not cover wide ranges of chemical compounds from various chemical families. Therefore, they are not so general and comprehensive.
3. The average deviations of the results are about 10 %. These deviations may lead to further unreliability of the calculations/predictions of the diffusivity amounts of solutes in desired solvent, which is a significant factor especially in wastewater treatment processes.
4. Calculations of the model parameters are not generally easy for complicated models such as QSPR ones.

Regarding the preceding drawbacks, more general, reliable, and comprehensive methods are needed to calculate or predict the diffusion coefficients of various chemical compounds from wide ranges of chemical families in water diffusing at infinite dilution. In this work, the artificial neural network-group contribution (ANN-GC) method is used for this purpose.

II. MATERIALS AND METHODS

A. Materials. The accuracy and reliability of models for the representation and prediction of physical properties, especially those dealing with large number of experimental data, directly depend on the quality and comprehensiveness of the applied data set for their development.[13] The aforementioned characteristics of such models include both the diversity in the investigated chemical families and the number of pure compounds available in the data set. In this work, we used the database prepared by Yaws,[14] which is one of the most comprehensive sources of physical property data for chemical species, for example, diffusion coefficients of pure nonelectrolyte organic compounds in water at infinite dilution and 298.15 K. The values of these diffusion coefficients for 4852 investigated pure compounds are available upon request to the authors.

B. Development of New Group Contributions. Having defined the data set, the chemical structures of all 4852 nonelectrolyte organic compounds have been analyzed. Consequently, 148 functional groups have been found to be more efficient for the representation and prediction of the diffusion coefficients of pure compounds in water diffusing at infinite dilution at 298.15 K. The functional groups used in this study are presented in Table 1. Besides, their numbers of occurrences in pure compounds used in this work are extensively presented as Supporting Information. These chemical groups are used as the proposed model parameters.

C. Generation of the ANN-GC. The next calculation step, and perhaps the most significant one, is to search for a relationship between the chemical functional groups and the molecular diffusivity of chemical compounds at infinite dilution and 298.15 K. The simplest method for this purpose is the assumption of the existence of a multilinear relationship between these groups and the desired property (here is the diffusion coefficients of pure compounds).[15−18] This technique is a similar method used in the most of classical group contribution methods.[19] Several calculations have shown that the application of the mentioned methodology for the current problem brings about poor results. Consequently, the nonlinear mathematical method of artificial neural networks (ANNs) is investigated. Artificial neural networks are extensively used in various scientific and engineering problems[13,18,20−46]

All of the 148 functional groups and also the diffusion coefficient values of pure compounds are normalized between −1 and +1 to decrease computational errors. This can be performed using maximum and minimum values of each functional group for input data and using maximum and minimum values of diffusion coefficients for output parameters. Because of the fact that we face with a large range of diffusion coefficient values for different compounds, these values are generally normalized between −1 and +1 to prevent truncation errors. Besides, this procedure, which is done in optimization process, is performed to obtain the parameters of the neural networks (W_1, W_2, b_1, b_2 as shown in Figure 1), and it has no effect on the model results. Later, these values are again changed to the original diffusivity coefficient values, which are finally used as the inputs and reported as outputs of the developed model. Later, the main data set is divided into three new subdata sets including the "training" set, the "validation" set, and the "test" 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 main data set into three subdata sets is performed randomly. For this purpose, about 80 %, 10 %, and 10 % of the main data set are randomly selected for the training set (3882 compounds), the validation set (485 compounds), and the test set (485 compounds). The effect of the allocation percent of the three subdata sets from the data of main data set on the accuracy of the ANN model has been studied elsewhere.[47,48]

As a matter of fact, generating an ANN model is the determination of the weight matrices and bias vectors.[15] As mentioned earlier and as shown in Figure 1, there are two weight matrices and two bias vectors in a three layer feed forward artificial neural network (FFANN): W_1 and W_2, b_1, and b_2.[13,18,20−45] These

Table 1. Functional Groups Used to Develop the Model

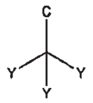
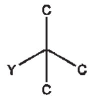
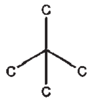
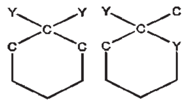
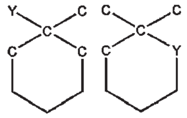
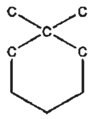
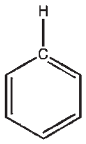
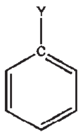
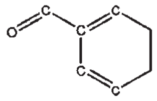
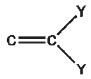
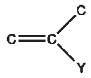
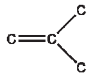
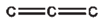
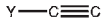
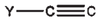
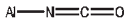
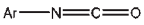
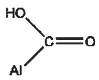
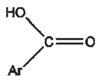
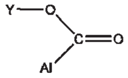
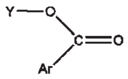
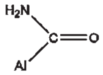
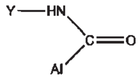
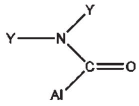
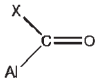
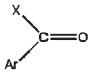
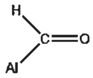
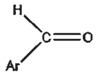
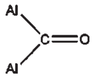
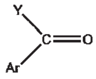
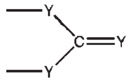
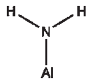
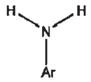
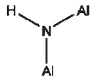
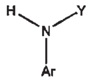
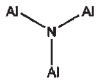
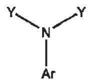
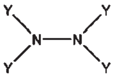
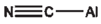
No.	ID	Functional Groups	Comments
1	DW001		terminal primary C(sp3) Y = any terminal atom or heteroaromatic group (i.e. H, X, OH, NH2, etc.)
2	DW002		total tertiary C(sp3) Y = H or any heteroatom
3	DW003		total quaternary C(sp3)
4	DW004		ring secondary C(sp3) Y = H or any heteroatom
5	DW005		ring tertiary C(sp3) Y = H or any heteroatom
6	DW006		ring quaternary C(sp3)
7	DW007		unsubstituted benzene C(sp2)
8	DW008		substituted benzene C(sp2) Y = carbon or any heteroatom
9	DW009		non-aromatic conjugated C(sp2)
10	DW010		terminal primary C(sp2) Y = any terminal atom or heteroaromatic group (i.e. H, X, OH, NH2, etc.)
11	DW011		aliphatic secondary C(sp2) Y = H or any heteroatom
12	DW012		aliphatic tertiary C(sp2)
13	DW013		allenes groups
14	DW014		terminal C(sp) Y = any terminal atom or heteroaromatic group (i.e. H, X, OH, NH2, etc.)
15	DW015		non-terminal C(sp) Y = C or any non-terminal heteroatom
16	DW016		isocyanates (aliphatic)
17	DW017		isocyanates (aromatic)
18	DW018		carboxylic acids (aliphatic)
19	DW019		carboxylic acids (aromatic)
20	DW020		esters (aliphatic) Y = Ar or Al (not H) Al = H or aliphatic group linked through C
21	DW021		esters (aromatic) Y = Al or Ar
22	DW022		primary amides (aliphatic) Al = H or aliphatic group linked through C
23	DW023		secondary amides (aliphatic) Y = Ar or Al (not H, not C = O) Al = H or aliphatic group linked through C
24	DW024		tertiary amides (aliphatic) Y = Ar or Al (not H, not C = O) Al = H or aliphatic group linked through C
25	DW025		acyl halogenides (aliphatic)
26	DW026		acyl halogenides (aromatic)
27	DW027		aldehydes (aliphatic)
28	DW028		aldehydes (aromatic)
29	DW029		ketones (aliphatic)
30	DW030		ketones (aromatic) Y = Al or Ar
31	DW031		carbonate (-thio) derivatives (Y = O or S)
32	DW032		primary amines (aliphatic) Al = aliphatic group linked through C (not C = O)
33	DW033		primary amines (aromatic)
34	DW034		secondary amines (aliphatic) Al = aliphatic group linked through C (not C = O)
35	DW035		secondary amines (aromatic) Y = Ar or Al (not C = O)
36	DW036		tertiary amines (aliphatic) Al = aliphatic group linked through C (not C = O
37	DW037		tertiary amines (aromatic) Y = Ar or Al (not C = O)
38	DW038		N hydrazines Y = C or H
39	DW039		nitriles (aliphatic)

Table 1. Continued

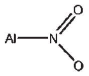
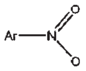
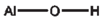
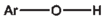
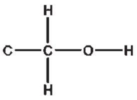
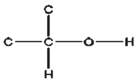
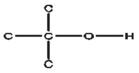
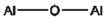
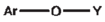
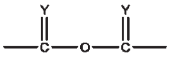
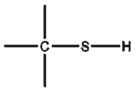
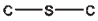
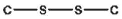
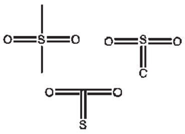
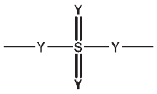
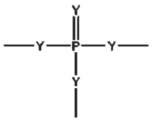
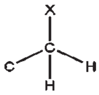
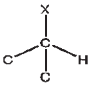
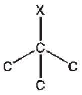
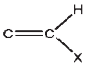
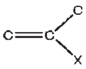
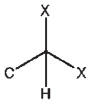
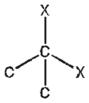
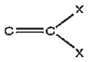
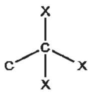
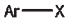
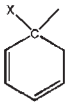
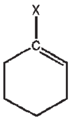
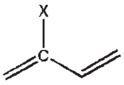
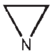
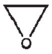
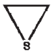
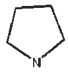
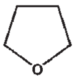
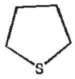
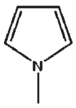
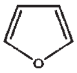
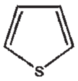
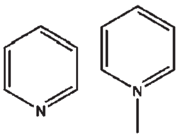
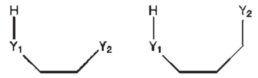
No.	ID	Functional Groups	Comments
40	DW040		nitro groups (aliphatic) Al = H or aliphatic group linked through carbon
41	DW041		nitro groups (aromatic) Ar = aromatic group linked through carbon
42	DW042		hydroxyl groups Al = aliphatic group linked through any atom
43	DW043		aromatic hydroxyls Ar = aromatic group linked through any atom
44	DW044		primary alcohols
45	DW045		secondary alcohols
46	DW046		tertiary alcohols
47	DW047		ethers (aliphatic) Al = aliphatic group linked through C (not C = O, not C # N)
48	DW048		ethers (aromatic) Y = Ar or Al (not C = O, not C # N)
49	DW049		anhydrides (thio-) Y = O or S
50	DW050		thiols
51	DW051		sulfides
52	DW052		disulfides
53	DW053		sulfones
54	DW054		sulfates (thio- / dithio-) (Y = O or S)
55	DW055		phosphates / thiophosphates (Y = O or S)
56	DW056		CH2RX
57	DW057		CHR2X
58	DW058		CR3X
59	DW059		R=CHX
60	DW060		R=CRX
61	DW061		CHRX2
62	DW062		CR2X2
63	DW063		R=CX2
64	DW064		CRX3
65	DW065		X on aromatic ring
66	DW066		X on ring C(sp3)
67	DW067		X on ring C(sp2)
68	DW068		X on exo-conjugated C
69	DW069		Aziridines
70	DW070		Oxiranes
71	DW071		Thiranes
72	DW072		Pyrrolidines
73	DW073		Oxolanes
74	DW074		tetrahydro-Thiophenes
75	DW075		Pyrroles
76	DW076		Furancs
77	DW077		Thiophenes
78	DW078		Pyridines
79	DW079	Sum of the hydrogens linked to all of the Os and Ns in the molecul	donor atoms for H-bonds (N and O)
80	DW080	Total Ns, Os and Fs in the molecule, excluding N with a formal positive charge, higher oxidation states and pyrrolyl form of N	acceptor atoms for H-bonds (N, O, F)

Table 1. Continued

No.	ID	Functional Groups	Comments
81	DW081		intramolecular H-bonds (Y1 = B, N, O, Al, P, S, Y2 = N, O, F.)
82	DW082	CH3R / CH4	
83	DW083	CH3X	
84	DW084	CH2RX	
85	DW085	CH2X2	
86	DW086	CHR2X	
87	DW087	CHX3	
88	DW088	CR3X	
89	DW089	CR2X2	
90	DW090	CX4	
91	DW091	=CH2	
92	DW092	=CRX	
93	DW093	R--CR--R	
94	DW094	R--CX--R	
95	DW095	R--CH--X	
96	DW096	R--CR--X	
97	DW097	R--CH..X	
98	DW098	R--CR..X	
99	DW099	R-C(=X)-X / R-C#X / X=C=X	
100	DW100	X-C(=X)-X	
101	DW101	H^a attached to C^0(sp3) no X attached to next C	
102	DW102	H^a attached to C^1(sp3) / C^0(sp2)	
103	DW103	H^a attached to C^2(sp3) / C^1(sp2) / C^0(sp)	
104	DW104	H^a attached to C^3(sp3) / C^2(sp2) / C^3(sp2) / C^3(sp)	
105	DW105	H attached to alpha-C^b	
106	DW106	H^a attached to C^0(sp3) with 1X attached to next C	
107	DW107	H^a attached to C^0(sp3) with 2X attached to next C	
108	DW108	H^a attached to C^0(sp3) with 3X attached to next C	
109	DW109	H^a attached to C^0(sp3) with 4X attached to next C	
110	DW110	alcohol	
111	DW111	phenol / enol / carboxyl OH	
112	DW112	=O	
113	DW113	Al-O-Al	
114	DW114	Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X	
115	DW115	O--^{c}	
116	DW116	R-O-O-R	
117	DW117	Al3-N	
118	DW118	Ar-NH-Al	
119	DW119	RCO-N< / >N-X=X	
120	DW120	Ar2NH / Ar3N / Ar2N-Al / R..N..R^c	
121	DW121	R#N / R=N-	
122	DW122	Al-NO2	
123	DW123	Ar-N=X / X-N=X	
124	DW124	F^a attached to C^1(sp3)	
125	DW125	F^a attached to C^2(sp3)	
126	DW126	F^a attached to C^3(sp3)	
127	DW127	F^a attached to C^1(sp2)	
128	DW128	F^a attached to C^2(sp2)-C^4(sp2) / C^1(sp) / C^4(sp3) / X	
129	DW129	Cl^a attached to C^1(sp3)	
130	DW130	Cl^a attached to C^2(sp3)	
131	DW131	Cl^a attached to C^3(sp3)	
132	DW132	Cl^a attached to C^1(sp2)	
133	DW133	Cl^a attached to C^2(sp2)-C^4(sp2) / C^1(sp) / C^4(sp3) / X	
134	DW134	Br^a attached to C^1(sp3)	
135	DW135	Br^a attached to C^2(sp3)	
136	DW136	Br^a attached to C^3(sp3)	
137	DW137	Br^a attached to C^1(sp2)	
138	DW138	Br^a attached to C^2(sp2)-C^4(sp2) / C^1(sp) / C^4(sp3) / X	
139	DW139	I^a attached to C^1(sp3)	
140	DW140	I^a attached to C^2(sp3)	
141	DW141	I^a attached to C^3(sp3)	
142	DW142	I^a attached to C^1(sp2)	
143	DW143	R-SH	
144	DW144	R2S / RS-SR	
145	DW145	R=S	
146	DW146	R-SO2-R	
147	DW147	>Si<	
148	DW148	X3-P=X (phosphate)	

[a] The superscript represents the formal oxidation number. R represents any group linked through carbon; X represents any electronegative atom (O, N, S, P, Se, halogens); Al and Ar represent aliphatic and aromatic groups, respectively; = represents a double bond; # represents a triple bond; -- represents an aromatic bond as in benzene or delocalized bonds such as the N—O bond in a nitro group; .. represents aromatic single bonds as the C—N bond in pyrrole.

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 (represented/predicted diffusion coefficients) and the target values (experimental diffusion coefficient values). This minimization is performed by the Levenberg–Marquardt (LM)[47] 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 converge to the global optimum. The LM[47] is the most-widely used algorithm for training due to being robust and accurate enough to deal with the considered system.[13,18,20−45]

In most cases, the number of neurons in the hidden layer (n) is firstly fixed, and then the main goal is to produce an ANN model, which is able to predict the target values as accurately as expected. This step is repeated until the best ANN 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.[13,18−45]

3. RESULTS AND DISCUSSION

An optimized FFANN has been obtained using the aforementioned procedure for the representation and prediction of the diffusion coefficients of 4852 pure nonelectrolyte organic compounds in water at infinite dilution and 298.15 K. For this purpose, several three-layer FFANN modules have been generated assuming numbers 1 through 50 for n (number of neurons in hidden layer) using the previously described procedure. The most accurate results without overfitting are observed for $n = 4$. It should be noted that this value is not the global value, because the optimization method used to train the ANN has great effects on the obtained value.[27] Therefore, the developed three-layer FFANN has the structure of 148-4-1.

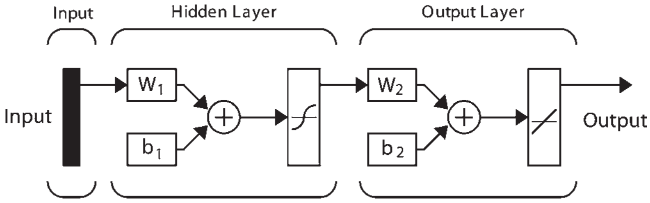


Figure 1. Schematic structure of the FFANN used in this study. W: weight; b: bias.

The represented and predicted diffusion coefficients are shown in Figure 2 in comparison with the experimental values.[14] More meticulous investigation of the results show that there are 58 compounds for which the presented model results lead to more than 13 % (based in Figure 2) absolute deviations from experimental values.[14] It seems that there is no relation between these compound structures to show some weaknesses in representing and predicting of the diffusion coefficient values of related chemical families. Therefore, we may suspect that corresponding experimental diffusion coefficient values may not be accurate or may be somehow erroneous because of the existing difficulties in experimental measurements especially those where complex chemical structures are involved. For further investigation of the reliability of such data, we have pursued the following procedure:

1. Eliminating the outlier data points (58 points) from the investigated experimental values.[14]
2. Developing a new ANN-GC model for the representation and prediction of the remaining diffusion coefficient values (4794).
3. Prediction of the eliminated outlier data point values using the new developed model for further checking the reliability of these values.

Figure 3 shows the eliminated outlier set from the main data set. The results of the new developed model are shown in Figure 4. More detailed results including the absolute deviation

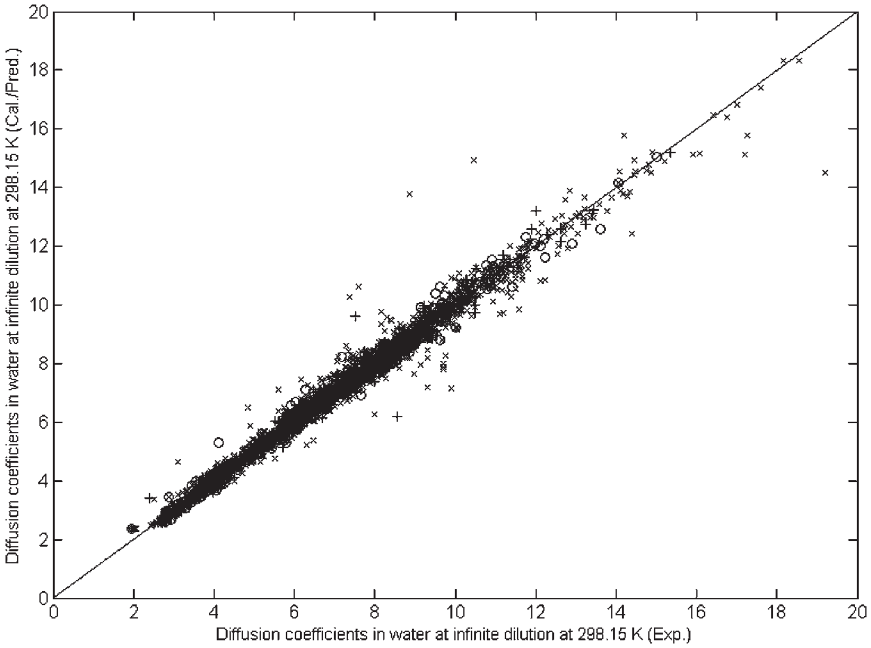


Figure 2. Comparison between the calculated and predicted results of the first model and experimental values[14] of diffusion coefficients of investigated pure nonelectrolyte organic compounds in water at infinite dilution and 298.15 K. ×, training set; +, validation set; ○, test set. The unit of the diffusion coefficient values reported in the figure is $(\mathrm{cm}^2\cdot\mathrm{s}^{-1})\cdot10^6$.

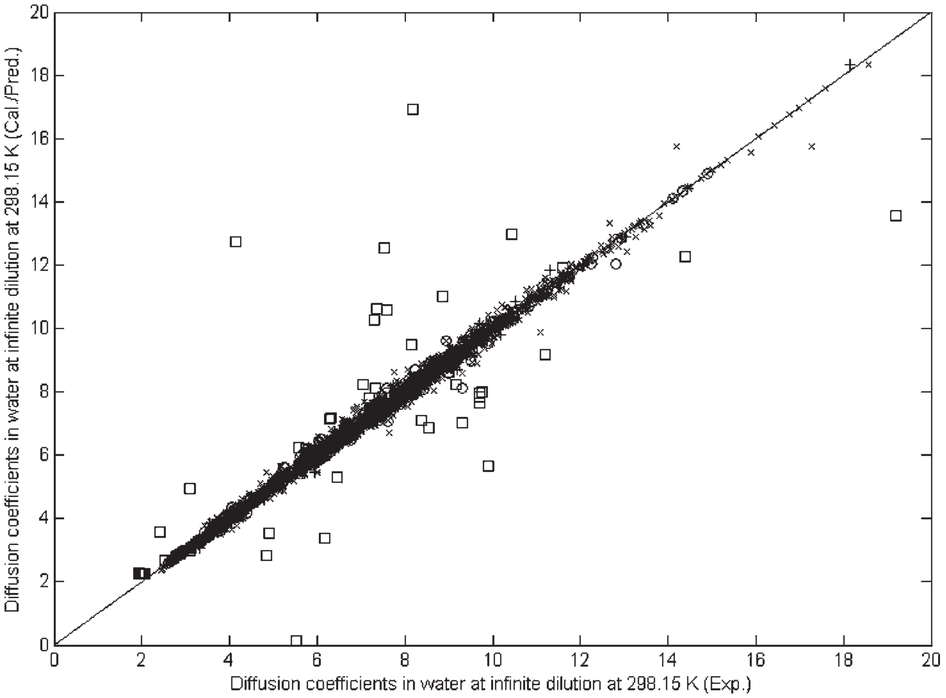


Figure 3. Definition of the outlier set eliminated from the main data set. ×, training set; +, validation set; ○, test set; □, outlier set; ○, test set. The unit of the diffusion coefficient values reported in the figure is $(\mathrm{cm}^2 \cdot \mathrm{s}^{-1}) \cdot 10^6$.

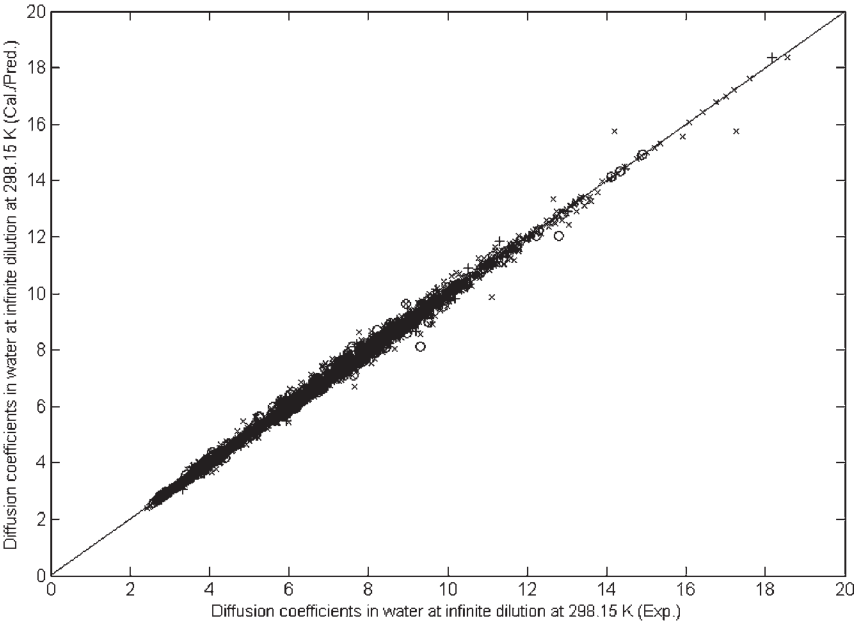


Figure 4. Comparison between the calculated and predicted results of the second model and experimental values[14] of diffusion coefficients of investigated pure nonelectrolyte organic compounds in water at infinite dilution and 298.15 K × training set; +, validation set; ○, test set. The unit of the diffusion coefficient values reported in the figure is $(\mathrm{cm}^2 \cdot \mathrm{s}^{-1}) \cdot 10^6$.

ranges of the represented and predicted diffusion coefficient values using the first and the second models are reported in Figure 5. Besides, the statistical parameters of both of the models are reported in Table 2. As can be observed, the new model leads to the absolute deviation ranges not to exceed 13 % for the new sets of data excluding the outliers. Consequently, the average absolute deviation of the model is about 1.4 % while this value is about 2.6 % regarding the previous model results. All of the calculated and predicted results, the number of occurrences of the 148 functional groups in all of investigated pure compounds, and the absolute deviations of the represented and predicted diffusion coefficients are available as Supporting Information. It is inferred from these results that most of the calculated and predicted diffusion coefficient values for outlier (eliminated) data points (about 59 %) bring about high deviations (over 13 %) even in the new developed model. Therefore, it can be implied that these

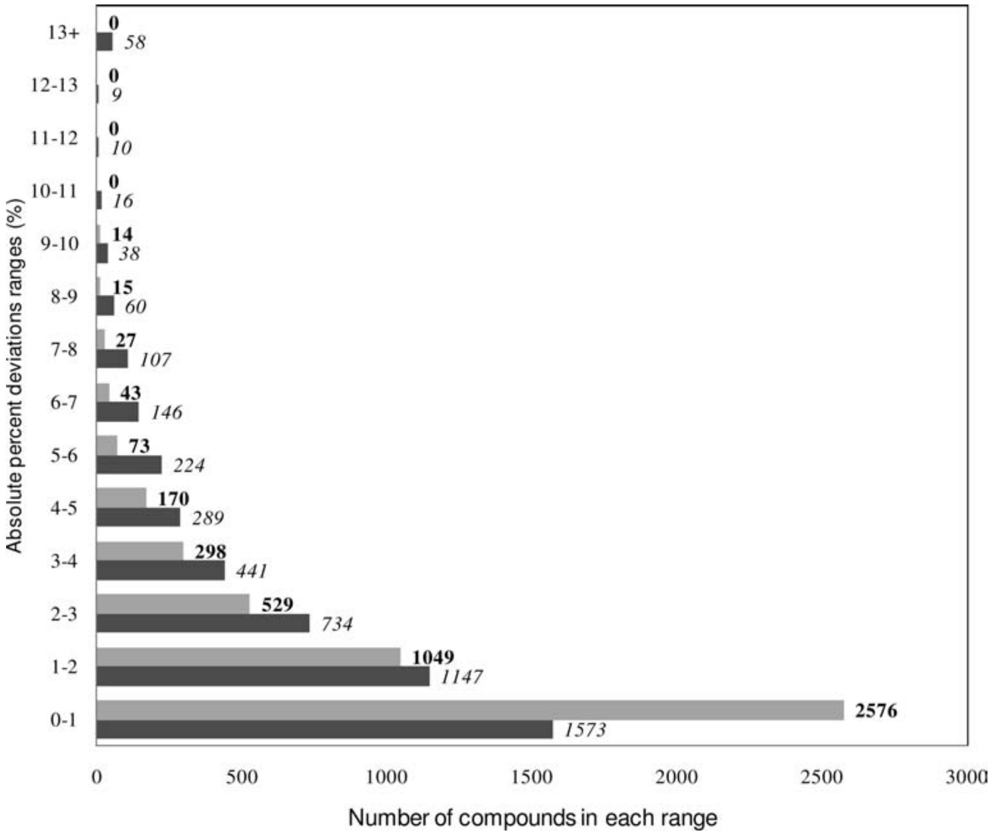


Figure 5. Deviation ranges of the results of the two developed models over all of the investigated compounds. The upper bars show the second model, and the other ones indicate the results of the first developed model.

Table 2. Statistical Parameters of the Presented Models

statistical parameter	value	value
training set	the first model	the second model
R^2[a]	0.982	0.996
average absolute deviation[b]	2.58 %	1.41 %
standard deviation error	2.2	2.2
mean square error	0.091	0.021
N[c]	3882	3836
validation set		
R^2	0.984	0.996
average absolute deviation	2.59 %	1.37 %
standard deviation error	2.2	2.1
mean square error	0.075	0.018
N	485	479
test set		
R^2	0.987	0.996
average absolute deviation	2.60 %	1.47 %
standard deviation error	2.1	2.1
mean square error	0.060	0.024
N	485	479
training + validation + test set		
R^2	0.982	0.996
average absolute deviation	2.58 %	1.41 %
standard deviation error	2.2	2.2
mean square error	0.086	0.021
N	4852	4794

[a] Squared correlation coefficient. [b] %AAD = $100/N\ \Sigma_i^N(|\mathrm{Calc.}(i)/\mathrm{Pred.}(i) - \mathrm{Exp.}(i)|)/(\mathrm{Exp.}(i))$. [c] Number of data points.

data points may be among the probable doubtful data with higher experimental uncertainties, as we have already expected from the first model results. All of the developed model results regarding the suspected outliers have been presented as Supporting Information. The mat file (MATLAB file format) of the new obtained ANN containing all parameters of the model is also available as Supporting Information. Also, the instructions for running this developed computer program have been presented in the Appendix.

To recapitulate, the results imply that the new obtained ANN-GC model is an accurate method to represent/predict the diffusion coefficients of pure chemical compounds in water diffusing at infinite dilution and 298.15 K. Besides, the comprehensiveness of the model that is imperative in representation/prediction of physical properties of large numbers of pure compounds is guaranteed because it is developed over a diverse set of 4852/4794 pure compounds from various chemical families. The two preceding points obviously demonstrate the capabilities of the proposed model in comparison with the previously presented one based on QSPR approach.[2]

4. CONCLUSION

In this study, a group contribution-based model was presented for representation and prediction of the molecular diffusivity of 4852 pure nonelectrolyte organic compounds in water at infinite dilution and 298.15 K. These conditions are of much interest for wastewater treatment processes. The model is the result of a combination of FFANN and GC methods. The required parameters of the model are the numbers of occurrences of 148 functional groups in each investigated molecule. It should be noted that most of these functional groups are not simultaneously available in a particular molecule. Therefore, the computation of the required parameters from the chemical structure of any molecule is simple. For developing the model, the experimental diffusivity in water values from a large data set[14] containing 4852 pure compounds from various chemical families were applied. As a consequence, a comprehensive model was developed to represent and predict the diffusion coefficients of many of pure compounds in water although there are still some limitations. The model has a wide range of applicability, but the prediction capability of the model is restricted to the compounds, which are similar to those ones applied to develop the model. The application of the model for the totally different compounds than the investigated ones is not recommended although it may be used for a rough estimation of the molecular diffusivity of these kinds of compounds.

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. It was found that experimental values of diffusion coefficients for 34 chemical compounds are among the real outliers of the model and we may consider them as the data with higher uncertainties than other experimental values in the data set.

Finally, the average absolute deviation of the model results from experimental values[14] demonstrates the accuracy of the presented model. It should be noted that the extension of the model to different temperature conditions requires adequate data of diffusion coefficients at these conditions. More meticulous experimental works are required to be done for this concept.

■ APPENDIX: INSTRUCTIONS FOR USING THE PROPOSED MODEL

The model is very easy to apply. Just drag and drop the mat file into the MATLAB environment (any version) workspace. One can follow the below example to get a response from the model step by step:

Assume that one is willing to predict the diffusion coefficient of methylcyclopropane in water at ambient conditions using the developed model. First of all, the group-contribution parameters should be defined from the chemical structure of methylcyclopropane (refer to the Supporting Information). Later, drag and drop the mat file, and the following commands should be entered in MATLAB workspace:

```
methylcyclopropane_GCs=[1   1   0   2   1   0   0   0   0   0   0   0   0   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 0   0   0   0   8   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]';

D_methylcyclopropane=sim(net,methylcyclopropane_GCs).
```

Therefore, one will observe the estimated log *D* (natural logarithm of the diffusivity) as follows: *11.61*, where the experimental value for this compounds is equal to 11.62 (approximately ARD = 0.1 %).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Calculated/predicted diffusion coefficients by the presented ANN-GC model accompanied with absolute deviations of the results from the experimental values and the number of occurrences of the functional groups in each molecule. Also, different types of subdata sets have been shown. Moreover, the outlier set and the developed computer program have been also presented. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: amir-hossein.mohammadi@mines-paristech.fr. Tel.: + (33) 1 64 69 49 70. Fax: + (33) 1 64 69 49 68.

Funding Sources

A.E. wishes to thank MINES ParisTech for providing him a Ph.D. scholarship.

■ REFERENCES

(1) Sattari, M.; Gharagheizi, F. Prediction of molecular diffusivity of pure components into air: A QSPR approach. *Chemosphere* **2008**, *72*, 1298–1302.

(2) Gharagheizi, F.; Sattari, M. Estimation of molecular diffusivity of pure chemicals in water: A quantitative structure-property relationship study. *SAR & QSAR Environ. Res.* **2009**, *20*, 67–285.

(3) Poling, B. E.; Prausnitz, J. M.; O'Connell, J. P. *Properties of Gases and Liquids*, 5th ed. McGraw-Hill: New York, 2001.

(4) Wilke, C. R.; Chang, P. Correlation of diffusion coefficients in dilute solutions. *AIChE J.* **1955**, *1*, 264–270.

(5) Amourdam, M. J.; Laddha, G. S. Diffusivity of some binary liquid systems using diaphragm cell. *J. Chem. Eng. Data* **1967**, *12*, 389–391.

(6) Lusis, M. A. Predicting liquid diffusion coefficients. *Chem. Process. Eng.* **1971**, *5*, 27–35.

(7) Olander, D. R. The diffusivity of water in organic solvents. *AIChE J.* **1961**, *7*, 175–176.

(8) Wise, D. L.; Houghton, G. The diffusion coefficients of ten slightly soluble gases in water at 10−60 °C. *Chem. Eng. Sci.* **1966**, *21*, 999–1010.

(9) Witherspoon, P. A.; Bonoli, L. Correlation of diffusion coefficients for paraffin, aromatic, and cycloparaffin hydrocarbons in water. *Ind. Eng. Chem. Fundam.* **1969**, *8*, 589–591.

(10) Tyn, M. T.; Calus, W. F. Diffusion coefficients in dilute binary liquid mixtures. *J. Chem. Eng. Data* **1975**, *20*, 106–109.

(11) Hayduk, W.; Minhas, B. S. Correlations for prediction of molecular diffusivities in liquids. *Can. J. Chem. Eng.* **1982**, *60*, 295–299.

(12) Nakanishi, K. Prediction of diffusion coefficient of nonelectrolytes in dilute solution based on generalized Hammond-Stokes plot. *Ind. Eng. Chem. Fundam.* **1978**, *17*, 253–256.

(13) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; 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*, 221–226.

(14) Yaws, C. L. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*; Knovel: Norwich, NY, 2003.

(15) Eslamimanesh, A.; Gharagheizi, F.; Mohammadi, A. H.; Richon, D. Artificial neural network modeling of solubility of supercritical carbon dioxide in 24 commonly used ionic liquids. *Chem. Eng. Sci.* **2011**, accepted manuscript.

(16) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Determination of parachor of various compounds using artificial neural network-group contribution approach. *Ind. Eng. Chem. Res.* **2011**, in press.

(17) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Determination of critical properties and acentric factors of pure compounds using artificial neural network-group contribution algorithm. *J. Chem. Eng. Data* **2011**, in press.

(18) Gharagheizi, F.; Abbasi, R.; Tirandazi, B. Prediction of Henry's law constant of organic compounds in water from a new group-contribution-based model. *Ind. Eng. Chem. Res.* **2010**, *49*, 12685–12695.

(19) Hou, T. J.; Wang, J. M. Structure - ADME relationship: still a long way to go. Expert. *Opin. Drug Metab. Toxicol.* **2008**, *4*, 759–771.

(20) Chouai, A.; Laugier, S.; Richon, D. Modeling of thermodynamic properties using neural networks: Application to refrigerants. *Fluid Phase Equilib.* **2002**, *199*, 53–62.

(21) Piazza, L.; Scalabrin, G.; Marchi, P.; Richon, D. Enhancement of the extended corresponding states techniques for thermodynamic modelling. I. Pure fluids. *Int. J. Refrig.* **2006**, *29*, 1182–1194.

(22) Scalabrin, G.; Marchi, P.; Bettio, L.; Richon, D. Enhancement of the extended corresponding states techniques for thermodynamic modelling. II. Mixtures. *Int. J. Refrig.* **2006**, *29*, 1195–1207.

(23) Chapoy, A.; Mohammadi, A. H.; Richon, D. Predicting the hydrate stability zones of natural gases using Artificial Neural Networks. *Oil Gas Sci. Technol. Rev. IFP* **2007**, *62*, 701–706.

(24) Mohammadi, A. H.; Richon, D. Hydrate phase equilibria for hydrogen + water and hydrogen + tetrahydrofuran + water systems: Predictions of dissociation conditions using an artificial neural network algorithm. *Chem. Eng. Sci.* **2010**, *65*, 3352–3355.

(25) Mohammadi, A. H.; Richon, D. Estimating sulfur content of hydrogen sulfide at elevated temperatures and pressures using an artificial neural network algorithm. *Ind. Eng. Chem. Res.* **2008**, *47*, 8499–8504.

(26) Mohammadi, A. H.; Richon, D. A Mathematical model based on artificial neural network technique for estimating liquid water - hydrate equilibrium of water - hydrocarbon System. *Ind. Eng. Chem. Res.* **2008**, *47*, 4966–4970.

(27) Mohammadi, A. H.; Afzal, W.; Richon, D. Determination of critical properties and acentric factors of petroleum fractions using artificial neural networks. *Ind. Eng. Chem. Res.* **2008**, *47*, 3225–3232.

(28) Mohammadi, A. H.; Richon, D. Use of artificial neural networks for estimating water content of natural gases. *Ind. Eng. Chem. Res.* **2007**, *46*, 1431–1438.

(29) Mohammadi, A. H.; Martínez-López, J. F.; Richon, D. Determining phase diagrams of tetrahydrofuran+methane, carbon dioxide or nitrogen clathrate hydrates using an artificial neural network algorithm. *Chem. Eng. Sci.* **2010**, *65*, 6059–6063.

(30) Mehrpooya, M.; Mohammadi, A. H.; Richon, D. Extension of an Artificial Neural Network algorithm for estimating sulfur content of sour gases at elevated temperatures and pressures. *Ind. Eng. Chem. Res.* **2010**, *49*, 439–442.

(31) Mohammadi, A. H.; Belandria, V.; Richon, D. Use of an artificial neural network algorithm to predict hydrate dissociation conditions for hydrogen + water and hydrogen + tetra-n-butyl ammonium bromide + water systems. *Chem. Eng. Sci.* **2010**, *65*, 4302–4305.

(32) Gharagheizi, F. A new group contribution-based method for estimation of lower flammability limit of pure compounds. *J. Hazard. Mater.* **2009**, *170*, 595–604.

(33) Gharagheizi, F. New neural network group contribution model for estimation of lower flammability limit temperature of pure compounds. *Ind. Eng. Chem. Res.* **2009**, *48*, 7406–7416.

(34) Gharagheizi, F. Prediction of standard enthalpy of formation of pure compounds using molecular structure. *Aust. J. Chem.* **2009**, *62*, 376–381.

(35) Gharagheizi, F.; Tirandazi, B.; Barzin, R. Estimation of aniline point temperature of pure hydrocarbons: A quantitative structure-property relationship approach. *Ind. Eng. Chem. Res.* **2009**, *48*, 1678–1682.

(36) Gharagheizi, F.; Mehrpooya, M. Prediction of some important physical properties of sulfur compounds using QSPR models. *Mol. Divers.* **2008**, *12*, 143–155.

(37) Gharagheizi, F.; Alamdari, R. F.; Angaji, M. T. A new neural network-group contribution method for estimation of flash point. *Energy Fuels* **2008**, *22*, 1628–1635.

(38) Gharagheizi, F.; Fazeli, A. Prediction of Watson characterization factor of hydrocarbon compounds from their molecular properties. *QSAR Comb. Sci.* **2008**, *27*, 758–767.

(39) Gharagheizi, F.; Alamdari, R. F. A molecular-based model for prediction of solubility of c60 fullerene in various solvents. *Fullerenes, Nanotubes, Carbon Nanostructures* **2008**, *16*, 40–57.

(40) Gharagheizi, F. A new neural network quantitative structure-property relationship for prediction of θ (Lower Critical Solution Temperature) of polymer solutions. *e-Polym.* **2007**, no. 114.

(41) Gharagheizi, F. QSPR studies for solubility parameter by means of genetic algorithm-based multivariate linear regression and generalized regression neural network. *QSAR Combin. Sci.* **2008**, *27*, 165–170.

(42) Gharagheizi, F. A chemical structure-based model for estimation of upper flammability limit of pure compounds. *Energy Fuels* **2010**, *24*, 3867–3871.

(43) Vatani, A.; Mehrpooya, M.; Gharagheizi, F. Prediction of standard enthalpy of formation by a QSPR Model. *Int. J. Mol. Sci.* **2007**, *8*, 407–432.

(44) Mehrpooya, M.; Gharagheizi, F. A Molecular approach for prediction of sulfur compounds solubility parameters, phosphorus sulfur and silicon and related elements. *Phosphorus Sulfur* **2010**, *185*, 204–210.

(45) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Representation/Prediction of Solubilities of Pure Compounds in Water using Artificial Neural Network-Group Contribution Method. Accepted manuscript. *J. Chem. Eng. Data* **2011**.

(46) Kalogirou, S. A. Artificial neural networks in renewable energy systems applications: A review. *Renewable Sustainable Energy Rev.* **2001**, *5*, 373–401.

(47) Hagan, M.; Demuth, H. B.; Beale, M. H. *Neural Network Design*; International Thomson Publishing: Boston, 2002.

(48) Gharagheizi, F. QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN. *Comput. Mater. Sci.* **2007**, *40*, 159.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on April 12, 2011. Equation 1 was updated. The revised paper was reposted on April 15, 2011.