ЖУРНАЛ СТРУКТУРНОЙ ХИМИИ

2016. Том 57, № 8 *Ноябрь – декабрь* *С. 1625 – 1632*

UDC 541.6:548.737

STRUCTURAL CHARACTERIZATION AND PREDICTION OF KOVATS RETENTION INDICES (RI) FOR ALKYLBENZENE COMPOUNDS

L.-M. Liao[1,2], J.-F. Li[1,2], G.-D. Lei[1]

[1]*College of Chemistry and Chemical Engineering, Neijiang Normal University, Neijiang, Sichuan, P. R. China*
E-mail: leigdnjtc@126.com
[2]*College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, P. R. China*

Received June, 10, 2015 *Revised — July, 09, 2015*

A new molecular structural characterization (MSC) method called the molecular vertex eigenvalue correlative index (MVECI) is constructed and used to describe the structures of 122 alkylbenzene compounds. Through multiple linear regression (MLR) and stepwise multiple regression (SMR), a quantitative structure-retention relationship (QSRR) model with correlation coefficient (R) of 0.995 is obtained. Through partial least-square regression (PLS), another QSRR model with correlation coefficient (R) of 0.991 is obtained. The estimation stability and prediction ability of the two models are strictly analyzed by both internal and external validations. For the internal validation, the cross-validation (CV) correlation coefficients (R_{CV}) of the two models are 0.993 and 0.988. For the external validation, the correlation coefficients (R_{test}) of the two models are 0.996 and 0.995, respectively. The results show that the stability and predictability of the models are good, and the molecular vertex eigenvalue correlative index can successfully describe the structures of alkylbenzene compounds.

DOI: 10.15372/JSC20160806

K e y w o r d s: alkylbenzene, retention index, structural descriptors, QSRR.

INTRODUCTION

Alkylbenzene compounds are widely applied in the chemical industry. Alkylbenzene compounds have a certain degree of toxicity and the number of them is huge. The compounds enter into the environment to become one of important pollutants and seriously endanger people′s health. Therefore, researches on their properties are essential. Usually, alkylbenzene compounds are determined by chromatography. The main qualitative index in the chromatographic analysis is the retention value. However, some data cannot be obtained through experiments, and the prediction of some retention data for organic compounds has become a simple and effective way of the qualitative analysis [1]. In addition, there is a certain correlation between the GC retention value and the octanol/water partition coefficient (lgKow) of compounds [2]. The octanol/water partition coefficient (lgKow) of compounds is a vital index in environmental chemistry. The prediction of retention values can also provide useful references for the researches of octanol/water partition coefficients (lgKow) of organic compounds. The structures of compounds firstly need to be parameterized to construct the corresponding structural descriptors for building up a quantitative structure — retention relationship model. In recent years, many meaningful works have been carried out by computational chemists in the study of QSRR or QSAR [3—8]. In this paper, a new molecular structural characterization (MSC) method called the molecular vertex eigenvalue correlative index (MVECI) is constructed based on preliminary studies and referring to the relevant literatures. The descriptors are employed to characterize 122 alkylbenzene

compounds, and the multiple linear regression and partial least squares regression methods are used to construct the quantitative structure — retention relationship models. It is concluded that there is good correlation between the molecular vertex eigenvalue correlative index (MVECI) and the chromatographic retention indices of the compounds, and the stability and predictability of the models are good.

DATA SETS

The retention indices (RI) of 122 alkylbenzenes are taken from the literature [9]. The compounds are listed in Table 1 based on the order of the retention index (from small to large). 24 samples are selected out as a test set (marked with "*" in the Table) for evaluating the external predictive ability of the models. The remaining 98 samples are used as a training set to build up the models.

PRINCIPLES AND METHDOS

Structural characterization. The properties of an organic compound directly relate to the molecular structure, such as the molecular flexibility, hydrogen bonding between the molecules, molecular polarity, molecular size, etc. For the molecular matrix skeleton, each non-hydrogen atom is a molecular vertex. Vertices are the basic units of the whole molecule, therefore, the external properties of a compound can be reflected from the level of its vertices. Firstly, based on the previous studies [10—14] in the classification, without regarding to hydrogen atoms, these vertex atoms are divided into four atomic types according to the number of other vertex atoms linked to them through chemical bond(s). If a vertex is linked to k (k = 1, 2, 3, 4) other vertices through chemical bonds, the atomic type belongs to the kth one. For example, a carbon atom connected with two other vertex atoms belongs to the second atomic type. Then, the eigenvalue of the vertex should be defined for the correlative index. According to the atomic structure and surrounding environment, referring to a large number of publications [15—17], Eq. (1) is adopted to calculate the eigenvalue (Z_i) of vertices:

$$Z_i = \left[m_i (n_i - 1) \left(\frac{X_C}{X_i} \right)^{1/2} - h_i \right]^{1/2}, \quad (1)$$

where n_i represents the number of atomic electronic shells of the i vertex; m_i represents the number of valence shell electrons of the i vertex, and h_i represents the number of hydrogen atoms linked to the i vertex; X_i is Pauling′s electronegativity of the i vertex, and X_C is Pauling′s electronegativity of the carbon atom. For example, Z_i of the oxygen atom in alcohol can be calculated as follows:

$$Z_i = \left[6(2-1) \left(\frac{2.55}{3.44} \right)^{1/2} - 1 \right]^{1/2} = 2.0410. \quad (2)$$

The properties of an organic compound mainly depend on different correlativeness between the vertices in the molecule. The main idea of this study is focused on an indirect reflection of the overall molecular structures via expressing correlativeness among different eigenvalues of the vertices. This correlativeness reflects two changing trends rather than a specific interaction manner. One varies in a positive trend with a change in the eigenvalue and the other in a negative trend with a change in interatomic distances. Countdown-type function Eq. (3) can meet the requirements.

$$x_r = m_{nl} = \sum_{i=n,\, j=l} \frac{Z_i \cdot Z_j}{r_{ij}^2}, \quad (3)$$

where n or l represents the atomic type of the i or j vertex; i or j is a code of a vertex in the molecular skeleton graph; in the MEDV [13, 14], Z_i and Z_j are the relative electronegativity of the i vertex and the j vertex to the C atom, but in this paper, they are the eigenvalues of i and j vertices, and they are calculated according to Eq. (1); r_{ij} represents the relative distance between the ith and jth atoms (*viz.* the sum of the experienced shortest path relative to the C—C single bond length). According to Eq. (3), the correlativeness in a molecule could be assembled as m_{11}, m_{12}, m_{13}, m_{14}, m_{22}, m_{23}, m_{24}, m_{33}, m_{34}, and m_{44} shortened as x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, and x_{10}, and they are all called the molecular

Table 1

Values of experimental retention indices, RI (Exp.) (*squalane* 100 °C)

No.	Compounds	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	RI(Exp.)	MLR	PLS
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Benzene	0.0000	0.0000	0.0000	0.0000	28.8518	0.0000	0.0000	0.0000	0.0000	0.0000	650.5	659.4	674.7
2	Methylbenzene	0.0000	2.6385	4.0000	0.0000	19.2346	12.8230	0.0000	0.0000	0.0000	0.0000	760.1	796.8	793.9
3	Ethylbenzene	0.0000	3.2620	1.0000	0.0000	21.8731	16.8230	0.0000	0.0000	0.0000	0.0000	850.0	852.2	862.2
4	1,4-Dimethylbenzene	0.0451	4.8405	8.5820	0.0000	10.0265	23.0201	0.0000	0.5457	0.0000	0.0000	864.6	893.1	888.5
5*	1,2-Dimethylbenzene	0.1187	3.6194	10.2102	0.0000	13.3005	15.8242	0.0000	4.9110	0.0000	0.0000	886.0	898.2	889.8
6	*iso*-Propylbenzene	0.2500	2.5240	8.0000	0.0000	19.2346	15.4616	0.0000	4.0000	0.0000	0.0000	908.4	907.8	890.5
7	*n*-Propylbenzene	0.0000	3.2459	0.4444	0.0000	25.1351	17.8230	0.0000	0.0000	0.0000	0.0000	938.0	917.8	942.7
8	1-Methyl-3-ethylbenzene	0.0433	5.3811	5.7847	0.0000	12.7953	26.4713	0.0000	1.2277	0.0000	0.0000	949.7	945.5	948.5
9	1-Methyl-2-ethylbenzene	0.0657	4.8343	6.5799	0.0000	15.1102	20.9293	0.0000	4.9110	0.0000	0.0000	966.0	957.8	957.6
10*	1,3,5-Trimethylbenzene	0.2072	5.6278	15.0503	0.0000	2.7624	28.5108	0.0000	3.6832	0.0000	0.0000	969.0	951.6	956.7
11	*tert*-Butylbenzene	0.7500	3.7860	3.0000	12.0000	19.2346	12.8230	2.6385	0.0000	4.0000	0.0000	971.7	965.6	956.3
12	1,2,4-Trimethylbenzene	0.2329	5.0588	15.8090	0.0000	5.0133	23.8727	0.0000	6.6844	0.0000	0.0000	987.0	989.2	978.1
13	*iso*-Butylbenzene	0.2500	2.4919	6.8889	0.0000	21.8731	20.0850	0.0000	1.0000	0.0000	0.0000	991.3	999.8	952.5
14	1-Methyl-3-*iso*-propylbenzene	0.3366	4.3668	13.1300	0.0000	10.5381	24.2201	0.0000	5.7361	0.0000	0.0000	1009.9	1004.0	983.8
15*	1-Methyl-4-*iso*-propylbenzene	0.3114	4.6735	12.6971	0.0000	10.0265	25.4403	0.0000	4.8367	0.0000	0.0000	1011.8	1001.6	1019.2
16	1,2,3-Trimethylbenzene	0.3065	3.8376	17.4373	0.0000	8.2872	16.3698	0.0000	11.0496	0.0000	0.0000	1012.9	1017.9	1038.0
17	1,3-Diethylbenzene	0.0297	6.2395	2.5526	0.0000	15.1907	30.9796	0.0000	1.2277	0.0000	0.0000	1027.4	1012.5	1036.5
18	1-Methyl-3-*n*-propylbenzene	0.0297	5.4856	5.1261	0.0000	15.8501	27.7475	0.0000	1.2277	0.0000	0.0000	1034.6	1009.1	1036.1
19	*n*-Butylbenzene	0.0000	3.2166	0.2500	0.0000	28.3810	18.2675	0.0000	0.0000	0.0000	0.0000	1036.8	1011.5	981.4
20*	1,2-Diethylbenzene	0.0416	6.0088	2.9496	0.0000	17.1572	26.0344	0.0000	4.9110	0.0000	0.0000	1039.9	1015.6	1033.9
21	1,4-Diethylbenzene	0.0222	6.3454	2.3610	0.0000	14.9573	31.6021	0.0000	0.5457	0.0000	0.0000	1040.5	1014.8	1028.5
22	1-Methyl-2-*n*-propylbenzene	0.0416	5.0430	5.8122	0.0000	18.0161	22.4041	0.0000	4.9110	0.0000	0.0000	1045.6	1054.1	1043.9
23	1,3-Dimethyl-5-ethylbenzene	0.1557	6.7376	11.5861	0.0000	4.6383	33.5276	0.0000	3.6832	0.0000	0.0000	1046.1	1037.8	1035.6
24	1-Methyl-3-*tert*-butylbenzene	0.8799	5.4216	8.3372	12.0691	10.5381	21.9629	2.2572	1.2277	4.5084	0.0000	1057.9	1014.9	1026.4
25*	1,3-Dimethyl-4-ethylbenzene	0.1654	6.5204	12.0682	0.0000	6.6047	28.9619	0.0000	6.6844	0.0000	0.0000	1066.3	1047.1	1043.1
26	*tert*-Pentylbenzene	0.4722	5.7699	2.4444	9.0000	20.4966	13.8230	6.6385	0.0000	4.0000	0.0000	1071.2	1009.4	1027.3
27	1,2-Dimethyl-4-ethylbenzene	0.1927	6.0534	12.4664	0.0000	7.0522	28.6721	0.0000	6.6844	0.0000	0.0000	1072.7	1063.0	1042.2
28	1,3-Dimethyl-2-ethylbenzene	0.2004	5.6440	13.1766	0.0000	9.2681	22.5801	0.0000	11.0496	0.0000	0.0000	1073.5	1052.9	1045.4
29	*sec*-Pentylbenzene	0.0625	4.6175	4.5833	0.0000	23.2425	20.6560	0.0000	4.0000	0.0000	0.0000	1078.2	1086.6	1042.3
30*	1-Ethyl-3-*iso*-propylbenzene	0.3094	5.3376	9.8722	0.0000	12.7953	28.8666	0.0000	5.7361	0.0000	0.0000	1078.2	1065.9	1049.1
31	1-Ethyl-2-*iso*-propylbenzene	0.3332	4.9804	10.4901	0.0000	15.1102	22.9764	0.0000	10.0161	0.0000	0.0000	1081.5	1056.2	1049.0
32	1,2-Dimethyl-3-ethylbenzene	0.2277	5.2957	13.5748	0.0000	9.7156	21.9833	0.0000	11.0496	0.0000	0.0000	1087.7	997.2	1170.4
33	1-Methyl-4-*sec*-butylbenzene	0.1640	6.4814	9.8339	0.0000	11.1532	29.6208	0.0000	4.8367	0.0000	0.0000	1092.7	1066.7	1051.5
34	1-Ethyl-4-*iso*-propylbenzene	0.2945	5.5027	9.5722	0.0000	12.4468	29.8216	0.0000	4.8367	0.0000	0.0000	1099.0	1071.7	1044.8
35*	*iso*-Pentylbenzene	0.2500	2.4332	6.5000	0.0000	25.1351	21.0690	0.0000	0.4444	0.0000	0.0000	1100.1	1084.5	1043.4
36	1,2,4,5-Tetramethylbenzene	0.4658	4.8405	23.6181	0.0000	0.4092	23.0201	0.0000	13.3687	0.0000	0.0000	1107.1	1104.4	1106.5
37	1,2,3,5-Tetramethylbenzene	0.4897	4.5145	24.0528	0.0000	0.9208	20.5987	0.0000	14.0508	0.0000	0.0000	1113.1	1073.8	1109.0
38	1,3-*diiso*-Propylbenzene	0.6187	4.2192	17.4516	0.0000	10.5381	26.4773	0.0000	10.4517	0.0000	0.0000	1120.5	1072.6	1106.3
39	1-Ethyl-2-*n*-propylbenzene	0.0287	6.1453	2.1819	0.0000	20.1945	27.5092	0.0000	4.9110	0.0000	0.0000	1120.8	1073.5	1046.7
40*	1,2-*diiso*-Propylbenzene	0.6664	3.6236	18.4245	0.0000	13.3005	19.4435	0.0000	15.4773	0.0000	0.0000	1121.9	1043.5	1008.0
41	1,3-Dimethyl-5-propylbenzene	0.1284	6.9627	10.8245	0.0000	7.4859	35.0801	0.0000	3.6832	0.0000	0.0000	1130.9	1095.0	1112.7
42	*n*-Pentylbenzene	0.0000	3.1995	0.1600	0.0000	31.5976	18.5175	0.0000	0.0000	0.0000	0.0000	1133.0	1069.5	1107.9
43	1-Methyl-3-*n*-butylbenzene	0.0216	5.5269	4.8771	0.0000	18.9661	28.3652	0.0000	1.2277	0.0000	0.0000	1134.1	1109.2	1125.1
44	1,2,3,4-Tetramethylbenzene	0.5394	3.6194	25.2463	0.0000	3.6832	15.8242	0.0000	17.7340	0.0000	0.0000	1138.6	1095.4	1067.7
45*	1,4-Dimethyl-2-propylbenzene	0.1164	6.8462	11.0759	0.0000	9.1403	31.2373	0.0000	6.6844	0.0000	0.0000	1145.2	1111.1	1115.6
46	1,3-Dimethyl-5-*tert*-butylbenzene	1.0907	5.7781	15.3800	12.4568	5.0133	23.5657	2.0390	6.6844	4.7994	0.0000	1150.8	1068.7	1049.5
47	1,4-*diiso*-Propylbenzene	0.5889	4.5065	16.9676	0.0000	10.0265	27.8606	0.0000	9.2630	0.0000	0.0000	1154.1	1128.9	1189.1
48	1,2-Dimethyl-4-propylbenzene	0.1706	6.2319	11.7501	0.0000	9.9717	30.1288	0.0000	6.6844	0.0000	0.0000	1154.8	1110.9	1113.2
49	1-Ethyl-4-*tert*-butylbenzene	0.8167	6.6907	4.7220	12.0307	12.4468	27.3111	2.5105	0.5457	4.2910	0.0000	1162.5	1126.1	1185.7
50*	1-Isopropyl-4-n-propylbenzene	0.2837	5.5130	8.9949	0.0000	15.5734	31.1393	0.0000	4.8367	0.0000	0.0000	1181.9	1105.3	1114.3
51	1,2-Dipropylbenzene	0.0210	6.3088	1.4142	0.0000	23.3149	28.9840	0.0000	4.9110	0.0000	0.0000	1183.2	1152.3	1131.6
52	1,3-Dipropylbenzene	0.0164	6.3858	1.2354	0.0000	21.5329	33.5322	0.0000	1.2277	0.0000	0.0000	1195.5	1128.2	1122.3
53	1,4-Dipropylbenzene	0.0132	6.4249	1.1345	0.0000	21.3778	33.9631	0.0000	0.5457	0.0000	0.0000	1212.5	1148.5	1131.0
54	1-Ethyl-4-*n*-butylbenzene	0.0132	6.3800	1.5194	0.0000	21.3436	33.3498	0.0000	0.5457	0.0000	0.0000	1227.1	1110.1	1069.7
55*	*n*-Hexylbenzene	0.0000	3.1912	0.1111	0.0000	34.7971	18.6775	0.0000	0.0000	0.0000	0.0000	1231.0	1140.9	1130.3
56	1-Methyl-4-*n*-pentylbenzene	0.0132	5.6842	4.5183	0.0000	21.8772	29.3977	0.0000	0.5457	0.0000	0.0000	1236.3	1166.4	1135.0
57	1,4-Diterc-butylbenzene	1.7000	6.7598	7.0830	24.4605	10.0265	23.0201	4.8405	0.5457	8.5820	0.1805	1281.3	1192.3	1129.3
58	1-Isobutyl-4-*tert*-butylbenzene	1.1010	5.7797	10.7427	12.1778	12.4468	30.4377	2.6008	1.7262	4.4138	0.0000	1290.6	1126.2	1188.1
59	1-*sec*-Butyl-4-*tert*-butylbenzene	0.9783	7.4770	9.1313	12.2117	11.1532	29.6208	2.5431	4.8367	4.4715	0.0000	1290.7	1165.1	1137.8

Continued Table 1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
60*	1,4-*disec*-Butylbenzene	0.2913	8.4281	11.2072	0.0000	12.3242	36.3751	0.0000	9.2630	0.0000	0.0000	1304.3	1173.2	1130.2
61	1,4-*diiso*-Butylbenzene	0.5528	4.7487	14.4036	0.0000	14.9573	38.0088	0.0000	2.9733	0.0000	0.0000	1307.7	1146.3	1119.2
62	1-Propyl-4-butylbenzene	0.0106	6.4161	0.9061	0.0000	24.6098	34.5303	0.0000	0.5457	0.0000	0.0000	1311.0	1171.3	1195.6
63	1-Ethyl-4-*n*-penthylbenzene	0.0106	6.3850	1.4078	0.0000	24.5272	33.6887	0.0000	0.5457	0.0000	0.0000	1322.1	1206.1	1202.9
64	n-Heptylbenzene	0.0000	3.1881	0.0816	0.0000	37.9883	18.7886	0.0000	0.0000	0.0000	0.0000	1328.1	1207.2	1204.8
65*	1-*n*-Butyl-4-*tert*-butylbenzene	0.7896	6.9484	3.8804	12.0168	18.7273	29.0588	2.6164	0.5457	4.2910	0.0000	1345.3	1206.1	1201.1
66	1-*n*-Butyl-4-*sec*-butyilbenzene	0.1349	7.4376	5.8535	0.0000	19.9934	35.8557	0.0000	4.8367	0.0000	0.0000	1358.4	1201.2	1203.1
67	1-*n*-Butyl-4-isobutylbenzene	0.2712	5.5958	7.5129	0.0000	21.3436	36.6857	0.0000	1.7262	0.0000	0.0000	1359.6	1199.5	1202.2
68	1,2-Dibutylbenzene	0.0126	6.3807	0.8329	0.0000	29.7360	30.3982	0.0000	4.9110	0.0000	0.0000	1371.9	1217.2	1208.7
69	1-Ethyl-2-hexylbenzene	0.0126	6.2391	1.6699	0.0000	29.6312	28.9076	0.0000	4.9110	0.0000	0.0000	1400.5	1232.2	1217.1
70*	1,4-Dibutylbenzene	0.0087	6.4217	0.6778	0.0000	27.8681	35.0976	0.0000	0.5457	0.0000	0.0000	1410.8	1301.7	1291.5
71	1-Propyl-4-pentylbenzene	0.0087	6.4133	0.7946	0.0000	27.8197	34.8692	0.0000	0.5457	0.0000	0.0000	1418.9	1263.6	1283.5
72	1-Ethyl-4-*n*-hexylbenzene	0.0087	6.3931	1.3444	0.0000	27.7026	33.9161	0.0000	0.5457	0.0000	0.0000	1421.1	1307.4	1296.4
73	1-Methyl-4-*n*-heptylbenzene	0.0087	5.7152	4.4151	0.0000	28.1778	29.7889	0.0000	0.5457	0.0000	0.0000	1433.9	1287.3	1276.1
74	1-Butyl-2-pentylbenzene	0.0102	6.3827	0.6912	0.0000	32.9592	30.8146	0.0000	4.9110	0.0000	0.0000	1465.5	1299.0	1353.8
75*	1-Propyl-2-hexylbenzene	0.0102	6.3544	0.9022	0.0000	32.8830	30.3824	0.0000	4.9110	0.0000	0.0000	1471.9	1320.8	1364.4
76	1-Butylhexylbenzene	0.0123	6.4172	0.5786	0.0000	38.0251	27.3646	0.0000	4.0000	0.0000	0.0000	1491.4	1337.5	1383.1
77	1-Ethyl-2-heptylbenzene	0.0102	6.2568	1.6205	0.0000	32.7915	29.1026	0.0000	4.9110	0.0000	0.0000	1502.5	1295.9	1291.4
78	1-Ethyloctylbenzene	0.0123	6.2017	1.3182	0.0000	37.8407	25.9829	0.0000	4.0000	0.0000	0.0000	1517.2	1291.4	1289.5
79	*n*-Nonylbenzene	0.0000	3.1893	0.0494	0.0000	44.3643	18.9327	0.0000	0.0000	0.0000	0.0000	1534.3	1305.5	1294.8
80*	1-Methylnonylbenzene	0.0123	4.9475	4.0963	0.0000	38.7022	22.1066	0.0000	4.0000	0.0000	0.0000	1550.7	1322.0	1303.5
81	1-Butyl-2-hexylbenzene	0.0084	6.3862	0.6115	0.0000	36.1719	31.0895	0.0000	4.9110	0.0000	0.0000	1563.2	1339.4	1302.2
82	1-Propyl-2-heptylbenzene	0.0084	6.3648	0.8528	0.0000	36.0686	30.5774	0.0000	4.9110	0.0000	0.0000	1572.7	1348.8	1364.5
83	1-Pentylhexylbenzene	0.0100	6.4509	0.4622	0.0000	41.3072	27.7121	0.0000	4.0000	0.0000	0.0000	1581.3	1368.9	1381.8
84	1-Butylheptylbenzene	0.0100	6.4199	0.5125	0.0000	41.2639	27.5957	0.0000	4.0000	0.0000	0.0000	1587.0	1397.2	1382.1
85*	1-Ethyl-2-octylbenzene	0.0084	6.2721	1.5878	0.0000	35.9567	29.2483	0.0000	4.9110	0.0000	0.0000	1604.4	1394.9	1379.6
86	1-Ethylnonylbenzene	0.0100	6.2185	1.2907	0.0000	41.0190	26.1066	0.0000	4.0000	0.0000	0.0000	1615.4	1389.7	1375.6
87	*n*-Decylbenzene	0.0000	3.1975	0.0278	0.0000	53.9397	19.0552	0.0000	0.0000	0.0000	0.0000	1637.3	1390.8	1380.9
88	1-Methyldecylbenzene	0.0100	4.9687	4.0770	0.0000	41.8602	22.2029	0.0000	4.0000	0.0000	0.0000	1650.1	1388.2	1379.1
89	1-Butyloctylbenzene	0.0083	6.4237	0.4712	0.0000	44.4932	27.7607	0.0000	4.0000	0.0000	0.0000	1679.8	1381.6	1376.4
90*	1-Propylnonylbenzene	0.0083	6.3858	0.6796	0.0000	44.3445	27.3011	0.0000	4.0000	0.0000	0.0000	1691.3	1393.3	1380.8
91	1-Ethyldecylbenzene	0.0083	6.2327	1.2715	0.0000	44.2018	26.2029	0.0000	4.0000	0.0000	0.0000	1710.4	1410.9	1389.7
92	*n*-Undecylbenzene	0.0000	3.1944	0.0331	0.0000	50.7453	19.0221	0.0000	0.0000	0.0000	0.0000	1725.7	1490.9	1470.7
93	1-Pentyloctylbenzene	0.0069	6.4452	0.3548	0.0000	47.8200	28.1082	0.0000	4.0000	0.0000	0.0000	1766.8	1486.3	1467.2
94	1-Butylnonylbenzene	0.0069	6.4282	0.4438	0.0000	47.7162	27.8844	0.0000	4.0000	0.0000	0.0000	1773.5	1529.8	1551.6
95*	1-Propyldecylbenzene	0.0069	6.3948	0.6604	0.0000	47.5473	27.3973	0.0000	4.0000	0.0000	0.0000	1785.0	1525.6	1547.0
96	1-Ethylundecylbenzene	0.0069	6.2449	1.2575	0.0000	47.3887	26.2799	0.0000	4.0000	0.0000	0.0000	1807.0	1479.5	1462.3
97	1-Methyldodecylbenzene	0.0069	5.0010	4.0526	0.0000	48.2017	22.3430	0.0000	4.0000	0.0000	0.0000	1843.3	1520.4	1539.6
98	*n*-Tridecylbenzene	0.0000	3.2005	0.0237	0.0000	57.1372	19.0829	0.0000	0.0000	0.0000	0.0000	1923.6	1480.8	1466.7
99	1,3-Dimethylbenzene	0.0691	4.5145	9.0168	0.0000	10.5381	21.9629	0.0000	1.2277	0.0000	0.0000	864.8	1563.8	1540.2
100*	1-Methyl-4-ethylbenzene	0.0307	5.5920	5.4715	0.0000	12.4468	27.3111	0.0000	0.5457	0.0000	0.0000	954.3	1584.7	1560.6
101	*sec*-Butylbenzene	0.1111	4.2579	5.1944	0.0000	20.4966	19.4616	0.0000	4.0000	0.0000	0.0000	990.2	1582.8	1558.6
102	1-Methyl-2-isopropylbenzene	0.3813	3.6215	14.1734	0.0000	13.3005	17.6339	0.0000	10.0161	0.0000	0.0000	1017.6	1576.4	1554.3
103	1-Methyl-4-*n*-propylbenzene	0.0222	5.6499	4.8582	0.0000	15.5734	28.4916	0.0000	0.5457	0.0000	0.0000	1040.2	1623.8	1642.5
104	1,4-Dimethyl-2-ethylbenzene	0.1541	6.5169	11.9466	0.0000	6.4416	29.4862	0.0000	6.6844	0.0000	0.0000	1062.2	1622.1	1640.0
105*	1-Methyl-4-*tert*-butylbenzene	0.8421	5.8002	7.8325	12.0451	10.0265	23.0201	2.4203	0.5457	4.2910	0.0000	1076.4	1616.1	1634.5
106	1-Methyl-2-*tert*-butylbenzene	0.9470	4.5274	9.5295	12.1187	13.3005	13.8462	1.8097	4.9110	5.1051	0.0000	1092.6	1568.4	1548.8
107	1-Ethyl-3-*n*-propylbenzene	0.0216	6.3031	1.8940	0.0000	18.3321	32.2559	0.0000	1.2277	0.0000	0.0000	1112.7	1609.6	1626.3
108	1-Ethyl-4-*n*-propylbenzene	0.0168	6.3779	1.7477	0.0000	18.1453	32.7826	0.0000	0.5457	0.0000	0.0000	1126.4	1742.7	1724.5
109	1-Methyl-4-*n*-butylbenzene	0.0168	5.6682	4.6299	0.0000	18.7273	29.0588	0.0000	0.5457	0.0000	0.0000	1138.8	1652.0	1626.3
110*	1,3-Dimethyl-2-propylbenzene	0.1523	6.0775	12.1968	0.0000	11.8179	24.5297	0.0000	11.0496	0.0000	0.0000	1157.6	1716.6	1731.3
111	1-Ethyl-2-butylbenzene	0.0210	6.1911	1.8912	0.0000	23.3266	28.2163	0.0000	4.9110	0.0000	0.0000	1204.3	1713.0	1727.9
112	1-Propyl-2butylbenzene	0.0160	6.3314	1.1235	0.0000	26.5045	29.6911	0.0000	4.9110	0.0000	0.0000	1278.1	1705.8	1721.6
113	1-Isobutyl-4-*sec*-butylbenzene	0.4212	6.4584	12.8046	0.0000	13.6348	37.2183	0.0000	6.1093	0.0000	0.0000	1305.8	1698.3	1712.8
114	1-Methyl-4-*n*-hexylbenzene	0.0106	5.7000	4.4549	0.0000	25.0262	29.6251	0.0000	0.5457	0.0000	0.0000	1333.4	1655.5	1638.5
115*	1-Propyl-2-pentylbenzene	0.0126	6.3437	0.9819	0.0000	29.6956	30.1076	0.0000	4.9110	0.0000	0.0000	1373.8	1740.0	1712.5
116	*n*-Octylbenzene	0.0000	3.1879	0.0625	0.0000	41.1764	18.8702	0.0000	0.0000	0.0000	0.0000	1431.1	1808.0	1819.5
117	1-Propylheptylbenzene	0.0123	6.3653	0.7483	0.0000	37.9369	27.0124	0.0000	4.0000	0.0000	0.0000	1499.1	1803.2	1815.4
118	1,2-Dipentylbenzene	0.0084	6.4062	0.5496	0.0000	36.2076	31.2311	0.0000	4.9110	0.0000	0.0000	1559.8	1795.0	1808.5
119	1-Propyloctylbenzene	0.0100	6.3760	0.7071	0.0000	41.1415	27.1774	0.0000	4.0000	0.0000	0.0000	1596.6	1786.7	1799.3
120*	1-Pentylheptylbenzene	0.0083	6.4467	0.3961	0.0000	44.5707	27.9432	0.0000	4.0000	0.0000	0.0000	1676.0	1742.7	1724.5
121	1-Methylundecylbenzene	0.0083	4.9862	4.0631	0.0000	45.0272	22.2799	0.0000	4.0000	0.0000	0.0000	1744.8	1827.8	1798.6
122	*n*-Dodecylbenzene	0.0000	3.1975	0.0278	0.0000	53.9397	19.0552	0.0000	0.0000	0.0000	0.0000	1820.4	1830.0	1810.4

vertex eigenvalue correlative index (MVECI) in this paper. m_{11} (x_1) represents correlativeness between vertices whose atomic type belongs to the first one, and m_{12} (x_2) represents correlativeness between the first type of vertices and the second type of vertices, etc. All MVECI values of the compounds are listed in Table 1.

Validation models. The statistical parameter correlation coefficient (R_{CV}) and the standard deviation (SD_{CV}), for the leave-one-out (LOO) cross-validation are usually used for indicating the predictive ability of a model. However, the recent study of Tropsha and co-workers shows that the evaluation of the actual predictive power of a QSAR model using only the correlation coefficient (R_{CV}) is not enough, and the external validation is required [18—20]. The predictive power of a model for the external data set (test set) can be expressed by R_{test} and SD_{test}.

$$R_{test} = \sqrt{1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}}, \quad (4)$$

$$SD_{test} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{test}(y_i - \hat{y}_i)^2}. \quad (5)$$

In Eq. (4) or Eq. (5), both y_i and $\hat{y}_i$ are the experimented and calculated values of the test set, and $\overline{y}_i$ is the mean value of experimental values of the test set. In this paper, all R_{CV}, SD_{CV}, R_{test}, and SD_{test} were calculated to evaluate the models.

RESULTS AND DISCUSSION

Multiple linear regression model. Multiple linear regression (MLR) is a classic modeling approach of a linear fit for independent and dependent variables to get the best results in the least squares sense. Firstly, the retention indices (RIs) are used as the dependent variables and the structural descriptors are used as the independent variables to construct a model through MLR. The screening of variables prior to modeling is particularly necessary to find their best combinations. This study used the stepwise regression (SMR) to select variables. The stepwise regression screening is one of the most commonly used variable selection methods and an effective way to find the optimal subspace. SMR is carried out based on the significant level value (P) of a partial F-test. In the candidate variables, when the largest partial F-test significant level value $P \le 0.05$, the corresponding variable is introduced into the model; in the already introduced variables, when the smallest partial F-test′s significant level value $P \ge 0.10$, the corresponding variable is removed from the model. The ″leave one out″ cross-validation is carried out to test the models of each step. The change in correlation coefficients (R/R_{CV}) and standard deviations (SD/SD_{CV}) in SMR are shown in Figs. 1 and 2.

In Figs. 1 and 2, the large inflexions of the multiple correlation coefficients (R/R_{CV}) and standard deviations (SD/SD_{CV}) appear when the step is 3 in SMR. The multiple correlation coefficients (R/R_{CV})

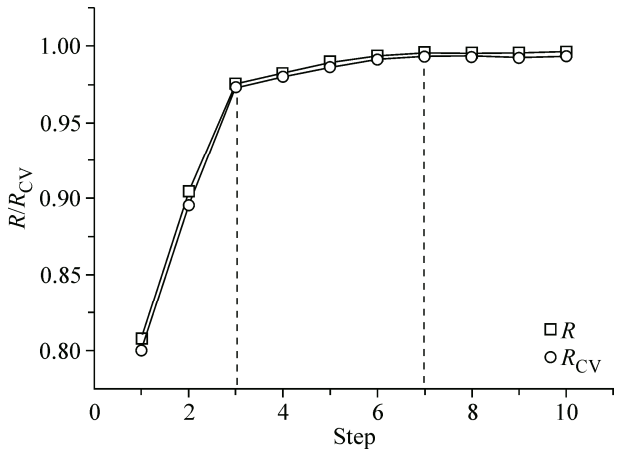

Fig. 1. Plot of R and R_{CV} change with SMR

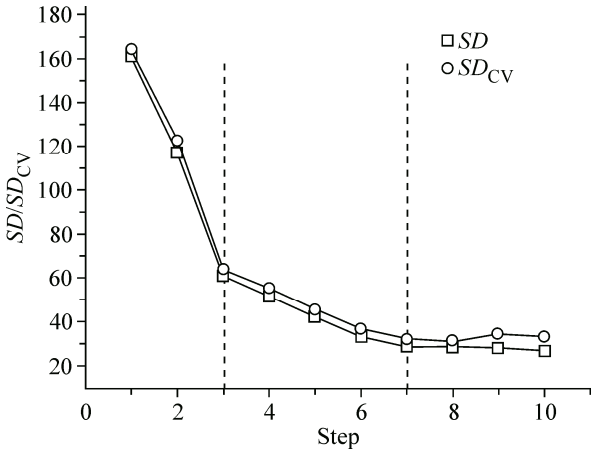

Fig. 2. Plot of SD and SD_{CV} change with SMR

reach the maximum and standard deviations (SD/SD_{CV}) reach the minimum when the step is 7. Therefore, subsets of these two steps should be chosen as the optimal subsets. x_2, x_3, and x_5 are selected when the step is 3, a 3-variable model (Eq. (6)) is constructed as follows:

$$\mathrm{RI} = 11.270 + 93.362x_2 + 26.214x_3 + 25.382x_5. \quad (6)$$

Model fitting: $N = 98$, $R = 0.975$, $SD = 60.822$, $F = 615.358$; cross-validation: $R_{CV} = 0.973$, $SD_{CV} = 63.962$, $F_{CV} = 553.431$; external prediction: $n = 24$, $R_{test} = 0.977$, $SD_{test} = 57.153$. N is the regression points, R is the multiple correlation coefficient, SD is the standard deviation of the estimate, F is the Fischer test value; R_{CV} is the correlation coefficient of the cross-validation, SD_{CV} is the standard deviation of the cross-validation, F_{CV} is the Fischer test value for the cross-validation; n is the number of samples of the test set, R_{test} is the multiple correlation coefficient of the external validation, SD_{test} is the standard deviation of the external validation.

x_1, x_2, x_3, x_4, x_5, x_6, and x_8 are selected when the step is 7, a 7-variable model (eq.(7)) is constructed as follows:

$$\mathrm{RI} = -123.709 - 532.449x_1 - 13.349x_2 + 33.690x_3 + 54.284x_4 + 27.144x_5 + 23.304x_6 + 14.097x_8. \quad (7)$$

Model fitting: $N = 98$, $R = 0.995$, $SD = 28.447$, $F = 1254.124$; cross-validation: $R_{CV} = 0.993$, $SD_{CV} = 32.289$, $F_{CV} = 970.569$; External prediction: $n = 24$, $R_{test} = 0.996$, $SD_{test} = 26.938$.

The multiple correlation coefficient (R), the cross-validated multiple correlation coefficient (R_{CV}), and the predictive correlation coefficient (R_{test}) of the two models are desirable, indicating good predictive ability and strong stability of the two models. The multiple correlation coefficient (R), the cross-validated multiple correlation coefficient (R_{CV}) and the predictive correlation coefficient (R_{test}) of the 7-variable model are significantly larger than those of the 3-variable model; the standard deviation (SD), the cross-validation standard deviation (SD_{CV}), and the predictive standard deviation (SD_{test}) of the 7-variable model are significantly smaller than those of the 3-variable model, indicating that the 7-variable model is significantly better than the 3-variable model.

Partial least squares regression model. To further explore the relationship between the molecular descriptors and Kovats RIs, then the partial least-squares (PLS) regression is used to construct another model. The Simca-P11.5 software (http://www.umetrics.com) is used to build the PLS model for the samples of the training set (software default parameters used). The ″leave one out″ cross-validation is used to test the model, and the test set is used for evaluating the external predictive ability of the model. A model is constructed from 10 descriptors (listed in Table 1) of training set samples. The number of principal components is 7. The correlation coefficient of the model (R), the cross-validation correlation coefficient (R_{CV}), and the predictive correlation coefficient (R_{test}) are 0.991, 0.988, and 0.995. The standard deviation (SD), the cross-validation standard deviation (SD_{CV}), and the predictive standard deviation (SD_{test}) are 36.831, 40.182, and 39.150, respectively. The R, R_{CV}, and R_{test} are higher than 0.95, showing good predictive ability and strong stability of the model.

Fig. 3 shows score distribution plots of 98 samples of the training set in the two front principal components. Most samples fall into the 95 % Hotelling T^2 confidence circle, and only three samples (< 4 %) are beyond this range. Statistical results indicate that the structural descriptors can successfully characterize a structural characteristic of organic compounds.

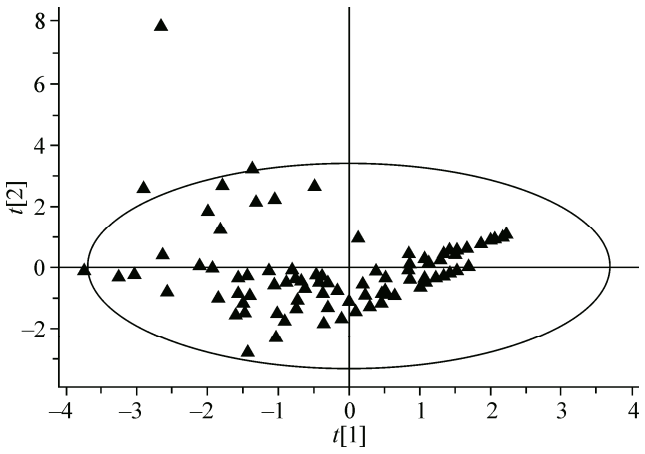

Fig. 3. Score distribution plots of the two front principal components of the PLS model

The variable importance in the projection (VIP) is an indicator that reflects the explanatory ability of each variable to Y (Fig. 4). Usually, the variable whose VIP value is greater

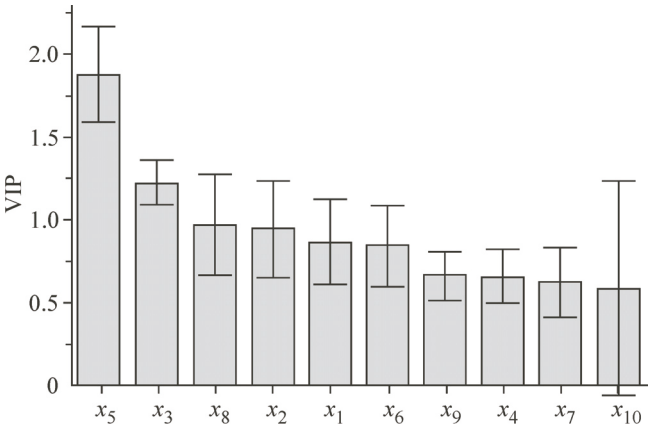

Fig. 4. Plot of variable importance in the projection

than 1, has greater correlation and larger explanatory power to *Y*. For this system, the VIP values of x_5 and x_3 are greater than 1, therefore the explanatory power of these two variables on the *Y* are relatively large. x_5 corresponds to the correlation between the second type of vertices; x_3 corresponds to the correlation between the first and third categories of vertices. The number of substituents in the benzene ring, the size and complexity of the substituents significantly affect the retention behavior of the compound.

Comparison of models. The Kovats retention indices of the training set are estimated and predicted by the MLR model (Eq. (7)) and the PLS model, and the results are listed in Table 1. Fig. 5 presents a plot of the observed RIs versus the estimated and predicted ones, and Fig. 6 presents the calculated results related to the residual distribution.

In Fig. 5, most sample points are close to the 45° diagonal line, indicating the good predictive ability and small prediction errors of the two models. Fig. 5 also shows that the MLR and PLS models have similar predictive abilities. In Fig. 6, ″±2*SD*″ lines of MLR model are closer to the middle ″0″ line, and only five sample points (4.10 %) beyond this range; ″±2*SD*″ lines of the PLS model are slightly farther from the middle ″0″ line, and seven sample points (5.74 %) beyond this range, indicating that the predictive ability of the MLR model is slightly better than that of the PLS model. The correlation coefficient (R_{test}) and the standard deviation (SD_{test}) of the two models also reflect that the predictive ability of the MLR model is slightly better than that of the PLS model. The spans of the molecular structures of the samples are considerably large, therfore the results obtained by the two models are satisfactory.

CONCLUSIONS

In this paper, a new molecular structural characterization method (MVECI) is constructed and successfully used to describe the molecular structures of a set of organic compounds. The models based on the descriptors have been developed to estimate and predict RIs of alkylbenzene components. The predictive ability of the models was evaluated by the leave-one-out cross-validation and the external sample test. The results presented above show that the descriptors developed in this work can be

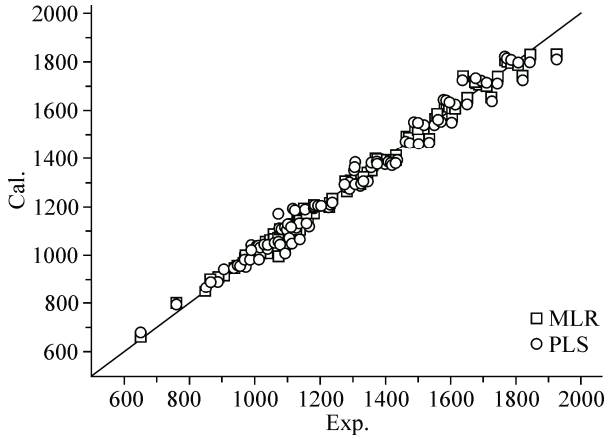

Fig. 5. Calculated vs. experimental RIs

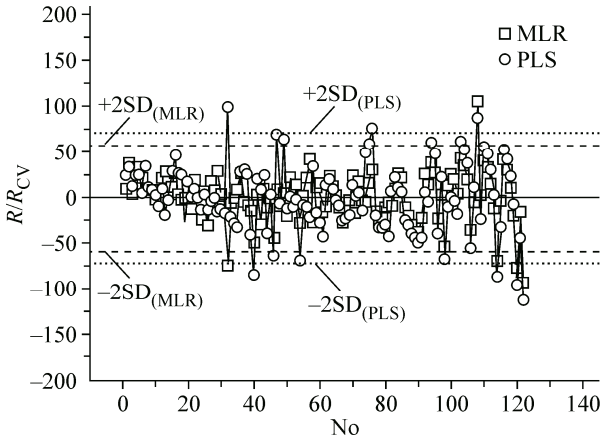

Fig. 6. Comparative residuals vs. compound No

used to study the structure — retention relationship of organic compounds, and their performance is satisfactory. The descriptors can be calculated directly from molecular 2D structures. As compared with the Comfa or Somfa methods [3—5], this method does not suffer from some limitations such as the requirement of molecular superposition, hence, it is simple and convenient. Furthermore, the descriptors can also be extended in other QSRR/QSAR investigations. This paper can present the favorable reference for the quantitative structure-retention relationship study of organic compounds.

This work was supported by the Foundation of Education Bureau, Sichuan Province (13ZB0003) and Technology Bureau, Sichuan Province (2012j13-141).

REFERENCES

1. *Luan F., Xue C.X., Zhang R.S. et al.* // Anal. Chim. Acta. – 2005. – **537**. – P. 101.
2. *Wang Z.S., Lin R.Z., Sun W.L. et al.* // Environ. Chem. – 2003. – **22**. – P. 85.
3. *Helgren T.R., Sciotti R.J., Lee P. et al.* // Bioorg. Med. Chem. Lett. – 2015. – **25**. – P. 327.
4. *Yu S.L., Yuan J.T., Shi J.H. et al.* // Chemometr. Intell. Lab. – 2015. – **146**. – P. 34.
5. *Arvind K., Anand Solomon K., Rajan S.S.* // Med. Chem. Res. – 2014. – **23**. – P. 1789.
6. *Musa Y., Ahmoda W., Al-Amiery A.A. et al.* // J. Struct. Chem. – 2013. – **54**. – P. 301.
7. *Qian J.-Z., Wang B.-C., Fan Y. et al.* // J. Struct. Chem. – 2015. – **56**. – P. 338.
8. *Yu X.L., Tan Z.D., Wang X.Y.* // J. Struct. Chem. – 2012. – **53**. – P. 443.
9. *Porto L.C., Souza E.S., Junkes B.S. et al.* // Talanta. – 2008. – **76**. – P. 407.
10. *Liao L.M., Zhu J., Li J.F. et al.* // Chin. J. Struct. Chem. – 2011. – **30**. – P. 105.
11. *Liao L.M., Li J.F., Lei G.D. et al.* // J. Struct. Chem. – 2011. – **52**. – P. 1111.
12. *Qin S., Li J.F., Liao L.M.* // Chin. J. Struct. Chem. – 2012. – **31**. – P. 665.
13. *Zhu W.P., Liang G.Z., Liao L.M. et al.* // Chin. J. Struct. Chem. – 2009. – **28**. – P. 391.
14. *Liao L.M., Mei H., Li J.F. et al.* // J. Mol. Struct.: THEOCHEM. – 2008. – **850**. – P. 1.
15. *Du X.H.* // J. Instrum. Anal. – 2003. – **22**. – P. 18.
16. *Tang Z.Q., Du X.H., Feng C.J. et al.* // J. Shanghai Univ. (Nat. Sci.). – 2003. – **9**. – P. 266.
17. *Qin Z.L.* // J. Xuzhou Normal Univ. (Nat. Sci.). – 2001. – **19**. – P. 50.
18. *Golbraikh A., Tropsha A.* // J. Mol. Graph. Model. – 2002. – **20**. – P. 269.
19. *Tropsha A., Gramatica P., Gombar V.K.* // QSAR Comb. Sci. – 2003. – **22**. – P. 69.
20. *Gramatica P., Pilutti P., Papa E.* // J. Chem. Inf. Comput. Sci. – 2004. – **44**. – P. 1794.