生态与农村环境学报
Journal of Ecology and Rural Environment
ISSN 1673-4831,CN 32-1766/X

《生态与农村环境学报》网络首发论文

题目：人工智能关键技术在化学物质毒性预测中的应用研究进展
作者：李思敏，张后虎，张佩雯，卜元卿
DOI：10.19741/j.issn.1673-4831.2025.0337
收稿日期：2025-05-07
网络首发日期：2025-07-18
引用格式：李思敏，张后虎，张佩雯，卜元卿．人工智能关键技术在化学物质毒性预测中的应用研究进展[J/OL]．生态与农村环境学报．https://doi.org/10.19741/j.issn.1673-4831.2025.0337

DOI：10.19741/j.issn.1673-4831.2025.0337

人工智能关键技术在化学物质毒性预测中的应用研究进展

李思敏[1]，张后虎[1]，张佩雯[1]，卜元卿[1,2①]（1.生态环境部南京环境科学研究所固体废物污染防治研究中心 南京 210033；2.南京信息工程大学江苏省大气环境与装备技术协同创新中心 南京 210044）

摘要：随着化学物质数量和种类的持续增加及其潜在环境风险问题的加剧，传统毒性测试方法已难以满足高通量筛查与系统性风险评估的需求。人工智能技术，特别是大数据与机器学习技术，在化学物质毒性预测中展现出显著的应用潜力。本文系统综述了人工智能关键技术在化学物质毒性预测模型构建全过程中的应用进展，涵盖数据收集、清洗与预处理、分子描述符计算、特征提取与选择、模型训练与验证、模型适用域界定与可解释性分析等核心环节。此外，还结合了国内在人工智能毒性预测领域的主要研究成果与本研究团队在人工智能辅助毒性预测领域的研究实践，展示了本团队在数据标准化、分子特征工程、模型开发、适用域与模型可解释性等方面的关键成果。最后，针对当前毒性预测模型在数据异质性、多模态数据融合、复杂毒性终点预测及预测结果认可度等方面存在的挑战，提出了未来发展方向。本文旨在推动人工智能技术在化学物质毒性预测中的深度应用，为高效、可靠的环境污染物风险评估提供理论基础与技术支持。

关键词：人工智能；大数据技术；机器学习；化学物质；毒性预测

Research Progress on the Application of Key Artificial Intelligence Technologies in Chemical Toxicity Prediction. *LI Si-min, ZHANG Hou-hu, ZHANG Pei-wen, BU Yuan-qing* (1. Nanjing Institute of Environmental Sciences, MEE, Nanjing 210033, 2. Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: With the continuous increase in the number and diversity of chemicals and the growing concerns over their potential environmental risks, traditional toxicity testing methods are no longer sufficient to meet the demands of high-throughput screening and systematic risk assessment. Artificial intelligence (AI) technologies, particularly big data and machine learning, have shown remarkable potential in chemical toxicity prediction. Therefore, this review systematically summarizes the recent advances in the application of key AI technologies throughout the entire process of toxicity prediction modeling, including data acquisition, cleaning and preprocessing, molecular descriptor calculation, feature extraction and selection, model training and validation, applicability domain definition, and interpretability analysis. Furthermore, this review incorporates recent advances on AI technologies based toxicity prediction by domestic research teams, and highlights the contributions in AI-assisted toxicity prediction by our research team, focusing on data standardization, molecular feature engineering, model development, applicability domain expansion, and enhancement of model interpretability. Finally, it discusses the current challenges faced by predictive models, such as data heterogeneity, multi-modal data integration, complex toxicity endpoint prediction, and lack of consensus in prediction reliability, and proposes directions for future research. This review aims to promote the in-depth application of AI technologies in chemical toxicity

收稿日期：2025-05-07

基金项目：国家重点研发计划课题（2023YFC3706603）；国家自然科学基金长江水科学研究联合基金项目（U2340202）；生态环境部预算项目：化学品与重金属污染防治监督管理、农村和农业环境保护管理

① 通信作者：卜元卿，E-mail: byq@nies.org

prediction and provide a theoretical and technical foundation for efficient and reliable environmental risk assessment of emerging contaminants.

Keywords: Artificial Intelligence; big data technology; machine learning; chemicals; toxicity prediction

1 引言

随着工业化和城市化进程的加速，环境中化学物质的种类和数量激增，其中，新污染物的广泛存在和持续释放，已成为亟待解决的全球性环境问题之一[1-2]。然而，大多数化学物质缺乏系统的生态毒理数据，且在复杂环境条件下的危害效应差异显著，给化学物质环境管理与风险防控带来了严峻挑战[2-3]。传统的实验室毒性评价方法虽然科学性高，但存在成本高、周期漫长以及动物实验伦理限制等问题，难以有效应对化学污染物种类繁多、结构复杂及长期累积效应等特征[4-5]。

定量结构-活性关系（Quantitative Structure-Activity Relationship, QSAR）模型是化学物质毒性预测的核心工具。该模型通过建立化学物质的分子结构或理化性质与其生物活性或毒性效应之间的定量关系，实现对化学物质毒性的快速预测与评估[5-7]。然而，传统 QSAR 模型在应对化学物质结构复杂性和数据异构性方面仍面临诸多挑战。一方面，传统 QSAR 模型依赖预定义的分子描述符和简单线性建模方法，难以有效捕捉结构特征与毒性效应之间潜在的非线性关系；另一方面，在处理高维、多源异构数据时易出现过拟合、欠拟合及适用域受限等问题，制约了模型的泛化能力与实际应用价值[6-7]。

人工智能技术的引入为 QSAR 模型的发展带来了新的突破[8-10]。其中，大数据技术能够从多源异构的数据中高效提取、存储和管理海量信息，为模型训练提供丰富的数据资源[11-12]。机器学习技术能够从大规模复杂数据中自动学习潜在规律，尤其擅长处理非线性关系、高维度特征、噪声信息以及非结构化数据等，可有效提升 QSAR 模型的预测精度与泛化能力[13-14]。此外，机器学习模型还能够持续优化与更新，在不断引入新数据的过程中提高模型预测性能和适用性，从而为化学物质毒性预测提供更加精准、高效和自动化的技术支持[15-16]。

本文将通过分析人工智能关键技术在化学物质毒性预测中的应用进展，系统探讨这些技术在毒性预测模型构建全过程体系的应用与优势，结合国内人工智能技术在毒性预测应用中的研究进展与本研究团队的研究成果，进一步推动人工智能技术在化学物质毒性预测中的深度应用。

2 关键人工智能技术及其应用

2.1 大数据技术在化学物质毒性预测中的应用

数据是化学物质毒性预测的核心基础。常见的毒性数据的获取来源包括开源数据库、公开报告和发表文献等。其中，从数据库或文献中获取毒性数据是最常见的途径，常用的数据库如表 1 所示，涵盖了急性毒性、慢性毒性、生态毒性、健康毒性及环境行为等多类信息。在此基础上，大数据技术通过对海量、多源、异构数据进行自动化抓取、清洗、标注、存储与可视化管理[17]，构成了内容丰富类型多样的高质量数据集或数据库（图 1），为毒性预测模型提供强有力的数据支撑[18-21]。

表 1 常见的毒性数据库及其介绍

Table 1 The common toxicity databases and introductions

数据库	数据类型	简要介绍
ECOTOX	生态毒性数据	ECOTOX 是一个综合性知识库，记载了 13,031 种化合物的毒性数据，包含 521,636 种陆生生物和 684,547 种水生生物（截至 2025.04.20）
PubChem	物理化学性质、生物活性、毒性信息等	PubChem 是全球最大的免费化学信息库，包含超过 1.18 亿种化合物和 2.95 亿项生物测定信息
ChEMBL	化学、生物活性、基因组数据	ChEMBL 是一个人工编辑的具有药物类似特性的生物活性分子数据库，含有超过 240 万种化合物和 160 万项生物测定信息
ToxCast	体外毒性数据	包含了对 2000 多个化学品进行的 700 多项高通量测试结果。
TOX21	体外毒性数据	包含超过 12,000 种化学物质的生物活性数据，主要用于评估化学物质对 12 种不同生物学终点的毒性，包括核受体活性和应激反应。
eChemPortal	物理化学性质、生态毒性和健康毒性数据	eChemPortal 是一个全球化合物信息门户，包含近 26,700 种物质以及超过 130 万条化学性质记录。
Integrated Chemical Environment (ICE)	体内、体外和计算数据	ICE 提供了精选的大量体内、体外和计算数据以及计算工具，数据库中包含大约 100 万种物质。
PPDB	物理化学性质、人类健康和生态毒理数据	PPDB 包含 1,958 种农药的化学特性、理化性质、人类健康和生态毒理数据。
EnviroTox	物理化学性质和毒性数据	EnviroTox 数据库包含 1,563 个物种的 91,217 条水生毒性记录。
ChemSpider	物理化学性质和毒性数据	ChemSpider 是一个免费的化学结构数据库，提供对来自数百个数据源的 1 亿多个结构的快速文本和结构搜索访问。
ToxRefDB	体内毒性数据	ToxRefDB 包含来自 1,100 多种化合物的 5,900 多项指南或类似指南研究的体内研究数据。
DSSTox	化学结构信息和毒性数据	DSSTox 包含映射到数据的精选化合物，包括化学标识符并在适当的情况下提供化学结构表示。

CompTox	化学性质、危害、生物活性信息等	CompTox Chemicals Dashboard 是一个广泛使用的资源库，提供了 100 多万种化合物的化学、毒性和暴露信息。
ECHA	物理化学性质、生态毒性和健康毒性等信息	ECHA 是关于欧洲制造和进口化合物的唯一信息来源，涵盖了它们的危险特性、分类和标签，以及如何安全使用它们的信息。

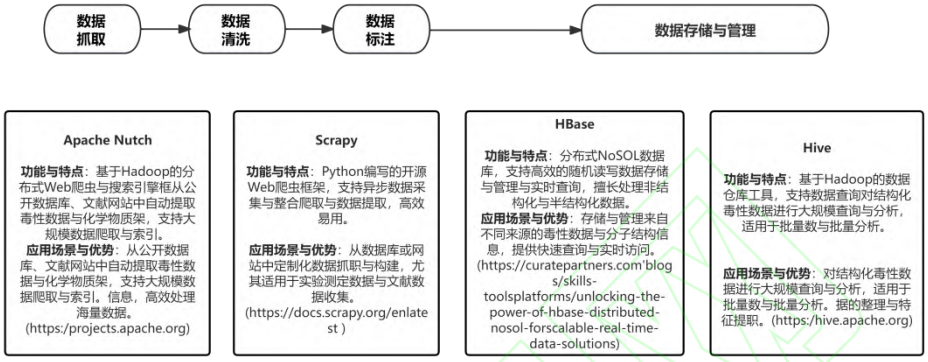

图 1 利用大数据技术的数据库系统构建流程及常见的大数据技术与功能

Figure 1 The construction process of database systems utilizing big data technology, and the common big data technologies and their functions

在数据收集与整合方面，Apache Nutch 和 Scrapy 是目前广泛使用的两种 Web 爬虫工具。其中，Apache Nutch 基于 Hadoop 架构，具备良好的拓展性与并行处理能力，能够实现大规模数据的自动爬取与索引；而 Scrapy 则擅长异步爬取与数据提取，适用于从特定的网站或资料中快速构建爬虫系统[18-19]。在数据存储与管理方面，HBase 和 Hive 提供了对结构化与半结构化数据的高效管理。HBase 作为分布式 NoSQL 数据库，支持实时查询与高并发访问，适用于大体量、动态更新的数据存储场景；而 Hive 则基于 Hadoop 架构，适合批量分析与查询结构化数据[20]。此外，Cassandra 与 MongoDB 等高性能 NoSQL 数据库在非结构化数据的存储与快速查询中表现出色，广泛应用于文献、报告与实验原始记录等数据管理场景[21]。通过这些大数据技术的协同应用，可显著提升数据采集效率、处理能力与存储能力，为后续机器学习技术在化学物质毒性预测模型构建提供坚实的数据基础与技术保障[22-23]。

2.2 机器学习技术在毒性预测中的应用

机器学习技术作为人工智能领域的核心手段，逐渐成为化学物质毒性预测的重要方法之一[3,9]。相较于传统的 QSAR 模型，机器学习技术具有更强的非线性建模能力与自动化特征学习能力，尤其在高维、多样化以及噪声显著的数据环境中表现出色，能够准确捕捉化学物质的结构特征与毒性效应之间的复杂关系，从而实现更为精准、高效的毒性预测与评估[13-

14,24-25]。

2.2.1 数据清洗与预处理

QSAR 模型建模的首要步骤是收集与整理化学物质的分子结构信息及其对应的毒性效应数据[26]。利用大数据技术收集的数据通常具有体量大、格式异构、非结构化显著等特点，难以直接用于模型训练与分析。机器学习技术被广泛应用于数据清洗、标准化与数据填补等预处理环节，显著提高了数据质量和可靠性[27]。

常见的毒性数据清洗与预处理方法主要包括异常值检测与去除、数据标准化、缺失值填补以及数据增强与类别平衡等（表 2）。在化学物质毒性预测中，原始数据集中可能包含由于实验误差、录入错误或数据噪声引入的异常值。若不加以识别与处理，容易导致模型训练中出现偏差，影响预测结果的准确性与稳健性。除了基于专业知识和统计分布方法（箱式图、四分位法等）排除异常值外，还可利用支持向量机（SVM）、孤立森林（Isolation Forest）与局部离群因子（LOF）等机器学习算法自动检测与剔除异常值，提高数据集的质量与模型稳定性[28]。在大规模数据收集中，不可避免地还会存在缺失数据的现象。为了减少缺失数据对模型训练的负面影响，可直接采用均值或中位数插补等简单方法进行初步处理，也可以借助 K 近邻算法（KNN）、随机森林（RF）等机器学习方法对缺失值进行自动填补，从而提高数据的完整性[29]。此外，在化学物质毒性预测中，不同毒性终点或类型（如有毒/无毒）的样本数量可能存在显著差异，从而造成数据集不平衡，导致模型对少数类别样本的预测精度与泛化能力可能受限[30]。如若有毒样本数量远高于无毒样本数量，模型可能在总体上表现出较高的准确度，但对无毒样本的预测效果却严重下降[30-31]。目前常通过 SMOTE 算法（Synthetic Minority Over-sampling Technique）、集成学习方法（如 Adaboost、Bagging）与随机欠采样技术等来增强少数样本数或删除多数样本数以调整数据结构，使各类别数据的数量趋于平衡，改善模型在不平衡数据下的预测精度与稳健性[32-33]。最后，还需要对数据进行标准化处理，以保证数据的可比性与统一性。常用的方法包括归一化（Min-Max Scaling）、标准差标准化（Z-score Normalization）以及数据平滑技术（如对数转换、Box-Cox 转换）等。这些方法能够有效消除不同测量尺度或单位带来的影响，降低特征值范围差异对模型训练的干扰，提高模型的训练效率与稳定性[34-35]。例如 Demir 和 Şahin 系统评估了多种数据预处理策略的效果后发现数据标准化不仅能显著加快模型的收敛速度，还有效提高了多种机器学习模型的预测精度与泛化能力[34]。

此外，机器学习方法的引入还提高了对多模态数据的处理能力，包括结构化数据（如分

子描述符、理化性质等）、半结构化数据（如分子指纹、分子结构图等）和非结构化数据（如文献文本、报告等）等[11-14,36]。对于结构化与半结构化数据，可采用上述方法进行数据清洗、填补与标准化等构建统一的数据集；对于非结构化数据，可结合自然语言处理技术（NLP）、自编码器（Autoencoder）等方法提取关键信息并将其转化为可识别的数值化特征，再通过数据清洗与预处理方法整合为统一的数据集。

基于上述机器学习技术，本研究团队针对不同类型的数据特征与处理需求，构建了一套系统化的数据预处理系统，可灵活适配不同类型数据的预处理流程，包括缺失值填补、异常值剔除、特征归一化与数据增强等模块，有效解决了多源数据的异构性问题。依托该体系，本研究团队已形成了适用于化学物质毒性预测的高质量应用数据集，具备良好的一致性、完整性与建模适应性。相关方法已被用于新污染物环境风险评估实践，为新污染物风险识别与数据支撑体系建设提供了可推广的技术路径。

表 2 基于机器学习技术的数据清洗与预处理方法及其应用

Table 2 The machine learning technology based data cleaning and preprocessing methods and their applications

类型	方法与算法	主要功能与特点	应用场景与优势
异常值检测与去除	支持向量机（SVM）、孤立森林（Isolation Forest）、LOF（Local Outlier Factor）	检测并去除噪声数据与异常点，提高数据的准确性与一致性。	在高维数据中识别异常样本或误差数据，有效避免模型训练过程中的偏差。
缺失值填补	K 近邻算法（KNN）、随机森林（RF）、自编码器（Autoencoder）、均值或中位数插补	自动填补缺失值，减少数据不完整对模型训练的影响。	在结构化与半结构化数据中填补缺失值，适用于毒性数据的完整性优化。
数据增强与平衡	SMOTE（Synthetic Minority Over-sampling Technique）、Adaboost、Bagging、随机欠采样（Random Undersampling）	改善数据集不平衡问题，提高少数类别样本的预测精度与泛化能力。	尤其适用于类别不平衡的毒性分类任务，提高模型的整体表现。
数据标准化	归一化（MinMax Scaling）、标准化（Z-score Normalization）、数据平滑（Log、Box-Cox 转换）	对不同来源的数据进行格式化处理，使其具有相同的尺度与分布。	提高数据的一致性与可比性，适用于多源数据的整合与分析。

2.2.2 分子特征工程

在化学物质毒性预测的建模过程中，化学物质的分子结构信息需要经过分子特征表征、降维提取与特征选择优化等步骤，才能作为模型的有效输入变量被引入模型[37-38]。按照处理流程，分子特征工程可分为分子描述符计算、特征提取与特征选择。分子描述符是将化学分子的结构或性质转化为数学形式的数值或符号，用以反映其化学反应性、生物活性或环境行为等特征。传统分子描述符通常包括物理化学性质参数、拓扑结构、电子分布特征等，可通过 PaDEL、RDKit、ChemAxon 等方法计算获得。然而，这些分子描述符通常维度较高，还存在大量冗余信息，增加了模型的训练难度与过拟合风险[39-40]。同时，这些方法对于分子结构图、指纹图谱等数据的适应性较差，难以充分挖掘分子结构中的深层次特征，限制了模型对复杂结构-毒性间非线性关系的识别能力。此外，传统描述符选择依赖于专家知识，存在一定的主观性和局限性，容易遗漏潜在重要特征，从而影响模型的预测性能和泛化能力[41]。

机器学习技术的引入显著提升了分子描述符计算、特征提取与选择的效率与准确性（表3）。在分子描述符计算阶段，相较于依赖人工设计的传统分子描述符，近年来图神经网络（GNN）和基于 Transformer 的分子预训练语言模型被逐渐用于从分子结构图或 SMILES 序列中直接学习特征[42-44]。其中，GNN 能够基于分子结构图中节点（原子）与边（化学键）之间的拓扑关系，进行端到端的特征学习，实现对分子性质或毒性效应的直接预测[45-47]。而 Transformer 架构通过对 SMILES（Simplified Molecular Input Line Entry Specification）序列进行语义建模，捕捉分子的全局结构信息，适用于大规模、高通量的毒性预测任务[48-49]。此外，卷积神经网络（CNN）和自编码器等方法也被用于从分子图像、分子指纹等二维、三维或高维数据中自动提取潜在特征，进一步丰富了分子结构表达的维度与层次[50-51]。例如 Matsuzaka 和 Uesawa 提出了基于三维分子结构图像的 DeepSNAP-深度学习方法，能够直接从分子空间结构中自动学习复杂几何特征，无需传统的分子描述符计算与特征提取，有效克服了传统分子描述符在高维度、冗余性及特征表达不足方面的缺陷，显著提升了预测模型在复杂化学结构空间中的适应性与准确性[46]。

在特征选择阶段，机器学习技术能够自动提取分子结构中的深层次特征并自动识别重要特征，无需依赖专家知识[3,9]。目前常利用 SVM 或 RF 等算法进行特征选择，可通过评估特征重要性逐步消除不重要的特征，实现模型性能优化。此外，线性判别分析（LDA）等降维方法也被用于处理高维分子描述符，可有效降低特征维度、缓解噪声干扰[52]。如李昕容等（2023）利用 RF 模型筛选出了影响重庆市 483 个采样点中土壤交换酸含量的最重要特征是

年平均降水、年日照时数、成土母质和年平均温度[53]。Roy 等人利用 LDA 构建了一个化学物质对蚯蚓毒性预测模型，不仅能够较好地预测化学物质对蚯蚓毒性，还通过去除冗余特征后识别出了八个对模型贡献显著的分子描述符[54]。

基于以上方法，本研究团队总结了一套面向结构化化学物质毒性数据建模的分子特征处理体系（图 2）。首先，根据美国化学会化学文摘社（Chemical Abstracts Service, CAS）规定的化学物质登记号（CAS 号）或国际通用名从 Pubchem 等权威数据库中检索并获取目标化学物质的 SMILES 码。SMILES 码是一种用 ASCII 字符串形式表示分子结构的规范格式，具有简洁、唯一、易于机器解析等优点，广泛应用于基于分子结构的建模任务中。对于数据库中未直接提供 SMILES 码的化合物，可借助 RDkit、Open Babel 等开源化学信息学工具通过化学结构式、国际化合物标识（InChI）等形式进行 SMILES 转换与验证。随后，利用 RDKit、PaDEL 等化学信息学计算平台基于 SMILES 表达式计算各类分子描述符和分子指纹，形成分子特征集。最后，通过机器学习技术如 RF、SVM 等方法评估特征重要性，保留对模型影响最为显著的特征，同时去除高相关系数特征（如 Pearson 相关吸收 $r > 0.95$），减少冗余信息。这一流程显著提升了建模过程的标准化程度与自动化水平，确保了数据获取、特征生成与筛选环节的可控性与一致性，具有通用推广潜力，能够为构建化学物质毒性预测模型提供结构清晰、质量可靠的建模输入。

表 3 基于机器学习技术的分子特征处理方法

Table 3 The machine learning technology based molecular feature processing method

类型	机器学习方法	功能
分子描述符计算	图神经网络（GNN）	从分子结构图直接学习节点（原子）与边（键）的特征，提取结构-性质关系信息
	Transformer	基于 SMILES 字符串建模，自动学习分子语言中的上下文语义与结构模式
特征提取	卷积神经网络（CNN）	从分子图像、分子指纹等二维或三维数据中提取空间特征与局部结构信息
	自编码器	对高维分子描述符进行无监督压缩，提取潜在空间中的关键结构特征
特征选择	SVM、RF	利用模型输出中的特征重要性排序评估变量贡献，剔除冗余或弱相关特征
	LASSO 回归	通过稀疏正则化线性建模过程选择少量关键变量，实现自动特征筛选

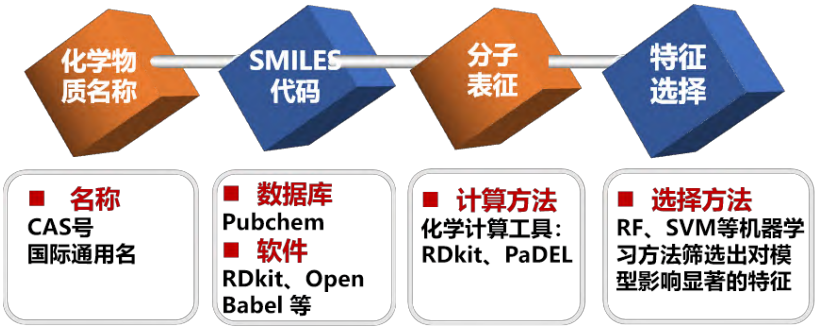

图 2 基于结构化毒性数据建模的分子特征处理体系

Figure 2 The molecular feature processing system based on structured toxicity data modeling

2.2.3 模型训练与验证

完成分子特征的提取与筛选后，就可以对筛选的分子特征与毒性效应数据进行学习，开展模型训练与验证。通过比较 QSAR 模型与机器学习算法[3-4,6,10]（表 4），可以发现 QSAR 模型主要适用于简单线性和非线性关系的化学分子活性预测，化学意义明确，可解释性强。相比之下，机器学习算法具有强大的精度和适用性，适合大数据集和复杂算法，还能多任务模型同时处理多种数据信息，多种线性非线性关系，具有高灵敏度和准确度。但是部分复杂模型的可解释性较差，限制了其在实际的广泛应用[4,36,55]。因此，通过将机器学习算法与 QSAR 建模策略相结合，可显著提高模型的准确性和泛化力。目前常见的算法类别与功能如表 5 所示[56-59]。总体来说，目前 RF、SVM、XGBoost、神经网络模型、GNN 等模型在化学物质毒性预测中应用最为广泛，其在可解释性、预测精度、数据适应规模、抗噪能力和复杂结构的建模能力如图 3 所示。其中，RF 和 SVM 具有较强的模型可解释性，适用于需要明确变量贡献与机制阐释的毒性评价任务，但当噪声和异常数据存在时 RF 的稳定性高于 SVM；XGBoost、DNN 和 GNN 在大数据和复杂非线性任务中表现出更高的预测准确性，而且 DNN 和 GNN 更适合高维和大规模数据集；此外，GNN 还能够直接处理分子图结构，适用于基于化学物质结构的毒性预测建模。因此，根据数据类型、样本规模和预测目标的不同，合理选择和组合上述模型，可显著提升 QSAR 模型的准确性、泛化能力与实用性。

表 4 QSAR 模型和机器学习算法的特点

Table 4 The characteristics of QSAR models and machine learning algorithms

特点	QSAR 模型	机器学习算法
基本概念	基于分子结构与生物活性之间的数学关系进行预测	使用算法从数据中自动学习模式和预测结果
输入数据类型	分子描述符和分子指纹	多种类型数据，包括分子描述特征

模型训练过程	依赖简单线性回归建立描述符与活性之间的方程	大数据，通过训练集和测试集进行模型优化
适用范围	主要用于简单线性或非线性关系的化学分子活性预测	适用于复杂数据模式，如高维、非线性数据
常用算法	多元回归、偏最小二乘法等	随机森林、支持向量机、神经网络等
解释性	具有较强的解释性，可明确说明哪些分子特性影响活性	解释性较弱，需要借助特征重要性分析工具
泛化能力	泛化能力相对较弱，依赖于具体数据集	泛化能力强，适用于多种不同数据集
计算复杂度	计算复杂度较低，适合小规模数据集	较高，尤其是深度学习模型需要大量计算资源
优缺点	优点：可解释性强；缺点：复杂数据或非线性关系下效果差	优点：灵活、准确率高；缺点：可解释性弱
应用领域	化学品毒性预测、药物活性评价、生态毒理学等	药物设计、毒性预测、环境科学等多领域应用

表 5 常见的机器学习算法类别及功能

Table 5 The categories and functions of common machine learning algorithms

算法类别	代表算法	功能与特点
线性模型	线性回归、逻辑回归	构建简洁、可解释性强的模型，适合特征清晰的任务
核方法	支持向量机（SVM）	适用于中小样本、高维空间建模，尤其擅长处理非线性边界问题
树模型	随机森林（RF）、XGBoost	可处理特征间交互，抗过拟合能力强，适合高维复杂数据
神经网络	DNN、CNN、GNN 等	自动捕捉高阶非线性关系，适合大规模训练与端到端建模
集成学习	Bagging、Boosting Stacking	提高模型鲁棒性与泛化能力，常用于提升整体性能

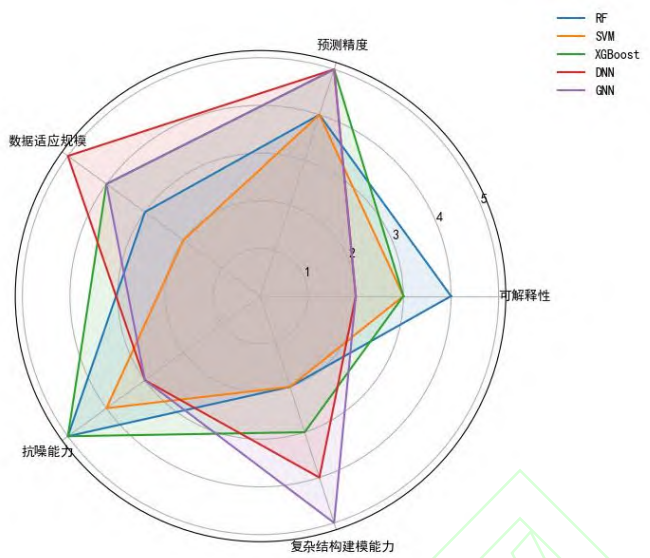

图 3 五种常用机器学习模型在可解释性、预测精度、数据适应规模、抗噪能力和复杂结构的建模能力上的性能雷达图

Figure 3 Radar chart of prediction performance in five common machine learning models on interpretability, predictive accuracy, data scalability, noise robustness, and complex structure modeling capability

在模型训练过程中，首先要将数据划分为训练集与测试集（7：3 或 8：2），还可以使用 K 折交叉验证，确保模型在有限数据条件下的稳定性与泛化能力。随后，可通过网格搜索、贝叶斯优化等方法对超参数（如学习率、树深度、正则化系数）进行优化，进一步提升模型性能[60-62]。在模型验证阶段，需要借助一系列评价指标对模型的拟合优度、预测能力及鲁棒性进行综合评估。对于回归模型，常采用均方误差（MSE）、均方根误差（RMSE）、决定系数（R^2）和平均绝对误差（MAE）等指标（表 6）。其中，MSE、RMSE 和 MAE 常用于衡量模型预测值与真实值之间的误差，一般而言，这些指标值越小，回归模型的拟合优度越好。然而，每个指标都有其局限性。MSE 能够比较灵敏的反映模型的预测误差，但其受异常值的影响较大；RMSE 可以避免较大误差值对拟合度的影响，但也受到异常值的限制；MAE 对异常值相对不敏感，但其不能反映误差的方向，并且对较大的误差值不敏感。R^2 能够直观反映回归模型的拟合效果，其取值范围在 0~1 之间。R^2 越接近 1，模型的拟合度越好。然而，R^2 不能判断回归模型是否发生过拟合，也不能保证模型的泛化性能。因此，在评估回归模型的预测性能时，需要综合考虑以上几个评价指标。对于分类模型，可根据混淆矩阵、准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 分数（F1-score）和受试者工作特征曲线下的面积（AUC）等指标综合判断（表 7-8）。混淆矩阵能够有效地说明单一模型的类别性能，是分类任务中常用的评估手段。通过将模型预测的结果与真实的类

别标签进行比较来展示模型在分类任务上的表现情况，全面反映分类模型的表现[63]。此外，为验证模型在实际应用中的可推广性，还需要引入独立外部测试集，检验模型对外部数据的预测能力，避免因训练过程过拟合而导致性能虚高的现象，确保模型具备良好的现实适应性与稳定性。

表 6 回归模型常用的评估指标

Table 6 The common evaluation indicators used in regression models

性能度量	公式
均方误差（MSE）	$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
均方根误差（RMSE）	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$
平均绝对误差（MAE）	$\frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$
决定系数（R2）	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$

y_i是真实值，$\bar{y}_i$是真实值的平均值，$\hat{y}_i$是预测值。

表 7 混淆矩阵（以二分类为例）

Table 7 Confusion matrix (binary classification)

真实值	预测值	
	阳性	阴性
阳性	TP（真阳）	FN（假阴）
阴性	FP（假阳）	TN（真阴）

表 8 分类模型常用的评估指标

Table 8 The common evaluation indicators used in classification models

性能度量	含义	公式
准确率（Acc）	预测正确的样本数占总样本数的比例	$\frac{TP+TN}{TP+FP+FN+TN}$
查准率（P）	预测为阳性的样本中，真正为阳性的样本所占比例。	$\frac{TP}{TP+FP}$
查全率（R）	所有阳性样本中被正确预测为阳性的比例。	$\frac{TP}{TP+FN}$
AUPR（P-R 曲线下面积）	描述查准率和查全率随着阈值变化的曲线。	——
F1 度量（F1 Score）	平衡查准率和查全率	$\frac{2\times P\times R}{P+R}$

灵敏度（TPR）	正确识别的阳性样本占实际所有阳性样本的比例	$\frac{TP}{TP+FN}$
特异性（TNR）	正确识别的阴性样本占实际所有阴性样本的比例	$\frac{TN}{TN+FP}$
AUROC（ROC 曲线下面积）	综合考虑灵敏度和特异性	——
马修斯相关系数	考虑了混淆矩阵中的所有四个类别的计数，适用于存在类别不平衡问题的数据集	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

利用以上机器学习算法和模型训练与验证步骤，本研究团队基于农药对 7 种环境敏感生物的毒性数据开发了 2 项回归模型，7 项分类模型。以家蚕急性毒性回归预测模型为例[64]，本研究团队首先将毒性数据按照 7：3 划分为 70%训练集与 30%测试集，并利用了 20 种经典的机器学习算法开展模型训练，包括 11 种线性回归模型和 9 种非线性回归模型。随后使用 4 折和 10 折进行交叉验证和模型超参数调优。为了更进一步提高模型的鲁棒性，减少模型的错误率，进一步构建了综合投票模型。结果显示综合投票模型的预测值与实测值相关性（$R^2=0.78$）显著高于传统的 QSAR 模型（$R^2<0.1$），且模型在独立测试集上的预测性能良好。本模型展示了集成学习在提升毒性预测准确性和稳定性方面的优势，为新污染物毒性建模提供了可借鉴的技术路径。

2.2.4 模型适用域与可解释性

模型的适用域与可解释性分析是保障模型科学性与实际应用价值的重要环节，共同决定了模型预测结果的可靠性、透明性及其在化学物质毒性识别与风险评估中的应用范围[65-67]。

目前，模型的适用域主要从描述符变化范围、结构相似性、机理相似性和代谢转化途径和产物等多个维度进行界定[68]。在实践中，较常采用的方法包括基于描述符分布范围、结构相似性以及基于距离的算法等[69]。以本研究团队在农药对家蚕的毒性建模为例，构建了基于 SMILES 结构相似性和关键描述符（LogP、分子量）分布范围的适用域评估方法[64]。此外，本研究团队在农药对蚯蚓的毒性建模中利用基于平均欧氏距离的算法判断新样本是否包含在适用性域内以表明模型预测结果的可靠性。模型适用域界定对于确保模型预测的合理性与可靠性具有重要意义，尤其适用于结构多样性高、机制尚不明的化学物质预测任务。

模型的可解释性是指对模型决策过程和预测依据的理解程度，尤其在化学物质环境风险评估中，高可解释性对于提升结果的接受度与决策透明度具有重要意义[66]。为了增强模型的可解释性，目前常使用局部可解释模型（如 LIME）、SHAP 值（SHapley Additive exPlanations）

分析等特征重要性表征工具[65]。这些方法能够揭示不同分子特征对预测结果的具体贡献，识别潜在的关键结构片段或理化性质特征，从而为模型输出提供解释[70]。如本研究团队利用SHAP 分析解析了影响商业农药在土壤环境中对蚯蚓毒性的关键因素，不仅能够量化各特征对预测毒性贡献的大小，还能揭示各特征对个体样本预测结果的正负影响方向。因此，将解释性算法集成于机器学习建模流程中，有助于提升模型透明性、增强结果的可接受性，并推动其在实际环境管理中的规范应用。基于以上机器学习技术，可大幅提高毒性预测模型构建的自动化水平、准确性和适用性。

3 国内在人工智能技术在毒性预测领域的主要研究成果

近年来，国内也有多个研究团队在人工智能毒性建模方面取得进展。例如，大连理工大学的陈景文教授团队长期从事生态毒理与健康毒理预测工作，通过汇集持久性、生物蓄积性、迁移性有毒(PBMT)化学品数据集，构建了基于深度学习模型的 PBMT 多物种、多终点集成筛查模型[71]；华东理工大学的唐赟教授团队采用人工智能技术和多模态分子表征方法，构建了免费的 ADMET 在线毒性预测和分子设计优化平台 admetSAR（https://lmmd.ecust.edu.cn/admetsar3/），可提供 119 个端点的属性预测[72]；浙江大学的庄树林教授也基于机器学习方法构建出了 PBMT 的筛选模型以及针对内分泌干扰效应的一系列预测模型[73]。这些研究在多模态数据融合、算法和模型创新以及模型适用性等方面取得了显著的进展。此外，国内“化学物质预测模型工具学组”的成立进一步推动我国化学物质预测模型工具的研究与应用。然而，值得注意的是，这些建模的数据资源多来自国外公开的数据库。一方面对我国毒性预测模型的数据安全与自主性构成潜在威胁，另一方面，这些数据库的毒性测试体系和评价标准与我国可能存在差异，在物种选择、实验设计、毒性终点等方面与中国典型生态环境和暴露场景不完全匹配，限制了模型在本土化情境下的适用性与推广性。

4 总结与展望

尽管大数据、机器学习等人工智能（AI）技术的蓬勃发展，为化学物质毒性预测提供了强有力工具[3,10]。然而，当前基于人工智能技术的化学物质毒性预测方面仍面临诸多挑战。首先，高质量、标准化的毒性数据依然有限，数据稀缺与异质性问题在一定程度上制约了模型性能的进一步提升，本土化的数据缺失限制了模型在我国本土化情境下的推广[74-75]。其次，目前机器学习和深度学习技术在多模态数据融合、复杂毒性预测等方面的垂直渗透率仍较

低。现有的模型仍以结构化数据为主，缺乏对多模态特征信息的有效整合，且对于多终点毒性、长期低剂量暴露、组合污染效应等复杂毒性终点的预测能力仍较弱，难以准确刻画化学物质在真实、复杂环境情境下的生态风险。同时，虽然机器学习模型具备强大的预测能力，但其“黑箱”性质仍然会影响结果的可解释性与应用可靠性[66,70]。更为重要的是，当前基于人工智能技术辅助的预测结果在实际应用中尚缺乏统一的认可标准，无论是在科学界还是在政策制定与风险管理实践中，对于人工智能技术辅助预测结果的可信度、可接受性仍存在较大争议。

因此，未来建议从以下四方面突破：

（一） 数据标准化与本土化建设。随着大数据平台建设与实验数据积累的加速，应进一步推动化学物质数据库的动态更新与标准化建设，加强我国本土毒性数据的采集与标准化整合工作，建立统一的化学物质数据筛选、评价和清洗技术指南，提升数据质量与一致性；

（二） 多模态数据融合建模。在建模技术层面，利用多模态数据融合、图神经网络、卷积图神经网络等方法，实现分子结构、理化性质、毒性效应的协同学习，充分挖掘多维度特征之间的复杂关联，提高模型的预测精度[3,10]；

（三） 建模算法与任务创新。引入多任务学习（Multi-task Learning）、迁移学习（Transfer Learning）等方法，提升模型对多终点、多过程、多物种复杂毒性效应的预测能力；

（四） 模型可解释性与透明度。在模型预测结果评价中，引入如 SHAP、LIME 等模型可解释性分析技术，提升模型决策透明度[70]。

总体而言，随着大数据、模型与应用场景的持续迭代，人工智能技术辅助的化学物质毒性预测将在未来环境管理领域发挥更加重要的作用。通过构建高质量数据基础、深入挖掘多源、多模态数据，发展可迁移与可解释的预测模型，未来有望实现高效、精准、可信的毒性预测与风险预警，为新污染物环境管理提供更强有力的科技支撑与决策依据。

附表:

附表 1 文中出现的术语及定义

Appendix Table 1 Terms and Definitions in the text

术语	定义
NoSQL 数据库	Not Only SQL，一类非关系型数据库，适用于大规模、非结构化或半结构化数据的高效存储与查询
Cassandra	一种高可扩展性、高可用性的分布式 NoSQL 数据库
MongoDB	文档型 NoSQL 数据库，支持灵活存储半结构化与非结构化数据
SMOTE 算法	Synthetic Minority Over-sampling Technique，一种用来平衡样本类别分布的方法
Adaboost	通过迭代方式重点训练前一轮错误分类的样本，提升模型分类能力
Bagging	通过对训练集进行有放回采样，构建多个子模型并投票/平均结果，降低模型方差
随机欠采样技术	一种通过随机减少多数类样本数量来平衡类别分布的数据处理方法
多模态数据	指来源不同、类型各异的数据，如结构化表格、图像、文本等
自然语言处理技术	NLP，一种用于处理和理解人类语言文本的人工智能技术
自编码器	Autoencoder，一种可从数据中自动提取主要特征的神经网络结构
图神经网络	GNN，一种可直接处理图结构数据的深度学习模型
Transformer	一种用于序列建模的深度学习结构
SMILES	Simplified Molecular Input Line Entry Specification，一种用 ASCII 字符串表示分子结构的规范格式
LIME	Local Interpretable Model-agnostic Explanations，一种用于解释单个预测结果的局部线性模型
SHAP 值	SHapley Additive exPlanations，一种基于博弈论的特征重要性解释方法
迁移学习	Transfer Learning，指在一个任务中获得的知识被迁移到另一个相关任务中，以提升新任务建模效率与准确性。

致谢：

国家重点研发计划课题（National Key Research and Development Program of China）：场地新污染物环境风险指标体系构建（Construction of Index System on Environmental Risk Assessment for New Pollutants in the Industrial Site, 2023YFC3706603)；国家自然科学基金长江水科学研究联合基金项目（Joint Fund Project of Yangtze River Water Science Research, National Natural Science Foundation of China）：长江流域生态环境风险溯源与预警研究（Study on the identification and early warning of ecological and environmental risks of typical industrial parks in the Yangtze River Basin，U2340202）；生态环境部预算项目（Department Budget Project of the Ministry of Ecology and Environment）：化学品与重金属污染防治监督管理——有毒有害化学物质慢性毒性预测与毒性识别研究（Project of Supervision and Management of Chemical and Heavy Metal Pollution Prevention and Control——Research on Chronic Toxicity Prediction and Toxicity Identification of Toxic and Harmful Chemicals）；生态环境部预算项目（Department

Budget Project of the Ministry of Ecology and Environment）：农村和农业环境保护管理（Supervision and Management of Rural and Agriculture Environmental Protection）。

参考文献：

[1] VALDUGA A T, GONÇALVES I L, SAORIN PUTON B M, et al. Anthraquinone as emerging contaminant: technological, toxicological, regulatory and analytical aspects[J]. Toxicological Research, 2024, 40:11-21.

[2] SCHYMANSKI E L, ZHANG J, THIESSEN P A, et al. Per- and polyfluoroalkyl substances (PFAS) in PubChem: 7 million and growing[J]. Environmental Science & Technology, 2023, 57:16918-16928.

[3] 王中钰, 陈景文, 乔显亮, 等. 面向化学品风险评价的计算(预测)毒理学[J]. 中国科学:化学, 2016, 46:222-240.

[4] KREWSKI D, ANDERSEN M E, TYSZENKO M G, et al. Toxicity testing in the 21st century: progress in the past decade and future perspectives[J]. Archives of Toxicology, 2020, 94:1-58.

[5] WANG Y, GAO X, CHENG Y, et al. Nano-TiO2 modifies heavy metal bioaccumulation in *Daphnia magna*: a model study[J]. Chemosphere, 2023, 312:137263.

[6] 李欢, 刘志永, 李存治, 等. 定量构效关系方法在毒性预测领域的应用[J]. 环境与健康杂志, 2024, 41:89-92.

[7] GAJEWICZ-SKRETNA A, KAR S, PIOTROWSKA M, et al. The kernel-weighted local polynomial regression (KwLPR) approach: an efficient, novel tool for development of QSAR/QSAAR toxicity extrapolation models[J]. Journal of Cheminformatics, 2021, 13.

[8] JORDAN M I, MITCHELL T M. Machine learning: trends, perspectives, and prospects science[M]. 2015:255-260.

[9] 郑雨, 李呈, 胡贵平, 等. 机器学习技术在环境健康领域中的应用进展[J]. 广西医科大学学报, 2024, 41:1558-1564.

[10] AGNIESZKA G, SCHAEUBLIN N, RASULEV B, et al. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: hints from nano-QSAR studies[J]. Nanotoxicology, 2015, 9:313-325.

[11] ZHOU H J, SHEN T T, LIU X L, et al. Survey of knowledge graph approaches and applications[J]. Journal on Artificial Intelligence, 2020, 2(2):89-101.

[12] LI X C, ZHANG J Q, FAN L A, et al. Construction and analysis of knowledge graphs for multi-source heterogeneous data of soil pollution[J]. Soil Use and Management, 2023, 39(3):1036-1039.

[13] DAGHIGHI A, CASANOLA-MARTIN G M, IDUOKU K, et al. Multi-endpoint acute toxicity assessment of organic compounds using large-scale machine learning modeling[J]. Environmental Science & Technology, 2024, 58(23):10116-10127.

[14] YAP X H, RAYMER M. Toxicity prediction using locality-sensitive deep learner[J]. Computational Toxicology, 2021, 21, 100210.

[15] NAKATSUGAWA M, CHENG Z, KIESS A P, et al. The needs and benefits of continuous model updates on the accuracy of RT-induced toxicity prediction models within a learning health system[J]. International Journal of Radiation Oncology, Biology, Physics, 2019, 103:460-467.

[16] LUNGHINI F, MARCOU G, AZAM P, et al. Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context[J]. SAR and QSAR in Environmental Research, 2019, 30:879-897.

[17] BOWER D, CROSS K, MYATT G. Chapter 4: organisation of toxicological data in databases[J]. Issues in Toxicology, 2019:108-165.

[18] HAN X, ZHENG L. Design and implementation of firmware data acquisition system based on Scrapy framework[C]. 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2020:168-174.

[19] NAIR D B, A R P, RAJ A. Web data mining and crawling: a detailed overview in the aspect of Googlebot, Apache Nutch and Bingbot[J]. Interantional Journal of Scientific Research in Engineering and Management, 2024, 274888962.

[20] SAADA O, DABA J. Automatic SQL to HQL-NoSQL querying using PostgreSQL and integrated Hive-Hbase[J]. WSEAS Transactions on Information Science and Applications, 2023, 20:16-27.

[21] BAO C, CAO M. Query optimization of massive social network data based on HBase[C]. 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), 2019:94-97.

[22] 冯建周, 宋沙沙, 孔令富. 物联网语义关联和决策方法的研究[J]. 自动化学报, 2016, 42(11):1691-1701.

[23] 赵又霖, 庞烁, 吴宗大. 社会感知数据驱动下突发事件应急管理的时空语义模型构建研究[J]. 情报科学, 2021, 39(2):44-53.

[24] BELFIELD S J, CRONIN M T D, ENOCH S J, et al. Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs)[J]. PLOS ONE, 2023, 18(5):e0282924.

[25] LIU C, ZONG C, CHEN S, et al. Machine learning-driven QSAR models for predicting the cytotoxicity of five common microplastics[J]. Toxicology, 2024:153918.

[26] MARAN U, SILD S. QSAR modeling of mutagenicity on non-congeneric sets of organic compounds[J]. Artificial Intelligence Review, 1970, 13:213-222.

[27] HELAL M, AL-REYASHI A. Exploring the effectiveness of different data cleaning techniques for improving data quality in machine learning[J]. Humanitarian and Natural Sciences Journal, 2023, 259520172.

[28] ACHMAD R M, YUHANA U L. Software defect prediction using outlier detection algorithm[C]. 2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT), 2024:204-209.

[29] JOHN M A J, BARHUMI I. Handling missing data in limited-view photoacoustic tomography using compressive sensing algorithm-based deep learning[C]. 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2024:1-6.

[30] SANTIAGO-GONZALEZ F, MARTÍNEZ-RODRÍGUEZ J L, GARCÍA-PEREZ C, et al. Hybrid class balancing approach for chemical compound toxicity prediction[J]. Current Computer-Aided Drug Design, 2024, 39318212.

[31] BASHA S J, MADALA S R, VIVEK K, et al. A review on imbalanced data classification techniques[C]. 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), 2022:1-6.

[32] GAO Q, JIN X, XIA E, et al. Identification of orphan genes in unbalanced datasets based on ensemble learning[J]. Frontiers in Genetics, 2020, 11:820.

[33] KUMARI C, ABULAISH M, SUBBARAO N. Using SMOTE to deal with class-imbalance problem in bioactivity data to predict mTOR inhibitors[J]. SN Computer Science, 2020, 1:1-7.

[34] DEMIR S, ŞAHIN E K. The effectiveness of data pre-processing methods on the performance of machine learning techniques using RF, SVR, Cubist and SGB: a study on undrained shear strength prediction[J]. Stochastic Environmental Research and Risk Assessment, 2024, 38(8):3273-3290.

[35] GRIFFITH M, WALKER J R, SPIES N C, et al. Informatics for RNA sequencing: a web resource for analysis on the cloud[J]. PLOS Computational Biology, 2015, 11(8):e1004393.

[36] 张涛, 朱明华, 傅志强, 等. 筛查全/多氟烷基化合物(PFASs)生物活性的卷积神经网络模型[J]. 生态毒理学报, 2023, 18(3):11-21.

[37] MANNHOLD R, KUBINYI H, FOLKERS G. Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references[M]. 2010.

[38] IDAKWO G, LUTTRELL J, CHEN M, et al. A review of feature reduction methods for QSAR-based toxicity prediction[M]. Challenges and Advances in Computational Chemistry and Physics, 2019, 119-139.

[39] STEPISNIK T, BLAŽ ŠKRLJ, WICKER J S, et al. A comprehensive comparison of molecular feature representations for use in predictive modeling[J]. Computers in Biology and Medicine, 2021, 130:104197.

[40] TANGADPALLIWAR S R, VISHWAKARMA S, NIMBALKAR R D, et al. ChemSuite: a package for chemoinformatics calculations and machine learning[J]. Chemical Biology & Drug Design, 2019, 93:960-964.

[41] MATSUZAKA Y, UESAWA Y. Ensemble learning, deep learning-based and molecular descriptor-based quantitative structure-activity relationships[J]. Molecules, 2023, 28(5):2410.

[42] 张若驰. 基于图神经网络的分子表征及性质预测算法的研究与应用[M]. 吉林: 吉林大学, 2024.

[43] RAJAN K, ZIELESNY A, STEINBECK C. DECIMER 1.0: deep learning for chemical image recognition using transformers[J]. Journal of Cheminformatics, 2021, 13:1-16.

[44] ZHOU J, CUI G, HU S, et al. Graph neural networks: a review of methods and applications[J]. AI Open, 2020, 1:57-81.

[45] CHEN Y, LEUNG C T, HUANG Y, et al. MolNexTR: a generalized deep learning model for molecular image recognition[J]. Journal of Cheminformatics, 2024, 16, 141.

[46] YOO S, KWON O, LEE H. Image-to-graph transformers for chemical structure recognition[C]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, 3393-3397.

[47] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. ArXiv, 2016.

[48] STAKER J, MARSHALL K, ABEL R, et al. Molecular structure extraction from documents using deep learning[J]. Journal of Chemical Information and Modeling, 2019, 59(3):1017-1029.

[49] XU Z, LI J, YANG Z, et al. SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer[J]. Journal of Cheminformatics, 2022, 14(1):41.

[50] LI Z, LIU F, YANG W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(12):6999-7019.

[51] LIM J, RYU S, KIM J W, et al. Molecular generative model based on conditional variational autoencoder for de novo molecular design[J]. Journal of Cheminformatics, 2018, 10(1):1-9.

[52] ZHOU W, WU S, DAI Z, et al. Nonlinear QSAR models with high-dimensional descriptor selection and SVR improve toxicity prediction and evaluation of phenols on Photobacterium phosphoreum[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 145:30-38.

[53] 李昕容, 杨超, 张鑫, 等. 基于随机森林模型与 SHAP 算法的渝东北烟区土壤交换酸含量影响因素分析研究[J]. 中国烟草学报, 2023, 30(2):52-60.

[54] ROY J, KUMAR OJHA P, CARNESECCHI E, et al. First report on a classification-based QSAR model for chemical toxicity to earthworm[J]. Journal of Hazardous Materials, 2020, 386:121660.

[55] BARROS R P C, SOUSA N F de, SCOTTI L, et al. Use of machine learning and classical QSAR methods in computational ecotoxicology[C]. 2020, 151-175.

[56] KAR S, PATHAKOTI K, TCHOUNWOU P B, et al. Evaluating the cytotoxicity of a large pool of metal oxide nanoparticles to *Escherichia coli*: mechanistic understanding through in vitro and in silico studies[J]. Chemosphere, 2021, 264(1):128428.

[57] REICHSTEIN M, CAMPS-VALLS G, STEVENS B, et al. Deep learning and process understanding for data-driven Earth system science[J]. Nature, 2019, 566:195-204.

[58] MCBRATNEY A, DE GRUIJTER J, BRYCE A. Pedometrics timeline[J]. Geoderma, 2019, 338:568-575.

[59] 仇皓雷, 王海燕. 机器学习在土壤性质预测研究中的应用进展[J]. 生态学杂志, 2023, 1:283-294.

[60] BELETE D M, HUCHAIAH M D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results[J]. International Journal of Computers and Applications, 2021, 44:875-886.

[61] LUO G. A review of automatic selection methods for machine learning algorithms and hyper-parameter values[J]. Network Modeling Analysis in Health Informatics and Bioinformatics, 2016, 5(1):18.

[62] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization[J]. Journal of Machine Learning Research, 2012, 13:281-305.

[63] POMMÉ L E, BOURQUI R, GIOT R, et al. Relative confusion matrix: an efficient visualization for the comparison of classification models[M]//KOVALERCHUK B, NAZEMI K, ANDONIE R, et al. Artificial Intelligence and Visualization: Advancing Visual Knowledge Discovery. Cham: Springer Nature Switzerland, 2024:223-243.

[64] LIU Y T, YU Y, WU B, et al. A comprehensive prediction system for silkworm acute toxicity assessment of environmental and in-silico pesticides[J]. Ecotoxicology and Environmental Safety, 2024, 282:116759.

[65] GALLEGOS M, VASSILEV-GALINDO V, POLTAVSKY I, et al. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors[J]. Nature Communications, 2024, 15(1):4345.

[66] 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.

[67] ROY K, KAR S, AMBURE P. On a simple approach for determining applicability domain of QSAR models[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 145:22-29.

[68] PÉREZ-SANTÍN E, DE-LA-FUENTE-VALENTÍN L, GARCÍA M G, et al. Applicability domains of neural networks for toxicity prediction[J]. AIMS Mathematics,2023, 8(11): 27858-27900.

[69] MORA J R, MARQUEZ E A, PÉREZ-PÉREZ N, et al. Rethinking the applicability domain analysis in QSAR models[J]. Journal of Computer-Aided Molecular Design, 2024, 38(1):9.

[70] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature Machine Intelligence, 2019, 1(5):206-215.

[71] WANG H B, LIU W J, CHEN J W, et al. Transfer Learning with a Graph Attention Network and Weighted Loss Function for Screening of Persistent, Bioaccumulative, Mobile, and Toxic Chemicals[J]. Environmental Science & Technology, 2024, 59(1): 578-590.

[72] Gu Y X, Yu Z H, Wang Y M, et al. admetSAR3.0: a comprehensive platform for exploration, prediction and optimization of chemical ADMET properties[J]. Nucleic Acids Research, 2024, 52(W1): W432-W438.

[73] Zhao Q M, ZHENG Y T, QIU Y, et al. Graph Convolutional Network-Enhanced Model for Screening Persistent, Mobile, and Toxic and Very Persistent and Very Mobile Substances[J]. Environmental Science & Technology, 2024, 58(14): 6149-6157.

[74] STEINMETZ F P, ENOCH S J, MADDEN J C, et al. Methods for assigning confidence to toxicity data with multiple values-identifying experimental outliers[J]. Science of The Total Environment, 2014, 482:358-365.

[75] OECD. Description of selected key generic terms used in chemical hazard/risk assessment[R]. Paris: OECD, 2003.

作者简介：李思敏（1994 年-），女，河南省洛阳市人，助理研究员，研究方向为固废与新污染物智能化防控。E-mail：lisimin@nies.org