Multi-objective Modeling and Assessment of Partition Properties: A GA-Based Quantitative Structure-Property Relationship Approach

YIN, Chun-Sheng[*,a,b] **(印春生)** **LIU, Xin-Hui**[b] **(刘新会)** **GUO, Wei-Min**[c] **(郭卫民)**
LIU, Shu-Shen[b] **(刘树深)** **HAN, Shuo-Kui**[b] **(韩朔睽)** **WANG, Lian-Sheng**[b] **(王连生)**

[a] *School of Environmental Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China*
[b] *State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu 210093, China*
[c] *Department of Applied Chemistry, University Science and Technology of China, Hefei, Anhui 230026, China*

In this work a multi-objective quantitative structure-property relationship (QSPR) analysis approach was reported based on the study on three partition properties of 50 aromatic sulfur-containing carboxylates. Here multi-objectives (properties) were taken as a vector for QSPR modeling. The quantitative correlations for partition properties were developed using a genetic algorithm-based variable-selection approach with quantum chemical descriptors derived from AM1-based calculations. With the QSPR models, the aqueous solubility, octanol/water partition coefficients and reversed-phase HPLC capacity factors of sulfur-containing compounds were estimated and predicted. Using GA-based multivariate linear regression with cross-validation procedure, a set of the most promising descriptors was selected from a pool of 28 quantum chemical semi-empirical descriptors, including steric and electronic types, to integrally build QSPR models. The selected molecular descriptors included the net charges on carboxyl group (Q_{OC}), the 2nd power of net charges on nitrogen atoms (Q_N^2), the net atomic charge on the sulfur atoms (Q_S), the van der Waals volume of molecule (V), the most positive net atomic charge on hydrogen atoms (Q_H) and the measure of polarity and polarizability (π), which were main factors affecting the distribution processes of the compounds under study. The statistically best QSPR models of six descriptors were simultaneously obtained by GA-based linear regression analysis. With the selected descriptors and the QSPR equations, mechanisms of partition action of the Sulfur-containing carboxylates were able to be investigated and interpreted.

Keywords multi-objective QSPR, partition property, quantum chemical semi-empirical descriptor, sulfur-containing carboxylate, genetic algorithm

Introduction

Some physicochemical properties of organic pollutants such as partition properties during an environmental process may be affected by the same factors and their action mechanisms can be discussed together. The useful means to assess this process is obviously the quantitative structure-activity/property relationship (QSAR/QSPR) approach. Indirect observation of chemical properties is the goal for the extensive use of QSPR approaches, for good reason such as speed and economy. The development of QSPR can be helpful in the understanding of environmental action mechanism and can obtain a reliable and predictive model for predicting the properties of new chemical substances. Here three partition properties, including the aqueous solubility, octanol/water partition coefficients, and capacity factors of reversed-phase high performance liquid chromatography (HPLC) of 50 aromatic sulfur-containing compounds are taken as examples. The reasons for the selection of these chemicals as investigated objects are: (1) These compounds, used extensively either as intermediates in the manufacture of pesticides, herbicides and drugs, or as floatation agents and extractants in the petrochemical and metallurgical industries,[1,2] are being introduced into the environment; (2) Their partition properties have been systematically measured and reported in our lab. Thereinto, some data of the latter 20 compounds are newly measured and reported in this paper; and (3) The data set for these chemicals are much complete and suitable for multi-objective modeling study.

In recent years, quantum chemical descriptors were used in the QSAR/QSPR study in environmental chemistry by some research workers.[3-7] The descriptors derived from quantum computation have several advantages: Firstly, quantum chemical descriptors can be applied to predict toxicity and physicochemical properties of new chemicals before they are synthesized and introduced into environment. They are favorable to the ecological risk assessment and management of toxic and harmful chemicals. Thus the pollution prevention can be realized. Secondly, the quantum chemical descriptors have explicit physical meaning,

* E-mail: csyin@sjtu.edu.cn; Tel.: 021-54741065; Fax: 021-54740825
Received January 28, 2002; revised April 7, 2003; accepted May 2, 2003.
Project supported by the Research Foundation for the Doctoral Program of Higher Education of China and China Postdoctoral Science Foundation.

which led to their convenient application in studying interaction modes between toxicants and acceptors, and molecular structure features affecting physicochemical (environmental) properties of organic pollutants with QSAR models. Especially the quantum chemical descriptors may be used to perform QSAR/QSPR studies of heterogeneous compounds. Furthermore, they are not restricted to closely related compounds, and can be easily obtained, and describe clearly defined molecular properties. For these reasons, there are quite a few examples using quantum semi-empirical descriptors in the environmental science studies.[8-15]

Genetic algorithms (GA) always allow the exploration of the whole experimental space: due to the occurrence of the mutations, each possible combination can occur at any moment. GAs, which are based on the principle of Darwinian evolution, appear as a powerful tool to optimize many problems associated with drug design in recent years. They can be used to develop one or more QSAR or QSPR models for problems involved in drug or molecular design.[16] The results obtained by GAs are a whole population (group) of solutions: the user can then choose the one (usually the statistically best one) he prefers, taking into the response (dependent) and independent variables account at the same time. The general steps for GA-based MLR involved in a QSAR/QSPR study include compilation of a data set, calculation of descriptors to numerically encode the structure features of the data set compounds, feature selection to choose a small subset of descriptors that relate molecular structure to physicochemical property, and validation of the models developed.

Arx and coworkers used GA-based methods to determine reaction rate constants with limited concentration.[17] Gramatica and coworkers employed a genetic algorithm variable subset selection strategy to select the most representative training set and the best descriptors subset for predicting degradation rate constants k_{OH} and k_{ON_3} of organic compounds in the troposphere.[18]

The aim of this study was to set up multi-objective quantitative relationships with the aqueous solubility, octanol/water partition coefficients, and reversed-phase high-performance liquid chromatographic (HPLC) capacity factors of 50 aromatic sulfur-containing carboxylates by using GA-based multiple linear regression (GA-MLR). The authors try to investigate common factors affecting the three partition properties. First, the theoretical bases of quantum semi-empirical descriptors and GA are briefly presented. The quantum chemical descriptors are used to select the best subset of descriptive variable combination and be related to experimental partition properties, giving a series of quantitative models. Then these quantitative correlations are used to try to explain the environmental action mechanisms of carboxylates in the partition process. The results should be valuable in evaluating the potential behavior of suchlike chemicals.

Experimental

Samples

A total of 50 aromatic sulfur-containing compounds were used as the investigated objects for the QSPR studies in this work. They were synthesized in our laboratory and mainly used as intermediates in drug or pesticide development. Their purity was monitored by HPLC to ensure that no interference peak had occurred.

Determination of aqueous solubility

The aqueous solubility of these compounds was determined by shake-flake according to OECD guideline[19] for testing of chemicals. All operations were conducted at atmospheric pressure and (25 ± 0.5) ℃. Samples of the solution were centrifuged at 15000 r/min, and were quantitatively measured with a UV-spectrophotometer against water blank.

Determination of partition coefficients

The octanol/water partition coefficients were determined by shake-flake method as described by the OECD guideline[19] for the testing of these chemicals at atmospheric pressure and (25 ± 0.5) ℃ followed by centrifuging and analysis of chemical in the water phase by the method used for the chemical solubility studies.

Determination of reversed-phase HPLC *capacity factors*

A Nucleosil C18 column, 15 cm × 4.6 mm (made by Dalian Institute of Chemical Physics, Chinese Academy of Science), was used with flow rate of 1.0 mL/min at ambient temperature (20 ± 2) ℃. The detector was set at 230 nm. The mobile phases consist of different volume fraction of methanol in water (100/0, 95/5, 90/10, 85/15, 80/20, 75/25, 70/30), respectively. The column dead time (t_0) was determined by the injection of $NaNO_3$ dissolved in methanol, and the capacity factor $k' = [k' = (t_R - t_0)/t_0]$ of these compounds were measured. t_R is retention time.

Calculation of geometric and electronic descriptors

The molecules were sketched as two-dimensional structures using the CS Chem3D 5.0[20] software to generate the starting geometry. Then geometric optimization was performed to generate three-dimensional representations of the compounds; geometric and electronic properties were determined by the AM1 method of the MOPAC 97 program. With an optimum geometry, the molecular surface (S_A in nm^2), volume (V in nm^3), and ovality (O) were calculated by Connolly method.[21,22] Molecular weight

(M_W) was also included. The electronic descriptors such as dipole moment (μ in D), polarizability (α in a.u.), energy of the highest occupied molecular orbital (E_{HOMO} in eV), energy of the lowest unoccupied molecular orbital (E_{LUMO} in eV), atomic charge (in a.c.u.), were achieved. All possible sums of squared charges for each given element, and that of the absolute values of atomic charges on different functional groups were generated. Because of the expected nonlinearity of the model, all squared and square-rooted descriptors were generated. The most negative atomic charge (Q in a.c.u.) and the most positive charge of a hydrogen atom (Q_H in a.c.u.) in the solute molecule were obtained. A total of 28 quantum chemistry-based steric and electronic descriptors were calculated for each sulfur-containing aromatic compounds. The data set of full descriptors for QSPR modeling is given in Table 1.

Variable selection and QSPR *model assessment*

Using GA-based MLR feature selection procedures, the dependent variables, *i.e.*, the three partition properties, were used at the same time to find subsets of quantum descriptors that provide a set of good relationships with the partition properties. All of the descriptors were subjected to multi-objective feature selection to remove those that did not contribute useful information to the variable pool. This multi-objective-based feature selection left a set of reduced and predictive descriptors for the studied compounds. The reduced dimension set of informative descriptors was then used to build the quantitative relationships between molecular structure and partition properties using the multivariate linear regression with leave-one-out cross validation (LOOCV) procedures. A series of subsets of various sizes were investigated and used to create statistically valid linear models, with the quality of the model based on the predicted results with LOOCV, the correlation coefficients and the root-mean-square error (*RMSE*) between the predicted and experimental values. When adding or omitting another descriptor to the descriptor combination did not obviously improve the statistics of the models, it was determined that the optimum subset and the best predictive QSPR models have been obtained for the MLR modeling.

Table 1 Data set of full descriptors discussed in this work

No.	Descriptor	Name
1	M_W	molecular weight;
2	S_A	van der Waals area of molecule;
3	S_A^2	the 2nd power of van der Waals area of molecule;
4	O	ovality of molecule;
5	O^2	the 2nd power of ovality of molecule;
6	μ	dipole moment of molecule;
7	Q_{ON}	sum of absolute values of atomic charge on oxygen and nitrogen atoms in the nitro-group;
8	Q_{OS}	sum of atomic charge on oxygen atoms in the sulfinyl/sulfonyl group;
9	Q_{OC}	sum of absolute values of atomic charge on carbon and oxygen atoms in the carboxyl group;
10	Q_N	the net charges on nitrogen atoms;
11	Q_N^2	the 2nd power of net charges on nitrogen atoms;
12	Q_N^4	the 4th power of net charges on nitrogen atoms;
13	Q_O	net atomic charge on the oxygen atoms;
14	Q_O^2	the 2nd power of net atomic charge on the oxygen atoms;
15	Q_O^4	the 4th power of net atomic charge on the oxygen atoms;
16	Q_S	net atomic charge on the sulfur atoms;
17	Q_S^2	the 2nd power of net atomic charge on the sulfur atoms;
18	Q_S^4	the 4th power of net atomic charge on the sulfur atoms;
19	V	van der Waals volume of molecule;
20	α	average molecular polarizability;
21	E_{HOMO}	energy of the highest occupied molecular orbital;
22	E_{LUMO}	energy of the lowest unoccupied molecular orbital;
23	E_{LUMO}^2	the 2nd power of energy of the lowest unoccupied molecular orbital;
24	Q	the largest negative atomic charge on an atom;
25	Q_H	the most positive net atomic charge on hydrogen atoms;
26	π	a measure of polarity and polarizability;
27	E_B	the magnitude of the difference between E_{HOMO} of the solute and E_{LUMO} of water, divided by 100;
28	E_A	the magnitude of the difference between E_{LUMO} of the solute and E_{HOMO} of water, divided by 100.

Results and discussion

Calculation (estimation and prediction)

The data set The data set was composed of 50 structurally diverse aromatic sulfur-containing carboxylates. Their experimentally measured S_W, K_{OW} and k_W values in logarithm mode were reported in our previous work (from No. 1 to 30 in Table 1)[22-24] and partly in this work (from No. 31 to 50), and are listed in Table 2 together with quantum descriptors.

Table 2 Experimental partition properties for 50 sulfur-containing aromatic esters with the quantum descriptors

No.	Compound	Q_{OC}	Q_{N2}	Q_S	V ($\times 10^3$)	Q_H	π	log S_W	log K_{OW}	log k_W
1	2-NO_2PhSCH_2CO_2Me	1.12800	0.38143	0.21490	1.69932	0.19880	0.66880	−0.76000	1.88000	2.28
2	4-Cl-2-NO_2PhSCH_2CO_2Me	1.12420	0.39816	0.41010	1.78056	0.20840	0.73230	−1.70000	2.24000	3.11
3	4-NO_2PhSO_2C(CH_2)$_2CO_2$Me	1.09990	0.39993	3.00180	2.06369	0.20100	0.63580	−3.38000	1.33000	1.06
4	4-NO_2PhSO_2C(CH_2)$_2CO_2$-*i*-Pr	1.13380	0.39917	2.99460	2.40682	0.20020	0.62010	−4.26000	2.05000	1.75
5	4-NO_2PhSO_2C(CH_2)$_3CO_2$-*i*-Pr	1.10030	0.39955	2.95840	2.62021	0.20040	0.59810	−3.76000	2.36000	2.08
6	4-NO_2PhSO_2C(CH_2)$_5CO_2$-*i*-Pr	1.06330	0.39904	2.91900	2.95836	0.20020	0.59450	−4.88000	2.84000	3.17
7	4-NO_2PhSO_2C(CH_2)$_6CO_2$-*i*-Pr	1.06310	0.39904	2.90830	3.12079	0.20050	0.58720	−5.07000	3.41000	3.55
8	4-BrPhSO_2C(CH_2)$_2CO_2$Me	1.13590	0.00000	3.02170	2.05327	0.18480	0.61350	−3.67000	2.32000	1.55
9	4-BrPhSO_2C(CH_2)$_3CO_2$Me	1.11290	0.00000	2.96250	2.24166	0.18590	0.59800	−3.55000	2.45000	1.69
10	4-BrPhSO_2C(CH_2)$_4CO_2$Me	1.09490	0.00000	2.93360	2.41210	0.18580	0.58830	−4.01000	2.73000	2.12
11	4-BrPhSO_2C(CH_2)$_5CO_2$Me	1.05210	0.00000	2.92460	2.58758	0.18680	0.58890	−4.48000	2.94000	2.55
12	4-ClPhSO_2C(CH_2)$_2CO_2$Me	1.13680	0.00000	3.02350	1.99421	0.18610	0.61560	−3.31000	2.03000	1.23
13	4-ClPhSO_2C(CH_2)$_3CO_2$Me	1.08390	0.00000	2.97370	2.17502	0.18710	0.60170	−3.00000	2.28000	1.40
14	4-ClPhSO_2C(CH_2)$_2CO_2$-*i*-Pr	1.13530	0.00000	2.99980	2.35581	0.18680	0.59500	−3.54000	2.64000	1.86
15	4-ClPhSO_2C(CH_2)$_2CO_2$-*t*-Bu	1.12010	0.00000	2.99990	2.54031	0.18660	0.58160	−4.12000	2.68000	2.19
16	4-ClPhSO_2C(CH_2)$_4CO_2$-*i*-Pr	1.07690	0.00000	2.93550	2.71206	0.18660	0.57220	−4.65000	3.16000	2.61
17	4-ClPhSO_2C(CH_2)$_5CO_2$-*i*-Pr	1.04590	0.00000	2.92410	2.90672	0.18760	0.56960	−5.54000	3.49000	3.01
18	4-ClPhSO_2C(CH_2)$_6CO_2$-*i*-Pr	1.04700	0.00000	2.91350	3.05509	0.18740	0.56620	−5.52000	3.83000	3.55
19	4-MePhSO_2C(CH_2)$_2CO_2$-*i*-Pr	1.12750	0.00000	3.00100	2.36910	0.17220	0.59790	−3.23000	2.52000	1.60
20	4-MePhSO_2C(CH_2)$_3CO_2$-*i*-Pr	1.07140	0.00000	2.95230	2.59201	0.17270	0.57650	−3.34000	2.78000	1.81
21	4-MePhSO_2C(CH_2)$_2CO_2$Me	1.13690	0.00000	3.02320	2.02214	0.17170	0.61630	−2.88000	1.77000	1.02
22	4-MePhSO_2C(CH_2)$_2CO_2$Et	1.14090	0.00000	3.00100	2.18927	0.17240	0.61050	−3.01000	2.23000	1.31
23	4-MePhSO_2C(CH_2)$_3CO_2$Et	1.10150	0.00000	2.97230	2.40522	0.17290	0.58500	−2.96000	2.31000	1.47
24	4-MePhSO_2C(CH_2)$_4CO_2$-*i*-Pr	1.07830	0.00000	2.93860	2.73863	0.17180	0.57310	−3.91000	2.88000	2.24
25	4-MePhSO_2C(CH_2)$_5CO_2$-*i*-Pr	1.04770	0.00000	2.92840	2.93338	0.17320	0.57020	−4.62000	3.21000	2.67
26	4-MePhSO_2C(CH_2)$_5CO_2$Me	1.05540	0.00000	2.93000	2.55559	0.17370	0.59020	−4.61000	2.54000	1.90
27	PhSO_2C(CH_2)$_2CO_2$Me	1.13620	0.00000	3.02230	1.85338	0.17290	0.61390	−2.26000	1.43000	0.66
28	PhSO_2C(CH_2)$_3CO_2$Me	1.11380	0.00000	2.96500	2.04110	0.17340	0.59620	−3.00000	1.63000	0.85
29	PhSO_2C(CH_2)$_4CO_2$Me	1.09700	0.00000	2.93660	2.21158	0.17310	0.58540	−2.55000	1.98000	1.13
30	PhSO_2C(CH_2)$_5CO_2$Me	1.05470	0.00000	2.92770	2.38745	0.17460	0.58510	−3.85000	2.30000	1.52
31	4-NO_2PhSO_2CH(Me)CO_2Me	1.10100	0.39917	2.96100	2.03624	0.22300	0.61064	−2.96000	1.06000	0.65
32	4-NO_2PhSO_2C(Me)$_2CO_2$Me	1.08400	0.39942	2.94000	2.19266	0.20200	0.60493	−3.39000	1.38000	0.94
33	4-NO_2PhSO_2C(Et)$_2CO_2$Me	1.05000	0.39942	2.91000	2.51677	0.18700	0.59942	−4.18000	2.24000	1.64
34	4-NO_2PhSO_2C(*n*-Bu)$_2CO_2$-Me	1.15800	0.31259	2.84700	3.24524	0.20100	0.58433	−5.55000	3.38000	3.13
35	4-NO_2PhSO_2C(CH_2Ph)$_2CO_2$Me	0.97100	0.31360	2.80200	3.48368	0.18170	0.68023	−6.24000	4.46000	3.38
36	4-NO_2PhSO_2C(*n*-Bu)$_2CO_2$Et	1.06900	0.39917	2.92200	3.42052	0.20160	0.55512	−5.76000	3.81000	3.47
37	4-NO_2PhSO_2C(Me)(CH_2Ph)CO_2-Et	1.03500	0.39930	2.91900	3.04362	0.20100	0.63707	−5.44000	3.40000	2.34
38	4-NO_2PhSO_2C(Me)(CH_2CH=CH_2)CO_2Et	1.07700	0.39917	2.93900	2.64507	0.20160	0.60505	−4.56000	2.30000	1.71

Continued

No.	Compound	Q_{OC}	Q_{N2}	Q_S	V ($\times 10^3$)	Q_H	π	log S_W	log K_{OW}	log k_W
39	4-NO_2PhSO_2C(Me)(CH_2-∝-Naph)CO_2Et	1.02800	0.39955	2.92000	3.40780	0.20180	0.69065	-5.83000	4.40000	3.18
40	4-NO_2PhSO_2C(*n*-Bu)$_2CO_2$-*i*-Pr	1.10200	0.39942	2.90600	3.62379	0.19930	0.54722	-5.85000	4.06000	3.62
41	4-NO_2PhSO_2CH(Me)CO_2CH-$(CH_2)_5$	0.93900	0.31394	2.83800	2.81509	0.18190	0.59170	-4.61000	2.82000	2.11
42	4-NO_2PhSO_2CH(CH_2CO_2Et) CO_2-Me	2.20200	0.39879	2.96600	2.55532	0.22900	0.62239	-3.04000	1.40000	1.03
43	4-NO_2PhSO_2CH(CH_2CO_2-*i*-Pr)-CO_2-*i*-Pr	1.93400	0.31304	2.83700	3.13158	0.19880	0.58207	-4.29000	2.18000	1.97
44	4-NO_2PhSO_2C(CH_2CO_2Et)$_2CO_2$-*i*-Pr	2.33700	0.39791	3.00600	3.83582	0.20060	0.59690	-4.81000	3.56000	3.06
45	4-NO_2PhSO_2C(=CHPh)CO_2Me	1.11700	0.39904	3.03600	2.53133	0.19870	0.72973	-4.57000	2.90000	1.55
46	4-NO_2PhSO_2C(=CHPh)CO_2Et	1.12400	0.39904	3.03500	2.72796	0.17270	0.70994	-4.62000	3.20000	1.95
47	4-NO_2PhSO_2C(=CHPh)CO_2-*i*-Pr	1.00600	0.31382	2.88500	2.91975	0.17810	0.68931	-5.07000	3.62000	2.26
48	4-NO_2PhSO_2C(=CHPh)CO_2-*i*-Bu	1.00300	0.31382	2.88400	3.08706	0.19830	0.68006	-5.28000	3.68000	2.42
49	4-$MePhSO_2C$(=CHPh)CO_2-*i*-Pr	1.12700	0.00000	3.03500	2.87511	0.18580	0.67250	-5.50000	3.92000	2.28
50	4-$ClPhSO_2C$(=CHPh)CO_2-*i*-Pr	1.12500	0.00000	3.03400	2.86916	0.18000	0.66828	-5.65000	4.18000	2.41

Variable selection and regression analysis The full sets of quantum parameters were used to select a set of promising common descriptive variables and correlate the partition properties, *i.e.* log S_W, log K_{OW} and log k_W, of these compounds using the GA-based MLR. As the GA-based regression procedure was manipulated, a variety of correlation models were obtained for these three partition properties. Here, seven, six and five-descriptor quantitative models for log S_W, log K_{OW} and log k_W respectively were investigated. The statistics of the 7-, 6- and 5-parameter models for the partition properties are listed in Table 3. According to Table 3, the correlation coefficients (R) and the root-mean-square errors ($RMSE$), and the LOOV correlation coefficients (R_{CV}) and root-mean-square errors ($RMSE_{CV}$) calculated for the data set derived from the 6-parameter models are much close to those derived from the 7-parameter models, and better than those derived from the 5-parameter models. Especially, the F statistics for the 6-parameter models are found the best. This resulted in the focus of this study on the 6-parameter models given as follows:

$$\log S_W = 8.2317 + 1.4708Q_{OC} + 1.2950Q_N^2 - 0.5938Q_S - 2.1681V - 20.1400Q_H - 5.0883\pi \quad (1)$$

$N = 50$, $R_{CV} = 0.9501$, $RMSE_{CV} = 0.3638$

$$\log K_{OW} = -5.0855 - 0.9577Q_{OC} - 2.1634Q_N^2 - 0.3022Q_S + 1.9338V + 5.9907Q_H + 6.4873\pi \quad (2)$$

$N = 50$, $R_{CV} = 0.9720$, $RMSE_{CV} = 0.2006$

$$\log k_W = -0.5670 - 0.9075Q_{OC} - 1.3535Q_N^2 - 0.8891Q_S + 1.734V + 13.1111Q_H - 0.9709\pi \quad (3)$$

$N = 50$, $R_{CV} = 0.9086$, $RMSE_{CV} = 0.3395$

Applying multiple linear regression procedure, the best models containing 6 descriptors were developed with root-mean-square error ($RMSE$) and correlation coefficient of 0.3203 and 0.9613 for log S_W, 0.1661 and 0.9806 for log K_{OW}, and 0.2644 and 0.9457 for log k_W. The partition properties estimated, for 50 compounds in the data set are listed in Table 4. The partition property values are plotted against the corresponding experimental values, given in Fig. 1.

A cross-validation technique was employed to verify the predicted performance of the routine models and predict the partition properties of each compound in data set. Average regression coefficients of 50 sulfur-containing aromatic carboxylates obtained by cross-validation are expressed as above.

Table 3 Statistics of 7-, 6- and 5-parameter models for the partition properties

Model	Dependent variable	R	$RMSE$	R_{CV}	$RMSE_{CV}$	F
7 parameter	log S_W	0.9660	0.3006	0.9549	0.3452	83.7672
	log K_{OW}	0.9832	0.1531	0.9715	0.1988	173.9329
	log k_W	0.9478	0.2594	0.9079	0.3409	52.9710
6 parameter	log S_W	0.9613	0.3203	0.9501	0.3638	87.3136
	log K_{OW}	0.9806	0.1661	0.9720	0.2006	179.0849
	log k_W	0.9457	0.2644	0.9086	0.3395	60.6480
5 parameter	log S_W	0.9536	0.3500	0.9418	0.3910	88.3206
	log K_{OW}	0.9393	0.2876	0.9251	0.3184	65.9745
	log k_W	0.9281	0.3028	0.8867	0.3759	54.6772

Table 4 Observed and calculated values of log S_W, log K_{OW} and log k_W, and quantum descriptors for the sulfur-containing compounds

No.	log S_W			log K_{OW}			log k_W		
	Calc.[a]	Res.[b]	CV[c]	Calc.[a]	Res.[b]	CV[c]	Calc.[a]	Res.[b]	CV[c]
1	−0.8341	0.0741	−0.9255	1.7600	0.1200	1.6119	2.6387	−0.3587	3.0813
2	−1.6265	−0.0735	−1.5545	2.2950	−0.0550	2.3489	2.6526	0.4574	2.2046
3	−3.1727	−0.2073	−3.1405	1.4083	−0.0783	1.4205	0.8610	0.1990	0.8301
4	−3.7675	−0.4925	−3.7233	1.9366	0.1134	1.9264	1.4441	0.3059	1.4167
5	−4.1495	0.3895	−4.1812	2.2499	0.1101	2.2410	1.9043	0.1757	1.8900
6	−4.8920	0.0120	−4.8929	2.9277	−0.0877	2.9345	2.5674	0.6026	2.5207
7	−5.2070	0.1370	−5.2197	3.1997	0.2103	3.1803	2.8729	0.6771	2.8102
8	−3.1871	−0.4829	−3.1511	1.9712	0.3488	1.9452	1.1429	0.4071	1.1126
9	−3.5375	−0.0125	−3.5368	2.2814	0.1686	2.2718	1.5762	0.1138	1.5697
10	−3.8650	−0.1450	−3.8572	2.5735	0.1565	2.5650	1.9252	0.1948	1.9147
11	−4.3263	−0.1537	−4.3160	2.9664	−0.0264	2.9682	2.2923	0.2577	2.2751
12	−3.0957	−0.2143	−3.0772	1.8769	0.1531	1.8637	1.0520	0.1780	1.0366
13	−3.4854	0.4854	−3.5182	2.2081	0.0719	2.2033	1.4879	−0.0879	1.4938
14	−3.7771	0.2371	−3.7907	2.4554	0.1846	2.4448	1.7376	0.1224	1.7306
15	−4.1273	0.0073	−4.1278	2.7386	−0.0586	2.7420	2.0852	0.1048	2.0791
16	−4.4772	−0.1728	−4.4641	3.0706	0.0894	3.0638	2.4919	0.1181	2.4830
17	−4.9449	−0.5951	−4.8798	3.4693	0.0207	3.4670	2.8872	0.1228	2.8737
18	−5.2374	−0.2826	−5.1989	3.7351	0.0949	3.7221	3.1564	0.3936	3.1028
19	−3.5388	0.3088	−3.5582	2.4195	0.1005	2.4132	1.5727	0.0273	1.5710
20	−3.9769	0.6369	−4.0186	2.7832	−0.0032	2.7834	2.0851	−0.2751	2.1031
21	−2.8695	−0.0105	−2.8684	1.8492	−0.0792	1.8572	0.9116	0.1084	0.9008
22	−3.1974	0.1874	−3.2114	2.1419	0.0881	2.1353	1.2356	0.0744	1.2300
23	−3.5868	0.6268	−3.6253	2.4435	−0.1335	2.4517	1.7068	−0.2368	1.7214
24	−4.2411	0.3311	−4.2670	3.0368	−0.1568	3.0491	2.3396	−0.0996	2.3474
25	−4.7157	0.0957	−4.7245	3.4354	−0.2254	3.4562	2.7391	−0.0691	2.7454
26	−3.9981	−0.6119	−3.9654	2.8297	−0.2897	2.8452	2.0554	−0.1554	2.0637
27	−2.5160	0.2560	−2.5472	1.5154	−0.0854	1.5258	0.6352	0.0248	0.6322
28	−2.8420	−0.1580	−2.8285	1.8054	−0.1754	1.8203	1.0594	−0.2094	1.0772
29	−3.1584	0.6084	−3.2024	2.0879	−0.1079	2.0957	1.4054	−0.2754	1.4253
30	−3.6254	−0.2246	−3.6132	2.4782	−0.1782	2.4879	1.7800	−0.2600	1.7941
31	−3.4034	0.4434	−3.5696	1.3367	−0.2767	1.4405	1.1621	−0.5121	1.3541
32	−3.3027	−0.0873	−3.2908	1.4985	−0.1185	1.5146	1.2003	−0.2603	1.2357
33	−3.7075	−0.4725	−3.6340	2.0413	0.1987	2.0103	1.6348	0.0052	1.6340
34	−5.4083	−0.1417	−5.3956	3.5394	−0.1594	3.5536	3.1858	−0.0558	3.1908
35	−6.2715	0.0315	−6.2784	4.6976	−0.2376	4.7500	3.4661	−0.0861	3.4851
36	−5.7150	−0.0450	−5.7069	3.5678	0.2422	3.5237	3.4263	0.0437	3.4184
37	−5.3508	−0.0892	−5.3441	3.4001	−0.0001	3.4001	2.7114	−0.3714	2.7396

Continued

No.	log S_W			log K_{OW}			log k_W		
	Calc.[a]	Res.[b]	CV[c]	Calc.[a]	Res.[b]	CV[c]	Calc.[a]	Res.[b]	CV[c]
38	-4.2862	-0.2738	-4.2655	2.3793	-0.0793	2.3853	1.9958	-0.2858	2.0173
39	-6.4397	0.6097	-6.5781	4.4626	-0.0626	4.4768	3.3135	-0.1335	3.3438
40	-6.0109	0.1609	-6.0485	3.8685	0.1915	3.8237	3.7442	-0.1242	3.7732
41	-4.4435	-0.1665	-4.4240	2.8505	-0.0305	2.8541	2.3790	-0.2690	2.4104
42	-3.0935	0.0535	-3.1379	1.3976	0.0024	1.3955	1.1364	-0.1064	1.2245
43	-3.9582	-0.3318	-3.8624	2.5506	-0.3706	2.6576	2.2639	-0.2939	2.3488
44	-4.9945	0.1845	-5.1955	3.3989	0.1611	3.2233	2.8771	0.1829	2.6778
45	-4.6145	0.0445	-4.6264	2.8834	0.0166	2.8790	1.5149	0.0351	1.5055
46	-4.4056	-0.2144	-4.3139	2.9732	0.2268	2.8762	1.5325	0.4175	1.3541
47	-5.0200	-0.0500	-5.0116	3.5853	0.0347	3.5794	2.3154	-0.0554	2.3248
48	-5.7464	0.4664	-5.8143	3.9730	-0.2930	4.0157	2.8862	-0.4662	2.9541
49	-5.3103	-0.1897	-5.2678	3.9537	-0.0337	3.9612	2.5360	-0.2560	2.5934
50	-5.1615	-0.4885	-5.0782	3.8823	0.2977	3.8315	2.4563	-0.0463	2.4642

[a] Calc.: the results calculated from the Eqs. (1)—(3); [b] CV: the cross-validated results. [c] Res.: differences between the observed and calculated values.

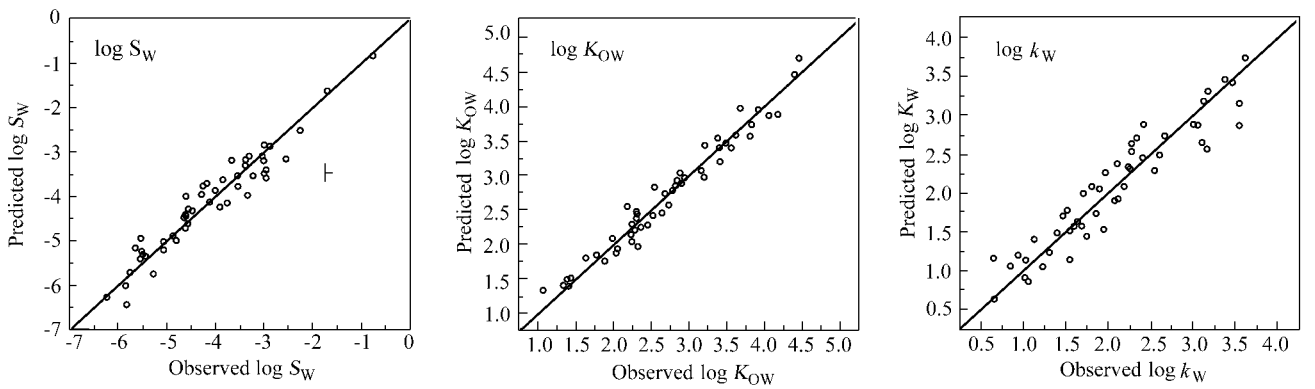

Fig. 1 Plots of observed vs. calculated partition properties for 50 aromatic carboxylates.

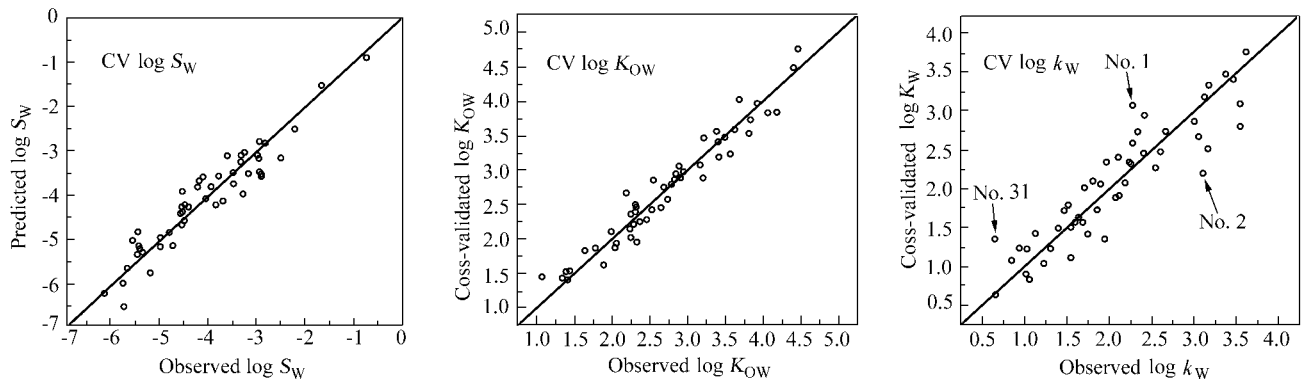

Fig. 2 Plots of observed vs. cross-validated partition properties for 50 aromatic carboxylates.

The predicted partition property values of 50 sulfur-containing aromatic carboxylates are also listed in Table 4. The plots of predicted partition properties by cross-validation versus the observations are given in Fig. 2. It was found that the cross-validated results are acceptable on the whole for the aqueous solubility and octanol-water partition coefficients. The reversed-phase HPLC capacity factors predicted by the cross-validation, however, seem not acceptable because the difference between R_{CV} and R is much high. The CV results in Fig. 2 show that 3 log k_W values (*i.e.* No. 1, No. 2, and No. 31) predicted by Eq. (6) has observable difference from log k_W values observed. This is because only No. 1 and No. 2 have phenylthio- groups while the others all have phenylsul-

fonyl- groups. And No. 31 is the least in size in the 48 phenylsulfonyl compounds. In other words, the partition properties of No. 31 may be boundary values in the full data set. So, the regression equation without these three samples (No. 1, No. 2, and No. 31) as well as statistics were obtained and expressed as follows:

$$\log k_W = -6.788 - 1.0896Q_{OC} - 1.0891Q_N^2 + 1.4461Q_S + 1.8382V + 14.7247Q_H - 2.6518\pi \quad (4)$$

$N = 47$, $R = 0.9595$, $RMSE = 0.225$, $R_{CV} = 0.9440$, $RMSE_{CV} = 0.263$, $F = 77.391$

The reversed-phase HPLC capacity factors predicted by Eq. (4) are better than those predicted by Eq. 3. Especially the cross-validated results and F statistics are obviously improved.

Variable correlation problems

According to the principle of statistics, a regression equation is of no relevance when the explanatory variables applied were mutually interrelated by simple or multiple correlations. Here, the bivariate correlation of two sets of variables, including the dependent variables (Table 5) was investigated. As it was shown by the correlation coefficients of the independent variables given in Table 5, the six-parameter models for modeling partition properties have cleared up the possibility.

Table 5 Correlation coefficients of variables selected for modeling

	Q_{OC}	Q_N^2	Q_S	V	Q_H	π
Q_{OC}	1.0000					
Q_N^2	0.2028	1.0000				
Q_S	0.0303	−0.2449	1.0000			
V	0.2089	0.3246	0.3123	1.0000		
Q_H	0.3775	0.7325	−0.2297	0.1659	1.0000	
π	−0.0637	0.3618	−0.3718	−0.1213	0.1553	1.0000

Mechanism analysis of partition action

According to this group of the statistically best equations [Eqs. (1)—(3)] with 6 descriptors and all the other results obtained, the partition mechanism can be explained. Here the Q_{OC} term is the sum of atomic charges of carbon and oxygen atoms in the carboxyl group. This case demonstrated that the carboxyl group played a dominant role in the partition mechanisms and may imply the polar interaction of hydrogen-bond interaction between the carboxyl group and the strong polar molecules, *i.e.* water, methanol molecules. Such hydrogen-bond interaction resulted in phenylsulfonyl acetate molecules existing a tendency to partition into the water and mobile phase. Q_N^2 is the 2nd power of net charges on nitrogen atoms. Similarly, the positive or negative effects of the Q_N^2 term show that log S_W increases, and log K_{OW} and log k_W decrease with increasing electrostatic interactions among the solvent and nitro-group in solute molecules, such as the hydrogen-bonding formation. This directly resulted in phenylsulfonyl acetate molecules to partition into the water and mobile phase. The V term denotes the van der Waals volume of the solute molecule. The negative or positive signs with V agree with theoretical expectation: log S_W decreases, log K_{OW} and log k_W increase with increasing cavity formation energy in water, or in water-methanol mixture phase, or increasing preference for solute-solvent dispersion interactions, resulting in solute molecules to tend to partition in to the weak polar phase, such as octanol phase or the immobile phase. The Q_H term denotes the most positive net atomic charge on hydrogen atoms that renders the ability to donor proton of the solute molecules. Because the ability to donor proton of the water molecules is much more than that of the solute molecules, it took a negative action in water solution process, and positive actions in octanol-water partition and chromatographic process tended to partition into the octanol phase and immobile phase to form weak hydrogen-bond interaction with the octanol molecules and oxygen atoms at the Si—O—Si skeleton in the immobile phase. The log S_W, log K_{OW}, and log k_W have negative, positive, and negative signs (−, +, and −) for the π term, respectively, which is a measure of polarity and polarizability. It is expected to be involved since it is in direct proportion to intrinsic molecular volume, and molecular volume is a measure of the energy to form a cavity in the solvent. The positive signs indicate that larger molecules tend to partition into the less polar phase, *i.e.* the octanol phase or the immobile phase. The particularly interesting thing is that the log S_W, log K_{OW}, and log k_W all have negative signs (−) for the Q_S term and this seems unreasonable. In Table 2, there are two compounds containing the thio (—S—) groups [2-NO_2PhSCH$_2$-COOCH$_3$ (**1**), 4-Cl-2-NO_2 PhSCH$_2$ COOCH$_3$ (**2**)] in the full data set, which are not homogeneous compounds, while the other 48 compounds contain sulfonyl (—SO_2—) groups. Their occurrence directly affected the interpretation of the Q_S term based on the MLR equations. When these two compounds are omitted from the whole data set, the rebuilt models are obtained and the constant coefficients for the new models are given in Table 6. The signs of the constant coefficients are the same as those in Eqs. (1)—(3) except for the constant coefficients for the Q_S term with negative for log S_W, both positive for log K_{OW} and log k_W. This result may indicate that there is a weak repulsive interaction between the sulfur atom in phenylsulfonyl acetate molecules and the oxygen atoms in the polar molecules, *i.e.* the water molecules or methanol molecules.

Table 6 Constant coefficients for QSPR modeling without samples **1** and **2**

Descriptor	*b*		
	log S_W	log K_{OW}	log k_W
b_0	7.667	−11.513	−6.557
Q_{OC}	1.458	−1.099	−1.035
Q_N^2	1.290	−1.980	−1.032
Q_S	−0.422	2.037	1.542
V	−2.156	2.078	1.892
Q_H	−19.916	5.636	10.855
π	−5.091	5.401	−2.671

Conclusion

The multi-objective quantitative structure-property relationships were successfully developed based on three partition properties in this study. And main factors affecting the partition process for these sulfur-containing aromatic compounds were discussed. Although some of these quantitative models built may not be more predictive than those theoretical models built based on individual partition properties as objectives, they may be more favorable and reasonable for the mechanism discussion on multi-properties and the selected (independent) descriptors were also more interpretative and informative.

Acknowledgement

We wish to thank Prof. Juguang Han for the valuable help that has led to an important improvement for the material representation.

References

1 Han, S. K.; Jiang, L. Q.; Wang, L. S.; Zhang, Z. *Chemosphere* **1992**, *25*, 643.
2 Pandeya, S. N.; Ojha, T. N.; Srivastava, V. *J. Sci. Ind. Res.* **1985**, *44*, 150.
3 Brasquet, C.; Bourges, B.; Cloirec, P. L. *Environ. Sci. Technol.* **1999**, *33*, 4226.
4 Nevalainen, T.; Kolehaunen, H. *Environ. Toxicol. Chem.* **1994**, *13*, 1699.
5 Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. *Environ. Toxicol. Chem.* **1994**, *13*, 841.
6 Baj, S.; Dawid, M. *J. Liq. Chromatogr.* **1994**, *17*, 3933.
7 Nendza, M.; Muller, M. *Quant. Struct.-Act. Relat.* **2001**, *19*, 581.
8 Dai, J. Y.; Jin, L. J.; Yao, S. C.; Wang, L. S. *Chemosphere* **2001**, *42*, 899.
9 Basak, S. C.; Grunwald, G. D.; Gute, B. D.; Balasubramanian, K.; Opitz, D. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 885.
10 Klopman, G.; Saiakhov, R.; Rosenkranz, H. S. *Environ. Toxicol. Chem.* **2000**, *19*, 441.
11 Basak, S. C.; Gute, B. D.; Grunwald, G. D. *SAR QSAR Environ. Res.* **1999**, *10*, 117.
12 Gusten, H. *Chemosphere* **1999**, *38*, 1361.
13 Huang, X. D.; Krylov, S. N.; Ren, L. S.; McConkey, B. J; Dixon, D. G.; Greenberg, B. M. *Environ. Toxicol. Chem.* **1997**, *16*, 2296
14 Ramos, E. U.; Vaes, W. H. J.; Verhaar, H. J. M.; Hermens, J. L. M. *Environ. Sci. Pollut. Res. Int.* **1997**, *4*, 83.
15 Mekenyan, O. G.; Veith, G. D.; Call, D. J.; Ankley, G. T. *Environ. Health Perspect.* **1996**, *104*, 1302.
16 Devillers, J. *Genetic Algorithms in Molecular Modeling*, Academic Press Ltd., London, **1996**, Chapter 2.
17 Arx, K. B. V.; Manock, J. J.; Huffman, S. W.; Messina, M. *Environ. Sci. Technol.* **1998**, *32*, 3207.
18 Gramatica, P.; Consonni, V.; Todeschini, R. *Chemosphere* **1999**, *38*, 1371.
19 OECD Guideline for Testing of Chemicals, Paris, **1981**.
20 CambridgeSoft Corp, **1999**.
21 Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548.
22 Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118.
23 Hong, H.; Han, S. K.; Wang, X. R.; Wang, L. S.; Zhang, Z.; Zou, G. W. *Environ. Sci. Technol.* **1995**, *29*, 3044.
24 Feng, L.; Han, S. K.; Wang, L. S.; Wang, Z. T.; Zhang, Z. *Chemosphere* **1996**, *32*, 353.
25 He, Y. B.; Wang, L. S.; Han, S. K.; Zhao, Y. H.; Zhang, Z.; Zou, G. W. *Chemosphere* **1995**, *30*, 117.

(E0201285 ZHAO, X. J.)