• 研究论文 •

电拓扑状态预测多氯二苯并噻吩及噻吩砜化合物的气相色谱保留指数

莫凌云^{*,a,b,c} 刘红艳^{a,b} 温焕宁^b

(^a 桂林理工大学广西环境工程与保护评价重点实验室 桂林 541004)

(^b 桂林理工大学化学与生物工程学院 桂林 541004)

(^c 广西师范大学珍稀濒危动植物生态与环境保护省部共建教育部重点实验室 桂林 541004)

摘要 以原子类型电拓扑状态指数(ETSI)有效表征了 135 个多氯二苯并噻吩(PCDT)和 135 个多氯二苯并噻吩砜(PCDTO₂)的分子结构,应用基于预测的变量选择与模型化(VSMP)方法建立 PCDT 和 PCDTO₂ 化合物在 DB-5 气相色谱柱上的气相色谱保留指数(RI)与分子结构(ETSI)的定量相关模型,模型的相关系数 r^2 分别为 0.9939 和 0.9729, LOO 交叉验证相关系数 q^2 分别为 0.9921 和 0.9692. 为验证模型稳定性和预测能力,应用 17 个 PCDT 和 PCDTO₂ 训练集样本构建的 QSRR 模型的 r^2 分别为 0.9959 和 0.9783, LOO 交叉验证相关系数 q^2 分别为 0.9921 和 0.9740, 说明模型具有良好的稳定性. 以此模型预测外部 8 个检验集及 110 个预测集的 RI 值, 8 个检验集样本的结果表明训练集模型具有良好预测能力.

关键词 多氯二苯并噻吩(PCDT); 多氯二苯并噻吩砜(PCDTO₂); 定量结构-保留相关(QSRR); 电拓扑状态指数(ETSI)

Prediction of Gas Chromatographic Retention Indices of Polychlorinated Dibenzothiophenes and Polychlorinated Dibenzothiophene Sulfones by Electrotopological State Index

Mo, Lingyun^{*,a,b,c} Liu, Hongyan^{a,b} Wen, Huanning^b

(^a The Guangxi Key Laboratory of Environmental Engineering, Protection and Assessment, Guilin University of Technology, Guilin 541004)

(^b College of Chemistry and Bioengineering, Guilin University of Technology, Guilin 541004)

(^c Key Laboratory of Ecology of Rare and Endangered Species and Environmental Protection, Ministry of Education of China, Guangxi Normal University, Guilin 541004)

Abstract Structures of 135 polychlorinated dibenzothiophenes (PCDTs) and 135 polychlorinated dibenzothiophene sulfones (PCDTO₂s) were characterized using electrotopological state index (ETSI) for atom types. Quantitative linear relationship models were built between gas chromatographic retention index (RI) from DB-5 GC column and ETSI descriptors were established for PCDT and PCDTO₂. The models derived from the variable selection and modeling based on prediction (VSMP) technique present correlation coefficient (r^2) of 0.9939 for PCDT and of 0.9729 for PCDTO₂, leave-one-out cross-validation coefficient (q^2) of 0.9921 for PCDT and of 0.9692 for PCDTO₂, respectively. In order to validate the stability and predictive power of the models, 17 PCDT and PCDTO₂ were used as training sets to build QSRR models. The models

* E-mail: molingyun123@126.com; Tel.: 13152582650

Received November 11, 2011; revised March 7, 2012; accepted March 20, 2012.

Project was supported by the Provincial Natural Science Foundation of Guangxi (No. 2011XNSFA018059), the Ministry of Education of Guangxi Zhuang Autonomous Region (Nos. 201010LX174, 200103YB065) and Key Laboratory of Ecology of Rare and Endangered Species and Environmental Protection (Guangxi Normal University), Ministry of Education, China (No. 1002K001).

广西自然科学基金(No. 2011XNSFA018059)、广西壮族自治区教育厅科研(Nos. 201010LX174, 200103YB065)、珍稀濒危动植物生态与环境保护省部共建教育部重点实验室研究基金(No. 1002K001)资助项目.

present r^2 of 0.9959 for PCDT and of 0.9783 for PCDT O_2 , q^2 of 0.9921 for PCDT and of 0.9740 for PCDT O_2 . The results demonstrate the robustness of the QSRR models. The models were used to predict the RI of test set with 8 compounds and predict set with 110 compounds, and the result of test set with 8 compounds show that the training set has a good predictive ability.

Keywords polychlorinated dibenzothiophene (PCDT); polychlorinated dibenzothiophene sulfone (PCDT O_2); quantitative structure-retention relationship (QSRR); electrotopological state index (ETSI)

多氯代二苯并噻吩(Polychlorinated Dibenzothiophenes, PCDTs)是结构类似于二噁英的含硫氯代多环芳烃化合物, 简称为“类二噁英(Dioxin Analog)”化合物^[1,2]. 由于其脂溶性和抗代谢, 导致平面氯代多环芳烃化合物具有在食物链中积聚并引起毒性作用的潜在可能性. 随着现代仪器检验技术的发展, 人们现已发现多氯代二苯并噻吩对环境具有潜在的危害^[3,4]. 多氯二苯并噻吩砜(PCDT O_2)类化合物, 是一类重要的有机合成中间体, 在农药、医药、生物、染料等合成方面起到重要的作用^[5], 而且在我国生产和使用量较大, 常通过多种途径进入生态环境, 从而对环境造成严重的危害.

定量分子结构-色谱保留相关(Quantitative structure-retention relationships, QSRR)作为 QSAR 研究中的重要组成部分, 对于建立分子结构与色谱保留的变化规律, 估计与预测化合物的物理化学性质与生物活性、选择分离条件及深入探索色谱保留机理具有重要意义^[6-9]. 本工作以原子类型电拓扑状态指数(ETSI)有效表征了 135 个多氯二苯并噻吩(PCDT)和 135 个多氯二苯并噻吩砜(PCDT O_2)的分子结构, 应用基于预测的变量选择与模型化(VSMP)^[7,10-14]方法建立 PCDT 和 PCDT O_2 化合物在 DB-5 气相色谱柱上的气相色谱保留指数(RI)与分子结构(ETSI)的定量相关模型. 以其中的

17 个 PCDT 和 17 个 PCDT O_2 训练集样本分别构建 QSRR 模型, 并以此模型预测外部 8 个检验集及 110 个预测集 RI , 从 8 个检验集样本的结果可以看出训练集模型具有良好预测能力.

1 实验数据和计算方法

1.1 数据来源

本工作所研究的 25 个多氯二苯并噻吩化合物来自文献[15], 文献中还报道了 25 个 PCDT O_2 的相对保留时间, 这些化合物的母体结构如图 1 所示, 图中数字 1~9 为氯原子取代位置, 所有化合物的色谱保留指数见表 1. 25 个 PCDT 和 PCDT O_2 化合物的 RI 分别介于 17.15~111.49 和 41.23~119.50 之间, 化合物的结构多样性以电拓扑状态指数有效表征.

图 1 PCDT 和 PCDT O_2 化合物的母体结构
Figure 1 Structure of PCDT and PCDT O_2

表 1 135 个 PCDT 与 PCDT O_2 化合物的实验值与计算值^a
Table 1 Observed values and calculated values of 135 PCDT and PCDT O_2

No.	取代位置	PCDT			PCDT O_2		
		x_{39}	实验值	预测值	x_{39}	实验值	预测值
1	1-	6.1840	17.88	19.77	6.0602	41.23	47.18
2	2-	5.9806	18.39*	19.34	5.8881	42.69*	46.89
3	3-	5.9529	18.46	19.28	5.8290	43.50	46.79
4	4-	6.1389	17.15	19.68	5.9651	49.41	47.02
5	1,2-	12.2386		32.79	12.0224		57.18
6	1,3-	12.1785	31.77*	32.66	11.9309	49.56*	57.02
7	1,4-	12.3496	31.95	33.03	12.0519	57.19	57.22
8	1,6-	12.3414	32.98	33.01	12.0438	58.22	57.21
9	1,7-	12.1505		32.60	11.9028		56.97
10	1,8-	12.1831		32.67	11.9668		57.08
11	1,9-	12.3947		33.13	12.1470		57.38
12	2,3-	12.0075	36.25	32.30	11.7912	55.24	56.79
13	2,4-	12.1611	31.09*	32.63	11.8949	56.85*	56.96

续表

No.	取代位置	PCDT			PCDTO ₂		
		x_{39}	实验值	预测值	x_{39}	实验值	预测值
14	2,6-	12.1632	32.72	32.63	11.8969	59.70	56.96
15	2,7-	11.9715		32.22	11.7552		56.73
16	2,8-	12.0024	35.55	32.29	11.8175	53.74	56.83
17	3,4-	12.1658		32.64	11.8682		56.92
18	3,6-	12.1054	34.09	32.51	11.8077	60.40	56.82
19	3,7-	11.9162		32.10	11.6685		56.58
20	4,6-	12.2963	32.68	32.92	11.9487	66.23	57.05
21	1,2,3-	18.3072		45.84	17.9671		67.14
22	1,2,4-	18.4459		46.14	18.0558		67.29
23	1,2,6-	18.4097		46.06	18.0195		67.23
24	1,2,7-	18.2155		45.65	17.5482		66.44
25	1,2,8-	18.2513		45.72	17.9426		67.10
26	1,2,9-	18.4678		46.19	18.1277		67.41
27	1,3,4-	18.4182		46.08	17.9967		67.19
28	1,3,6-	18.3496	45.36*	45.93	17.9281	65.28*	67.07
29	1,3,7-	18.1554	47.19*	45.52	17.7840	59.27*	66.83
30	1,3,8-	18.1880		45.59	17.8479		66.94
31	1,3,9-	18.4028		46.05	18.0313		67.25
32	1,4,6-	18.5255		46.31	18.0540		67.29
33	1,4,7-	18.3296		45.89	17.9082		67.04
34	1,4,8-	18.3623		45.96	17.9721		67.15
35	1,4,9-	18.5787		46.43	18.1573		67.46
36	1,6,7-	18.3820		46.00	17.9605		67.13
37	1,6,8-	18.3822		46.00	17.9921		67.18
38	1,7,8-	18.2236		45.66	17.8835		67.00
39	2,3,4-	18.2621		45.75	17.8720		66.98
40	2,3,6-	18.1736		45.56	17.7835		66.83
41	2,3,7-	17.9365		45.05	17.5964		66.52
42	2,3,8-	17.9168	51.95	45.00	17.6080	64.66	66.54
43	2,4,6-	18.3321	44.82*	45.90	17.8920	72.41*	67.01
44	2,4,7-	18.1380		45.48	17.7479		66.77
45	2,4,8-	18.1689	46.44	45.55	17.8102	65.97	66.88
46	2,6,7-	18.3369		45.91	17.8654		66.97
47	3,4,6-	18.1427		45.49	17.7213		66.73
48	3,4,7-	18.1704		45.55	17.7803		66.83
49	1,2,3,4-	24.5885		59.35	24.0746		77.38
50	1,2,3,6-	24.4919		59.14	23.9779		77.22
51	1,2,3,7-	24.2946		58.72	23.8306		76.97
52	1,2,3,8-	24.3304		58.79	23.8978		77.08
53	1,2,3,9-	24.5500		59.27	24.0861		77.40
54	1,2,4,6-	24.6354		59.45	24.0715		77.37
55	1,2,4,7-	24.4364		59.02	23.9224		77.12
56	1,2,4,8-	24.4722		59.10	23.9896		77.23
57	1,2,4,9-	24.6936		59.58	24.1796		77.55
58	1,2,6,7-	24.4606		59.07	23.9467		77.16
59	1,2,6,8-	24.4640		59.08	23.9814		77.22

续表

No.	取代位置	PCDT			PCDTO ₂		
		x_{39}	实验值	预测值	x_{39}	实验值	预测值
60	1,2,6,9-	24.6655		59.52	24.1516		77.51
61	1,2,7,8-	24.3023		58.73	23.8697		77.03
62	1,2,7,9-	24.4864		59.13	24.0224		77.29
63	1,2,8,9-	24.5546		59.28	24.1220		77.46
64	1,3,4,6-	24.6077		59.39	24.0124		77.27
65	1,3,4,7-	24.4087		58.96	23.8634		77.02
66	1,3,4,8-	24.4413		59.03	23.9273		77.13
67	1,3,4,9-	24.6610		59.51	24.1156		77.45
68	1,3,6,7-	24.4005		58.95	23.8552		77.01
69	1,3,6,8-	24.4007		58.95	23.8868		77.06
70	1,3,6,9-	24.6005		59.38	24.0552		77.34
71	1,3,7,8-	24.2390		58.60	23.7751		76.88
72	1,3,7,9-	24.4214		58.99	23.9261		77.13
73	1,4,6,7-	24.5796		59.33	23.9843		77.23
74	1,4,6,8-	24.5798		59.33	24.0159		77.28
75	1,4,6,9-	24.7813		59.76	24.1860		77.56
76	1,4,7,8-	24.4164		58.98	23.9025		77.09
77	1,6,7,8-	24.4968		59.15	23.9828		77.22
78	2,3,4,6-	24.4468		59.05	23.8828		77.06
79	2,3,4,7-	24.2495		58.62	23.7355		76.81
80	2,3,4,8-	24.2803		58.69	23.7978		76.91
81	2,3,6,7-	24.2214		58.56	23.7075		76.76
82	2,3,6,8-	24.2199	61.74	58.56	23.7373	76.72	76.81
83	2,3,7,8-	24.0599		58.21	23.6273		76.63
84	2,4,6,7-	24.3831		58.91	23.8192		76.95
85	2,4,6,8-	24.3816		58.90	23.8490		77.00
86	3,4,6,7-	24.3878		58.92	23.7925		76.90
87	1,2,3,4,6-	30.7917		72.69	30.1039		87.48
88	1,2,3,4,7-	30.5513		72.17	29.9135		87.16
89	1,2,3,4,8-	30.5668		72.21	29.9604		87.24
90	1,2,3,4,9-	30.7733		72.65	30.1356		87.54
91	1,2,3,6,7-	30.4157		71.88	29.7779		86.94
92	1,2,3,6,8-	30.5566		72.18	29.9502		87.23
93	1,2,3,6,9-	30.7613		72.62	30.1236		87.52
94	1,2,3,7,8-	30.5533		72.18	29.9155		87.17
95	1,2,3,7,9-	30.4752		72.01	29.8688		87.09
96	1,2,3,8,9-	30.6472		72.38	30.0908		87.46
97	1,2,4,6,7-	30.7000		72.49	30.0122		87.33
98	1,2,4,6,8-	30.7034		72.50	30.0470		87.39
99	1,2,4,6,9-	30.9098		72.94	30.2220		87.68
100	1,2,4,7,8-	30.5367	74.40	72.14	29.9303	84.64	87.19
101	1,2,4,7,9-	30.7257		72.55	30.0880		87.46
102	1,2,4,8,9-	30.7102		72.51	30.1038		87.48
103	1,2,6,7,8-	30.5890		72.25	29.9827		87.28
104	1,2,6,7,9-	30.7581		72.62	30.1204		87.51
105	1,3,4,6,7-	30.6723		72.43	29.9531		87.23

续表

No.	取代位置	PCDT			PCDTO ₂		
		x_{39}	实验值	预测值	x_{39}	实验值	预测值
106	1,3,4,6,8-	30.6725		72.43	29.9847		87.28
107	1,3,4,6,9-	30.8772		72.87	30.1580		87.57
108	1,3,4,7,8-	30.5058		72.07	29.8681		87.09
109	1,3,4,7,9-	30.6931		72.48	30.0240		87.35
110	1,3,6,7,8-	30.5258		72.12	29.8880		87.12
111	1,4,6,7,8-	30.7081		72.51	30.0203		87.34
112	2,3,4,6,7-	30.5081		72.08	29.8204		87.01
113	2,3,4,6,8-	30.5066		72.08	29.8502		87.06
114	2,3,4,7,8-	30.3417		71.72	29.7353		86.87
115	1,2,3,4,6,7-	36.8667		85.75	36.0551		97.46
116	1,2,3,4,6,8-	36.8359		85.69	36.0557		97.46
117	1,2,3,4,6,9-	37.0262		86.10	36.2146		97.73
118	1,2,3,4,7,8-	36.6418		85.27	35.9116		97.22
119	1,2,3,4,7,9-	36.8321		85.68	36.1308		97.59
120	1,2,3,4,8,9-	36.9606		85.95	36.2304		97.75
121	1,2,3,6,7,8-	36.6921		85.38	35.9619		97.30
122	1,2,3,6,7,9-	36.8644		85.75	36.1028		97.54
123	1,2,3,7,8,9-	36.7503		85.50	36.0700		97.48
124	1,2,4,6,7,8-	36.8420		85.70	36.0618		97.47
125	1,2,4,6,7,9-	37.0160		86.07	36.2044		97.71
126	1,2,4,6,8,9-	37.0518		86.15	36.2716		97.82
127	1,2,4,7,8,9-	36.9002		85.82	36.1699		97.65
128	1,3,4,6,7,8-	36.8111		85.63	35.9995		97.37
129	1,3,4,6,7,9-	36.9834		86.00	36.1404		97.60
130	2,3,4,6,7,8-	36.6421		85.27	35.8618		97.14
131	1,2,3,4,6,7,8-	43.0191	98.70*	98.98	42.1151	110.50*	107.62
132	1,2,3,4,6,7,9-	43.1963	98.07	99.36	42.2609	108.37	107.86
133	1,2,3,4,6,8,9-	43.2321	97.71	99.44	42.3280	106.56	107.97
134	1,2,3,4,7,8,9-	43.0773	102.94	99.11	42.2232	105.19*	107.80
135	1,2,3,4,6,7,8,9-	49.4228	111.49*	112.75	48.3949	119.50	118.14

^a 带“*”的为检验集化合物。

1.2 计算方法

PCDT 化合物的分子结构采用修正的原子类型电拓扑状态指数(Electrotopological State Indices for atom type, 简称 ETSI) 进行描述^[16~19]。首先利用 Chemioffice2005 软件中的 ChemiDraw 模块构建所有化合物的分子结构, 然后将所有化合物的分子结构依据非氢原子的原子属性及邻接关系输入电脑形成一个文本文件, 采用 VSMP^[11,20]程序共产生 41 个电拓扑描述子, 删除全部为零的描述子后共得到 4 个非零电拓扑描述子, 分别是 x_6 , x_{11} , x_{35} 和 x_{39} , 对应的原子类型分别为: aCHa, aCa—, —S—和—Cl。再次应用 VSMP 方法进行变量优化选择, 直至得到最佳模型, 最后以此模型预测未知 PCDT 类化合物的 RI 值。

对于 PCDTO₂ 类化合物采用同样的方法得到 5 个非零电拓扑描述子 x_6 , x_{11} , x_{25} , x_{38} 和 x_{39} , 对应的原子类型分别为: aCHa, aCa—, =O, $\geq S \leq$ 和—Cl。也应用 VSMP 方法进行变量优化选择, 直至得到最佳模型, 并以此模型预测未知 PCDTO₂ 类化合物的 RI 值。

2 结果与讨论

2.1 变量筛选

以 6 个非零电拓扑描述子为自变量(X), 以色谱保留指数为因变量(Y)构建数据集, 应用基于预测的最佳子集回归变量选择与模型化(VSMP)方法建立最佳子集回归模型。不同变量数下的模型相关系数(r^2)、交互验证相关系数(q^2)及均方根误差(RMSEP)随变量数的变化趋势

如表 2 所示. 从表中结果我们不难看出, 对于 PCDTO₂ 化合物来说, 当变量数为 1 时, 模型的交互验证相关系数达到最大, 均方根误差已达到最小, 因此对该数据集建立 1 变量即描述子 x_{39} 模型; 而对于 PCDT 化合物, 虽然 2 变量时模型的 q^2 达到最大, RMSEP 达到最小, 但是和 1 变量相比变化不大, 所以我们仍然建立其 1 变量即描述子 x_{39} 模型.

表 2 QSRR 模型的统计参数

Table 2 The practical parameters of the QSRR models

m	r^2	r	RMSEE	q^2	q	RMSEP	Variable
PCDT							
1	0.99390	0.9969	2.27	0.9921	0.9961	2.59	39
2	0.99450	0.9972	2.15	0.9925	0.9962	2.53	11, 39
PCDTO ₂							
1	0.97290	0.9863	3.74	0.9692	0.9845	4.00	39
2	0.97350	0.9867	3.69	0.9671	0.9834	4.14	11, 39

2.2 模型与验证

根据变量筛选结果, 分别建立了 PCDT 和 PCDTO₂ 的 1 变量即描述子 x_{39} 与 RI 的 QSRR 模型如下.

PCDT 化合物(模型 1):

$$RI = (6.760379 \pm 0.852399) + (2.159092 \pm 0.0353473)x_{39} \quad (1)$$

$n=25, m=1, r^2=0.9939, r=0.9969,$
 $RMSEE=2.27, F=3731.102$ (建模)
 $n=25, m=1, q^2=0.9921,$
 $q=0.9961, RMSEP=2.59$ (LOO 检验)

PCDTO₂ 化合物(模型 2):

$$RI = (35.811734 \pm 1.403109) + (1.706400 \pm 0.059424)x_{39} \quad (2)$$

$n=25, m=1, r^2=0.9929, r=0.9863,$
 $RMSEE=3.74, F=824.595$ (建模)
 $n=25, m=1, q^2=0.9692, q=0.9845,$
 $RMSEP=4.00$ (LOO 检验)

式中 n 为样本数, m 为最佳变量数, F 为 Fisher 统计量, $\pm$ 号前后的数值表示模型的回归系数及其相应的标准偏差. 由上述两个方程我们不难看出, 模型 1 和模型 2 均具有良好的估计能力. 对于这两类化合物来说, 其色谱保留指数均和二苯并噻吩环上的取代原子氯原子相关, 模型 2 的相关性低于模型 1 可能是因为 PCDTO₂ 类化合物结构中多了两个氧原子, 氧原子的电负性对氯原子的

取代有影响, 影响了其拓扑结构.

一个好的 QSRR 模型不仅应该具有良好的估计能力, 还应该具有较好的稳定性和对外部样本的预测能力. 为了验证模型的稳定性和预测能力, 我们将全部已知的 RI 进行从小到大排序, 均匀选取其中的 2/3 作为训练集样本进行建模, 然后用模型来预测以其余 1/3 化合物建立的外部样本的 RI . 同样应用 VSMP 方法, 以 q^2 和 LOO 检验均方根误差为优化目标, 对训练集进行不同变量数的最佳变量选择, 结果见表 3. 由表 3 不难看出, 化合物最佳模型的变量数仍然为 1, 选择的最佳量子集与应用全部 25 个化合物所得结果相同, 即仍是描述子 x_{39} , 这说明 17 个训练集与 25 个 PCDT 及 PCDTO₂ 的整体数据集具有相同的 QSRR 规律. 两个训练集模型的有关统计量与全部 25 化合物建立的模型统计量之间没有显著性差异, 说明训练集模型同样具有良好估计能力与稳定性. 所得模型见方程 3 和 4.

PCDT 化合物的训练集模型(模型 3):

$$RI = (6.915293 \pm 0.851903) + (2.143315 \pm 0.035715)x_{39} \quad (3)$$

$n=17, m=1, r^2=0.9959, r=0.9979,$
 $RMSEE=1.85, F=3601.396$ (建模)
 $n=17, m=1, q^2=0.9921, q=0.9960,$
 $RMSEP=2.74$ (LOO 检验)

PCDTO₂ 化合物的训练集模型(模型 4):

$$RI = (30.022547 \pm 1.490945) + (1.674930 \pm 0.064372)x_{39} \quad (4)$$

$n=17, m=1, r^2=0.9783, r=0.9891,$
 $RMSEE=3.29, F=677.006$ (建模)
 $n=17, m=1, q^2=0.9740, q=0.9869,$
 $RMSEP=3.67$ (LOO 检验)

表 3 训练集化合物模型的统计参数

Table 3 The practical parameters of the QSRR models of the training set

m	r^2	r	RMSEE	q^2	q	RMSEP	Variable
PCDT							
1	0.9959	0.9979	1.85	0.9921	0.9960	2.74	39
2	0.9973	0.9986	1.49	0.9940	0.9970	2.45	11, 39
PCDTO ₂							
1	0.9783	0.9891	3.29	0.9740	0.9869	3.67	39
2	0.9788	0.9894	3.25	0.9697	0.9848	4.00	11, 39

由 Eqs. 3, 4 得到的 17 个化合物的计算值和实验值一起列于表 1, 以及由 Eqs. 3, 4 估计的 8 个化合物的估计值以及由这两个方程预测的 111 个化合物的预测值均列于表 1. 以训练集化合物的实验值对估计值作相关图见图 2 和图 3, 17 个化合物的实验值和估计值除个别化合物(比如 12 号)均位于直线附近, 12 号化合物的估计值和实验值误差较大, 可能是因为氯原子取代于 2,3 位上, 二者相邻, 强电负性对化合物的拓扑结构有一定影响, 另外这个化合物的实验值和同样是 2 个氯取代的化合物相比也偏大一些. 因此由图 2 和图 3 及表 1 也可以看出, 该模型具有良好的预测能力.

图 2 PCDT 化合物的训练集 QSRR 模型估计及预测与实验 RI 相关图

Figure 2 Plot of estimated and predicted versus observed RI of PCDTs in training set

图 3 PCDTO₂ 化合物的训练集 QSRR 模型估计及预测与实验 RI 相关图

Figure 3 Plot of estimated and predicted versus observed RI of PCDTO₂s in training set

从得到的 4 个方程来看, 对于 PCDT 和 PCDTO₂ 化合物来说, 它们的色谱保留指数都和氯原子电拓扑指数

有着密切的关系, 从表 1 我们也不难看出, 随着氯原子取代个数的增加, 氯原子电拓扑指数逐渐增加, 化合物的色谱保留指数逐渐增加, 进一步说明氯原子是影响这类化合物色谱保留指数的重要因素.

为了进一步验证预测模型的质量, 我们参照文献 [21~25] 的相关计算方法计算了相关检验集的统计参数, 如下所示:

PCDT:

$$q^2=0.9987, r^2=0.9987, k=0.9915, k'=1.0084, r_0^2=0.9997, r_0'^2=0.9997, [(r^2-r_0^2)/r^2]=0.00096$$

PCDTO₂:

$$q^2=0.9592, r^2=0.9982, k=0.9810, k'=1.0156, r_0^2=0.9956, r_0'^2=0.9975, [(r^2-r_0^2)/r^2]=0.0026$$

其中, q^2 表示交互验证相关系数, r^2 表示相关系数, k 表示估计值与实验值相关回归直线的斜率, k' 表示预测值与实验值相关回归直线的斜率, r_0^2 表示估计值对实验值的拟合相关系数, $r_0'^2$ 表示预测值对实验值的拟合相关系数. 相应的计算公式如下:

$$q^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(\hat{y}_i-\bar{y})^2}, r^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}, k=\frac{\sum y_i \hat{y}_i}{\sum \hat{y}_i^2},$$

$$k'=\frac{\sum y_i \hat{y}_i}{\sum y_i^2}, r_0^2=1-\frac{\sum(\hat{y}_i-k\hat{y}_i)^2}{\sum(\hat{y}_i-\bar{y})^2},$$

$$r_0'^2=1-\frac{\sum(y_i-k'\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$$

从以上结果我们可以看出, 检验集的相关参数均满足: (i) $q^2>0.5$, (ii) $r^2>0.6$, (iii) r_0^2 接近 r^2 , 并且也满足条件 $[(r^2-r_0^2)/r^2]<0.1$, $0.85\leq k\leq 1.15$ 或 $0.85\leq k'\leq 1.15$. 因此, 认为采用氯原子的电拓扑指数建立的 QSAR 模型不仅具有很高的估计质量, 也具有很好的预测能力.

3 结论

原子类型 E-状态指数(ETSI)能有效表征 135 个 PCDT 和 135 个 PCDTO₂ 类化合物的分子结构. 通过 VSMP 优化选择的 1 个 ETSI 描述子, 即二苯并噻吩环上的取代原子氯原子的电拓扑指数与化合物的色谱保留指数具有高度相关性, 随着氯原子取代个数的增加, 氯原子电拓扑指数逐渐增加, 化合物的色谱保留指数逐渐增加. ETSI 与保留指数之间的定量相关模型不仅对内部样本具有良好的估计能力, 同时对外部样本具有良好的预测能力.

References

- 1 Wang, Y.; Zeng, X.-L.; Chen, H.-J.; Wang, H.-J. *J. Chem. Eng. Data* **2007**, *52*, 1442.
- 2 Sielex, K.; Andersson, J.-T. *Fresenius J. Anal. Chem.* **1997**, *359*, 261.
- 3 Sinkkonen, S. *Chemosphere* **1997**, *34*, 2585.
- 4 Wiedmann, T.; Riehle, U.; Kurz, J.; Ballschmi, K. *Fresenius J. Anal. Chem.* **1997**, *359*, 176.
- 5 Ma, J.; Wang, C.-S.; Zhang, X.-L.; Zhou, S.-X. *Chemistry* **2007**, *8*, 604 (in Chinese).
(马洁, 王长水, 张秀兰, 周生学, 化学通报, **2007**, *8*, 604.)
- 6 Wang, L.-S. *Chemistry of Organic Pollution*, Higher Education Press, Beijing, **2004**, pp. 412~483 (in Chinese).
(王连生, 有机污染化学, 高等教育出版社, 北京, **2004**, pp. 412~483.)
- 7 Wang, Y.; Liu, S.-S.; Zhao, J.-S.; Wang, X.-D.; Wang, L.-S. *Acta Chim. Sinica* **2006**, *64*(10), 1043 (in Chinese).
(王宇, 刘树深, 赵劲松, 王晓栋, 王连生, 化学学报, **2006**, *64*(10), 1043.)
- 8 Liu, L.; Mei, H.; Pi, X.-T.; Long, R.-C.; Li, J.; Li, Z.-L. *Acta Chim. Sinica* **2008**, *66*(16), 1873 (in Chinese).
(刘丽, 梅虎, 皮喜田, 龙瑞才, 李建, 李志良, 化学学报, **2008**, *66*(16), 1873.)
- 9 Li, Z.-H.; Cheng, F.-S.; Xia, Z.-N. *Chin. J. Chromatogr.* **2011**, *29*(1), 63 (in Chinese).
(李正华, 程凡圣, 夏之宁, 色谱, **2011**, *29*(1), 63.)
- 10 Liu, S.-S.; Yin, C.-S.; Cai, S.-X.; Li, Z.-L. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321.
- 11 Liu, S.-S.; Liu, H.-L.; Yin, C.-S.; Wang, L.-S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*(3), 964.
- 12 Liu, H.-Y.; Liu, S.-S.; Qin, L.-T. *J. Mol. Model.* **2007**, *13*, 611.
- 13 Liu, H.-Y.; Yi, Z.-S.; Mo, L.-Y. *Acta Chim. Sinica* **2009**, *67*(14), 1626 (in Chinese).
(刘红艳, 易忠胜, 莫凌云, 化学学报, **2009**, *67*(14), 1626.)
- 14 Zhang, Y.-H.; Liu, S.-S.; Xiao, Q.-F.; Qin, L.-T.; Xiao, Z.-L. *Acta Chim. Sinica* **2011**, *69*(10), 1232 (in Chinese).
(张永红, 刘树深, 肖乾芳, 覃礼堂, 夏之宁, 化学学报, **2011**, *69*(10), 1232.)
- 15 Sielex, K.; Andersson, J.-T. *J. Chromatogr. A* **2000**, *866*, 105.
- 16 Liu, S.-S. *Structural Characterization of Organic Compounds by the Molecular Electronegativity Distance Vector*, Higher Education Press, Beijing, **2005**, pp. 39~47 (in Chinese).
(刘树深, 有机物分子电性距离矢量表征及其应用, 高等教育出版社, 北京, **2005**, pp. 39~47.)
- 17 Kier, L.-B.; Hall, L.-H. *Pharm. Res.* **1990**, *7*, 801.
- 18 Hall, L.-H.; Kier, L.-B. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039.
- 19 Hall, L.-H.; Story, C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004.
- 20 Liu, S.-S.; Yin, C.-S.; Wang, L.-S. *J. Chem. Inf. Comput. Sci.* **2002**, *42*(3), 749.
- 21 Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.
- 22 Golbraikh, A.; Tropsha, A. *J. Comput. Aided Mol. Des.* **2002**, *16*, 357.
- 23 Golbraikh, A.; Tropsha, A. *Mol. Diversity* **2000**, *5*, 231.
- 24 Afantitis, A.; Melagraki, G.; Sarimveis, H.; Igglessi M.-O.; Kollias, G. *Eur. J. Med. Chem.* **2009**, *44*(2), 877.
- 25 Roy, P.-P.; Roy, K. *QSAR Comb. Sci.* **2008**, *27*, 302.

(A1111113 Zhao, C.)