刘 冰,时凯歌,陈燕敏,等.基于可解释机器学习的地表水 DOM 光化学性质预测 [J]. 中国环境科学, 2025,45(7):3925-3934.
Liu B, Shi K G, Chen Y M, et al. Prediction of DOM photochemical properties in surface water based on interpretable machine learning [J]. China Environmental Science, 2025,45(7):3925-3934.

基于可解释机器学习的地表水 DOM 光化学性质预测

刘 冰[1*],时凯歌[1],陈燕敏[1],卢 鑫[1],韩帅军[1],古 励[2] (1.郑州师范学院化学化工学院,河南 郑州 450044；2.重庆大学环境与生态学院,重庆 400044)

摘要：以 DOM 光化学产物三重激发态 DOM(^{3}DOM*)、单线态氧($^1O_2$)、羟基自由基(•OH)作为研究目标,分别使用 3 种回归算法构建机器学习(ML)模型,评估环境因素、紫外-可见光参数和荧光参数作为特征的重要性.结果表明,分类梯度提升算法 ML 模型在测试集中对预测目标的表现更好,R^2 分别为 0.9344、0.8091 和 0.7775.温度、DOC 和 pH 值分别是预测 ^{3}DOM*、$^1O_2$ 和•OH 的最重要特征.应用分类梯度提升算法链式 ML 模型时,对 $^1O_2$ 和•OH 预测性能有所优化,$^1O_2$ 在测试集的 R^2 由 0.8091 增加至 0.8233,RMSE 由 1.322 降至 0.418; •OH 在测试集的 R^2 由 0.7775 上升至 0.8094,RMSE 由 2.982 降至 0.312.另外,ML 模型对 ^{3}DOM*、$^1O_2$ 和•OH 的预测表现均优于偏最小二乘法回归模型,表明 ML 模型对非线性关系具有较强的处理能力.基于 ML 的特征重要性分析识别出 DOM 光化学反应中的关键特征,可为众多影响因素的集约化处理提供直接证据.

关键词：溶解性有机物；可解释；机器学习；光化学；预测
中图分类号：X52 **文献标识码**：A **文章编号**：1000-6923(2025)07-3925-10

Prediction of DOM photochemical properties in surface water based on interpretable machine learning. LIU Bing[1*], SHI Kai-ge[1], CHEN Yan-min[1], LU Xin[1], HAN Shuai-jun[1], GU Li[2] (1.School of Chemistry and Chemical Engineering, Zhengzhou Normal University, Zhengzhou 450044, China；2.College of Environment and Ecology, Chongqing University, Chongqing 400044, China). *China Environmental Science*, 2025,45(7)：3925~3934

Abstract：Dissolved organic matter (DOM) in surface water, which serves as a significant component within the carbon cycle, plays a vital role in the biogeochemical cycle of aquatic ecosystems. The triplet excited state of DOM (^{3}DOM*), singlet oxygen ($^1O_2$), along with hydroxyl radicals (•OH), which are the photochemical products generated from DOM, were taken as the research targets. Three regression algorithms were employed to constructed the machine learning (ML) models predictive models for evaluating importance of features including environmental factors, ultraviolet-visible light parameters and fluorescence parameters. The ML model based on the categorical boosting algorithm exhibited better performance for the prediction targets in the test set, with R^2 values of 0.9344, 0.8091, and 0.7775, respectively. Temperature, DOC, and pH were identified as the most important features for predicting ^{3}DOM*, $^1O_2$ and •OH, respectively. When applying the chained ML model incorporating with the categorical boosting algorithm, the prediction performance for $^1O_2$ and •OH was further optimized. For $^1O_2$ the R^2 value in the test set increased from 0.8091 to 0.8233, and the RMSE decreased from 1.322 to 0.418. For •OH the R^2 value in the test set rose from 0.7775 to 0.8094, and the RMSE dropped from 2.982 to 0.312. Additionally, the ML models perform better than the partial least squares regression model in predicting ^{3}DOM*, $^1O_2$ and •OH, indicating that the ML models possessed strong capability to handle non-linear relationships. The feature importance analysis based on ML successfully distinguished the key features in the DOM photochemical reaction. This finding can provide direct evidence for the intensive management of numerous influencing factors.

Key words：dissolved organic matter；explainable；machine learning；photochemistry；prediction

地表水中溶解性有机物(DOM)是碳循环中重要组成部分,也是水生生物的重要营养来源[1].由于 DOM 化学成分复杂和结构多样,直接和间接光化学反应是影响 DOM 转化的重要因素.太阳光辐射可改变地表水中 DOM 的化学性质和组成[2],例如:地表上层水中 DOM 受到阳光照射后,导致 DOM 直接光化学降解,生成无机物盐和 CO_2[3];也可通过间接光化学反应改变 DOM 分子结构,导致其消毒副产物生成势增加[4].另一方面,DOM 的光化学反应在保护水生生物免受紫外线辐射伤害和增强有机污染物转化

收稿日期：2024-11-21
基金项目：国家自然科学基金资助项目(51208448);河南省高等学校重点科研项目(24B610016);河南省科技厅科技攻关项目(222102910508)
* 责任作者, 副教授, liubing_982002@163.com

等方面起到关键作用,具有重要的水环境意义[5-6].

自然水体中 DOM 包含的发色溶解性有机物(CDOM)由吸光有机分子组成,是重要的光化学反应组分[7].CDOM 在光照下可敏化许多光化学反应,产生几种活性中间体,如 ^{3}DOM*、$^1O_2$、•OH 和过氧化氢(H_2O_2)等,统称为光生活性中间体(PPRIs)[7].部分 PPRIs 由 ^{3}DOM*引导产生,例如:^{3}DOM*可以通过能量转移将水体中的溶解氧转化为 $^1O_2$,$^1O_2$ 是一种以富电子官能团为目标的活性氧,并通过依赖/不依赖 H_2O_2 两种途径生成•OH[8-9].^{3}DOM*既是 PPRIs 的生产者,也是重要的氧化剂[8].目前,DOM 的光化学特性即生成 PPRIs 的能力一般用表观量子产率(Φ_{PPRIs})来表示[8-9].由于 PPRIs 在水溶液中存在时间较短(10^{-6}~1s),Φ_{PPRIs} 直接测定方法比较繁琐,需借助于大型仪器设备或特殊试剂.另外,DOM 的光化学性质受许多因素影响,如环境因素(温度、pH 值、DOC)、结构和来源($SUVA_{254}$、E_2/E_3、FIX、BIX 和 HIX)等.有文献研究结果•OH 表观量子产率($\Phi_{\bullet OH}$)与 $SUVA_{254}$ 呈正相关[10];另有研究表明 $SUVA_{254}$ 与 Φ_{PPRIs} 之间呈负相关[11].所以,需要不断发展精确、稳定的预测方法.

ML 可以处理复杂和多维度数据,并从中学习构建预测模型,已经应用在材料吸附、生物降解、气体排放、水质评估、微塑料和光降解等领域[12-16].例如:随机森林算法 ML 模型用来预测城市固体废物焚烧厂 CO_2 排放量和浓度[12];梯度提升树和人工神经网络算法 ML 模型预测溶解性有机碳吸附去除效率[13];广义加法和多层感知器算法 ML 模型预测大气中含碳气溶胶和识别它们的来源[14].随机森林和人工神经网络算法 ML 模型还被用于监测和预测地下水中污染物,并评估地下水污染的风险程度[15].此外,ML 模型可准确地预测生物炭的产率,以及对金属和微污染物的吸附去除率[16].

本研究选择 Φ_{PPRIs}(Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\bullet OH}$)作为研究对象,采用分类梯度提升器(CB)、自适应提升器(AB)和极限梯度提升器(XB)3 种树模型算法,对 DOM 光化学反应影响因素的重要性进行排序,揭示水环境参数、紫外-可见光参数和荧光参数作为对 DOM 光化学反应的影响规律,旨在构建一个准确率和可靠性较高的预测模型,为 DOM 光学特性与光化学特性之间的联系提供参考.

1 材料与方法

1.1 数据选择

地表水(仅河流、湖泊)中 DOM 光化学特性数据收集选用数据库为 web of science,keyword 选用 dissolved organic matter、river、lakes、photochemistry、reactive intermediate、excited triplet state、singlet oxygen 和 hydroxylradical 等.(1)考虑参考文献实验条件和方法有一定差异性,首先保持数据独立性,不做前期处理和公式换算;(2)尽量参考同一篇文献中同时测量 3 种预测目标(Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\bullet OH}$),筛查和避免重复值;(3)根据 ^{3}DOM*、$^1O_2$ 和•OH 生成机理,选择环境因素(温度、pH 值、DOC)、紫外-可见光参数($SUVA_{254}$、E_2/E_3)和荧光参数(FIX、BIX、HIX)作为特征;(4)DOM 光化学特性研究文献中,氙灯、汞灯和 UVA 灯都可作为光源,由于氙灯光谱与太阳光相似,为了更好地预测地表水体中 DOM 光化学性质,只选择氙灯实验数据;(5)收集 11a 中(2013~2024)的 45 份同行评议出版物,经初步筛选提取了 531 个数据点,组成 531 行、11 列数据集.

1.2 数据填充

数据集缺失的数据主要集中在预测目标(Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\bullet OH}$),缺失占比分别为 9.06%、3.02%和 16.04%.首先使用均值插补、贝叶斯多重插补、K 最近邻插补和 RF 插补,再通过 CB 算法对插补后的数据集进行建模,用平均绝对误差(MAE)评估预测精度,RF 插补后数据集 MAE 较小,后续处理使用 RF 插补的数据集.

1.3 数据标准化处理

为消除数据集中不同特征之间量纲的影响,研究中采用 Z-score 标准化处理方法.将原始数据的均值(mean)和标准差(standard deviation)进行数据的标准化,使用公式(1)进行计算:

$$Z_i = \frac{(X_i - \mu)}{\sigma} \tag{1}$$

式中:Z_i 是标准化后的数据;X_i 为需要标准化的数据;μ 是数据的均值;σ 是数据的标准差.

1.4 相关性分析和偏最小二乘法回归(PLSR)模型

采用 SPSS 27 计算特征变量和目标变量之间 Spearman 等级相关系数,Spearman 等级相关矩阵对

变量之间的相关性进行分析,考察变量之间的相互依赖关系.PLSR 常用于研究多元线性回归中自变量间的相关问题,把特征变量(温度、pH 值、 DOC、$SUVA_{254}$、E_2/E_3、FIX、BIX 和 HIX)和目标变量(Φ_{3DOM^*}、$\Phi_{^1O2}$ 和 $\Phi_{\bullet OH}$)进行 PLSR 建模,可识别出最优的正交主成分(SPSS 27).

1.5　ML 模型建立

使用 Anaconda.3 和 python.3.9 中相关库建立 ML 模型,主要包括算法的筛选以及最优参数调整.首先,70%数据集划分为训练集,30%数据集划分为测试集,训练集用于模型的构建,测试集评价模型的预测性能和泛化能力.采用 CB、AB 和 XB 算法构建 ML 模型,模型开发过程采用 5 折交叉验证.最后用 R^2 和 RMSE 两个指标评价模型的预测性能,R^2 越高和 RMSE 越小表示模型预测性能越好.

1.6　ML 模型解释

ML 模型解释采用排列重要性考察特征的重要程度.(1)树模型基于基尼指数或熵选择分割节点,节点的最佳选择导致最小的基尼指数,在此基础上计算每个特征重要性得分[17],判断特征之间的相对重要程度.(2)Shapley 加和解释(SHAP)通过计算每个样本中各个特征的 SHAP 值来衡量各特征对该样本模型预测值的贡献,能进行全局和局部解释[18].本文选择 SHAP 值和特征重要性相结合对 ML 模型分析,为理解每个特征对模型预测的重要程度提供解释.

2　结果与讨论

2.1　数据集分布范围

如表 1 所示,变异系数(CV)范围为 9.139%~152.573%,温度和 pH 值变异系数(CV)相对较小(13.024%和 9.139%).但是,总体数据差异显著,分异程度较高.从 Φ_{3DOM^*}、$\Phi_{^1O2}$ 和 $\Phi_{\bullet OH}$ 数据可以知道,Φ_{3DOM^*}和 $\Phi_{^1O2}$ 较为相近,$\Phi_{\bullet OH}$ 值最小(相比 Φ_{3DOM^*} 和 $\Phi_{^1O2}$ 值小两个数量级).研究表明,易于产生 ^{3}DOM*的 DOM 类物质,也易于生成 $^1O_2$,Φ_{3DOM^*}和 $\Phi_{^1O2}$ 之间相关系数约为 0.89[19].•OH 具有较高的亲电性和氧化还原电位,反应速率较快.另外,•OH 氧化活性比其它活性氧物质强,但其稳定性低和存在时间很短,还可以无选择性地与−C=C−、C−S 和 C−H 键反应,容易被化学反应清除和导致其浓度不高[20].

表 1　数据集中预测目标和特征统计值

Table 1　Statistical values of targets and feaures in the datasets

指标	最小值	最大值	均值	标准差	变异系数(%)
温度(℃)	20.000	30.000	24.900	3.24305	13.024
pH 值	5.950	9.700	7.372	0.67369	9.139
DOC(mg/L)	1.344	49.200	8.824	7.74418	87.763
$SUVA_{254}$[L/(mg C·m)]	0.119	27.314	5.828	4.28625	73.546
E_2/E_3	1.685	30.000	5.839	2.82746	48.421
FIX	0.590	3.110	1.616	0.33979	21.026
BIX	0.240	0.990	0.813	0.11826	14.548
HIX	0.080	1.559	0.689	0.21161	30.728
Φ_{3DOM^*}(10^{-2}mol/mol−photons)	0.141	26.000	3.106	3.13574	100.969
$\Phi_{^1O2}$(10^{-2}mol/mol−photons)	0.230	15.590	3.430	2.36345	68.898
$\Phi_{\bullet OH}$(10^{-4}mol/mol−photons)	0.013	7.430	0.582	0.88758	152.573

2.2　特征变量和预测目标相关性分析

从图 1 中可知,目标变量 Φ_{3DOM^*}、$\Phi_{^1O2}$ 和 $\Phi_{\bullet OH}$ 之间存在显著的正相关性.其中,Φ_{3DOM^*}和 $\Phi_{^1O2}$ 之间的 Spearman 等级相关系数为 0.57(P≤0.001),表示为中等相关程度;Φ_{3DOM^*}和 $\Phi_{\bullet OH}$ 之间 Spearman 等级相关系数为 0.36(P≤0.001),$\Phi_{^1O2}$ 和 $\Phi_{\bullet OH}$ 之间的 Spearman 等级相关系数为 0.40(P≤0.001).这主要与 ^{3}DOM*、$^1O_2$ 和•OH 的生成机制相关,^{3}DOM*是产生 $^1O_2$ 的主要前体物,^{3}DOM*通过能量转移把 O_2 形成 $^1O_2$,^{3}DOM*量子产率系数(f_{TMP})和 $\Phi_{^1O2}$ 相互关联[21-22].•OH 则是通过电子转移反应形成,^{3}DOM*、硝酸盐和 O_2 等可能都是•OH 的前体物[21-22],这导致 Φ_{3DOM^*}、$\Phi_{^1O2}$ 和 $\Phi_{\bullet OH}$ 之间的相关性差异.

从图 2 可知,Φ_{3DOM^*}、$\Phi_{^1O2}$ 和 $\Phi_{\bullet OH}$ 与 pH 值呈正相关,相关系数分别为 0.08(P≤0.001)、0.34(P≤0.001)和 0.21(P≤ 0.001).当水体环境为酸性时,H^+会与 DOM 结构外层的负电荷中和,H^+增加到一定程度后,DOM 分子间相互作用力的降低,促使分子的聚合从而导致 DOM 分子变大,促进 DOM 内部结构的荧光淬灭反应[23-24].在碱性条件下,随着 pH 值的增加,由于去质子化的酚部分可增强内部电荷转移,OH^-可作为•OH 的引发剂和促进剂[24-25].这说明 pH 值在 6~9 时,随着 pH 值增加可提高 DOM 光化学反应.

与紫外−可见光参数($SUVA_{254}$、E_2/E_3)相关系数可以看出,Φ_{3DOM^*}、$\Phi_{^1O2}$ 和 $\Phi_{\bullet OH}$ 与 $SUVA_{254}$ 呈负相关,相关系数分别为−0.02(P≤0.001)、−0.11(P≤0.001)和−0.53(P≤0.001).$SUVA_{254}$ 表示 DOM 芳香性,其值与 DOM 所含芳香性化合物的含量呈正相关[26].这

表明随着 $SUVA_{254}$ 的增大,DOM 的 PPRIs 光生成效率降低,这与先前研究相一致[27-28].Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 与 E_2/E_3 呈正相关,相关系数分别为 0.31($P\leqslant$0.001)、0.39($P\leqslant$0.001)和 0.16($P\leqslant$0.001).E_2/E_3 表示 DOM 相对分子量大小,其值与分子量大小成反比[26].先前研究发现,E_2/E_3 与 Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 之间呈正相关,小分子量 DOM 更能激发 ^{3}DOM*和 $^1O_2$ 产生[26].与 ^{3}DOM*和 $^1O_2$ 相比,E_2/E_3 与 $\Phi_{\cdot OH}$ 相关系数较小[26].另外,E_2/E_3 与 Φ_{3DOM*}和 Φ_{1O2} 的相关系数(0.31 和 0.39)比先前文献研究中值要小[28-29],这可能主要是数据集中 DOM 的类型和来源比较复杂,降低了分子量对 Φ_{3DOM*}和 Φ_{1O2} 的影响.

与荧光光谱参数(FIX、BIX、HIX)相关系数可以知道,Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 与 FIX 呈正相关,相关系数分别为 0.02($P\leqslant$0.001)、0.11($P\leqslant$0.001)和 0.53($P\leqslant$0.001).Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 与 BIX 也呈正相关,相关系数分别为 0.05、0.26 和 0.62($P\leqslant$0.001).FIX 确定陆地和微生物源 DOM 所占比重,其值越大表示 DOM 中所包含的微生物来源 DOM 含量越高[30];BIX 代表新生微生物源 DOM 所占比重,其值越大表示新生微生物源 DOM 含量越大[30].这说明与陆源 DOM 相比,微生物源 DOM 中的新生 DOM 更能提高 ^{3}DOM*、$^1O_2$ 和•OH 光生成效率[27-30].这与文献研究相一致,微生物源 DOM 的光反应活性比陆源 DOM 强[27-30].Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 与 HIX 呈负相关,相关系数分别为 −0.003($P\leqslant$0.001)、−0.18($P\leqslant$0.001)和 −0.50($P\leqslant$0.001).HIX 代表了 DOM 中腐殖质含量或腐殖化程度,其值较高时表示 DOM 腐殖化程度高或所含腐殖质丰富[27-30].这说明 DOM 腐殖化程度高或所含腐殖质丰富时,导致 Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 光生成率降低.

另外,FIX 和 BIX 呈正相关,相关系数为 0.69($P\leqslant$0.001),这主要是 FIX 和 BIX 都与微生物源 DOM 相关.FIX 和 BIX 与 $\Phi_{\cdot OH}$ 呈正相关(相关系数分别 0.59 和 0.62),也表明荧光光谱参数改变可以影响到 Φ_{PPRIs} 的变化,特别是对 $\Phi_{\cdot OH}$ 的影响更为明显.

2.3 PLSR 模型预测 DOM 光化学性质

从图 1 可知,温度、pH 值、DOC、$SUVA_{254}$、E_2/E_3、FIX、BIX、HIX、Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 之间都具有一定线性关系.因此,可使用表观量子产率(Φ_{3DOM*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$)作为预测目标,温度、pH、DOC、$SUVA_{254}$、E_2/E_3、FIX、BIX 和 HIX 作为特征变量进行 PLSR 建模,结果如图 2 和图 3 所示.

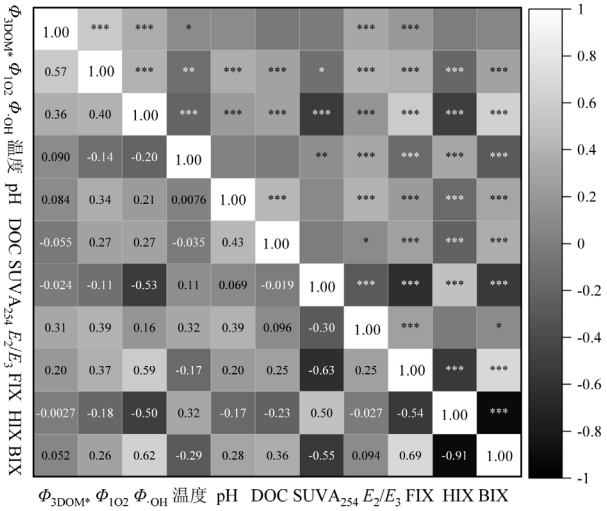
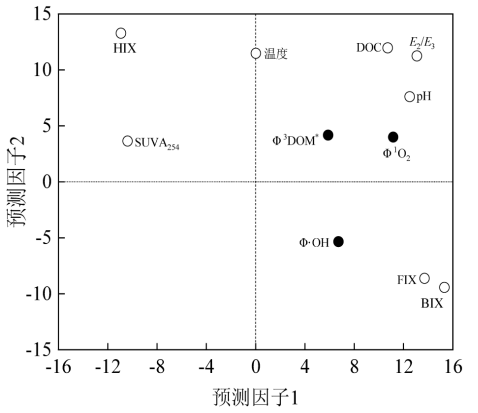


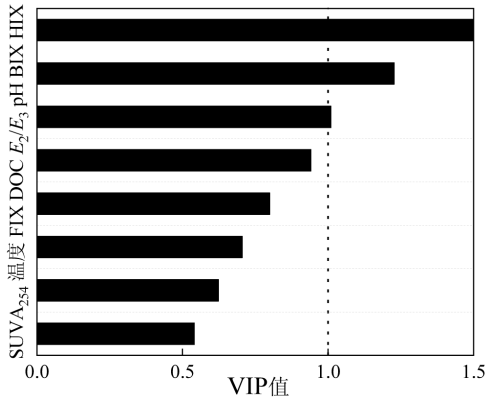
图 1 数据集的 Spearman 等级相关性矩阵

Fig.1 Spearman rank correlation matrix for the datasets

显著性标记"*":$P\leqslant$0.05;"**":$P\leqslant$0.01;"***":$P\leqslant$0.001



(a) 载荷散点图



(b) VIP 得分图

图 2 PLSR 分析得出的载荷散点图和特征变量 VIP 得分

Fig.2 Loading scatter plots and the VIP score of feature variables for the PLS analysis

从 PLSR 分析的变量载荷散点图(图 2(a))中可知,Φ_{3DOM^*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 分别在第一预测主成分的正半轴,Φ_{1O2} 比 Φ_{3DOM^*}和 $\Phi_{\cdot OH}$ 更靠近第一预测主成分轴,说明 Φ_{1O2} 主要受第一预测主成分的正影响.Φ_{3DOM^*}比 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 更靠近第二预测主成分轴,说明 Φ_{3DOM^*}主要受第二预测主成分的正影响.特征变量 VIP 得分图如图 2(b)所示,HIX、BIX 和 pH 值的 VIP 得分大于 1.其中,HIX 的 VIP 得分最高,对 PLSR 模型性能的影响最大,这也表明 DOM 荧光特性对 PLSR 模型预测的重要性.

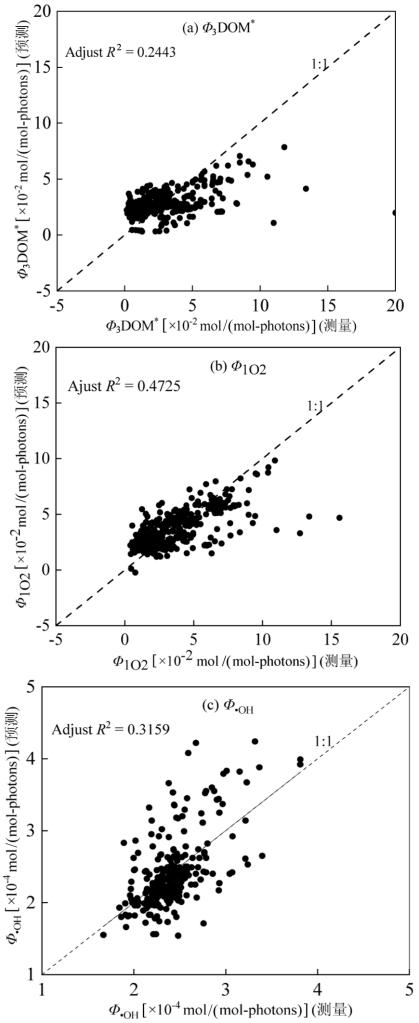


图 3　测量值 Φ_{PPRIs} 与 PLSR 模型预测值 Φ_{PPRIs} 散点图

Fig.3　Scatter plots of measured Φ_{PPRIs} versus Φ_{PPRIs} as predicted by PLSR model

如图 3 所示,PLSR 模型对 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 预测性能 (R^2=0.4725 和 0.3159) 比 Φ_{3DOM^*} 高 (R^2=0.2443),这表明数据集中特征变量的选择更能说明地表水中 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的变化.Φ_{3DOM^*}和 Φ_{1O2} 的 PLSR 模型预测能力呈现明显差距(R^2=0.2443 和 0.4725),这较大差异表示基于特征变量的 PLSR 模型更倾向于解释 $^3DOM^*$的能量转移途径而不是电子转移途径[30].另外,需要更多与 $^3DOM^*$ 的电子转移途径有关的特征才能进一步解释 $^3DOM^*$变化的主导因素.

2.4　ML 模型预测 DOM 光化学性质

2.4.1　ML 模型性能评价　数据集采用切分比例 7:3 和交叉验证 5 次,由表 2、表 3 可知,3 种算法 ML 模型学习能力较强,模型基本学习到训练样本中的信息,对预测目标 Φ_{3DOM^*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的 R^2 都比较接近 1.其中,CB 算法在测试集对预测目标 Φ_{3DOM^*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的表现都比 AB 和 XB 算法更好,R^2 分别为 0.9344、0.8091 和 0.7775.这说明 CB 算法 ML 模型预测效果更加精确和具有较强泛化能力.3 种算法 ML 模型在训练集或测试集中,Φ_{3DOM^*}的 R^2 都比 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 要高,这可能主要与 $^3DOM^*$、$^1O_2$ 和•OH 生成机理有关[7-9](图 4).从图 5 可知, DOM 在光照条件下,能够吸收辐射转变为 $^1DOM^*$, $^1DOM^*$可以通过荧光发射或无辐射跃迁等过程释放能量回到基态,或者通过自旋轨道耦合发生系间窜越到达 $^3DOM^*$,$^3DOM^*$与 O_2 之间通过能量转移生成 $^1O_2$;通过电子转移生成•OH[7-9].因为数据集中选择的特征与 $^3DOM^*$的生成都有直接关系,对 Φ_{3DOM^*}的预测表现更好.

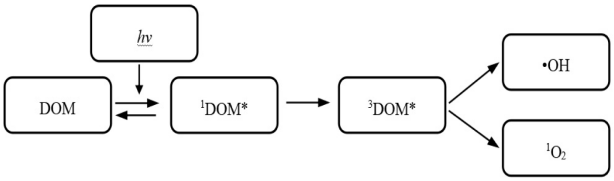


图 4　光生成 PPRIs 机理示意

Fig.4　Schematic of PPRIs photogeneration mechanism

相对 AB、XB 算法,CB 算法 ML 模型预测性能更好,以下模型解释都以 CB 模型为例.

从表 3 可见,数据集 CB 模型中 Φ_{3DOM^*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的 RMSE,测试集都比验证集要高,这表明 CB 模型的预测值离散程度较高,结合图 5 可以看出,测试集的预测值分布与测量值相比有所收窄,具体表现为 CB 模型对较大/较小测量值的预测时存在较大的偏差,这也是导致 RMSE 较高和预测值离散程度较大的主要原因.这种表现在 $\Phi_{\cdot OH}$ 测试集中更

为明显,对 $\Phi_{\cdot OH}$ 的预测值存在一定的波动变化.在 Φ_{3DOM^*}、$\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 数据分布较为集中的区间,CB 模型的预测表现较好,这是因为数据集中的区间数据量大,较多相似数据值能被 CB 模型重复学习,并取得较好的预测结果.这说明 ML 的数据集存在有效训练区间,由训练集数据分布决定.当测试集落在有效训练区间时,ML 模型可以表现出较好预测效果.所以,对有效区间大小的评价也可以在某种程度上表示模型的泛化能力.

表 2 三种 ML 模型优化后参数

Table 2 Optimize parameters of three ML models

模型	参数名	参数个数	参数值
CB	迭代次数	5	100
	学习率		0.1
	L2 正则项		1.5
	树的最大深度		10.0
	达成优化以后继续迭代的次数		20
AB	基分类器数量	4	100
	损失函数		linear
	基分类器		决策树分类器
	学习率		1.0
XB	基学习器	9	gbtree
	基学习器数量		100
	学习率		0.1
	L1 正则项		1.0
	L2 正则项		1.0
	样本征采样率		1.0
	树特征采样率		1.0
	节点特征采样率		1.0
	树的最大深度		10.0

2.4.2 CB 算法 ML 模型解释 利用 SHAP 值和树模型特征重要性对 CB 模型的结果进行解释,如图 5 和图 6 所示,在 SHAP 图中纵轴上特征顺序是根据 SHAP 值分析得到的特征重要性大小进行排列,从图 6(a)可知,对于 Φ_{3DOM^*} 的 CB 模型中,温度、FIX 和 pH 值是预测 Φ_{3DOM^*} 时的重要特征,其结果与基于树模型的特征重要性分析(图 7(a))相似.温度在 SHAP 图和树模型的特征重要性中排名第 1,表现为最高影响特性.这可能是当温度上升时,光化反应的速率也随之提高[29].FIX 在 SHAP 图和树模型的特征重要性中排名为第 2 和第 3,这说明 DOM 来源类型是 ^{3}DOM* 产生的主要影响因素,因为陆源/微生物源 DOM 能降低/提高 ^{3}DOM* 产生[26].多数数据点对应的 SHAP 值为负数,对 ^{3}DOM* 的产生有负贡献,这主要是多数数据点 FIX 中位数为 1.57(平均值为 1.62,表 1),小于 1.9,说明多数样品 DOM 中陆源 DOM 所占比例较大,对 Φ_{3DOM^*} 产生表现负影响.pH 值在 SHAP 图和树模型的特征重要性中排名为第 3 和第 2,这主要与 ^{3}DOM* 产生机制有关.H^+ 会与 DOM 结构外层的负电荷中和,促使分子的聚合从而导致 DOM 分子变大,导致 DOM 内部结构的荧光淬灭反应,影响 DOM 光化学反应[23-34].

表 3 Φ_{3DOM^*}, $\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 对应的 3 种 ML 模型评价结果

Table 3 Evaluation results of three ML models for Φ_{3DOM^*}, $\Phi_{^1O_2}$ and $\Phi_{\cdot OH}$

预测目标	模型	R^2 (训练)	MSE (训练)	RMSE (训练)	R^2 (测试)	MSE (测试)	RMSE (测试)
Φ_{3DOM^*}	CB	0.9828	0.405	0.636	0.9344	2.928	2.632
	AB	0.9889	0.194	0.441	0.8511	3.309	2.883
	XB	0.9899	0.026	0.161	0.8840	4.746	3.57
$\Phi_{^1O_2}$	CB	0.9947	0.032	0.179	0.8091	1.748	1.322
	AB	0.9986	0.010	0.101	0.7333	1.556	1.247
	XB	0.9996	0.026	0.161	0.7029	3.746	3.570
$\Phi_{\cdot OH}$	CB	0.9214	0.274	0.523	0.7775	3.894	2.982
	AB	0.9889	0.057	0.239	0.6829	4.846	3.584
	XB	0.9987	0.021	0.146	0.5687	4.228	3.351

对于 $\Phi_{^1O_2}$ 的 SHAP 图 6(b)中,DOC、E_2/E_3 和 FIX 是预测 Φ_{3DOM^*} 时的重要特征.其中,DOC 在 SHAP 图和树模型的特征重要性中(图 7(b))排名都是第 1,表现为最高影响特性,这说明 DOM 浓度水平是对 $\Phi_{^1O_2}$ 预测最重要的影响因素.据先前文献研究,随着 DOC 浓度不断增加,DOM 能够猝灭 ^{3}DOM*,导致 $^1O_2$ 产生率下降和 $\Phi_{^1O_2}$ 水平降低[27,30].E_2/E_3 在 SHAP 图和树模型的特征重要性中排名第 2,这也与 $^1O_2$ 产生的前体物 ^{3}DOM* 有关[9-10].E_2/E_3 表示 DOM 相对分子量,较小分子量 DOM 更能激发 ^{3}DOM* 和 $^1O_2$ 产生[26-27].另外,FIX 在 SHAP 图和树模型的特征重要性中排名分别第 3 和第 5,这主要是当 FIX 值增加到一定程度时,对 $^1O_2$ 产生的影响程度会趋于平缓,导致 FIX 对 $\Phi_{^1O_2}$ 的重要性也相应下降[27,30].

对于 $\Phi_{\cdot OH}$ 的 SHAP 图 6(c)中,pH 值、DOC 和温度是预测 $\Phi_{\cdot OH}$ 时的重要特征,从 $\Phi_{\cdot OH}$ 的树模型的特征重要性分析图 7(c)可知,pH 值在 SHAP 图和树模型的特征重要性中排名都是第 1,大多数据点对应的 SHAP 值为正数,表现为最高影响特征,这主要与 •OH 产生机制有关,同时 OH^- 是 •OH 的引发剂和促进

剂[23-25].从图 1 可以看出,pH 值的均值为 7.37,中位数为 7.20,大多数数据对•OH 的产生形成正向影响.DOC 和温度在 SHAP 图和树模型的特征重要性中排名比较一致,都是第 2 和第 3.

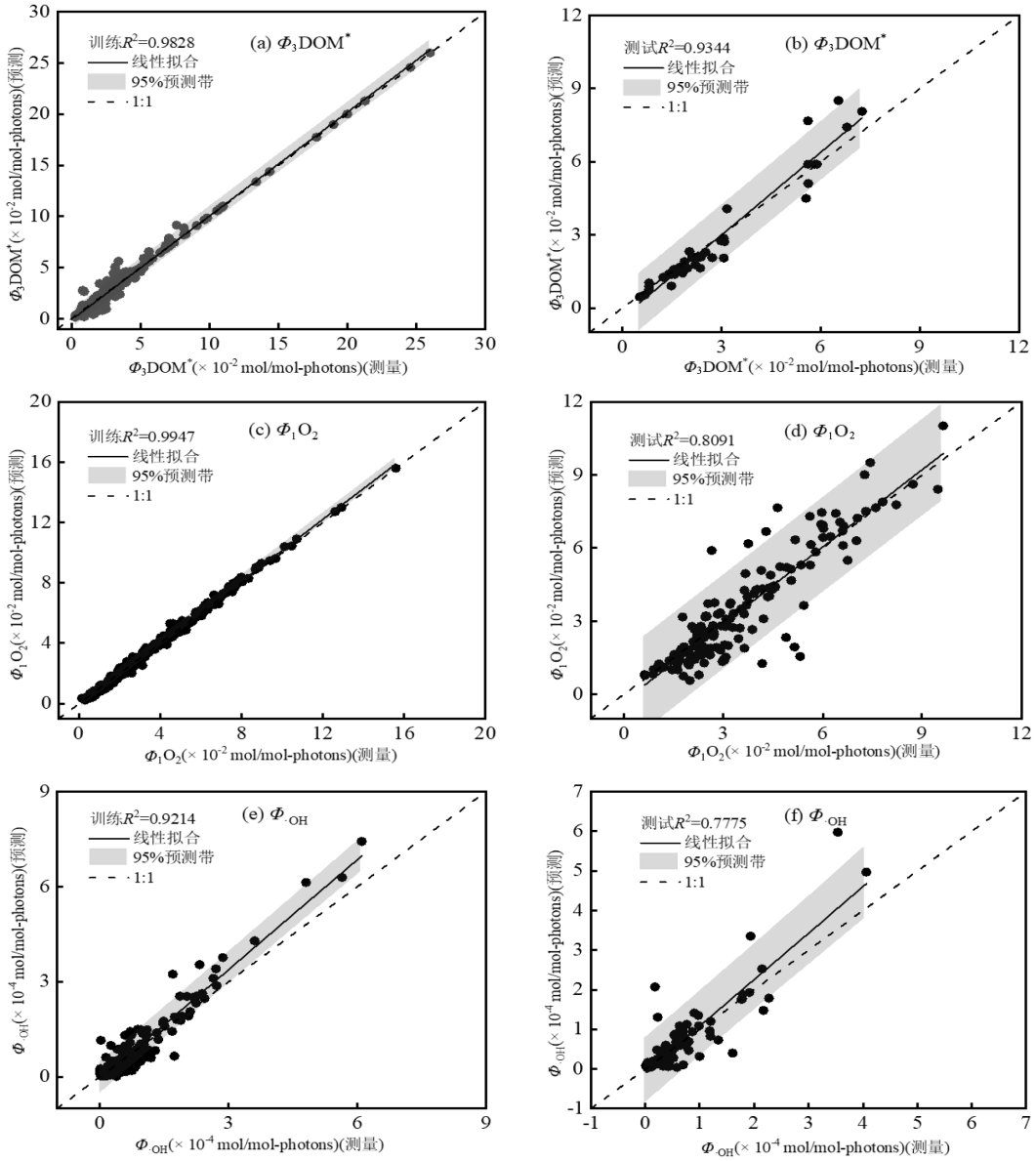
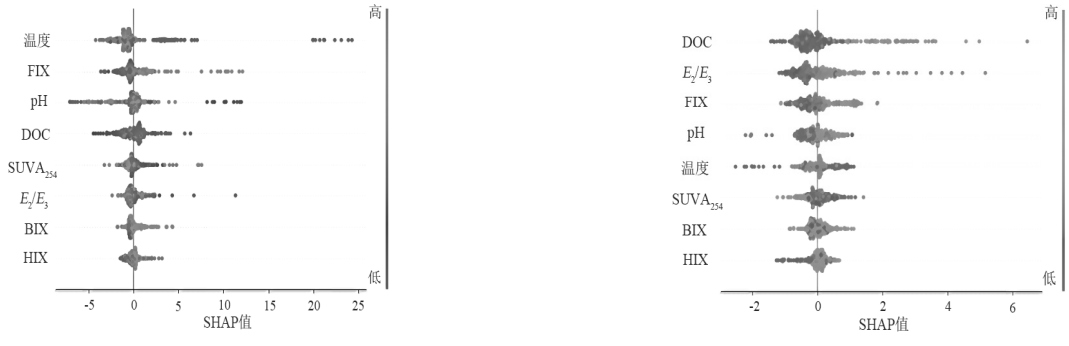
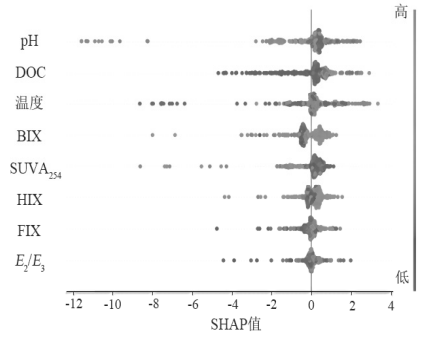


图 5　Φ_{3DOM*},Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的 CB 模型预测表现

Fig.5　CB model prediction of Φ_{3DOM*}, Φ_{1O2} and $\Phi_{\cdot OH}$



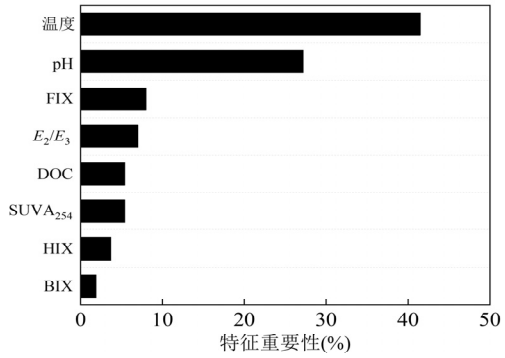


(c) $\Phi_{\cdot OH^*}$

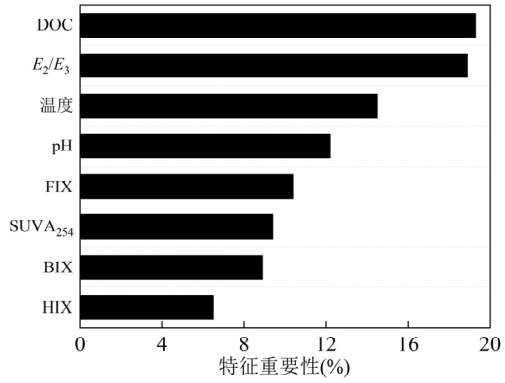
图 6 CB 模型的 SHAP 值

Fig.6 SHAP summary interpretation plots of the CB model

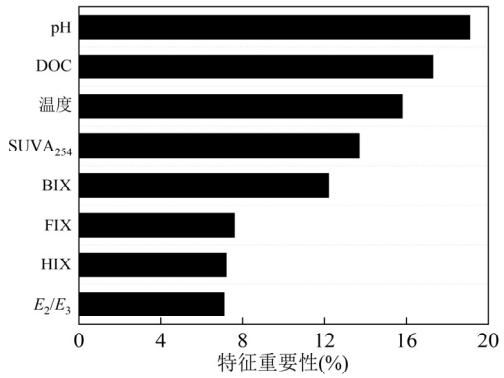
因此,结合 SHAP 值和树模型特征重要性两个方面来解释特征对 CB 模型的影响程度,在对模型的阐释起到关键作用.对于 Φ_{3DOM^*}、Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的 CB 模型表现中,虽然排名第 1 的重要特征各不相同,但是也有相同的特征,例如:DOC、温度、pH 值和 FIX.这说明对于 DOM 光化学性质而言,环境因素(温度、pH 值、DOC)始终对 PPRIs 产生有着重要影响.另外还得到了 Φ_{PPRIs} 与 DOM 的来源特性指标 FIX 之间的关联,表明 DOM 的来源组成变化对 CB 模型的影响程度也是一个重要特征.



(a) Φ_{3DOM^*}



(b) Φ_{1O2}



(c) $\Phi_{\cdot OH}$

图 7 CB 模型的树模型特征重要性

Fig.7 Feature importance based on tree model of the CB model

2.4.3 CB 链式 ML 模型对 DOM 光化学预测 在 ML 建模过程中加入已知关联特征,可使模型的性能进一步优化.从图 4 中 PPRIs 的生成机理可知,^{3}DOM*、$^1O_2$ 和•OH 之间存在一定的关联[7-8].根据以上关联特征,将预测目标 Φ_{3DOM^*}值作为关联特征来进一步对 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的 CB 模型预测进行优化,在预测过程中利用预测目标之间的关联来构建 CB 链式 ML 模型.

表 4 预测目标 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 对应的 CB 链式 ML 模型评价结果

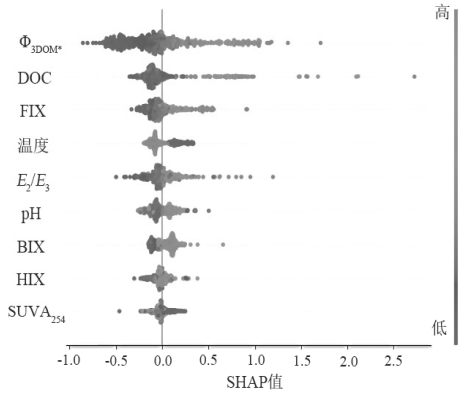
Table 4 Performance of CB chain ML model for Φ_{1O2} and $\Phi_{\cdot OH}$

预测目标	模型	R^2 (训练)	MSE (训练)	RMSE (训练)	R^2 (测试)	MSE (测试)	RMSE (测试)
Φ_{1O2}	CB	0.9959	0.008	0.089	0.8233	0.175	0.418
$\Phi_{\cdot OH}$	CB	0.9970	0.003	0.059	0.8094	0.098	0.312

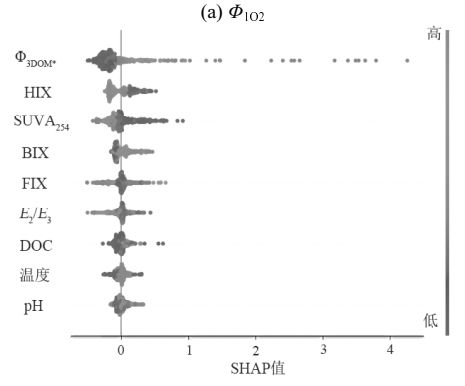
如表 4 和图 8 所示,与 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 的 CB 模型相比,链式 ML 模型预测性能得到优化.其中 Φ_{1O2} 在测试集的 R^2 由 0.8091 增加到 0.8233,RMSE 由 1.322 降到 0.418;$\Phi_{\cdot OH}$ 在测试集的 R^2 由 0.7775 上升到 0.8094,RMSE 由 2.982 降到 0.312.这说明 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 链式 ML 模型在学习过程中,更好地学到了 Φ_{3DOM^*}与 Φ_{1O2} 和 $\Phi_{\cdot OH}$ 之间的关系.两组链式 ML 模型的性能差异(0.8233 和 0.8094)主要与 PPRIs 的生成路径有关,其中 ^{3}DOM*被证实是产生 $^1O_2$ 的主要前体物,而•OH 产生还存在其它路径[7-8].从 SHAP 值图 8 可以看出,Φ_{3DOM^*}预测值对两组链式 ML 模型都

表现出最重要影响,在 SHAP 值的特征重要性中排第 1.

因此,结合 PPRIs 的生成机理,对 $\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 的 CB 模型优化是可行的,表明 CB 模型可以更好地学习变量之间的关系,如能把 PPRIs 生成路径加入模型中,也可以更准确地进行 DOM 光化学性质预测.



(a) $\Phi_{^1O_2}$



(b) $\Phi_{\cdot OH}$

图 8　CB 链式模型 SHAP 值图

Fig.8　SHAP summary interpretation plots of CB chain model

2.4.4　PLSR 模型与 ML 模型的性能对比　如图 3 和表 3 所示,PLSR 模型对 Φ_{3DOM^*} 预测表现较差,R^2 为 0.2443.对 $\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 的表现适中,R^2 分别为 0.4725 和 0.3159.与 ML 模型相比,PLSR 模型预测性能表现欠佳,因为其主要依靠特征与预测目标之间的相关性.从图 1 可知,Φ_{3DOM^*} 与多数特征变量之间的相关性较低,导致 Φ_{3DOM^*} 预测表现较差,这表明 PLSR 模型不能准确解释特征与 Φ_{3DOM^*} 之间的非线性关系.$\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 与多数特征的线性关系较高,导致其预测表现适中.

通过 PLSR 模型与 ML 模型的预测性能对比可知,数据集中的特征与 Φ_{3DOM^*}、$\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 之间存在着线性和非线性关系,PLSR 模型只能对线性关系作出解释,存在一定的局限性.ML 模型可以对线性和非线性关系进行解释,表现出更强的学习能力.另外,ML 模型结合特征分析(SHAP 值、特征重要性等)能对特性与 Φ_{3DOM^*}、$\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 之间的非线性关系作出更加详细的解释.

3　结论

3.1　对 Φ_{PPRIs} 预测时,CB 算法 ML 模型表现出较好预测性能和较强泛化能力.其中,Φ_{3DOM^*} 的 R^2 比 $\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 高,R^2 分别为 0.9344、0.8091 和 0.7775.温度、FIX 和 pH 值是预测 Φ_{3DOM^*} 的重要特征,DOC、E_2/E_3 和 FIX 是预测 Φ_{3DOM^*} 的重要特征,pH 值、DOC 和温度是预测 $\Phi_{\cdot OH}$ 的重要特征.

3.2　CB 算法链式 ML 模型对 $\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 预测性能有所优化,$^1O_2$ 在测试集的 R^2 值由 0.8091 增加至 0.8233,RMSE 由 1.322 降至 0.418; •OH 在测试集的 R^2 值由 0.7775 上升至 0.8094,RMSE 由 2.982 降至 0.312,表明链式 ML 模型可以学习 Φ_{3DOM^*} 与 $\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 之间生成关系.另外,ML 模型对 Φ_{3DOM^*}、$\Phi_{^1O_2}$ 和 $\Phi_{\cdot OH}$ 的预测表现均优于 PLSR 模型.

参考文献:

[1] Leenheer J A, Croué J P. Peer reviewed: characterizing aquatic dissolved organic matter [J]. Environmental Science & Technology, 2003,37(1):18–26.

[2] Wilske C, Herzsprung P, Lechtenfeld O J, et al. New insights into the seasonal variation of DOM quality of a humic-rich drinking-water reservoir-coupling 2D-Fluorescence and FTICR MS measurements [J]. Water, 2021,13(12):1703.

[3] Stephanie M B, Kristine H W, Christina K R. Dissolved organic matter photoreactivity is determined by its optical properties, redox activity, and molecular composition [J]. Environmental Science & Technology, 2023,57:6703–6711.

[4] Andersson A, Lavonen E, Harir M, et al. Selective removal of natural organic matter during drinking water production changes the composition of disinfection by-products [J]. Environmental Science: Water Research & Technology, 2020,6(3):779–794.

[5] Creed I F, Bergström A K, Trick C G, et al. Global change-driven effects on dissolved organic matter composition: implications for food webs of northern lakes [J]. Global Change Biology, 2018,24(8):3692–3714.

[6] Lipczynska-Kochany E. Effect of climate change on humic substances and associated impacts on the quality of surface water and groundwater: a review [J]. Science of the Total Environment, 2018, 640:1548–1565.

[7] Du P H, Tang K X, Yang B W. Reassessing the quantum yield and

reactivity of triplet-state dissolved organic matter via global kinetic modeling [J]. Environmental Science & Technology, 2024,58:5856–5865.

[8] 郭忠禹.海水溶解性有机质对磺胺氯哒嗪光降解的影响 [D]. 大连: 大连理工大学, 2020.
Guo Z Y. Photodegradation of sulfachloropyridazine in presence of seawater dissolved organic matter [D]. Dalian: Dalian University of Technology, 2020.

[9] Zhou H, Yan S, Lian L, et al. Triplet-state photochemistry of dissolved organic matter: triplet-state energy distribution and surface electric charge conditions [J]. Environmental Science & Technology, 2019, 53(5):2482–2490.

[10] Berg S M, Whiting Q T, Herrli J A, et al. The role of dissolved organic matter composition in determining photochemical reactivity at the molecular level [J]. Environmental Science & Technology, 2019,53 (20):11725–11734.

[11] Mckay G, Huang W X, Romera-Castillo C, et al. Predicting reactive intermediate quantum yields from dissolved organic matter photolysis using optical properties and antioxidant capacity [J]. Environmental Science & Technology, 2017,51(10):5404–5413.

[12] Ma Y F, He J P, Lu H F, et al. Machine learning-based prediction of the CO_2 concentration in the flue gas and carbon emissions from a waste incineration plant [J]. American Chemical Society Environmental Science & Technology Engineering, 2024,4(3):737–747.

[13] Zhu X, He M, Sun Y, et al. Insights into the adsorption of pharmaceuticals and personal care products (PPCPs) on biochar and activated carbon with the aid of machine learning [J]. Journal of Hazardous Materials, 2022,423:127060.

[14] Zhu J J, Chen Y C, Shie R H, et al. Predicting carbonaceous aerosols and identifying their source contribution with advanced approaches [J]. Chemosphere, 2021,266:128966.

[15] Salomé D M, Guimarães P, Pedro O. A review of computational modeling in wastewater treatment processes [J]. American Chemical Society Environmental Science & Technology Water, 2024,4(3):784–804.

[16] Zhong S F, Zhang K, Bagheri M. Machine learning: new ideas and tools in environmental science and engineering [J]. Environmental Science & Technology, 2021,55(19):12741–12754.

[17] Wang C, Pan Y, Chen J, et al. Indicator element selection and geochemical anomaly mapping using recursive feature elimination and random forest methods in the Jingdezhen region of Jiangxi Province, South China [J]. Applied Geochemistry, 2020,122:104760

[18] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees [J]. Nature Machine Intelligence, 2020,2(1):56–67.

[19] 赵健辰.溶解性有机质光致生成活性物种机理及量子产率的定量构效关系模型研究 [D]. 长春:东北师范大学, 2020.
Zhao J C. Photochemical generation mechanisms and quantum yield quantitative structure-activity relationship models of reactive species from dissolved organic matter [D]. Changchun: Northeast Normal University, 2020.

[20] 唐建伟,李 孟,李肇东.光电催化氧化处理中各类活性物质的氧化机理 [J]. 中国环境科学, 2019,39(5):2048–2054.
Tang J W, Li M, Li Z D. Oxidation mechanisms of all kinds of active substances by hybrid photoelectrocatalytic treatment [J]. China Environmental Science, 2019,39(5):2048–2054.

[21] Stephanie M B, Kristine H W, and Christina K R. Dissolved organic matter photoreactivity is determined by its optical properties, redox activity, and molecular composition [J]. Environmental Science & Technology, 2023,57:6703-6711.

[22] He H, Sun N N, Li L F. Photochemical transformation of dissolved organic matter in surface water augmented the formation of disinfection byproducts [J]. Environmental Science & Technology, 2024,58:3399–3411.

[23] Zhang Y N, Zhao J, Zhou Y, et al. Combined effects of dissolved organic matter, pH, ionic strength and halides on photodegradation of oxytetracycline in simulated estuarine waters [J]. Environmental Science-Processes&Impacts, 2019,21:155–162.

[24] Pace M L, Reche I, Cole J J, et al. pH change induces shifts in the size and light absorption of dissolved organic matter [J]. Biogeochemistry, 2012,108:109–118.

[25] Song G, Li Y, Hu S, et al. Photobleaching of chromophoric dissolved organic matter (CDOM) in the Yangtze River estuary kinetics and effects of temperature, pH, and salinity [J]. Environmental Science-Processes&Impacts, 2017,19:861–873.

[26] Du Z Y, He Y S, Fan J N, et al. Predicting apparent singlet oxygen quantum yields of dissolved black carbon and humic substances using spectroscopic indices [J]. Chemosphere, 2018,194:405–413

[27] Mccabe A J, Arnold W A. Reactivity of triplet excited states of dissolved natural organic matter in stormflow from mixed-use watersheds [J]. Environmental Science & Technology, 2017,51(17): 9718–9728.

[28] Bodhipaksha L C, Sharpless C M, Chin Y P, et al. Triplet photochemistry of effluent and natural organic matter in whole water and isolates from effluent-receiving rivers [J]. Environmental Science & Technology, 2015,49(6):3453–3463.

[29] Garrett M, Fernando L R. Temperature dependence of the photochemical formation of hydroxyl radical from dissolved organic matter [J]. Environmental Science & Technology, 2015,49(7):4147–4154.

[30] O'connor M, Helal S R, Latch D E, et al. Quantifying photo-production of triplet excited states and singlet oxygen from effluent organic matter [J]. Water Research, 2019,156:23–33.

作者简介：刘 冰(1978-),男,河南许昌人,博士,副教授,主要研究方向为水体中含氮消毒副产物及前体物识别和控制城市污水处理原理与技术.发表论文 30 余篇.liubing_982002@163.com.