Macedonian Journal of Chemistry and Chemical Engineering, Vol. **35**, No. 1, pp. 53–62 (2016)

MJCCA9 – 704

ISSN 1857-5552
e-ISSN 1857-5625
UDC:547:543.062
DOI:10.20450/mjcce.2016.848

Received: December 21, 2015
Accepted: April 8, 2016

Original scientific paper

A SIMPLE 2D-QSPR MODEL FOR THE PREDICTION OF SETSCHENOW CONSTANTS OF ORGANIC COMPOUNDS

Qi Xu, Lingling Fan, Jie Xu*[*]*

*College of Materials Science & Engineering, College of Textile Science & Engineering,
Wuhan Textile University, 430200, Wuhan, China*

[*]xujie0@ustc.edu

A quantitative structure-property relationship (QSPR) analysis of the Setschenow constants (K_{salt}) of organic compounds in a sodium chloride solution was carried out using only two-dimensional (2D) descriptors as input parameters. The whole set of 101 compounds was split into a training set of 71 compounds and a validation set of 30 compounds by means of the Kennard and Stones algorithm. A general four-parameter equation, with correlation coefficient (R) of 0.887 and standard error of estimation (s) of 0.031, was obtained by stepwise multilinear regression analysis (MLRA) on the training set. The reliability and robustness of the present model was verified with leave-one-out cross-validation, randomization tests, and the external validation set. All of the descriptors contained in this model are calculated directly from the molecular 2D structures; thus, this model can be used to easily predict the K_{salt} of other compounds not involved in the present dataset.

Keywords: QSPR; Setschenow constants; 2D descriptor; multilinear regression analysis

ЕДНОСТАВЕН 2D-QSPR-МОДЕЛ ЗА ПРЕДВИДУВАЊЕ НА КОНСТАНТИТЕ НА SETSCHENOW ЗА ОРГАНСКИТЕ СОЕДИНЕНИЈА

Анализата на квантитативната зависност на структурата и својствата (QSPR) на константите на Setschenow (K_{salt}) на органските соединенија во раствор од натриум хлорид е извршена користејќи само дводимензионални (2D) дескриптори како влезни параметри. Целото множество од 101 соединение беше поделено во множество за подготовка од 71 соединение и множество за валидација од 30 соединенија според алгоритамот на Kennard и Stones. Од множеството за подготовка со мултилинеарна регресиона анализа MLRA е добиена општа четирипараметарска равенка со коефициент на корелација $R = 0{,}887$ и стандардна грешка на проценатa $s = 0{,}031$. Веродостојноста и робусноста на овој модел беше верифицирана со повеќе тестови: вкрстена валидација со испуштање на еден параметар, тест на рандомизација, како и екстерно множество за валидација. Сите дескриптори содржат модел што се пресметува директно од молекулските 2D-структури, и така овој модел може да се користи за едноставно предвидување на K_{salt} на други соединенија што не се вклучени во ова множество на податоци.

Клучни зборови: QSPR; константи на Setschenow; 2D-дескриптори;
мултилинеарна регресиона анализа

1. INTRODUCTION

The aqueous solubility of organic compounds is an important molecular property that plays a key role in pharmaceutical, environmental, and other physical and biological processes. The aqueous solubility has been found to be dependent on the concentration and type of salt present in solution. The salt effect can be described by the Setschenow equation:

$$\log\left(S_{\text{salt}} / S_{\text{water}}\right) = -K_{\text{salt}} C_{\text{salt}} \, ,$$

where S_{salt} and S_{water} are the solubilities of the organic compound in aqueous salt solution and water, respectively, C_{salt} is the molar concentration of electrolyte, and K_{salt} is the empirical Setschenow constant.

The theoretical prediction of K_{salt} has been carried out using several methods [1–6]; however, they require experimental physicochemical properties or ambiguously determined parameters, resulting in limited predictive ability. Therefore, predicting K_{salt} directly from the molecular structure of the compound concerned it is of particular interest.

Alternatively, the quantitative structure-property relationship (QSPR) provides a promising method for the prediction of K_{salt} using descriptors derived solely from the molecular structure to fit experimental data. The QSPR method is based on the assumption that the variation in the behavior of compounds, as expressed by any measured physicochemical properties, can be correlated with numerical changes in structural features of all compounds, termed “molecular descriptors” [7–21]. The advantage of this method lies in the fact that it requires only knowledge of the chemical structure and is not dependent on any experimental properties. Once a correlation is established, it can be applicable for the prediction of the property of new compounds that have not yet been synthesized or identified. Thus, the QSPR method can expedite the process of development of new molecules and materials with desired properties. The QSPR method has been successfully applied to predict the chemical, physical, biochemical, and pharmacological properties of compounds; however, there have been relatively few attempts to correlate and predict K_{salt}. Zhong et al. [22] correlated the K_{salt} values of 101 compounds with their connectivity indices and developed a relatively good model with standard error of estimation (*s*) of 0.042 and 0.040 for the training and validation set, respectively. However, the zero-order variable connectivity index $^0\chi^f$ in this model were calculated from the optimal weights for the non-hydrogen atoms (including carbon, nitrogen, sulfur, oxygen, chlorine, and fluorine) by fitting the data of the training set, which makes this model inapplicable for compounds containing other non-hydrogen atoms. In our previous work [23], three-dimensional (3D) QSPR models were developed to predict the K_{salt} values of 101 compounds using multilinear regression analysis (MLRA) and artificial neural network (ANN) with *s* of 0.034 and 0.029 for the training set, respectively. We also developed QSPR models for the K_{salt} with the combination of two-dimensional (2D) and 3D descriptors using support vector machine [24]. However, it was found that the proposed models containing 3D descriptors were difficult to use because of their complex calculations.

The goal of this study was to produce a QSPR model based on two-dimensional (2D) descriptors, which is expected to predict the K_{salt} values of various organic compounds directly from their molecular structures. The 2D-QSPR approach is simple and less error prone compared to 3D-QSPR, as it neither requires conformational search nor structural alignment [25]. Furthermore, 2D methods also have some basic advantages such as structural key type descriptors, which implicitly encode much chemical information that might otherwise be difficult to explicitly calculate [26].

2. MATERIALS AND METHOD

2.1. *Dataset*

The experimental K_{salt} data for 101 compounds in aqueous NaCl solution (Table 1) were taken from the article by Ni and Yalkowsky [6]. The reported K_{salt} values ranged from –0.068 to 0.354.

Kennard and Stones algorithm [27] has been widely used for splitting datasets into two subsets. This algorithm starts by finding two samples, based on the input variables that are the farthest apart from each other. These two samples are removed from the original dataset and put into the calibration set. This procedure is repeated until the desired number of samples has been selected in the calibration set. The advantages of this algorithm are that the calibration samples always map the measured region of the input variable space completely with respect to the induced metric and that the no validation samples fall outside the measured region. The Kennard and Stones algorithm has been considered one of the best ways to build training and validation sets [28, 29]. Using Kennard and Stones algorithm, the entire dataset was divided into two subsets: a training set of 71 compounds, and a validation set including the remaining 30 compounds.

T a b l e 1

Experimental and calculated K_{salt} data for 101 organic compounds

No.	Compound	Expt.	Calc.		h_i
			This work	Zhong et al.	
1	2	3	4	5	6
1	1,2,3-Trimethylbenzene [a]	0.321	0.274	0.291	0.0239
2	1,2,4-Trichlorobenzene [a]	0.250	0.187	0.217	0.0055
3	1,2,4-Trimethylbenzene [a]	0.293	0.274	0.291	0.0239
4	1,2-Benzanthracene [a]	0.354	0.325	0.334	0.0504
5	1,3,5-Trimethylbenzene	0.318	0.274	0.291	0.0239
6	1-Ethyl anthracene	0.313	0.300	0.334	0.0338
7	1-Ethyl naphthalene	0.273	0.264	0.290	0.0177
8	1-Methyl naphthalene	0.200	0.265	0.276	0.0184
9	1-Naphthol [a]	0.207	0.215	0.224	0.0099
10	2,4,6-Trichlorophenol	0.228	0.209	0.195	0.3468
11	2,4-Dichlorophenol	0.218	0.194	0.190	0.1835
12	2-Methyl anthracene [a]	0.336	0.302	0.320	0.0369
13	2-Naphthol	0.220	0.215	0.224	0.0099
14	5-Fluorouracil [a]	0.014	0.089	0.066	0.0253
15	6-Mercaptopurine	0.048	0.103	0.071	0.0250
16	Acenaphthene [a]	0.238	0.256	0.261	0.0158
17	Acetic acid [a]	0.064	0.110	0.133	0.0160
18	Acetone [a]	0.110	0.204	0.184	0.0062
19	Aniline [a]	0.136	0.165	0.163	0.0261
20	Anthracene	0.326	0.294	0.290	0.0293
21	Benzene	0.195	0.213	0.202	0.0244
22	Benzo[*a*]-pyrene [a]	0.328	0.338	0.350	0.0652
23	Benzoic acid	0.177	0.164	0.176	0.0144
24	Benzylamine	0.112	0.177	0.178	0.0293
25	Biphenyl	0.276	0.280	0.275	0.0214
26	Bipyridyl	0.251	0.212	0.191	0.0129
27	Butane [a]	0.217	0.207	0.216	0.0333
28	Butanoic acid	0.166	0.164	0.161	0.0140
29	Butyl acetate [a]	0.224	0.223	0.208	0.0117
30	Caffeine	0.128	0.153	0.134	0.1057
31	Chlorobenzene [a]	0.198	0.218	0.207	0.0085
32	Chrysene	0.336	0.325	0.334	0.0504
33	Cycloheptane [a]	0.343	0.271	0.214	0.0265
34	Cyclohexane	0.277	0.252	0.200	0.0276
35	Cyclohexanone	0.202	0.237	0.182	0.0144
36	Cyclopentane	0.182	0.230	0.186	0.0313
37	Cystine [a]	-0.068	-0.025	-0.040	0.4040
38	Cytosine [a]	-0.005	0.069	0.058	0.0405
39	Ethane [a]	0.162	0.182	0.188	0.0309
40	Ethylacetate	0.172	0.164	0.179	0.0140
41	Ethylbenzene	0.234	0.220	0.246	0.0161
42	Ethylene	0.127	0.124	0.176	0.0769
43	Fluorene	0.267	0.271	0.275	0.0199
44	Fluroanthene	0.339	0.309	0.306	0.0420
45	Glycine [a]	0.002	0.041	0.079	0.0651
46	Heptanoic acid	0.242	0.238	0.203	0.0136
47	Hexanoic acid	0.220	0.217	0.189	0.0120
48	Hexyl acetate [a]	0.312	0.257	0.236	0.0163
49	iso-Butyl acetate	0.225	0.246	0.221	0.0136
50	iso-Propylbenzene	0.316	0.246	0.273	0.0145
51	Leucine	0.114	0.153	0.163	0.0143
52	Lindane	0.166	0.243	0.219	0.0133
53	*m*-Chlorobenzoic acid	0.180	0.209	0.181	0.2867
54	*m*-Cresol	0.182	0.180	0.209	0.0111
55	*m*-Dichlorobenzene	0.226	0.199	0.212	0.0056
56	*m*-Dinitrobenzene	0.109	0.103	0.085	0.0843
57	Methane	0.127	0.115	0.157	0.0000
58	Methyl acetate [a]	0.185	0.160	0.165	0.0118
59	Methylcyclohexane	0.274	0.248	0.228	0.0237
60	Methylcyclopentane	0.273	0.227	0.214	0.0256

1	2	3	4	5	6
61	*m*-Nitrophenol [a]	0.147	0.134	0.121	0.0124
62	*m*-Xylene	0.248	0.244	0.261	0.0132
63	Naphthalene	0.220	0.254	0.246	0.0155
64	*n*-Hexane	0.276	0.248	0.245	0.0263
65	*n*-Hexanol	0.232	0.209	0.207	0.0325
66	*n*-Pentane	0.221	0.229	0.231	0.0290
67	*o*-Chlorobenzoic acid	0.182	0.196	0.181	0.1835
68	*o*-Dichlorobenzene	0.247	0.199	0.212	0.0056
69	*o*-Dinitrobenzene [a]	0.124	0.116	0.085	0.0537
70	*o*-Hydroxylbenzoic acid [a]	0.172	0.149	0.154	0.0129
71	*o*-Nitrophenol	0.136	0.138	0.121	0.0120
72	*o*-Xylene	0.227	0.244	0.261	0.0132
73	*p*-Dichlorobenzene	0.240	0.199	0.212	0.0056
74	*p*-Dinitrobenzene [a]	0.097	0.089	0.085	0.1225
75	Pentyl acetate [a]	0.283	0.241	0.222	0.0136
76	Phenanthrene [a]	0.272	0.294	0.290	0.0293
77	Phenol	0.111	0.168	0.180	0.0220
78	Phenylacetic acid	0.190	0.175	0.190	0.0145
79	Phenylthiourea	0.184	0.184	0.158	0.0054
80	Phenytoin	0.191	0.241	0.208	0.0174
81	Phthalic acid	0.178	0.174	0.150	0.0058
82	Piperidine	0.156	0.172	0.155	0.0562
83	*p*-Nitrophenol	0.165	0.131	0.121	0.0133
84	*p*-Nitrotoluene	0.163	0.176	0.173	0.0074
85	Progesterone	0.288	0.343	0.378	0.0643
86	Propane	0.194	0.186	0.202	0.0379
87	Propionic acid	0.132	0.132	0.147	0.0182
88	Propyl acetate	0.201	0.204	0.194	0.0105
89	*p*-Toluidine	0.170	0.176	0.193	0.0141
90	*p*-Xylene	0.251	0.244	0.261	0.0132
91	Pyrene	0.320	0.309	0.306	0.0420
92	*sec*-Butyl acetate	0.241	0.241	0.221	0.0126
93	*sec*-Butylbenzene	0.288	0.248	0.288	0.0156
94	Sulfanilamide	0.124	0.088	0.012	0.0490
95	*tert*-Butyl acetate	0.269	0.282	0.240	0.0378
96	*tert*-Butylbenzene	0.243	0.281	0.306	0.0246
97	Testosterone	0.326	0.316	0.330	0.0418
98	Theobromine [a]	0.056	0.092	0.095	0.0845
99	Theophylline	0.100	0.092	0.092	0.0845
100	Toluene	0.228	0.217	0.231	0.0138
101	Tyrosine [a]	0.048	0.114	0.142	0.0378

[a] Data used for the validation set

2.2. *Descriptor generation*

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [30]. Then, the molecular structures were used as input for the generation of 578 empirical 2D descriptors using the Dragon software [31]. These descriptors include topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, and eigenvalue-based indices.

In order to reduce redundant and non-useful information, constant or near constant values and descriptors, which have been found to be highly correlated pairwise (one of any two descriptors with a correlation greater than 0.99 [32]), were excluded in a pre-reduction step. Thus, 319 descriptors were remained to undergo subsequent descriptor selection.

2.3. *Model development and validation*

Stepwise multilinear regression analysis (MLRA) with Leave-One-Out (LOO) cross-validation was used to select descriptors for the linear QSPR models on the training set. *F*-to-enter and *F*-to-remove were 4 and 3, respectively. The models were justified by the correlation coefficient *R*, the cross-validated *R*, the adjusted *R*, the standard error of estimation *s*, the *F* ratio values, and the significance level value *p*. The adjusted R^2 is calculated using the following formula:

$$R^2_{adj} = 1 - \left[\left(\frac{N-1}{N-M-1}\right)(1-R^2)\right] \qquad (1)$$

where N is the number of members of the training set and M is the number of descriptors involved in the correlation. The adjusted R^2 is a better measure of the proportion of variance in the data explained by the correlation than R^2 (especially for correlations developed using small datasets) because R^2 is somewhat sensitive to changes in N and M. The adjusted R^2 corrects for the artificiality introduced when M approaches N through the use of a penalty function which scales the result. A variance inflation factor (VIF) was calculated to test whether multicollinearities existed among the descriptors, which is defined as:

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

where R^2_j is the squared correlation coefficient between the jth coefficient regressed against all the other descriptors in the model. Models would not be accepted if they contain descriptors with VIFs above a value of five [33].

Randomization tests were also carried out to prove the possible existence of chance correlation. To do this, the dependent variable was randomly scrambled and used in the experiment. Models were then investigated with all members in the descriptor pool to determine the most predictive models. The resulting models obtained on the training set with the randomized K_salt values should have significantly lower R^2 values than the proposed one because the relationship between the structure and property is broken. This is proof of the proposed model's validity as it can be reasonably excluded that the originally proposed model was obtained by chance correlation.

Validation of the model was further performed by using the external validation set composed of data not used to develop the prediction model. The external $R^2_\mathrm{CV,ext}$ for the validation sets is determined with Eq. (3):

$$R^2_\mathrm{CV,ext} = 1 - \frac{\sum (y_i - \tilde{y}_i)^2}{\sum (y_i - \bar{y}_{test})^2} \qquad (3)$$

where y_i and $\tilde{y}_i$ are the observed and the calculated values, respectively; and $\bar{y}_{test}$ is the averaged value for the response variable of the validation set. According to Golbraikh and Tropsha [34], a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_\mathrm{CV,ext} > 0.5 \qquad (4a)$$

$$r^2 > 0.6 \qquad (4b)$$

$$(r^2 - r_0^2)/r^2 < 0.1 \text{ or } (r^2 - r'^2_0)/r^2 < 0.1 \qquad (4c)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \qquad (4d)$$

Here:

$$r = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \qquad (5a)$$

$$r_0^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \qquad (5b)$$

$$r'^2_0 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \qquad (5c)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \qquad (5d)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \qquad (5e)$$

where $\tilde{y}^{r_0}$ and y^{r_0} are defined as $\tilde{y}^{r_0} = ky$ and $y^{r_0} = k'\tilde{y}$, respectively.

The applicability domain of a QSPR model [28, 35] must be defined if the model is to be used for screening new compounds. Predictions for only those compounds that fall into this domain may be considered reliable. Extent of extrapolation [28] is a simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i for each compound, where the QSPR model is used to predict its property.

$$h_i = x_i^T (X^T X)^{-1} x_i \qquad (6)$$

where x_i is the descriptor row-vector the i-th compound, x_i^T is the transpose of x_i, X is the descriptor matrix, X^T is the transpose of X. The warning leverage h^* is, generally, fixed at 3(m+1)/n, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation. A leverage greater than the warning leverage h^* means that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

3. RESULTS AND DISCUSSION

Stepwise MLRA with LOO cross-validation was applied on the training set to select the descriptors for the best model and the number of descriptors in the final QSPR model was determined on the basis of the dataset size and on the basis of the correlation coefficient *R*, the adjusted *R*, the significance test *F* and the standard error *s*. The *R* and *s* results during the stepwise MLRA are shown in Figure 1.

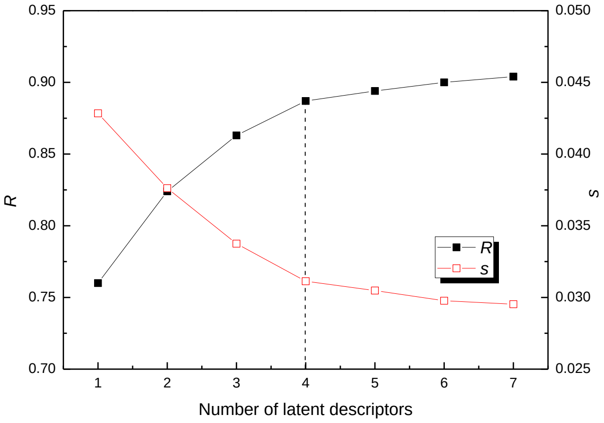

Fig. 1. *R* and *s* vs. number of latent descriptors in the best MLRA equation

The *R* increases gradually with the increased number of descriptors. When adding another descriptor did not significantly improve the statistics of a model, it was determined that the optimum subset size had been achieved. To avoid over-parameterization of the models, such as those which contain an excess of descriptors and are difficult to interpret in terms of physical interactions, an increase of the *R* value of less than 0.01 was chosen as the breakpoint criterion. Thus, a four-parameter model with *R* of 0.887 and R_{CV} of 0.870 was obtained, which is as follows:

$$K_{\text{salt}} = -0.003443[\text{T(N..O)}] + 0.006697[\text{T(O..Cl)}] + 0.09306[\text{CIC1}] - 0.1168[\text{GATS2m}] + 0.1148 \tag{7}$$

$$N = 71,\ R = 0.887,\ R_{\text{CV}} = 0.870,\ R_{\text{adj}} = 0.879,\ s = 0.031,\ F = 60.7,\ p < 0.00001$$

Here, T(N..O) is the sum of topological distances between N..O; T(O..Cl) is the sum of topological distances between O..Cl; CIC1 is the complementary information content (neighborhood symmetry of 1-order); and GATS2m is the Geary autocorrelation – lag 2/weighted by atomic masses, respectively. More information about these descriptors can be found in the Dragon software user guide [31] and the references therein.

Table 2

Results of randomization test

Iteration	R^2	R^2_{CV}	Iteration	R^2	R^2_{CV}	Iteration	R^2	R^2_{CV}	Iteration	R^2	R^2_{CV}
1	0.141	0	6	0.000	0	11	0.309	0	16	0.261	0
2	0.168	0	7	0.106	0	12	0.187	0	17	0.335	0
3	0.278	0	8	0.272	0	13	0.318	0	18	0.275	0.021
4	0.137	0	9	0.100	0	14	0.092	0	19	0.120	0
5	0.290	0	10	0.087	0	15	0.276	0	20	0.382	0

Table 3

Characteristics of descriptors in the best MLRA model

Descriptor	Descriptor type	X	DX	*t*-value	*t*-probability	VIF
Constant		1.148E-01	1.611E-02	7.128	0.000	
CIC1	Information indices	9.306E-02	6.958E-03	13.374	0.000	2.034
GATS2m	2D autocorrelations	-1.168E-01	1.994E-02	-5.856	0.000	1.847
T(N..O)	Topological descriptors	-3.443E-03	9.127E-04	-3.772	0.000	1.127
T(O..Cl)	Topological descriptors	6.697E-03	1.878E-03	3.566	0.001	1.200

Table 4

Correlation matrix between the selected descriptors and K_{salt}.

	CIC1	GATS2m	T(N..O)	T(O..Cl)	K_{salt}
CIC1	1.000				
GATS2m	0.673	1.000			
T(N..O)	–0.266	–0.248	1.000		
T(O..Cl)	–0.367	–0.263	–0.071	1.000	
K_{salt}	0.760	0.275	–0.419	–0.041	1.000

The large F ratio of 60.7 indicates that Eq. (7) does a good job of predicting the K_{salt} values. The cross-validated correlation coefficient R_{CV} = 0.870 illustrates the reliability of the model by focusing on the sensitivity of the model to the elimination of any single data point [36]. Eq. (7) has an adjusted R value of 0.879, which indicates a satisfied agreement between the correlation and variation in the data. The model was further validated by applying the randomization tests; several results are shown in Table 2. The low R^2 and $R_{CV}{}^2$ values indicate that the good results of the original model are not due to chance correlation or structural dependency of the training set. The statistical characteristics of the four descriptors are given in Table 3, which indicate that all descriptors are highly significant from the t-test values. The VIF values (less than five) and the correlation matrix as shown in Table 4 suggest that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

The calculated results of the K_{salt} values from Eq. (7) for the training and validation set are shown in Table 1 and Figure 2. The distributions of errors for the entire dataset are given in Figure 3. As the errors are distributed on both sides of the zero line, one may conclude that there is no systematic error in the model development. The following statistical parameters were obtained for the validation set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present model:

$$R^2_{\text{CV,ext}} = 0.867 > 0.5$$

$$r^2 = 0.905 > 0.6$$

$$(r^2 - r_0^2)/r^2 = (0.905 - 0.980)/0.905 < 0.1$$

$$\text{or } (r^2 - r'^2_0)/r^2 = (0.905 - 0.988)/0.905 < 0.1$$

$$0.85 \le k = 1.0234 \le 1.15 \text{ or } 0.85 \le k' = 0.9396 \le 1.15$$

It needs to be pointed out that no matter how robust and validated a QSPR model may be, it cannot be expected to reliably predict the modeled property for the entire universe of compounds. Therefore, before a QSPR model is put into use for screening compounds, its applicability domain must be defined and predictions for only those compounds that fall in this domain can be considered reliable. The extent of extrapolation method was applied to the 101 compounds that constitute the entire dataset. The leverages for all compounds were computed (as listed in Table 1) and only three compounds (2,4,6-Trichlorophenol, Cysteine and m-Chlorobenzoic acid) were found to fall outside the domain of the model (warning leverage limit 0.2113).

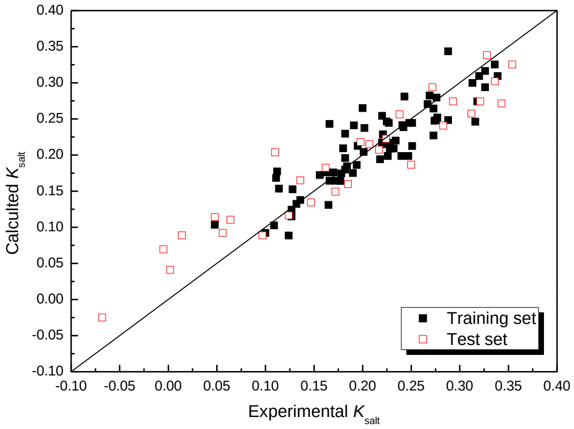

Fig. 2. Experimental vs. calculated K_{salt} values

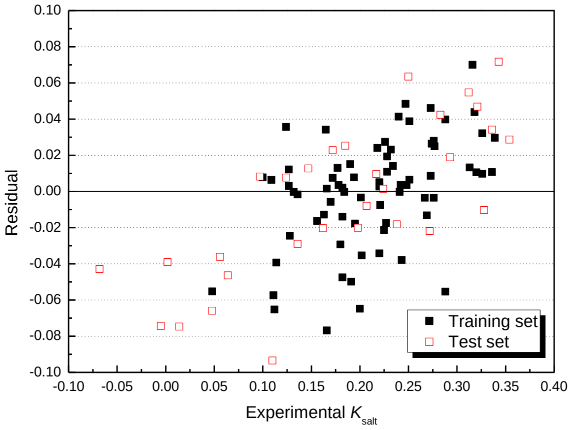

Fig. 3. Residuals vs. experimental K_{salt}

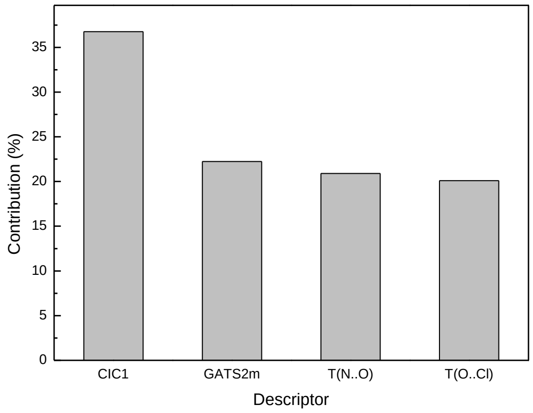

Fig. 4. Relative contributions of the descriptors to the QSPR model

To further test the suitability of the QSPR model developed in our study, the obtained statistical parameters were compared with those calculated from previously reported models [22, 23]. It can be seen that the performance of the present model (R = 0.887 and s = 0.031) is a little better than that of Zhong's model (R = 0.887 and s = 0.042). Moreover, all of the descriptors in the present model could be directly obtained from the molecular structure; while the zero-order variable connectivity index $^{0}\chi^{f}$ in Zhong's model should be calculated from the optimal weights for the non-hydrogen atoms by fitting the data of the training set. In our previous work [23], a five-parameter linear model based on 3D descriptors was obtained, with an average absolute error (AAE) of 0.023 for the entire dataset. The AAE of the present 2D-QSPR model (0.026) is compared to that of the 3D-QSPR model, while the 2D-QSPR models normally imply a quicker calculation process.

Based on a previously described procedure [37, 38], the relative contributions of the four descriptors to the present model were determined and are plotted in Figure 4. The significance of the descriptors involved in the model decreases in the following order: CIC1 (36.8%) > GATS2m (22.2%) > T(N..O) (20.9%) > T(O..Cl) (20.1%).

The first important descriptor is the 1-order complementary information content CIC1, which explains 36.8% contribution of the total and correlates relatively high (R = 0.760) with the target experimental K_{salt} values. The descriptor CIC1 [31] is defined by Eq. 7(a), where A_g is the cardinality of the g*th* equivalence class, *nAT* is the total number of atoms, and IC1 is the 1-order information content itself defined by Eq. 7(b). CIC1 describes the atomic connectivity in the molecule and encodes the size and atomic constitution of the compound. These parameters directly affect the intermolecular interaction. The positive coefficient of CIC1 indicates that the compounds with larger values for this descriptor would have larger K_{salt} values. Thus, this descriptor could be an indicator for compounds that have a large K_{salt} value.

$$CIC1 = \log_2 nAT - IC1 \tag{7a}$$

$$IC1 = -\sum_{i=1}^{1} \frac{A_g}{nAT} \log_2 \frac{A_g}{nAT} \tag{7b}$$

The second important descriptor is the Geary autocorrelation GATS2m, which explains 22.2% of the contributions. The descriptor GATS2m [31] is defined by Eq. (8), where m is the atomic mass, $\bar{m}$ is its average value on the molecule, *nSK* is the number of non-hydrogen atoms, and δ_{ij} is the Kronecker delta (δ_{ij} = 1 if $d_{\mathrm{ij}} = k$, zero otherwise, d_{ij} being the topological distance between two considered atoms). Δ is the sum of the Kronecker deltas, i.e. the number of atom pairs at distance equal to k. The negative sign of GATS2m in Eq. (6) indicates that the compounds containing atoms with larger atomic masses would possess higher K_{salt}, because this descriptor increases with increased atomic masses.

$$GATS2m = \frac{\frac{1}{2\Delta} \cdot \sum_{i=1}^{nSK}\sum_{j=1}^{nSK} \delta_{ij} \cdot (m_i - m_j)^2}{\frac{1}{(nSK-1)} \cdot \sum_{i=1}^{nSK} (m_i - \bar{m})^2} \tag{8}$$

The presence of T(N..O) and T(O..Cl) in the model reflects the influence of topological distances between certain atoms on the K_{salt} values. The coefficient of T(N..O) is negative, suggesting that a smaller value of T(N..O) would be beneficial to the K_{salt} values. The positive sign of T(O..Cl) indicates that the compounds with a larger sum of topological distances between O..Cl would have larger K_{salt} values.

4. CONCLUSIONS

In this work, a general QSPR model with good statistical parameters (R = 0.887 and s = 0.031) was reported for the prediction of Setschenow constants of a variety of organic compounds. The results of leave-one-out cross-validation, randomization tests, and validation through the validation set illustrated the reliability of the proposed model. The most significant descriptor in the model is the 1-order complementary information content (CIC1), which encodes the size and atomic constitution of the compound and shows the importance of the intermolecular interaction to the Setschenow constants. The proposed model is predictive because all of the descriptors involved are two-dimensional and can be calculated easily as long as the molecular structure of the compound concerned is known.

Acknowledgements. This work was supported by the Educational Commission of Hubei Province (T201507) and the Natural Science Foundation of China (No. 51003082). The authors gratefully wish to express their thanks to the reviewers for critically reviewing the manuscript and making important suggestions.

REFERENCES

[1] B. E. Conway, J. E. Desnoyers, A. C. Smith, On the Hydration of Simple Ions and Polyions, *Philos. Trans. R. Soc.* **131**, 389–437 (1964).

[2] W. L. Masterton, T. P. Lee, Salting coefficients from scaled particle theory, *J. Phys. Chem.* **74**, 1776–1782 (1970).

[3] S. Miyazaki, M. Oshiba, T. Nadai, Precaution on use of hydrochloride salts in pharmaceutical formulation, *J. Pharm. Sci.* **70**, 594–596 (1981).

[4] P. L. Gould, Salt Selection for Basic Drugs, *Int. J. Pharm.* **33**, 201–217 (1986).

[5] N. Ni, M. M. El-Sayed, T. Sanghvi, S. H. Yalkowsky, Estimation of the effect of NaCl on the solubility of organic compounds in aqueous solutions, *J. Pharm. Sci.*, 1620–1625 (2000).

[6] N. Ni, S. H. Yalkowsky, Prediction of Setschenow constants, *Int. J. Pharm.* **254**, 167–172 (2003).

[7] X. J. Yao, Y. W. Wang, X. Y. Zhang, R. S. Zhang, M. C. Liu, Z. D. Hu, B. T. Fan, Radial basis function neural network-based QSPR for the prediction of critical temperature, *Chemom. Intell. Lab. Syst.* **62**, 217–225 (2002).

[8] J. Xu, B. Guo, B. Chen, Q. Zhang, A QSPR treatment for the thermal stabilities of second-order NLO chromophore molecules, *J. Mol. Model.* **12**, 65–75 (2005).

[9] J. Xu, L. Liu, W. Xu, S. Zhao, D. Zuo, A general QSPR model for the prediction of θ(lower critical solution temperature) in polymer solutions with topological indices, *J. Mol. Graph. Model.* **26**, 352–359 (2007).

[10] J. Xu, H. Liang, B. Chen, W. Xu, X. Shen, H. Liu, Linear and nonlinear QSPR models to predict refractive indices of polymers from cyclic dimer structures, *Chemom. Intell. Lab. Syst.* **92**, 152–156 (2008).

[11] J. Xu, H. Zhang, L. Wang, W. Ye, W. Xu, Z. Li, QSPR analysis of infinite dilution activity coefficients of chlorinated organic compounds in water, *Fluid Phase Equilib.* **291**, 111–116 (2010).

[12] G. Liang, J. Xu, L. Liu, QSPR analysis for melting point of fatty acids using genetic algorithm based multiple linear regression (GA-MLR), *Fluid Phase Equil.* **353**, 15–21 (2013).

[13] X. Wang, Y. Sun, L. Wu, S. Gu, R. Liu, L. Liu, X. Liu, J. Xu, Quantitative structure-affinity relationship study of azo dyes for cellulose fibers by multiple linear regression and artificial neural network, *Chemom. Intell. Lab. Syst.* **134**, 1–9 (2014).

[14] D. Wang, Y. Yuan, S. Duan, R. Liu, S. Gu, S. Zhao, L. Liu, J. Xu, QSPR study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural network, *Chemom. Intell. Lab. Syst.* **143**, 7–15 (2015).

[15] Y. Yuan, Y. Sun, D. Wang, R. Liu, S. Gu, G. Liang, J. Xu, Quantitative structure-property relationship study of liquid vapor pressures for polychlorinated diphenyl ethers, *Fluid Phase Equil.* **391**, 31–38 (2015).

[16] X. Yu, B. Yi, X. Wang, Prediction of refractive index of vinyl polymers by using density functional theory, *J. Comput. Chem.* **28**, 2336–2341 (2007).

[17] X. Yu, Support Vector Machine-based QSPR for the Prediction of Glass Transition Temperatures of Polymers, *Fiber. Polym.* **11**, 757–766 (2010).

[18] X. Yu, R. Yu, Setschenow constant prediction based on the IEF-PCM calculations, *Ind. Eng. Chem. Res.* **52**, 11182–11188 (2013).

[19] A. Afantitis, G. Melagraki, K. Makridima, A. Alexandridis, H. Sarimveis, O. Iglessi-Markopoulou, Prediction of high weight polymers glass transition temperature using RBF neural networks, *J. Mol. Struc. Theochem* **716**, 193–198 (2005).

[20] A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Prediction of intrinsic viscosity in polymer-solvent combinations using a QSPR model, *Polymer* **47**, 3240–3248 (2006).

[21] G. Melagraki, A. Afantitis, Enalos KNIME nodes: exploring corrosion inhibition of steel in acidic medium, *Chemom. Intell. Lab. Syst.* **123**, 9–14 (2013).

[22] Y. Li, Q. Hu, C. Zhong, Topological modeling of the Setschenow constant, *Ind. Eng. Chem. Res.* **43**, 4465–4468 (2004).

[23] J. Xu, L. Wang, L. Wang, G. Liang, X. Shen, W. Xu, Prediction of Setschenow constants of organic compounds based on a 3D structure representation, *Chemom. Intell. Lab. Syst.* **107**, 178–184 (2011).

[24] J. Xu, L. Wang, L. Wang, X. Shen, W. Xu, QSPR Study of Setschenow Constants of Organic Compounds Using MLR, ANN, and SVM Analyses, *J. Comput. Chem.* **32**, 3241–3252 (2011).

[25] M. Shen, A. LeTiran, Y. Xiao, A. Golbraikh, H. Kohn, A. Tropsha, Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods, *J. Med. Chem.* **45**, 2811–2823 (2002).

[26] D. Rogers, A. J. Hopfinger, Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships, *J. Chem. Inf. Comput. Sci.* **34**, 854–866 (1994).

[27] R. W. Kennard, L. A. Stone, Computer aided design of experiments, *Technometrics* **11**, 137–148 (1969).

[28] 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**, 69–77 (2003).

[29] W. Wu, B. Walczak, D. L. Massart, S. Heuerding, F. Erni, I. R. Last, K. A. Prebble, Artificial neural networks in classification of NIR spectral data: Design of the training set, *Chemom. Intell. Lab. Syst.* **33**, 35–46 (1996).

[30] C. T. Klein, D. Polheim, H. Viernstein, P. Wolschann, Predicting the free energies of complexation between cyclodextrins and guest molecules: linear versus nonlinear models, *Pharm. Res.* **17**, 358–365 (2000).

[31] R. Todeschini, V. Consonni, A. Mauri, M. Pavan. *TALETE* srl, Milan, 2006.

[32] H. Liu, P. Gramatica, QSAR study of selective ligands for the thyroid hormone receptor β, *Bioorgan. Med. Chem.* **15**, 5251–5261 (2007).

[33] A. J. Holder, D. M. Yourtee, D. A. White, A. G. Glaros, R. Smith, Chain melting temperature estimation for phosphatidyl cholines by quantum mechanically derived quantitative structure property relationships, *J. Comput. Aid. Mol. Des.* **17**, 223–230 (2003).

[34] A. Golbraikh, A. Tropsha, Beware of q^2!, *J. Mol. Graph. Model.* **20**, 269–276 (2002).

[35] M. Shen, C. Béguin, A. Golbraikh, J. P. Stables, H. Kohn, A. Tropsha, Application of predictive QSAR models to database mining: Identification and experimental validation of novel anticonvulsant compounds, *J. Med. Chem.* **47**, 2356–2364 (2004).

[36] J. Xu, B. Chen, Q. Zhang, B. Guo, Prediction of refractive indices of linear polymers by a four-descriptor QSPR model, *Polymer* **45**, 8651–8659 (2004).

[37] F. Zheng, E. Bayram, S. P. Sumithran, J. T. Ayers, C.-G. Zhan, J. D. Schmitt, L. P. Dwoskin, P. A. Crooks, QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release, *Bioorg. Med. Chem.* **14**, 3017–3037 (2006).

[38] R. Guha, P. C. Jurs, Interpreting computational neural network QSAR models: A measure of descriptor importance, *J. Chem. inf. Model.* **45**, 800–806 (2005).