DOI:10.7524/j.issn.0254-6108.2018062001
郑珊珊, 李田田, 王晶,等.芳香族化合物与水合电子水相反应速率常数的 QSAR 模型研究[J].环境化学,2019,38(5):1005-1013.
ZHENG Shanshan, LI Tiantian, WANG Jing, et al. QSAR models for predicting the aqueous reaction rate constants of aromatic compounds with hydrated electrons[J].Environmental Chemistry,2019,38(5):1005-1013.

芳香族化合物与水合电子水相反应速率常数的 QSAR 模型研究*

郑珊珊 李田田 王 晶 胡雅君 张会霞 赵淑娴 赵元慧 李 超**

(东北师范大学环境学院,国家环境保护湿地生态与植被恢复重点实验室, 长春, 130117)

摘 要 基于水合电子(e^-_{aq})的高级还原技术(ARPs)是一种有效地去除水中痕量有机污染物的技术.有机物与水合电子反应的速率常数($k_{e^-_{aq}}$)是评价高级还原处理体系中有机物去除效率的一个重要参数.然而化学品种类繁多,通过实验方法一一获取 $k_{e^-_{aq}}$ 是不现实的.因此,需要建立一种能够快速预测 $k_{e^-_{aq}}$ 的方法,填补现有数据的缺失.本研究搜集整理了 94 种芳香族化合物的 $k_{e^-_{aq}}$ 实验值,采用逐步多元线性回归(MLR)和支持向量机(SVM)方法分别构建了预测化合物的 $k_{e^-_{aq}}$ 线性和非线性定量结构活性关系(QSAR)模型.两个模型均具有良好的拟合度($R^2_{adj,tr}>0.800$)、稳健性($Q^2_{LOO}=0.782$)和预测能力($Q^2_{ext}>0.790$).机理分析表明,最低未占据分子轨道能(E_{LUMO})和偶极矩加权的来自扩增边缘邻接处的最大特征值(SpMAD_AEA(dm))是影响有机物同 e^-_{aq} 反应活性的重要参数,此外,$k_{e^-_{aq}}$ 还与化合物的键级和极化率有关.元分析结果表明具有吸电子官能团的芳香化合物比含供电子基团的化合物倾向于与 e^-_{aq} 有更高的反应活性.

关键词 芳香族化合物, 水合电子(e^-_{aq}), 反应速率常数, 定量结构活性关系(QSAR).

QSAR models for predicting the aqueous reaction rate constants of aromatic compounds with hydrated electrons

*ZHENG Shanshan LI Tiantian WANG Jing HU Yajun ZHANG Huixia ZHAO Shuxian ZHAO Yuanhui LI Chao***

(State Key Laboratory of Wetland Ecology and Vegetation Restoration, School of Environment, Northeast Normal University, Changchun, 130117, China)

Abstract: Hydrated electron (e^-_{aq})-based reduction processes are promising for degrading organic pollutants in water engineering systems. The second order rate constant ($k_{e^-_{aq}}$) of e^-_{aq} with organic compounds is an important parameter for evaluating the removal efficiency of organic pollutants in advanced reduction proless. However, experimental determination of $k_{e^-_{aq}}$ seems fairly unrealistic because of the large number of organic chemicals. Thus, it is necessary to develop an effective method to predict $k_{e^-_{aq}}$. In this study, the experimental $k_{e^-_{aq}}$ values of 94 aromatic compounds were collected. The quantitative structure-activity relationship (QSAR) models for predicting $k_{e^-_{aq}}$ were constructed by stepwise multiple linear regression (MLR) and support vector machines (SVM)

2018 年 6 月 20 日收稿(Received: June 20, 2018).
* 国家自然科学基金(21607022)和吉林省科学技术发展项目(20180520078JH)资助.
Supported by the National Natural Science Foundation of China (21607022) and Jilin Province Science and Technology Development Project (20180520078JH).
* * **通讯联系人**,Tel: 0431-89165610, E-mail: lic932@nenu.edu.cn
Corresponding author, Tel: 0431-89165610, E-mail: lic932@nenu.edu.cn

methods, respectively. Both two models had satisfactory goodness-of-fit ($R^2_{adj,tr}$ > 0.800), robustness (Q^2_{LOO} = 0.782) and good predictability (Q^2_{ext} > 0.790). Mechanistic analysis revealed that the energy of the lowest unoccupied molecular orbital (E_{LUMO}) and the spectral mean absolute deviation from augmented edge adjacency mat weighted by dipole moment (SpMAD_AEA(dm)) were the most important descriptors. Additionally, $k_{e^-_{aq}}$ was found to be related to the bond order and polarizability of the compounds. The meta-analysis showed that aromatic compounds with electron-withdrawing functional groups tended to have higher reactivity with e^-_{aq} than those containing electron-donating groups.

Keywords: aromatic compounds, hydrated electron (e^-_{aq}), reaction rate constants, quantitative structure-activity relationship (QSAR).

高级还原技术(ARPs)因在去除难降解和持久性有机物上颇有成效而引起了人们的广泛关注.在高级还原体系中,利用活化技术与还原剂相结合产生还原性自由基达到高效降解目标有机污染物的目的[1-2].目前,被证实产生的还原性自由基有水合电子(e^-_{aq})、氢原子和亚硫酸根自由基等[3-5].其中,e^-_{aq}是一种极强的还原性物质,其氧化还原电位低至-2.9 V.e^-_{aq}与有机物的反应机理主要为电子转移,它与难生物降解的卤代有机物反应能提高其脱卤效率,因此在水处理领域有着广泛的应用前景[6-8].

有机化合物与e^-_{aq}的二级反应速率常数($k_{e^-_{aq}}$)是衡量有机化合物反应活性的重要参数,可用于评价水中有机物的去除效率.鉴于此,获取$k_{e^-_{aq}}$不仅对于污染物的预防和控制具有重要意义,还可以为该技术的实用可行性评估提供科学依据.然而,有机化学品的数量及种类众多,单纯地通过实验方法逐一获取有机物的$k_{e^-_{aq}}$几乎是不现实的,不但需要消耗大量的人力、物力,且在时间上滞后于污染预防的需求[9].因此,亟需发展一种快捷高效的预测方法.定量结构活性关系(QSAR)技术反映和揭示有机污染物的分子结构与其"活性"(理化性质、环境行为参数或毒理学参数)之间的内在联系[10-11],具有弥补测试数据缺失、降低测试费用等优点,在有机化学品的生态风险评价领域得到了越来越广泛的重视和应用[12].作为一种可信的技术工具,QSAR 技术已被成功应用于预测有机物各类反应动力学速率常数(例如与·OH, NO_3·, ClO_2和SO_4·$^-$反应)[13-17].然而,目前有关有机物的$k_{e^-_{aq}}$值的 QSAR 模型尚未见报道.此外,相比于已有大量的研究围绕有机物与氧化活性物质(如·OH 和SO_4·$^-$)的反应速率常数开展,目前人们对于有机物与还原性物质这方面的研究甚少.例如,目前人们对具有何种分子结构的化合物与e^-_{aq}有较高的反应活性没有深入的认识.因此,本文的另一目的是揭示分子结构(供电子和吸电子官能团)对有机物与e^-_{aq}反应活性的影响.

本研究搜集整理了 94 个芳香族化合物的水相$k_{e^-_{aq}}$实验值,采用逐步多元线性回归(MLR)和支持向量机(SVM)方法,依据经济合作与发展组织(OECD)有关 QSAR 模型发展和验证导则,构建了用于预测含有不同取代基的芳香族化合物的$k_{e^-_{aq}}$的线性和非线性模型,采用 Williams 法表征了模型的应用域,并采用元分析方法分析分子结构特征对有机物与e^-_{aq}反应活性的影响.

1 实验方法(Materials and methods)

1.1 数据的获取与整理

通过查阅文献和数据库 NDRL/NIST (http://kinetics.nist.gov/solution/),获得了 94 个芳香族化合物(涵盖—COOH、—COO^-、—CH_3、—NH_2、—X (F、Cl、Br、I)、—NO_2、—OH/O^-、—SO_3H、—CN 等不同官能团的化合物)的$k_{e^-_{aq}}$值($M^{-1}s^{-1}$,298 K) (表 1).所有的$k_{e^-_{aq}}$值都经对数转换为$\lg k_{e^-_{aq}}$.图 1 是数据集的分布图,$\lg k_{e^-_{aq}}$值涵盖的范围为 6.95—10.65.此外,本文还考虑了含有可解离基团的化合物的形态,将具有不同形态(离子态和分子态)的化合物均收录到数据表中并用于构建模型.将数据集按照 4:1 的比例随机拆分为训练集和验证集.训练集用于构建模型,验证集用于对所构建的模型进行验证.

1.2 分子结构描述符的计算

在模型的建立过程中考虑了量子化学和 DRAGON 两类分子结构描述符.使用 Gaussian 09 软件中的

B3LYP 方法对有机物结构进行优化, 采用基于自洽反应场的极性连续介质模型(IEFPCM)模拟水的效应.对于 H、C、N、O、F、Cl 和 Br 原子,采用 6-311+G(d,p)基组进行优化,对于 I 原子,采用 LANA2DZ 基组进行优化.基于优化好的结构进行频率计算,确保没有虚频.之后从高斯结果文件里提取量子化学描述符,包括最高占据分子轨道能(E_{HOMO})、最低非占据轨道能(E_{LUMO})、前线分子轨道能极差($E_{LUMO}-E_{HOMO}$)、电离势(IP)、电子亲和势(EA)、分子体积(V)、电子能(E_{ee})、硬度(η)、软度(S)、偶极矩(μ)、极化率(a)及各种原子电荷.基于以上优化好的结构,采用 Dragon 计算 Dragon 描述符.按照以下原则对描述符进行剔除:①所有缺失的描述符数值;②常数项和近似常数项的描述符.

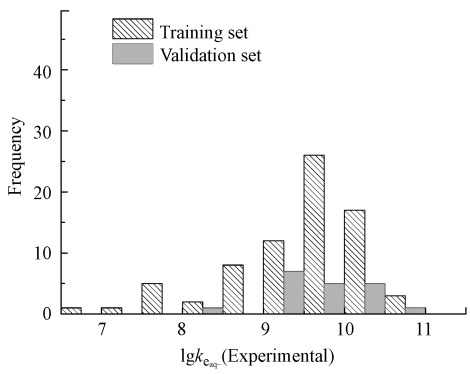

图 1　$lgk_{e^-_{aq}}$ 实验数据分布图

Fig.1　Distribution of experimental $lgk_{e^-_{aq}}$ in the data set

表 1　芳香类化合物的 CAS、名字、价态、$lgk_{e^-_{aq}}$ 实验值、预测值及残差值

Table 1　CAS, names, charge, experimental and predicted $logk_{e^-_{aq}}$ and standardized residual of aromatic chemicals

序号 Number	CAS	化合物名字 Name	价态 Charge	$lgk_{e^-_{aq}}$ (Exp.)	$lgk_{e^-_{aq}}$ (MLR)	$lgk_{e^-_{aq}}$ (SVM)	标准残差 Standardized Residual (MLR)	标准残差 Standardized Residual (SVM)
训练集 Training set								
1	62-53-3	Aniline	0	7.52	7.61	7.35	−0.25	0.51
2	65-85-0	Benzoic Acid	0	9.85	9.98	10.10	−0.33	−0.75
3	69-72-7	2-Hydroxybenzoic Acid	−2	9.51	9.03	9.14	1.27	1.09
4	71-43-2	Benzene	0	6.95	7.00	6.65	−0.12	0.92
5	83-42-1	2-Chloro-6-Nitrotoluene	0	10.53	10.57	10.27	−0.11	0.80
6	88-67-5	2-Iodobenzoic Acid	−1	9.66	9.77	9.84	−0.29	−0.53
7	88-75-5	2-Nitrophenolate	−1	10.30	10.06	9.99	0.63	0.93
8	89-56-5	5-Methyl-Salicylic Acid	−1	9.67	8.92	9.37	2.01	0.92
9	89-98-5	2-Chlorobenzaldehyde	0	10.34	10.31	10.35	0.08	−0.03
10	95-50-1	1,2-Dichlorobenzene	0	9.67	9.39	9.53	0.75	0.44
11	98-07-7	Benzotrichloride	0	9.92	9.61	9.96	0.81	−0.12
12	98-08-8	Benzotrifluoride	0	9.26	9.38	9.56	−0.34	−0.92
13	98-11-3	Benzenesulfonic Acid	−1	9.60	10.07	9.91	−1.25	−0.93
14	99-04-7	3-Methylbenzoic Acid	−1	9.41	9.19	9.30	0.61	0.33
15	99-06-9	3-Hydroxybenzoic Acid	−2	9.04	9.11	9.17	−0.17	−0.40
16	99-94-5	4-Methylbenzoic Acid	−1	9.48	9.06	9.17	1.10	0.93
17	99-96-7	4-Hydroxybenzoic Acid	−2	8.60	8.90	8.91	−0.79	−0.93
18	100-01-6	4-Nitroaniline	0	9.26	9.92	10.17	−1.76	−2.75
19	100-02-7	4-Nitrophenol	−1	10.40	9.71	9.91	1.83	1.48
20	100-21-0	Terephthalic acid	−2	9.86	9.71	9.79	0.41	0.22
21	100-39-0	Benzyl bromide	0	9.18	9.98	10.05	−2.13	−2.64
22	100-46-9	Benzenemethanamine	0	7.90	8.28	8.19	−0.99	−0.86

续表1

序号 Number	CAS	化合物名字 Name	价态 Charge	$\lg k_{e^-_{aq}}$ (Exp.)	$\lg k_{e^-_{aq}}$ (MLR)	$\lg k_{e^-_{aq}}$ (SVM)	标准残差 Standardized Residual (MLR)	标准残差 Standardized Residual (SVM)
23	100-52-7	Benzaldehyde	0	10.65	10.16	10.35	1.32	0.93
24	104-85-8	4-Methylbenzonitrile	0	10.11	9.56	9.81	1.47	0.92
25	106-39-8	1-Bromo-4-chlorobenzene	0	9.62	9.59	9.72	0.09	−0.28
26	106-41-2	4-Bromophenol	0	9.80	9.40	9.47	1.07	1.00
27	106-41-2	4-Bromophenolate	−1	9.46	9.21	9.16	0.67	0.92
28	106-43-4	4-Chlorotoluene3-Chlorophenol	0	8.65	8.73	8.75	−0.20	−0.30
29	106-44-5	4-Methylphenol	0	7.62	8.25	8.13	−1.67	−1.52
30	106-46-7	1,4-Dichlorobenzene	0	9.70	9.17	9.39	1.39	0.92
31	106-47-8	4-Chloroaniline	0	8.72	8.65	8.69	0.18	0.07
32	106-48-9	4-Chlorophenol	0	9.18	8.99	9.02	0.49	0.48
33	106-48-9	4-Chlorophenolate	−1	8.81	8.66	8.50	0.39	0.92
34	106-50-3	1,4-Benzenediamine	0	7.98	8.07	7.93	−0.24	0.15
35	108-36-1	1,3-Dibromobenzene	0	10.00	10.28	10.14	−0.73	−0.41
36	108-37-2	1-Bromo-3-chlorobenzene	0	9.71	10.03	10.01	−0.84	−0.92
37	108-43-0	3-Chlorophenol	0	9.28	9.40	9.49	−0.31	−0.65
38	108-43-0	3-Chlorophenolate	−1	8.70	9.03	9.16	−0.87	−1.39
39	108-86-1	Bromobenzene	0	9.58	9.14	9.27	1.17	0.93
40	108-88-3	Toluene	0	7.04	7.62	7.35	−1.54	−0.93
41	108-90-7	Chlorobenzene	0	8.70	8.70	8.70	0.00	−0.01
42	118-90-1	o-Toluic Acid	−1	9.48	9.06	9.17	1.12	0.92
43	118-91-2	2-Chlorobenzoic Acid	−1	9.15	9.55	9.59	−1.07	−1.35
44	120-83-2	2,4-Dichlorophenol	0	8.70	9.56	9.66	−2.29	−2.89
45	121-14-2	2,4-Dinitrotoluene	0	10.38	10.16	10.14	0.59	0.74
46	121-57-3	P-aminophenylsulfonic acid	0	9.77	9.91	9.93	−0.36	−0.49
47	121-57-3	P-aminophenosulfonate	−1	8.66	9.05	9.29	−1.02	−1.90
48	121-73-3	1-Chloro-3-Nitrobenzene	0	10.49	10.86	10.62	−0.97	−0.39
49	363-72-4	Pentafluorobenzene	0	10.20	10.32	10.16	−0.31	0.15
50	371-41-5	4-Fluorophenol	−1	8.08	8.36	8.25	−0.74	−0.53
51	372-20-3	3-Fluorophenol	−1	8.30	8.61	8.61	−0.81	−0.92
52	392-56-3	Hexafluorobenzene	0	9.93	9.89	10.15	0.09	−0.66
53	455-38-9	3-Fluorobenzoic Acid	−1	9.83	9.58	9.69	0.66	0.42
54	456-22-4	4-Fluorobenzoic Acid	−1	9.58	9.41	9.48	0.46	0.30
55	462-06-6	Fluorobenzene	0	7.78	8.31	8.08	−1.41	−0.93
56	535-80-8	3-Chlorobenzoic Acid	−1	9.74	9.77	9.77	−0.07	−0.08
57	540-36-3	1,4-Difluorobenzene	0	9.30	8.98	9.13	0.85	0.51
58	541-73-1	1,3-Dichlorobenzene	0	10.00	9.79	9.84	0.57	0.50
59	551-62-2	1,2,3,4-Tetrafluorobenzene	0	10.41	10.39	10.11	0.07	0.93
60	554-84-7	3-Nitrophenol	−1	10.40	10.35	10.51	0.13	−0.33
61	586-76-5	4-Bromobenzoic acid	−1	9.89	9.37	9.54	1.37	1.05
62	591-50-4	Iodobenzene	0	10.08	9.30	9.62	2.05	1.38
63	611-20-1	2-Cyanophenol	−1	9.91	10.00	9.86	−0.23	0.16
64	618-51-9	3-Iodobenzoic Acid	−1	10.11	10.15	9.90	−0.09	0.63
65	619-58-9	4-Iodobenzoic Acid	−1	9.96	9.67	9.82	0.78	0.42
66	619-65-8	4-Cyanobenzoic Acid	−1	10.00	10.37	10.25	−0.97	−0.75
67	695-96-5	2-Bromo-4-chlorophenol	0	9.92	9.89	9.84	0.08	0.23
68	771-60-8	2,3,4,5,6-Pentafluoroaniline	0	9.94	9.97	9.79	−0.06	0.46

续表1

序号 Number	CAS	化合物名字 Name	价态 Charge	$\lg k_{e^-_{aq}}$ (Exp.)	$\lg k_{e^-_{aq}}$ (MLR)	$\lg k_{e^-_{aq}}$ (SVM)	标准残差 Standardized Residual (MLR)	标准残差 Standardized Residual (SVM)
69	771-61-9	Pentafluorophenol	0	10.38	10.32	10.08	0.16	0.89
70	773-82-0	Pentafluorobenzonitrile	0	10.43	10.86	10.27	−1.14	0.49
71	873-62-1	3-Cyanophenol	−1	9.68	9.95	9.94	−0.72	−0.79
72	880-78-4	Pentafluoronitrobenzene	0	10.64	10.77	10.34	−0.34	0.93
73	1194-02-1	4-Fluorobenzonitrile	0	10.20	9.84	10.13	0.97	0.23
74	3964-56-5	4-Bromo-2-chlorophenol	0	9.95	9.54	9.65	1.09	0.92
75	3964-56-5	4-Bromo-2-chlorophenolate	−1	9.65	10.01	9.84	−0.95	−0.57
验证集 (Validation set)								
76	74-11-3	4-Chlorobenzoic acid	−1	9.78	9.47	9.60	0.94	0.66
77	83-40-9	3-Methylsalicylic acid	−1	9.73	9.24	9.33	1.49	1.52
78	88-73-3	2-Nitrochlorobenzene	0	10.38	10.83	10.55	−1.37	−0.63
79	95-75-0	3,4-Dichlorotoluene	0	9.23	9.51	9.51	−0.85	−1.08
80	98-10-2	Benzenesulfonamide	0	10.20	9.91	9.99	0.91	0.81
81	98-95-3	Nitrobenzene	0	10.58	10.62	10.70	−0.12	−0.46
82	100-44-7	Benzylchloride	0	9.18	9.39	9.48	−0.65	−1.16
83	100-47-0	Benzonitrile	0	10.28	10.07	10.21	0.63	0.26
84	100-51-6	Benzyl alcohol	0	8.30	8.49	8.44	−0.59	−0.51
85	100-53-8	Benzylmercaptan	0	9.94	9.86	9.71	0.24	0.86
86	100-65-2	N-Phenylhydroxylamine	0	9.26	9.61	9.43	−1.08	−0.67
87	104-15-4	p-Toluenesulfonic acid	−1	9.22	9.14	9.39	0.24	−0.65
88	104-88-1	4-Chlorobenzaldehyde	0	10.34	10.47	10.67	−0.40	−1.26
89	367-11-3	1,2-Difluorobenzene	0	9.08	9.05	9.24	0.08	−0.62
90	445-29-4	2-Fluorobenzoic acid	−1	9.49	9.61	9.65	−0.37	−0.62
91	587-04-2	3-Chlorobenzaldehyde	0	10.34	10.83	10.56	−1.49	−0.82
92	591-20-8	3-Bromophenol	0	9.78	9.82	9.85	−0.11	−0.28
93	694-80-4	2-Bromochlorobenzene	0	9.77	9.55	9.62	0.66	0.56
94	767-00-0	4-Cyanophenol	−1	9.30	9.24	9.38	0.18	−0.30

1.3　模型的建立和评价

使用 SPSS 19.0 软件中的逐步 MLR 方法筛选变量,构建模型.MLR 是一种简洁透明的线性回归算法[18],有利于模型的机理解释.基于筛选出的描述符,采用了 SVM 方法构建非线性的模型[19].本文的 SVM 算法在 MATLAB 2014 软件上执行.基于遗传算法对模型参数进行优化,种群个数设置为 20,终止代数为 100,最终选取方差最小的模型.SVM 模型由容量参数(C)、不敏感损失函数(ε)和可影响模型预测能力的参数(γ) 3 个参数决定[20-21].为了获得最好的泛化能力,在建模过程中需要选择合适的核函数,调节相应的参数(C, γ 和 ε)组合[22-22].

模型建立后,对其进行内部、外部验证及评价.使用以下统计学参数评价模型的性能,包括校正后相关系数(R^2_{adj})、均方根误差(RMSE)、去一法交叉验证系数(Q^2_{LOO})和由 bootstraping 计算的 Q^2_{boot}(每次随机从训练集剔除 20%的化合物,重复 5000 次).基于以下原则获取最优模型:模型的描述符数目要少,拥有较高的 R^2_{adj}、Q^2_{LOO}、Q^2_{boot}和低的 RMSE,所有描述符的方差膨胀因子(VIF)小于 10,自变量矩阵 Mx 的 K 相关系数(K_x)小于自变量与因变量矩阵 Mxy 的 K 相关系数(K_{xy}).采用 Williams 法表征模型的应用域.将训练集和验证集中每个化合物的杠杆值(h)和标准残差(δ)作图.化合物的$|\delta|$值大于 3 被认为是离群点.

1.4　元分析

为了描述不同取代基对 e^-_{aq}与有机物反应速率常数的影响,本文进行了元分析.依据 Luo 等[23]的方

法,将取代基按照供电子和吸电子强弱依次分为强供、中供、弱供、强吸、中吸和弱吸等 6 类.当一种物质含有多个取代基时,则分别归为所在类别进行分析.最终,以 $\lg k_{e^-_{aq}}$ 对这 6 类官能团作箱型图,从而分析官能团对 $\lg k_{e^-_{aq}}$ 的影响.

2 结果与讨论(Results and discussion)

2.1 $\lg k_{e^-_{aq}}$ 预测模型

构建的最佳 MLR 预测模型如下:

$\lg k_{e^-_{aq}} = 3.877 - 12.435\ E_{LUMO} + 0771\ ATS4m + 5.089\ SpMAD\_AEA(dm) - 1.640\ Eig03\_AEA(bo) + 0.268\ GATS5m - 1.610\ HATS4p$

$n_{tr}=75$, $R^2_{adj,tr}=0.801$, $RMSE_{tr}=0.359$, $Q^2_{LOO}=0.810$, $Q^2_{BOOT}=0.761$, $Kx=0.443$, $Kxy=0.489$, $n_{ext}=19$, $R^2_{adj,ext}=0.820$, $RMSE_{ext}=0.260$, $Q^2_{ext}=0.792$

该模型包含 6 个描述符,其含义见表 2.每个描述符的 VIF 值均小于 5,且 $K_X < K_{XY}$,表明该模型不存在多重相关性(表 2).训练集的 $R^2_{adj,tr}$ 大于 0.8,表明模型具有很好的拟合能力.$R^2_{adj,tr}$ 和 Q^2_{LOO} 与 Q^2_{BOOT} 的差值均远小于 0.3,表明模型具有良好的稳健性,不存在过拟合问题.从验证集的统计学参数 $R^2_{adj,ext}$、Q^2_{ext} 和 $RMSE_{ext}$ 可以得出,所构建的模型具有良好的预测能力.MLR 模型中化合物的 $\lg k_{e^-_{aq}}$ 的预测值见表 1,其实验值和预测值的拟合情况见图 2a.

采用上述筛选出的描述符构建 SVM 模型.最终选取的 SVM 模型中 $C=4.733$、$\varepsilon=0.083$、$\gamma=0.476$.统计学参数分别为 $n_{tr}=75$, $R^2_{adj}=0.861$, $RMSE_{tr}=0.315$, $n_{ext}=19$, $R^2_{adj,ext}=0.880$, $RMSE_{ext}=0.209$, $Q^2_{ext}=0.884$.由此可以看出,SVM 模型的拟合能力显著高于 MLR 模型,表明非线性关系能更好的描述 $\lg k_{e^-_{aq}}$ 和描述符之间的关系.但是,MLR 方法具有更好的透明性,便于应用和机理解释. SVM 模型中化合物 $\lg k_{e^-_{aq}}$ 的预测值见表 1,实验值和预测值的拟合情况见图 2b.

表 1 QSAR 模型中描述符的含义、VIF 值、t 检验值和显著性水平 P 值

Table 1 Meanings, VIF, t and P values of descriptors in the QSAR model

描述符 Descriptor	含义 Definition	VIF	t	P
E_{LUMO}	最低未占据分子轨道能	1.344	−7.109	$< 1\times10^{-3}$
ATS4m	质量加权的位于 lag4 处的 Broto-Moreau 自相关描述符	3.589	6.276	$< 1\times10^{-3}$
SpMAD_AEA(dm)	偶极矩加权的来自于扩增边缘邻接处的光谱绝对偏差	3.623	6.338	$< 1\times10^{-3}$
Eig03_AEA(bo)	键级加权的来自于扩增边缘邻接处的特征值	4.626	−5.408	$< 1\times10^{-3}$
GATS5m	质量加权的位于 lag5 处的 Geary 自相关描述符	1.229	4.043	$< 1\times10^{-3}$
HATS4p	极化率加权的于 lag4 处的基于杠杆的自相关描述符	2.494	−2.766	$< 7\times10^{-3}$

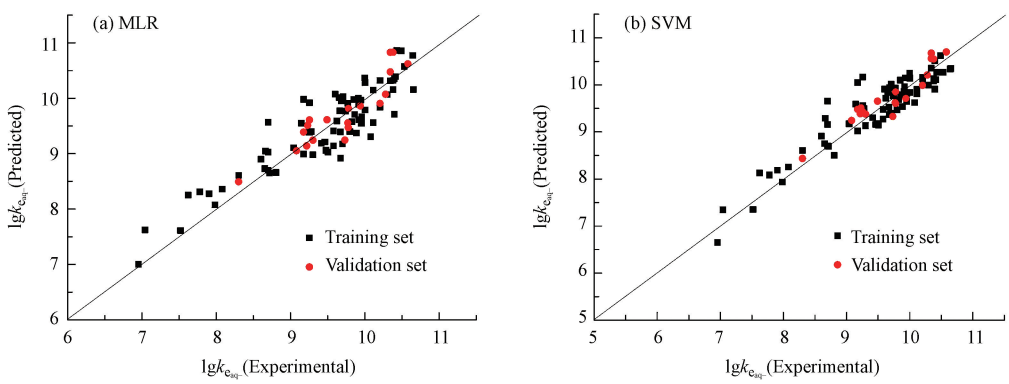

图 2 基于 MLR (a)和 SVM (b)构建的 QSAR 模型的 $\lg k_{e^-_{aq}}$ 实验值与预测值的拟合图

Fig.2 Plots of the predicted versus experimental $\lg k_{e^-_{aq}}$ values for QSAR models developed by the MLR (a) and SVM (b) methods

2.2　应用域表征

采用 Williams 方法对所构建的两个模型的应用域进行了表征(图 3).从图 3 可以看出,94 个化合物的 h 值均小于警戒值(h^*)($h^*=0.280$),且基于 MLR 和 SVM 方法构建的模型中化合物的|δ|均小于 3,表明没有离群点,说明本研究所构建的模型整体上具有很好的预测能力.从图 3 还可以看出,两个模型均对训练集中 2,4-二氯苯酚(No. 44)、4-硝基苯胺(No. 18)和苄基溴(No. 21)这 3 个物质的 $\lg k_{e_{aq}^-}$ 的预测误差较大,模型均高估了它们的 $\lg k_{e_{aq}^-}$ 值.其中,苄基溴包含一个 CH_2Br—基团,而训练集中仅有一个含有该官能团的化合物,说明其结构特征在数据集中体现的不明显,故其预测误差较大.而对于 2,4-二氯苯酚和 4-硝基苯胺的预测误差较大,我们推测很大程度上是由实验测定误差导致的.原因如下:对比了与这两种物质具有相似结构的化合物,发现与 2,4-二氯苯酚具有相似结构的其他 4 个含有两个 Cl 原子取代的芳香化合物(1,2-二氯苯、3,4-二氯甲苯、1,4-二氯苯和 1,3-二氯苯)的 $\lg k_{e_{aq}^-}$ 实验值范围为 9.23—10.00,而 2,4-二氯苯酚的实测值仅为 8.69.同理,本研究中同 4-硝基苯胺结构相似的其它 9 个含有—NO_2的芳香化合物(2-氯-6-硝基甲苯、2-硝基氯苯、2-硝基苯酚、硝基苯、4-硝基苯酚、2,4-二硝基甲苯、3-硝基氯苯、3-硝基苯酚和 1-硝基 5-氟苯)的 $\lg k_{e_{aq}^-}$ 实验值的范围在 10.36—10.64,特别是同样含有吸电子基团和供电子基团的 4-硝基苯酚的 $\lg k_{e_{aq}^-}$ 实验值也大于 10.0,因此推测 4-硝基苯胺的预测误差较大归因于实验测定误差导致.

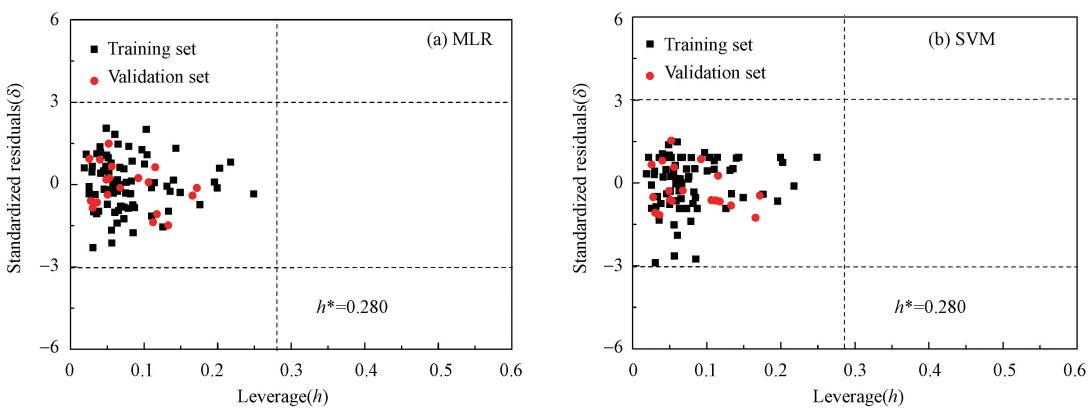

图 3　基于 Williams 方法表征的模型应用域,其中(a)和(b)分别是基于 MLR 和 SVM 方法构建的模型

Fig.3　Williams plots of the QSAR models developed by the MLR (a) and SVM (b) methods

2.3　模型机理分析

从表 1 可以看出,E_{LUMO}($t=-7.109$, $P<1\times10^{-3}$)和 SpMAD_AEA(dm)($t=6.276$, $P<1\times10^{-3}$)是对模型贡献最大的描述符.E_{LUMO} 是衡量分子亲电性的重要参数,E_{LUMO} 值越小,分子亲电性越强[24].因此具有低的 E_{LUMO} 值的化合物越易从 e_{aq}^- 获取电子而被还原.SpMAD_AEA(dm)是指偶极矩加权的来自于扩增边缘邻接处的光谱绝对偏差,与 $\lg k_{e_{aq}^-}$ 呈正相关.前人[25]在研究苯酚类化合物与·OH 的反应时,得出偶极矩与其反应速率常数(k_{OH})呈负相关,因此相对于 e_{aq}^- 参加的还原反应,该类参数同 $\lg k_{e_{aq}^-}$ 呈正相关.ATS4m 和 GATS5m 均是质量加权的 2D 自相关描述符(表示对分子中所有路径长度的终端原子质量进行求和)[26],同 $\lg k_{e_{aq}^-}$ 呈正相关关系,表明芳香类化合物的分子量越大,趋于与 e_{aq}^- 有更高的反应活性.进一步研究发现,芳香类化合物的 $\lg k_{e_{aq}^-}$ 与化合物分子量(MW)呈显著正相关关系(图 4),而描述符 ATS4m 和 GATS5m 与 MW 也存在显著正相关关系($P<0.05$),证实了以上结论.

Eig03_AEA(bo)和 HATS4p 分别表示键级加权的来自于扩增边缘邻接矩阵的特征值[25]和极化率加权的位于 lag4 处的杠杆加权的自相关描述符[26].键级是描述分子中相邻原子之间的成键强度的物理量,键级越大,表示键越强,越不易与 e_{aq}^- 发生反应.芳香化合物的分子极化率越高说明共轭 π 电子的离域性越大,而电子的离域效应可以增加体系的共轭程度,分子的共轭程度越高,其体系越稳定[27].因此,在模型中,这两种描述符的系数为负.

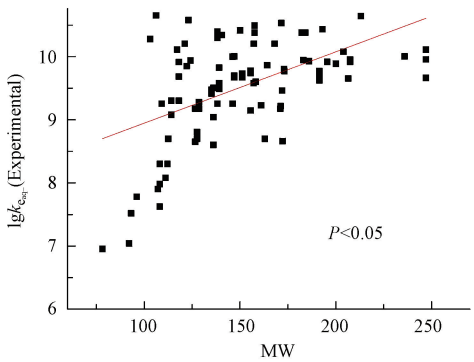

图 4 化合物的实验值 lg$k_{e^-_{aq}}$ 与分子量(MW)的相关关系图

Fig.4 Plot of the experimental lg$k_{e^-_{aq}}$ values versus molecular weight (MW)

2.4 元分析

图 5 展示了芳香类化合物 lg$k_{e^-_{aq}}$ 与官能团的统计结果.总体来看,含有吸电子官能团的化合物的 lg$k_{e^-_{aq}}$ 值普遍高于供电子基团的 lg$k_{e^-_{aq}}$ 值.而且,从图 5 还可以看出,随着吸电子能力的增强,化合物与 e^-_{aq} 的反应活性越高.这是因为 e^-_{aq} 是亲核试剂,它倾向于与电子密度低的位点反应,而不易与电子密度高的位点反应.因此,带有吸电子基团的芳香化合物倾向于比带有供电子基团的化合物与 e^-_{aq} 有更高的反应活性.例如,苯胺的 lg$k_{e^-_{aq}}$ 为 7.51,而硝基苯的 lg$k_{e^-_{aq}}$ 为 10.58.同时,我们发现带有供电子基团的化合物的 lg$k_{e^-_{aq}}$ 随供电子强度变化趋势并不明显,可能是由于每一类供电子基团的化合物覆盖的 lg$k_{e^-_{aq}}$ 的范围较大,导致趋势并不明显.此外,一些化合物除了含供电子基团还含有其它吸电子基团,导致整体上供电子效应不显著.

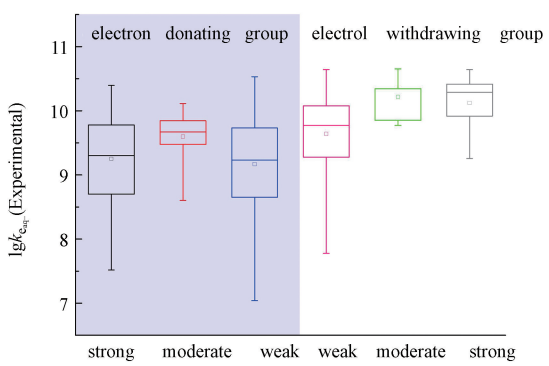

图 5 lg$k_{e^-_{aq}}$ 箱型图(对于含有多个官能团的化合物在每一类别里都进行数)

Fig.5 Box plot of lg$k_{e^-_{aq}}$ values in the study (A compound containing multiple functional groups are counted in the stack of each functional group)

3 结论(Conclusion)

采用 MLR 和 SVM 两种方法,分别构建了用于预测水相中芳香族化合物与 e^-_{aq} 反应的速率常数的线性和非线性 QSAR 模型.所构建的两个模型均具有良好的拟合能力、稳健性和预测能力,可用于预测含有不同官能团的芳香族化合物的 lgk$_{e^-_{aq}}$ 值,为预测污染物在城市污水、饮用水等处理过程中的转化速率提供基本工具,并为筛选及评价污染物是否可以通过 e^-_{aq} 还原降解从水中去除提供理论依据.机理分析表明:最低未占据分子轨道能(E_{LUMO})和偶极矩加权的来自于扩增边缘邻接点光谱的平均绝对偏差(SpMAD_AEA(dm))是影响有机物与 e^-_{aq} 反应活性的重要参数.此外,$k_{e^-_{aq}}$ 还与化合物的键级和极化率有

关.元分析结果表明,具有吸电子官能团的芳香化合物比含供电子基团的化合物倾向于与 e_{aq}^- 有更高的反应活性.

参考文献(References)

[1] VELLANKI B P, BATCHELOR B, ABDEL-WAHAB A. Advanced reduction processes: A new class of treatment processes[J]. Environmental Engineering Science, 2013, 30(5): 264-271.

[2] JUNG B, NICOLA R, BATCHELOR B, et al. Effect of low-and medium-pressure Hg UV irradiation on bromate removal in advanced reduction process[J]. Chemosphere, 2014, 117(1): 663-672.

[3] YAN Q, ZHANG C J, FEL L, et al. Photo-reductive defluorination of perfluorooctanoic acid in water[J]. Water Research, 2010, 44(9): 2939-2947.

[4] LI X, MA J, LIU G, et al. Efficient reductive dechlorination of monochloroacetic acid by sulfite/UV process[J]. Environmental Science & Technology, 2012, 46(13): 7342-7349.

[5] YU H, NIE E, XU J, et al. Degradation of diclofenac by advanced oxidation and reduction processes: Kinetic studies, degradation pathways and toxicity assessments[J]. Water Research, 2013, 47(5): 1909-1918.

[6] ROSSKY P J, SCHNITKER J. The hydrated electron: Quantum simulation of structure, spectroscopy, and dynamics[J]. The Journal of Physical Chemistry, 1988, 92(15): 4277-4285.

[7] THOMAS-SMITH T E, BLOUGH N V. Photoproduction of hydrated electron from constituents of natural waters[J]. Environmental Science & Technology, 2001, 35(13): 2721-2726.

[8] MEZYK S P, HELGESON T, COLE S K, et al. Kinetics of hydrated electron and hydroxyl radical reactions with halonitromethanes in water[J]. The Journal of Physical Chemistry, 2006, 110(6): 2176-2180.

[9] JUDSON R, RICHARD A, DIX D J, et al. The toxicity data landscape for environmental chemicals[J]. Environmental Health Perspectives, 2009, 117(5): 685-695.

[10] ZHUANG S L, WANG H F, DING K K, et al. Interactions of benzotriazole UV stabilizers with human serum albumin: Atomic insights revealed by biosensors, spectroscopies and molecular dynamics simulations[J]. Chemosphere, 2016, 144: 1050-1059.

[11] QU R J, LIU H X, FENG M B, et al. Investigation on intra molecular hydrogen bond and some thermodynamic properties of polyhydroxylated anthraquinones[J]. Journal of Chemical & Engineering Data. 2012, 57(9): 2442-2455.

[12] OECD,Guidance doucument on the validation of (quantitative) structure-activity relationships models[R]. OECD, 2007.

[13] LUO X, YANG X, QIAO X, et al. Development of a QSAR model for predicting aqueous reaction rate constants of organic chemicals with hydroxyl radicals[J]. Environmental Science Processes & Impacts, 2017, 19(3): 350-356.

[14] 徐童,陈景文,李超,等. 气相有机化学品与羟基自由基反应速率常数的 QSAR 模型[J]. 环境化学,2017, 36(4): 703-709.
XU T, CHEN J W, LI C, et al. QSAR models for predicting hydroxyl radical reaction rate constants with organic chemicals in the atmosphere[J]. Environmental Chemistry, 2017, 36(4): 703-709 (in Chinese).

[15] LEE Y, VON GUNTEN U. Quantitative structure-activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment[J]. Water Research, 2012, 46(19): 6177-6195.

[16] POUTSMA M L. Evolution of structure-reactivity correlations for the hydrogen abstraction reaction by hydroxyl radical and comparison with that by chlorine atom[J]. The Journal of Physical Chemistry A, 2013, 117(30): 6433-6449.

[17] XIAO R, YE T, WEI Z, et al. Quantitative structure-activity relationship(QSAR)for the oxidation of trace organic contaminants by sulfate radical[J]. Environmental Science & Technology, 2015, 49(22):13394-13402.

[18] LI X, ZHAO W, LI J, et al. Development of a model for predicting reaction rate constants of organic chemicals with ozone at different temperatures[J]. Chemosphere, 2013, 92(8): 1029-1034.

[19] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[20] LIU H X, XUE C X, ZHANG R S, et al. Quantitative prediction of logk of peptides in high-performance liquid chromatography based on molecular descriptors by using the heuristic method and support vector machine[J]. Journal of Chemical Information & Computer Scienences, 2004, 44(6): 1979-1986.

[21] YAO X J, PANAYE A, DOUCET J P, et al. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression[J]. Journal of Chemical Information & Computer Sciences, 2004, 44(4): 1257-1266.

[22] REN Y, LIU H, YAO X, et al. Prediction of ozone tropospheric degradation rate constants by projection pursuit regression[J]. Analytica Chimica Acta, 2007, 589(1): 150-158.

[23] LUO S, WEI Z, DIONYSIOU D D, et al. Mechanistic insight into reactivity of sulfate radical with aromatic contaminants through single-electron transfer pathway[J]. Chemical Engineering Journal, 2017, 327: 1056-1065.

[24] KARELSON M, LOBANOV V S, KATRIZKY A R. Quantum-chemical descriptors in QSAR/QSPR studies[J]. Chemical Reviews. 1996, 96(3), 1027-1043

[25] WANG Y, CHEN J, LI X, et al. Estimation of aqueous-phase reaction rate constants of hydroxyl radical with phenols, alkanes and alcohols[J]. Molecular Informatics, 2009, 28(11-12): 1309-1316

[26] TODESCHINI R. DRAGON-Software for Molecular Descriptor Calculations (Version 5.4 for Windows). Milan, Italy: Talete srl, 2006.

[27] 桑兰芬,杨滢,杜秀华,等. 苯、萘、蒽芳烃分子稳定性和分子极化率规律的理论研究[J]. 化学研究与应用,1996, 8(1): 70-72.
SANG L F, YANG Y, DU X H. Theoretical study of regularity of stability and polarizability of benzene, naphthalene and anthracene[J]. Chemical Research and Application, 1996, 8(1): 70-72(in Chinese).