中国科学: 技术科学 2024 年 第 54 卷 第 10 期: 1838 ~ 1853
SCIENTIA SINICA Technologica techcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述 环境中新污染物专题

机器学习在新污染物环境风险识别与防控上的研究进展与挑战

胡献刚[1*], 王张佳[1], 邓鹏[1], 于福波[1], 穆莉[2], 王赛[3], 周启星[1]

1. 南开大学环境科学与工程学院, 环境污染过程与基准教育部重点实验室, 天津 300350;
2. 农业农村部环境保护科研监测所, 农业农村部农产品质量安全环境因子控制重点实验室, 天津 300191;
3. 海南大学南海海洋资源利用国家重点实验室, 海口 570228
* E-mail: huxiangang@nankai.edu.cn

收稿日期: 2024-01-27; 接受日期: 2024-04-02; 网络版发表日期: 2024-08-28
国家重点研发计划项目(编号: 2020YFC1807000)和国家自然科学基金(批准号: U22A20615)资助项目

摘要 随着技术的不断发展, 越来越多的新污染物(emerging contaminants, ECs)产生并进入环境中. 这些ECs往往具有生物毒性和环境持久性, 对生态环境和人类健康构成风险, 因此越来越多的研究开始关注ECs的风险识别与防控. 自然环境中ECs的存在情况复杂, 传统的分析方法在ECs风险的识别与预测上费时费力. 机器学习(machine learning, ML)是一种数据驱动的研究方法, 使用已有的数据训练模型, 以更好地预测和发现研究对象的发展变化规律. 由于ML能够深入了解参数之间复杂的关系, 近年来, ML在ECs风险识别与防控领域的应用愈加广泛. 传统技术与ML结合能够降低计算成本, 并通过优化和减少实验次数来节省实验时间和能源消耗. ML在ECs中的应用大致可以分为毒性预测、识别分类、性质评估与辅助去除. 本文梳理了ML在典型ECs(纳米材料、微纳米塑料、抗生素抗性基因、全氟烷基和多氟烷基物质、内分泌干扰物和持久性有机污染物)中的应用情况与挑战. 在毒性研究中, ML的使用能够减少动物实验, 但是该领域目前普遍存在数据较少的情况, 部分模型使用的数据集较小, 数据质量较低, 模型应用范围有限. 并且并不是所有的环境问题都适用ML, 没有必要过度依赖ML. 复杂的ML模型通常伴随着可解释性的不足, 使得研究者难以理解ML如何得出特定的预测结果, 这对于解释和理解ECs的环境行为和风险是一个挑战. 在ML中, 除了模型算法本身, 训练数据的质量决定着模型的准确性和预测的准确性. 因此今后应重点提升ECs数据质量, 实现数据共享, 并建立统一完善的数据库, 形成适合ECs研究的研究体系或框架.

关键词 机器学习, 新污染物, 纳米材料, 抗生素, 微塑料, 内分泌干扰物

1 引言

全球范围内注册生产和使用的化学品已经超过了350000种[1]. 随着技术不断的发展, 越来越多的污染物排放到环境中. 某些污染物之前研究较少或新近发现, 且具有生物毒性、环境持久性和生物累积性等特征, 由于缺少相应的法律法规和标准予以监管, 具有较大的环境风险, 此类污染物称为新污染物(emerging con-

引用格式: 胡献刚, 王张佳, 邓鹏, 等. 机器学习在新污染物环境风险识别与防控上的研究进展与挑战. 中国科学: 技术科学, 2024, 54: 1838–1853
Hu X G, Wang Z J, Deng P, et al. Application of machine learning in environmental risk identification and control of emerging contaminants: Research progress and challenges (in Chinese). Sci Sin Tech, 2024, 54: 1838–1853, doi: 10.1360/SST-2024-0037

© 2024 《中国科学》杂志社 www.scichina.com

taminants, ECs)[2]. 微塑料(microplastics, MPs)、全氟烷基和多氟烷基物质(per- and polyfluoroalkyl substances, PFAS)、内分泌干扰物(endocrine disrupting chemicals, EDCs)、抗生素抗性基因(antibiotics and resistance genes, ARGs)、纳米材料(nano materials, NMs)、药品和个人护理产品(pharmaceutical and personal care products, PPCPs)等均属于ECs. ECs可能从各种环节中产生并进入环境中, 如农用化学品、工业活动、个人护理品和雨水径流等[3]. 地表水、地下水、城市废水及农业用地等许多环境基质中均已检测出ECs[4], 甚至在南极港湾的浮游植物中也发现了阻燃剂和多氯联苯等多达70种不同的污染物[5]. 作为海洋食物网的基础, 浮游植物中检测到如此多的ECs意味着整个生态系统已经受到了威胁. 研究表明, ECs会对生物的生殖、神经、发育和免疫系统造成危害. 在人体中甚至也已经检测出了部分ECs的存在, 如多氯联苯、六氯苯、全氟辛烷磺酸、全氟辛酸、聚对苯二甲酸乙二醇酯和聚乙烯等[6]. 尽管ECs在环境中普遍存在, 并对人类和环境健康产生了影响, 但人们对它们在生物和非生物基质中的发生情况缺乏足够的认知. 因此需要更多的数据和方法以评估ECs带来的相关风险. 环境中的ECs浓度广、数量大、类别多, 分析样本数量巨大, 此时使用传统分析方法显得费时费力, 因此开发快速准确有效的分析方法是十分必要的[7].

机器学习(machine learning, ML)是人工智能的一个子集, 是计算机利用已有的历史数据, 通过算法得出某种模型, 并利用此模型描述或预测未来的一系列方法[8]. ML的主要特点是使用计算机程序通过数据进行“训练”, 以更好地发现研究对象的发展规律. 同时, 也需要采取某些验证措施, 确保训练后的模型能够对未知数据进行预测分析, 从而不会过度拟合训练数据. 深度神经网络、支持向量机、k-近邻、决策树、随机森林、朴素贝叶斯和梯度提升等算法属于常见的ML算法. 表1总结了这几种常见的ML算法的部分优缺点[9~18]. 每个ML算法都由3个关键部分组成: 算法的结构(如随机森林和深度神经网络等)、待实现的目标(如预测精度和平方误差)以及实现目标的训练方法(如随机梯度下降)[19]. 作为ECs与ML研究的一般流程, 首先需建立ECs相关数据集, 再对数据进行筛选以及预处理, 以满足模型训练的要求; 然后选择或创建ML相关算法, 进行模型优化, 获得预测准确性高、误差小、泛化能力强的预测模型, 最终对未知信息进行预测. ML作为对先验知识依赖性较低的数据分析方法, 因其具有强大的拟合能力, 在解决复杂的数据模式或格式方面具有巨大的应用前景. 与传统数学模型相比, ML不需要考虑过程的科学背景和机制来预测相关数据. 因此ML过程被广泛用于模拟不同应用中各种独立和依赖性较弱参数之间的相关性[20].

近些年来, ML在ECs风险识别与防控领域的应用也愈加广泛. 在Web of Science上以各ML算法和ECs名

表 1　几种常见机器学习模型的优缺点

Table 1　Advantages and disadvantages of common machine learning models

算法	优点	缺点	参考文献
深度神经网络	可以分析大型数据集以及各种其他数据集, 并提取其特征; 可以使用不同的输入数据集进行训练, 并且可以从隐藏层中提取其特征	对数据集开发具有较高要求; 可能产生不满足预期性质的输出结果	[9,10]
支持向量机	高泛化能力、良好的鲁棒性以及在相对少量的样本数据下也能获得出色结果; 高维数据处理效率高	经典支持向量机求解速度慢、精度不足	[11]
k-近邻	可以在大维度上执行穷举搜索; 算法易于实现、简单有效; 易于解释和理解结果; 具有预测离散变量和分类变量的能力; 新到达的样本不需要训练过程, 支持增量数据	只搜索测试实例的最近邻, 易受到噪声样本的影响, 影响邻域概念的结果, 降低了其分类能力, 从而降低了预测准确性	[12]
决策树	对所有类型的数据集都具有简单和直观的可视化能力	容易出现过拟合, 无法给出全局最优解	[13]
随机森林	容易适应分类和回归, 可以简单并行化, 可以处理混合变量类型(连续和分类), 不受单调变换影响, 对异常值不敏感; 对噪声变量具有鲁棒性; 只需很少或无需调整即可产生良好的结果, 并且可以应用于各种领域	随机森林在决策树的构建中使用随机性, 从而避免了对训练样本的过度拟合. 因此, 必须构建大量的决策树才能获得高度的通用性, 增加决策树的数量会增加内存需求	[14,15]
朴素贝叶斯	具有显著的分类精度; 对缺失数据不太敏感, 算法较简单	条件独立性假设在现实中很少成立; 忽略了类变量之间的依赖关系	[16,17]
梯度提升	允许通过梯度下降法优化任意可微损失函数	算法是顺序训练学习器, 很难并行化	[18]

称为关键词检索, 时间段选择2000年1月1日至2024年1月21日, 对所得结果使用Web of Science自带工具进行分析, 发现近年来相关论文数量显著增长, 如图1(a)所示. 对作者所属地进行分析后发现, 美国和中国发表文献最多, 是如今该领域研究的中心, 如图1(b)所示. 图2展示了发表文献量最大的前15个研究领域, 结果显示研究最多的2个领域分别是药理学与传染病学, 以上结果可能与抗生素抗性基因的普遍关注有关. 生态学与环境科学排列第3位, 以上结果说明人体与生态环境健康是目前研究的主要关注方向. 现有ML和ECs的综述偏向于关注单一ECs, 主要研究重心集中在MPs和NMs. 此种局面致使观点单一, 导致缺少ML在整个

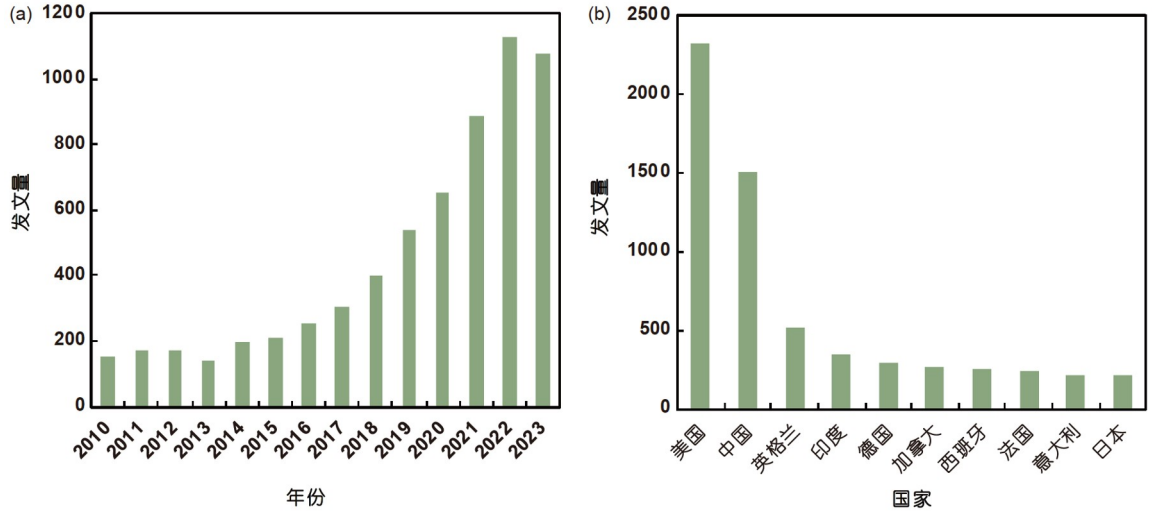


图 1 (网络版彩图)机器学习-新污染物相关研究的文献发表情况. (a) 2010~2023年发表的机器学习-新污染物相关研究的文献数量; (b) 机器学习-新污染物相关研究的文献归属国家分布

Figure 1 (Color online) Publication status of literature on machine learning-emerging contaminants related research. (a) Publication numbers of machine learning-emerging contaminants related research from 2010 to 2023; (b) publication numbers of machine learning-emerging contaminants related according to countries.

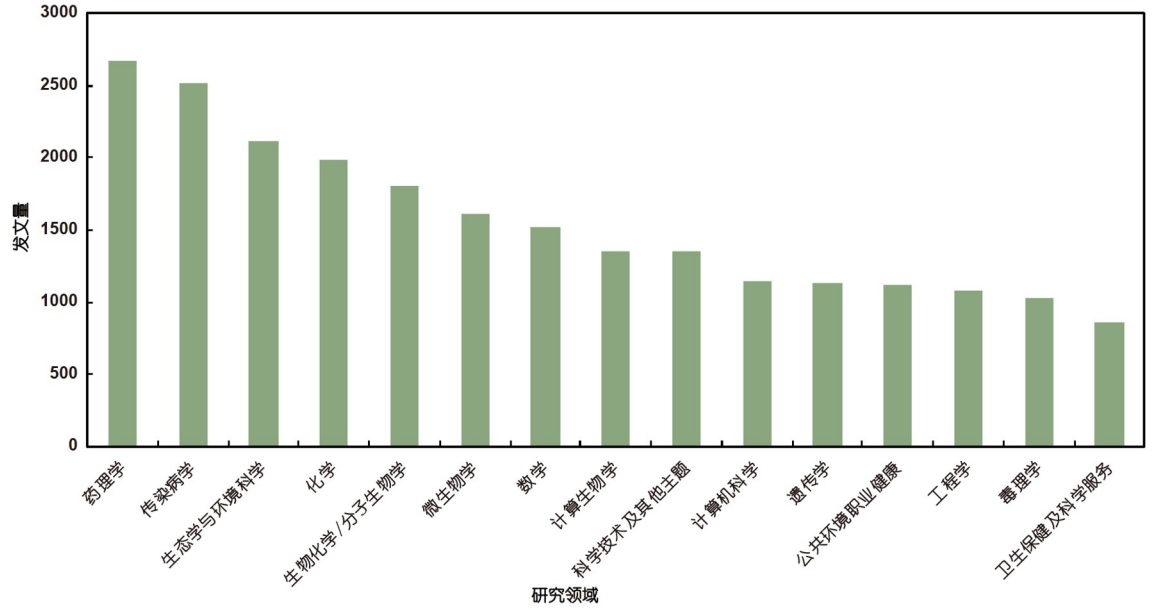


图 2 (网络版彩图)机器学习-新污染物相关文献在不同研究领域发表情况

Figure 2 (Color online) Numbers of the paper published about machine learning-emerging contaminants related researches in different fields.

ECs领域应用的综合认识[21~23]. 由于ML对数据量具有一定的要求, 受研究热度或发文量的影响, 当前ML-ECs的研究还是更多地介绍NMs, NPs和ARGs, 对于其他ECs的介绍较少[24]. 本文旨在系统性综述ML在各类ECs风险防控领域的应用现状, 帮助理解主要研究方向, 并对未来发展趋势做出探讨.

2 机器学习在新污染物风险识别与防控中应用

ML与传统技术的结合是目前重要研究趋势之一. 在一项预测海洋硅藻在不同剂量下对不同ECs的暴露程度的研究中, ML算法与传统的荧光高通量数据相结合, 形成了生态毒性快速评价的有力工具, 预测污染物类型的准确率达97.65%[25].

需要注意的是, 真实环境中的ECs常常不会单独存在, 污染物间会发生复杂的相互作用, 影响风险评估的准确性[26]. 单个污染物的风险结果无法反映混合物对生物体的综合影响[27]. 现有大多数分析方法的关注点局限在小部分的确定数量的化学品上, 忽略了复合污染的整体效应, 因此对复杂混合物进行非靶向分析具有较好的科学性[28]. 有研究对100个随机组合的ECs进行混合风险测试, 使用贝叶斯神经网络开发了可以预测ECs混合物对藻类风险的ML模型, 该研究帮助研究人员发现了ECs数量与藻类生物学功能之间的关系, 为后续研究人员开发复合污染风险模型提供了理论基础[29]. 在未来, 应有组织地管理各ECs的化学信息, 包括典型毒性官能团、异质性、生物学相关信息、基本化学结构及其描述符等信息, 建立综合的数据库, 为ML在ECs风险评估上提供高质量数据支撑.

ML也被应用在ECs去除、寻找最佳参数、提升去除效率等方面. 高级氧化、膜分离、生物修复以及吸附分离等均是现有的常用技术[4]. 传统生物炭吸附研究存在以下问题[30]: 生物炭与ECs的吸附受多种因素影响; 传统吸附模型通常是静态模型, 只能解释一组固定操作条件下的吸附机制; 批量实验成本高且耗时, 无法描述参数间相互作用的影响. 因此建立通用的可解释数据驱动的数学预测模型来研究ECs吸附是至关重要的. Haider Jaffari等人[31]构建了10个基于树的ML模型, 使用24个输入变量及3757个数据点以研究每种吸附条件和生物炭特性对ECs吸附的特征重要性和影响. Categorical boosting模型在此方面表现最好, 具有最高的试验决定系数(0.9433)以及最低的平均绝对误差(4.95 mg/g).

ML的出现为ECs传统风险防控技术提供了新的发展方向和机遇. 将ML与传统技术相结合以提高实验效率并降低成本, 也可使用ML模型对ECs数据进行分析以发现深层次关系, 在复杂(多维、多变量以及非线性)数据中识别出内在趋势, 预测风险的演变趋势. 下文将具体介绍ML在典型ECs中的研究进展与挑战.

2.1 纳米材料

随着纳米技术的快速发展, NMs的使用量逐渐增大, 在催化剂、化妆品、化肥、农药及医疗等行业中均有应用[32]. 大量使用导致其不可避免地暴露在环境中, NMs的环境和健康风险倍受关注. 暴露在NMs中的生物体有可能发生生长抑制、生殖毒性、光合作用抑制甚至死亡等现象. NMs虽然在水生环境中容易团聚, 但在环境相关浓度下还是会引发斑马鱼的发育畸形、氧化应激和代谢紊乱[33]. 浓度分别为0.25, 2.5和25 mg/kg植物鲜重的石墨烯基纳米材料注射进了辣椒植株的茎中, 石墨烯基纳米材料的存在使植物过氧化氢酶活性最高降低了28.9%, 致使H_2O_2含量增加, 诱导氧化损伤, 辣椒果实中丙二醛的含量也有所增加, 出现了膜脂质过氧化现象[34]. 经过70 mg/L纳米CuO处理10 d后, *Myriophyllum spicatum*的生长速度降低了30%[35]. 也有研究发现纳米ZnO可以干扰孵化酶的结构和功能, 抑制其活性, 从而降低胚胎在孵化过程中溶解绒毛膜的能力, 最终导致孵化延迟[36]. 鉴于以上生物毒性研究结果, 开展NMs风险防控, 避免或减少其对环境和人体的健康危害具有重要的环境意义.

ML技术可用于处理具有异质性、不确定性以及具有良好容错性的数据, 并且能够在没有输入和输出参数之间预定义关系的情况下学习领域中的底层模式, 是破译和阐明NMs动态聚集过程的现有非线性关系的有力工具[37]. Derjaguin-landau-vervey-overbeek (DLVO)理论是研究NMs在水环境中行为的经典模型, 然而该方法存在部分局限, 如仅适用于描述短距离范围内(0.1~10 nm)的相互作用[38].

物质的结构影响其活性, 因此相似的结构往往表现出相似的活性. 定量构效关系(quantitative structure-activity relationship, QSAR)模型即利用此原理预测污

染物的毒性. 基于统计学和ML方法的QSAR模型, 以多壁碳纳米管物理化学性质为输入特征, 能够预测其遗传毒性[39]. 通过物理化学性质预测金属氧化物纳米材料炎症潜力的QSAR模型, 预测准确度超过90%[40].

尽管ML在构建模拟复杂关系的模型方面具有很好的潜力, 其较差的可解释性一直是目前研究所面临的困难. 为了提高模型的可解释性, Yu等人[41]提出了一个基于树的随机森林特征重要性和特征交互网络分析框架. 该模型取得了良好的结果, 准确预测了NMs的肺免疫反应和肺负荷, 所有训练集的相关系数均大于0.9. 该研究为理想NMs的设计和应用提供了指导. 数据库的构建能够对相关领域研究产生很大影响. 在NMs风险防控领域, 为了进一步指导NMs的设计, 需要建立一个公开可用的NMs数据库. Yan等人[42]构建了一个包含705种独特的NMs数据库, 纳米结构注释程序为NMs生成2142个纳米描述符, 助力ML在NMs领域的进一步发展.

以往NMs的毒性研究成本高且可重复性低, 但前期实验积累的数据为ML提供了数据支持. 作为数据驱动的方法, ML具有强大的揭示复杂关系和预测的能力, 但NMs的高反应性为其环境行为预测带来了很大的不确定性[43]. 今后的研究应着重于基于ML方法解释NMs的普遍环境行为或规律, 这是实验和传统的模型(QSAR)难以完成的. 传统的动植物实验成本高, 周期长, 并且所得数据难以满足风险评估的需求. 计算机建模可以作为体内外实验测试的更快、更廉价的替代方案, 用于污染物的危害表征, 也能够减少动物实验. 同时, NMs在环境中会发生聚集、沉降、溶解、吸附等复杂的环境行为. 使用数据驱动的ML模型, 即使不了解NMs的环境行为与生物作用机制, 也可直接捕获物理化学性质与生物毒性之间的关系[23].

目前将ML应用在NMs去除或纳米生物效应优化上的研究相对较少, 而这恰恰是在风险防控中较为重要的一环. 将来应当有更多的研究关注NMs的表面改性等提高其生物相容性的方法, 或者提升NMs去除技术的效率. 在毒性预测方面, 尽管NMs的细胞毒性问题已得到广泛研究, 但其血液循环寿命和生物降解性仍然是风险防控途中真正的挑战. 因此, 建议在研究中考虑NMs在体内长期反应, 例如对寿命和繁殖的影响. 随着NMs领域迅速发展, 新型纳米结构不断出现, 为ML在NMs风险识别与防控的应用带来了挑战. 对于新型NMs, 有限的风险数据是限制ML精准识别其风险的首要因素. 现有ML模型定义域通常局限于常见NMs, 且由于描述符的限制难以实现对新型NMs的迁移学习. 随着新型NMs研究的不断深入, 现有模型可能需要不断更新以适应新的数据.

2.2 微纳米塑料

塑料具有多功能性, 被广泛应用于包装、建筑或电子工业等领域, 其产量也在近几十年呈指数级增长[44]. 与1950年相比, 全球塑料产量增加了200倍; 目前全球塑料年产量超过3.35亿吨, 如此巨大的生产和使用量使其不可避免地被排放到环境中[45]. 这些塑料在使用和降解的过程中会形成MPs, 即粒径小于5 mm的塑料颗粒. MPs体积小, 容易被生物摄入并在食物链中积累, 较大的比表面积也使其容易吸附其他有毒物质, 产生难以预料的毒性结果. 此外, 塑料具有化学惰性, 其在自然环境中不易降解, 可以长期存在, 从而增加其对生物的潜在风险. 已有许多研究报道了MPs对各种生物的影响, 例如海洋变暖情况下聚苯乙烯MPs对海洋浮游植物中聚球菌属的影响, 发现MPs使聚球菌的细胞生长速率和固碳能力被削[46]. 聚苯乙烯等MPs在水相中也会释放邻苯二甲酸二乙基己酯和双酚A等化学添加剂, 抑制蓝细菌等自养生物[47]. 在聚氯乙烯MPs对人淋巴细胞毒性研究中, 观察到MPs会诱导人淋巴细胞的氧化应激和细胞器损伤[48]. MPs进入人体后也会诱发炎症、神经退行性疾病、免疫紊乱以及癌症的发生[49]. 因此开展高效的MPs风险防控具有重要的生态与环境健康意义.

MPs的分析方法主要包括目视检查和振动光谱技术两种[50]. 目视检查方法常常用到扫描电子显微镜和透射电子显微镜等显微技术, 其能提供MPs大小、颜色和形状的物理性质, 缺点是无法展示MPs化学成分信息[51]. MPs由高分子量的长链烃组成, 通过产生相应样品的独特分子振动, 可以很容易地通过振动光谱法检测其化学成分[52]. 当MPs所处环境条件复杂时, 使用传统的目视检查和光谱技术进行分析是耗时耗力的. ML与现有MPs分析检测方法的结合能够克服这些缺点, 提高分析准确性并缩短检测时间, 帮助更好地评估环境中MPs污染.

在未来, 实现各个红外光谱数据库的共享, 是强化ML应用的关键. 针对光谱技术公开可用的光谱数据库

有限、频谱特征可变等问题, Lei等人[53]创建了高分辨率全窗口拉曼光谱数据库, 该数据库能够适应不同的实验设置和目标. 高光谱成像技术和卷积神经网络模型的结合是一种新颖、高效、稳健的方法, 用于农田土壤中各种单一MPs聚合物(聚乙烯、聚丙烯及聚氯乙烯)分类, Ai等人[54]采集了600个光谱数据, 最终分类准确率达92.6%. 该方法能够准确、快速地识别农田土壤中不同类型的MPs, 为MPs检测提供了无损检测方法.

MPs在水生环境中会发生矿化和尺寸变化, 释放出各种化学物质(添加剂、单体和氧化中间体等). 这些化学物质对环境具有潜在毒性风险, 如邻苯二甲酸盐能够干扰内分泌, 诱导斑马鱼胚胎产生卵黄囊异常、免疫反应改变和脂质水平降低等现象[55]. 因此有必要了解MPs的降解风化过程, 以对其中化学物质的释放有更精确的预测. 与传统分析方法相比, ML能够考虑MPs中污染物浸出的许多影响因素, 可以分析和整合MPs特性、环境条件、流体动力学因素等各种数据, 从而进行更全面的分析. Kida等人[56]采用了多元回归、人工神经网络、支持向量机和随机森林来预测MPs中增塑剂和其他污染物的浸出情况, 对比发现多元回归分析结果的决定系数R^2最低, 说明ML方法在污染物建模方面确实比传统方法更具优势.

纳米塑料为粒径小于1 μm(或100 nm)的塑料颗粒, 与MPs相比, 纳米塑料更小的体积使其能够穿透生物屏障并在组织间转移, 对健康构成风险. MPs检测常常使用拉曼光谱法, 然而当涉及纳米尺寸的颗粒时, 识别准确性往往较低. 随着尺寸减小, 许多复杂情况会出现, 例如塑料结晶度的变化, 以及一些特征拉曼峰变得太弱或发生移动[57]. 纳米塑料中较弱的拉曼光谱信号很容易被其他因素掩盖[58]. 因此, 需要建立有效的纳米塑料微弱信号的获取方法和纳米塑料拉曼光谱数据库. 随机森林能够处理嘈杂和高维数据, 实现高精度, 并提供一定程度上的可解释性[59]. 拉曼光谱可与随机森林结合以识别纳米塑料, 实验室测试中该方法识别准确率超过97%, 表明随机森林在识别和区分纳米塑料以及其他环境颗粒之间具有应用前景[60].

2.3　抗生素与抗性基因

2000~2010年间, 全球人类抗生素消费量增长了35%, 尤其是中等收入国家, 抗生素使用量显著增加[61]. 当微生物能够在药物存在的环境中适应和生长时, 就产生了抗生素耐药性, 此类微生物具有ARGs. 抗生素过量和不当的使用会引发ARGs的传播. 由于全球大多数水体都是相连的, ARGs可以通过径流从污泥和垃圾填埋厂渗入其他环境, 从而使水生微生物群落或人群暴露在不同来源的ARGs中[62]. 河口、污泥处理湿地以及蔬菜水果中均检测到了ARGs的存在[63,64]. 不同种类和来源的ARGs会在细菌间传播, 使细菌产生更高的耐药性. 而世界卫生组织的调查显示, 这些细菌每年会造成全球近70万人死亡[65]. 抗生素滥用和ARGs传播已经成全球的重要卫生问题. 而抗生素本身也存在毒性, 双氯芬酸是一种广泛使用的抗生素药物, 其被发现能够延迟大型溞的产卵天数, 影响生长、发育和繁殖相关基因的表达[66]. 宏基因组学、实时荧光定量PCR和传统PCR常常被用来研究水生环境中ARGs的浓度. 不过此类方法由于非特异性扩增产物和较高的成本, 难以广泛应用到常规监测中[67].

可动遗传因子(mobile genetic elements, MGEs)介导的水平基因转移是抗性基因在微生物群落之间传播的方式之一, 减轻ARG/MGE向环境的传播已成为全球关注的焦点[68]. 厌氧消化中的沼渣是ARGs进入生态系统的重要途径[69]. 相较于传统的厌氧消化1号模型(anaerobic digestion model No 1, ADM1), ML算法可以深入识别各种过程参数之间的潜在非线性相互作用[70]. Haffiez等人[71]比较了三种不同的ML方法(随机森林、极端梯度提升和人工神经网络)预测厌氧消化期间ARGs和MGEs丰度变化, 其中人工神经网络在此方面更加可靠. Zheng等人[72]使用ML算法(支持向量机、k-近邻、梯度提升决策树和随机森林)研究1088个土壤宏基因组学样品, 定量模拟ARGs在环境中的转移, 生成了全球土壤ARGs丰度图, 揭示了抗生素耐药性的热点, 并确定了全球土壤抗生素耐药性控制的优先区域.

缺乏数据成为目前限制研究者开展相关研究的主要瓶颈. 在全球海洋抗生素风险评估方面, Zhang等人[73]使用海洋栖息地的712个样本建立了预测模型, 使用随机森林算法结合10倍交叉验证进行ML分析, 绘制了全球海洋抗生素耐药性威胁状况图. 该研究发现纬度、紫外线辐射变化、海平面上升、地表温度上升以及海洋酸化等是全球海洋抗性基因分布的重要影响因素. 为了确定ARGs在海洋中的分布、宿主和表达, Xu等人[74]基于随机森林、支持向量回归和套索回归

等算法分析了来自海洋样本的347个宏基因组和182个宏转录组, 发现ARGs的多样性和丰度随纬度和深度而变化, 并预测出到2100年超过97%的海洋ARG将改变表达方式, 说明海洋环境中的ARGs可能非常活跃. 在抗生素去除领域, 传统方法利用生物炭吸附去除环境中的四环素[75]. ML算法可利用吸附实验的数据模拟四环素在生物炭上的吸附情况, 分析出何种条件下的生物炭具有最高的吸附效率, 也能够探索四环素在生物炭上吸附的驱动因素; 在算法选择方面, 随机森林、梯度提升决策、极限梯度提升和人工神经网络四种算法相比, 随机森林具有更高的预测准确性(R^2 = 0.9625)[76].

2.4 全氟烷基和多氟烷基物质

PFAS是一类广泛的人为有机化学品, 其特征都是不同长度的全氟脂肪族碳链和酸性头基, 在各种商业产品和工业应用中以复杂混合物使用[77]. 此类产品的生产和使用导致了PFAS在自然环境中的释放和大量积累, 人口稠密的城市和工业区周围的河流和港口水排放是PFAS进入海洋的主要输入途径. PFAS已在各种地方检出, 包括河流、饮用水甚至人类血液[78,79]. 2022年8月, 美国环保署将PFAS指定为有害物质, 认为其“可能对人类健康、福利或环境构成重大危险”. 由于PFAS具有疏水性并且拥有独特的形态和复杂的化学成分, 其可在环境中长期存在, 造成持久性的影响. PFAS被发现与人类哮喘、高血压、乳腺癌、肝癌等疾病存在关联. 进入体内的PFAS会积聚在生物的肾、肝、脾和脑等器官中[80].

PFAS具有许多独特的化学指纹, 将指纹数据与环境命运和传输模型以及地理空间统计技术相结合, 可用于污染源跟踪[81]. ML与环境检测数据的结合研究越来越受到关注, 其在PFAS的环境取证中也能发挥重要作用. Stults等人[82]使用三种ML算法(k-近邻、决策树和支持向量机)对鱼类中文献报道的PFAS指纹进行分类, 分类准确率最高达91.4%, 并指出鱼类组织中的PFAS指纹可能是地表水系统中PFAS来源追踪的有效手段. PFAS被植物根系吸收后会在植物的可食用部位积累, 这也是人类接触PFAS的重要途径之一[83]. 基于随机森林回归、梯度提升回归树、多层感知器网络和支持向量回归预测作物中PFAS的根系浓度因子, 发现碳链长度是影响PFAS根系吸收的关键分子结构[84].

为了深入认识新兴的PFAS的持久性、生物蓄积性和毒性(persistent, bioaccumulative, and toxic, PBT)等特性, 可编制从非靶向分析中获得的新兴PFAS数据库, 并使用ML模型评估了它们的PBT特性[85]. 定性和定量ML模型的结合可以提供一种方法工具, 用于确定其他新兴有机污染物的优先级. 不过该领域目前仍存在一定限制[86], 如目前可用于预测PBT特性的模型之间存在差异, 假阳性和假阴性预测结果可能导致PBT错误分类, 并且PBT筛查标准在体现PFAS独特理化性质上不具备筛查优势.

在PFAS去除方面, 以光催化降解为例, 其被用于处理全氟辛酸, 形成低毒性的中间体[87]. 为了预测光催化对有机污染物的降解效率, 传统实验通常需要进行数小时甚至数天. 并且实验所用半导体较为昂贵, 并不具有成本效益. 数学模型可以通过优化和减少实验次数来节省实验时间、成本和能源消耗, 是光催化降解效率预测的有力工具[20]. 而与传统数学模型相比, ML不需要考虑过程的科学背景和机制来预测相关性, 并且可以识别数据集中的隐式关系, 处理高维问题和数据. 总的来说, 在对众多过程变量之间的复杂相关性进行建模方面ML优于传统数学模型[88].

2.5 内分泌干扰物

EDCs是指能够改变激素正常功能的化合物, 是一类广泛分布在环境中的合成化学物质或重金属. 这些外源性化学物质会干扰生物体中的天然激素受体, 导致荷尔蒙失衡和异常, 即使在非常低的浓度下也是如此. 雌性化、免疫系统功能异常、神经缺陷、性行为变化和肿瘤的出现被认为与EDCs暴露有关[89]. 因此, 开发评估环境中潜在EDCs的有效方法是十分重要的. 信号通路分析与细胞增殖、配体结合、蛋白质和基因检测等传统的体外方法用于确定EDCs的潜在影响[90]. 此类方法的优点是可以鉴定由受体相互作用介导的化学物质的生物活性并减少对全动物毒性测试的依赖, 同时通常比体内检测更便宜、快速. 但是待评估化学物质范围广泛, 此种方法难以在EDCs与人类健康之间建立剂量反应关系, 并且当处理量大时其成本高昂且耗时长[91]. QSAR是一种具有前景的替代计算技术, 广泛用于预测EDCs对人类和野生动物健康的毒理学影响. 目前通过QSAR建模分析的数据集也越来越庞大、高维和复杂[92]. Heo等人[91]开发了一种基于深度

学习(堆叠自编码、深度置信网或深度神经网)的QSAR模型, 评估EDCs对人类健康的不利影响, 该模型具有较高的分类和预测性能以及令人满意的拟合优度. 深度学习是人工神经网络概念的延伸, 具有许多隐藏层的分层结构, 可以有效地捕获数据集中的非线性属性和复杂模式, 用于改善QSAR的数据管理和建模结果[92].

EDCs主要是通过与内分泌激素平衡密切相关的受体如雄激素受体和雌激素受体结合来影响内分泌系统[93]. 雌激素受体是一种细胞蛋白, 可触发对内分泌系统至关重要的基因产物的表达[94]. EDCs与雌激素受体的结合会导致内质网信号传导改变[95]. Russo等人[96]使用伯努利朴素贝叶斯、自适应提升算法、随机森林、支持向量机和深度神经网络等算法对化合物与雌激素受体的结合进行了准确预测, 表明经典的ML算法足以开发高质量的雌激素受体预测模型. Zhang等人[97]使用线性判别分析、分类与回归树以及支持向量机三种ML方法开发了一个评估平台, 旨在对EDCs进行有效分类, 最终在测试集上的总体预测准确率达75.2%. Chen等人[93]通过使用7个指纹图谱结合4个ML方法(k-近邻、决策树、朴素贝叶斯和支持向量机)预测了雄激素受体和雌激素受体的结合物, 发现PubChem Fingerprint-SVM是预测雄激素受体结合物的最佳模型(准确度0.84), Extended Fingerprint-SVM是预测雌激素受体结合物的最佳模型(准确度0.79).

在EDCs去除方面, 菌藻共生系统是一种利用微藻和细菌之间的协同作用处理废水和污染物的环保技术, 能够以环境友好的方式有效消除污染物, 不会造成二次污染[98]. Fu等人[99]使用人工神经网络预测双酚A降解率, 进而优化菌藻共生系统对双酚A的去除, 提高去除效率. 菌藻共生系统对双酚A的去除过程遵循Monod的动力学模型, 对双酚A的去除率达57.5%. 此研究也为应用ML辅助ECs降解研究提供了新的视角以及研究策略.

2.6　持久性有机污染物

持久性有机污染物(persistent organic pollutants, POPs)会在自然界中长期存在, 并具有生物积累性, 对野生动植物和人类造成严重伤害, 在国内外已经有较长的研究历史. 流行病学研究表明, 接触某些有机氯持久性有机污染物可能会引发子宫内膜异位, 导致出现慢性盆腔疼痛、痛经等非特异性症状[100]. 化学物质数量的众多以及混合物的复杂性限制了污染物同时暴露的研究. ML的出现给此类研究带来了希望, 这些算法能够评估个体变量之间的关联, 同时调整共暴露, 解决共线性问题, 其已在流行病学中被广泛应用[101]. Matta等人[102]使用了五种不同的ML模型(正则化逻辑回归、人工神经网络、支持向量机、自适应提升和偏最小二乘判别分析), 揭示了与深部子宫内膜异位症最相关的POPs: 八氯二苯并呋喃、顺式环氧化七氯、多氯联苯77以及反式九氯. 研究结果显示, ML可以对高度相关的暴露生物标志物及其与健康结果的关联进行建模, 具有应用前景.

POPs的被动采样技术使污染物集中在采样材料中, 通过吸附平衡时的分配系数计算污染物浓度[103]. 为了简化实验中分配系数的测量, Zhu等人[104]使用7种算法和2种描述符选择方法构建更稳定的定量结构-性质关系模型(quantitative structure-property relationship, QSPR), 将化学物质分子特征与计算技术相结合. 模型在拟合、鲁棒性、预测能力和应用领域方面都表现良好. 由于模型开发考虑了多类别化合物, 因此可以用来填补应用范围内相关分配系数的数据空白. QSPR模型也可使用不同的理论描述符(分子及结构)用于对POPs进行分类. Vakarelska等人[105]采用了多元统计方法, 使用了10个最合适的描述符, 将149种代表性POPs划分为4类. 此模型揭示了POPs之间特征的联系与差异, 借此可以根据未知的POPs的结构特性对其进行分类.

3　挑战

虽然ML在ECs研究中具有优势与应用前景, 但是其仍面临着不少挑战. 数据缺少以及数据质量不高是目前ECs领域面临的主要挑战, 这使部分模型的应用范围具有局限性, 只有在特定条件下才具有较高的效率. ML模型相当于一个黑箱, 难以对最终预测做出解释. 这使缺乏可解释性成为了ML领域的限制因素. 尽管通用ML模型可以指导污染物毒理预测并提高效率, 但ML目前也并非完全可信, 应谨慎解释结果. 在模型选择方面, 尽管深度神经网络等深度模型, 已被证明在区分目标之间的潜在关联和因果关系方面比浅层模型更可行[106]. 但考虑到深度模型受到对数据或计算资源

的高需求、复杂的可调参数和较差的可解释性的限制, 加上ECs领域数据较为缺乏, 更好的选择是将浅层模型和深度学习结合.

需要注意的是, ML模型相当于一个黑盒或灰盒, 难以以人类可理解的方式对最终预测做出解释. Iftikhar等人[62]希望在ARGs领域创建可解释的ML模型. 该研究首先通过ML模型预测ARGs的发生, 然后使用六种解释方法(敏感性分析、排列特征重要性、部分依赖图、SHapley加性解释、莫里斯法和傅里叶振幅灵敏度测试)对其进行解释以深入了解模型; 最终发现SHapley加性解释不仅可以解释ML模型的行为, 还可以提供有关ML模型在新条件下的行为. 可解释性的缺乏使研究者无法确切知道模型的分析依据以及该分析可靠性. 因此提升ML的可解释性能够建立研究者与模型间的信任关系, 并清除模型在实际应用中的潜在威胁[107]. 可解释的ML技术能够帮助理解ML的工作原理进而促进模型在ECs领域的应用.

目前MPs污染已经成为全球性问题, 但仍然缺乏利用ML探究MPs大尺度环境命运、迁移行为及毒性效应的相关研究. 此外, 不同的测量与表征方法所获取的MPs信息之间通常存在差异, 这导致MPs大数据存在不可避免的异质性, 从而为MPs-ML研究带来了挑战. 纳米塑料的时空差异明显, 难以通过实验采样等方法对纳米塑料的分布及风险进行大尺度评估. 目前, 纳米塑料的实地研究急剧增加, ML方法强大的预测能力为大尺度纳米塑料的评估提供了解决方案, 但采样和检测方法的不统一给模型带来了较大的不确定性.

抗生素领域中, 目前ML仍然在多种抗生素混合物的定性分析时存在较大局限性, 并且较难准确预测低浓度的抗生素, 这是未来研究应关注的方向之一. 在进行PFAS预测时需要注意, PFAS化合物由大量类似物或同系物组成, 且在亲蛋白性和离子性等物理化学性质上与传统有机污染物不同. 因此, 用于预测的数据点不应太少, 否则会使预测结果具有不确定性. ML算法可以用公共大数据来研究复杂的生物现象以及气候环境与ECs可能产生的耦合风险, 减少测试成本, 预测复杂环境情景下污染物风险.

4 结论与展望

NMs, MPs, PFAS, EDCs, 抗生素以及POPs等均属于ECs. 本文介绍了ML在各类ECs中的应用案例, 并给出了其在ECs领域研究中的优势与局限, 如图3所示. ML的应用围绕毒性预测、识别分类、性质评估、辅助去除这几点展开. 作为数据驱动的数学模型, ML能够解决输入与输出变量之间的复杂关系, 并且能够考虑多种影响因素, 进行全面分析. 将ML应用于传统风险防控技术中能够降低实验成本并节省实验时间.

统一原位的ECs采集方法, 规范ECs数据库格式以及增加时空匹配数据能提高ML模型识别环境风险的准确性. ML在生态环境领域的研究不同于其他学科, 问题导向性清晰, 前提是基于数据因果关系发现、评估生态环境风险, 最终是寻求降低风险的方法策略. ML在ECs风险调控上的研究, 尤其是涉及多目标优化的研究, 比如减污降碳, 需要投入更多的精力. 当前ECs数据大量来自于实验室模拟, 缺乏真实环境的情景考虑. 气候变化与ECs环境健康效应是当前相关研究的弱点. 基于ML的泛化性及对数据类型的广泛接受度, 未来ML在气候变化-ECs研究上可能会发挥重要作用.

目前大多数研究使用的ML算法较为老旧, 运用或发展新兴的算法是更好应对生态环境风险的亟需要求. 因果学习、因果推断及因果网络分析等是提高模型可信任度的重要方法, 是当前及未来的生态环境大数据分析的研究重点之一. 借助于因果关系推理, 相较于传统算法, ML的可解释性与泛化能力将得到提升. 目前, 针对于ECs的ML研究通常局限于常见的随机森林、支持向量机、人工神经网络等小型模型. 与此同时, AI领域的大模型, 如Transformer等, 在其他领域取得了显著的成果, 且有望从新的视角解决ECs风险问题. 大模型可以精细刻画环境中污染物浓度, 辅助决策者寻找效果更好的风险防控方案, 降低成本; 但是由于数据量的限制, 大模型在ECs研究中的应用仍然相对有限, 需要进一步深入探索和完善. 传统的单模态ML模型依赖于大量准确标记的数据和单一数据类型. 今后研究建议发展多模态ML模型处理来自不同来源的数据, 提升模型对不同数据的鲁棒性. 需要注意的是, 生态环境数据具有自身的特点, 包括非线性、尺度性、时空异质性和系统性等. 不必在具体算法上一味求新, 建立适合ECs数据特征的系统性研究体系或研究框架可能更有利于ECs领域的发展.

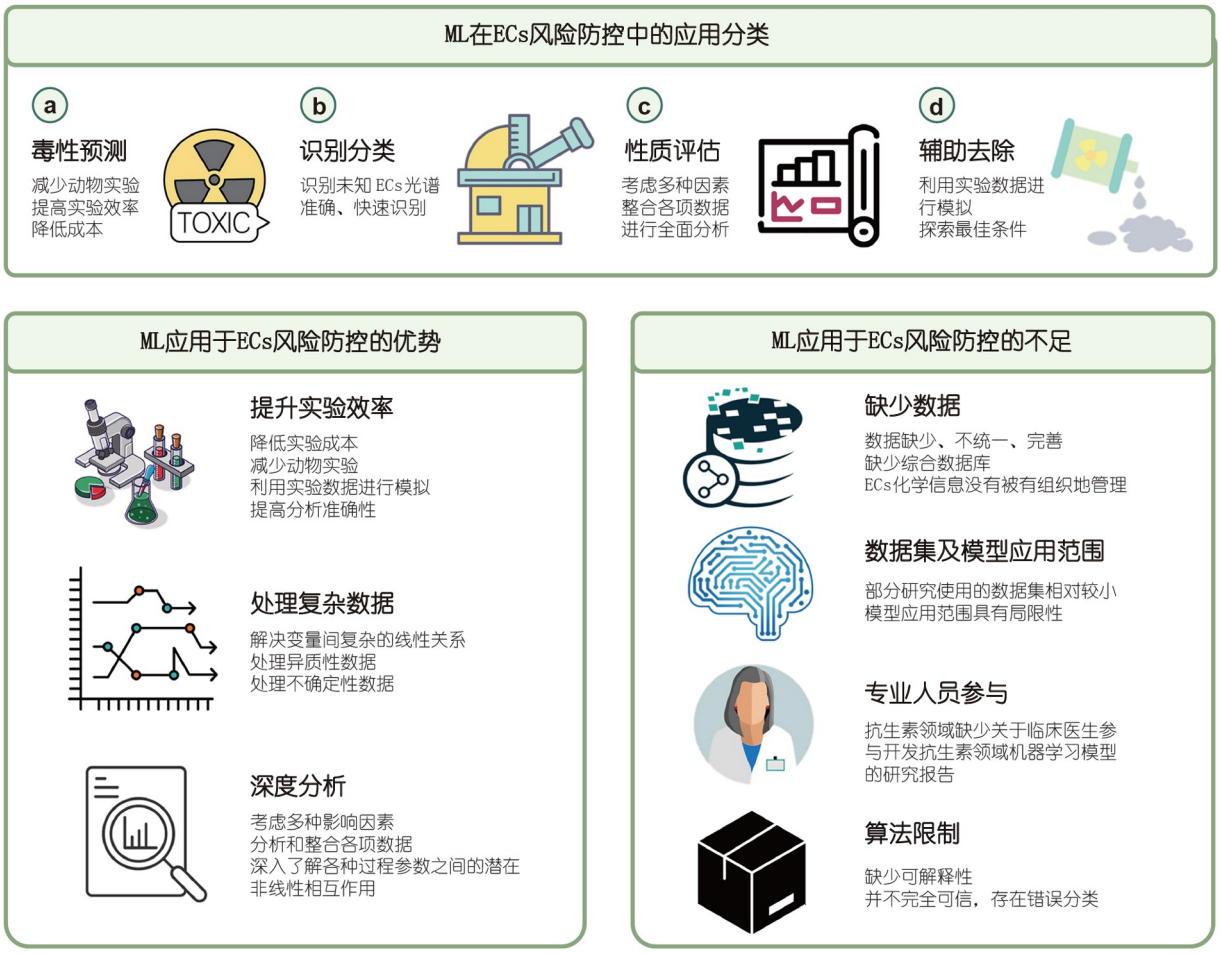


图 3　(网络版彩图)机器学习在新污染物风险防控中的应用分类、优势及不足

Figure 3　(Color online) Application classification, advantages, and disadvantages of machine learning in risk prevention and control of emerging contaminants.

建议重视ECs数据的开发，建立ECs数据共享机制，建立有组织的统一完善的数据库，提升ECs研究的数据质量. 2021年，我国成立了“可持续发展大数据国际研究中心”，是全球首个以大数据服务联合国可持续发展目标的机构. 针对生态环境问题包括ECs等，发展高质量数据库与应用平台势在必行. 在ECs研究中，数据缺乏是常见的限制因素. 然而样本量并不是越大越好，特征量也起着决定性的作用，样本量与特征量之比(sample-size to the feature-size ratio, SFR)是决定数据集是否足以解决所定义的研究问题的重要指标. 在数据数量上，应满足所用数据的SFR超过10，这被认为是必要的. 在简单的分类问题中，SFR应大于10，而对于计算机视觉等更复杂问题的稳定建模，需要更大的SFR[108]. 由于受具体问题及使用算法的影响，样本量与特征量之比的标准尚未形成共识. 对于未来的研究，在缺少具体参考依据情况下，建议所选用数据的SFR值应不低于10，或者基于模型运行效果来定. 当样本数量较小或特征数量较大时，可以使用特征选择方法保留重要特征，减少噪声和计算负荷，提高模型学习效率. 在模型解释性方面，使用高效的解释方法或者其他途径，开发具有可解释性的ML模型也是未来的研究重点. 应该注意的是，每个预测模型都有不同程度的不确定性，而它们的组合会给最终的风险评估带来更大的不确定性. 因此，应基于全面的毒性实验和化学标准品的准确定量，建立稳健的风险评估. ML也不是万能的，其有很多限制因素. 同其他依赖大型数据集的建模应用程序一样，ML算法也难以避免过拟合问题. 面对稀疏、零膨胀或高度删减的数据集，需要仔细

调整数据结构与所选的算法. 因此, 在应用ML时, 丰富的经验以及过硬的专业知识是必不可缺的. ML在真实环境中的可验证性以及对数据的依赖性是当前研究的重要争议点或关注点. 上述提到的ML不足, 也正是其区别于其他分析手段的重要优势所在. ML在生态环境领域的发展应该是在争议中不断改进、提升, 就像从20世纪提出“图灵计算机”到当前ChatGPT一样, 科学需要在包容中不断修正、前进.

参考文献

1 Wang Z, Walker G W, Muir D C G, et al. Toward a global understanding of chemical pollution: A first comprehensive analysis of national and regional chemical inventories. Environ Sci Technol, 2020, 54: 2575–2584

2 Wang X H, Yu X X, Wang S Q, et al. Environmental process, effects and risks of emerging contaminants in the estuary-coastal environment (in Chinese). Environ Sci, 2022, 43: 4810–4821 [王新红, 于晓璇, 王思权, 等. 河口-近海环境新污染物的环境过程、效应与风险. 环境科学, 2022, 43: 4810–4821]

3 Naeem M, Gill R, Gill S S, et al. Editorial: Emerging contaminants and their effect on agricultural crops. Front Plant Sci, 2023, 14: 1296252

4 Morin-Crini N, Lichtfouse E, Fourmentin M, et al. Removal of emerging contaminants from wastewater using advanced treatments: A review. Environ Chem Lett, 2022, 20: 1333–1375

5 Duarte B, Gameiro C, Matos A R, et al. First screening of biocides, persistent organic pollutants, pharmaceutical and personal care products in Antarctic phytoplankton from Deception Island by FT-ICR-MS. Chemosphere, 2021, 274: 129860

6 Krawczyk K, Marynowicz W, Gogola-Mruk J, et al. A mixture of persistent organic pollutants detected in human follicular fluid increases progesterone secretion and mitochondrial activity in human granulosa HGrC1 cells. Reproductive Toxicol, 2021, 104: 114–124

7 Kedzierski M, Falcou-Préfol M, Kerros M E, et al. A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the Mediterranean Sea. Chemosphere, 2019, 234: 242–251

8 Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349: 255–260

9 Chakraborty C, Bhattacharya M, Dhama K, et al. Deep learning research should be encouraged more and more in different domains of surgery: An open call—Correspondence. Int J Surg, 2022, 104: 106749

10 Liang Z, Liu W W, Wu T R, et al. Survey on repair strategies for deep neural network (in Chinese). J Softw, 2024, 35: 1231–1256 [梁震, 刘万伟, 吴陶然, 等. 深度神经网络修复策略综述. 软件学报, 2024, 35: 1231–1256]

11 Wang F, Xie K, Han L, et al. Research on support vector machine optimization based on improved quantum genetic algorithm. Quantum Inf Process, 2023, 22: 380

12 Ozturk Kiyak E, Ghasemkhani B, Birant D. High-level k-nearest neighbors (HLKNN): A supervised machine learning model for classification analysis. Electronics, 2023, 12: 3828

13 Panhalkar A R, Doye D D. A novel approach to build accurate and diverse decision tree forest. Evol Intel, 2022, 15: 439–453

14 Rhodes J S, Cutler A, Moon K R. Geometry- and accuracy-preserving random forest proximities. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 10947–10959

15 Mishina Y, Murata R, Yamauchi Y, et al. Boosted random forest. IEICE Trans Inf Syst, 2015, E98.D: 1630–1636

16 Chen S L, Webb G I, Liu L Y, et al. A novel selective naïve Bayes algorithm. Knowledge-Based Syst, 2020, 192: 105361

17 Sethi J K, Mittal M. Efficient weighted naive bayes classifiers to predict air quality index. Earth Sci Inform, 2022, 15: 541–552

18 Li B, Yu Q, Peng L. Ensemble of fast learning stochastic gradient boosting. Commun Stat-Simul Computat, 2019, 51: 40–52

19 Zhong S F, Zhang K, Bagheri M, et al. Machine learning: New ideas and tools in environmental science and engineering. Environ Sci Technol, 2021, 55: 12741–12754

20 Navidpour A H, Hosseinzadeh A, Huang Z G, et al. Application of machine learning algorithms in predicting the photocatalytic degradation of perfluorooctanoic acid. Catal Rev Sci Eng, 2022, 66: 687–712

21 Lin J, Liu H, Zhang J. Recent advances in the application of machine learning methods to improve identification of the microplastics in environment. Chemosphere, 2022, 307: 136092

22 Wu Z H, Zhang S Y, Dong S P, et al. Overview of application of machine learning in field of nanomaterials risk assessment (in Chinese). Asian J Ecotoxicol, 2022, 17: 139–151 [武子豪, 张司雨, 董仕鹏, 等. 机器学习在纳米材料风险评估中的应用. 生态毒理学报, 2022, 17: 139–151]

23 Hu X G, Li J W, Li J Q, et al. Scientific questions on the biological effects of nanomaterials based on machine learning (in Chinese). Environ Eng, 2022, 40: 171–181 [胡献刚, 李佳薇, 李佳蔷, 等. 基于机器学习的纳米材料生物效应研究的科学问题. 环境工程, 2022, 40: 171–181]

24 Lei L, Pang R, Han Z, et al. Current applications and future impact of machine learning in emerging contaminants: A review. Crit Rev Environ Sci Tech, 2023, 53: 1817–1835

25 Rodrigues N M, Batista J E, Mariano P, et al. Artificial intelligence meets marine ecotoxicology: Applying deep learning to bio-optical data from marine diatoms exposed to legacy and emerging contaminants. Biology, 2021, 10: 932

26 Bradley P M, Journey C A, Romanok K M, et al. Multi-region assessment of chemical mixture exposures and predicted cumulative effects in USA wadeable urban/agriculture-gradient streams. Sci Total Environ, 2021, 773: 145062

27 Johnson A C, Jin X, Nakada N, et al. Learning from the past and considering the future of chemicals in the environment. Science, 2020, 367: 384–387

28 Feng X, Li D, Liang W, et al. Recognition and prioritization of chemical mixtures and transformation products in Chinese estuarine waters by suspect screening analysis. Environ Sci Technol, 2021, 55: 9508–9517

29 Chen J, Wang B, Huang J, et al. A machine-learning approach clarifies interactions between contaminants of emerging concern. One Earth, 2022, 5: 1239–1249

30 Shin J, Lee Y G, Lee S H, et al. Single and competitive adsorptions of micropollutants using pristine and alkali-modified biochars from spent coffee grounds. J Hazard Mater, 2020, 400: 123102

31 Haider Jaffari Z, Jeong H, Shin J, et al. Machine-learning-based prediction and optimization of emerging contaminants' adsorption capacity on biochar materials. Chem Eng J, 2023, 466: 143073

32 Hochella Jr. M F, Mogk D W, Ranville J, et al. Natural, incidental, and engineered nanomaterials and their impacts on the Earth system. Science, 2019, 363: eaau8299

33 Kang W L, Li X K, Mu L, et al. Nanoscale colloids induce metabolic disturbance of zebrafish at environmentally relevant concentrations. Environ Sci-Nano, 2019, 6: 1562–1575

34 Li X, Sun S, Guo S, et al. Identifying the phytotoxicity and defense mechanisms associated with graphene-based nanomaterials by integrating multiomics and regular analysis. Environ Sci Technol, 2021, 55: 9938–9948

35 Roubeau Dumont E, Elger A, Azéma C, et al. Cutting-edge spectroscopy techniques highlight toxicity mechanisms of copper oxide nanoparticles in the aquatic plant *Myriophyllum spicatum*. Sci Total Environ, 2022, 803: 150001

36 Bai W, Zhang Z Y, Tian W J, et al. Toxicity of zinc oxide nanoparticles to zebrafish embryo: A physicochemical study of toxicity mechanism. J Nanopart Res, 2009, 12: 1645–1654

37 Sun A Y, Scanlon B R. How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. Environ Res Lett, 2019, 14: 073001

38 Yalezo N, Musee N. Meta-analysis of engineered nanoparticles dynamic aggregation in freshwater-like systems using machine learning techniques. J Environ Manage, 2023, 337: 117739

39 Kotzabasaki M I, Sotiropoulos I, Sarimveis H. QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: An integrated study from data curation to model development. RSC Adv, 2020, 10: 5385–5391

40 Huang Y, Li X, Xu S, et al. Quantitative structure-activity relationship models for predicting inflammatory potential of metal oxide nanoparticles. Environ Health Perspect, 2020, 128: 67010

41 Yu F, Wei C, Deng P, et al. Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. Sci Adv, 2021, 7: abf4130

42 Yan X, Sedykh A, Wang W, et al. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. Nat Commun, 2020, 11: 2519

43 Geitner N K, Zhao W, Ding F, et al. Mechanistic insights from discrete molecular dynamics simulations of pesticide-nanoparticle interactions. Environ Sci Technol, 2017, 51: 8396–8404

44 Brandon J, Goldstein M, Ohman M D. Long-term aging and degradation of microplastic particles: Comparing in situ oceanic and experimental weathering patterns. Mar Pollution Bull, 2016, 110: 299–308

45 Lv X M, Dong Q, Zuo Z Q, et al. Microplastics in a municipal wastewater treatment plant: Fate, dynamic distribution, removal efficiencies, and control strategies. J Cleaner Product, 2019, 225: 579–586

46 Zeng H, Hu X, Ouyang S, et al. Microplastics weaken the adaptability of Cyanobacterium *Synechococcus* sp. to ocean warming. Environ Sci Technol, 2023, 57: 9005–9017

47 Liu X J, Wang S T, Mu L, et al. Microplastics reshape the fate of aqueous carbon by inducing dynamic changes in biodiversity and chemodiversity. Environ Sci Technol, 2023, 57: 10415–10425

48 Salimi A, Alavehzadeh A, Ramezani M, et al. Differences in sensitivity of human lymphocytes and fish lymphocytes to polyvinyl chloride microplastic toxicity. Toxicol Ind Health, 2022, 38: 100–111

49 Prata J C, da Costa J P, Lopes I, et al. Environmental exposure to microplastics: An overview on possible human health effects. Sci Total Environ, 2020, 702: 134455

50 Nguyen M K, Hadi M, Lin C, et al. Microplastics in sewage sludge: Distribution, toxicity, identification methods, and engineered technologies. Chemosphere, 2022, 308: 136455

51 Luo Y, Su W, Xu D, et al. Component identification for the SERS spectra of microplastics mixture with convolutional neural network. Sci Total Environ, 2023, 895: 165138

52 Patil S M, Rane N R, Bankole P O, et al. An assessment of micro- and nanoplastics in the biosphere: A review of detection, monitoring, and remediation technology. Chem Eng J, 2022, 430: 132913

53 Lei B, Bissonnette J R, Hogan Ú E, et al. Customizable machine-learning models for rapid microplastic identification using Raman microscopy. Anal Chem, 2022, 94: 17011–17019

54 Ai W, Chen G L, Yue X J, et al. Application of hyperspectral and deep learning in farmland soil microplastic detection. J Hazard Mater, 2023, 445: 130568

55 Mu X, Huang Y, Li J, et al. New insights into the mechanism of phthalate-induced developmental effects. Environ Pollut, 2018, 241: 674–683

56 Kida M, Pochwat K, Ziembowicz S. Assessment of machine learning-based methods predictive suitability for migration pollutants from microplastics degradation. J Hazard Mater, 2024, 461: 132565

57 Winkler A, Fumagalli F, Cella C, et al. Detection and formation mechanisms of secondary nanoplastic released from drinking water bottles. Water Res, 2022, 222: 118848

58 Xu G, Cheng H, Jones R, et al. Surface-enhanced raman spectroscopy facilitates the detection of microplastics <1 μm in the environment. Environ Sci Technol, 2020, 54: 15594–15603

59 Balakrishnan V, Kherabi Y, Ramanathan G, et al. Machine learning approaches in diagnosing tuberculosis through biomarkers: A systematic review. Prog Biophys Mol Biol, 2023, 179: 16–25

60 Xie L, Luo S, Liu Y, et al. Automatic identification of individual nanoplastics by raman spectroscopy based on machine learning. Environ Sci Technol, 2023, 57: 18203–18214

61 Van Boeckel T P, Gandra S, Ashok A, et al. Global antibiotic consumption 2000 to 2010: An analysis of national pharmaceutical sales data. Lancet Infect Dis, 2014, 14: 742–750

62 Iftikhar S, Karim A M, Karim A M, et al. Prediction and interpretation of antibiotic-resistance genes occurrence at recreational beaches using machine learning models. J Environ Manage, 2023, 328: 116969

63 Zhu Y G, Zhao Y, Li B, et al. Continental-scale pollution of estuaries with antibiotic resistance genes. Nat Microbiol, 2017, 2: 16270

64 Sun Y, Guo G, Tian F, et al. Antibiotic resistance genes and bacterial community on the surfaces of five cultivars of fresh tomatoes. Ecotoxicology, 2021, 30: 1550–1558

65 Han X, Ma X L, Chao S L, et al. Influence of nanomaterials on the spread of environmental antibiotic resistance genes: A review (in Chinese). Asian J Ecotoxicol, 2019, 14: 46–54 [韩雪, 马晓琳, 晁韶良, 等. 纳米材料对环境抗生素抗性基因污染扩散影响的研究进展. 生态毒理学报, 2019, 14: 46–54]

66 Liu Y, Wang L, Pan B, et al. Toxic effects of diclofenac on life history parameters and the expression of detoxification-related genes in *Daphnia magna*. Aquat Toxicol, 2017, 183: 104–113

67 Jang J, Abbas A, Kim M, et al. Prediction of antibiotic-resistance genes occurrence at a recreational beach with deep learning models. Water Res, 2021, 196: 117001

68 Sun W, Qian X, Gu J, et al. Mechanism and effect of temperature on variations in antibiotic resistance genes during anaerobic digestion of dairy manure. Sci Rep, 2016, 6: 30237

69 Sun C, Li W, Chen Z, et al. Responses of antibiotics, antibiotic resistance genes, and mobile genetic elements in sewage sludge to thermal

hydrolysis pre-treatment and various anaerobic digestion conditions. Environ Int, 2019, 133: 105156
70 Batstone D J, Keller J, Angelidaki I, et al. The IWA anaerobic digestion model No 1 (ADM1). Water Sci Technol, 2002, 45: 65–73
71 Haffiez N, Chung T H, Zakaria B S, et al. Exploration of machine learning algorithms for predicting the changes in abundance of antibiotic resistance genes in anaerobic digestion. Sci Total Environ, 2022, 839: 156211
72 Zheng D, Yin G, Liu M, et al. Global biogeography and projection of soil antibiotic resistance genes. Sci Adv, 2022, 8: abq8015
73 Zhang Z, Zhang Q, Wang T, et al. Assessment of global health risk of antibiotic resistance genes. Nat Commun, 2022, 13: 1553
74 Xu N, Qiu D, Zhang Z, et al. A global atlas of marine antibiotic resistance genes and their expression. Water Res, 2023, 244: 120488
75 Phoon B L, Ong C C, Mohamed Saheed M S, et al. Conventional and emerging technologies for removal of antibiotics from wastewater. J Hazard Mater, 2020, 400: 122961
76 Zhang P, Liu C, Lao D, et al. Unveiling the drives behind tetracycline adsorption capacity with biochar through machine learning. Sci Rep, 2023, 13: 11512
77 Glüge J, Scheringer M, Cousins I T, et al. An overview of the uses of per- and polyfluoroalkyl substances (PFAS). Environ Sci-Processes Impacts, 2020, 22: 2345–2373
78 Liu L Q, Qu Y X, Huang J, et al. Per- and polyfluoroalkyl substances (PFASs) in Chinese drinking water: Risk assessment and geographical distribution. Environ Sci Eur, 2021, 33: 6
79 Qin X D, Qian Z, Vaughn M G, et al. Positive associations of serum perfluoroalkyl substances with uric acid and hyperuricemia in children from Taiwan. Environ Pollut, 2016, 212: 519–524
80 Manojkumar Y, Pilli S, Rao P V, et al. Sources, occurrence and toxic effects of emerging per- and polyfluoroalkyl substances (PFAS). Neurotoxicol Teratol, 2023, 97: 107174
81 Charbonnet J A, Rodowa A E, Joseph N T, et al. Environmental source tracking of per- and polyfluoroalkyl substances within a forensic context: Current and future techniques. Environ Sci Technol, 2021, 55: 7237–7245
82 Stults J F, Higgins C P, Helbling D E. Integration of per- and polyfluoroalkyl substance (PFAS) fingerprints in fish with machine learning for PFAS source tracking in surface water. Environ Sci Technol Lett, 2023, 10: 1052–1058
83 Xiang L, Li Y W, Yu P F, et al. Food safety concerns: Crop breeding as a potential strategy to address issues associated with the recently lowered reference doses for perfluorooctanoic acid and perfluorooctane sulfonate. J Agric Food Chem, 2020, 68: 48–58
84 Xiang L, Qiu J, Chen Q Q, et al. Development, evaluation, and application of machine learning models for accurate prediction of root uptake of per- and polyfluoroalkyl substances. Environ Sci Technol, 2023, 57: 18317–18328
85 Han B C, Liu J S, Bizimana A, et al. Identifying priority PBT-like compounds from emerging PFAS by nontargeted analysis and machine learning models. Environ Pollut, 2023, 338: 122663
86 Howard P H, Muir D C G. Identifying new persistent and bioaccumulative organics among chemicals in commerce. Environ Sci Technol, 2010, 44: 2277–2285
87 Li Z, Zhang P, Shao T, et al. Different nanostructured In_2O_3 for photocatalytic decomposition of perfluorooctanoic acid (PFOA). J Hazard Mater, 2013, 260: 40–46
88 Hosseinzadeh A, Zhou J L, Altaee A, et al. Machine learning modeling and analysis of biohydrogen production from wastewater by dark fermentation process. Bioresource Tech, 2022, 343: 126111
89 Vilela C L S, Bassin J P, Peixoto R S. Water contamination by endocrine disruptors: Impacts, microbiological aspects and trends for environmental protection. Environ Pollut, 2018, 235: 546–559
90 Zhang Q, Lu M, Dong X, et al. Potential estrogenic effects of phosphorus-containing flame retardants. Environ Sci Technol, 2014, 48: 6995–7001
91 Heo S K, Safder U, Yoo C K. Deep learning driven QSAR model for environmental toxicology: Effects of endocrine disrupting chemicals on human health. Environ Pollut, 2019, 253: 29–38
92 Safder U, Nam K J, Kim D, et al. Quantitative structure-property relationship (QSPR) models for predicting the physicochemical properties of polychlorinated biphenyls (PCBs) using deep belief network. Ecotoxicol Environ Saf, 2018, 162: 17–28
93 Chen Y, Cheng F, Sun L, et al. Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors. Ecotoxicol Environ Saf, 2014, 110: 280–287
94 Yoh K, Ikeda K, Horie K, et al. Roles of estrogen, estrogen receptors, and estrogen-related receptors in skeletal muscle: Regulation of

mitochondrial function. Int J Mol Sci, 2023, 24: 1853
95 Shanle E K, Xu W. Endocrine disrupting chemicals targeting estrogen receptor signaling: Identification and mechanisms of action. Chem Res Toxicol, 2011, 24: 6–19
96 Russo D P, Zorn K M, Clark A M, et al. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. Mol Pharm, 2018, 15: 4361–4370
97 Zhang Q, Yan L, Wu Y, et al. A ternary classification using machine learning methods of distinct estrogen receptor activities within a large collection of environmental chemicals. Sci Total Environ, 2017, 580: 1268–1275
98 Li Y, Cao P, Wang S, et al. Research on the treatment mechanism of anthraquinone dye wastewater by algal-bacterial symbiotic system. Bioresource Tech, 2022, 347: 126691
99 Fu W, Li X, Yang Y, et al. Enhanced degradation of bisphenol A: Influence of optimization of removal, kinetic model studies, application of machine learning and microalgae-bacteria consortia. Sci Total Environ, 2023, 858: 159876
100 Cano-Sancho G, Ploteau S, Matta K, et al. Human epidemiological evidence about the associations between exposure to organochlorine chemicals and endometriosis: Systematic review and meta-analysis. Environ Int, 2019, 123: 209–223
101 Tomiazzi J S, Pereira D R, Judai M A, et al. Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke. Environ Sci Pollut Res, 2019, 26: 6481–6491
102 Matta K, Vigneau E, Cariou V, et al. Associations between persistent organic pollutants and endometriosis: A multipollutant assessment using machine learning algorithms. Environ Pollut, 2020, 260: 114066
103 Alvarez D A, Cranor W L, Perkins S D, et al. Chemical and toxicologic assessment of organic contaminants in surface water using passive samplers. J Env Qual, 2008, 37: 1024–1033
104 Zhu T, Tao C. Prediction models with multiple machine learning algorithms for POPs: The calculation of PDMS-air partition coefficient from molecular descriptor. J Hazard Mater, 2022, 423: 127037
105 Vakarelska E, Nedyalkova M, Vasighi M, et al. Persistent organic pollutants (POPs)-QSPR classification models by means of machine learning strategies. Chemosphere, 2022, 287: 132189
106 Oh E, Liu R, Nel A, et al. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. Nat Nanotech, 2016, 11: 479–486
107 Ji S L, Li J F, Du T Y, et al. Survey on techniques, applications and security of machine learning interpretability (in Chinese). J Comput Res Devel, 2019, 56: 2071–2096 [纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56: 2071–2096]
108 Zhu J J, Yang M, Ren Z J. Machine learning in environmental research: Common pitfalls and best practices. Environ Sci Technol, 2023, 57: 17671–17689

Application of machine learning in environmental risk identification and control of emerging contaminants: Research progress and challenges

HU XianGang[1], WANG ZhangJia[1], DENG Peng[1], YU FuBo[1], MU Li[2], WANG Sai[3] & ZHOU QiXing[1]

[1] *Key Laboratory of Pollution Processes and Environmental Criteria (Ministry of Education), College of Environmental Science and Engineering, Nankai University, Tianjin 300350, China;*
[2] *Key Laboratory for Environmental Factors Control of Agro-Product Quality Safety (Ministry of Agriculture and Rural Affairs), Institute of Agro-Environmental Protection, Ministry of Agriculture and Rural Affairs, Tianjin 300191, China;*
[3] *State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University, Haikou 570228, China*

As technology advances, an increasing number of emerging contaminants (ECs) are being generated and released into the environment. These ECs exhibit biological toxicity and environmental persistence, posing risks to ecological systems and human health. Consequently, many studies have focused on preventing and managing the risks associated with ECs. However, the detection and control of EC risks in natural environments through traditional analysis methods are complex, time-consuming, and labor-intensive. Machine learning (ML) is a data-driven research method that utilizes existing data to train models and forecast the trends of interest. Recently, ML has been widely adopted in the identification and control of risks associated with ECs owing to its ability to comprehend intricate parameter relationships within datasets. The integration of traditional technology and ML can reduce computational costs and save experimental time and energy consumption by optimizing and reducing the number of experiments required. The application of ML in addressing ECs can be categorized into several categories, such as toxicity prediction, identification/classification, property assessment, and assisted removal. This review outlines the application of ML in addressing various types of ECs, such as nanomaterials, microplastics, antibiotic-resistant genes, perfluoroalkyl and polyfluoroalkyl substances, endocrine-disrupting chemicals, and persistent organic pollutants. ML can reduce the need for animal experiments in toxicity studies. Nonetheless, there is limited data available in this field, and certain models rely on small datasets with low data quality that limit the ML application scope. However, not all environmental problems are suitable for ML. Hence, the over-reliance on this approach is not recommended. Complex ML models often lack interpretability, making it challenging to understand how ML generates specific predictive outcomes. The absence of interpretability in complex ML poses a challenge in explaining and understanding the environmental behavior and risks associated with ECs. In addition to the intricacies within ML models, the accuracy of ML models heavily relies on the quality of the input data. Therefore, the focus should be on enhancing the quality of EC data, promoting data sharing, establishing a comprehensive and unified database, and developing suitable frameworks for managing ECs.

machine learning, emerging contaminants, nanomaterials, antibiotics, microplastics, endocrine-disrupting chemicals

doi: 10.1360/SST-2024-0037