2026 年 4 月 色 谱 Vol.44 No.4
April 2026 Chinese Journal of Chromatography 444~452

研究论文 DOI: 10.3724/SP.J.1123.2025.02012

基于机器学习的全氟及多氟化合物电离效率预测模型开发及其在半定量分析中的应用

孙沈正[1,2]，李瑶瑶[2]，高 燕[2*]，李康聪[2]，陈 智[1*]，李秀琴[2*]，张庆合[2*]

（1. 中国计量大学材料与化学学院，浙江 杭州 310018；2. 中国计量科学研究院化学计量与分析科学研究所，北京 100029）

摘要：全氟和多氟烷基化合物（PFASs）具有持久性、生物累积性和潜在毒性，在环境科学和食品安全等领域备受关注。尽管基于高分辨率质谱（HRMS）的 PFASs 筛查方法发展迅速，但 PFASs 种类繁多，无标准品的 PFASs 定量困难。本研究对 50 种 PFASs 进行液相色谱-高分辨质谱分析，计算其电离效率（IE），并基于机器学习建立 PaDEL 分子描述符与 PFASs 电离效率的定量构效关系（QSAR）预测模型，该模型的构建能够预测无标准品 PFASs 的电离效率，通过预测的电离效率进行 PFASs 的半定量预测。研究采用递归特征消除（RFE），从 1 444 个 PaDEL 描述符中选取了 18 个特征变量；比较了弹性网络线性回归、随机森林和极致梯度提升（XGBoost）3 种不同算法所建立的 QSAR 模型性能。结果表明，XGBoost 模型性能最优，其训练集的均方根误差（RMSE）为 0.052 1，决定系数（R^2）为 0.992 0；测试集的 RMSE 为 0.118 4，R^2 为 0.871 3。50 种 PFASs 预测的 IE 误差在 1.67 倍以内，中位值为 1.04 倍，RMSE 为 1.06。采用预测的电离效率值对不同梯度浓度的 PFASs 标准溶液进行半定量预测来验证模型性能，浓度预测误差倍数为 0.12~4.90 倍，中位值为 0.96 倍，RMSE 为 0.94；且随着溶液浓度升高，浓度半定量预测的准确度提高。将该电离效率预测模型应用于鱼肉中 9 种 PFASs 的半定量预测，预测误差倍数为 0.79~1.81 倍。该模型能够较为准确地进行 PFASs 的电离效率预测，在无标准品 PFASs 的可疑和非靶向筛查、风险评估中具有良好的应用前景。

关键词：全氟及多氟烷基化合物；定量构效关系；机器学习；半定量预测模型

中图分类号：O658 **文献标识码**：A

Development of a machine learning-based ionization efficiency prediction model for per- and polyfluoroalkyl substances and its application in semi-quantitative analysis

SUN Shenzheng[1,2], LI Yaoyao[1,2], GAO Yan[2*], LI Kangcong[2], CHEN Zhi[1*], LI Xiuqin[2*], ZHANG Qinghe[2*]

(*1. China Jiliang University, College of Materials and Chemistry, Hangzhou 310018, China; 2. Division of Chemical Metrology and Analytical Science, National Institute of Metrology, China, Beijing 100029, China*)

Abstract: Per- and polyfluoroalkyl substances (PFASs) represent a category of emerging contaminants of global concern in fields such as environmental science and food safety, due to their persistence, bioaccumulative properties and potential toxicity. Although screening methods for PFASs using high resolution mass spectrometry (HRMS) have been developed rapidly, the diversity of PFASs and the absence of standards pose significant challenges for quantitative analysis. In this study, 50 PFASs were analyzed by HPLC-HRMS. The ionization efficiency (IE) was calculated as

收稿日期：2025-02-24

* **通信联系人**. E-mail: gaoyan@nim.ac.cn（高燕）；E-mail: zchen@cjlu.edu.cn（陈智）；E-mail: zhangqh@nim.ac.cn（张庆合）；E-mail: lixq@nim.ac.cn（李秀琴）

基金项目：国家重点研发计划项目（2023YFF0612601）；国家自然科学基金（22006145）.

Foundation item: National Key Research and Development Program of China (No. 2023YFF0612601); National Natural Science Foundation of China (No. 22006145).

the slope of the calibration curve. A quantitative structure-activity relationship (QSAR) model was developed employing machine learning to predict the ionization efficiencies of PFASs using PaDEL molecular descriptors. The model enables semi-quantitative estimation of PFASs concentrations in the absence of reference standards by incorporating predicted IE values. Eighteen critical descriptors were selected from a total of 1 444 PaDEL descriptors through the application of recursive feature elimination (RFE). These selected descriptors encompassed topological descriptors, geometrical descriptors, autocorrelation descriptors, electrostatic and polarity descriptors. These individual descriptors including VE1_Dzv, GATS6i, JGI10, GATS1p and MATS4m were of great importance. Three algorithms including elastic net linear regression, random forest (RF), and XGBoost were evaluated for model performance. In the elastic net linear regression model, the root mean square error (RMSE) for the training dataset was 0.049 0, and the coefficient of determination (R^2) was 0.993 0; for the test dataset, the RMSE was 0.163 0, with an R^2 of 0.756 1. In the RF model, the RMSE for the training dataset was 0.163 1, and the R^2 was 0.921 9; for the test dataset, the RMSE was 0.131 6, with an R^2 of 0.840 9. In the XGBoost model, the RMSE for the training dataset was 0.052 1, and the R^2 was 0.992 0; for the test dataset, the RMSE was 0.118 4, with an R^2 of 0.871 3. Nonlinear algorithms of random forest and XGBoost demonstrated superior predictive performance compared to the elastic net linear regression, with XGBoost exhibiting best performance. Random forest, a bagging-based approach, trains individual decision trees independently and aggregates predictions through averaging. In contrast, XGBoost employs gradient boosting methodology, iteratively optimizing the model by sequentially training new trees in order to address residuals from previous iterations. The independent training mechanism of random forest inherently lacks the iterative optimization framework that is characteristic of gradient boosting. Specifically, XGBoost systematically enhances predictive accuracy by generating new trees that target residual errors from preceding models, thereby progressively refining predictive performance. This fundamental difference in optimization strategy enables XGBoost to more effectively correct prediction errors compared to the ability of random forest. Based on the results of a comprehensive evaluation of the three models, the XGBoost algorithm was ultimately selected for its demonstrated performance advantages. The prediction errors of ionization efficiency (IE) for the 50 PFASs were within 1.67-fold, with a median value of 1.04-fold and RMSE of 1.06. The established XGBoost model was further applied for the semi-quantitative concentration prediction of 50 PFASs across concentration gradients, where the prediction errors ranged from 0.12 to 4.90-fold, with a median value of 0.96-fold and RMSE of 0.94. The accuracy of the prediction improved as the concentrations increased. Furthermore, the model was applied to predict concentrations of PFASs in fish tissue. After sample extraction and cleanup using solid-phase extraction, the samples were analyzed using HPLC-HRMS. The concentrations of PFASs were semi-quantified using the predicted IEs, yielding prediction errors ranging from 0.79-fold to 1.81-fold. These findings highlight the robustness of the IE prediction model for PFASs. Notably, the performance of the developed model was better than or comparable to the performance of previous studies. In conclusion, this study introduces a machine learning-based QSAR model for

引用本文:孙沈正,李瑶瑶,高燕,李康聪,陈智,李秀琴,张庆合. 基于机器学习的全氟及多氟化合物电离效率预测模型开发及其在半定量分析中的应用. 色谱,2026,44(4):444-452.

SUN Shenzheng, LI Yaoyao, GAO Yan, LI Kangcong, CHEN Zhi, LI Xiuqin, ZHANG Qinghe. Development of a machine learning-based ionization efficiency prediction model for per- and polyfluoroalkyl substances and its application in semi-quantitative analysis. Chinese Journal of Chromatography, 2026, 44(4): 444-452.

the prediction of ionization efficiency. This approach illustrates the ability to estimate the concentrations of PFASs in the absence of standards, thereby presenting considerable potential for the risk assessment of compounds lacking standards in suspect and non-targeted screening.

Keywords: per- and polyfluoroalkyl substances (PFASs); quantitative structure-activity relationship (QSAR); machine learning; semi-quantitative predictive model

全氟及多氟烷基化合物(per- and polyfluoroalkyl substances, PFASs)是一类具有独特理化性质的有机化合物,因其卓越的耐热性、化学稳定性及疏水疏油特性,被广泛应用于工业生产和日常消费品中,如防水防油涂层、电子产品制造、灭火泡沫及食品包装材料等[1,2]。然而,这类化合物也具有环境持久性和生物累积性,成为全球关注的持久性有机污染物[3,4]。近年来,PFASs 在全球环境介质和生物体中广泛检出[5-7],以鱼肉为例,PFASs 在鱼肉中被广泛检出[8],且研究发现人体血清中 PFASs 浓度水平和鱼肉食用量具有关联性[9]。PFASs 被证实具有潜在的生态毒性和健康危害,包括致癌性、内分泌干扰及免疫系统影响等[10-12]。这些问题促使国际社会对其使用和排放实施严格的管控政策,同时推动对替代品和治理技术的研发。

全氟辛烷磺酸盐(perfluorooctanesulfonate, PFOS)、全氟辛酸(perfluorooctanoic acid, PFOA)、全氟己烷磺酸盐(perfluorohexanesulfonate, PFHxS)等传统 PFASs 的生产和使用受到限制,但仍有大部分结构类似物作为替代品使用[1]。目前对于 PFASs 的分析既包括对特定目标物的靶向分析,也包括可疑和非靶向分析。基于液相色谱-三重四极杆质谱(LC-MS/MS)的靶向分析是目前检测 PFASs 最主要的分析方法[13-15],LC-MS/MS 分析依赖于标准品,由于 PFASs 标准品种类不足,LC-MS/MS 对于无标准品 PFASs 的分析有一定的局限性。基于液相色谱-高分辨质谱(LC-HRMS)的靶向、可疑和非靶向筛查近年来发展迅速,成为 PFASs 全面分析的重要工具[16],开发无标准品的 PFASs 半定量预测方法也是研究的重点。定量构效关系(QSAR)模型是一种基于化学分子结构和性质预测目标化合物行为的数学模型,QSAR 模型近年来逐渐应用于高分辨质谱的非靶向筛查半定量预测中,通过化合物的分子质量、电负性、拓扑结构等分子描述符与电离效率等变量之间的数学关系,建立预测模型[17-19]。

本研究采用液相色谱-高分辨质谱分离了 50 种 PFASs,获得其电离效率,采用弹性网络、随机森林和 XGBoost(Extreme Gradient Boosting)等机器学习算法建立了基于分子描述符和电离效率的 QSAR 半定量预测模型,比较了不同模型的性能并将所建立 QSAR 模型应用于鱼肉中 PFASs 的半定量预测。

1 实验部分

1.1 仪器与试剂

Vanquish-Q Exactive Plus orbitrap 液相色谱-高分辨质谱联用仪(LC-HRMS),美国 ThermoFisher 公司。Milli-Q 超纯水仪,德国 Merck 公司。ME614S 电子天平,德国 Sartorius 公司;CR21GIII 离心机,日本 Hitachi 公司;KS260 型振荡摇床,德国 IKA 公司;N-EVAP 111 氮吹仪,美国 Organomation Associates 公司。

50 种 PFASs 单标准溶液(50 μg/mL)(表 1)及 9 种 PFASs 同位素内标,加拿大 Wellington Laboratories 公司。GBW(E)100740 鱼肉粉中 9 种全氟化合物成分分析标准物质,中国计量科学研究院;甲醇、乙腈,色谱级,美国 ThermoFisher 公司;醋酸铵,色谱级,德国 Sigma 公司;OASIS WAX(6 mL, 150 mg)固相萃取柱,美国 Waters 公司。

1.2 实验过程

1.2.1 样品前处理

参考实验室前期开发的方法[20],适当优化后,进行鱼肉粉样品的前处理。准确称取 0.5 g 鱼肉粉样品(精确至 0.000 1 g)至离心管,加入 1.5 g 水复溶至肉浆,立即手摇混匀,涡旋 2 min,静置 30 min。准确加入同位素内标(精确至 0.000 1 g),涡旋 2 min,静置 30 min 平衡。加入 10 mL 乙腈,振荡提取 30 min,超声提取 10 min,3 000 r/min 离心,转移上清液至离心管。剩余残渣重复提取 1 次,8 000 r/min 离心,合并提取液并氮吹浓缩至 3 mL,加入 20 mL 水稀释。WAX 固相萃取柱经 4 mL 0.1% 氨水甲醇、4 mL 甲醇、4 mL 水活化。样品稀释液过 WAX 柱,4 mL 25 mmol/L 的醋酸铵溶液(pH=

4)淋洗、4 mL 水淋洗,抽真空 30 min 去除水分,固相萃取柱采用 4 mL 甲醇和 4 mL 0.1% 氨水甲醇洗脱,收集洗脱液,40 °C 水浴下氮气吹至约 1 mL,定容至 1 mL。溶液在−30 °C 冷冻 2 h,冷冻离心机 10 000 r/min 离心 10 min,取上清液,过 0.2 μm 微孔滤膜,转移至进样瓶待测。

1.2.2 分析条件

采用 Acquity UPLC BEH C18 色谱柱(100 mm×2.1 mm,1.7 μm,美国 Waters 公司);流动相由 5 mmol/L 醋酸铵水溶液(A)和乙腈(B)组成;流速 0.2 mL/min;色谱柱温度 40 °C;样品温度 10 °C;进样体积:5 μL。梯度洗脱程序:0~3.5 min,90%A~40%A;3.5~9 min,40%A~5%A;9~11 min,5%A;11~11.5 min,5%A~90%A;11.5~15 min,90%A。

采用电喷雾离子源,在负离子模式下进行分析。喷雾电压−2 kV;离子传输管温度 350 °C;鞘气流速 50 arb;辅助气流速 12 arb;辅助气加热温度 350 °C;吹扫气流速 2 arb;S lens RF 水平 80。数据采集采用全扫描-数据依赖采集模式(Full MS-ddMS^2),一级质谱分辨率为 70 000;扫描范围为 *m/z* 80~1 000;自动增益控制进入轨道阱中的离子数(AGC target)为 1.6×10^5;最大驻留时间为 50 ms;半峰宽设为 10 s。ddMS^2 扫描参数为分辨率 35 000;AGC target 为 1×10^5;最大驻留时间为 50 ms;TopN 设为 5;复合碰撞能设置为 10、30、70 eV;质量隔离窗口 *m/z* 为 0.4;动态排除时间为 10 s。

1.2.3 溶液配制

移取 50 种 PFASs 单标准溶液(50 μg/mL),用甲醇配制成 100 ng/mL 的 50 种 PFASs 混合标准溶液。用甲醇对上述混合标准溶液逐级稀释,配制质量浓度为 0.2、0.5、1、2、5、10 ng/mL 的 PFASs 系列混合标准溶液,进行液相色谱-高分辨质谱分析。采用一级全扫描的峰面积数据绘制标准曲线,计算标准曲线的斜率作为电离效率(表 1)。

表 1　50种PFASs的信息

Table 1　Information of the 50 per- and polyfluoroalkyl substances (PFASs)

Compound	Abbreviation	Molecular formula	M_W	S_{PFAS}	EXP.logIE
Perfluorobutanoic acid	PFBA	$C_4HF_7O_2$	214.04	1154942	0.5126
Perfluoropentanoic acid	PFPeA	$C_5HF_9O_2$	264.05	1574622	0.5560
Perfluorohexanoic acid	PFHxA	$C_6HF_{11}O_2$	314.05	1570712	0.4796
Perfluoroheptanoic acid	PFHpA	$C_7HF_{13}O_2$	364.06	1910129	0.5004
Perfluorooctanoic acid	PFOA	$C_8HF_{15}O_2$	414.07	2282327	0.5218
Perfluorononanoic acid	PFNA	$C_9HF_{17}O_2$	464.08	1527177	0.2978
Perfluorodecanoic acid	PFDA	$C_{10}HF_{19}O_2$	514.08	1643246	0.2852
Perfluoroundecanoic acid	PFUnDA	$C_{11}HF_{21}O_2$	564.09	1404931	0.1768
Perfluorododecanoic acid	PFDoDA	$C_{12}HF_{23}O_2$	614.10	1159583	0.0566
Perfluorotridecanoic acid	PFTrDA	$C_{13}HF_{25}O_2$	664.10	1012491	−0.0363
Perfluorotetradecanoic acid	PFTeDA	$C_{14}HF_{27}O_2$	714.11	793649	−0.1736
Perfluorohexadecanoic acid	PFHxDA	$C_{16}HF_{31}O_2$	814.13	1096572	−0.0902
Perfluorooctadecanoic acid	PFOcDA	$C_{18}HF_{35}O_2$	914.10	1149002	−0.1202
Perfluoropropanesulfonate	PFPrS	$C_3F_7NaO_3S$	272.08	11867290	1.4202
Perfluorobutanesulfonate	PFBS	$C_4F_9KO_3S$	338.19	4856766	0.9377
Perfluoropentanesulfonate	PFPeS	$C_5F_{11}SO_3Na$	372.09	5094298	0.9169
Perfluorohexanesulfonate	PFHxS	$C_6F_{13}hO_3S$	422.10	5916332	0.9271
Perfluoroheptanesulfonate	PFHpS	$C_7H_{15}FhO_3S$	450.12	3224561	0.6356
Perfluorooctanesulfonate	PFOS	$C_8F_{17}NaO_3S$	522.11	3628964	0.6225
Perfluorononanesulfonate	PFNS	$C_9F_{19}NaO_3S$	572.12	3092414	0.5133
Perfluorodecanesulfonate	PFDS	$C_{10}F_{21}NaO_3S$	622.13	3011099	0.4654
Perfluorododecanesulfonate	PFDoS	$C_{12}F_{25}NaO_3S$	722.14	2936978	0.3898
1*H*,1*H*,2*H*,2*H*-Perfluorohexanesulfonate	4:2FTS	$C_6H_5F_9O_3S$	328.15	1710828	0.4976
1*H*,1*H*,2*H*,2*H*-Perfluorooctanesulfonate	6:2FTS	$C_8H_5F_{13}O_3S$	428.17	1216809	0.2341
1*H*,1*H*,2*H*,2*H*-Perfluorodecanesulfonate	8:2FTS	$C_{10}H_5F_{17}O_3S$	528.18	955746	0.0381

表 1 （续）
Table 1 （Continued）

Compound	Abbreviation	Molecular formula	M_W	S_{PFAS}	EXP.logIE
1*H*,1*H*,2*H*,2*H*-Perfluorododecanesulfonate	10∶2FTS	$C_{12}H_5F_{21}O_3S$	628.20	771748	−0.1301
Perfluorobutanesulfonamide	FBSA	$C_4H_2F_9NO_2S$	299.12	7780847	1.1957
Perfluorohexanesulfonamide	FHxSA	$C_6H_2F_{13}NO_2S$	399.13	7125577	1.0322
Perfluorooctanesulfonamide	FOSA	$C_8H_2F_{17}NO_2S$	499.15	4692375	0.7537
Perfluorodecanesulfonamide	FDSA	$C_{10}H_2F_{21}NO_2S$	599.16	2006669	0.3054
N-Methylperfluorobutanesulfonamide	*N*-MeFBSA	$C_5H_4F_9NO_2S$	313.14	2373360	0.6601
N-Methylperfluorooctanesulfonamide	*N*-MeFOSA	$C_9H_4F_{17}NO_2S$	513.17	4125792	0.6858
N-Ethylperfluorooctanesulfonamide	*N*-EtFOSA	$C_{10}H_6F_{17}NO_2S$	527.20	4389897	0.7010
Perfluorooctanesulfonamidoacetic acid	FOSAA	$C_{10}H_4F_{17}NO_4S$	557.18	4360824	0.6741
N-Methylperfluorooctanesulfonamidoacetic acid	*N*-MeFOSAA	$C_{11}H_6F_{17}NO_4S$	571.21	1183535	0.0969
N-Ethylperfluorooctanesulfonamidoacetic acid	*N*-EtFOSAA	$C_{12}H_8F_{17}NO_4S$	585.24	1378411	0.1526
2-Perfluorohexyl ethanoic acid	FHEA	$C_8H_3F_{13}O_2$	378.09	47959	−1.1162
2-Perfluorooctyl ethanoic acid	FOEA	$C_{10}H_3F_{17}O_2$	478.10	97802	−0.9087
2-Perfluorodecyl ethanoic acid	FDEA	$C_{12}H_3F_{21}O_2$	578.12	85756	−1.0483
2*H*-Perfluorooctenoic acid	FHUEA	$C_8H_2F_{12}O_2$	358.08	593572	0
2*H*-Perfluorodecenoic acid	FOUEA	$C_{10}H_2F_{16}O_2$	458.10	929333	0.0877
2*H*-Perfluorododecenoic acid	FDUEA	$C_{12}H_2F_{20}O_2$	558.11	760903	−0.0849
Perfluoro-4-oxapentanoic acid	PF4OPeA	$C_4HF_7O_3$	230.04	2770477	0.8613
Perfluoro-5-oxahexanoic acid	PF5OHxA	$C_5HF_9O_3$	280.04	3818766	0.9152
Perfluoro-2-methyl-3-oxahexanoic acid	HFPO-DA	$C_6HF_{11}O_3$	330.05	150303	−0.5611
Perfluoro-3,6-dioxaheptanoic acid	PF3,6,OHpA	$C_5HF_9O_4$	296.04	622356	0.1032
Dodecafluoro-3*H*-4,8-dioxanonanoate	NaDONA	$C_7HF_{12}O_4Na$	400.05	7485739	1.0526
6∶2 Chlorinated polyfluorinated ether sulfonate	6∶2ClPFAES	$C_8ClF_{16}KO_4S$	570.67	5863862	0.7923
Perfluoro-4-ethylcyclohexanesulfonate	PFECHS	$C_8F_{15}SO_3K$	500.22	8249681	0.9978
Perfluoro-2-ethoxyethanesulfonate	PFEESA	$C_4F_9SO_4K$	354.19	11596492	1.2956

M_W: the molecular weight; S_{PFAS}: the slope of the calibration curve; EXP.logIE: the relative ionization efficiency determined experimentally.

1.2.4 QSAR 模型构建方法

通过比较各化合物与校正化合物 2*H*-全氟-2-辛烯酸（FHUEA）的响应值计算得到其实验相对电离效率（EXP.logIE）。EXP.logIE 计算如公式（1）所示，

$$\text{EXP.logIE} = \log\frac{S_{PFAS}}{S_{FHUEA}} \times \frac{M_{W\text{-}FHUEA}}{M_{W\text{-}PFAS}} \quad (1)$$

其中，S_{FHUEA}、$M_{W\text{-}FHUEA}$ 分别为校正化合物 FHUEA 的斜率和相对分子质量，S_{PFAS}、$M_{W\text{-}PFAS}$ 为各化合物标准曲线的斜率和相对分子质量。本研究 PFASs 标准曲线斜率基于质量浓度（ng/mL），而在 QSAR 模型建立中，通常应该以物质的量浓度单位以消除分子质量的影响[21]，因此，公式（1）中将质量浓度转化为物质的量浓度。这些通过实验算得的 EXP.logIE 作为建模时的响应变量，通过 PaDEL 软件计算得到的各化合物的分子描述符作为输入变量构建 QSAR 模型。当后续出现没有标准品的全氟化合物即可通过该 QSAR 模型得到此化合物的相对电离效率（Pred.logIE）。通过预测相对电离效率，可以计算预测的化合物的浓度（C_{pred}），C_{pred} 计算如公式（2）所示，其中 A_{PFAS} 为对应的峰面积。

$$C_{pred} = \frac{A_{PFAS} \times M_{W\text{-}FHUEA}}{[10^{\text{Pred.logIE}}] \times S_{FHUEA} \times M_{W\text{-}PFAS}} \quad (2)$$

使用 PaDEL 计算各化合物对应的描述符作为 QSAR 模型的特征变量，上述计算得到的相对电离效率作为 QSAR 模型的响应变量。对于特征变量先进行数据预处理和特征清洗，排除当中的异常值和对模型没有贡献的值。再将这 50 种化合物划分为训练集和测试集，训练集采用不同的算法构建模型，测试集用来评价模型的性能。最终选取模型性

能最好的模型用于实际样品的预测。以上 QSAR 模型的建立通过 R 和 Rstudio 软件实现。

2　结果与讨论

2.1　数据预处理

化合物在 ESI-MS 中的绝对电离效率受流动相组成、质谱参数、电离源的结构等多种因素的影响[22-24],并且这些参数的影响程度相互关联,其重复性很难控制,因此各化合物的绝对电离效率难以用于浓度预测,因此本研究采用相对电离效率进行浓度预测。本研究选择 FHUEA 作为校正化合物,一方面,FHUEA 质谱响应灵敏度高,在实验的较低浓度水平仍然有很好的质谱信号;另一方面,FHUEA 响应具有良好的重复性,其作为校正化合物具有较好的耐受性和稳定性,是良好的电离效率基线指标[25]。

PaDEL 由 Chun Wei Yap 发明,用来快速计算化合物从 1D 到 3D 的多种分子特征和分子指纹[26]。本研究通过 PaDEL 计算各化合物的 2D 分子描述符,得到 1 444 个描述符。因为本研究的 PFASs 样本只有 50 个,而计算出的分子描述符数量远大于样本量,引起了“维数灾难”,导致这些数据在空间中分布变得极为稀疏,使得很多算法不再有效,因此,建模之前需要进行特征清洗,首先移除数据集中的缺失值、接近零方差的特征变量和相关系数大于 0.8 的变量,然后使用递归特征消除(RFE)选取了最佳的特征变量。经过处理后的特征变量有 121 个,这 121 个特征变量经过 RFE 筛选后有 18 个特征变量保留下来。

2.2　电离效率预测模型建立

将 50 种化合物随机划分为训练集和验证集,训练集中有 42 种化合物,用于构建模型,验证集中有 8 种化合物,用于对模型进行外部验证。本研究比较了弹性网络线性回归算法、随机森林及 XGBoost 非线性回归算法的预测效果。使用训练集决定系数(R^2)和均方根误差(RMSE)来表征模型的拟合优度,使用测试集 R^2 和 RMSE 表征模型的预测能力。以上功能在 R 和 Rstudio 上实现。

弹性网络是一种多元线性回归算法,结合了岭回归和套索回归的优点,适用于处理高度相关的特征或特征数量大于样本数量的情形。预测结果如图 1a 所示,在训练集上的 RMSE 为 0.049 0,R^2 为 0.993 0,测试上的 RMSE 为 0.163 0,R^2 为 0.756 1;该方法出现了较严重的过拟合现象,可能是电离效率和特征变量之间并不存在明显的线性关系。

进一步采用随机森林和 XGBoost 两种多元非线性回归算法进行模型建立。随机森林算法能通过集成多个决策树,每棵树的分裂可以基于不同的特征组合,从而能够有效捕捉变量之间的非线性关系。不像线性回归或其他传统回归方法,随机森林不需要提前假设变量之间的关系,因此特别适用于无法事先定义的复杂数据结构。使用随机森林在训练集上的 RMSE 为 0.163 1,R^2 为 0.921 9,测试上的 RMSE 为 0.131 6,R^2 为 0.840 9,如图 1b 所示。XGBoost 是一种基于梯度提升树的高效实现,它能通过集成多个决策树的预测来提高精度,同时通过正则化和剪枝避免过拟合。使用 XGBoost 在训练集上的 RMSE 为 0.052 1,R^2 为 0.992 0,测试上的 RMSE 为 0.118 4,R^2 为 0.871 3,如图 1c 所示。

综上,两种非线性方法预测效果均优于弹性网络,且 XGBoost 的预测效果优于随机森林。随机森林是基于 Bagging 的方法,每棵树是独立训练的,

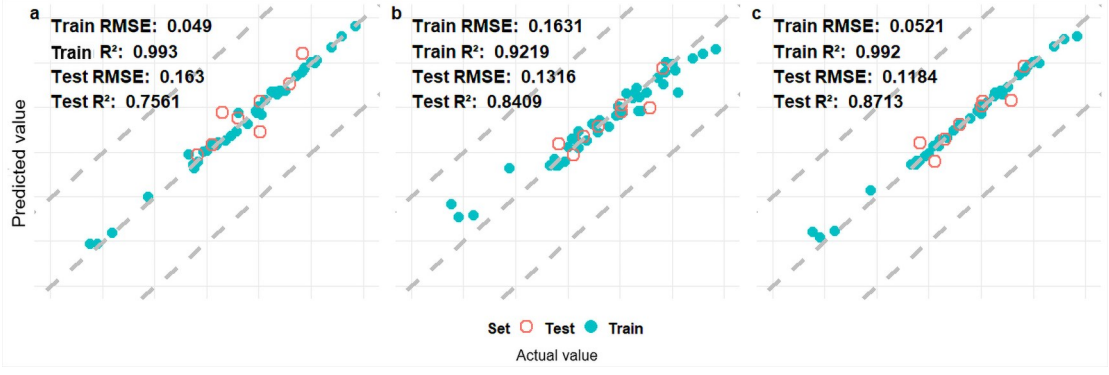


图 1　(a)弹性网络、(b)随机森林和(c)XGBoost 模型的预测结果

Fig. 1　Prediction results of (a) elastic net, (b) random forest, and (c) XGBoost models

最终通过平均方式得到预测结果；XGBoost 使用梯度提升(Boosting)方法，在每一轮训练中逐步优化模型。由于随机森林模型每棵树都是独立训练的，随机森林的模型学习过程并没有梯度提升那样的迭代优化机制，而 XGBoost 每次新加入的树是根据前一轮模型的残差来进行训练的，这样能够更精准地纠正之前的预测错误。通过以上 3 种模型的比较，选择 XGBoost 算法。

2.3 模型机理分析

XGBoost 模型最终选取的描述符有 18 个，主要分为拓扑描述符(topological descriptors)、几何描述符(geometrical descriptors)、自相似性描述符(autocorrelation descriptors)和电性与极性描述符(electrostatic and polarity descriptors)4 个类别。拓扑描述符反映分子的拓扑结构特征，描述分子中原子、键和分支的关系。几何描述符为量化分子的几何形状、大小或空间分布。自相似性描述符是基于分子中原子间的物理化学性质的相关性计算。电性与极性描述符描述分子的电性、极性或反应性特征[27]。本研究中选择的 18 个描述符的重要性如图 2 所示，其中重要性较高的描述符为 VE1_Dzv、GATS6i、JGI10、GATS1p 和 MATS4m。其中 VE1_Dzv 是基于分子拓扑结构计算的描述符，衡量分子图的电子复杂性和结构特征，它反映了分子中原子和化学键的连接方式和分布。GATS6i 是一种基于自相关函数的几何描述符，表示 6 步长内的分子内部原子间相互作用，且与原子的离子化势加权相关。JGI10 是基于分子结构中原子之间路径长度的拓扑描述符，用于量化分子结构的锯齿性或连接的不规则性。GATS1p 是基于分子内 1 步长的几何自相关性描述符，加权因子为原子的偶极矩。MATS4m 是基于 4 步长的分子内原子间距离的几何自相关描述符，权重为原子的质量[27]。这 5 个描述符都与分子的结构和拓扑相关，并且可以反映分子内原子和键的连接性和空间特征，表明分子整体的结构复杂性和键合模式是影响电离效率的主要因素。

2.4 预测误差分析

使用公式(3)计算 IE 的预测误差，其中 IE_{pred}、IE_{actual} 分别为通过模型预测的 IE 和通过公式(1)计算出的实际 IE。IE_{error} 为这两个值的比值，用来表示预测的误差。误差累计图如图 3 所示，结果表明，模型预测的 IE 误差都在 1.67 倍以内，误差中位值为 1.04 倍，平均值为 1.10 倍，均方根误差为 1.06。误差最大的化合物是 2-全氟己烷乙酸(FHEA)，发现该化合物 IE 值仅为 0.076 5，是本研究中 IE 最低的 PFASs，其次是 6∶2 氯代多氟烷基醚磺酸盐(6∶2ClPFAES)，误差为 1.65 倍，该化合物中含有氯原子，碳氯键和碳氟键极性和键能有一定差异，本研究 50 种化合物仅该化合物含氯原子，模型对该化合物的 IE 的预测能力相对较差。

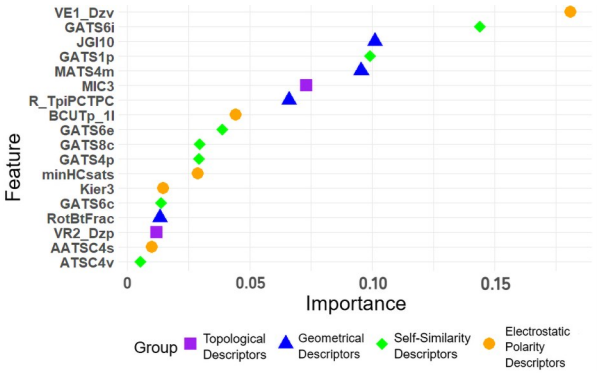


图 2 描述符的重要性

Fig. 2 Importance of descriptors

$$IE_{error}=\begin{cases}\dfrac{IE_{pred}}{IE_{actual}} & IE_{pred}\geqslant IE_{actual}\\ \dfrac{IE_{actual}}{IE_{pred}} & IE_{pred}<IE_{actual}\end{cases} \quad (3)$$

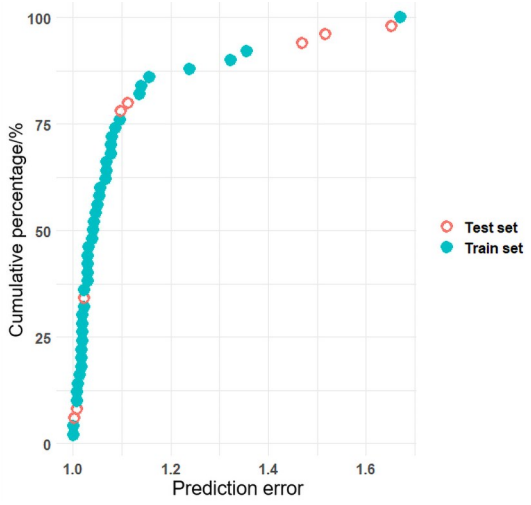


图 3 IE 预测误差累计图

Fig. 3 Cumulative plot of ionization efficiency (IE) prediction error

将预测的电离效率代入公式(2)，计算 50 种不同梯度浓度的 PFASs 预测浓度(实际质量浓度为 0.2、0.5、1、2、5、10 ng/mL)。用公式(4)计算预测误差(error)，其中 C_{actual} 为实际浓度。结果表明，总体

浓度预测误差倍数在 0.12~4.09 范围内,误差中位值为 0.96,平均值为 1.03,均方根误差为 0.94。且浓度越高离散程度越低,预测的结果越稳定,如图 4 所示。图 4 中的离群点为该梯度浓度下预测结果较差的化合物。

$$\text{Error} = \frac{C_{\text{pred}}}{C_{\text{actual}}} \tag{4}$$

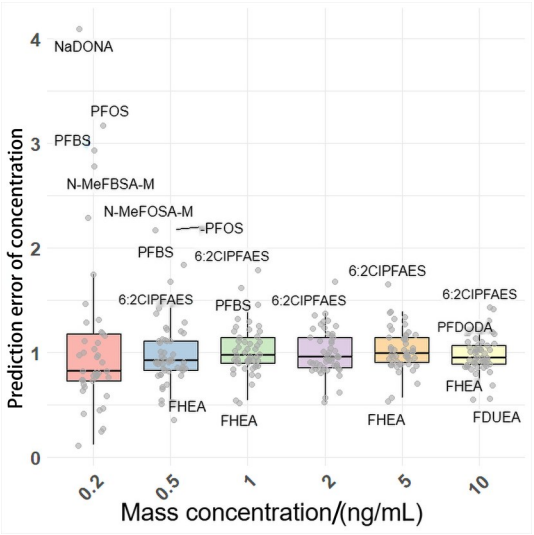


图 4　不同梯度浓度的预测误差

Fig. 4　Prediction errors of concentrations across gradients

对于不同 PFASs,各浓度点预测误差中位值为 0.53~1.66 内,平均值在 0.51~1.73,如图 5 所示。预测误差相对较大的化合物进行分析发现,部分化合物预测浓度整体偏高或偏低,例如,6∶2ClPFAES 的预测误差为 1.43~1.79,中位值为 1.67,平均值为 1.62,其预测浓度整体偏高,其误差主要来源于电离效率预测值的误差,FHEA、2*H*-全氟-2-十二烯酸(FDUHA)等也具有相似的误差来源。而对于 4,8-二氧杂-3*H*-全氟壬酸盐(NaDONA)、全氟丁基磺酸盐(PFBS)、全氟辛基磺酸盐(PFOS),其 IE 预测误差为 1.14、1.04 和 1.01,但其 0.2 ng/mL 和 0.5 ng/mL 低浓度点的预测误差倍数较大,分析其标准曲线结果发现该两种化合物截距值相对较大,对低浓度预测结果有较大的影响。但本研究整体上对 50 种 PFASs 化合物有较好的预测结果。

2.5　模型比较

近年来,QSAR 模型逐渐应用于高分辨质谱可疑或非靶向分析中无标准品化合物的半定量预测。本研究采用了 XGBoost 算法,已有研究中采用的算法主要包括随机森林、梯度提升决策树(xgbTree)、遗传算法-多元线性回归(GA-MLR)、人工神经网络(ANN)等[28]。对于 PFASs 的半定量预测,Hu 等[29]采用随机森林算法,建立了 50 种 PFASs 半定量预测模型,对于电离效率预测误差,Hu 等[29]模型整体上预测误差最高为 2.7 倍,中位值为 1.1,略高于本研究的预测误差。Lauria 等[18]在不同类别化合物半定量预测模型的基础上,新加入了 33 种 PFASs,将原模型拓展为能对 PFASs 进行半定量预测的模型,测试集的最大预测误差为 2.3 倍,略高于本研究。Liigand 等[17]构建了药物、代谢物、氨基酸、酚类化合物、杂环化合物等不同类别化合物的电离效率通用预测模型,ESI^+模式下化合物电离效率的平均预测误差为 2.2 倍,ESI^-模式下化合物电离效率的平均预测误差为 2.0 倍,与该模型相比,本研究也具有良好的预测结果。

2.6　模型应用

将本研究建立的模型应用于 1 种鱼肉标准物

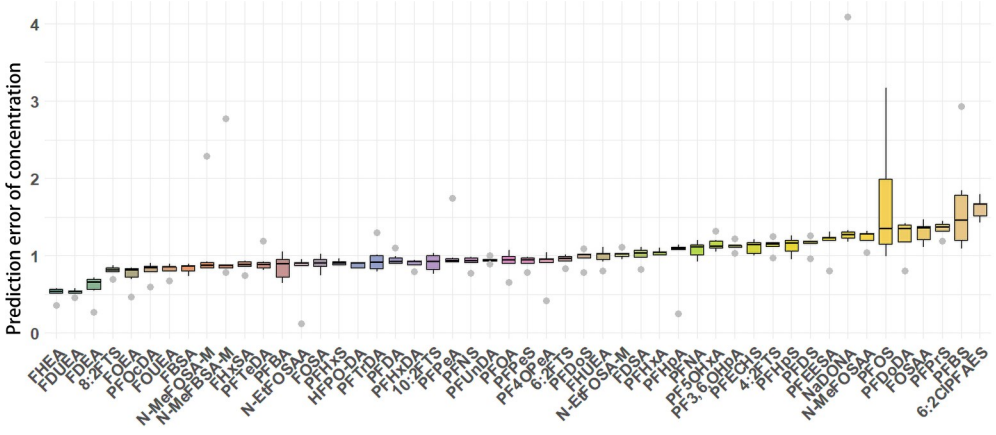


图 5　不同化合物的预测浓度误差

Fig. 5　Prediction errors of concentrations for different compounds

质中 9 种 PFASs 的半定量预测。经前处理后的样品溶液进行 HPLC-HRMS 分析后，分别采用同位素稀释质谱法（IDMS）[20]和本研究建立的 QSAR 半定量预测模型进行 PFASs 定量，同位素稀释质谱法的 PFASs 定量结果均在标准物质的不确定度范围内（表 2）。QSAR 模型预测结果与 IDMS 结果的误差在 0.79~1.81 倍范围内，证明建立的预测模型具有良好的实际应用价值。

表 2 鱼肉粉中 PFASs 的预测结果

Table 2 Prediction results of PFASs in fish powder

Compound	Certified value	Uncertainty	IDMS result	QSAR predicted value	Prediction error of concentration
PFBS	3.45	0.45	3.76	4.38	1.16
PFHxS	5.04	0.63	5.17	4.93	0.95
PFOS	6.41	0.93	6.39	6.95	1.09
PFOA	4.47	0.61	4.64	3.65	0.79
PFNA	4.78	0.36	4.59	5.15	1.12
PFDA	4.64	0.49	4.75	4.19	0.88
PFUnDA	4.83	0.69	5.00	4.76	0.95
PFDoDA	4.27	0.55	4.01	7.24	1.81
PFTeDA	3.69	0.51	3.22	3.93	1.22

3 结论

本研究开发了一种基于 PaDEL 描述符和机器学习算法的多类型 PFASs 半定量预测模型，比较了弹性网络、随机森林、XGBoost 算法的预测性能，选取了性能最佳的 XGBoost 用于电离效率预测。本研究在可疑和非靶向筛查中无标准品化合物的风险评估中具有良好的应用前景，能够为相关研究和应用提供参考。本研究还存在一定的局限性，建模的方法具有良好的适用性，但模型结果只适用于该类型仪器和色谱条件，不同仪器条件应用该方法进行半定量预测时，需进行模型结果的重新输出。未来的研究在建模过程中考虑流动相组成、流动相的比例、pH、黏度等经验描述符能够进一步增强 QSAR 模型的移植能力和适用性。

参考文献：

[1] Wang Z Y, DeWitt J, Higgins C P, et al. Environ Sci Technol, 2017, 51(5): 2508

[2] Wang J L, Lin Z H, He X X, et al. Environ Sci Technol, 2022, 56(9): 5355

[3] Rudin E, Gluege J, Scheringer M, et al. Sci Total Environ, 2023, 877: 162618

[4] Liu Z Y, Liu S, Xiao F, et al. J Hazard Mater, 2024, 465: 133184

[5] Xing Y N, Li Q, Chen X, et al. Agric Food Chem, 2023, 71(3): 1259

[6] Li W Z, Li H Z, Zhang D N, et al. Environ Sci Technol, 2022, 56(6): 3492

[7] Gao Y, Song B Y, He A E, et al. Sci Total Environ, 2023, 865: 161125

[8] Zhang M, Cai D, Chen X X, et al. Environmental Chemistry, 2024, 43(7): 2224
张敏, 蔡丹, 陈小霞, 等. 环境化学, 2024, 43(7): 2224

[9] Bjermo H, Darnerud P O, Pearson M, et al. Mol Nutr Food Res, 2013, 57(12): 2206

[10] Zhang R J, Lin Y S, Tu L Y, et al. Chinese Journal of Chromatography, 2025, 43(1): 96
张睿佳, 林颖诗, 涂蓝尹, 等. 色谱, 2025, 43(1): 96

[11] Park S Y, Kong S H, Kim K J, et al. Endocrinol Metab, 2024, 39(4): 539

[12] Zhang L P, Louie A, Rigutto G, et al. Environ Res, 2023, 220: 115188

[13] Zhou X Y, Lu Y, Luo Y D, et al. Journal of Instrumental Analysis, 2024, 43(8): 1166
周鑫月, 卢瑶, 罗雅丹, 等. 分析测试学报, 2024, 43(8): 1166

[14] Chen L, Dai Y Y, Zhou C, et al. Agric Food Chem, 2020, 68(30): 8026

[15] Frigerio G, Cafagna S, Polledri E, et al. Bioanal Chem, 2022, 414(3): 1259

[16] Zhu L Y, Bossi R, Carvalho P N, et al. Environ Pollut, 2024, 360: 124605

[17] Liigand J, Wang T T, Kellogg J, et al. Sci Rep, 2020, 10(1): 5808

[18] Aalizadeh R, Panara A, Thomaidis N S, et al. J Am Soc Mass Spectrom, 2021, 32(6): 14

[19] Lauria M Z, Sepman H, Ledbetter T, et al. Environ Sci Technol, 2024, 58(5): 2458

[20] Gao Y, Li X M, Li X Q, et al. J Chromatogr B, 2018, 1084: 45

[21] Cherkasov A, Muratov E N, Fourches D, et al. J Med Chem, 2014, 57(12): 4977

[22] Liigand P, Liigand J, Kaupmees K, et al. Anal Chim Acta, 2021, 1152: 238117

[23] Kruve. ARapid Commun Mass Spectrom, 2019, 33(S3): 54

[24] Kruve A, Kaupmees K. Anal Chem, 2017, 89(9): 5079

[25] Aalizadeh R, Nikolopoulou V, Alygizakis N, et al. Anal Bioanal Chem, 2022, 414(25): 7435

[26] Yap C W. J Comput Chem, 2011, 32(7): 1466

[27] Roberto T, Viviana C. Handbook of Molecular Descriptors. Weinheim: WILEY-VCH Verlag GmbH, 2000

[28] Abrahamsson D P, Park J S, Singh R R, et al. J Chem Inf Model, 2020, 60(6): 2718

[29] Hu J R, Lyu Y, Chen H, et al. Water Res, 2023, 233: 119735