Eco-Environmental Knowledge Web

环 境 化 学
ENVIRONMENTAL CHEMISTRY

第 43 卷第 1 期 2024 年 1 月
Vol. 43, No. 1 January 2024

DOI:10.7524/j.issn.0254-6108.2023020306

王紫维, 韩民, 金彪. 机器学习在化合物属性预测中的应用[J]. 环境化学, 2024, 43(1): 69-81.

WANG Ziwei, HAN Min, JIN Biao. Applications of machine learning on compound property prediction[J]. Environmental Chemistry, 2024, 43 (1): 69-81.

机器学习在化合物属性预测中的应用*

王紫维[1,2,3] 韩 民[1,2,3] 金 彪[1,2,3] **

（1. 中国科学院广州地球化学研究所，有机地球化学国家重点实验室，广州，510640；2. 中国科学院深地科学卓越创新中心，广州，510640；3. 中国科学院大学，北京，100049）

摘 要 化合物的属性预测是药物研发、毒理学研究、环境行为预测等工作的核心任务. 目前，人工合成的化学物质层出不穷，相关的实验研究数据在持续扩充，但实验研究数据远无法赶超新型化学物质的研发速度. 近年来，机器学习算法及模型在化合物属性预测方面展现了独特的优势和巨大的潜力，尤其在实验数据匮乏的情况下，提供了可靠的模型预测数据. 本文介绍了机器学习应用于化合物属性预测的主要流程步骤和相应的模块的内容，涵盖数据集、分子描述方法、模型性能评估指标和评估方法等. 同时，本文系统总结了机器学习方法在化合物物理化学性质预测、生物活性预测和毒性预测方面的应用实例，并从数据集、分子特征化、模型解释等方面分析并讨论了相关研究工作现存问题与未来挑战.

关键词 机器学习，化合物属性，分子结构，模型预测.

Applications of machine learning on compound property prediction

WANG Ziwei[1,2,3] *HAN Min*[1,2,3] *JIN Biao*[1,2,3] **

（1. State Key Laboratory of Organic Geochemistry, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou, 510640, China；2. CAS Center for Excellence in Deep Earth Science, Guangzhou, 510640, China；3. University of Chinese Academy of Sciences, Beijing, 100049, China）

Abstract Compounds property prediction is an essential task in drug development, toxicology, and environmental behavior prediction. Along with an increasing number of synthetic chemicals, the corresponding experimental research data are expanding. However, the experimental data are still far away from rapid invention of novel chemicals. In recent years, machine learning algorithms and models have shown advantages and great potential in compound property prediction, especially in case of lacking experimental data, providing reliable model-predicted data. Our study outlines the main procedures and corresponding modules related to applications of machine learning tools for compound property prediction, specifically including datasets, molecular description methods, model performance evaluation metrics, and methods. Furthermore, this work systematically summarizes progress and advances in compound property prediction based on machine learning approaches, and also introduces specific examples on compounds predictions of physical and chemical properties, bioactivity, and toxicity. To end, the existing problems and challenges are discussed based on data sets, molecular characterization, and model outcome interpretation.

Keywords machine learning，compound property，molecular structure，model prediction.

2023 年 2 月 3 日收稿(Received：February 3，2023).

* 国家重点研发计划重点专项（2019YFC1805500，2019YFC1805503）资助.

Supported by National Key Research and Development Plan （2019YFC1805500，2019YFC1805503）.

** **通信联系人 Corresponding author**，E-mail: jinbiao@gig.ac.cn; Tel: +86-20-83274209

化合物的属性预测在药物研发、材料设计、毒理学研究等领域发挥了重要的作用，与人类生活息息相关[1－2]. 化合物属性预测的相关研究可追溯到药物合成的早期研究，当时主要是化学家通过重复实验，进行测试和验证并获取各类化学信息，合成目标分子[3]. 由于重复实验耗时长、成本高，科学家基于构效关系(SAR)发展出了定量-构效关系，为化合物结构与其性质之间建立了数学关系框架. 1962 年，Hansch 等首次实践了定量-构效关系(QSAR)，成为该领域具有里程碑意义的事件，也是化合物属性预测研究迈入新阶段的标志[4]. 随后，Hansch 在 1964 年提出了 Hansch 方程，这个发现为 QSAR 模型运行提供了一种新方法. 但传统 QSAR 模型一般使用一些常见的分子描述符来预测化合物属性，然而化合物结构多样，少量的分子描述符很难全面地描述化合物的结构信息，这使得模型很难精准预测化合物性质. 同时，随着研究数据集增大、描述符增多，传统的方法难以拟合化学结构与性质之间的复杂关系. 因此，需要比传统统计工具更先进、更强大的计算和数据分析方法.

机器学习(特别是深度学习)，由于其强大的计算和数据分析能力，已被用于解决以上 QSAR 研究中的问题. 例如，研究人员通过机器学习或深度学习方法将三维甚至更高维分子结构与其属性联系起来，弥补了传统的化合物属性预测方法的不足之处，大力推动了化合物属性研究的发展[5－6].

近年来，机器学习在化合物属性的预测研究上表现出不俗的潜力，因此这方面的研究也逐年增多. 比如在理化性质方面，在机器学习的帮助下，预测分子的原子化能、振动频率、溶剂化自由能、计算键离能等，成本更低，结果准确可靠，计算速度更快[7－11]；在生物活性方面，建模方面逐步引入了神经网络算法、分子图等，所构建模型性能更优异，结果可靠[12－14]；在毒性方面，根据机器学习建立的模型可以非常有效地识别有毒分子和预测特定毒性，可筛选确认之前未曾识别出的危险化学品[15－17]. 本文主要介绍机器学习在化合物属性预测方面的应用过程及相应的模块内容，并结合应用实例总结和展望机器学习在该应用方面现存的问题和机遇.

1 机器学习预测化合物属性的流程(The process of machine learning on compound property prediction)

在实际应用中，用机器学习预测化合物属性的整体过程如下所述，见图 1.

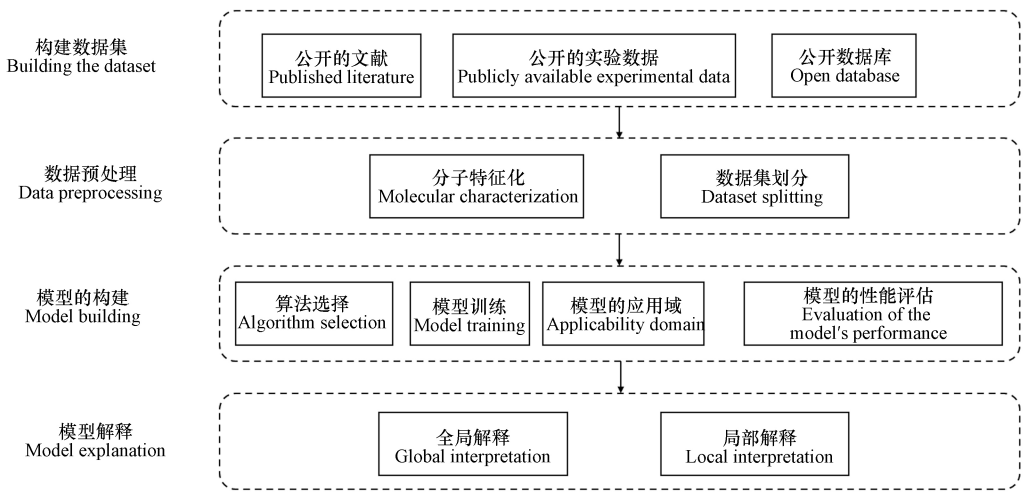

图 1 机器学习进行化合物属性预测的流程

Fig.1 Process for compound property prediction based on machine learning

(1)构建数据集：数据集一般来源于公开的数据库、实验记录数据、研究收集的文献；

(2)数据预处理：主要包括分子特征化和数据集划分；

(3)模型构建：主要包括模型训练、算法选择、模型的应用域、模型评估这四方面内容；

(4)模型解释：解释机器学习模型的预测机制.

1.1　构建数据集

构建数据集是构建模型进行化合物属性预测之前的必要准备, 模型的工作主要基于数据运行, 数据集的质量影响了模型预测结果的可靠性以及准确性. 目前众多研究工作一般通过以下几种途径进行数据集的构建: (1)根据实验所得数据构建数据集; (2)从公开的数据库中下载研究所需要的数据, 并整理成数据集; (3)提取他人发表在期刊、专利中的数据, 整理成数据集; (4)前述 3 种方式的组合形式.

实验室所记录的数据比较全面, 是构建数据集的来源之一. 实验所得数据一般分为纸质记录数据和电子记录数据, 其中, 电子记录数据可用于数据集构建和数据挖掘, 但由于实验数据为实验者所有, 大多用于进行数据存档和知识产权维权, 难以获得全面的数据来进行数据挖掘. 同时, 他人发表在期刊、专利中的数据相对于实验所记录的数据大多数都经过文章作者精心筛选, 没有展示相关实验失败的数据. 失败数据的缺乏可能会造成信息缺失, 从而导致构建的模型不能进行准确地预测.

构建数据集最常用的方法是从公开的数据库获取研究要用的数据. 经过多年的发展, 目前也有许多可免费获取化学数据的公开数据库, 部分较常见的公开数据库可见表 1.

表 1　常见的公开数据库

Table 1　Common public databases

数据库 Database	简要介绍 Brief introduction	参考文献 References
PubChem	有机小分子生物活性数据库, 包含3个子数据库PubChem BioAssay , PubChem Compound, PubChem Substance.	[18]
Chemspider	小分子信息整合数据库, 包含了分子简介、实验测定和实时估算的理化性质、毒性、Smile字符串等信息.	[19]
GDB-13	有机小分子数据库, 含有多达13个C、N、O、S和Cl原子的小有机分子.	[20]
GDB-17	有机小分子数据库, 含有1664亿个分子, 组成中多达17个C、N、O、S和卤素等原子.	[21]
FreeSolv	提供实验和计算的水合自由能的数据库.	[22]
ZINC	可以进行虚拟筛选的数据库, 汇总了化合物的购买信息和来自其他数据库的注释化合物.	[23 − 25]
ChEMBL	一个大型的生物活性数据库, 数据主要来源于文献提取和PCBA数据库.	[26]
DrugBank	综合性药物数据库, 包含有关药物的机制、相互作用和靶标的全面分子信息.	[27]
ToxCast	来自EPA开展的一个利用高通量筛选方法和计算毒理学方法预测化合物毒性并进行优先级排序的研究项目, 拥有约1800种化学品的数据.	[28]
PDBbind	提供生物分子复合物的结合亲和力数据.	[29]
BindingDB	收集了生物活性数据注释的小分子.	[30]

1.2　数据预处理

1.2.1　分子特征化

分子特征化是把化合物的化学结构编码成机器学习算法能识别的模式. 不同的分子特征化方式提取的分子信息有所差异, 直接影响模型的预测效果, 因此是化合物属性预测的重要部分. 常见的分子特征化方法有分子描述符、分子图、分子线性表示、分子图像[31], 以及三维分子表面点云[32].

(1)分子描述符

分子描述符与分子结构的关系密切, 可以有效地表示相应的化学信息[33]. 分子描述符按照复杂程度, 可分为零维、一维、二维、三维等(见图 2)[34]. 零维描述符是最简单的分子描述符, 其信息含量低, 可表示原子数、原子性质总和、分子量等; 一维描述符表示一些官能团、分子片段、取代基等信息, 如分子量、摩尔折射率、辛醇/水分配系数的对数等; 二维描述符可描述从二维分子表示计算得到的性质; 三维描述符信息含量很高, 可描述原子的性质、连通性以及分子的空间构型, 可用于确定化合物的活性构象等问题; 四维描述符可以定量识别和描述分子与受体活性位点之间的相互作用[34].

分子描述符按照定量和定性分类, 可分为定量分子描述符和定性分子描述符. 定量分子描述符有分子场描述符、分子形状描述符、物理化学描述符、基于组成信息的描述符等[35]. 定性分子描述符一般指分子指纹, 分子指纹又称二元指纹, 采用二进制编码相关的化学信息, 指纹所具有的化学信息内容一般为化学图中的原子、键类型和距离等, 是化学结构的表示, 常被用于分子相似性/多样性问题[34, 36].

常见的分子指纹可主要分为基于子结构的指纹、基于拓扑或路径的指纹和圆形指纹、药效团指纹

等[37]. 基于子结构的指纹主要有 MACCS 指纹[38]、PubChem 指纹、BCI 指纹、TGD 和 TGT 指纹等. 基于拓扑或路径的指纹主要有 Daylight 指纹(Daylight fingerprint)和 Tree 指纹(Tree fingerprint). 圆形指纹主要有扩展连通性指纹(ECFP/Morgan Fingerprint)[39]、FCFP(Functional-Class Fingerprints)、Molprint2D[40].

常用于计算分子指纹的软件或工具包有 alvaDesc[41]、RDkit、Open Babel[42]、CDK[43]、ChemFP、OEChem TK、Molecular Operating Environment (MOE)、JChem from ChemAxon、Pipeline Pilot from Accelrys 等.

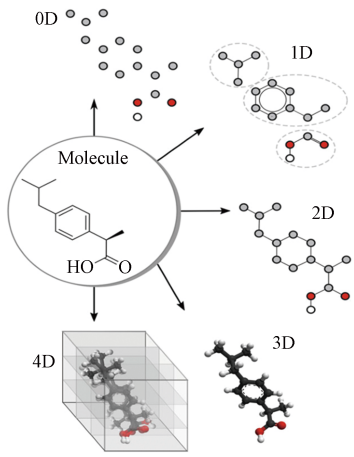

图 2 同分子的不同分子表示的示例[34]

Fig.2 An example of different molecular representations of the same molecule[34]

(2)分子图

分子图是指化合物用图进行表示,是化合物的拓扑表示. 在分子图中,原子用节点表示,分子键用边表示,示例可见图 3. 分子图降低了分子结构表示的复杂性,可以捕捉到分子中原子核与电子间的关键的相互作用. 此外,图神经网络(GNN)模型从分子图进行学习表示可以得到很好的处理效果,减少了相应的特征工程的工作,能进行更好的分子性质预测,如 Attentive FP[44]、D-MPNN[8].

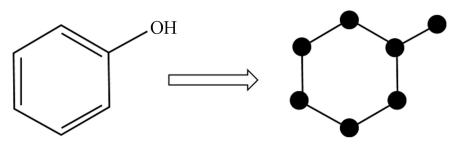

图 3 化合物转换成分子图的示例

Fig.3 An example of a compound turning into a molecular graph

(3)分子线性表示

分子线性表示最常用的有两种:简化分子线性输入规范(Simplified molecular input line entry system, SMILES)和国际化合物标识符(International Chemical Identifier, InChI). SMILES 是一种利用 ASCII 编码表示分子结构的线性符号,是化学家为了进行化学方面的机器计算而设计的化学符号语言,是根据相应的规则对化学结构简化的二维价键图[45]. SMILES 既可以与化学数据库使用,又可以节省存储空间,为化学数据的输入提供了一种更简便的方式. 化合物的“SMILES”字符串可通过一些软件或程序获得,如 ChemDraw、OpenBabel、CIRpy [46](https://github.com/mcs07/CIRpy)等,同时也可以通过网站获得化合物的“SMILES”字符串,如 PubChem. “SMILES”字符串除了可以直接作为模型的输入,也可以通过一些软件或程序转换为其他分子特征化形式,再作为模型的输入[6,46−48]. 通用的 SMILES 基于 CANGEN 算法衍生了规范的 SMILES(Canonical SMILES),但其算法具有盈利性质,从而存在无法自由使用的问题. InChI[49] 是一个非盈利的、免费的化学标识,在描述分子方面具有严格的唯一性,在层状设计时考虑了分子结构,容易获得和生成,可以由 InChI 软件或者利用通用的化学绘图软件生成. 因此,InChI 也被许多化学数据库使用.

(4)分子图像

分子图像是将分子结构或坐标映射到图像上后,作为模型的输入数据用于模型训练,从而进行分子性质预测[50]. 比如,可以通过 OpenBabel、Pybel 和 RDKit 等化学信息软件将 SMILES 解码为对应的

分子二维结构, 再将其生成的坐标映射到网格上, 形成分子图像, 示例可见图 4. 对于所生成的图像可再进行一个“灰色编码”或者更为复杂的“颜色编码”, 表示出原子/键属性, 再用于卷积神经网络(CNN)算法进行训练[50－51].

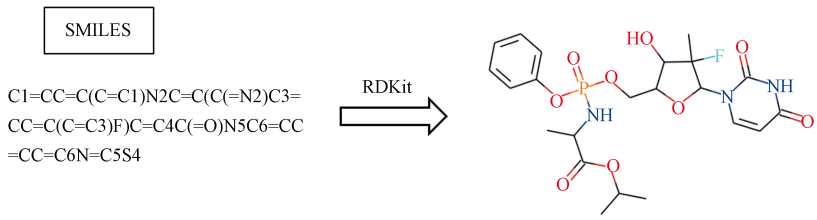

图 4　RDkit 将化合物 SMILES 转换为分子图像的示例

Fig.4　An example of RDkit transforming SMILES into the molecular image

1.2.2　数据集划分

整理好研究所需的数据集之后, 应及时对数据进行划分, 数据划分对于所建机器学习模型的泛化能力有一定影响. 一般, 数据集按比例随机划分为 3 部分: 训练集、测试集、验证集, 其中, 训练集一般用于模型的训练, 测试集用于模型的性能评估, 验证集用于超参数的优化[52－53]. 但是, 研究过程中数据集划分的具体的分配比例应按照研究需要进行选择, 比如, 有研究按 8:1:1 的比例划分成训练集、验证集和测试集[47]; 也有研究先将数据集按 4:1 的比例随机分成两部分: 训练集和测试集, 随后在训练过程中随机抽取 10% 的训练集数据作为验证集数据[52].

当机器学习应用于分类问题时, 可能会碰到数据集中各类别样本数量分布不均衡的问题, 即数据集中某一类别的样本数量远远高于另一类. 不平衡数据会影响模型的性能, 因此, 在数据集划分后需对训练集数据进行不平衡处理. 目前进行不平衡数据处理的方法主要有数据重构和分类模型改进.

数据重构策略可分为特征选择和重采样技术[54]. 特征选择方法主要分成 3 类: 过滤式、包裹式和嵌入式. 重采样技术是通过调整多数类和少数类的分布, 削弱数据集不平衡的程度, 主要包括欠采样、过采样、混合采样. 欠采样通过减少多数类中的样本数量, 以平衡多数类和少数类; 过采样通过增加少数类中的样本数量, 以均衡数据集; 混合采样通过将过采样和欠采样组合在一起, 提高分类性能[52].

分类模型改进策略主要从分类算法和分类思想对不平衡数据集进行优化和改进. 分类算法主要有 K 最近邻、支持向量机、决策树、朴素贝叶斯、神经网络算法; 分类思想主要有代价敏感学习、集成学习、单类学习、主动学习[54].

1.3　模型的构建

1.3.1　算法选择

模型训练是指通过训练集数据进行拟合模型、学习模型参数的过程. 在这个过程, 需要选择合适的算法进行训练. 算法是机器学习的核心内容, 没有算法, 机器学习模型将无法运行. 目前, 机器学习的算法按照是否有人类监督训练, 可分为监督式学习、无监督式学习、半监督式学习、强化学习这 4 个主要类型. 监督式学习可以处理有标记的训练数据, 其算法主要有 K-近邻算法、线性回归、逻辑回归、支持向量机(SVM)[55]、神经网络、决策树[56] 和随机森林(RF)[57]. 无监督式学习使用的数据是没有标记的, 其算法主要可分为聚类算法、可视化和降维算法、关联规则学习算法. 聚类算法主要有 k-平均算法、分层聚类分析、最大期望算法等; 可视化和降维算法主要有主成分分析(PCA)、核主成分分析(Kernel PCA)、局部线性嵌入(LLE)、t-分布随机近临嵌入(t-SNE); 关联规则学习算法主要有 Apriori、Eclat. 半监督式学习可以处理部分标记(大量未标记和少量标记)的数据, 其主要为监督式学习算法与无监督式学习算法的结合. 强化学习[58] 是学习到行动的一种映射, 通过不断试错, 寻找到能够最大化预期的路径, 并对能最大化预期的行动进行奖励, 主要涉及到的策略是试错搜索和延迟奖励, 这两种策略是强化学习的最明显最重要的特征. 在化合物属性预测中, 常用的是监督学习模式[59]、半监督式学习; 常用的算法有随机森林[57]、支持向量机[55]、朴素贝叶斯、神经网络、梯度提升决策树(GBDT)、极限梯度提升算法(XGBoost)、线性回归、决策树[56]、逻辑回归等算法.

1.3.2 模型应用域

在化学信息研究中，为了更高效地进行化合物属性的预测，通常将机器学习方法应用于定量构效关系中，其中涉及到的模型应用域（AD）一般指化学空间中由描述符和模型响应定义的一个理论域，其任务是定义一个模型可以被使用的边界，并提供可靠的预测[60－61]. 当要预测的分子在所定义的 AD 范围内时，使用该模型进行预测才可靠，否则该模型不适用.

对于模型应用域的表征，已有学者在这个方面进行过总结，不同的学者考虑的角度不同，进行的归类方式和描述称呼各有不同. Kar 等[61] 根据不同假设将 AD 的定义方法分为 6 大类：描述符空间中基于范围的方法、基于距离的方法、几何方法、概率密度分布、响应变量的范围、其他方法，而王中钰等[62] 根据 AD 的概念将其分为描述符域、结构域、机理域 3 个大类，再从这 3 大类中对 AD 的表征方法进行细分. 此外，也有一些学者提出或开发了新的应用域表征方法或应用域算法，如 Wang 等人[63] 基于指纹特异性相似性阈值，开发了新的 AD 表征方法—$AD_{fingerprint}$，并证明其性能优于一些传统的 AD 表征方法；Berenger 等[64] 对于创建的应用域较为复杂并难以理解的现状，提出了基于距离的 Boolean 应用域算法（DBBAD）. 虽然 AD 表征方法的描述各有异同，但最常用的几种表征方法一般为欧式距离、Tanimoto 指数、杠杆方法、马氏距离、核密度估计（KDE）、基于范围的超矩形等方法. 连续数据的研究，一般选用欧式距离定义模型应用域[65]；而针对二进制数据或者想要进行分子相似度的比较的研究，一般选用 Tanimoto 指数定义模型应用域[47－48].

1.3.3 模型性能评估

模型的性能评估是对所构建的模型的泛化能力进行评估，有助于判定模型的工作性能和开发适合研究数据的最佳模型，主要包括了性能度量、评估方法、过拟合或欠拟合、超参数调优、泛化能力这几个方面的内容. 模型泛化能力是指所构建的模型经过在训练集数据的训练之后，在新数据上的适应能力. 过拟合和欠拟合都是模型泛化能力不好的行为表示. 过拟合指模型过度学习训练集数据，使得模型过为复杂，不能在除训练集外的数据集上得到好表现；欠拟合指模型过为简单，学习能力差，无法学到数据的内在特点，无法判定其潜在的趋势[66].

模型性能的评估方法常见的有留出法（hold-out）、交叉验证法（cross validation）、自助法（bootstrapping）. 在模型评估的实际情况中，一般先用评估方法对模型进行数据集划分，再在测试集上用评估指标对模型性能进行评估. 比如，Korkmaz[52] 在研究中先用留出法将数据集划分为 80% 的训练集和 20% 的测试集，之后用计算了几个性能指标对模型进行了性能评估.

在化合物属性预测方面的应用，主要可将机器学习任务分为回归问题和分类问题两方面. 在性能度量涉及到的性能指标方面，回归问题和分类问题所用到的性能评价指标并不完全一致. 回归问题常用到的评价指标有均方根误差（root mean squared error, RMSE）、平均绝对误差（mean absolute error, MAE）、均方误差（mean square error, MSE）、均方根对数误差（root mean squared logarithmic error, RMSLE）、决定系数（coefficient of determination, R^2）和预测平方相关系数（predictive squared correlation coefficient, Q^2）等. 分类问题常用到的评价指标有准确率（accuracy, acc）、错误率（error）、精确率（precision rate, p）、召回率（recall rate, r）、F1 分数（F1-score）、ROC 曲线（receiver operating characteristic）、AUC（area under curve）等. 分类问题的数据集并不一定平衡，在大多应用情况下都会出现数据分布不均导致数据不平衡的现象，这种情况下，首先要在训练集上进行数据不平衡处理，之后再用测试集对其进行评估. 针对不平衡数据集，准确率往往无法作为主要的判断指标，因此一般可采用前文所提到的精确率、召回率、F1 分数以及均衡准确率（balanced accuracy）和 G-mean[54]. 除了上述的指标外，还有一些其他的指标，如鲁棒性、PRC（精确-召回曲线）等. 在实际的应用情况中，指标的选择应根据数据的情况和研究需要来进行选择.

1.4 模型解释

模型解释是对模型的预测机制进行解析的过程，有利于研究者做出更好的决策，并建立起对模型的理解和信任[67－68]. 模型根据解释的难易程度，可以分为“白盒”模型和“黑盒”模型. “白盒”模型又可称为可解释性模型，创建其模型的算法透明度低，解释简单，更易被人们理解. 可解释性模型一般指由线性回归、逻辑回归、其他线性扩展、决策树等算法构建的模型. 建立“黑盒”模型后再进行解释这一行为

也可称为事后可解释性, 进行事后可解释的方法主要可以分为两大类: 全局解释和局部解释, 全局解释是对模型整体进行解释, 而局部解释是对单个预测进行解释[69].

全局解释的方法主要有部分依赖图(partial dependence plot, PDP)、累积局部效应(accumulated local effects plot, ALE)、规则提取[70]、模型蒸馏[71]、稀疏集团套索(sparse group lasso, SGL)[72]、全局 Shapley 值等. 局部解释的方法主要有个体条件期望(individual conditional expectation, ICE)[73]、敏感性分析、局部可解释的模型无关阐释(local interpretable model-agnostic explanations, LIME)[67]、Anchor[74]、基于局部规则的黑盒模型的分层相关性传播(LRP)[75]、类激活映射(class activation mapping, CAM)、梯度加权类激活映射(Grad-CAM)[76]、SHAP(shapley additive exPlanations)[77] 等. 在化合物属性预测方面, 比较常用的解释方法有 PDP、ALE、ICE、Grad-CAM、Shapley Value、SHAP 等. 如 Zhong 等[6] 利用 Grad-CAM 来解释构建的 CNN 模型通过选择分子图像的哪些特征来进行预测. Sanches-Neto 等[46] 在预测水中有机污染物自由基氧化过程的反应速率常数的研究中, 利用 SHAP 方法解释了反应过程中相关的结构分子特征, 将氧原子所做的贡献从氧原子与碳原子的比例(#O:C)的贡献区分出来.

2　机器学习在化合物属性预测中的应用进展(The application progress of machine learning on compound property prediction)

2.1　理化性质预测

机器学习中的神经网络算法可被用于量子化学性质预测. 比如, 2017 年, 由 Gilmer 等[7] 提出来的应用于分子图的监督学习框架——消息传递神经网络(message passing neural networks, MPNNs), 更易理解图的结构数据与模型之间的关系. 他们基于 MPNNs 进行建模, 采用 QM9 数据集的数据, 对分子的原子化能、振动频率、最高占据分子轨道(HOMO)、最低未占据分子轨道(LUMO)、偶极矩等性质进行了预测, 结果表明利用机器学习进行分子性质预测的成本比密度泛函理论(DFT)计算低且计算速度更快, 计算样本误差比 DFT 小, 在大型图中应用良好. 之后, 有学者在 MPNNs 的基础上进行改动, 提出了知识嵌入消息传递神经网络(KEMPNN)[78]. KEMPNN 在 MPNN 中的消息传递阶段添加了知识注意机制作为一项加权项, 采用两个数据集共同训练 MPNN, 并在 ESOL, FreeSolv, Lipophilicity 以及聚合物性能数据集上进行了测评, 与 MPNN 进行了对比. 结果表明, KEMPNN 比 MPNN 的模型的预测精度更高, 并且发现了 KEMPNN 在小数据集上的预测效果可与基于描述符的方法相当甚至更好.

溶剂化自由能与许多物理化学性质密切相关, 在药物发现方面有重要的影响, 但溶剂化自由能的实验数据较少, 且实验成本昂贵. 尽管已经有一些相应的溶剂模型可预测溶剂化自由能, 使得费用成本有所降低, 但其准确性较低. 相比之下, 机器学习在溶剂化自由能预测方面更具优势, 既不会产生昂贵的费用, 又保证了较高的溶剂化自由能预测准确率[10, 79]. 如, Yang 等[8] 在 MPNN 的基础上构建了一个基于定向键的消息传递方式, 并结合分子水平特征和分子式构建了新的模型 D-MPNN, 在 FreeSolv 数据集上表现出比其他基准模型更好的性能; Weinreich 等[9] 提出了一个以核岭回归(KRR)算法作为监督机器方法的自由能机器学习模型(FML), 并在 FreeSolv 数据集和 QM9 数据集上进行了溶剂化自由能预测, 模型误差与最好的物理预测方法相当, 但计算成本更低, 并且可在较小数据集上达到溶剂化的实验不确定度. 需要指出的是, 机器学习在溶剂化自由能预测方面存在数据稀缺的问题, 深度学习的模型在小数据集上容易过拟合, 性能差. 鉴于此, Vermeire 等[80] 基于 D-MPNN 构建了一个模型, 通过引入一种结合量子化学和实验数据的迁移学习方法使模型在溶剂化自由能预测方面的性能得到了显著提升; Zhang 等[10] 提出一个基于 GNN 和 3D 原子特征的深度学习(DL)模型构架, GNN 以主领域聚合 PNAConv 作为编码器, 并将其与迁移学习策略相结合, 进行模型微调后在 FreeSolv 数据集进行溶剂化自由能预测并得到了目前最好的性能, RMSE 为 0.719 kcal·mol^{-1}, MAE 为 0.417 kcal·mol^{-1}, 显著提高了 GNN 模型在溶剂化自由能预测方面的学习能力, 为处理小型实验数据集提供了思考方向.

此外, 机器学习在预测化合物的其他性质方面也有不错的表现, 以全氟化合物(PFASs)理化性质预测为例. 在全氟化合物(PFASs)理化性质预测方面, Raza 等[11] 在 2019 年提出了第一个利用机器学习来预测各种 PFAS 结构中的 C—F 键解离能的应用. 这个应用高效可靠准确, 训练数据时间短, 预测 C—F 键解离能的时间不超过 1 s, 偏差小于 0.70 kcal·mol^{-1}, 不需量子力学计算, 计算成本更低, 有助于

PFAS 和高效处理与去除. 之后, 有学者[81] 于 2021 年构建了一个数据库框架, 所构建 PFAS-Map 可以预测未测定的 PFAS 化学品的基本物理性质, 可视化 PFAS 活性/性质关系的实验数据趋势, 发现隐藏的结构 - 毒性关系.

2.2 生物活性预测

机器学习在上世纪就开始用于进行生物活性预测. 在 20 世纪 90 年代, 神经网络算法广泛应用于定量结构-活性关系, 但由于其算法的局限性, 在 2000 年早期被 SVM 和 RF 取代. 近些年, 神经网络算法逐步改进, 引起了人们的关注, 发现改进后的神经网络算法在生物活性预测方面颇具优势. 2015 年, Ma 和 Dahl 等[82] 采用“原子对”描述符和“供体-受体对”描述符的并集作为描述符来训练模型, 并将深度神经网络(DNN)的性能评估参数 R^2 与 RF 模型在 15 个数据集(Merck 公司内部的数据集)上进行比较, 结果表明 DNN 在大多数情况下预测性能都优于 RF 模型, 在计算时间和成本方面甚至比 RF 更有优势, 可作为一种实用的 QSAR 方法. 但需要指出的是, 该项研究也存在局限性, 无法阐明分子间未完成的潜在相互作用. 针对这些缺点, Wallach 等[12] 建立了第一个基于结构的深度卷积神经网络—AtomNet, 可应用于小分子生物活性预测. 他们将 AtomNet 与 DNN 技术进行对比, 发现 AtomNet 可为目标预测出新的活性分子, 所构建的模型能发现任意的分子特征, 可描述配体和目标之间的相互作用; 同时, 在 3 个基准上做了应用, 结果表明 AtomNet 表现出色, 在 DUDE 基准测试中有一半的目标的 AUC 为 0.9, 远超以前的对接方法.

此外, 2019 年, Cheng 和 Ng[13] 在前人的基础上建立了 ML-QSAR 模型预测全氟化合物(PFASs)的生物活性, 引入了基于图的模型, 预测了 OECD 名单中未经测试的 PFASs 的生物活性. 在整个过程中, 基于自行收集整理的 PFASs 数据库训练和评估了 5 种机器学习模型, 采用了 ECFP、图卷积、weave 特征 3 种方法进行分子特征化, 网格搜索和贝叶斯优化技术进行超参数调优, 基于距离的方法确定 QSAR 模型的 AD 值, 结果表明, 多任务神经网络模型和基于图的图卷积模型性能优异, 但构建的模型不能提供有关效应强度或剂量反应的信息, 有进一步发展的空间. 此外, 不同于常用于化合物活性预测的结构-活性关系(SAR)模型, Bertoni 等[14] 于 2021 年构建了一个深度神经网络的集合—SigAR (signature–activity relationship)模型预测分子的生物活性, 让机器学习从化合物的 CC signatures(基于一个小分子生物活性特征集合开发的分子表征方法)中学习活性特征, 并用 MoleculeNet 中的 9 个数据集评估了 SigAR. 其结果表明, 相较于基于化学描述符的方法, SigAR 的性能更好.

2.3 毒性预测

对化合物的毒性进行预测, 是药物研发的一部分, 对于药物研发的成本和成功率有重要影响. 同时, 化合物毒性预测也是化学品风险评估的一部分内容, 但基于动物实验的毒性预测, 时间周期长, 成本开支大. 此外, 人工合成化合物的种类在逐渐增多, 在日常生活中随处可见, 识别危险化学品的潜在毒性是有必要的, 对化合物进行毒性预测的需求在持续增长. 机器学习应用于化合物的毒性预测具有降低成本和加快研究速度的特点, 因此, 机器学习在化合物毒性预测方面的研究一直以来都是热点研究领域, 相关的研究也比较多.

2008 年, 美国的 EPA、NIH 和 FDA 开展了 Tox21 计划, 这个计划汇总了许多化合物的毒性数据, 推动了机器学习在预测化学品的潜在毒性和评估化学品风险的进程. 2016 年, Mary 等[83] 开发了适用于毒性预测的集成模型—DeepTox, 并将其运用于 Tox21 挑战赛上. 他们采用了化合物的大量的静态特征(如, MACCS 指纹、PubChem 子结构指纹等)和动态特征(如, ECFP 指纹、径向 2D 指纹等)作为机器学习的输入, 并对 DeepTox 中的每个机器学习算法模型进行了性能评估, 比较了各算法的 AUC 值, 结果表明 DNN 优于 SVM、RF、弹性网(ElNet). 同时, 由 DNN 主导的 DeepTox 应用于预测化合物毒性, 取得了 Tox21 大挑战的冠军. 2019 年, Pu 等[84] 基于机器学习技术开发了一个新的程序—eToxPred, 可以直接从分子指纹预测小型化合物的毒性. eToxPred 采用额外树(Extra Trees, ET)算法作为毒性预测的默认分类器, 并在不同的数据集上与线性判别分析(LDA), 多层感知器(MLP), 随机森林(RF)算法进行了性能对比. 结果表明, 使用分子指纹作为输入, 基于 ET 的分类器性能普遍高于 LDA 和 MLP, 仅在一个组合数据集上略低于 RF, 可以非常有效地识别有毒分子和预测特定毒性.

在化合物毒性评估方面, 常用结构警报(structural alerts, SAs)作为识别危险化学品的潜在毒性的

方法, 但 SAs 的准确性有限, 有时在无毒化合物中也会发现 SAs[85]. Mukherjee 等[15] 引入了一个新概念——“关键结构图案”(critical structural motif, CSM), CSM 包含了 SAs 的特异性. 同时, 他们用 SMILES 字符串作为模型输入, 开发了一个基于卷积神经网络(CNN)的多输出分类的深度学习模型--VisualTox, 并在不同的化学数据上进行了训练, 通过识别 CSM 来预测内分泌干扰物质(ECD)的毒性, 提供了一种理解化学毒性来源的新方法.

此外, 持久性有机污染物(POPs)和持久性、生物累积性和毒性物质(PBT)对生态环境和人类健康都有重大影响, PBT/POP 类化学品也备受人们的关注. Sun 等[16] 于 2020 年采用基于 2424 个分子描述的二维表示矩阵(MDRM)作为模型输入, 开发了一个深度卷积神经网络(DCNN)模型来筛选化学品库中潜在的 PBT/POP 类物质, 并采用 k 折交叉验证法和专家经验判断方法对模型性能进行评价, 得到模型的预测精度可达 90.4%. 但需要指出的是, DCNN 模型是一个“黑盒”模型, 基本不可得到有效的解释. 最近, Wang 等[17] 利用一个包含 14994 种 PBT 和 non-PBT 物质的化学数据库, 基于图注意力网络(graph attention networks, GATs)架构, 构建了可筛选 PBT 化学品并具有可解释性的 GAT 模型. GATs 是一种较先进的 GNN, 为分子图的每个节点引入了注意权重参数(P_{AW}), 可反映节点对预测端点的贡献, 关注与目标任务相关的重要局部结构, 具有模型可解释性. 他们在 AD 表征方面, 提出并采用了一种新的方法—$AD_{FP\text{-}AC}$, 使 GAT 模型更加可靠; 在模型性能方面, 将具有 $AD_{FP\text{-}AC}$ 表征的 GAT 模型与 DCNN 模型、传统的机器学习方法(如随机森林、支持向量机)和根据不同分子特征化方法建立的 QSAR 模型进行性能对比, 发现 GAT 模型的性能最佳. 在建立好 GAT 模型之后, 他们还将其应用在中国现有化学物质清单(IECSC)上, 从中确定了 8 类之前未确认的化合物类别为 PBT 化学品.

3　机器学习在化合物属性预测中的挑战(The challenges of machine learning on compound property prediction)

3.1　数据集

目前, 在构建数据集的过程中, 研究者往往面临以下 3 个问题, 包括数据量不足、数据质量不高以及数据不平衡. 针对数据量不足问题, 虽然前文介绍了一些相关的公开数据库, 但这些数据库对于研究人员来说, 数量还是较少, 而且数据不够全面, 很多重要的化学信息被收集在商业数据库中或其他难以获取的数据库中. 此外, 虽然公开的一些大型化学数据库数据多, 规模大, 但是拥有的标签数据并不多. 这种情况限制了监督学习在化合物属性预测方面进行更深入的研究[51]. 这些都使得研究人员无法得到足够的数据, 利用机器学习在化合物属性层面进行一个更好的突破. 面临的问题之二是数据的质量不高. 有些数据来自于实验记录, 虽然实验记录数据能得到更多、更为全面的数据, 但公开的实验数据如何保证质量, 也是值得思考的问题. 面临的问题之三是数据不平衡问题. 虽然机器学习有许多经典的分类算法, 如朴素贝叶斯、KNN、基于神经网络的分类算法等, 这些算法尽可能地保留了原数据所有的信息, 但是由于这些算法的假设都是基于平衡的样本数据, 所以当数据有少数类和多数类的情况出现时, 这些算法皆会更倾向于多数类数据[86]. 对此, 许多学者提出了一些数据不平衡处理方法, 如过采样、欠采样、混合采样和特征选择等, 这些处理方法在一定程度上能够缓解不平衡问题, 但都存在不足. 比如, 欠采样方法虽然简单又效果好, 但是容易忽略多数类数据的内在特征信息, 影响模型的泛化能力.

3.2　分子特征化

分子特征化方法是化合物属性预测中的重点之一, 决定了模型的性能和解释. 目前, 分子特征化方法能够表征的信息很多, 比如, 定量分子描述符可以量化 Hammett 常数、偶极矩、HOMO 和 LUMO 能量等信息, 为化合物的性质预测提供了良好的输入信息. 但目前还没有可以完整表达原始分子信息的特征化方式[87].

此外, 虽然分子指纹种类也颇多, 但目前主要还是用二维(2D)分子指纹来做相应的研究, 高维度的分子指纹设计较为困难, 这导致了现有的分子指纹种类缺少对分子立体结构描述的三维结构信息. 对于此类问题, 近些年也有学者提出了代数图、代数拓扑、微分几何等分子三维结构信息的表示方法, 但是这些方法较为依赖分子结构的可用性[88]. 分子特征化方法在描述分子的立体化学信息方面还有许多空间可以提升.

3.3 模型的可解释性

模型的可解释性是模型的重要部分,是可信性的前提,如何让模型的工作机制更为透明,获得人们的理解和信任,这是值得讨论和重视的. 可解释的机器学习模型没有“黑盒”模型的特征,更易被理解,透明度高. 相较于可解释的机器学习模型,具有“黑盒”特征的机器学习模型虽然更难进行直观的解释,但是其性能更高,预测效果更好. 因此,如何对“黑盒”模型进行更好的解释,增加模型的可解释性,需要更多的研究来进行探究. 同时,现在缺乏明确的模型的可解释性基准,没有严格的方法来评估和比较模型解释方法[89].

4 总结(Conclusion)

机器学习在化合物属性预测方面的应用不断拓展,不仅提高了预测结果的准确性,而且为评估新化学物质的环境风险提供了新方法. 其中,深度学习算法更适用于大数据集,而机器学习算法应用在小数据集更具优势. 但是,机器学习在化合物属性预测中的应用仍存在未知和挑战,这些亟待解决的问题将是未来研究工作的焦点. 机器学习(特别是深度学习)将会与量子力学、毒理学、量子化学、电化学等深度融合,在药物研发、毒理学研究、环境行为预测、材料研发等领域继续发挥重要作用.

参考文献 (References)

[1] MARTIN Y C, KOFRON J L, TRAPHAGEN L M. Do structurally similar molecules have similar biological activity? [J]. Journal of Medicinal Chemistry, 2002, 45(19): 4350-4358.

[2] PANDEY S, QU J X, STEVANOVIĆ V, et al. Predicting energy and stability of known and hypothetical crystals using graph neural network [J]. Patterns, 2021, 2(11): 100361.

[3] WALTERS W P, BARZILAY R. Applications of deep learning in molecule generation and molecular property prediction [J]. Accounts of Chemical Research, 2021, 54(2): 263-270.

[4] HANSCH C, MALONEY P P, FUJITA T, et al. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients [J]. Nature, 1962, 194(4824): 178-180.

[5] CHERKASOV A, MURATOV E N, FOURCHES D, et al. QSAR modeling: Where have you been? Where are you going to? [J]. Journal of Medicinal Chemistry, 2014, 57(12): 4977-5010.

[6] ZHONG S F, HU J J, YU X, et al. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation [J]. Chemical Engineering Journal, 2021, 408: 127998.

[7] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for Quantum chemistry[C]//Proceedings of the 34th International Conference on Machine Learning - Volume 70. August 6-11, 2017, Sydney, NSW, Australia. New York: ACM, 2017: 1263–1272.

[8] YANG K, SWANSON K, JIN W G, et al. Analyzing learned molecular representations for property prediction [J]. Journal of Chemical Information and Modeling, 2019, 59(8): 3370-3388.

[9] WEINREICH J, BROWNING N J, von LILIENFELD O A. Machine learning of free energies in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation [J]. The Journal of Chemical Physics, 2021, 154(13): 134113.

[10] ZHANG D D, XIA S, ZHANG Y K. Accurate prediction of aqueous free solvation energies using 3D atomic feature-based graph neural network with transfer learning [J]. Journal of Chemical Information and Modeling, 2022, 62(8): 1840-1848.

[11] RAZA A, BARDHAN S, XU L H, et al. A machine learning approach for predicting defluorination of per- and polyfluoroalkyl substances (PFAS) for their efficient treatment and removal [J]. Environmental Science & Technology Letters, 2019, 6(10): 624-629.

[12] WALLACH I, DZAMBA M, HEIFETS A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery [J]. Mathematische Zeitschrift, 2015, 47(1): 34-46.

[13] CHENG W X, NG C A. Using machine learning to classify bioactivity for 3486 per- and polyfluoroalkyl substances (PFASs) from the OECD list [J]. Environmental Science & Technology, 2019, 53(23): 13970-13980.

[14] BERTONI M, DURAN-FRIGOLA M, BADIA-I-MOMPEL P, et al. Bioactivity descriptors for uncharacterized chemical compounds [J]. Nature Communications, 2021, 12(1): 3932.

[15] MUKHERJEE A, SU A, RAJAN K. Deep learning model for identifying critical structural motifs in potential endocrine disruptors [J]. Journal of Chemical Information and Modeling, 2021, 61(5): 2187-2197.

[16] SUN X F, ZHANG X M, MUIR D C G, et al. Identification of potential PBT/POP-like chemicals by a deep learning approach based on

2D structural features [J]. Environmental Science & Technology, 2020, 54(13): 8221-8231.
[17] WANG H B, WANG Z Y, CHEN J W, et al. Graph attention network model with defined applicability domains for screening PBT chemicals [J]. Environmental Science & Technology, 2022, 56(10): 6774-6785.
[18] KIM S, CHEN J, CHENG T J, et al. PubChem in 2021: New data content and improved web interfaces [J]. Nucleic Acids Research, 2021, 49(D1): D1388-D1395.
[19] PENCE H E, WILLIAMS A. ChemSpider: An online chemical information resource [J]. Journal of Chemical Education, 2010, 87(11): 1123-1124.
[20] BLUM L C, REYMOND J L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13 [J]. Journal of the American Chemical Society, 2009, 131(25): 8732-8733.
[21] RUDDIGKEIT L, van DEURSEN R, BLUM L C, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17 [J]. Journal of Chemical Information and Modeling, 2012, 52(11): 2864-2875.
[22] MOBLEY D L, GUTHRIE J P. FreeSolv: A database of experimental and calculated hydration free energies, with input files [J]. Journal of Computer-Aided Molecular Design, 2014, 28(7): 711-720.
[23] IRWIN J J, STERLING T, MYSINGER M M, et al. ZINC: A free tool to discover chemistry for biology [J]. Journal of Chemical Information and Modeling, 2012, 52(7): 1757-1768.
[24] IRWIN J J, TANG K G, YOUNG J, et al. ZINC20-a free ultralarge-scale chemical database for ligand discovery [J]. Journal of Chemical Information and Modeling, 2020, 60(12): 6065-6073.
[25] IRWIN J J, SHOICHET B K. ZINC: A free database of commercially available compounds for virtual screening [J]. Journal of Chemical Information and Modeling, 2005, 45(1): 177-182.
[26] BENTO A P, GAULTON A, HERSEY A, et al. The ChEMBL bioactivity database: An update [J]. Nucleic Acids Research, 2014, 42(D1): D1083-D1090.
[27] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018 [J]. Nucleic Acids Research, 2018, 46(D1): D1074-D1082.
[28] DIX D J, HOUCK K A, MARTIN M T, et al. The ToxCast program for prioritizing toxicity testing of environmental chemicals [J]. Toxicological Sciences, 2007, 95(1): 5-12.
[29] LIU Z H, LI Y, HAN L, et al. PDB-wide collection of binding data: Current status of the PDBbind database [J]. Bioinformatics, 2015, 31(3): 405-412.
[30] GILSON M K, LIU T Q, BAITALUK M, et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology [J]. Nucleic Acids Research, 2016, 44(D1): D1045-D1053.
[31] SHEN W X, ZENG X, ZHU F, et al. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations [J]. Nature Machine Intelligence, 2021, 3(4): 334-343.
[32] WANG L G, ZHAO L, LIU X, et al. SepPCNET: Deeping learning on a 3D surface electrostatic potential point cloud for enhanced toxicity classification and its application to suspected environmental estrogens [J]. Environmental Science & Technology, 2021, 55(14): 9958-9967.
[33] TODESCHINI R, CONSONNI V. Molecular Descriptors for Chemoinformatics[M]. Wiley-VCH, 2009.
[34] GRISONI F, CONSONNI V, TODESCHINI R. Impact of molecular descriptors on computational models [J]. Methods in Molecular Biology (Clifton, N. J.), 2018, 1825: 171-209.
[35] 吴萍, 孔德信. 分子相似性与MOLPRINT 2D的本地化 [J]. 计算机与应用化学, 2008, 25(4): 505-508.
WU P, KONG D X. Molecular similarity and localization of MOLPRINT 2D [J]. Computers and Applied Chemistry, 2008, 25(4): 505-508(in Chinese).
[36] KHAN A U. Descriptors and their selection methods in QSAR analysis: Paradigm for drug design [J]. Drug Discovery Today, 2016, 21(8): 1291-1302.
[37] CERETO-MASSAGUÉ A, OJEDA M J, VALLS C, et al. Molecular fingerprint similarity search in virtual screening [J]. Methods, 2015, 71: 58-63.
[38] DURANT J L, LELAND B A, HENRY D R, et al. Reoptimization of MDL keys for use in drug discovery [J]. Journal of Chemical Information and Computer Sciences, 2002, 42(6): 1273-1280.
[39] ROGERS D, HAHN M. Extended-connectivity fingerprints [J]. Journal of Chemical Information and Modeling, 2010, 50(5): 742-754.
[40] BENDER A, MUSSA H Y, GLEN R C, et al. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance [J]. Journal of Chemical Information and Computer Sciences, 2004, 44(5): 1708-1718.
[41] MAURI A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints[M]// Ecotoxicological QSARs. New York: Humana, 2020: 801-820.
[42] O'BOYLE N M, BANCK M, JAMES C A, et al. Open Babel: An open chemical toolbox [J]. Journal of Cheminformatics, 2011, 3: 33.
[43] STEINBECK C, HAN Y Q, KUHN S, et al. The Chemistry Development Kit (CDK): An open-source *Java* library for Chemo- and

Bioinformatics [J]. Journal of Chemical Information and Computer Sciences, 2003, 43(2): 493-500.

[44] XIONG Z P, WANG D Y, LIU X H, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism [J]. Journal of Medicinal Chemistry, 2020, 63(16): 8749-8760.

[45] WEININGER D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules [J]. Journal of Chemical Information & Computer Sciences, 1988, 28(1): 31-36.

[46] SANCHES-NETO F O, DIAS-SILVA J R, KENG QUEIROZ L H Jr, et al. "*py*SiRC": Machine learning combined with molecular fingerprints to predict the reaction rate constant of the radical-based oxidation processes of aqueous organic contaminants [J]. Environmental Science & Technology, 2021, 55(18): 12437-12448.

[47] ZHONG S F, ZHANG K, WANG D, et al. Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds [J]. Chemical Engineering Journal, 2021, 405: 126627.

[48] ZHONG S F, HU J J, FAN X D, et al. A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants [J]. Journal of Hazardous Materials, 2020, 383: 121141.

[49] HELLER S R, McNAUGHT A, PLETNEV I, et al. InChI, the IUPAC international chemical identifier [J]. Journal of Cheminformatics, 2015, 7: 23.

[50] GOH G B, SIEGEL C, VISHNU A, et al. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models[J]. ArXiv, 2017, abs/1706.06689

[51] GOH G B, SIEGEL C, VISHNU A, et al. ChemNet: A transferable and generalizable deep neural network for small-molecule property prediction[J]. ArXiv, 2017, abs/1712.02734

[52] KORKMAZ S. Deep learning-based imbalanced data classification for drug discovery [J]. Journal of Chemical Information and Modeling, 2020, 60(9): 4180-4190.

[53] WU Z Q, RAMSUNDAR B, FEINBERG E N, et al. MoleculeNet: A benchmark for molecular machine learning [J]. Chemical Science, 2017, 9(2): 513-530.

[54] 徐玲玲, 迟冬祥. 面向不平衡数据集的机器学习分类策略 [J]. 计算机工程与应用, 2020, 56(24): 12-27.
XU L L, CHI D X. Machine learning classification strategy for imbalanced data sets [J]. Computer Engineering and Applications, 2020, 56(24): 12-27(in Chinese).

[55] NOBLE W S. What is a support vector machine? [J]. Nature Biotechnology, 2006, 24(12): 1565-1567.

[56] QUINLAN J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81-106.

[57] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[58] SUTTON R S, BARTO A G. Reinforcement learning: An introduction [J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054.

[59] MITCHELL J B O. Machine learning methods in chemoinformatics [J]. Wiley Interdisciplinary Reviews. Computational Molecular Science, 2014, 4(5): 468-481.

[60] GRAMATICA P. Principles of QSAR models validation: Internal and external [J]. QSAR & Combinatorial Science, 2007, 26(5): 694-701.

[61] KAR S, ROY K, LESZCZYNSKI J. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling[M]// Computational Toxicology, Part of the Methods in Molecular Biology book series. New York: Humana Press, 2018: 141-169.

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

[63] WANG Z Y, CHEN J W, HONG H X. Developing QSAR models with defined applicability domains on PPARγ binding affinity using large data sets and machine learning algorithms [J]. Environmental Science & Technology, 2021, 55(10): 6857-6866.

[64] BERENGER F, YAMANISHI Y. A distance-based Boolean applicability domain for classification of high throughput screening data [J]. Journal of Chemical Information and Modeling, 2019, 59(1): 463-476.

[65] 郑玉婷, 乔显亮, 于洋, 等. 有机化学品生物富集因子定量结构-活性关系模型 [J]. 生态毒理学报, 2019, 14(2): 214-221.
ZHENG Y T, QIAO X L, YU Y, et al. Quantitative structure-activity relationship model for bioconcentration factors of organic chemicals [J]. Asian Journal of Ecotoxicology, 2019, 14(2): 214-221(in Chinese).

[66] 杨真真, 匡楠, 范露, 等. 基于卷积神经网络的图像分类算法综述 [J]. 信号处理, 2018, 34(12): 1474-1489.
YANG Z Z, KUANG N, FAN L, et al. Review of image classification algorithms based on convolutional neural networks [J]. Journal of Signal Processing, 2018, 34(12): 1474-1489(in Chinese).

[67] RIBEIRO M T, SINGH S, GUESTRIN C. "why should I trust You?": Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13-17, 2016, San Francisco, California, USA. New York: ACM, 2016: 1135-1144.

[68] BARREDO ARRIETA A, DÍAZ-RODRÍGUEZ N, del SER J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies,

opportunities and challenges toward responsible AI [J]. Information Fusion, 2020, 58: 82-115.
[69] PETCH J, DI S, NELSON W. Opening the black box: The promise and limitations of explainable machine learning in cardiology [J]. Canadian Journal of Cardiology, 2022, 38(2): 204-213.
[70] ANDREWS R, DIEDERICH J, TICKLE A B. Survey and critique of techniques for extracting rules from trained artificial neural networks [J]. Knowledge-Based Systems, 1995, 8(6): 373-389.
[71] LIU X, WANG X G, MATWIN S. Improving the interpretability of deep neural networks with knowledge distillation[C]//2018 IEEE International Conference on Data Mining Workshops (ICDMW). November 17-20, 2018, Singapore. IEEE, 2019: 905-912.
[72] MASHAYEKHI M, GRAS R. Rule extraction from decision trees ensembles: New algorithms based on heuristic search and sparse group lasso methods [J]. International Journal of Information Technology & Decision Making, 2017, 16(6): 1707-1727.
[73] GOLDSTEIN A, KAPELNER A, BLEICH J, et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation [J]. Journal of Computational and Graphical Statistics, 2015, 24(1): 44-65.
[74] RIBEIRO M T, SINGH S, GUESTRIN C. Anchors: High-precision model-agnostic explanations[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 1527-1535.
[75] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation [J]. PLoS One, 2015, 10(7): e0130140.
[76] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]// 2017 IEEE International Conference On Computer Vision (ICCV). 2017: 618-626.
[77] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017: 4768-4777.
[78] HASEBE T. Knowledge-embedded message-passing neural networks: Improving molecular property prediction with human knowledge [J]. ACS Omega, 2021, 6(42): 27955-27967.
[79] MATOS G D R, KYU D Y, LOEFFLER H H, et al. Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the FreeSolv database [J]. Journal of Chemical and Engineering Data, 2017, 62(5): 1559-1569.
[80] VERMEIRE F H, GREEN W H. Transfer learning for solvation free energies: From quantum chemistry to experiments [J]. Chemical Engineering Journal, 2021, 418: 129307.
[81] SU A, RAJAN K. A database framework for rapid screening of structure-function relationships in PFAS chemistry [J]. Scientific Data, 2021, 8(1): 1-10.
[82] MA J S, SHERIDAN R P, LIAW A, et al. Deep neural nets as a method for quantitative structure-activity relationships [J]. Journal of Chemical Information and Modeling, 2015, 55(2): 263-274.
[83] MAYR A, KLAMBAUER G, UNTERTHINER T, et al. DeepTox: Toxicity prediction using deep learning [J]. Frontiers in Environmental Science, 2016, 3: 80.
[84] PU L M, NADERI M, LIU T R, et al. *e*ToxPred: A machine learning-based approach to estimate the toxicity of drug candidates [J]. BMC Pharmacology and Toxicology, 2019, 20(1): 2.
[85] ALVES V, MURATOV E, CAPUZZI S, et al. Alarms about structural alerts [J]. Green Chemistry, 2016, 18(16): 4348-4360.
[86] WU X D, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining [J]. Knowledge and Information Systems, 2008, 14(1): 1-37.
[87] JIMÉNEZ-LUNA J, GRISONI F, SCHNEIDER G. Drug discovery with explainable artificial intelligence [J]. Nature Machine Intelligence, 2020, 2(10): 573-584.
[88] CHEN D, GAO K F, NGUYEN D D, et al. Algebraic graph-assisted bidirectional transformers for molecular property prediction [J]. Nature Communications, 2021, 12(1): 1-9.
[89] RODRÍGUEZ-PÉREZ R, BAJORATH J. Explainable machine learning for property predictions in compound optimization [J]. Journal of Medicinal Chemistry, 2021, 64(24): 17744-17752.