Chemical Engineering Journal 163 (2010) 195–201

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Prediction of infinite-dilution activity coefficients of organic solutes in ionic liquids using temperature-dependent quantitative structure–property relationship method

Lili Xi[a], Huijun Sun[a], Jiazhong Li[a], Huanxiang Liu[b], Xiaojun Yao[a,*], Paola Gramatica[c]

[a] State Key Laboratory of Applied Organic Chemistry, Department of Chemistry, Lanzhou University, 222, Tianshui South Road, Lanzhou 730000, China
[b] School of Pharmacy, Lanzhou University, Lanzhou 730000, China
[c] QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Via Dunant 3, 21100 Varese, Italy

ARTICLE INFO

Article history:
Received 9 February 2010
Received in revised form 13 May 2010
Accepted 12 July 2010

Keywords:
Ionic liquid
Infinite-dilution activity coefficient
Genetic algorithm-variables subset selection
Ordinary least-square regression
Temperature-dependent quantitative structure–property relationship

ABSTRACT

Ionic liquids (ILs) are a type of potential green solvents, which can be used as a media for reaction and separation. The infinite-dilution activity coefficient is an important parameter to measure the interaction between ILs and solutes. In this work, we proposed a new method to predict infinite-dilution activity coefficients of ILs at different temperatures. A temperature-dependent quantitative structure–property relationship (QSPR) model was developed for a series of organic solutes in the ionic liquid trihexyl(tetradecyl)phosphonium bis(trifluoromethylsulfonyl)imide. By using genetic algorithm-variables subset selection (GA-VSS) and ordinary least-square regression (OLS) methods, six variables, including temperature and five significant molecular descriptors, were selected and used to build the temperature-dependent prediction model. The satisfactory results of the internal and external validations proved the reliability, stability and predictive ability of the built model.

1. Introduction

Ionic liquids (ILs), solely composed by organic cations and organic or inorganic anions, are a type of novel salts that maintain stable liquid state at or close to room temperature. Compared with conventional organic solvents, ILs possess many special features, such as low-saturated vapor pressure, wide liquid range, excellent thermal and chemical stability, high electrochemical stability, low viscosity, etc. [1–3]. The most important feature of ILs is their designability, i.e. the physicochemical property can be designed by suitable choice of anion and cation for a specific application. Hence, ILs are generally termed as "designable green solvents" [4,5]. ILs are widely used as reaction media for synthesis of potential new pharmaceutical drug molecules, biomolecules and polymers [6,7], reservoirs for the controlled release of drug molecules in pharmaceutical formulations [8], extraction solvent for the removal of sulfur compounds [9] and organic contaminants from petroleum crude oils and soil samples [10].

In addition to the features of pure ILs, the knowledge of the properties of their mixtures is also very important for the synthesis and process design. It is essential to know how they interact with each other. The infinite-dilution activity coefficient γ_i^∞ is a quantitative measure to describe the degree of non-ideality for a solute in a mixture. It gives key information for the understanding of many separation processes that employ ionic liquids, and it also provides useful information about solute-ionic liquid intermolecular interactions. These information are helpful to screen the solutes for extraction and extractive distillation and other potential applications [11].

The experimental methods to measure the infinite-dilution activity coefficients include gas–liquid chromatography (GLC) [12–14], dilution method [15] and vapor–liquid equilibria (VLE) measurements [16,17]. Actually, it is unfeasible to experimentally measure all the possible combinations of anions and cations in ILs with all the possible solutes. Therefore, the use of prediction methods for the properties of ILs is very helpful. Recently, Diedenhofen et al. [18] investigated the infinite-dilution activity coefficients for 38 solutes in three different ionic liquids by using COSMO-RS, which is a general and fast method for the prediction of thermophysical properties of liquids. They concluded that COSMO-RS method could predict the infinite-dilution activity coefficients in various ionic liquids with good accuracy. Freire et al. [19] also employed COSMO-RS method to predict binary liquid–liquid equilibria (LLE) and VLE measurements in several alcohol-ILs systems based on quantum chemistry calculations. They obtained a reasonable quali-

* Corresponding author. Tel.: +86 931 891 2578; fax: +86 931 891 2582.
E-mail address: xjyao@lzu.edu.cn (X. Yao).

1385-8947/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cej.2010.07.023

tative agreement between the model predictions and experimental data. Eike et al. [20] established several QSPR models to predict the infinite-dilution activity coefficients of 38 solutes in three ionic liquids, and the squared correlation coefficients ranged from 0.90 to 0.99.

Most of the published studies focused on predicting infinite-dilution activity coefficients at single temperature. However, the values of infinite-dilution activity coefficients highly depend on the temperature even for the same solute-IL system, i.e. the solute behaves differently when it interacts with IL at different temperatures. Therefore it is desirable and useful to capture the effect of both the structural information and temperature on the infinite-dilution activity coefficients. Eike et al. [20] developed a correlation of $\ln\gamma_i^\infty$ with four descriptors for each IL at four different temperatures. The coefficients for the individual descriptors were weighted by temperature. Though the results are satisfactory, the calculation process is inconvenient, because for each solute-IL system, each descriptor needs to be weighted by certain temperature. Furthermore, according to the OECD principles [21], the built QSPR model should be validated both internally and externally to evaluate the model stability and predictive ability [22,23].

Recently, Revelli et al. [24] described the relationship between infinite-dilution activity coefficient and temperature as $\ln\gamma_i^\infty = A_i + (B_i/T)$, which is derived from the Gibbs Helmholtz equation. Here, A_i and B_i are coefficients of the equation for the ith solute. For different solutes, A_i and B_i are different and need to be redefined. Inspired the success of these reported works, we aim to propose a new equation to predict infinite-dilution activity coefficients of different solutes in ILs at any temperatures from know b, X and c values for a particular solute, which can be described as:

$$\ln\gamma_i^\infty = \frac{b}{T} + X_i + c \quad (X_i = a_1x_{i,1} + a_2x_{i,2} + \ldots + a_nx_{i,n}) \tag{1}$$

where b and c are constants, and X_i is the molecular descriptors representing structural parameters of the ith solute mostly related to infinite-dilution activity coefficient. Eq. (1) considers the effect of temperature and structural information. By using Eq. (1), one can calculate the infinite-dilution activity coefficient values at given temperature.

The studied dataset includes 39 polar and nonpolar solutes in the ionic liquid trihexyl(tetradecyl)phosphonium bis(trifluoromethylsulfonyl)imide at three different temperatures measured by GLC [24]. The theoretical descriptors were calculated by CODESSA [25] program to describe molecular structures. By using genetic algorithm-variables subset selection (GA-VSS) method, six descriptors, including temperature and the significant structural descriptors related to the infinite-dilution activity coefficient, were selected, and then a QSPR model was developed. The model was internally validated by leave-one-out (LOO) cross-validation, bootstrap [26] and Y-scrambling [27] techniques, and externally validated by one external test set. The applicability domain (AD) of the built model was also defined.

2. Materials and methods

2.1. Dataset

In this work, two datasets were employed, the original set and external test set. The original set included 39 solutes whose experimental values of infinite-dilution activity coefficients in the ionic liquid trihexyl(tetradecyl)phosphonium bis(trifluoromethylsulfonyl)imide at three different temperatures (302.45 K, 322.35 K and 342.45 K) were collected from the literature [24]. The type of 39 solutes includes alkanes, alkenes, alkynes, cycloalkanes, aromatic compounds (benzene, alkyl-substituted benzene, pyridine and thiophene), ketone, ether, chlorinated methane, acetonitrile, nitroalkane, alcohols and aldehyde. The external test set contained 18 solutes, whose experimental values of infinite-dilution activity coefficients in the same ionic liquid at four temperatures (303.15 K, 308.15 K, 313.15 K and 318.15 K) measured by GLC, were collected from the literature [28]. Among these 18 solutes, 12 solutes were presented in original set, and the rest of 6 solutes (pentane, cyclopentane, pent-1-ene, hept-1-ene, oct-1-ene and oct-1-yne) were not contained in the original set.

The original set was composed of 117 samples (39 solutes at 302.45 K, 322.35 K and 342.45 K, respectively) whose experimental value of infinite-dilution activity coefficient was expressed by $\ln\gamma_i^\infty$, listed in Table 1. The experimental $\ln\gamma_i^\infty$ values ranged from −2.813 to 1.278. By fully considering model validation, the 117 samples in the original set were randomly divided into a 94-sample training set and a 23-sample test set. The training set was used to construct model, while test set was used to validate the built model. The external test set was composed of 72 samples (18 solutes at 303.15 K, 308.15 K, 313.15 K and 318.15 K, respectively) listed in Table 2, which was used to externally validate the predictive ability of the built model. This external test set contained 6 new structures and their infinite-dilution activity coefficients were measured at different temperatures. Thus this external test can be used to evaluate the prediction ability of the model.

2.2. Descriptors generation

The molecular structures were drawn in HyperChem7.0 program [29] and pre-optimized using MM+ molecular mechanics force field. A more precise optimization was done with semi-empirical PM3 method in MOPAC7.0 [30]. All calculations were carried out at restricted Hartree–Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.01. The MOPAC output files were transferred into CODESSA program to calculate five types of 282 descriptors: constitutional, topological, geometrical, electrostatic and quantum chemical. Constant descriptors and highly correlated descriptors (one of any two descriptors with a correlation coefficient greater than 0.95) were excluded to reduce redundant and useless information. The variable T^{-1} was also added in the descriptor pool. Finally, 127 descriptors were remained for the next step.

2.3. Descriptor selection and model construction

Many applications have proved genetic algorithm (GA) to be a very effective tool in solving feature selection problems [31–34]. Therefore, GA was employed to select the significant descriptors in this work. Selection procedure was performed in MOBYDIGS [35] program by using GA-VSS. The used fitness function was LOO cross-validation. The important parameters of GA were as follows: population size 100 and reproduction/mutation trade-off 0.5.

The initial population (i.e. a set of models) with the minimum number of allowed descriptors is developed by all-subset-method procedure to explore all the low dimension combination. Then these models are evaluated by fitness function, and ranked according to the fitness score (the squared correlation coefficient Q_{loo}^2). If this population meets the conditions of convergence, it can be considered as potential solution; otherwise, individual selection, crossover and mutation are operated to produce new population. Then the fitness function is used again to evaluate the new population. This process continues until the new population matches the conditions of convergence. When increasing the model size does not increase the Q_{loo}^2 value to any significant degree, the best solution is obtained.

Table 1
The experimental and predicted infinite-dilution activity coefficients ln γ^{∞} in IL at different temperatures of training set and test set.

Solutes	No.	T = 302.45 K		No.	T = 322.35 K		No.	T = 342.45 K	
		Exp.	Pred.		Exp.	Pred.		Exp.	Pred.
Hexane	1	0.086	0.171	40[a]	0.039	0.082	79	0.01	0.002
3-Methylpentane	2	0.039	0.135	41	−0.02	0.046	80	−0.062	−0.030
Heptane	3[a]	0.239	0.282	42	0.199	0.193	81	0.174	0.113
Octane	4[a]	0.392	0.451	43	0.351	0.362	82	0.336	0.282
Nonane	5	0.626	0.617	44	0.588	0.528	83	0.554	0.448
Decane	6	0.678	0.825	45	0.637	0.735	84[a]	0.61	0.656
Undecane	7	0.842	1.041	46[a]	0.779	0.952	85	0.742	0.872
Methylcyclopentane	8[a]	−0.163	0.036	47	−0.223	−0.054	86	−0.261	−0.130
Cyclohexane	9	−0.186	−0.003	48	−0.248	−0.092	87[a]	−0.288	−0.170
Methylcyclohexane	10	−0.083	0.120	49	−0.117	0.030	88	−0.151	−0.050
Cycloheptane	11	−0.186	0.045	50	0.593	−0.044	89	1.278	–
Benzene	12[a]	−0.942	−0.760	51	−0.942	−0.849	90	−0.942	−0.930
Toluene	13	−0.777	−0.523	52	−0.777	−0.613	91	−0.755	−0.690
Ethylbenzene	14[a]	−0.616	−0.412	53[a]	−0.562	−0.501	92	−0.528	−0.580
1-Hexene	15	−0.094	−0.291	54	−0.139	−0.380	93	−0.163	−0.460
1-Hexyne	16	−0.357	−0.328	55	−0.371	−0.417	94	−0.400	−0.500
1-Heptyne	17	−0.261	−0.236	56[a]	−0.248	−0.326	95	−0.223	−0.410
2-Butanone	18[a]	−1.273	−1.065	57	−1.309	−1.155	96	−1.470	−1.230
2-Pentanone	19	−1.171	−0.958	58[a]	−1.139	−1.048	97	−1.109	−1.130
1,4-Dioxane	20	−0.673	−0.921	59	−0.713	−1.011	98[a]	−0.734	−1.090
Methanol	21	0.255	0.110	60	−0.02	0.021	99	−0.151	−0.060
Ethanol	22	0.445	0.323	61	0.207	0.234	100	0.01	0.154
1-Propanol	23	0.425	0.302	62[a]	0.191	0.212	101	0	0.133
2-Propanol	24	0.438	0.276	63	0.199	0.187	102[a]	−0.010	0.107
2-Methyl-1-propanol	25	0.399	0.284	64	0.14	0.194	103	−0.051	0.115
1-Butanol	26[a]	0.482	0.315	65	0.231	0.226	104	0.039	0.146
Diethyl ether	27[a]	−0.478	−0.266	66	−0.528	−0.356	105	−0.562	−0.440
Diisopropyl ether	28[a]	−0.073	−0.047	67	−0.105	−0.136	106[a]	−0.139	−0.220
Chloroform	29	−1.238	−1.013	68	−1.171	−1.103	107	−1.139	−1.180
Dichloromethane	30[a]	−1.47	−1.084	69	−1.386	−1.173	108	−1.309	−1.250
Tetrachloromethane	31	−0.598	−0.516	70	−0.598	−0.605	109	−0.598	−0.690
Acetonitrile	32	−0.580	−0.838	71	−0.693	−0.928	110	−0.777	−1.010
Nitromethane	33	−0.261	−0.335	72	−0.446	−0.424	111	−0.545	−0.500
1-Nitropropane	34	0.784	0.308	73	0.482	0.219	112	0.300	0.139
Pyridine	35	−0.654	−0.822	74	−0.892	−0.911	113	−1.079	−0.990
Thiophene	36[a]	−0.868	−0.742	75	−0.916	−0.832	114	−0.968	−0.910
Formaldehyde	37	−2.813	−2.646	76	−2.813	−2.735	115	−2.659	−2.820
Propionaldehyde	38[a]	−1.05	−1.022	77	−1.109	−1.111	116	−1.171	−1.190
Butyraldehyde	39	−0.968	−0.972	78	−1.022	−1.061	117	−1.050	−1.140

[a] Samples in test set.

2.4. Model validation and performance

The internal predictive ability and robustness of the built model are evaluated by LOO cross-validation and bootstrap approaches. The LOO procedure involves removing one sample from the original training set and constructing the model only based on the remaining samples and then testing on the removed one sample. In this form, all the samples in training set are tested, and Q^2_{loo} is calculated.

Table 2
The experimental and predicted infinite-dilution activity coefficients ln γ^{∞} in IL at different temperatures of external test set.

Solutes	No.	T = 303.15 K		No.	T = 308.15 K		No.	T = 318.15 K		No.	T = 313.15 K	
		Exp.	Pred.		Exp.	Pred.		Exp.	Pred.		Exp.	Pred.
Pentane[a]	1	−0.062	0.051	19	−0.041	0.028	37	−0.01	0.005	55	0.020	−0.017
Hexane	2	0.086	0.168	20	0.095	0.144	38	0.122	0.122	56	0.157	0.100
Heptane	3	0.231	0.279	21	0.239	0.255	39	0.262	0.232	57	0.285	0.210
Octane	4	0.278	0.448	22	0.307	0.425	40	0.336	0.402	58	0.378	0.380
Cyclopentane[a]	5	−0.386	−0.124	23	−0.371	−0.148	41	−0.342	−0.170	59	−0.315	−0.192
Cyclohexane	6	−0.248	−0.006	24	−0.223	−0.03	42	−0.198	−0.052	60	−0.163	−0.074
Cycloheptane	7	−0.128	0.042	25	−0.094	0.018	43	−0.083	−0.004	61	−0.062	−0.026
Pent-1-ene[a]	8	−0.248	−0.383	26	−0.223	−0.406	44	−0.198	−0.429	62	−0.163	−0.451
Hex-1-ene	9	−0.128	−0.294	27	−0.105	−0.318	45	−0.083	−0.340	63	−0.03	−0.362
Hept-1-ene[a]	10	0	−0.177	28	0.02	−0.201	46	0.039	−0.223	64	0.086	−0.245
Oct-1-ene[a]	11	0.104	−0.022	29	0.131	−0.045	47	0.148	−0.068	65	0.191	−0.09
Hex-1-yne	12	−0.416	−0.331	30	−0.386	−0.355	48	−0.342	−0.377	66	−0.301	−0.399
Hept-1-yne	13	−0.528	−0.240	31	−0.462	−0.263	49	−0.416	−0.286	67	−0.357	−0.308
Oct-1-yne[a]	14	−0.635	−0.110	32	−0.598	−0.133	50	−0.545	−0.156	68	−0.462	−0.178
Benzene	15	−0.968	−0.763	33	−0.942	−0.787	51	−0.916	−0.809	69	−0.868	−0.831
Methanol	16	0.231	0.107	34	0.215	0.083	52	0.207	0.061	70	0.207	0.039
Ethanol	17	0.307	0.320	35	0.285	0.296	53	0.270	0.273	71	0.262	0.251
1-Propanol	18	0.351	0.298	36	0.336	0.275	54	0.329	0.252	72	0.322	0.230

[a] Solutes that were not contained in training set.

In the bootstrap procedure, K n-dimensional groups are generated by a randomly repeated selection of n-objects from the original dataset. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2_{boot} is calculated for each model. The bootstrapping was repeated 5000 times for each validated model. In addition, to avoid chance correlation, Y-scrambling technique is used [36]. In Y-scrambling procedure, the response vector **Y** is randomly reordered (Y-scrambling), and new models were recalculated based on randomized responses. The resulted models should have significantly lower values of the squared correlation coefficient than the proposed one because the relationship between the structure and response is broken. This procedure is repeated 500 times and the mean value of $Q^2_{Y\text{-}scrambling}$ is reported.

It is worthwhile to point out that all the validation methods mentioned above just assess the internal predictive ability of models [23,37,38]. It is necessary to externally validate the built model, because compared with internal validation methods, external validation can provide a more rigorous evaluation of the model's predictive capability. Here, one external test set was employed (as mentioned in Section 2.1).

Model performance was evaluated by the following parameters: the squared correlation coefficient (R^2), which can be interpreted mathematically as the proportionate reduction of total variation associated with the independent variable; the root mean squared error ($RMSE$), which measures the difference between the actual and estimated values. $RMSE$ is calculated as below:

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

where i represents the ith sample, y_i is the experimental values, and $\hat{y}_i$ is the predicted value by the model; n is the number of samples in the dataset.

For the external test set, the validation parameter Q^2_{ext} is calculated:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_{tr})^2} \qquad (3)$$

where $\bar{y}_{tr}$ is the averaged value of the response variable for the training set; m is the number of the samples in the external test set.

2.5. Applicability domain (AD)

In this work, the applicability domain was verified by the leverage approach to verify prediction reliability [23]. The leverage (h) is defined as follows:

$$h_i = x_i(X^TX)^{-1}x_i^T \quad (i = 1, 2, \ldots, n) \qquad (4)$$

where x_i is the descriptor row-vector of the query sample i; n is the number of query samples; X is the $n \times p$ matrix of dataset (p is the number of selected descriptors). The threshold value of warning leverage h^* is defined as $3p'/n$, where p' is the number of the model descriptors plus one. In fact, leverage can be used as a quantitative measure of the model applicability domain suitable for evaluating the degree of extrapolation. It represents a sort of compound distance from the model experimental space.

Williams plot [39,22] (leverage values versus standardized residuals) is used to visualize the model AD. In Williams plot, the two horizontal lines indicate the limit of normal values for Y outliers (i.e. samples with standardized residuals greater than 3.0 standard deviation units, $\pm 3.0\sigma$); the vertical straight lines indicate the limits of normal values for X outliers (i.e. samples with leverage values greater than the threshold value, $h > h^*$). For a sample in external test set whose leverage value is greater than h^*, its prediction is considered unreliable, because the prediction is the result of substantial extrapolation of the model.

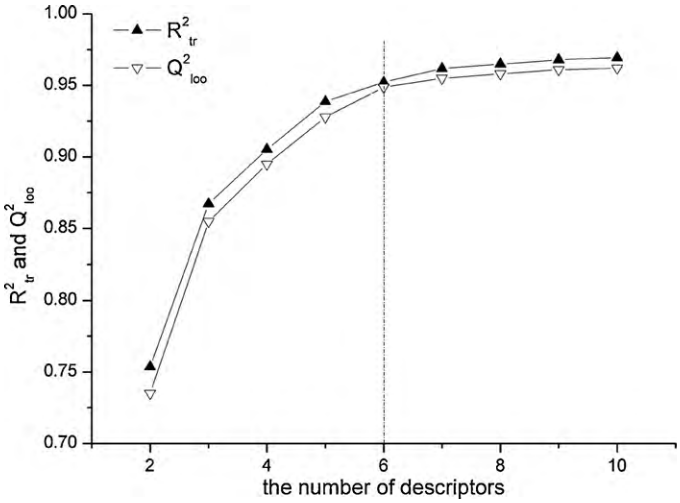

Fig. 1. The squared correlation coefficients (R^2_{tr}, Q^2_{loo}) versus the number of descriptors.

3. Results and discussions

3.1. Analysis of dataset

Generally, for any QSPR model, quality of the dataset has great influence on performance of the built model. In order to build a well generalized QSPR model, a preliminary analysis for the dataset, mainly detecting outliers, was performed.

A number of models for the whole dataset were developed using OLS regression based on the descriptors selected by GA-VSS. By analyzing the applicability domain of most of the models, sample **89** (cycloheptane) was shown to be a suspected response outlier (Y outlier) by most of the models. This sample was the infinite-dilution activity coefficient of cycloheptane at 342.45 K. By analyzing the experimental values of different cycloalkanes, we found that the infinite-dilution activity coefficient of methylcyclopentane, cyclohexane and methylcyclohexane decreased with the rise of temperature. But for cycloheptane, the trend was contrary and significantly different from others. The experimental value for this sample may be susceptible. In order to build a reliable and general model, this sample was removed from the dataset.

3.2. Model construction and internal validation

In GA process, when adding another descriptor does not improve the statistics of a model to any significant degree, the optimum subset size is achieved. To avoid the "over-parameterization" of the model, the increase of the Q^2_{loo} value less than 0.02 was chosen as the breakpoint criterion. At last, the model with six descriptors was considered as the best one, which can be found from Fig. 1 (the number of descriptors versus the statistical parameters R^2_{tr} and Q^2_{loo}). With the selected descriptors, the following linear equation was obtained:

$$\ln \gamma^{\infty} = \frac{437.594}{T} + (2.356 \times \text{RPCG} + 0.113 \times \Delta E + 0.003 \times \text{PPSA\_2} + 7.942 \times V_C^{\min} + 13.604 \times V_C^{\max}) - 89.337 \qquad (5)$$

$$n_{tr} = 93, \quad R^2_{tr} = 0.952, \quad RMSE_{tr} = 0.159, \quad Q^2_{loo} = 0.945, \quad RMSE_{loo} = 0.171, \quad Q^2_{boot} = 0.939 \quad \text{and} \quad Q^2_{Y\text{-}scrambling} = 0.030$$

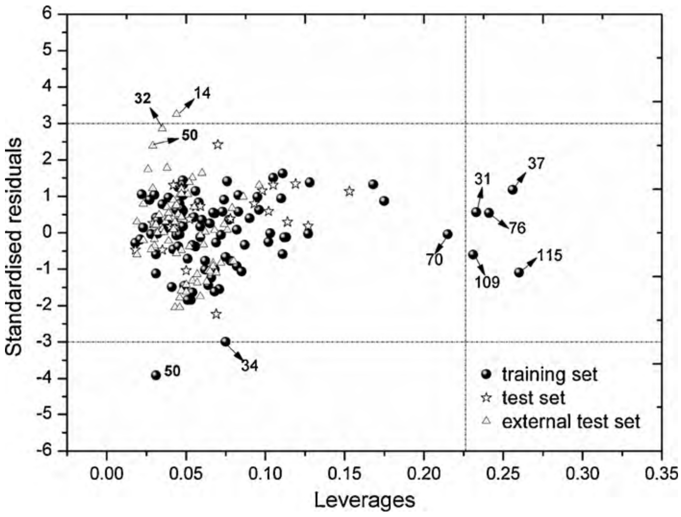

Fig. 2. Williams plot for the model with six descriptors. The values of training set, test set and external test set were labeled differently. The two short dashed horizontal lines are the ±3.0σ limit and one short dashed vertical line is the threshold value of leverage ($h^* = 0.226$).

The high statistical parameter values indicate that the model is robust and stable. In addition, the low value of $Q^2_{Y\text{-}scrambling}$ excluded the probability of chance correlation. The predicted values of training set were listed in Table 1.

3.3. Applicability domain (AD)

Analyzing AD of the built model can provide information about reliability of the prediction. Only predictions for samples that fall in the domain can be considered reliable. The model AD was analyzed in the Williams plot (shown in Fig. 2). The majority of the samples within the model AD were calculated accurately.

Generally, if the σ value is larger than ±3.0σ, the sample can be considered as response outlier. Thus, samples **50** and **34** would be recognized as response outliers. The sample **50** was the infinite-dilution activity coefficient of cycloheptane at 322.35 K. From the analysis of Section 3.1, we thought that the experimental value of sample **50** may be susceptible. The sample **34** is the infinite-dilution activity coefficient of 1-nitropropane at 302.45 K. It is poorly predicted. Samples **73** (1-nitropropane at 322.35 K) and **112** (1-nitropropane at 342.45 K) were well predicted. It is difficult to find a reason why the prediction for **34** is poor, however, it is important to keep in mind that the quality of the input experimental data could be arguable for some of these samples.

From Williams plot, we also can find that five samples (**31**, **37**, **76**, **109** and **115**) were *X* outliers (with leverage value higher than the threshold value of 0.226). In this study, these five samples were well predicted, being close to regression line of Fig. 2. Samples **31**, **70**, **109** were tetrachloromethane at three different temperatures. Sample **70** was within the domain and well predicted, and for similar samples **31** and **109**, their predictions were also good. Samples **37**, **76**, **115** were formaldehyde at three different temperatures.

3.4. Results of test set and external validation

In the present investigation, the built model was validated by test set and one external test set. The predicted results for test set were list in Table 1. The performance of the built model for test set was shown in Table 3. The plot of the predicted versus experimental $\ln\gamma^\infty$ of training set and test set was shown in Fig. 3. From Table 1 and Fig. 3, it can be found that samples in test set were all well predicted, indicating the good predictive ability of the built model for these samples. From Fig. 2, the samples of test set all fall in the applicability domain of the built model, indicating that the predictions of these samples were reliable.

Table 3
The performance of predictive ability of the built model.

		Q^2_{ext}	$RMSE_{tst}$
Test set		0.950	0.163
External test set	72 samples	0.938	0.181
	48 samples[a]	0.967	0.132
	24 samples[b]	0.880	0.252

[a] 12 solutes (12 × 4 samples) whose chemical structures were contained in training set.

[b] 6 solutes (6 × 4 samples) were new structures that were not contained in training set.

The predicted results for external test set were list in Table 2. The performance of external validation was shown in Table 3. For external test set, the predictions of most samples were satisfactory (Q^2_{ext} equalled to 0.938). As mentioned in Section 2.1, the chemical structures of 12 solutes in external test set were the same as those of training set, and for these 48 samples (12 solutes at four different temperatures), Q^2_{ext} was 0.967. The results illustrated that the built model can give accurate prediction for samples that had common structures with the training set but at different temperatures. For the rest of 6 solutes (24 samples in total) that were not contained in training set, Q^2_{ext} was 0.880. The results illustrated that the built model can also give good prediction for samples that were completely new structures and at different temperatures.

The Williams plot (Fig. 2) shows that there is one response outlier in external test set, sample **14**. Samples **14**, **32** and **50** correspond to the $\ln\gamma^\infty$ of the 1-octyne at 303.15 K, 308.15 K and 313.15 K, respectively, and their standard residuals were relative large (more than 2σ). By analyzing their experimental values, it can be found that there is no obvious change with temperature increasing, being different from homologues 1-hexyne and 1-heptyne. By comparing the experimental values of the training set and external test set, we can find that infinite-dilution activity coefficients of the training set at 302.45 K were in agreement with those of external test set at 303.15 K, expect for the *n*-alkyne solutes. Maybe that is the reason why sample **14** in external test set is poorly predicted. The built model has good fitting power and good prediction results for the external dataset. Furthermore, the majority of samples were within the applicability domain of the proposed model and were predicted correctly.

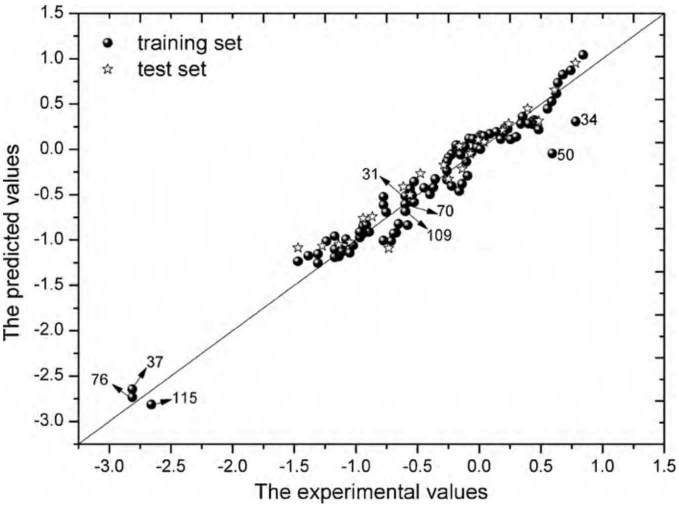

Fig. 3. The plots of the predicted versus experimental $\ln\gamma^\infty$ values of training set and test set.

Table 4
The detailed information of the selected descriptors.

Symbol	Descriptors	Regression Coeff.	Std. Reg. Coeff.	VIF
	Intercept	−89.337	–	–
T^{-1}	T^{-1}	437.594	0.093	1.018
RPCG	Relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]	2.356	0.623	2.150
ΔE	HOMO–LUMO energy gap	0.113	0.269	1.420
PPSA-2	Total charge weighted PPSA [quantum-chemical PC]	0.003	0.438	1.639
$V_C^{\min}$	Min valency of a C atom	7.942	0.421	1.494
$V_C^{\max}$	Max valency of a C atom	13.604	0.560	2.011

3.5. Interpretation of the built model

The six descriptors employed in the model were T^{-1}, relative positive charge (QMPOS/QTPLUS) [Zefirov's PC] (RPCG), HOMO–LUMO energy gap (ΔE), total charge weighted PPSA [quantum chemical] (PPSA-2), min valency of a C atom ($V_C^{\min}$) and max valency of a C atom ($V_C^{\max}$). The detail information was summarized in Table 4. Multi-colinearity between the six descriptors was detected by calculating their variation inflation factor (VIF), described as

$$\mathrm{VIF} = \frac{1}{1 - r^2} \qquad (6)$$

where r is the correlation coefficient of multiple regression between one descriptor and the others. If VIF equals to one, no inter-correlation exists for each descriptor; if VIF maintains within the range 1.0–5.0, the corresponding model is acceptable; if VIF is larger than 10.0, the corresponding model is unstable [40]. The r^2 values of these six descriptors were 0.018, 0.535, 0.296, 0.390, 0.331 and 0.503, respectively. Accordingly, their VIF values were 1.018, 2.150, 1.420, 1.639, 1.494 and 2.011, respectively. For each descriptor, its VIF value was less than five, which indicates that the built model was obvious statistic significance.

The correlation between the infinite-dilution activity coefficients with T^{-1} is positive, i.e. infinite-dilution activity coefficient will decrease with the rise of temperature. Both RPCG and PPSA-2 belong to charged partial surface area (CPSA) descriptors. RPCG is defined as the most relative positive charge divided by the sum of all of the relative positive charges of the molecule, and it represents the effect of the polar intermolecular interactions. PPSA-2 means total charge weighted partial positive surface area. This descriptor depends directly on the quantum-chemically calculated charge distribution in the molecule. Therefore, it encodes features responsible for polar interactions between molecules and the hydrogen-bond interaction as well. The intermolecular interactions between ionic liquid and solute become stronger with the increase of polar interaction, thus the infinite-dilution activity coefficient decreases. ΔE, the difference in energy between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), helps to estimate the relative reactivity of the ionic liquid. It relates to the activation energy of the corresponding chemical reaction as well as polarization. A large ΔE implies high stability for the solute in the ionic liquid. $V_C^{\min}$ and $V_C^{\max}$ belong to valency-related descriptors that relate to the strength of intramolecular bonding interactions and characterize the stability of the molecule, the conformational flexibility, and other valency-related properties.

From the above discussion, it can be seen that the five molecular descriptors, i.e. the polar intermolecular interaction, hydrogen-bond interaction, stability together with temperature can account for the factors influencing the infinite-dilution activity coefficient of solutes. As indicated by the standard regression coefficients in Table 4, RPCG is the most relevant descriptor related the infinite-dilution activity coefficient. In this sense, the polar intermolecular interaction is the preferential determination of infinite-dilution activity coefficients.

4. Conclusion

The infinite-dilution activity coefficient is an important parameter to measure the interaction between different systems of solute-ILs. In this investigation, temperature variable and five structural descriptors calculated by CODESSA were selected by GA-VSS method for development of a fast and accurate temperature-dependent QSPR model. The developed model was validated internally by LOO, bootstrap and *Y*-scrambling approaches and externally by one external test set. The obtained results indicated that our model was robust and stable, and had good predictive ability. To use the temperature as a variable to build QSPR model enable us to predict the infinite-dilution activity coefficient at other temperatures for different systems of solute-ILs. The applicability domain of the built model was also defined. Furthermore, the meanings of these selected structural descriptors were analyzed in detail. The polar intermolecular interaction is the preferential determination of infinite-dilution activity coefficients. In summary, the built QSPR model proved to be feasible and promising for the infinite-dilution activity coefficients prediction in ionic liquids at different temperatures.

Acknowledgements

This work was supported by the Program for New Century Excellent Talents in University (Grant No. NCET-07-0399) and the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] T. Welton, Room-temperature ionic liquids. solvents for synthesis and catalysis, Chem. Rev. 99 (1999) 2071–2084.

[2] R. Sheldon, Catalytic reactions in ionic liquids, Chem. Commun. 23 (2001) 2399–2407.

[3] J.G. Huddleston, A.E. Visser, W.M. Reichert, H.D. Willauer, G.A. Broker, R.D. Rogers, Characterization and comparison of hydrophilic and hydrophobic room temperature ionic liquids incorporating the imidazolium cation, Green Chem. 3 (2001) 156–164.

[4] M.J. Earle, K.R. Seddon, Ionic liquids. Green solvents for the future, Pure Appl. Chem. 72 (2000) 1391–1398.

[5] M. Freemantle, Ionic liquids show promise for clean separation technology, Chem. Eng. News 76 (1998).

[6] G. Viswanathan, S. Murugesan, V. Pushparaj, O. Nalamasu, P.M. Ajayan, R.J. Linhardt, Preparation of biopolymer fibers by electrospinning from room temperature ionic liquids, Biomacromolecules 7 (2006) 415–418.

[7] S. Park, R.J. Kazlauskas, Biocatalysis in ionic liquids – advantages beyond green technology, Curr. Opin. Biotechnol. 14 (2003) 432–437.

[8] V. Jaitely, A.T. Florence, 1-Alkyl-3-methyl-imidazole hexafluoro-phosphate water-immiscible ionic liquids as reservoirs for controlled release of drugs, in: Abstracts of Papers, 226th ACS National Meeting, New York, United States, September, 2003, pp. 7–11.

[9] J. Planeta, P. Karásek, M. Roth, Distribution of sulfur-containing aromatics between [hmim][Tf2N] and supercritical CO_2: a case study for deep desulfurization of oil refinery streams by extraction with ionic liquids, Green Chem. 8 (2006) 70–77.

[10] A.P. Khodadoust, S. Chandrasekaran, D.D. Dionysiou, Preliminary assessment of imidazolium-based room-temperature ionic liquids for extraction of organic contaminants from soils, Environ. Sci. Technol. 40 (2006) 2339–2345.

[11] T.M. Letcher, P. Reddy, Determination of activity coefficients at infinite dilution of organic solutes in the ionic liquid, trihexyl(tetradecyl)-phosphonium tris(pentafluoroethyl)trifluorophosphate, by gas–liquid chromatography, Fluid Phase Equilibr. 235 (2005) 11–17.

[12] T.M. Letcher, B. Soko, D. Ramjugernath, N. Deenadayalu, A. Nevines, P.K. Naicker, Activity coefficients at infinite dilution of organic solutes in 1-hexyl-3-methylimidazolium hexafluorophosphate from gas–liquid chromatography, J. Chem. Eng. Data 48 (2003) 708–711.

[13] D. Warren, T.M. Letcher, D. Ramjugernath, J.D. Raal, Activity coefficients of hydrocarbon solutes at infinite dilution in the ionic liquid, 1-methyl-3-octyl-imidazolium chloride from gas–liquid chromatography, J. Chem. Thermodyn. 35 (2003) 1335–1341.

[14] A. Heintz, D.V. Kulikov, S.P. Verevkin, Thermodynamic properties of mixtures containing ionic liquids. 2. Activity coefficients at infinite dilution of hydrocarbons and polar solutes in 1-methyl-3-ethyl-imidazolium bis(trifluoromethyl-sulfonyl) amide and in 1,2-dimethyl-3-ethyl-imidazolium bis(trifluoromethyl-sulfonyl) amide using gas–liquid chromatography, J. Chem. Eng. Data 47 (2002) 894–899.

[15] K. Marsh, A. Deev, A. Wu, E. Tran, A. Klamt, Room temperature ionic liquids as replacements for conventional solvents – a review, Korean J. Chem. Eng. 19 (2002) 357–362.

[16] T.V. Vasiltsova, S.P. Verevkin, E. Bich, A. Heintz, R.B. Lukasik, U. Domanska, Thermodynamic properties of mixtures containing ionic liquids. Activity coefficients of ethers and alcohols in 1-methyl-3-ethyl-imidazolium bis(trifluoromethyl-sulfonyl)imide using the transpiration method, J. Chem. Eng. Data 50 (2004) 142–148.

[17] R. Kato, M. Krummen, J. Gmehling, Measurement and correlation of vapor–liquid equilibria and excess enthalpies of binary systems containing ionic liquids and hydrocarbons, Fluid Phase Equilibr. 224 (2004) 47–54.

[18] M. Diedenhofen, F. Eckert, A. Klamt, Prediction of infinite dilution activity coefficients of organic compounds in ionic liquids using COSMO-RS, J. Chem. Eng. Data 48 (2003) 475–479.

[19] M.G. Freire, L.M.N.B.F. Santos, I.M. Marrucho, J.A.P. Coutinho, Evaluation of COSMO-RS for the prediction of LLE and VLE of alcohols + ionic liquids, Fluid Phase Equilibr. 255 (2007) 167–178.

[20] D.M. Eike, J.F. Brennecke, E.J. Maginn, Predicting infinite-dilution activity coefficients of organic solutes in ionic liquids, Ind. Eng. Chem. Res. 43 (2004) 1039–1048.

[21] http://www.oecd.org/document/23/0,2340,en_2649_201185_33957015_1_1_1_1,00.html.

[22] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.

[23] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.

[24] A.L. Revelli, L.M. Sprunger, J. Gibbs, W.E. Acree, G.A. Baker, F. Mutelet, Activity coefficients at infinite dilution of organic compounds in trihexyl(tetradecyl) phosphonium bis(trifluoromethylsulfonyl)imide using inverse gas chromatography, J. Chem. Eng. Data 54 (2009) 977–985.

[25] A.R. Katritzky, V.S. Lobanov, M. Karelson, Reference Manual, Semichem and the Comprehensive Descriptors for Structural and Statistical Analysis, Version 2.0 (Workstation) and 2.13 (PC), CODESSA, University of Florida, 1995–1997.

[26] B. Efron, The Jackknife, the Bootstrap, and Other Resampling Plans, Society for Industrial Mathematics, Philadelphia, PA, 1982.

[27] P. Gramatica, E. Giani, E. Papa, Statistical external validation and consensus modeling: a QSPR case study for KOC prediction, J. Mol. Graph. Model. 25 (2007) 755–766.

[28] T.M. Letcher, D. Ramjugernath, M. Laskowska, M. Królikowski, P. Naidoo, U. Domańska, Determination of activity coefficients at infinite dilution of solutes in the ionic liquid, trihexyltetradecylphosphonium bis(trifluoromethylsulfonyl)imide, using gas–liquid chromatography at T = (303.15, 308.15, 313.15, and 318.15) K, J. Chem. Eng. Data 53 (2008) 2044–2049.

[29] HyperChem 7.0, Hypercube Inc., 2002.

[30] J.P.P. Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange (QCPE), Indiana University, Bloomington, IN, 1989, Program 455.

[31] J.Z. Li, B.L. Lei, H.X. Liu, S.Y. Li, X.J. Yao, M.C. Liu, P. Gramatica, QSAR study of malonyl-CoA decarboxylase inhibitors using GA-MLR and a new strategy of consensus modeling, J. Comput. Chem. 29 (2008) 2636–2647.

[32] H.X. Liu, E. Papa, P. Gramatica, QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles, Chem. Res. Toxicol. 19 (2006) 1540–1548.

[33] L.L. Xi, J. Du, S.Y. Li, J.Z. Li, H.X. Liu, X.J. Yao, A combined molecular modeling study on gelatinises and their potent inhibitors, J. Comput. Chem. 31 (2010) 24–42.

[34] J.B. Ghasemi, A. Abdolmaleki, N. Mandoumi, A quantitative structure property relationship for prediction of solubilization of hazardous compounds using GA-based MLR in CTAB micellar media, J. Hazard. Mater. 161 (2009) 74–80.

[35] R. Todeschini, V. Consonni, M. Pavan, MOBYDIGS, Version 1.2 for Windows, Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Talete SRL, Milan, Italy, 2002.

[36] F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, L. Eriksson, Model validation by permutation tests: applications to variable selection, J. Chemometr. 10 (1996) 521–532.

[37] K. Baumann, Cross-validation as the objective function for variable-selection techniques, TrAC, Trends Anal. Chem. 22 (2003) 395–406.

[38] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.

[39] P. Gramatica, E. Papa, QSAR modeling of bioconcentration factor by theoretical molecular descriptors, QSAR Comb. Sci. 22 (2003) 374–385.

[40] R.F. George, A.P. Carl, Y.W. Leland, Using theoretical descriptors in quantitative structure activity relationships: some physicochemical properties, J. Phys. Org. Chem. 5 (1992) 395–408.