Korean J. Chem. Eng., 34(5), 1405-1415 (2017)
DOI: 10.1007/s11814-017-0018-0

pISSN: 0256-1115
eISSN: 1975-7220

Quantitative structure-property relationship (QSPR) for prediction of CO_2 Henry's law constant in some physical solvents with consideration of temperature effects

Ali Ebrahimpoor Gorji, Zahra Eshaghi Gorji, and Siavash Riahi†

Institute of Petroleum Engineering, Faculty of Chemical Engineering, College of Engineering,
University of Tehran, Tehran, Iran
(*Received 16 August 2016 • accepted 6 February 2017*)

Abstract−Different types of physical solvents have been utilized for CO_2 removal from natural gas in the sweetening process. In this work, quantitative structure-property relationship (QSPR) method is suggested to build powerful models to predict Henry's law constant (H_{LC}) for CO_2 in physical solvents. Modeling the H_{LC} for CO_2 as a function of molecular descriptors was achieved by multiple linear regression and descriptor selection was by genetic algorithm. The main proposed model has two simple descriptors, including the number of hydroxyl groups and molecular weight of solvents at fixed temperature. Also, the effect of temperature was studied, and this operational variable was added to the mentioned simple descriptors. In this case, the data set is comprised of 77 H_{LC} for CO_2 in solvents and at different temperatures. Several internal and external validation methods demonstrated the excellent ability for prediction, and the average relative deviation of main model was 6.48.

Keywords: Henry's Law Constant, CO_2, QSPR, Physical Solvent, Temperature

INTRODUCTION

Carbon dioxide (CO_2) is an acidic impurity in natural gas which has to be removed from industrial gas streams, because its presence in pipelines leads to hydrate formation and corrosion in transportation and storage processes. Furthermore, CO_2 is a greenhouse gas which has substantial environmental impacts, such as global warming. The most common methods for capturing CO_2 are physical absorption, chemical absorption and membrane separation [1].

Absorption by a liquid solvent is the most common technique in the gas separation industry. Based on the type of solute-solvent interaction, the solvents used in this process are classified into two categories: physical solvents, which have no chemical reaction with gas, and chemical solvents, which are associated with chemical reactions. When the partial pressure of solute and the amount of gas stream are high, physical solvents are a better choice than chemical solvents. In addition, physical solvents are non-corrosive and non-toxic and can be regenerated only by reducing the pressure and just a little heating [2].

CO_2 can be absorbed in physical solvents in accordance with Henry's law. Therefore, the solubility capacity of a physical solvent is proportional to the gas partial pressure [3]. Henry's law states that at a constant temperature (T), the solubility of gas in liquid is directly proportional to the partial pressure of that gas in equilibrium with that liquid. The constant of proportionality is called Henry's law constant (H_{LC}), which is a macroscopic property of a dilute solution of gas in a solvent. This thermodynamic parameter is defined as the ratio of the fugacity of component i in gas phase (f_i) to the mole fraction of component i in liquid phase (x_i), as the mole fraction of i approaches zero. Henry's law formula is mathematically expressed as [4]:

$$H_i = \lim_{x_i \to 0} f_i / x_i \tag{1}$$

Some theoretical and experimental methods have been developed to determine H_{LC} for CO_2 in different solvents. Gui et al. [5] experimentally measured the solubility and H_{LC} of CO_2 in different physical solvents including alcohols, glycols, ethers and ketones at high pressures. Experimental apparatus is based on a constant volume method; accordingly, a constant volume equilibrium vessel was applied. The temperature and pressure (P) were recorded in equilibrium state and the mole fraction of CO_2 (x) was calculated from material balance. The slope of the P−x curve drawn at constant temperature was equal to H_{LC}. They have found that ketones are the best solvents to absorb CO_2 and indicated that the hydroxyl functional group prevents the solubility of CO_2 in different solvents.

Glycol ethers are another group of physical solvents that have attracted a great deal of attention, due to their high potential for gas absorption. Henni et al. [6] determined the solubility of CO_2 in polyethylene glycol ethers and compared it to commercial solvents such as selexol and sulfolane. The absorption process took place in an autoclave glass cell, and the H_{LC} for CO_2 was calculated by analyzing a small amount of liquid picked up from the bottom of the cell at equilibrium condition. It was demonstrated that polyethylene glycol dimethyl ethers have the greatest ability for CO_2 removal.

Even though many experimental works have been done to determine solubility, it may be difficult to obtain the experimental values. Costly and time-consuming studies along with unavailability of materials have led to the development of different theoretical methods [7]. Some authors have used equations of state (EoS) to

†To whom correspondence should be addressed.
E-mail: riahi@ut.ac.ir

predict the solubility of CO_2 in physical solvents. Jou et al. [8] determined the solubility of carbon dioxide in diethylene glycol (DEG) at 298.15, 323.15, 348.15, 373.15 and 398.15 K and pressures up to 20 MPa. In the beginning, the data were obtained experimentally to correlate them by Peng-Robinson equation of state. The Krichevsky-Ilinskaya equation was used for the calculation of H_{LC} for CO_2 in DEG at different temperatures. Acceptable results were achieved compared to other experimental works. Further investigations of the CO_2 solubility were performed in other physical solvents, e.g., propylene carbonate [9], N-methyl-2-pyrrolidone [10], ester mixture [11], solvents containing carbonyl, acetate and ether groups [12] and ionic liquids [13], which were predicted using different modeling techniques.

One of the most common methods for H_{LC} calculation is the group contribution method, which is used when experimental data are not available. In this complex computational method, a molecule is separated into small fragments and properties of the molecule are predicted by summing up all the fragment contributions [14]. Majer et al. [15] calculated the H_{LC} for different compounds including CO_2, H_2S and CH_4 in water, using the Gibbs energy of hydration. After sophisticated calculations, required parameters for each fragment were obtained to predict H_{LC} of compounds at a temperature range from 300 K to 600 K and three pressure levels (0.1, 40 and 80 MPa). Comparisons indicated good results for the model.

A promising method for estimating a wide number of molecular properties is based on quantitative structure-property relationship (QSPR) theory. In this approach, descriptors are used to create a quantitative relationship between properties of compounds such as Henry's law constant [16], salvation enthalpy [17], aqueous solubility [18] and some other properties [19] and molecular characteristics. Descriptors are numerical values related to the structure and shape of the molecule. Therefore, in QSPR technique a knowledge of molecular structure is sufficient. Many researchers demonstrated that this method provides a powerful tool for predicting H_{LC} for organic chemicals [16,20-25]. English and Carroll [22] presented a QSPR model to predict H_{LC} for 357 pure chemicals in water at ambient temperature. Two models with 10 and 12 descriptors were developed, and for each model, a multivariate linear regression (MLR) and neural network analysis were employed. The comparisons reveal a good agreement between experimental and calculated values of H_{LC}, but the neural network model has a better performance.

Golzar et al. [26] studied solubility of CO_2 and N_2 in five common polymers at different temperatures and pressures, using QSPR model. Genetic algorithm (GA) was applied for descriptor selection, among more than 1600 descriptors. They presented a linear model with seven variables including temperature, pressure and five molecular descriptors, which two of them are related to gas and the other three are related to the polymer. Non-linear models were developed using artificial neural network (ANN) [27] and adaptive neuro fuzzy inference system (ANFIS) [28] techniques. Their results show that the statistical parameters of linear model were not appropriate but these parameters were improved in non-linear models. For example, squared correlation coefficients (R^2) for linear models, ANN and ANFIS are 0.5896, 0.9493 and 0.999 and their absolute relative deviation (ARD) values are 139.07%, 23.49% and 20%, respectively.

Our aim was to develop a robust QSPR model by multiple linear regression and to investigate which one of the molecular characterizations has more influence on determining the H_{LC} of CO_2 values in alcohol, ether, ketone and glycol solvents. Researchers usually attempt to determine the solubility or H_{LC} of compounds in a particular class of physical solvents which have similar properties. Therefore, using various types of physical solvents are one of the novelties of this study. Generally, few studies have been done on Henry's law constant, because this parameter is difficult to measure precisely. Our models are very accurate and powerful, and include simple and interpretable descriptors, which can be calculated without complex computations. Adding temperature as an operational variable to molecular descriptors is another novelty of the work, since there are limited methods to predict H_{LC} as a function of T [15]. To the best of our knowledge, this study is the first research on the H_{LC} of CO_2 in different class of physical solvents using QSPR method, so each result of this study can be considered as a new achievement.

MATERIAL AND METHOD

1. Data Set and Descriptors Generation

Experimental values of H_{LC} for CO_2 in 22 solvents at various temperatures were taken from Gui et al. [5] and Henni et al [6]. Three out of fourteen experimental data were disregarded from Henni et al. [6]. The H_{LC} in one of the solvents (ethylene glycol monomethyl ether) has the same value in both mentioned papers. This value was noy considered in the Henni et al. work, because it was reported in a smaller temperature range. One of the eliminated H_{LC} values was related to a polymer solvent (Polyethelene glycol dimethyl ether) and another solvent had a nitrogen atom in its structure. Note that this investigation was done for solvents containing only carbon, oxygen and hydrogen atoms and compounds that do not have cyclic structures. In this modeling, experimental values were converted to natural logarithm (ln H), and their values ranged between 1 and 4. First, a model was developed for 22 compounds at constant temperature, and in the next step, by adding T as an operational variable, the number of data changes from 22 to 77 and H_{LC} can be predicted at different temperatures.

The molecular structures of compounds were drawn using Hyperchem software [29] and pre-optimized using RM1 molecule mechanics Force Field (Polak-Ribiere algorithm) and with a gradient norm limit of 0.01 kcal/(A^o·mol). Further optimization was done by using Gaussian [30] on the basis of density functional Ttheory (DFT) at level of B3LYP and 6-31 ++ G (d. p). In this study, output files of Gaussian were fed into the Dragon 6 software [31] for descriptors generation. This software can calculate 4885 molecular descriptors, but since only five kinds of descriptors, including constitutional indices, topological indices, functional group counts, atom-centered fragments and molecular properties were selected, the number of descriptors decreased to 70.

Variable selection was done using GA [32], which led to the creation of several linear models. Descriptors that appeared in the ten best models, along with their types and definitions, have been

Table 1. The type and definition of descriptors that appear in different QSPR models

Descriptors	Type	Definition
ZM1V	Topological indices	First Zagreb index by valance vertex degrees
ZM2V	Topological indices	Second Zagreb index by valance vertex degrees
BAC	Topological indices	Balaban centric index
DELS	Topological indices	Molecular electrotopological variation
nROH	Functional groups	Number of hydroxyl groups
nROR	Functional groups	Number of ethers
MW	Constitutional indices	Molecular weight
nO	Constitutional indices	Number of oxygen atoms
nHet	Constitutional indices	Number of heteroatoms
SCBO	Constitutional indices	Sum of conventional bond orders (H-depleted)
TPSA (NO)	Molecular properties	Topological polar surface area using N, O polar contribution

summarized in Table 1.

2. Model Development and Validation

Different QSPR models for prediction of H_{LC} for CO_2 in physical solvents were developed under two conditions: at fixed temperature and at variable temperatures. In the first condition, the best descriptors for the solvent molecules are selected at constant temperature, and temperature as an operational variable is added to these descriptors in the second condition.

The performance and predictive potential of the developed models were assessed via several internal and external validation techniques, including leave one out-cross validation (LOO-CV), classification of data into training set and test set, clustering and application of main model on new external data set.

The LOO-CV is an internal validation method by which a compound is held out of the data set and a model is developed by the remaining data to predict H_{LC} for the eliminated compound. This procedure was repeated for all the solvents and predicted values were reported. In the second technique, experimental data in overall state (22 compounds) were classified into training set (80% of data or 17 compounds) and test set (20% of data or five compounds) at constant temperature by principal component analysis (PCA) method [33]. Finally, a model was built based on the training set and validated using a test set.

After classification of data into training and test sets, the best models with one and two descriptors were developed. Eq. (2) is a model with one descriptor and Eq. (3) to Eq. (5) are models with two descriptors at fixed temperature (298.15 K):

$$\ln(H)=0.543nROH^2+1.65069 \quad (2)$$

$$\ln(H)=0.46894nROH^2-0.00515ZM1v+2.21168 \quad (3)$$

$$\ln(H)=0.45935nROH^2-0.00838ZM2v+2.17652 \quad (4)$$

$$\ln(H)=0.47252nROH^2-0.0046MW+2.23264 \quad (5)$$

The next step is adding temperature to the Eq. (5), which is a simple and interpretable model. A new model with three descriptors, which is the main model of this study at variable temperature (Eq. (6)) was obtained by MLR method:

$$\ln H=0.01439T-0.00479MW+0.46378nROH^2-2.03786 \quad (6)$$

Some of the compounds in the data set that are more similar to each other, form a group which is called cluster, so clustering technique aims to divide compounds into different groups. In this study, the data are located in three clusters and an equation has been provided for each cluster to predict H_{LC} for CO_2 (Eq. (7) to Eq. (9)).

$$\ln H=-0.00749ZM2v+2.0954 \quad (7)$$

$$\ln H=-0.00414ZM1v+2.0631 \quad (8)$$

$$\ln H=-0.00563ZM1v+2.7275 \quad (9)$$

In the last technique, an external data set was taken from different literatures to check the ability of the main model for the prediction of H_{LC}.

Different statistical parameters such as R^2, adjusted R^2 (R^2_{adj}), leave-one-out cross validated correlation coefficient (Q^2_{LOO}), fisher function (F) and standard error (S) are defined by Eq. (10) to Eq. (14). The best model is the one which has the closest values of R^2, R^2_{adj}, Q^2_{LOO} to one, highest value of F and closest value of S to zero. In addition, relative deviation (RD%) value was calculated to estimate predictive ability of models, which is defined by Eq. (15).

$$R^2=1-\frac{RSS}{TSS}=1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y})^2} \quad (10)$$

$$R^2_{adj}=1-(1-R^2)\left(\frac{n-1}{n-p-1}\right) \quad (11)$$

$$Q^2_{LOO}=1-\frac{PRESS}{TSS}=1-\frac{\Sigma(y_i-\hat{y}_{cal-CV})^2}{\Sigma(y_i-\bar{y})^2} \quad (12)$$

$$F=\frac{MSS/df_m}{RSS/df_e}=\frac{\Sigma(\hat{y}_i-\bar{y})^2/p}{\Sigma(y_i-\hat{y}_i)^2/(n-p-1)} \quad (13)$$

$$s=\sqrt{\frac{\Sigma_{i=1}^{n}(\hat{y}_i-y_i)^2}{n-p-1}} \quad (14)$$

$$RD(\%)=\frac{|(y_{exp})-(y_{pre})|}{y_{exp}}\times 100 \quad (15)$$

Reliability of the proposed model is further determined by Golbraikh and Tropsha criteria [34] which can be stated as follows:

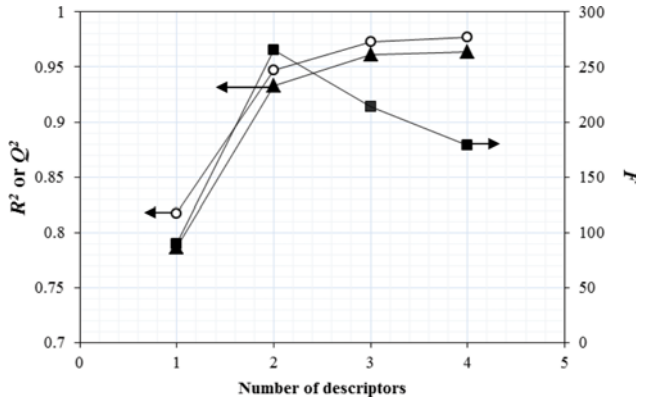

Fig. 1. Effect of number of descriptors on the R^2 (-○-), Q^2 (-▲-) and F (-■-) values.

$$R^2 > 0.6 \quad (16)$$

$$Q^2_{LOO} > 0.5 \quad (17)$$

$$\frac{(R^2 - R^2_0)}{R^2} < 0.1 \quad (18)$$

$$k = \frac{\Sigma y_i^{exp} y_i^{pre}}{\Sigma (y_i^{pre})^2}, \quad 0.85 < k < 0.15 \quad (19)$$

where R_0 is the correlation coefficient when the regression line passes through the origin and k represents the corresponding slope.

RESULTS AND DISCUSSION

Breaking point plot is an appropriate tool which can be used to optimize the number of descriptors in QSPR model. In this plot, the effect of the number of descriptors on statistical parameters such as R^2, Q^2_{LOO} and F is investigated as shown in Fig. 1. Developing a model with two descriptors is sufficient, while adding further descriptors has a little effect on R^2 and Q^2_{LOO} and strongly decreases F values.

Our main objective was to construct a model with simple descriptors by MLR method. First, a mono-variable model was presented. According to Eq. (2), there is a direct relationship between H_{LC} and square of number of hydroxyl functional group ($nROH^2$). The plot of lnH versus $nROH^2$ values shows that the compounds are divided into three clusters (see Fig. 2). The solvents that are in each cluster have equal number of hydroxyl functional groups (nROH). The first cluster, which contains seven molecules, has nROH=0 and second cluster with nROH=1 has 13 molecules.

The last cluster with nROH=2 was not used in modeling because it only contains two molecules. It is clear that H_{LC} values increase with the increasing of nROH. The descriptors that appear in Eq. (7) to Eq. (9) (zM1v and ZM2v) are similar to overall equation descriptors. This means that the overall and cluster equations confirm each other.

As mentioned, Eq. (3) to Eq. (5) represent the models with two simple descriptors to predict H_{LC} for CO_2 in physical solvents by MLR method. These descriptors do not require any computational software and can be calculated easily.

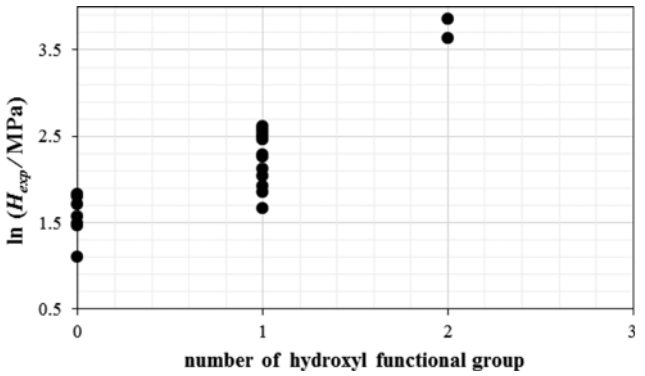

Fig. 2. Plot of lnH_{xep} vs. number of hydroxyl functional group in the molecular structure.

Table 2. Statistical parameters for different models in overall state

Equations	R^2	R^2_{adj}	Q^2_{LOO}	F	S
Eq. (2)	0.8172	0.808	0.7863	89.39	0.291
Eq. (3)	0.9664	0.9629	0.9563	273.49	0.128
Eq. (4)	0.9654	0.9618	0.9578	265.08	0.130
Eq. (5)	0.947	0.9414	0.9332	169.61	0.161
Eq. (6)	0.9318	0.929	0.9251	332.61	0.170
Eq. (7)	0.9529	0.9435	0.87	101.13	0.06
Eq. (8)	0.9336	0.9203	0.8264	70.28	0.071
Eq. (9)	0.8004	0.7823	0.7374	44.11	0.146

Table 2 reports statistical parameters of different models, when all compounds are in overall state. Although mono-variable equation (Eq. (2)) has acceptable statistical parameters, it is not an appropriate model, because some of the compounds have the same values of nROH, whereas their experimental values are different.

The results of experimental and predicted values of different QSPR models at fixed temperature are presented in Table 3.

Prior to the calculation of predicted values, PC2 versus PC1 values were plotted in Fig. 3 to select a test set. The PCA values were calculated from 11 descriptors listed in Table 1. Each axis of Fig. 3 shows the distribution of descriptors, and 84.4 and 12.8 percent of all data have been located in PC1 and PC2, respectively. Five molecules were selected for testing based on these PCA values at fixed temperature, which are marked with an asterisk in Table 3.

Since the goal of this study was a QSPR model with simple and interpretable descriptors, Eq. (5) was selected as the main model at constant temperature, and T was added to descriptors in this equation.

In a QSPR model, descriptors should not be related to each other, and for this reason correlations between ZM1v, ZM2v, MW and $nROH^2$ descriptors were examined and reported in Table 4. There is no linear relationship between these descriptors in an equation and all descriptors are independent of $nROH^2$.

Another issue that helps to interpret the main model is determining the mean effect, which shows the effect of each descriptor in model [35]. As can be seen from Fig. 4, the square of number of hydroxyl functional group has higher influence on predicting H_{LC} for CO_2 than the molecular weight descriptor (MW), at con-

Table 3. Experimental and predicted values at fixed temperature (298.15 K)

No.	Solvent	ln (H_{exp}/MPa)	Eq. (2)	Eq. (3)	Eq. (4)	Eq. (5)
			ln (H_{pre}/MPa)	ln (H_{pre}/MPa)	ln (H_{pre}/MPa)	ln (H_{pre}/MPa)
1	Methanol	2.612	2.19	2.55	2.59	2.56
2	Ethanol	2.587	2.19	2.53	2.54	2.49
3	n-Propanol*	2.553	2.19	2.51	2.50	2.43
4	n-Butanol	2.510	2.19	2.49	2.47	2.36
5	n-Pentanol*	2.464	2.19	2.46	2.43	2.30
6	Ethylene glycol (EG)	3.864	3.82	3.79	3.81	3.84
7	Propylene glycol (PG)	3.641	3.82	3.76	3.73	3.77
8	Acetone	1.826	1.65	1.93	1.91	1.97
9	2-Butanone	1.801	1.65	1.91	1.86	1.90
10	Ethylene glycol monomethylether (EGMME)	2.258	2.19	2.32	2.37	2.36
11	Ethylene glycol monoethylether (EGMEE)*	2.125	2.19	2.30	2.30	2.29
12	Diethylene glycol monomethylether (DGMME)	1.856	2.19	2.09	2.13	2.15
13	Triethylene glycol monomethylether (TrGMME)*	1.917	2.19	1.87	1.90	1.95
14	Ethylene glycol dimethylether (EGDME)	1.705	1.65	1.75	1.71	1.69
15	Diethylene glycol dimethylether (DGDME)	1.568	1.65	1.56	1.61	1.62
16	Triethylene glycol dimethylether (TrGDME)	1.482	1.65	1.34	1.37	1.41
17	Tetraethylene glycol dimethylether (TeGDME)	1.098	1.65	1.11	1.14	1.21
18	Diethylene glycol monoethylether (DGMEE)	2.041	2.19	2.07	2.07	2.09
19	Diethylene glycol diethylether (DGDEE)*	1.458	1.65	1.52	1.47	1.49
20	Ethylene glycol monobutylether (EGMBE)	2.493	2.19	2.26	2.23	2.16
21	Diethylene glycol monobutylether (DGMBE)	2.282	2.19	2.03	2.00	1.96
22	Triethylene glycol monobutylether (TrGMBE)	1.667	2.19	1.81	1.76	1.76

*Data used for the test set

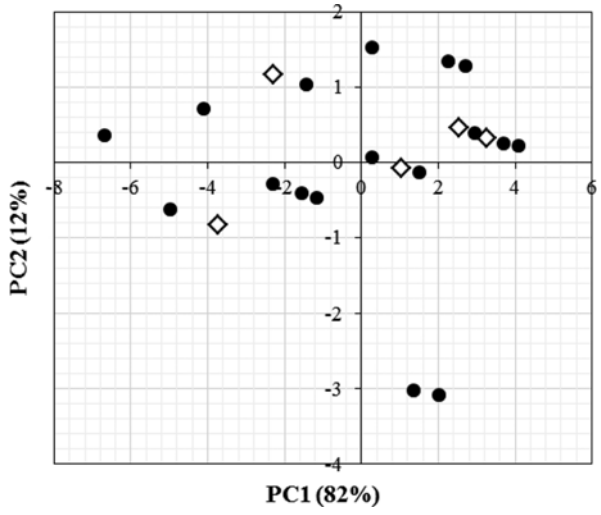

Fig. 3. PC2 versus PC1 in training (●) and test sets (◇).

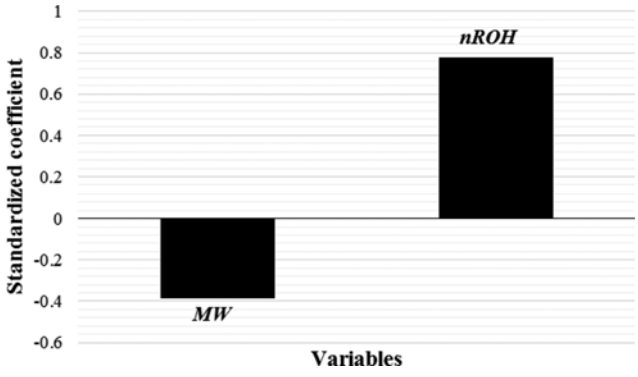

Fig. 4. Mean effect of main model descriptors at fixed temperature (298.15 K).

Table 4. Correlation matrix of descriptors that are appeared in different models

Descriptor	$nROH^2$	MW/Kg·mol^{-1}	ZM1v	ZM2v
$nROH^2$	1.000	0.117	0.101	0.131
MW/Kg·mol^{-1}	0.117	1.000	0.939	0.965
ZM1v	0.101	0.939	1.000	0.982
ZM2v	0.131	0.965	0.982	1.000

stant temperature.

One of the most important factors which help to determine the quality and reliability of a QSPR model and to identify outliers is applicability domain. For this reason, Williams plot (the plot of standardized residuals versus leverage values or hat values (h)) has been illustrated in Fig. 5. The difference between the observed value and the predicted value indicates the residual (E). The standard residual error (E′) and hat value are defined as follows:

$$E' = \frac{E}{RMSE\sqrt{1-h}} \tag{20}$$

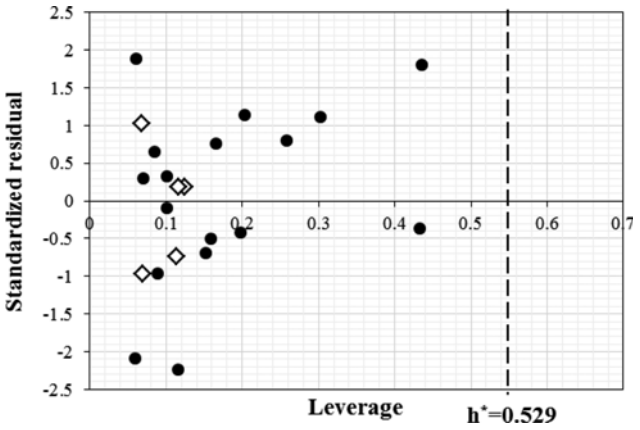

Fig. 5. Williams plot of main model at fixed temperature (training set: ●, test set: ◇).

$$h = X(X^T X)^{-1} X^T \quad (21)$$

where X and X^T are descriptor matrix and transposed version of descriptor matrix, respectively. Critical hat value is also calculated, using the following equation:

$$h^* = 3(p+1)/n \quad (22)$$

where n is the number of data in training set and p is the number of model descriptors. When the hat value of a compound is higher than its critical value ($h>h^*$) and the standardized residual value is higher than +3 or −3, the predicted value for this compound is unreliable. Fig. 5 shows that all compounds are in the applicability domain and there are no outliers [36,37].

Addition of temperature to Eq. (5) is the great advantage of the proposed study. This new model is shown in Eq. (6) and its statistical parameters are reported in Table 2. This table indicates that H_{LC} for different solvents at various temperatures can be predicted with high accuracy. The predicted values of Eq. (6) have been reported and compared with experimental data in Table 5. Test set molecules were randomly selected in a way that each compound was presented in one of the temperatures.

The high predictive abilities of Eqs. (5) and (6) are shown in Figs. 6 and 7, respectively. In these figures, natural logarithms of experimental values of H_{LC} (ln (H_{exp})) are plotted versus natural logarithms of predicted values of H_{LC} (ln (H_{pre})).

Validation and statistical parameters of Eqs. (5) and (6) for both training and test sets are given in Table 6 and it can be proven that these models have very good statistical qualities.

The plots of residual as a function of (H_{exp}) are depicted in Figs. 8 and 9, respectively, for Eqs. (5) and (6). The residual data are nor-

Table 5. Experimental and predicted values at variable temperature

Solvent	T/K	ln (H_{exp}/MPa)	ln (H_{pre}/MPa)	Solvent	T/K	ln (H_{exp}/MPa)	ln (H_{pre}/MPa)
Methanol	288.15	2.449	2.42	EGMME	288.15	1.856	2.21
	298.15	2.612	2.56*		298.15	2.258	2.35
	308.15	2.785	2.71		308.15	2.517	2.49*
	318.15	2.990	2.85		318.15	2.727	2.64
Ethanol	288.15	2.364	2.35	EGMEE	288.15	1.789	2.14*
	298.15	2.587	2.49		298.15	2.125	2.28
	308.15	2.758	2.64		308.15	2.355	2.43
	318.15	2.915	2.78*		318.15	2.505	2.57
n-Propanol	288.15	2.358	2.28	DGMME	298.15	1.856	2.14
	298.15	2.553	2.43		313.15	2.128	2.35*
	308.15	2.713	2.57*		333.15	2.312	2.64
	318.15	2.878	2.71				
n-Butanol	288.15	2.333	2.22	TrGMME	298.15	1.917	1.93
	298.15	2.510	2.36		313.15	2.041	2.14*
	308.15	2.686	2.50		333.15	2.104	2.43
	318.15	2.805	2.65*				
n-Pentanol	288.15	2.268	2.15*	EGDME	298.15	1.705	1.68*
	298.15	2.464	2.29		313.15	1.775	1.90
	308.15	2.631	2.44		333.15	1.960	2.19
	318.15	2.767	2.58				
EG	288.15	3.708	3.67	DGDME	298.15	1.569	1.61
	298.15	3.864	3.81		313.15	2.104	1.82
	308.15	3.983	3.95		333.15	2.476	2.11*
	318.15	4.105	4.10*				
PG	288.15	3.409	3.60	TrGDME	298.15	1.482	1.40*
	298.15	3.641	3.74*		313.15	1.792	1.61
	308.15	3.821	3.89		333.15	2.186	1.90
	318.15	3.962	4.03				

Table 5. Continued

Solvent	T/K	ln (H_{exp}/MPa)	ln (H_{pre}/MPa)	Solvent	T/K	ln (H_{exp}/MPa)	ln (H_{pre}/MPa)
Acetone	288.15	1.642	1.83	TeGDME	298.15	1.099	1.18
	298.15	1.826	1.97*		313.15	1.459	1.40
	308.15	1.979	2.12		333.15	1.872	1.69*
	318.15	2.101	2.26				
2-Butanone	288.15	1.608	1.76	DGMEE	298.15	2.041	2.07
	298.15	1.801	1.91*		313.15	2.219	2.29*
	308.15	1.928	2.05		333.15	2.370	2.57
DGDEE	298.15	1.459	1.47*	DGMBE	298.15	2.282	1.94*
	313.15	1.686	1.69		313.15	2.416	2.15
	333.15	1.974	1.98		333.15	2.549	2.44
EGMBE	298.15	2.493	2.15	TrGMBE	298.15	1.667	1.72
	313.15	2.603	2.36*		313.15	1.917	1.94
	333.15	2.754	2.65		333.15	2.104	2.23*

*Data used for the test set

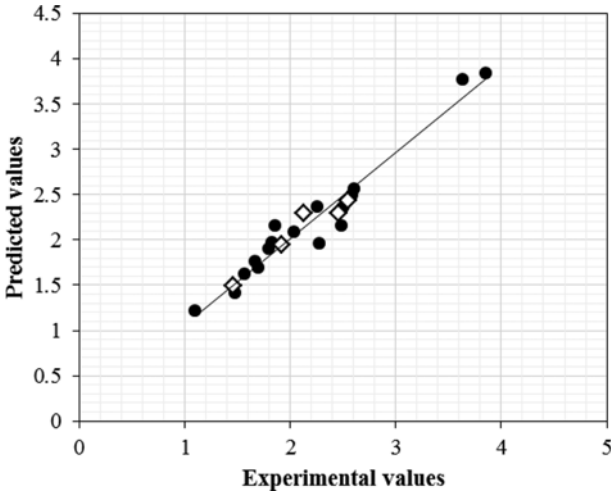

Fig. 6. Plot of experimental values vs. predicted values at fixed temperature (298.15 K) (training set: ●, test set: ◇).

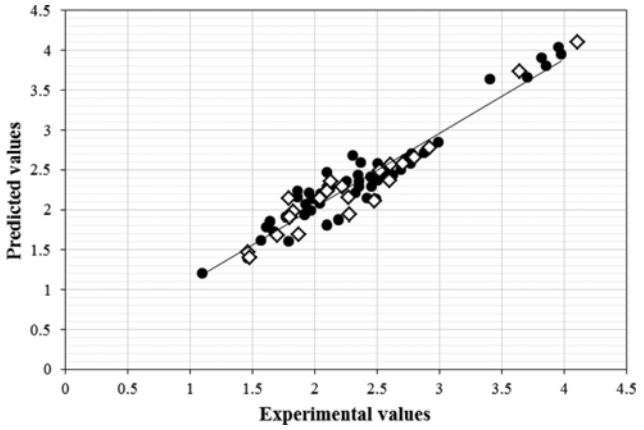

Fig. 7. Plot of experimental values vs. predicted values at variable temperatures (training set: ●, test set: ◇).

mally distributed above and below the zero line, and this issue reflects the lack of systematic error in the presented models. According to Fig. 8, EGMBE, DGMBE and DGMME have the highest residual (prediction error), in comparison with other compounds.

In addition to above methods, new models which were independent of the number of hydroxyl functional group were developed

Table 6. Validation and statistical parameters of Eqs. (5) and (6)

Equation	set	n	R^2	R^2_{adj}	F	S	k	RMSE*	ARD**
Eq. (5)	Train	17	0.961	0.955	133.25	0.174	0.9996	0.155	6.04
	Test	5	0.93	0.86	8.8	0.183	1.008	0.116	4.46
	Overall	22	0.947	0.941	169.61	0.161	1	0.149	6.29
Eq. (6)	Train	55	0.934	0.93	240.24	0.169	0.9998	0.161	6.03
	Test	22	0.93	0.915	76.87	0.193	1.0103	0.174	6.38
	Overall	77	0.93	0.93	332.61	0.17	0.9997	0.165	6.48

$$^{*}\mathrm{RMSE}=\sqrt{\frac{\Sigma_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}}$$

$$^{**}\mathrm{ARD}=\frac{1}{n}\Sigma\left(\frac{y_i^{exp}-y_i^{pre}}{y_i^{exp}}\right)\times 100$$

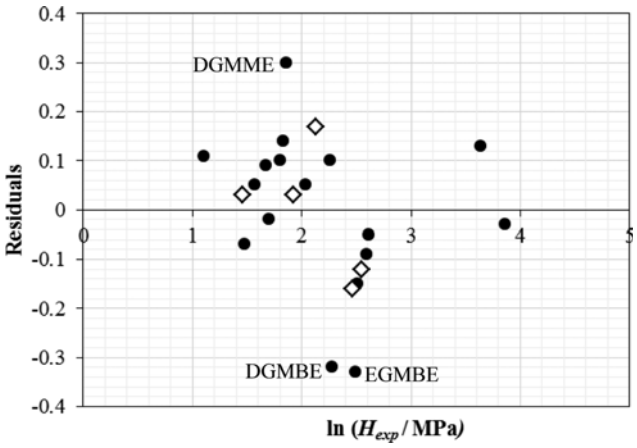

Fig. 8. Plot of residual values vs. lnH_{exp} at fixed temperature (training set: ●, test set: ◇).

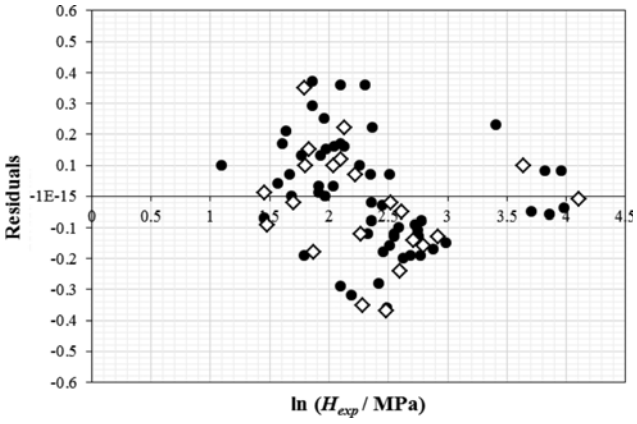

Fig. 9. Plot of residual values vs. lnH_{exp} at variable temperatures (training set: ●, test set: ◇).

for cluster 1 (Eqs. (7) and (8)) and cluster 2 (Eq. (9)) with only one descriptor, and their statistical parameters have also been listed in Table 2. According to Tables 7 and 8, which show ln (H_{exp}) and ln (H_{pre}) values for clusters 1 and 2, predicted values are near to their experimental values; therefore, these models have high prediction capability.

External validation is another method that helps to analyze the model. Some experimental data for CO_2 Henry's law constant have been reported in the literature [3,38,39] in different solvents and various operating conditions. Note that these data have not been used in the development of the model. The estimation results of Eqs. (5) and (6) are tabulated in Table 9. These main equations have enough ability and accuracy to predict H_{LC} for various solvents at different temperatures.

Table 7. Experimental and predicted values of cluster 1

		Eq. (7)	Eq. (8)
Solvent	ln (H_{exp}/MPa)	ln (H_{pre}/MPa)	ln (H_{pre}/MPa)
Acetone	1.83	1.85	1.84
2-Butanone	1.80	1.81	1.82
EGDME	1.70	1.67	1.69
DEGDME	1.57	1.58	1.54
TEGDME	1.48	1.37	1.35
TeEGDME	1.10	1.16	1.17
DEGDEE	1.46	1.46	1.50

Y-Randomization test is a simple method for the evaluation of a model and figuring out whether the proposed model was obtained by chance. This technique is based on rearranging the connection between target variable (ln H) and descriptors [40]. This procedure was repeated in ten iterations, and statistical parameters (R^2) are reported in Table 10. If at any stage the R^2 value is less than the original model, the model will be applicable and reliable. Therefore, the selected QSPR model does not suffer from chance correlation.

As mentioned, in addition to being simple, a QSPR model should be interpretable. Researchers proved that the hydroxyl functional group in the molecular structure of amine solvents is an effective parameter on the absorption capacity of CO_2 in different solvents [41], which also has an important role in the final model of this work. Singh et al. [42] reached the result that as nROH in the molecular structure increases, absorption capacity of CO_2 in chemical solvents also increase, but in this work by increasing nROH, H_{LC} of CO_2 in physical solvents increases and absorption capacity decreases; the reason is that gas absorption in physical solvents occurs in two steps. At first, solvent-solvent molecules must break and provide the required cavities; in the second step, these cavities must be filled with gas molecules [5]. In this study, when the hydroxyl functional group increased more intermolecular hydrogen bonds could be formed in solvents, and this matter prevents the creation of required cavities. Thus, H_{LC} increases by increasing the number of hydroxyl functional group.

The molecular weight is another simple descriptor in the main

Table 8. Experimental and predicted values of cluster 2

Solvent	ln (H_{exp}/MPa)	ln (H_{pre}/MPa)	Solvent	ln (H_{exp}/MPa)	ln (H_{pre}/MPa)
Methanol	2.61	2.58	DEGMME	1.86	2.08
Ethanol	2.59	2.55	TEGMME	1.92	1.84
Propanol	2.55	2.53	DEGMEE	2.04	2.06
Butanol	2.51	2.51	EGMBE	2.49	2.26
Pentanol	2.46	2.49	DEGMBE	2.28	2.02
EGMEE	2.13	2.31	TEGMBE	1.67	1.77
EGMME	2.258	2.33			

Table 9. External validation using Eqs. (5) and (6)

Condition	Solvent	T/K	ln (H_{exp}/MPa)	ln (H_{Pre}/MPa)	AD*	RD
Fixed temperature	Dimethyl carbonat	298.47	1.837	1.818	0.018	1.019
	Propylene carbonat	298	2.143	1.763	0.379	17.73
	Methyl cyanoacetate	298	2.189	1.777	0.412	18.83
	N-formyl morpholine	298	1.966	1.703	0.263	13.37
	Selexol	298	1.272	1.009	0.263	20.67
Variable temperature	1,4 Butylene glycol	303.15	4.874	3.748	1.126	23.10
		323.15	5.019	4.036	0.983	19.59
		373.15	5.315	4.755	0.559	10.53
		398.15	5.414	5.115	0.299	5.523
		423.15	5.518	5.475	0.043	0.784
						ARD=11.9
	Dimethyl carbonat	280.7	1.409	1.569	0.161	11.42
		289.49	1.617	1.696	0.079	4.911
		298.47	1.837	1.825	0.011	0.618
		307.84	2.009	1.960	0.048	2.415
		317.86	2.155	2.104	0.050	2.336
		327.66	2.288	2.245	0.042	1.849
						ARD=3.92
	Propylene carbonat	298	2.143	1.761	0.382	17.81
		303	2.253	1.833	0.419	18.63
		313	2.434	1.977	0.457	18.76
		323	2.517	2.121	0.396	15.73
		333	2.734	2.265	0.469	17.15
		343	2.778	2.408	0.369	13.28
						ARD=16.9
	Methyl cyanoacetate	298	2.189	1.776	0.413	18.88
		303	2.206	1.847	0.358	16.24
		313	2.416	1.991	0.424	17.56
		323	2.602	2.135	0.466	17.93
		333	2.660	2.279	0.381	14.31
						ARD=17
	N-formyl morpholine	298	1.966	1.698	0.267	13.58
		303	2.092	1.771	0.321	15.35
		313	2.223	1.915	0.308	13.86
		323	2.407	2.058	0.348	14.47
		333	2.587	2.202	0.384	14.86
		343	2.701	2.346	0.354	13.13
						ARD=14.2
	Selexol	298	1.272	0.976	0.295	23.25
		303	1.374	1.048	0.326	23.71
		313	1.541	1.192	0.348	22.64
		323	1.726	1.336	0.390	22.59
		333	1.879	1.479	0.399	21.24
						ARD=22.7

*$AD=|y_{exp}-y_{pre}|$

model. The researchers have proven that increasing molecular weight within each class of solvents changes the polarity of solvents towards the lower ones [43]. On the other hand, CO_2 is a non-polar molecule and dissolves best in non-polar solvents. It is expected that molecular weight and solubility have a direct relationship [44] and solubility decreases with increasing solvent polarity [45]. Since physical solvents in this data set are in different classes such as alcohol, glycol, ketone and glycol ether, the change of H_{LC} as a function of MW has been shown in Fig. 10. In alcohols, glycols and ketones, H_{LC} of CO_2 declines (and solubility increases) with increasing molec-

Table 10. Randomization test results

Iteration	R^2_{train}
1	0.023
2	0.018
3	0.024
4	0.017
5	0.009
6	0.005
7	0.018
8	0.003
9	0.005
10	0.009

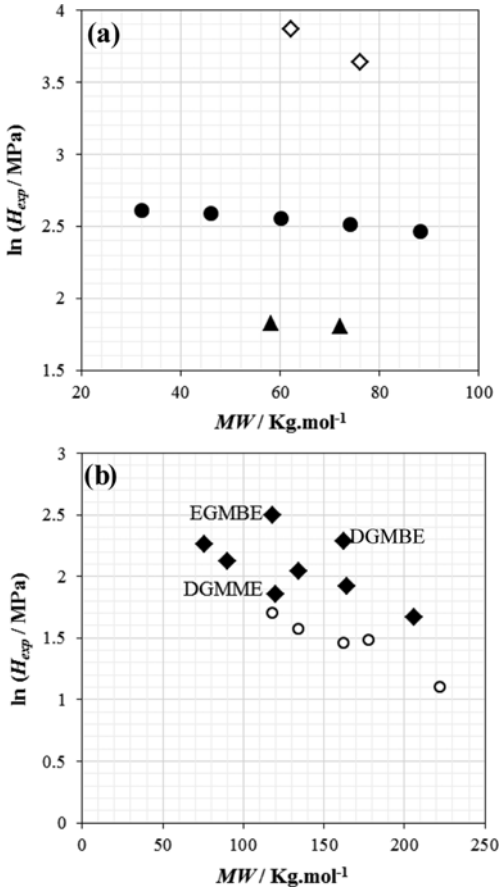

Fig. 10. Plot of ln H_{exp} vs. molecular weight for (a) ketone (▲), alcohol (●) and glycol (◇) and (b) glycol ether (nROH=0) (○) and glycol ether (nROH=1) (◆).

ular weight, as illustrated in Fig. 10(a). Similar behavior is observed in glycol ethers with nROH=0, but there are some contradictions in this class of solvent with nROH=1 (see Fig. 10(b)). Three of the solvents (EGMBE, DGMBE and DGMME) do not follow the above-mentioned principle, which can be attributed to the experimental error. These solvents have shown the most deviation in our model (Eq. (5)), which is indicated in Fig. 8. In Fig. 10, the effect of nROH is also clear and the group of solvents with higher hydroxyl functional group have higher H_{LC} values.

CONCLUSIONS

A QSPR approach was performed to predict H_{LC} for CO_2 in 22 physical solvents at different temperatures, by MLR method. $nROH^2$ and MW are molecular descriptors that appeared in the main models which the former had more effect on H_{LC} prediction than the latter. In mono-variable model, $nROH^2$ had strong positive correlation with H_{LC} (R^2=0.8172). Both molecular descriptors were uncomplicated and could be calculated easily. According to nROH values, the data set was divided into three clusters of which seven compounds were placed in the first cluster with no hydroxyl functional group and 13 compounds were placed in the second cluster which had one hydroxyl functional group in their structures. The obtained results from different validation methods reveal that the proposed model has high predictive ability in prediction of H_{LC} for CO_2. ARD values in training and test set were 6.04% and 4.46%, respectively. The effect of temperature on H_{LC} was investigated when T was added to the other two descriptors. In this model, ARD values were equal to 6.03% and 6.38% for training and test set, respectively, which represents the high accuracy of the model. According to the proposed models, the best physical solvents for capturing CO_2 are the ones that have the highest molecular weight and no hydroxyl functional group. In addition, external validation method, which is a powerful tool for demonstrating the applicability of the model, was done. Applying our model, the predicted values for all the external data sets were acceptable. Therefore, it can be used confidently for different solvents.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of Institute of Petroleum Engineering, University of Tehran.

LIST OF SYMBOLS

CO_2 : carbon dioxide
QSPR: quantitative structure property relationship
H_{LC} : Henry's law constant
MLR : multilinear regression
H : Henry's law constant for CO_2
ARD : absolute relative deviation
RD : relative deviation
RMSE : root mean square error
LOO-CV : leave one out- cross validation
PCA : principal component analysis
MW : molecular weight
nROH : number of hydroxyl functional group
ZM1v : first Zagreb index by valance vertex degrees
ZM2v : second Zagreb index by valance vertex degrees
GA : genetic algorithm
f : fugacity
x : mole fraction
T : temperature
P : pressure
F : fisher function
S : standard error

n : number of compound
p : number of descriptor
h : hat value
h* : critical hat value
X : descriptor matrix
E : Residual
E′ : standard residual

Subscripts
i : compound
LOO : leave one out
exp : experimental value
pre : predicted value

REFERENCES

1. A. Comite, C. Costa, R. D. Felice, P. Pagliai and D. Vitiello, *Korean J. Chem. Eng.*, **32**, 2 (2015).
2. A. Davea, M. Dave, Y. Huanga, S. Rezvani and N. Hewitt, *Int. J. Greenh Gas Con.*, **49**, 436 (2016).
3. X. Gui, Z. G. Tang and W. Fei, *J. Chem. Eng. Data*, **55**, 3736 (2010).
4. H. DeVoe, *Thermodynamics and chemistry*, 2nd Ed., Prentice-Hall (2014).
5. X. Gui, Z. G. Tang and W. Fei, *J. Chem. Eng. Data*, **56**, 2420 (2011).
6. A. Henni, P. Tontiwachwuthikul and A. Chakma, *Can. J. Chem. Eng.*, **83**, 358 (2005).
7. B. Gwinner, D. Roizard, F. Lapicque, E. Favre, R. Cadours, P. Boucot and P. L. Carrette, *Ind. Eng. Chem. Res.*, **45**, 5044 (2006).
8. F. Y. Jou, F. D. Otto and A. E. Mather, *Fluid Phase Equilibr.*, **175**, 53 (2000).
9. F. Y. Jou, A. E. Mather and K. A. G. Schmidt, *J. Chem. Eng. Data*, **60**, 3738 (2015).
10. M. R. Bohloul, A. Vatani and S. M. Peyghambarzadeh, *Fluid Phase Equilibr.*, **365**, 106 (2014).
11. X. Gui, Z. Tang and W. Fei, *LCE*, **2**, 26 (2011).
12. M. B. Miller, D. L. Chen, D. R. Luebke, J. K. Johnson and R. M. Enick, *J. Chem. Eng. Data*, **56**, 1565 (2011).
13. S. Baj, T. Krawczyk, A. Dąbrowska, A. Siewniak and A. Sobolewski, *Korean J. Chem. Eng.*, **32**, 11 (2015).
14. S. A. Brockbank, N. F. Giles, R. L. Rowley and W. V. Wilding, *J. Chem. Eng. Data*, **59**, 1052 (2014).
15. V. Majer, J. Sedlbauer and G. Bergin, *Fluid Phase Equilib.*, **272**, 65 (2008).
16. P. R. Duchowicz, J. C. M. Garro and E. A. Castro, *Chemometr. Intell. Lab.*, **91**, 133 (2008).
17. H. Golmohammadi, Z. Dashtbozorgi and W. E. Acree Jr., *Struct. Chem.*, **24**, 1799 (2013).
18. J. Kim, D. H. Jung, H. Rhee, S. H. Choi, M. J. Sung and W. S. Choi, *Korean J. Chem. Eng.*, **25**(4), 873 (2008).
19. M. Shacham, M. Elly, I. Paster and N. Brauner, *Chem. Eng. Sci.*, **97**, 186 (2013).
20. D. R. O'Loughlin and N. J. English, *Chemosphere*, **127**, 1 (2015).
21. F. Ghargheizi, P. Ilani-Kashkouli, S. A. Mirkhani, N. Farahani and A. H. Mohammadi, *Ind. Eng. Chem. Res.*, **51**, 4764 (2012).
22. N. J. English and D. G. Carroll, *J. Chem. Inf. Comput. Data*, **41**, 1150 (2001).
23. J. Xu, H. Zhang, L. Wang, W. Ye, W. Xu and Z. Li, *Fluid Phase Equilib.*, **291**, 111 (2010).
24. S. Sahoo, S. Patel and B. K. Mishra, *Thermochim. Acta*, **512**, 273 (2011).
25. M. Goodarzi, E. V. Ortiz, L. S. Coelho and P. R. Duchowicz, *Atmos. Environ.*, **44**, 3179 (2010).
26. K. Golzar, S. Amjad-Iranagh and H. Modarress, *Measurement*, **46**, 4206 (2013).
27. D. Shiffman, *The nature of code: Simulating Natural Systems with Processing*, The nature of code, first Ed., U.S.A. (2012).
28. S. Rajasekaran and G. A. Vijayalakshmi Pai, *Neural Networks, Fuzzy Logic and Genetic Algorithms: Synthesis and Applications (Computer)*, PHI Learning (2011).
29. http://www.hyper.com.
30. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, € O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian 98." Gaussian Inc., Pittsburgh PA, Gaussian Inc., Pittsburgh PA (1998).
31. S. R. L. Talete, *Dragon for Windows* (Software for Molecular Descriptor Calculation), Version 6 (2013).
32. S. N. Sivanandam and S. N. Deepa, *Introduction to Genetic Algorithms*, Springer Science & Business Media (2007).
33. I. T. Jolliffe, *Principal Component Analysis*, Springer, New York (2002).
34. A. Golbraikh and A. Tropsha, *J. Mol. Graph. Model*, **20**, 269 (2002).
35. XLSTAT software, *XLSTAT-CCR module*, Trial Version (2013).
36. P. Gramatica, *QSAR Comb. Sci.*, **26**, 694 (2007).
37. N. Minovski, S. Zuperl, V. Drgan and M. Novi, *Anal. Chim. Acta*, **759**, 28 (2013).
38. A. C. Galvão and A. Z. Francesconi, *J. Supercrit. Fluid*, **51**, 123 (2009).
39. Y. Xu, R. P. Schutte and L. G. Hepler, *Can. J. Chem. Eng.*, **70**, 569 (1992).
40. C. H. Rücker, G. Rücker and M. Meringer, *J. Chem. Inf. Model*, **47**, 2345 (2007).
41. P. Singh, J. P. M. Niederer and G. F. Versteeg, *Int. J. Greenh Gas Control*, **1**, 5 (2007).
42. P. Singh, J. P. M. Niederer and G. F. Versteeg, *Chem. Eng. Res. Des.*, **87**, 135 (2009).
43. B. Kanegsberg and E. Kanegsberg, Handbook for Critical Cleaning: *Cleaning Agents and Systems*, 2nd Ed., CRC Press (2011).
44. K. Osmialowski and R. Piekos, *J. Solution. Chem.*, **20**, 241 (1991).
45. R. G. Makitra, R. E. Pristanski and R. I. Flyunt, *Russ. J. Gen. Chem.*, **73**, 1227 (2003).