于林堂,陈咚咚,陶翠翠,等.基于可解释性机器学习的芬顿工艺降解有机污染物速率的研究 [J]. 中国环境科学, 2025,45(8):4294-4302.
Yu L T, Chen D D, Tao C C, et al. Research on the degradation rate of organic pollutants in fenton process based on explainable machine learning [J]. China Environmetal Science, 2025,45(8):4294-4302.

基于可解释性机器学习的芬顿工艺降解有机污染物速率的研究

于林堂,陈咚咚,陶翠翠,朱腾义* (扬州大学环境科学与工程学院,江苏 扬州 225127)

摘要：采用机器学习模型,包括多元线性回归(MLR)和轻量级梯度提升机(LGB)算法,预测芬顿工艺中 57 种有机污染物的降解效率.通过 SHAP 方法对模型进行机理解释,识别了影响降解效率的关键因素.研究结果表明,LGB 模型在预测精度上(R^2_{adj} = 0.969, Q^2_{LOO}= 0.925, R^2_{ext}= 0.844)优于 MLR 模型(R^2_{adj}=0.831, Q^2_{LOO}=0.802, R^2_{ext}=0.861).SHAP 分析揭示了温度、分子三维结构和原子电离能力是影响降解效率的主要因素.本研究为优化芬顿工艺的操作条件和提升降解效率提供了科学依据,对水处理领域的研究和实践具有重要的指导意义.

关键词：芬顿氧化工艺；污染物降解速率常数 log *K*；有机污染物；机器学习模型；SHAP 分析
中图分类号：X131 **文献标识码**：A **文章标号**：1000-6923(2025)08-4294-09

Research on the degradation rate of organic pollutants in fenton process based on explainable machine learning. YU Lin-tang, CHEN Dong-dong, TAO Cui-cui, ZHU Teng-yi* (College of Environmental Science and Engineering, Yangzhou University, Yangzhou 225127, China). *China Environmetal Science*, 2025,45(8)：4394~4302

Abstract：Machine learning models, including multiple linear regression (MLR) and decision tree machine learning (LGB) algorithms, were used to predict the degradation efficiency of 57 organic pollutants in the Fenton process. Mechanistic interpretation of the models was performed by SHAP method to identify the key factors affecting the degradation efficiency. The results showed that the LGB model (R^2_{adj} = 0.969, Q^2_{LOO} = 0.925, R^2_{ext} = 0.844) outperformed the MLR model (R^2_{adj} = 0.831, Q^2_{LOO} = 0.802, R^2_{ext} = 0.861) in terms of prediction accuracy. The SHAP analysis revealed that temperature, molecular three-dimensional structure, and atomic ionization capacity were the main factors affecting the degradation efficiency. This study provides a scientific basis for optimizing the operating conditions of the Fenton process and enhancing the degradation efficiency, which is of great significance in guiding the research and practice in the field of water treatment.

Key words：fenton oxidation process；pollutant degradation rate constant log *K*；organic pollutants；machine learning model；SHAP analysis

随着人口数量的不断攀升和工业化进程的加速,水体污染问题日益严峻.有机废水中的有毒有害污染物种类繁多,来源广泛,其中有机污染物尤为突出.这些污染物包括各种新型污染物.它们不仅对人类健康构成了直接威胁,也对生态系统的平衡和生物多样性造成了破坏[1-3].因其成分复杂,常规水处理技术难以有效处理.研究发现芬顿工艺凭借其高效的氧化能力和易于操作的优势,在处理难以分解的有机废水中表现出色[4].

芬顿氧化技术[5]是一种广泛应用的高级氧化工艺,利用羟基自由基的高效氧化能力,可降解多种有机废水中的污染物[6].该技术显著提升了污染物降解效率,有效降低废水的化学需氧量,改善出水水质,满足排放标准.芬顿氧化工艺其核心原理是利用二价铁离子(Fe^{2+})与过氧化氢(H_2O_2)在酸性条件下反应生成具有强氧化能力的羟基自由基(·OH),羟基自由基可以氧化大部分的有机物[7],其反应速率常数是臭氧的 10^6~10^{12} 倍[8-9].

在芬顿氧化工艺中,反应速率常数(log *K*)是评价降解污染物效率的非常重要的一项参数[10].然而,每年新出现的污染物种类繁多[11],导致传统的测定方法在确定 log *K* 值时耗时且效率不高.因此,寻求新的方法来测定污染物的 log *K* 变得十分重要.影响降解污染物的反应速率常数的因素有多种,因此改进确定 log *K* 的方法对提高芬顿工艺的去污效率起到关键作用.

机器学习(ML)[12]能够开发出高效的预测模型,

收稿日期：2025-01-03
基金项目：国家自然科学基金资助项目(42077331)
* 责任作者, 教授, tyzhu@yzu.edu.cn

目前有相关的已开发的 ML 模型对于水中降解有机污染物的机理进行研究与分析[13].如何家乐[14]等基于环境水体中化学品光化学持久性的预测模型,为源头防范化学品成为新污染物提供了有效的基础性工具.Qin, LT 等人[15]利用模型对饮用水中消毒副产物毒性进行研究预测,在预测消毒副产物毒性方面具有可靠表现.但现已开发的ML模型存在数据集总量少,描述符单一,拟合度、鲁棒性和预测性都有待提高等问题[16].此外,多元线性回归(MLR)算法是常见的线性回归算法.MLR 模型提供了输入和输出之间明确的线性关系,易于使用[17].由于化合污染物在反应过程中的影响因素过于复杂,并非和 log K 呈现严格的线性关系,其线性关系会导致在遇到工艺反应过程等非线性问题时无法解决而对预测结果产生影响.因此,为了合理有效地模拟芬顿工艺在污水处理中的应用,采用非线性算法的机器学习模型十分必要.

本研究结合了 153 个数据点,涵盖 57 种有机污染物,采用两种机器学习算法来预测芬顿工艺中对这些污染物的降解效率.这些算法包括 MLR 和非线性的轻量级梯度提升机(LGB),为了更进一步理解模型的工作原理,本研究运用了 Shapley Additive explanation(SHAP)方法对模型进行机理解释,为提高模型的有效性和可行性提供了新的思考方向,旨在为优化降解污染物的实验条件和提升降解效率提供科学依据.

1　材料与方法

1.1　数据集

通过查阅文献并整理[16,18-20],一共收集了 57 种有机污染物的共 153 个 log K 实测值(表 1).此外,为了评估模型的稳定性和预测力,采用了内部验证和外部预测的双重验证策略.把数据集随机化按照 4:1 的比例分为训练集和测试集,用训练集(占全部数据的 80%)建立模型并进行内部验证,测试集(占全部数据的 20%)不用于模型构建,仅用来评价模型的外部预测能力.

表 1　有机污染物的 log K 实验值和预测值

Table 1　Experimental and predicted log K values of organic pollutants

序号	化合物 CAS 号	温度 (℃)	Log K 实验值	MLR	LGB	序号	化合物 CAS 号	温度 (℃)	Log K 实验值	MLR	LGB	序号	化合物 CAS 号	温度 (℃)	Log K 实验值	MLR	LGB
1	76-60-8	25	0.054	-0.531	-0.130	52	108-73-6	60	0.478	0.521	0.476	103	606-20-2	25	-0.764	-1.268	-0.985
2	98-95-3	50	0.372	0.259	0.489	53	1936-15-8	50	0.533	0.515	0.651	104	1064-48-8	34.8	-1.277	-1.004	-1.312
3	1064-48-8	20	-1.332	-1.477	-1.464	54	98-95-3	25	-0.498	-0.360	-0.530	105	633-96-5	25	-0.104	-0.509	-0.029
4	95-76-1	40	0.035	-0.326	-0.079	55	5413-75-2	25	0.081	-0.415	-0.235	106	80-05-7	20	-0.575	-0.726	-0.623
5	28920-43-6	25	-0.889	-1.067	-0.877	56	1787-61-7	25	0.329	0.021	0.323	107	88-74-4	30	-0.082	-0.027	-0.120
6	10127-27-2	50	0.217	0.540	0.259	57	132-53-6	25	-0.449	-0.160	-0.471	108	106-44-5	40	-1.527	-1.473	-1.597
7	121-57-3	25	-0.548	-0.855	-0.630	58	80-05-7	25	-0.636	-0.534	-0.623	109	2538-85-4	30	-0.042	-0.115	-0.150
8	121-57-3	30	-0.476	-0.701	-0.489	59	121-57-3	60	0.243	-0.041	0.222	110	1122-91-4	25	-0.347	0.081	-0.307
9	5001-72-9	50	-0.597	-0.438	-0.562	60	62-53-3	60	-0.012	-0.242	-0.223	111	91-56-5	20	-0.127	-0.355	-0.151
10	121-57-3	40	-0.182	-0.450	-0.101	61	609-99-4	25	-1.878	-1.704	-1.523	112	98-95-3	30	-0.137	-0.206	-0.235
11	2538-85-4	50	0.163	0.350	0.220	62	119-26-6	50	0.089	0.139	0.113	113	591-35-5	25	-0.640	-0.310	-0.630
12	547-58-0	30	-0.423	0.203	-0.090	63	5001-72-9	30	-0.875	-0.903	-0.882	114	2602-46-2	30	-0.744	-0.616	-0.751
13	108-39-4	30	-1.588	-1.377	-1.674	64	95-48-7	50	-1.542	-1.409	-1.437	115	582-17-2	25	0.225	-0.033	0.198
14	88-74-4	50	0.499	0.438	0.551	65	91-56-5	30	0.016	-0.010	0.145	116	547-58-0	60	1.301	0.863	0.820
15	66-71-7	20	-1.174	-0.873	-0.908	66	2538-85-4	40	0.084	0.136	0.092	117	100-10-7	25	-0.388	-0.330	-0.294
16	5001-72-9	20	-1.087	-1.248	-1.021	67	10127-27-2	60	0.220	0.735	0.259	118	1064-48-8	15.8	-1.678	-1.693	-1.464
17	108-73-6	50	0.448	0.273	0.474	68	860-22-0	25	0.168	0.261	0.226	119	62-53-3	50	-0.178	-0.437	-0.223
18	115-39-9	20	-0.551	-0.214	-0.312	69	62-53-3	40	-0.747	-0.651	-0.527	120	582-17-2	50	0.424	0.586	0.479
19	10127-27-2	40	0.205	0.326	0.203	70	2303-01-7	30	0.043	-0.262	0.009	121	10127-27-2	25	0.097	-0.079	0.101
20	569-64-2	20	0.467	0.714	0.295	71	632-99-5	25	-1.051	-0.624	-0.826	122	547-58-0	40	0.272	0.454	0.352
21	1064-48-8	43.5	-1.294	-0.803	-1.196	72	582-17-2	30	0.285	0.121	0.289	123	61-73-4	25	-0.031	-0.268	-0.125
22	108-73-6	30	0.345	0.105	0.433	73	99-04-7	25	-0.910	-0.608	-0.916	124	62-53-3	25	-1.252	-1.056	-1.082
23	91-56-5	40	0.607	0.242	0.679	74	88-21-1	25	-1.421	-1.498	-1.413	125	62-53-3	30	-1.244	-0.902	-0.853
24	587-98-4	25	-0.360	-0.511	-0.365	75	2303-01-7	40	0.094	-0.011	0.068	126	66-71-7	50	0.264	-0.063	0.206
25	2602-46-2	20	-0.857	-0.773	-0.832	76	582-17-2	60	0.436	0.781	0.479	127	80-05-7	40	-0.504	-0.129	-0.339
26	66-71-7	30	-0.472	-0.528	-0.623	77	94-75-7	25	-0.107	-0.372	-0.301	128	80-05-7	50	-0.127	0.085	0.038

续表 1

序号	化合物 CAS 号	温度 (℃)	Log K 实验值	MLR	LGB	序号	化合物 CAS 号	温度 (℃)	Log K 实验值	MLR	LGB	序号	化合物 CAS 号	温度 (℃)	Log K 实验值	MLR	LGB
27	95-76-1	25	−0.326	−0.336	−0.332	78	95-76-1	20	−0.733	−0.754	−0.639	129	89-62-3	25	−0.296	−0.842	−1.133
28	108-39-4	50	−1.461	−0.912	−1.335	79	91-56-5	25	−0.009	−0.164	−0.151	130	91-56-5	50	0.781	0.455	0.943
29	88-74-4	25	−0.419	−0.181	−0.414	80	66-71-7	60	0.377	0.132	0.206	131	95-48-7	20	−1.818	−2.219	−1.788
30	88-74-4	60	0.567	0.633	0.551	81	1064-48-8	30	−1.255	−1.132	−1.317	132	95-76-1	60	0.154	0.157	0.129
31	100-02-7	25	−1.090	−1.041	−0.992	82	108-73-6	25	0.320	−0.030	0.178	133	98-95-3	40	0.024	0.046	0.218
32	99-09-2	25	−1.263	−0.277	−1.028	83	100-10-7	50	0.383	0.289	0.345	134	100-01-6	10	−1.488	−1.056	−0.452
33	2303-01-7	25	−0.043	−0.416	0.028	84	10127-27-2	30	0.163	0.075	0.075	135	100-01-6	30	−0.012	0.011	−0.187
34	1239-45-8	25	−0.272	0.145	−0.140	85	106-44-5	30	−1.627	−1.724	−1.723	136	100-01-6	40	0.172	0.262	0.099
35	1936-15-8	25	0.378	0.269	0.343	86	1936-15-8	30	0.502	0.381	0.581	137	100-01-6	50	0.479	0.476	0.440
36	304-88-1	25	−0.519	−0.656	−0.640	87	95-48-7	30	−1.686	−1.874	−1.766	138	100-01-6	60	0.542	0.671	0.440
37	95-48-7	40	−1.587	−1.622	−1.639	88	66-71-7	25	−0.836	−0.682	−0.908	139	100-10-7	30	−0.186	−0.176	−0.040
38	121-47-1	25	−1.308	−1.027	−1.275	89	98-95-3	60	0.483	0.454	0.489	140	100-10-7	60	0.460	0.484	0.345
39	91-56-5	60	1.074	0.650	0.943	90	66-71-7	40	−0.350	−0.277	−0.275	141	106-44-5	20	−1.764	−2.069	−1.705
40	121-57-3	50	0.200	−0.236	0.222	91	119-26-6	60	0.146	0.334	0.113	142	106-48-9	25	−0.664	−0.731	−0.808
41	106-44-5	50	−1.485	−1.259	−1.319	92	95-76-1	30	−0.103	−0.224	−0.107	143	108-39-4	40	−1.502	−1.126	−1.553
42	2602-46-2	40	−0.608	−0.571	−0.686	93	95-76-1	50	0.121	−0.090	0.129	144	119-26-6	25	−0.304	−0.479	−0.571
43	2303-01-7	50	0.133	0.202	0.173	94	582-17-2	40	0.359	0.372	0.330	145	119-26-6	30	−0.136	−0.325	−0.451
44	100-10-7	40	0.228	0.076	0.137	95	108-39-4	20	−1.676	−1.723	−1.680	146	119-26-6	40	0.012	−0.074	−0.018
45	69-72-7	25	−1.499	−0.741	−1.437	96	108-95-2	25	−1.165	−1.261	−1.169	147	120-83-2	25	−1.203	−1.233	−0.774
46	117-82-8	25	−1.509	−1.477	−1.441	97	1936-15-8	40	0.532	0.279	0.417	148	547-58-0	25	−0.455	0.049	−0.341
47	477-73-6	25	−0.375	−0.840	−0.412	98	80-05-7	60	−0.011	0.279	0.038	149	547-58-0	50	1.286	0.668	0.820
48	5001-72-9	40	−0.723	−0.651	−0.683	99	98-11-3	25	−1.540	−1.353	−1.591	150	1936-15-8	60	0.543	0.762	0.651
49	88-74-4	40	0.360	0.224	0.328	100	80-05-7	30	−0.600	−0.380	−0.576	151	2538-85-4	25	−0.190	−0.269	−0.265
50	2303-01-7	60	0.207	0.397	0.173	101	100-01-6	25	−0.279	−0.143	−0.452	152	2538-85-4	60	0.202	0.545	0.220
51	85-44-9	25	−1.125	−0.945	−1.087	102	108-73-6	40	0.424	0.037	0.375	153	2602-46-2	50	−0.581	−0.222	−0.629

1.2 描述符的计算与筛选

首先在计算分子结构描述符前,数据集中的化合物需要利用 Chem3D 软件生成初始有机物的分子结构,其次采用本软件中的 Minimize Energy 模块对分子结构进行最小能量化[21].使用 Gaussian 16W 软件对分子结构进行几何优化,其次通过(DFT B3LYP/6−311G)基组进一步计算量子描述符(QCD)(偶极矩(μ)、分子总能量(E_{B3LYP})、分子最低占据轨道能(E_{LUMO})、分子最高占据轨道能(E_{HOMO})等)[22].键级(BO)和福井指数(f (+),f (−),f (0))则采用 Materials Studio 软件中的 Dmol 3 模块进行计算得出[23],其余的 5290 个分子结构描述符(MD)通过 alvaDesc 软件生成[24].

利用 R 软件对分子结构描述符进行筛选(剔除相关性大于等于 0.9 的描述符、为 0 的值、缺失值 NA 和接近常数的变量)得到 440 个描述符(MD+QCD),再使用 SPSS 软件对其进行 MLR 分析,方差膨胀因子(VIF)是评估 MLR 建模过程中输入变量之间多重共线性关系的重要参数.VIF<5 是得到高性能质量的 MLR 模型的前提,在满足多重共线性 VIF<5 的条件下,选择出描述符数量最少,且决定系数(R^2_{adj})最大的(MLR)预测模型.

1.3 机器学习(ML)模型

在本文研究中,选择了两种流行 ML 算法.MLR 作为一种在水处理领域内广泛应用的线性算法[25],其核心目标在于从多变量数据集中识别出最具预测力的关键变量.该算法通过分析多个自变量与因变量之间的线性关联,实现了高维数据集向低维数据集的转换,有效降低了变量间的冗余信息,进而提升了模型的预测精确度和解释能力.在使用 SPSS 软件执行 MLR 降维分析的过程中,首先对自变量间的多重共线性进行检验,排除 VIF 大于 5 的变量.随后,基于筛选后的自变量集合,选择能够最优化模型的决定系数(R^2)及预测决定系数(Q^2)调整值的变量组合,以此作为最终的预测模型.LGB 算法是 GBDT[26]的高效实现[27],该算法的提出是为了解决 GBDT 在数据量增长中遇到的问题.LGB 算法在处理复杂数据集时表现出色,它能够识别并利用数据中的关键特征,从而有效降低数据维度,可以通过加权或采样技术有效应对分类问题中的样本不平衡问题,并对

数据进行预处理[28].它包含了两种新技术:独占特征捆绑[29]和基于梯度的单边采样[30],这些技术可以处理大量的记录和特征.在 LGB 中设置以下超参数及取值范围(表 2):

表 2　LGB 模型超参数

Table 2　Hyperparameters of the LGB model

模型名称	参数	取值范围
LGB	n_estimators	[50,100,200]
	learning_rate	[0.01,0.001,0.1]
	num_leaves	[40,50,80]
	max_dept	[4,6,8]
	feature_fraction	[0.5,0.6,0.7,0.8,0.9]

通过网格搜索遍历给定的参数网格来寻找最优参数组合,最后确定模型的超参数为 n_estimators=200, learning_rate=0.1, num_leaves=50, max_dept=8, feature_fraction=0.8. 两种算法均在 Python 3.12 环境下实现.

1.4　模型的表征

本研究中,所采用的最优模型的拟合程度通过两个关键参数来衡量,分别是经自由度调整后的决定系数(R^2_{adj})和均方根误差(RMSE).模型的稳健性由去一法交叉验证系数(Q^2_{LOO})进行表征[31].用测试集决定系数 R^2_{ext} 和外部验证系数 Q^2_{ext} 表征模型预测能力.其中部分的计算参数如下:

$$R^2_{adj}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n_{tra}-m-1)}{\sum_{i=1}^{n}(y_i-\hat{y})^2/(n_{tra}-1)} \tag{1}$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)}{n}} \tag{2}$$

$$Q^2_{ext}=1-\frac{\sum_{i=1}^{n_{ext}}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n_{ext}}(y_i-\overline{y}_{ext})^2} \tag{3}$$

式中:y_i和 $\hat{y}_i$ 分别表示在训练集中第 i 个物质的 log K 实测值和预测值;$\overline{y}$ 表示训练集中物质的 log K 实测值的平均值;n_{tra} 表示训练集中物质的个数;m 表示描述符个数;$\overline{y}_{ext}$ 表示测试集中物质 log K 实测值的平均值;n_{ext} 表示测试集中物质的个数.

采用预测模型过程中计算得到的标准残差(δ)和杠杆值(h)定义的 Willianms 图[32]表征模型的应用域,通过使用 h 和标准残差值 δ 两个参数来确定化合物的合理区间.δ、h_i 和警戒值 h^*的计算公式如下:

$$\delta=\frac{y_i-\hat{y}_i}{\sqrt{\sum_{i+1}^{n}(y_i-\hat{y}_i)^2/n-m-1}} \tag{4}$$

$$h_i=x_i^T(X^TX)^{-1}x_i \tag{5}$$

$$h^*=3(m+1)/n \tag{6}$$

式中:n 代表训练集中化合物的个数;m 为描述符的个数;X 表示 $n\times m$ 矩阵,表征了模型内训练集物质的描述符空间;h_i 为不同化合物的杠杆值;h^*定义为警戒值.

1.5　基于 SHAP 解释和特征选择

近些年来,在机器学习和深度学习(DL)[33]领域,模型解释性是一个重要的课题,尽管模型的精度和预测性表现日渐优异,但由于非线性算法的复杂程度越来越高导致难以解释模型内部决策过程.SHAP 是一种解决这一问题的工具,通过分配特征的重要性值来解释模型的输出.SHAP 基于合作博弈论的 Shapley 值[34],在机器学习模型解释中,SHAP 运用这一概念来计算每个特征对预测结果的贡献[35].化合物的 SHAP 值的绝对值越大说明输入特征量的贡献度就更高,正值表示变量有助于反应速率的提升,负值表示变量会抑制反应的速率.最后通过 SHAP 值图,可以直观地阐明特征变量对模型预测 log K 结果的影响.

2　结果与讨论

2.1　变量筛选

经过初步数据筛选处理,得到了 440 个自变量,用逐步增加的输入变量个数重新训练 MLR 模型.由图 1 看出当变量达到 15 的时候,模型的拟合度(R^2_{adj})和预测精度(Q^2_{ext})达到峰值.在最重要的变量达到 15 个之后,VIF>5,这表明 15 个描述符变量都有助于模型性能的提升.在相关性分析图中(图 2)可以看到输入特征量的相关程度,相关性系数均小于 0.95[36],所以这些变量可以用于模型的开发.最终筛选出的特征变量分别为:温度(T)、4 阶平均拓扑电荷指数(JGI4)、一阶平均电负性(E1m)、未加权 H 滞后 0 的自相关性(H0u)、以电离势加权的滞后 5 的 Moran 自相关(MATS5i)、以电离势加权的滞后 4 的 Moran 自相关(MATS4i)、原子极化率加权的第二分量无障碍定向 WHIM 指数(E2p)、按原子极化率加权的 3D-MoRSE 信号 15(Mor15p)、拓扑二元距离(TDB01i)、质量加权滞后 6 的杠杆加权自相关

(HATS6m)、由 Sanderson 电负性加权的滞后 2 的 Geary 自相关(GATS2e)、温度参数倒数(1/*T*)、增强边缘邻接矩阵的特征值 n3(Eig03_AEA(*dm*))、基于拓扑距离的极化性加权滞后 3(3TDB03p)、未加权的 *R* 滞后的最大自相关(R5u^{+}).

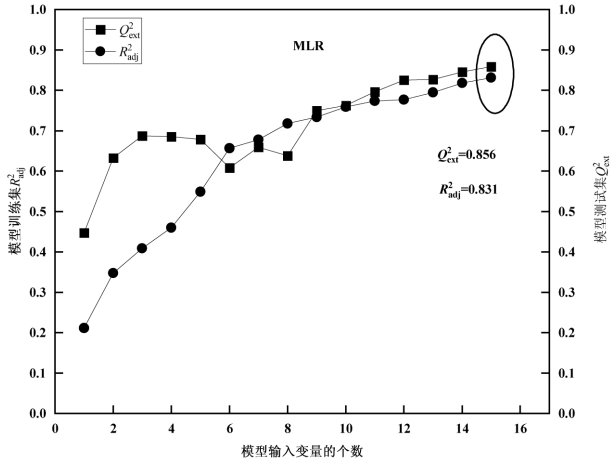


图 1 1~15 个自变量的 MLR 模型的 R^2_{adj} 和 Q^2_{ext} 线型图

Fig.1 R^2_{adj} and Q^2_{ext} line graphs of the MLR model with 1~15 independent variables

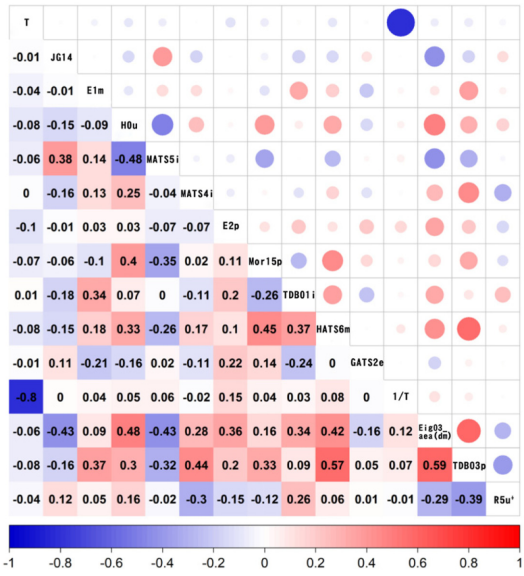


图 2 15 个自变量之间的相关性热力学图

Fig.2 Thermodynamics diagram of the correlation between 15 independent variables

2.2 预测模型

多元回归算法(MLR)建立的 log *K* 预测为:

log *K*=0.016*T*−3.557JGI4+0.974E1m−0.828H0u−1.898MATS5i+2.983MATS4i−4.033E2p+0.967Mor15p+4.514TDB01i−0.391HATS6m+0.756GATS2e−11.2911/*T*+0.204Eig03_AEA(dm) −0.948TDB03p−4.120R5u^{+} n_{tra}=122,R^2_{adj}=0.831,Q^2_{LOO}=0.802, $RMSE_{tra}$=0.271,*P*<0.001;n_{ext}=31,R^2_{ext}=0.861, Q^2_{ext}=0.859,$RMSE_{ext}$=0.286

LGB 算法建立的模型中:

n_{tra}=122,R^2_{adj}= 0.969,Q^2_{LOO}= 0.925,$RMSE_{tra}$= 0.162, *P*<0.001;n_{ext}= 31,R^2_{ext}= 0.844,Q^2_{ext}= 0.842, $RMSE_{ext}$= 0.420.

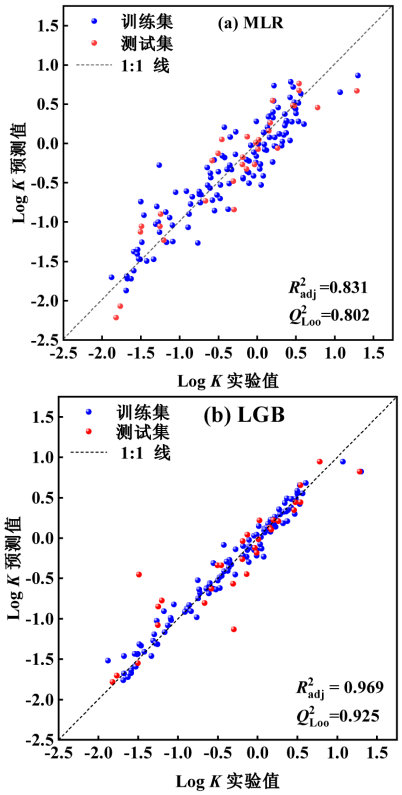


图 3 两个模型 log *K* 实测值和预测值的拟合关系

Fig. 3 The fitting relationship between the measured and predicted log *K* values of two models

本文所采用的两个模型的评价参数均具有较高的相关系数 R^2_{adj}、R^2_{ext} 值,符合 Gramatica 等[37]提出的模型标准(R^2>0.7,Q^2>0.6)[37],表明了预测模型均具有良好的拟合精度和预测能力,所以本文建立的模型是可靠的.依据 Kiralj 等[38]的观点,如果决定系数(R^2)与交叉验证的决定系数(Q^2)之差超过 0.3,则认为模型存在过度拟合和包含非显著性预测变量.上文两模型的 R^2_{adj} 和 Q^2_{Loo} 之差小于 0.3,因此模型并不存在过拟合现象.在 MLR 模型中 VIF 值均在 1~5 之间,表明输入变量之间不存在多重共线性.并

且由于 Q^2_{LOO} 的高分,所以模型也具有很好的稳定性.较低的误差参数 *RMSE* 值也表明模型有着较强的泛化能力.图 3 展示了 log *K* 的实验值与模型预测值之间的相关性,观测数据点紧密地分布在 1:1 线的近旁,表明模型具有较高的拟合度.

2.3　应用域表征

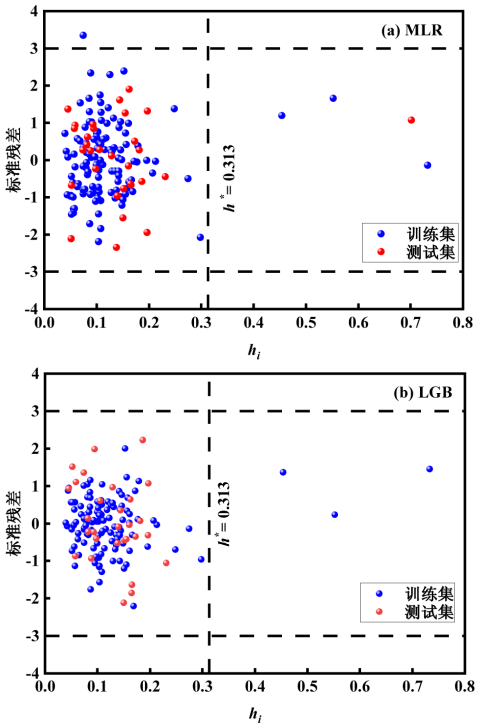


图 4　MLR 和 LGB 模型的 Williams 图

Fig.4　Williams diagram of MLR and LGB models

即使模型已经证明具有稳健的性能,也不能确保其对所有化合物的预测都是准确可靠的,所以有必要对模型定义应用范围.应用域是用于确定哪些化合物能被模型有效预测的边界,它定义了模型适用化合物的范围[39].当化合物的标准残差的绝对值 $|\delta|\leqslant 3$,则说明模型中没有离散点.由图 4(a)可看出 MLR 模型只有训练集中一个化合物为离群点,其杠杆值 $h<h^*$ 且 $|\delta|>3$,表明该物质对于模型的构建有较大影响,该合物的预测值与实验值相近,表明它们可以增强该模型的稳定性和可靠性,可以较准确地反映它们的实际性质.其他化合物的标准残差 δ 都落在了(−3,3)区间之内.并且绝大部分化合物的杠杆值小于警戒值(h^*=0.313),模型中存在几处高杠杆值点,其产生高杠杆点原因为位置相对孤立或数据收集过程中的特殊情况使得这些点在回归分析中具有较高的影响.高杠杆点与其他大多数点所形成的趋势一致,称这种点为“无害”高杠杆点,本不会对模型的稳定性和可靠性产生负面影响,而且可以提高模型的精度[40]、模型的整体性能和适用范围.因此,文中所建立的两种模型可以准确预测应用域内其他种类的化合物的 log *K*.

2.4　机理分析

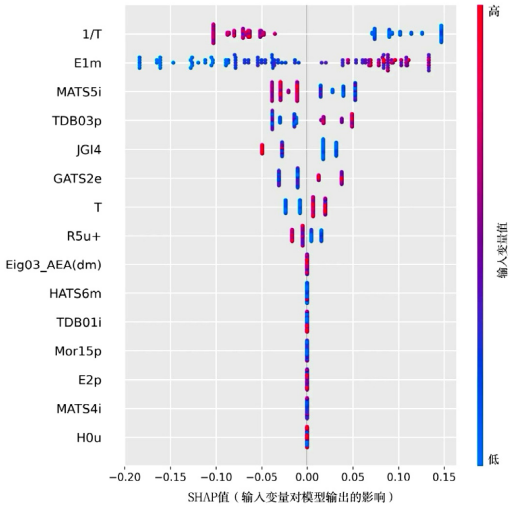


图 5　SHAP 特征概要图

Fig.5　SHAP feature summary diagram

为了进一步探究降解污染物 log *K* 的机理,分析了各个特征量对 log *K* 的影响,本研究计算分析了每个 SHAP 值以给予模型的可解释性.图 5 中, *JGI4*、GATS2e、MATS4i 和 MATS5i 为 2D 自相关描述符.作为平均拓扑电荷描述符,JGI4 与分子的疏水性有关.当 JGI4 值增加时分子的疏水性同时会增加,降解污染物的 log *K* 会减小.反之当 JGI4 较低时,log *K* 也会随之提高[41].GATS2e 用于描述分子结构的拓扑 2D 自相关描述符.它可以根据原子的 Sanderson 电负性加权来计算不同距离下的自相关性,在 Singh 等[42]的研究中发现 GATS2e 都对抑制活性有负增加作用,和图 5 中 GATS2e 的相关性也表示一致.MATS5i 和 MATS4i 表示 Moran 自相关指数,反映了原子的电离能力.值越高表明分子中具有较高电离势的区域在空间上的自相关性越强,失去电子所需的最小能量越大的分子结构更加稳定[43].*E1m* 和 *E2p* 属于 *WHIM* 描述符[44],表示分子可接近性的首个方向性分量,并按照原子质量进行加权.当 *E1m* 增加,分子部分结构更容易被接近,会增强分子的活性

或者改变其与受体的结合方式,和反应速率成正相关有利于 log *K* 的提高[45].R5u$^+$、H0u 和 HATS6m 属于 GETAWAY 描述符.R5u+表示无权重的最大自相关性,值增大意味着在给定的滞后 5 下,分子中原子的几何分布的相似性或聚集程度增加,在 SHAP 图中可以看出 R5u+与 log *K* 成负相关,可能是由于分子中几何分布可能会影响反应速率[46].3TDB03p 是一种基于三维拓扑距离和极化率的 3D 自相关描述符[47],其表示基于分子的拓扑距离矩阵和考虑原子间的连接关系,考虑了分子中原子间的距离,并以原子的极化率作为权重.图 5 中 TDB03p 右半轴为红色,表明其对反应速率有正贡献.温度对 log *K* 值的影响是正相关,温度升高会加剧分子热运动从而增加分子间碰撞频率,促进污染物降解反应的进行.Mor15p 属于 3D−MoRSE 描述符,3D−MoRSE 用于化学信息学的三维结构编码框架.分子的特征可能与范德华体积和极化率有关,Mor15p 代表化合物的性质,计算时考虑了原子的极化率作为权重[48].此外,E2p、Eig03_AEA(dm)、HATS6m、TDB01i、Mor15p、MATS4i 以及 H0u 在内的描述符对实验 log *K* 的影响方向不显著.

2.5 模型比较

本文构建的模型也与前人所研究的一些模型进行了比较,如表 3 所示,结果表明,此前的模型涵盖的化合物的范围较小,未达到预期的满意水平因此难以广泛应用,所以有必要对其进行进一步的优化和改进.本研究中所构建的模型涉及到的有机化合物数据集数量和种类更多(n=153)并分析了多种描述符类型,且获得了较高的拟合优度(R^2_{MLR}=0.831, R^2_{LGB}=0.969).

表 3 模型比较

Table 3 Model Comparison

模型	n	R^2	方程公式	文献
Log K	93	0.772	Log K=−2.813+0.015T−15.6081/T−0.013#C+0.361#O:C+7.3292$q(CH)_x^+$+1.181$q(C^-)_x$−0.167E_{gap}+4.120E^2_{gap}−2.678$q(H)^+$+1.475BO_n−15.675$f(+)_n$+0.383$f(0)_x$	[16]
ln K	19	0.898	ln K=4.009+0.239$_\mu$−30.829f(+)$_x$−3.329$q(CH)_x^+$	[18]
lg K	33	0.807	ln K=−1.281+0.004SAG+11.678$f(0)n$−5.839E_{gap}	[19]
Log K^*	153	0.831	Log K=0.016T−3.557JGI4+0.974E1m−0.828H0u−1.898MATS5i+2.983MATS4i−4.033E2p +0.967Mor15p+4.514TDB01i−0.391HATS6m+0.756GATS2e−11.2911/T+0.204Eig03_AEA(dm) −0.948TDB03p−4.120R5u$^+$	本研究

3 结论

3.1 使用两种机器学习模型模拟并分析了芬顿工艺降解废水中有机污染物的效率,模型均具有拟合度高,稳健性良好和预测能力强等特点.

3.2 两模型的统计指标 Q^2 和 R^2 均超过标准值,仅有一个异常值且其他化合物都在给出的应用域范围内,即标准残差$|\delta|\leqslant 3$,模型预测准确.但 LGB 方法建立的模型(R^2_{adj}=0.969)比用 MLR 方法建立的模型(R^2_{adj}=0.831)更加优异.

3.3 采用 SHAP 值解释法,证明影响芬顿工艺中降解有机污染物 log *K* 的主要因素是温度、分子三维结构和原子电离能力.为提高降解有机污染物 log *K* 可控能力提供了有效的理论支持.

参考文献:

[1] Alvarez P J J, Chan C K, Elimelech M, et al. Emerging opportunities for nanotechnology to enhance water security [J]. Nature Nanotechnology, 2018,13(8):634−641.

[2] Eliasson J. The rising pressure of global water shortages [J]. Nature, 2015,517(7532):6.

[3] Schwarzenbach R P, Escher B I, Fenner K, et al. The challenge of micropollutants in aquatic systems [J]. Science, 2006,313(5790): 1072−1077.

[4] Bello M M, Raman A A A, Asghar A. A review on approaches for addressing the limitations of Fenton oxidation for recalcitrant wastewater treatment [J]. Process Safety and Environmental Protection, 2019,126:119−140.

[5] 谢谨裕,王卓雯,高伟杰,等.钴基气体扩散电极强化电芬顿处理磺胺噻唑钠 [J]. 中国环境科学, 2023,43(1):122−131.
Xie J Y ,Wang Z W,Gao W J, et al. Cobalt−based gas diffusion electrodes enhance the E−Fenton for sulfathiazole sodium treatment [J]. China Environmental Science, 2023,43(1):122−131.

[6] 杜茂华,李皓芯,任 婧,等.改性阴极生物电芬顿系统降解罗丹明 B [J]. 中国环境科学, 2021,41(4):1681−1688.
Du M H, Li H X, Reng J, et al. Electricity production and rhodation B degradation in bioelectric fenton system with modified cathode [J]. China Environmental Science, 2021,41(4):1681−1688.

[7] 辛丽红,晏彩霞,聂明华,等.微量铜离子联合碳酸氢盐类芬顿体系降

解水中双酚 A [J]. 中国环境科学, 2023,43(3):1186−1196.
Xin L H, Yan C X, Nie M H, et al. Degradation of bisphenol A by Fenton−like system using trace copper ions combined with bicarbonate in water [J]. China Environmental Science, 2023,43(3):1186−1196.

[8] Cheng M, Zeng G M, Huang D L, et al. Hydroxyl radicals based advanced oxidation processes (AOPs) for remediation of soils contaminated with organic compounds: A review [J]. Chemical Engineering Journal, 2016,284:582−598.

[9] Hou X J, Huang X P, Jia F L, et al. Hydroxylamine promoted goethite surface fenton degradation of organic pollutants [J]. Environmental Science & Technology, 2017,51(9):5118−5126.

[10] Jubb A M, Gierczak T, Baasandorj M, et al. Methyl−perfluoroheptene−ethers ($CH_3OC_7F_{13}$): Measured OH radical reaction rate coefficients for several isomers and enantiomers and their atmospheric lifetimes and global warming potentials [J]. Environmental Science & Technology, 2014,48(9):4954−4962.

[11] 高　歌,张文晴,魏歆倪,等.新型优先管控污染物筛选研究进展 [J]. 中国环境科学, 2024,44(11):6472−6483.
Gao G, Zhang W Q, Wei X N, et al. Research progress on screening of emerging priority controlled contaminants [J]. China Environmental Science, 2024,44(11):6472−6483.

[12] Louridas P, Ebert C. Machine Learning [J]. Ieee Software, 2016,33(5): 110−115.

[13] Fedik N, Zubatyuk R, Kulichenko M, et al. Extending machine learning beyond interatomic potentials for predicting molecular properties (vol 6, pg 653, 2022) [J]. Nature Reviews Chemistry, 2022, 6(12):916−916.

[14] 何家乐,陈景文,王杰琼,等.环境水体中化学品光化学持久性的预测模型 [J]. 科学通报, 2024,69(6):731−745.
He J L, Chen J W, Wang J Q, et al. Prediction models for photochemical persistence of chemicals in aquatic environments (in Chinese). Chin Sci Bull, 2024,69:731–745.

[15] Qin L T, Zhang X, Chen Y H, et al. Predictive QSAR models for the toxicity of disinfection byproducts [J]. Molecules, 2017,22(10):11.

[16] Cheng Z W, Yang B W, Chen Q C, et al. Quantitative relationships between molecular parameters and reaction rate of organic chemicals in Fenton process in temperature range of 15.8°C−60°C [J]. Chemical Engineering Journal, 2018,350:534−540.

[17] Ghanbari−Adivi E. A new machine learning model for predicting the water quality index [J]. Modeling Earth Systems and Environment, 2024,10(4):5635−5667.

[18] Cheng Z W, Yang B W, Chen Q C, et al. 2D−QSAR and 3D−QSAR simulations for the reaction rate constants of organic compounds in ozone−hydrogen peroxide oxidation [J]. Chemosphere, 2018,212:828−836.

[19] Jia L J, Shen Z M, Guo W M, et al. QSAR models for oxidative degradation of organic pollutants in the Fenton process [J]. Journal of the Taiwan Institute of Chemical Engineers, 2015,46:140−147.

[20] Cheng Z W, Yang B W, Chen Q C, et al. Characteristics and difference of oxidation and coagulation mechanisms for the removal of organic compounds by quantum parameter analysis [J]. Chemical Engineering Journal, 2018,332:351−360.

[21] Fizir M, Richa A, Touil S, et al. Comparing adsorption properties of algerian kaolinite towards phosphate with those of activated carbon: adsorption experiments, molecular modeling and an initiative toward hydroponic wastewaters remediation [J]. Water Air and Soil Pollution, 2024,235(6):19.

[22] Elmorsy M R, Abdel−Latif E, Badawy S A, et al. Molecular geometry, synthesis and photovoltaic performance studies over 2−cyanoacetanilides as sensitizers and effective co−sensitizers for DSSCs loaded with HD−2 [J]. Journal of Photochemistry and Photobiology a−Chemistry, 2020,389:8.

[23] Zhu H C, Guo W M, Shen Z M, et al. QSAR models for degradation of organic pollutants in ozonation process under acidic condition [J]. Chemosphere, 2015,119:65−71.

[24] Mauri A. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints [M]//Roy K. Ecotoxicological QSARs. New York, NY: Springer US. 2020:801−820.

[25] Bieroza M, Baker A, Bridgeman J. New data mining and calibration approaches to the assessment of water treatment efficiency [J]. Advances in Engineering Software, 2012,44(1):126−135.

[26] Sperandei S. Understanding logistic regression analysis [J]. Biochemia Medica, 2014,24(1):12−18.

[27] Lyu J, Zheng P J, Qi Y, et al. LightGBM−LncLoc: A LightGBM−based computational predictor for recognizing long non−coding RNA subcellular localization [J]. Mathematics, 2023,11(3):13.

[28] 魏佳妹,袁书娟,孔闪闪,等.轻梯度提升机算法的发展与应用 [J]. 计算机工程与应用, 2025,61(5):32−42.
Wei J M, Yuan S J, Kong S S, et al. Development and application of light gradient boosting machine [J]. Computer Engineering and Applications, 2025,61(5):32−42.

[29] Gupta V, Kumar E. H^3O−LGBM: hybrid Harris hawk optimization based light gradient boosting machine model for real−time trading [J]. Artificial Intelligence Review, 2023,56(8):8697−8720.

[30] Zhang D Y, Gong Y C. The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure [J]. Ieee Access, 2020,8:220990−221003.

[31] Qin H, Chen J W, Wang Y, et al. Development and assessment of quantitative structure−activity relationship models for bioconcentration factors of organic pollutants [J]. Chinese Science Bulletin, 2009,54(4): 628−634.

[32] Hanser T, Barber C, Guesne S, et al. Applicability domain: Towards a more formal framework to express the applicability of a model and the confidence in individual predictions [M]. Springer International Publishing Ag, Gewerbestrasse 11, Cham, Ch−6330, Switzerland, 2019.

[33] Khoei T T, Slimane H O, Kaabouch N. Deep learning: systematic review, models, challenges, and research directions [J]. Neural Computing & Applications, 2023,35(31):23103−23124.

[34] Rodríguez−Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values [J]. Journal of Medicinal Chemistry, 2020,63(16):8761−8777.

[35] Baptista M L, Goebel K, Henriques E M P. Relation between prognostics predictor evaluation metrics andlocal interpretability SHAP values [J]. Artificial Intelligence, 2022,306:22.

[36] Khosravi A, Koury R N N, Machado L, et al. Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms [J]. J. Clean Prod, 2018,176:63−75.

[37] Chirico N, Gramatica P. Real external predictivity of QSAR Models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection [J]. Journal of Chemical Information and Modeling, 2012,52(8):2044−2058.

[38] Kiralj R, Ferreira M M C. Basic Validation procedures for regression models in QSAR and QSPR studies: Theory and application [J]. Journal of the Brazilian Chemical Society, 2009,20(4):770−787.

[39] 王中钰,陈景文,傅志强,等.QSAR 模型应用域的表征方法 [J]. 科学通报, 2022,67(3):255−266.
Wang Z Y, Chen J W, Fu Z Q, et al. Characterization of applicability domains for QSAR models [J]. Chinese Science Bulletin, 2022,67(3): 255−266.

[40] 赵清波,徐勇勇,夏结来.logistic 回归中高杠杆点的检测 [J]. 中国卫生统计, 1997,(2):19−22.
Zhao Q B, Xu Y Y, Xia L J. Detection of high leverage points in logistic regression [J]. Chinese Journal of Health Statistics, 1997,(2): 19−22.

[41] Li Q, Yang H, Hao N, et al. Biodegradability analysis of dioxins through in silico methods: Model construction and mechanism analysis [J]. Journal of Environmental Management, 2023,345:9.

[42] Singh P, Kumar R, Sharma B K, et al. Topological descriptors in modeling malonyl coenzyme A decarboxylase inhibitory activity: N−Alkyl−N−(1,1,1,3,3,3−hexafluoro−2−hydroxypropylphenyl)amide derivatives [J]. Journal of Enzyme Inhibition and Medicinal Chemistry, 2009,24(1):77−85.

[43] Abdulrahman H L, Uzairu A, Uba S. Computer modeling of some anti−breast cancer compounds [J]. Structural Chemistry, 2021,32(2): 679−687.

[44] Mamy L, Patureau D, Barriuso E, et al. Prediction of the fate of organic compounds in the environment from their molecular properties: A review [J]. Critical Reviews in Environmental Science and Technology, 2015,45(12):1277−1377.

[45] Che Z P, Zhang S Y, Shao Y H, et al. Synthesis and quantitative structure−activity relationship (QSAR) study of novel n−arylsulfonyl−3−acylindole arylcarbonyl hydrazone derivatives as nematicidal agents [J]. Journal of Agricultural and Food Chemistry, 2013,61(24):5696−5705.

[46] Prajapati L M, Parmar V K, Patel M J, et al. Development of QSAR model for prediction of fold selectivity of phenyl benzoxazole as estrogen receptor inhibitors [J]. Medicinal Chemistry Research, 2012, 21(11):3681−3686.

[47] Klein C T, Kaiser D, Ecker G. Topological distance based 3D descriptors for use in QSAR and diversity analysis [J]. Journal of Chemical Information and Computer Sciences, 2004,44(1):200−209.

[48] Devinyak O, Havrylyuk D, Lesyk R. 3D−MoRSE descriptors explained [J]. Journal of Molecular Graphics & Modelling, 2014,54: 194−203.

作者简介：于林堂(1974−),男,江苏扬州人,副研究员,硕士,主要从事水污染化学方面研究.发表论文 20 余篇.ltyu@yzu.edu.cn.