

灯盏花挥发性组分结构与保留时间关系研究

李建凤^{1,2}, 廖立敏^{1,2*}

¹内江师范学院化学化工学院; ²“果类废弃物资源化”四川省高等学校重点实验室, 内江 641100

摘要:通过对有机化合物非氢原子进行分类、参数化染色、建立非氢原子之间的关系得到新的结构描述符。对灯盏花的 64 种挥发性有机化合物结构进行了参数化表征,运用多元线性回归(MLR)和偏最小二乘回归(PLS)方法构建了化合物结构与色谱保留时间的关系模型。通过“留一法”交互检验对模型的稳定性进行了评价,利用外部样本集对模型的预测能力进行了检验。两模型的相关系数(R^2)、交互检验的相关系数(R_{cv}^2)、外部预测的相关系数(R_{test}^2)值均较为理想,表明所建模型具有良好的拟合能力、稳定性和外部预测能力。分析了影响化合物色谱保留时间的结构因素,结果表明化合物的伯、仲碳原子越多,该化合物可能具有较大的色谱保留时间(t_R)值。本文对于天然产物中的挥发性有机化合物结构与性质关系研究具有一定的参考价值。

关键词:灯盏花; 挥发性化合物; 结构描述符; 结构表征; 模拟预测

中图分类号:R629

文献标识码:A

文章编号:1001-6880(2021)2-0236-10

DOI:10.16333/j.1001-6880.2021.2.008

Study on the relationship between structure and retention time of volatile component from *Erigeron breviscapus*

LI Jian-feng^{1,2}, LIAO Li-min^{1,2*}

¹College of Chemistry and Chemical Engineering, Neijiang Normal University;

²Key Laboratory of Fruit Waste Treatment and Resource Recycling of Sichuan Provincial College, Neijiang 641100, China

Abstract: By classifying non-hydrogen atoms of organic compounds, parametric dyeing, and establishing the relationship between them, new structure descriptors were constructed. The structure of 64 volatile organic compounds of *Erigeron breviscapus* was parametrically characterized. The multiple linear regression (MLR) and partial least squares regression (PLS) methods were used to build two models of the relationship between compound structure and chromatographic retention time. The stability of the models was evaluated by the "leave-one-out" cross tests, and the predictive ability of the models was tested by an external sample set. The correlation coefficients (R^2), cross-test correlation coefficients (R_{cv}^2), and external prediction correlation coefficients (R_{test}^2) of the two models are excellent, which indicates the good fitting ability, stability, and external prediction ability of the models. The structural factors affecting the chromatographic retention time of the compounds were analyzed. The results show that the more first and second type of non-hydrogen atoms in a compound, the compound may has a larger chromatographic retention time (t_R) value. This paper has certain reference value for the study on the relationship between the structure and properties of volatile organic compounds.

Key words: *Erigeron breviscapus*; volatile organic compound; structure descriptor; structure characterization; simulation prediction

灯盏花是菊科植物短葶飞蓬(*Erigeron breviscapus*)的全草,又名灯盏细辛、东菊^[1]。灯盏花主要生长在我国云南、广西、四川、贵州、西藏等高海拔地

区。灯盏花性寒、微苦、甘温辛,具有散寒解表、祛风除湿、活血化瘀、通经活络、消炎止痛等功效^[2]。目前,灯盏花在临幊上主要用于心脑血管疾病、糖尿病、肾病和眩晕等疾病的治疗。Xu 等^[3]利用固相微萃取、气相色谱-质谱联用技术对灯盏花中的挥发性成分进行了定性分析。化合物结构与色谱保留时间关系研究,对于分析化合物的色谱保留行为、辅助鉴

收稿日期:2020-03-27 接受日期:2020-06-02

基金项目:四川省教育厅科研项目(18ZB0323);内江师范学院校
级重点项目(18ZA04)

*通信作者 E-mail:liaolimin523@126.com

定化合物等具有重要的意义^[4]。化合物结构参数化表征是建立化合物结构与性质关系的关键步骤之一,目前研究者在这方面做过许多工作,如 He 等^[5,6]利用分子连接性指数和分子拓扑指数对果酒、香水百合头中部分挥发性成分进行了结构-保留关系研究,Zhang 等^[7]利用分子连接性指数对燃料油中有机硫化物进行了结构-保留关系研究,均取得较好的结果。分子连接性指数和分子拓扑指数都属于分子二维结构描述符,它的优点就是计算快速、简便、易懂,缺点是不能区分顺反异构、光学异构等现象;又如 Jing 等^[8]利用三维全息原子场作用矢量对部分脂肪酸进行了结构-保留关系研究,Liu 等^[9]利用化合物三维结构计算结构参数值对饱和醇类化合物进行了结构-保留关系研究,均取得较好的结果。分子三维结构描述符,它的优点就是基于化合物分子三维立体结构计算,更加接近化合物分子实际存在的状态,可以区分顺反异构、光学异构等现象,缺点是需要进行分子结构优化、计算复杂、难懂、工作量大,有时还可能涉及到探针选取、网格划分、分子重叠等不确定因素^[10]。本文在前人的基础上构建

新的二维结构描述符,对灯盏花中出现的多类有机化合物进行结构表征,进而运用多元线性回归(MLR)和偏最小二乘回归(PLS)两种方法建立化合物结构与保留时间关系模型,分析影响化合物色谱保留时间的结构因素,为挥发性化合物结构-性质关系研究提供参考。

1 材料和方法

1.1 实验材料

以灯盏花中的 64 种挥发性有机化合物为研究样本,化合物保留时间是以 HP-5 ms(30 m × 250 μm × 0.25 μm)为色谱柱分离得到,柱箱升温程序为:初始温度 50 °C,保持 1 min,以 5 °C/min 升至 220 °C 保持 6 min,再以 10 °C/min 升至 260 °C 保持 15 min。进样口温度 280 °C。采用脉冲不分流进样模式。载气使用高纯氮气,流速为 1.0 mL/min。按照色谱保留时间(t_R)^[3]大小顺序列于表 1,取序号个位数为“0”和“5”的化合物(用“*”标注,共 12 个)为测试样本集,测试样本集不参与建模,仅用于评价模型的外部预测能力,其余 52 个化合物为训练样本集用于建立模型。

表 1 64 种挥发性有机化合物及其色谱保留时间

Table 1 64 volatile organic compounds and their chromatographic retention times

序号 No.	化合物 Compound	分子式 Molecular formula	保留时间						
			t_R (min)	Cal. 1	Err. 1	Err. 1%	Cal. 2	Err. 2	Err. 2%
1	Bicyclo[3.1.0]hexane,6-isopropylidene-	C ₉ H ₁₄	6.60	6.62	0.02	0.30	6.12	-0.48	-7.27
2	d- α -Pinene	C ₁₀ H ₁₆	8.35	8.83	0.48	5.75	8.11	-0.24	-2.87
3	Benzaldehyde	C ₇ H ₆ O	9.23	9.16	-0.07	-0.76	10.28	1.05	11.38
4	L- β -Pinene	C ₁₀ H ₁₆	9.72	9.82	0.10	1.03	10.56	0.84	8.64
5*	O-Cymene	C ₁₀ H ₁₄	11.16	11.35	0.19	1.70	11.04	-0.12	-1.08
6	Limonene	C ₁₀ H ₁₆	11.31	11.20	-0.11	-0.97	10.93	-0.38	-3.36
7	Benzeneacetaldehyde	C ₈ H ₈ O	11.76	11.92	0.16	1.36	12.81	1.05	8.93
8	p-Cymenene	C ₁₀ H ₁₂	13.20	13.81	0.61	4.62	14.34	1.14	8.64
9	Bicyclo[3.1.1]heptan-2-one,6,6-dimethyl-,(1R)-	C ₉ H ₁₄ O	14.70	15.42	0.72	4.90	15.48	0.78	5.31
10*	Cyclohexanone,5-methyl-2-(1-methylethyl)-,cis-	C ₁₀ H ₁₈ O	15.20	15.83	0.63	4.14	15.85	0.65	4.28
11	Bicyclo[2.2.1]heptan-3-one,6,6-dimethyl-2-methylene-	C ₁₀ H ₁₄ O	15.40	15.95	0.55	3.57	16.66	1.26	8.18
12	Ethanone,1-(3-methylphenyl)-	C ₉ H ₁₀ O	16.05	16.07	0.02	0.12	16.29	0.24	1.50
13	Benzenemethanol, $\alpha,\alpha,4$ -trimethyl-	C ₁₀ H ₁₄ O	16.13	15.21	-0.92	-5.70	16.53	0.40	2.48
14	Estragole	C ₁₀ H ₁₂ O	16.44	16.78	0.34	2.07	16.53	0.09	0.55
15*	Benzene,2-methoxy-4-methyl-1-(1-methylethyl)-	C ₁₁ H ₁₆ O	17.33	18.07	0.74	4.27	18.54	1.21	6.98
16	Pulegone	C ₁₀ H ₁₆ O	17.60	17.09	-0.51	-2.90	18.02	0.42	2.39
17	2-Cyclohexen-1-one,2-methyl-5-(1-methylethenyl)-	C ₁₀ H ₁₄ O	17.77	17.94	0.17	0.96	18.88	1.11	6.25

续表1(Continued Tab. 1)

序号 No.	化合物 Compound	分子式 Molecular formula	保留时间 <i>t_R</i> (min)	保留时间					
				Cal. 1	Err. 1	Err. 1%	Cal. 2	Err. 2	Err. 2%
18	1-Pentanone, 1-(2-furanyl)-	C ₉ H ₁₂ O ₂	17.91	17.87	-0.04	-0.22	19.71	1.80	10.05
19	(-) - <i>cis</i> -Myrtanol	C ₁₀ H ₁₈ O	18.35	18.14	-0.21	-1.14	16.78	-1.57	-8.56
20*	Acetic acid, 1,7,7-trimethyl-bicyclo[2.2.1]hept-2-yl ester	C ₁₂ H ₂₀ O ₂	18.90	19.90	1.00	5.29	19.12	0.22	1.16
21	Benzene, 1-methoxy-4-(1-propenyl)-	C ₁₀ H ₁₂ O	18.96	17.07	-1.89	-9.97	17.87	-1.09	-5.75
22	Benzene, 1,2,4-tripropyl-	C ₁₅ H ₂₄	20.15	20.92	0.77	3.82	19.29	-0.86	-4.27
23	β -Maaliene	C ₁₅ H ₂₄	20.41	20.50	0.09	0.44	20.79	0.38	1.86
24	α -Gurjunene	C ₁₅ H ₂₄	20.78	21.11	0.33	1.59	21.03	0.25	1.20
25*	Benzene, 1,3,5-tris(1-methylethyl)-	C ₁₅ H ₂₄	20.88	22.34	1.46	6.99	21.57	0.69	3.30
26	3-Allyl-6-methoxyphenol	C ₁₀ H ₁₂ O ₂	20.99	20.74	-0.25	-1.19	21.16	0.17	0.81
27	α -Fenchene	C ₁₀ H ₁₆	21.68	22.86	1.18	5.44	19.29	-2.39	-11.02
28	α -Cubebene	C ₁₅ H ₂₄	22.03	21.75	-0.28	-1.27	23.50	1.47	6.67
29	9-Methyltetracyclo[7.3.1.0(2.7).1(7.11)]tetradecane	C ₁₅ H ₂₄	22.37	22.79	0.42	1.88	23.50	1.13	5.05
30*	Caryophyllene	C ₁₅ H ₂₄	23.78	24.65	0.87	3.66	24.00	0.22	0.93
31	α -Selinene	C ₁₅ H ₂₄	24.09	25.37	1.28	5.31	24.72	0.63	2.62
32	Bicyclosesquiphellandrene	C ₁₅ H ₂₄	24.23	24.78	0.55	2.27	24.97	0.74	3.05
33	α -Bergamotene	C ₁₅ H ₂₄	24.39	24.88	0.49	2.01	25.75	1.36	5.58
34	Ethanone, 1-(2-hydroxy-4-methoxyphenyl)-	C ₉ H ₁₀ O ₃	24.67	25.33	0.66	2.68	22.35	-2.32	-9.40
35*	5,9-Undecadien-2-one, 6,10-dimethyl-, (<i>E</i>)-	C ₁₃ H ₂₂ O	25.11	25.49	0.38	1.51	26.37	1.26	5.02
36	Humulene	C ₁₅ H ₂₄	25.52	26.36	0.84	3.29	23.59	-1.93	-7.56
37	1,6-Dihydroxynaphthalene	C ₁₀ H ₈ O ₂	26.03	25.23	-0.80	-3.07	25.65	-0.38	-1.46
38	Naphthalene, 1,2,4a,5,6,8a-hexahydro-4,7-dimethyl-1-(1-methylethyl)-	C ₁₅ H ₂₄	26.44	26.80	0.36	1.36	26.07	-0.37	-1.40
39	Curcumene	C ₁₅ H ₂₂	26.76	27.84	1.08	4.04	27.43	0.67	2.50
40*	β -Eudesmene	C ₁₅ H ₂₄	27.09	26.32	-0.77	-2.84	27.10	0.01	0.04
41	γ -Muurolene	C ₁₅ H ₂₄	27.21	27.53	0.32	1.18	26.78	-0.43	-1.58
42	γ -Selinene	C ₁₅ H ₂₄	27.37	26.21	-1.16	-4.24	25.48	-1.89	-6.91
43	α -Muurolene	C ₁₅ H ₂₄	27.49	26.80	-0.69	-2.51	26.07	-1.42	-5.17
44	Pentadecane	C ₁₅ H ₃₂	27.70	27.09	-0.61	-2.20	26.84	-0.86	-3.10
45*	β -Bisabolene	C ₁₅ H ₂₄	27.97	28.66	0.69	2.47	28.03	0.06	0.21
46	δ -Cadinene	C ₁₅ H ₂₄	28.33	26.80	-1.53	-5.40	26.13	-2.20	-7.77
47	Calamenene	C ₁₅ H ₂₂	28.45	27.36	-1.09	-3.83	27.52	-0.93	-3.27
48	β -Sesquiphellandrene	C ₁₅ H ₂₄	28.61	28.30	-0.31	-1.08	27.92	-0.69	-2.41
49	α -Calacorene	C ₁₅ H ₂₀	29.19	28.25	-0.94	-3.22	28.68	-0.51	-1.75
50*	Pentadecane, 2-methyl-	C ₁₆ H ₃₄	30.12	29.57	-0.55	-1.83	29.87	-0.25	-0.83
51	N-(7-Methylbenzo(b)thien-3-yl)acetamide	C ₁₁ H ₁₁ NOS	30.61	31.03	0.42	1.37	31.47	0.86	2.81
52	Hexadecane	C ₁₆ H ₃₄	31.35	30.53	-0.82	-2.62	30.23	-1.12	-3.57
53	But-3-enal, 2-methyl-4-(2,6,6-trimethyl-1-cyclohexenyl)-	C ₁₄ H ₂₂ O	31.45	30.52	-0.93	-2.96	29.99	-1.46	-4.64
54	Cedrol	C ₁₅ H ₂₆ O	31.52	30.82	-0.70	-2.22	29.55	-1.97	-6.25
55*	Tricyclo[6.3.0.0(1,5)]undec-2-en-4-one, 2,3,5,9-tetramethyl-	C ₁₅ H ₂₂ O	31.79	30.16	-1.63	-5.13	31.65	-0.14	-0.44

续表1(Continued Tab. 1)

序号 No.	化合物 Compound	分子式 Molecular formula	保留时间 t_R (min)	Cal. 1	Err. 1	Err. 1%	Cal. 2	Err. 2	Err. 2%
56	(+)-Ledene	C ₁₅ H ₂₄	32.04	31.79	-0.25	-0.78	31.20	-0.84	-2.62
57	Hexadecane, 7,9-dimethyl-	C ₁₈ H ₃₈	32.77	34.04	1.27	3.88	34.13	1.36	4.15
58	Hexadecane, 2-methyl-	C ₁₇ H ₃₆	33.27	33.01	-0.26	-0.78	33.23	-0.04	-0.12
59	Patchouli alcohol	C ₁₅ H ₂₆ O	33.38	32.56	-0.82	-2.46	34.26	0.88	2.64
60*	Guaiazulene	C ₁₅ H ₁₈	33.45	32.38	-1.07	-3.20	32.02	-1.43	-4.28
61	Heptadecane	C ₁₇ H ₃₆	34.30	33.97	-0.33	-0.96	33.61	-0.69	-2.01
62	Pentadecane, 2,6,10,14-tetramethyl-	C ₁₉ H ₄₀	34.39	36.11	1.72	5.00	34.56	0.17	0.49
63	Hexadecane, 2,6,10,14-tetramethyl-	C ₂₀ H ₄₂	36.99	38.50	1.51	4.08	39.18	2.19	5.92
64	2-Pentadecanone, 6,10,14-trimethyl-	C ₁₈ H ₃₆ O	37.78	36.84	-0.94	-2.49	38.37	0.59	1.56

注: * 测试集样本。

Note: * Test set sample.

多元线性回归使用的是 SPSS13.0 软件、偏最小二乘回归使用的是 SIMCA-P 11.5 软件。

1.2 实验方法

1.2.1 化合物分子结构表征

构建化合物结构与性质之间的关系模型, 化合物结构参数化表征是关键步骤之一。认为化合物中的非氢原子以及它们之间的关系对化合物色谱保留时间产生影响, 而氢原子影响通常被忽略。不同类型的非氢原子以及不同类型非氢原子之间的关系对化合物色谱保留时间影响不同, 参照文献^[11-13]方法将化合物中的非氢原子按照其连接的其它非氢原子数分为 4 类, 与 k 个其它非氢原子相连的非氢原子属于第 k 类非氢原子, 例如与 2 个其它非氢原子相连的仲碳原子属于第 2 类非氢原子。在参考文献^[14-16]的基础上, 对化合物中的非氢原子进行参数化染色, 见式(1)。

$$Z_i = \{(n_i - 1) \times [(m_i + 2)/(d_i + 2)] \times L_i\}^{1/2} \quad (1)$$

式中 i 为非氢原子在分子中的编码; n_i 为非氢原子 i 的原子核外电子层数; m_i 为价电子数, d_i 为非氢原子 i 的成键电子数, 也就是氧化数; L_i 为非氢原子 i 与相邻非氢原子直接连接的化学键数 (δ 键取值为 1, π 键取值为 0.5)。非氢原子 i 的 n_i 越大, 其半径就越大, 相应的原子体积就越大, 相应的 Z_i 值也就越大。

非氢原子自身对化合物色谱保留时间的影响, 按照式(2)进行分类累加。

$$x_k = \sum_{i \in k} Z_i \quad (k = 1, 2, 3, 4) \quad (2)$$

式中, k 表示非氢原子 i 的原子类型, Z_i 按式

(1)计算。根据非氢原子的分类, 对于一个有机化合物分子中最多含有 4 类非氢原子, 因此最终可得到 4 个非氢原子自身对化合物保留时间的影响项, 用 x_1 、 x_2 、 x_3 和 x_4 表示。

4 类非氢原子之间可以有 10 种不同的组合, 即 m_{11} 、 m_{12} 、 \dots 、 m_{44} , 简写为 x_5 、 x_6 、 \dots 、 x_{14} , m_{13} 表示第 1 类非氢原子与第 3 类非氢原子之间的关系, 以此类推。非氢原子之间的关系并非某种具体的相互作用, 而是要反映出相关程度随着非氢原子间的距离的增大而减小, 随着非氢原子自身某种性质的增大而增大, 式(3)可以满足此要求。

$$x_r = m_{nl} = \sum_{i \in n, j \in l} Z_i \times Z_j \times \exp(-\alpha \times d_{ij}^2) \quad (n = 1, 2, 3, 4; n \leq l \leq 4) \quad (3)$$

Z 按式(1)计算; d_{ij} 为非氢原子 i 、 j 之间的相对距离(即键长之和与碳碳单键键长的比值, 如果 i 、 j 之间有多条路径, 则以最短的为准); n 和 l 为原子所属类型, $\alpha = 0.5$ 。这样一来, 对于一个化合物经参数化表达后最多可得 14 个变量(结构描述符)。

1.2.2 建模与评价

运用多元线性回归(MLR)和偏最小二乘回归(PLS)方法构建化合物结构与色谱保留时间的关系模型, 以方差膨胀因子(VIF)^[17]判断变量间的共线性, $VIF = (1 - r^2)^{-1}$, r 为某自变量与其它自变量之间的相关系数, 并且认为 VIF 小于 10, 变量间共线性不明显, 所建模型可以接受。利用相关系数(R^2)及标准偏差(SD)判断模型拟合效果, $R^2 \geq 0.60$ 、SD 与训练集的数值范围之比 $\leq 10\%$ 表明模型具有良好的拟合能力^[18]; 通过“留一法”交互检验评价模型的稳

定性, $R_{cv}^2 \geq 0.50$ 表明模型具有良好的稳定性^[17]。采用测试样本集相关系数(R_{test}^2)及标准偏差(SD_{test})评价模型的外部预测能力^[19], $R_{test}^2 \geq 0.50$ 、 SD_{test} 与测试集的数值范围(value range, V_r)之比 $\leq 10\%$, 表明模型预测准确性高、预测能力强。

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

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

式(4)和式(5)中, y_i 、 \hat{y}_i 分别测试集样本的实验值和预测值, \bar{y} 为测试集样本实验值的平均值。

2 结果与讨论

2.1 分子结构表征

化合物经分子结构表征后得到 14 个变量, 由于变量数较多, 在此没有全部列出。部分用到的变量列于表 2。

表 2 化合物的结构参数化表征结果

Table 2 Structural parameterized characterization results of compounds

序号 No.	x_1	x_2	x_3	x_4	x_5	x_7	x_8	保留时间 t_R (min)
1	2.000 0	4.102 6	7.205 8	0.000 0	0.355 3	3.033 2	0.000 0	6.60
2	3.224 7	4.122 6	5.334 9	2.000 0	0.636 7	2.716 7	2.466 0	8.35
3	1.732 1	9.286 8	1.870 8	0.000 0	0.000 0	0.729 0	0.000 0	9.23
4	3.224 7	3.242 6	5.334 9	2.000 0	0.136 7	2.218 7	2.466 0	9.72
5	3.000 0	6.224 6	5.473 7	0.000 0	0.136 3	3.129 5	0.000 0	11.16
6	3.224 7	5.423 8	5.473 7	0.000 0	0.213 1	4.443 2	0.000 0	11.31
7	1.732 1	10.901 0	1.870 8	0.000 0	0.000 0	0.065 7	0.000 0	11.76
8	3.224 7	6.324 6	5.612 5	0.000 0	0.188 3	4.437 6	0.000 0	13.20
9	3.732 1	4.142 6	5.334 9	2.000 0	0.137 9	3.951 2	2.496 4	14.70
10	4.732 1	4.142 6	7.067 0	0.000 0	0.139 3	6.624 8	0.000 0	15.20
11	4.732 1	2.228 4	7.205 8	2.000 0	0.171 2	5.147 7	2.451 0	15.40
12	3.732 1	6.624 6	5.612 5	0.000 0	0.347 7	5.900 1	0.000 0	16.05
13	4.414 2	4.576 5	7.483 3	0.000 0	0.004 0	4.746 4	0.000 0	16.13
14	2.224 7	11.319 9	3.741 7	0.000 0	0.000 0	0.370 8	0.000 0	16.44
15	4.000 0	6.743 4	7.344 5	0.000 0	0.135 4	4.215 0	0.000 0	17.33
16	4.732 1	4.242 6	7.344 5	0.000 0	0.139 6	6.940 5	0.000 0	17.60
17	4.956 8	4.409 6	7.344 5	0.000 0	0.248 3	7.630 9	0.000 0	17.77
18	2.732 1	9.986 1	3.741 7	0.000 0	0.000 0	2.697 3	0.000 0	17.91
19	3.414 2	6.656 9	5.196 2	2.000 0	0.135 4	2.391 2	2.427 4	18.35
20	5.732 1	5.242 6	5.034 9	4.000 0	0.505 2	3.383 5	4.456 9	18.90
21	2.000 0	11.486 8	3.741 7	0.000 0	0.000 0	0.363 9	0.000 0	18.96
22	3.000 0	12.228 7	5.012 5	0.000 0	0.000 0	0.064 2	0.000 0	20.15
23	4.000 0	7.652 2	3.464 1	6.121 3	0.136 6	1.092 8	5.062 3	20.41
24	4.000 0	5.056 9	10.669 9	2.000 0	0.135 4	2.008 9	2.027 2	20.78
25	6.000 0	4.043 4	10.808 6	0.000 0	0.406 0	7.838 7	0.000 0	20.88
26	3.639 0	9.738 8	5.612 5	0.000 0	0.001 6	2.689 4	0.000 0	20.99
27	3.224 7	6.042 6	5.334 9	2.000 0	0.175 2	4.986 5	2.852 4	21.68
28	3.000 0	8.952 2	8.799 0	2.000 0	0.135 3	3.929 7	0.022 2	22.03
29	1.000 0	11.313 7	8.660 3	2.000 0	0.000 0	1.305 4	0.022 2	22.37

续表2(Continued Tab. 2)

序号 No.	x_1	x_2	x_3	x_4	x_5	x_7	x_8	保留时间 $t_R(\text{min})$
30	4.224 7	7.652 2	7.205 8	2.000 0	0.135 4	2.616 8	2.427 5	23.78
31	4.224 7	8.652 2	7.205 8	2.000 0	0.213 5	4.934 8	1.235 3	24.09
32	4.224 7	7.038 0	10.669 9	0.000 0	0.135 3	4.456 4	0.000 0	24.23
33	4.000 0	8.819 1	7.205 8	2.000 0	0.135 7	4.129 1	1.235 3	24.39
34	2.914 2	8.157 6	9.483 3	0.000 0	0.000 0	4.070 5	0.000 0	24.67
35	5.732 1	9.819 1	5.612 5	0.000 0	0.482 9	6.906 6	0.0000	25.11
36	4.000 0	10.981 4	3.741 7	2.000 0	0.135 3	2.271 6	2.427 3	25.52
37	2.828 4	9.486 8	7.483 3	0.000 0	0.000 0	4.063 0	0.000 0	26.03
38	4.000 0	7.404 9	10.669 9	0.000 0	0.135 3	5.163 8	0.000 0	26.44
39	4.000 0	10.400 3	7.344 5	0.000 0	0.135 3	4.709 9	0.000 0	26.76
40	4.000 0	8.952 2	7.067 0	2.000 0	0.135 7	4.896 6	1.235 3	27.09
41	4.224 7	7.238 0	10.669 9	0.000 0	0.135 3	5.748 5	0.000 0	27.21
42	4.224 7	8.485 3	7.344 5	2.000 0	0.136 0	5.118 0	1.252 9	27.37
43	4.000 0	7.404 9	10.669 9	0.000 0	0.135 3	5.163 8	0.000 0	27.49
44	2.000 0	18.384 8	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	27.70
45	4.224 7	10.233 3	7.344 5	0.000 0	0.135 3	5.343 6	0.000 0	27.97
46	4.000 0	7.238 0	10.808 6	0.000 0	0.135 3	5.264 1	0.000 0	28.33
47	4.000 0	7.571 8	10.808 6	0.000 0	0.135 3	5.120 2	0.000 0	28.45
48	4.224 7	10.400 3	7.205 8	0.000 0	0.135 3	5.036 2	0.000 0	28.61
49	4.000 0	7.738 8	10.947 4	0.000 0	0.135 3	5.266 8	0.000 0	29.19
50	3.000 0	16.970 6	1.732 1	0.000 0	0.135 3	2.101 1	0.000 0	30.12
51	3.732 1	10.826 3	9.354 1	0.000 0	0.347 6	5.099 8	0.000 0	30.61
52	2.000 0	19.799 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	31.35
53	5.732 1	9.386 1	5.473 7	2.000 0	0.171 4	4.856 6	2.455 8	31.45
54	5.414 2	9.485 3	5.196 2	4.000 0	0.355 6	4.575 0	3.808 4	31.52
55	5.732 1	5.656 9	9.076 6	4.000 0	0.079 3	7.771 9	2.493 1	31.79
56	4.224 7	5.956 9	10.531 1	2.000 0	0.135 3	5.927 4	2.427 0	32.04
57	4.000 0	16.970 6	3.464 1	0.000 0	0.000 3	2.139 6	0.000 0	32.77
58	3.000 0	18.384 8	1.732 1	0.000 0	0.135 3	2.101 1	0.000 0	33.27
59	5.414 2	7.071 1	5.196 2	6.000 0	0.263 1	2.196 3	7.106 3	33.38
60	4.000 0	8.905 7	11.086 2	0.000 0	0.135 4	5.873 2	0.000 0	33.45
61	2.000 0	21.213 2	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	34.30
62	4.000 0	14.042 1	6.928 2	0.000 0	0.000 3	4.140 7	0.000 0	34.39
63	6.000 0	14.142 1	6.928 2	0.000 0	0.146 4	5.487 2	0.000 0	36.99
64	6.732 1	12.727 9	7.067 0	0.000 0	0.482 9	7.704 6	0.000 0	37.78

2.2 化合物结构与色谱保留关系

2.2.1 多元线性回归模型

本研究的训练集样本数为 52 个,而变量数多达

14 个,不符合 $N/n \geq 5$ 的经验规则,因而在确定最优模型之前应该对变量进行筛选,将与化合物色谱保留时间(t_R)相关性不大的变量进行剔除。逐步回归

(SMR) 是筛选变量的常用方法, 建模前运用逐步回归依据变量显著性大小顺序逐步引入候选变量, 根据每一步所得模型的相关系数 (R^2) 及标准偏差 (SD) 进行综合考量后选出最佳变量组合进行回归建模。对选定的模型变量间的共线性进行评价, 计算方差膨胀因子 (VIF), 如果某变量的方差膨胀因子 (VIF) 大于或等于 20, 则应减少变量建模。在逐步回归中, 相关系数 (R^2) 和标准偏差 (SD) 随变量的引入而发生变化, 为便于直观分析将其变化情况绘图于图 1 及图 2 中。

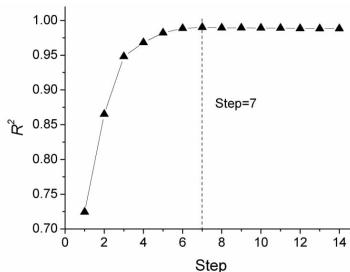


图 1 R^2 在逐步回归中的变化情况

Fig. 1 Changes of R^2 in stepwise regression

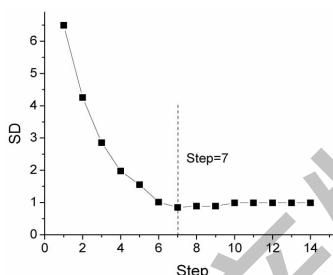


图 2 SD 在逐步回归中的变化情况

Fig. 2 Changes of SD in stepwise regression

从图 1 中可以发现, 相关系数 (R^2) 随着变量的引入而增大, 起初增大的趋势明显, 当逐步回归进行到第 7 步时相关系数 (R^2) 接近最大值, 之后相关系数 (R^2) 增大趋势放缓。同样, 图 2 中可以发现, 标准偏差 (SD) 随着变量的引入而逐渐减小, 起初减小的趋势明显, 当逐步回归进行到第 7 步时标准偏差 (SD) 达到最小值, 之后标准偏差 (SD) 略有增大。对 7 变量模型进行回归诊断, 发现变量 x_8 的方差膨胀因子 (VIF) 最大, 且值仅为 13.045 2, 变量之间没有明显的共线性。当变量继续引入, 逐步回归进行到第 8 步, 虽然相关系数 (R^2) 略有增大, 然而此时最大变量方差膨胀因子 (VIF) 远大于 20, 说明 8 变量模型不可靠。综合各方面的考量后, 认为应该选

择逐步回归第 7 步所得变量组合 ($x_1, x_2, x_3, x_4, x_5, x_7, x_8$, 列入表 2) 进行建模 (M1), 如式(6)。

$$\begin{aligned} t_R = & -18.6293 + 0.4794 \times x_1 + 2.4344 \times x_2 + \\ & 1.6406 \times x_3 - 1.6015 \times x_4 - 8.0066 \times x_5 + 1.7558 \times x_7 + \\ & 4.3227 \times x_8 \quad (6) \end{aligned}$$

$$\begin{aligned} N = 52, R_1^2 = 0.9901, R_{CV1}^2 = 0.9862, SD_1 = \\ 0.8483, F_1 = 629.8538; R_{test1}^2 = 0.9809, SD_{test1} = \\ 0.9623 \end{aligned}$$

模型 (M1) 的相关系数 (R_1^2) 高达 0.9901, 远大于 0.60 的标准; 标准偏差 (SD_1) 为 0.8483, 52 个训练集样本色谱保留时间值范围为 $37.78 - 6.60 = 31.18$, $0.8483 / 31.18 = 2.72\%$, 远小于 10% 的标准, 说明模型具有良好的拟合能力。交互检验的相关系数 (R_{CV1}^2) 为 0.9862, 远大于 0.50 的标准, 说明模型具有良好的稳定性。测试样本集相关系数 (R_{test1}^2) 为 0.9809, 远大于 0.50 的标准; 测试集样本标准偏差 (SD_{test1}) 为 0.9623, 测试集样本色谱保留时间值范围为 $33.45 - 11.16 = 22.29$, $0.9623 / 22.29 = 4.32\%$, 远小于 10% 的标准, 表明模型预测能力强, 预测准确性高。

2.2.2 偏最小二乘回归模型

偏最小二乘回归 (PLS) 也是化合物结构和性质关系研究中常用的方法之一, 特别适合于样本数较少而变量数较多的情况下建模。变量 $x_1, x_2, x_3, \dots, x_{14}$ 作为自变量 X , 化合物色谱保留时间 (t_R) 作为因变量 Y , 建立偏最小二乘回归模型 (M2)。相关系数 (R^2) 及交互检验的相关系数 (R_{CV}^2) 随主成分数 (A) 的变化情况见图 3。图 3 中可以看出当主成分数 (A) 达到 6 个时, 相关系数 (R^2) 接近最大值, 交互检验的相关系数 (R_{CV}^2) 达到最大值, 应该选择 6 个

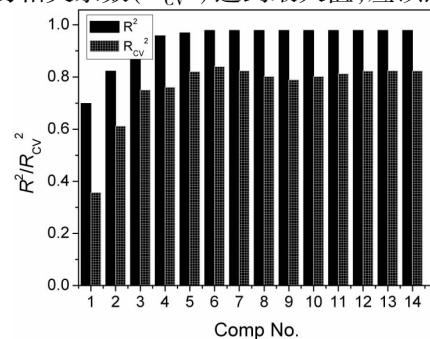


图 3 相关系数 (R^2/R_{CV}^2) 随主成分数的变化情况

Fig. 3 Correlation coefficient (R^2/R_{CV}^2) change with the number of principal components

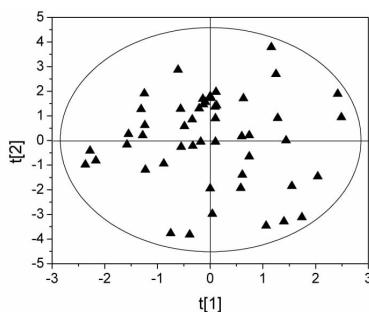


图 4 样本在前 2 个主成分得分分布

Fig. 4 Distribution of the top 2 principal component scores of the sample

主成分进行建模(M2)。52 个训练集样本在 PLS 前 2 个主成分得分空间散点分布绘于图 4, 图 4 中可以发现全部样本得分点都落在 95% 置信度的椭圆置信圈内, 没有出现一个异常点, 反映出构建的结构描述符能较好地表现挥发性有机化合物的分子结构特征, 并在统计模型中得到正确的表现。

此时所建 PLS 模型(M2)的相关系数(R^2)为 0.978 6, 远大于 0.60 的标准; 标准偏差(SD_2)为 $1.155 2, 1.155 2/31.18 = 3.70\%$, 远小于 10% 的标准, 说明模型具有良好的拟合能力。交互检验的相关系数(R_{CV}^2)为 0.840 4, 远大于 0.50 的标准, 说明模型具有良好的稳定性。测试样本集相关系数(R_{test2}^2)为 0.988 4, 远大于 0.50 的标准; 测试集样本标准偏差(SD_{test2})为 0.750 3, 测试集样本色谱保留时间值范围为 $22.29, 0.750 3/22.29 = 3.37\%$, 远小于 10% 的标准, 表明模型预测能力强, 预测准确性高。

为验证模型结果是否为偶然所得, 对模型进行 20 次 Y 向量随机排序验证。以 Y 原始向量和经过随机排序的 Y 向量相关系数对模型的 R^2 和 R_{CV}^2 作图于图 5。根据 Andersson 等^[20]提出的判断标准, R^2 和 R_{CV}^2 在纵轴上的截距分别不得超过 0.300 和

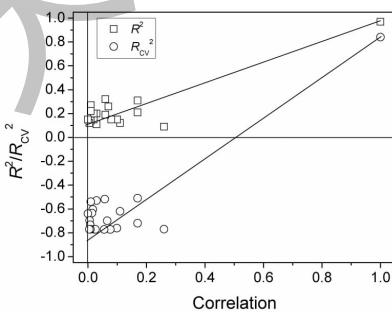


图 5 Y 向量随机排序验证结果

Fig. 5 Y vector random sorting verification results

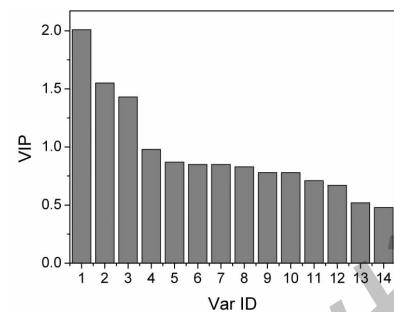


图 6 变量重要性投影图

Fig. 6 Projection of variable importance

0.050。从图 5 中可以发现所建 PLS 模型的 R^2 和 R_{CV}^2 的截距分别为 0.105 和 -0.882, 因而认为模型的优良结果并非偶然, 偏最小二乘回归模型(M2)可以用于分析化合物结构对色谱保留时间的影响。

变量重要性可以反映出变量与 Y 之间的相关程度, 通常认为变量重要性投影(VIP)值大于 1 的变量与化合物色谱保留时间(t_R)相关性大。变量重要性投影见图 6, 图 6 可以看出变量 x_2, x_1, x_9 这 3 个变量的 VIP 值大于 1, 说明这 3 个变量与化合物色谱保留时间(t_R)相关性大。 x_1, x_2 分别为第 1、2 类非氢原子自身对化合物色谱保留时间的影响, x_9 为第 2 类非氢原子之间的关系对化合物色谱保留时间的影响。第 1 类非氢原子对应取代基的末端原子, 而第 2 类非氢原子的多少决定了链的长度, 说明化合物支链越多、链越长, 即伯、仲碳原子越多, 化合物可能具有较大的色谱保留时间(t_R)值, 这与表 1 中的数据特征基本吻合。

2.2.3 模型结果与比较

多元线性回归(MLR)模型(M1)和偏最小二乘回归(PLS)模型(M2)对训练集化合物的色谱保留时间进行了计算, 对不参与建模的预测集化合物色谱保留时间进行了预测, M1 和 M2 对化合物的计算值及预测值分别列于表 1 的 Cal. 1 和 Cal. 2 列, Err. 1 和 Err. 2 分别为误差, Err. 1%、Err. 2% 分别为百分误差。两模型对化合物保留时间的计算值与实验值相关性, 见图 7。图 7 中容易看出所有样本点都落在正方形 45° 对角线附近, 说明两模型对化合物保留时间计算值与实验值相关性好, 两者间误差不大。另外, 图上可以看出 Cal. 1 的样本点与 Cal. 2 的样本点分布大体相似, 说明模型(M1)与模型(M2)的质量大体相当。

模型对化合物保留时间计算值的误差可以反映

模型预测的准确性,两模型计算误差分布见图8。图8中可以发现大部分样本的预测误差值都处于2倍标准偏差($\pm 2SD$)范围以内,对于模型(M1)、模型(M2)均仅有2个化合物(3.13%)超出 $\pm 2SD_1$ 、 $\pm 2SD_2$,说明两模型对化合物保留时间预测较为准确,误差都处于可以接受的范围内。另外,绝大多数样本的Err. 1 小于 Err. 2, $\pm 2SD_1$ 线处于 $\pm 2SD_2$ 线内侧并更加靠近中的“0”线,说明模型(M1)的质量略优于模型(M2)。

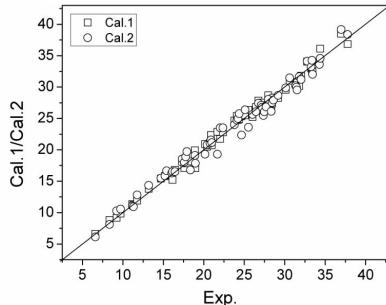


图7 模型预测值与实验值的相关图

Fig. 7 Correlation diagram between model predicted values and experimental values

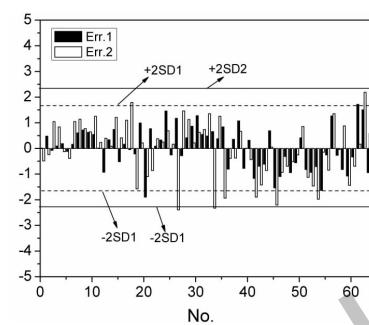


图8 模型对样本保留时间预测误差

Fig. 8 Model prediction error of samples' t_R

本文所研究的化合物结构跨度大,包括了33种烃类、9种酮类、8种芳香族类、4种醇类、3种醛类、3种醚类、2种酚类和2种酸类等类型化合物,含有氧、硫、氮等杂原子,含有单键、双键、三元环、四元环、五元环、六元环、七元环、八元环、十一元环等结构,说明构建的结构描述符普适性强。对于复杂样本体系的定量结构与保留时间关系研究已有一些报道,为便于对比,现将本文结果与部分文献结果列于表3,可以发现本文所取得的结果明显优于文献。

表3 模型比较

Table 3 Comparisons among different QSRR models

No.	结构描述符 Descriptor	训练集 样本数 <i>N</i>	入选变量数/ 主成分数 <i>m/A</i>	建模方法 Method	R^2	<i>R</i>	R_{test}^2	V_r	SD/V_r	<i>F</i>
1 ^[21]	MEDV ^R	55	8	MLR	0.978	0.989	-	34.28	3.98%	-
2 ^[21]	MEDV ^R	55	4	PLS	0.974	0.987	-	34.28	4.08%	-
3 ^[22]	I-MEDV	37	9	MLR	0.960	0.980	-	23.34	4.25%	74.389
4 ^[22]	I-MEDV	37	4	MLR	0.951	0.975	-	23.34	4.42%	152.688
5 ^[23]	MEDV	46	10	MLR	0.821	0.906	-	38.339	-	-
6 ^[23]	MEDV	46	6	MLR	0.815	0.903	-	38.339	-	-
7*	MVVR	52	7	MLR	0.990 1	0.995 0	0.980 9	31.18	2.72%	629.854
8*	MVVR	52	6	PLS	0.978 6	0.989 2	0.988 4	31.18	3.70%	-

注:“*”表示本文所取得的结果;“-”表示无法获取。

Note: “*” indicates the results obtained in this article; “-” indicates that it cannot be obtained.

3 结论

通过构建非氢原子之间的关系得到新的结构描述符,对灯盏花中的64种挥发性有机化合物结构进行了参数化表征,运用多元线性回归(MLR)和偏最小二乘回归(PLS)方法构建了化合物结构与色谱保留时间的关系模型。经检验模型具有通良好的拟合能力、稳定性和外部预测能力。分析了影响化合物色谱保留时间的结构因素,结果表明化合物的伯、仲碳原子越多,色谱保留时间(t_R)值就越大。构建的

化合物结构描述符为二维结构描述符,计算简便、快速、易懂、通用性强,对于化合物结构与性质关系研究具有一定的参考价值。

参考文献

- Liu H, Yang XL, Xu HB. Advances in studies on *Erigeron breviscapus*[J]. Chin Tradit Herb Drugs (中草药), 2001, 33:566-568.
- Zhang HX, Yang HJ, Bao GL, et al. Research progress in pharmacodynamics of breviscapine[J]. Yunnan J Tradit Chin

- Med Mater Med(云南中医中药杂志),2016,37(2):75-78.
- 3 Xv H,Gao XG,Li L,et al. Analysis on chemical components of essential oil from *Erigeron breviscapus* by SPME/GCMS [J]. Contemp Chem Ind(当代化工),2019,48:1354-1357.
- 4 Huang YF,Cai Z,Wu JZ,et al. Quantitative structure-retention relationship assisted gas chromatography-mass spectroscopy/gas chromatography-infrared spectroscopy for analysis of aldehydes,ketones and esters in fragrances[J]. Chin J Anal Chem(分析化学),2015,43:1558-1564.
- 5 He Q,Huang BJ,Chen JR,et al. Quantitative structure retention relationship study on aroma components of blackberry wine using neural network[J]. J Anal Sci(分析科学学报),2016,32:697-700.
- 6 He Q,Zhang D,Zhu L. Quantitative structure retention relationship study on the fragrance compounds of *Lilium* spp. using neural network[J]. J Xuchang Univ(许昌学院学报),2018,37(2):39-43.
- 7 Zhang XT,Shi LH,Song LJ,et al. QSRR models to predict retention indices of organic sulfur compounds in fuel oil on different GC columns[J]. Pet Process Petrochem(石油炼制与化工),2017,48(8):94-99.
- 8 Jing JH,Zhang QX,Li ZL. Study on QSRR of fatty acids by 3D-Hovaif[J]. Chem Anal Meterage(化学分析计量),2008,17(2):13-15.
- 9 Liu FP,Liu WQ,Xie WL,et al. Research the effect of the column polarity on QSRR model for saturated alcohol compounds[J]. J Hunan Univ Sci Technol; Nat Sci(湖南科技大学学报:自然科学版),2017,32(1):79-84.
- 10 Ren C,Yu DY,Wei L,et al. Three-dimensional quantitative structure-activity relationship studies on coumarins as agonists of G protein-coupled receptor 35[J]. Nat Prod Res Dev(天然产物研究与开发),2018,30:1066-1072.
- 11 Li JF. Study on acute toxicity for halogenated phenols by using molecular vertex electronegativity interaction vector[J]. Comput Appl Chem(计算机与应用化学),2015,32:1399-1403.
- 12 Liao LM,Li JF,Lei GD. Structural characterization and chromatographic retention index prediction for aroma components of lemon peels[J]. Nat Prod Res Dev(天然产物研究与开发),2016,28:90-95.
- 13 Liao LM,Huang X,Lei GD. Structural characterization and octanol/water partition coefficient(LogP) prediction for oxygen-containing organic compounds[J]. Chin J Struct Chem,2017,36:1243-1250.
- 14 Qin ZL. A new connectivity index for QSPR/QSAR study of alcohol[J]. J Xuzhou Normal Univ:Nat Sci(徐州师范大学学报:自然科学版),2001,19(3):50-52.
- 15 Du XH. Predicting the $\lg K_{ow}$ of PCDDs using novel topological parameter[J]. J Wuhan Univ Technol(武汉理工大学学报),2007,29(1):40-44.
- 16 Chen Y. Prediction of aqueous solubility, hydrophobic parameter for esters and ketones with novel valence connectivity index[J]. J Nanjing Univ Technol; Nat Sci(南京工业大学学报:自然科学版),2005,27(4):41-43.
- 17 Xu Q,Fan LL,Xu J. A Simple 2D-QSPR model for the prediction of setschenow constants of organic compounds[J]. Maced J Chem Chem En,2016,35(1):53-62.
- 18 Sung-sun S,Karplus M. A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors[J]. J Comput Aid Mol Des,1999,13:243-258.
- 19 Gramatica P,Pilotti P,Papa E. A tool for the assessment of VOC degradability by tropospheric oxidants starting from chemical structure[J]. J Chem Inf Comput Sci,2004,44:1794-1802.
- 20 Andersson PM,Sjstrom M,Lundstedt T. Preprocessing peptides sequences for multivariate sequence-property analysis[J]. Chemom Intell Lab Syst,1998,42(1-2):41-50.
- 21 Liao LM,Li JF,Qing DH,et al. Structural characterization and retention time prediction for components of essential oil of meconopsis integrifolia flowers[J]. Chin J Struct Chem,2010,29:1638-1645.
- 22 Liao LM,Zhu J,Li JF,et al. QSRR study on components of *styrax japonicus* sieb flowers using improved molecular electronegativity-distance vector(I-MEDV)[J]. Chin J Struct Chem,2011,30(1):105-110.
- 23 Zhu WP,Mei H,Shu M,et al. Structural characterization of some components from essential oils of *rosa banksiae* ait for estimation and prediction of their linear retention index and retention times[J]. Chin J Chin Mater Med(中国中药杂志),2008,33:609-611.