科学通报 2023 年 第 68 卷 第 17 期: 2184 ~ 2196

《中国科学》杂志社
SCIENCE CHINA PRESS

香山科学会议专栏 评 述

第716次学术讨论会·单分子科学与技术

面向分子科学的数据智能

李淹博[1,2], 江俊[2,3*], 罗毅[2,3*]

1. 材料科学姑苏实验室, 苏州 215123;
2. 中国科学技术大学化学与材料科学学院, 合肥 230026;
3. 合肥微尺度物质科学国家研究中心, 合肥 230026
* 联系人, E-mail: jiangj1@ustc.edu.cn; yiluo@ustc.edu.cn

2022-11-09 收稿, 2023-01-05 修回, 2023-01-09 接受, 2023-01-30 网络版发表
国家自然科学基金(22025304, 22033007)和中国科学院稳定支持基础研究领域青年团队计划(YSBR-005)资助

摘要 分子科学是化学的核心, 也是生物、材料、药学等学科的基础. 传统的分子科学研究通过实验或理论手段进行, 研究成本高、周期长, 难以处理高复杂度体系. 随着大数据时代的到来, 数据驱动的人工智能研究已成为继实验、理论和模拟之后的第4种科学研究范式. 数据驱动的机器学习凭借其快速高效的数据处理能力, 在分子科学领域展现出巨大的发展潜力. 尤其是在分子性质预测、分子设计、化学反应预测及逆合成、量子化学计算、自动化合成等领域获得了广泛应用. 本文首先介绍面向分子科学数据智能研究过程中的3个关键部分, 即分子科学开放数据集、分子描述符和机器学习算法; 然后, 列举机器学习在不同分子科学研究方向中的重要应用案例; 最后, 分析讨论该研究领域可能存在的挑战及潜在发展方向.

关键词 分子科学, 数据智能, 分子性质预测, 化学反应预测, 自动化合成

分子科学主要研究分子的结构、性质以及分子间的相互作用, 是化学、生物、材料、药学等学科的重要基础. 传统的实验研究手段无法对所有分子的性质及反应过程进行逐一探究. 薛定谔方程的提出为分子科学构建了清晰的结构-性质关系底层逻辑, 对于给定的化学结构, 通过求解薛定谔方程, 理论上可以准确描述电子的分布和广泛的物理响应. 但是薛定谔方程的直接求解极为复杂且计算量巨大, 这对复杂体系来说, 几乎不可能. 20世纪60年代, 随着Gaussian等基于量子力学求解薛定谔方程软件的出现[1], 降低了量子化学计算的门槛, 使得理论研究手段被广泛使用, 极大促进了分子科学领域的发展. 进入21世纪以来, 计算化学展现出更好的预见性, 其应用范围也更加广泛, 在结构化学、药物化学和材料科学领域扮演越来越重要的角色, 高通量计算筛选也成为常见的研究手段. 然而, 计算化学的瓶颈本质上是计算能力和资源的不足, 现代很多理论方法依赖于各种近似来大幅度降低计算量, 但计算结果与真实体系的差距不得而知.

随着大数据时代的到来, 大规模云计算机并行计算能力和存储能力显著提升, 高通量数据不断积累, 人工智能技术日趋成熟, 数据驱动的人工智能研究已成为继实验、理论和模拟之后的第4种科学研究范式. 近年来, 深度学习[2]——一种以神经网络为架构对数据进行表征学习的机器学习算法, 在计算机视觉、语音识别、自然语言处理与生物信息学等重要领域取得了惊人的进展[3~11]. 基于深度学习技术开发的人工智能围棋程序AlphaGo[12]和AlphaGo Zero的出现, 证实了机器学习方法可以快速处理规则清晰、数据高度复杂科学问题的可行性. 这种技术适用于处理涉及大量空间组合或非线性过程的复杂问题, 这些问题或不能通过传统

引用格式: 李淹博, 江俊, 罗毅. 面向分子科学的数据智能. 科学通报, 2023, 68: 2184–2196
Li Y B, Jiang J, Luo Y. Data intelligence for molecular science (in Chinese). Chin Sci Bull, 2023, 68: 2184–2196, doi: 10.1360/TB-2022-1152

© 2023《中国科学》杂志社

方法求解, 或只能牺牲巨大的计算成本进行求解. 在分子科学领域, 底层构效关系(元素周期表、物理常数、薛定谔方程)相对清晰且稳定, 但由于其原子和分子的组合自由度大、从微观量子态到化学性质的演化复杂度高, 使得该构效关系扑朔迷离. 这意味着通过数据驱动的人工智能机器学习手段开展分子科学研究具有很高的可行性.

机器学习在分子科学领域的应用可以追溯到20世纪70年代, Hiller等人[13]使用三层感知机成功将一系列烷基取代和醇烷基取代的1,3-二恶烷按药理活性或非活性进行分类. 20世纪90年代, 神经网络在计算机辅助药物设计中广泛使用, 特别是在定量结构-活性关系(quantitative structure-activity relationship, QSAR)研究中. 随着近些年化学信息学的发展, 机器学习在分子性质预测、分子设计、化学反应预测与逆合成、量子化学以及自动化合成等领域均有重要应用.

理解如何使用机器学习方法解决特定的分子科学问题, 需要介绍其中涉及的机制. 一个标准的机器学习工作流通常包括以下步骤: (1) 数据集构建, 包括数据获取、清理和预处理; (2) 分子描述符选取; (3) 模型建立(训练、验证和测试). 本文简要介绍这几个步骤及一些面向分子科学的数据智能研究案例, 并提出现有研究中可能存在的挑战与机遇.

1 数据集构建

可用数据的质量、数量和多样性给模型的准确性和通用性施加了上限. 数据可以是实验数据或计算数据, 可以来自公开来源, 也可以通过使用各种软件平台进行高通量计算自行生成. 所需的数据量取决于机器学习模型的选择, 但一般来说, 一个合理的模型至少需要50个数据点, 而某些模型(如神经网络)需要的数据量要大得多. 本节根据分子科学的几个重要研究类别将可公开获取的数据源进行如下分类, 对一些常用的数据源进行重点介绍, 并对所有常用数据源进行了列表统计(表S1).

1.1 分子基础信息

NIST(https://webbook.nist.gov/chemistry)是美国国家标准技术研究所的标准参考数据库, 是一个非常全面的气相分子数据库, 提供了热化学数据、凝相热化学数据、相变数据、反应热化学数据、相位离子能量数据、红外光谱数据、质谱数据、紫外可见光谱和气相色谱数据等信息. PubChem(https://pubchem.ncbi.nlm.nih.gov/)是美国国立卫生研究院的开放式化学数据库. 其条目涵盖分子的2D结构、3D结构、名称和识别代号、理化性质、相关记录、药品相关信息、食品相关信息、毒性信息和生物信息等方面. ChemSpider(http://www.chemspider.com/)是英国皇家化学会免费提供的化学结构数据库, 可快速访问超过6300万个结构, 提供分子的基本信息、理化性质、光谱等内容.

除了上述国外机构的数据库, 国内近年来也在分子基础信息数据库方面取得了较大进展. 机数大材库(http://www.dcaiku.com/)是由机数量子建立的综合性创新材料数据库平台. 它包含了9448万个分子基础化合物、40万个材料性能数据、90万个含磷化合物, 同时又扩展了1120万个化合物分子数据, 包含原子坐标参数、基本化学性质, 以及经计算获得的理化性质参数等. 该数据库整合了分子结构检索功能, 用户可以通过开放画图板, 直观快捷地进行结构搜索服务. 与此同时, 用户还可以通过其自主开发的Atom-Bond指纹生成算法, 利用查找迅速定位所寻找的结构. 网站还提供基于WebGL技术的分子3D模型展示, 用于实时交互式观察分子结构.

1.2 燃烧

RMG数据库(http://rmg.mit.edu/database/)是由麻省理工学院William H. Green团队建立的能够预测分子热力学、动力学、溶剂化效应和反应输运性质的数据库. RMG可以自动生成包括化学激活在内的压力依赖的反应机理. ATcT(https://atct.anl.gov/)是一种新的范式, 为稳定、有反应活性和瞬态的化学物质提供热化学值(如生成焓、吉布斯生成能、键解离能、反应焓等). 国内方面, 中国科学技术大学国家同步辐射实验室提供了众多小分子的电离能数据及光电离效率曲线数据库, 对燃烧产物的定性和定量分析十分重要.

1.3 大气、星际

KIDA(http://kida.astrophy.u-bordeaux.fr/)是一个气相天体化学数据库, 它包含了实验测量或理论计算得到的气相反应的反应速率常数. 除了提供星际介质分子的动力学信息外, KIDA还计划提供行星大气层和恒星环层的相关数据. MCM(http://mcm.york.ac.uk/home.htt)包含对流层降解过程中一系列初级挥发性有机化合物(volatile organic compound, VOC)的详细气相化学

反应过程, 其中包括降解甲烷和142种非甲烷VOC的相关数据. 同时, 它也提供了一些其他信息, 例如在VOC降解过程中形成的自由基和闭壳中间体的分布.

1.4 生物、蛋白质

PDBbind数据库(http://www.pdbbind.org.cn/)是蛋白质数据库中所有生物分子复合物相关联的实验测量数据集合, 提供了这些复合物的能量与结构信息之间的基本关系, 帮助科研人员进行分子识别、药物发现等方面的计算和统计研究. 该数据库每年更新一次, 实时跟进蛋白质数据库的发展. UniProt(https://www.uniprot.org/)是蛋白质序列和注释数据的综合资源数据库, 免费提供全面、高质量的蛋白质序列与功能信息, 其中很多内容来自基因组计划. 此外, 还包含了大量来自文献关于蛋白的生物学功能信息.

1.5 药品、有机

SDBS(http://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi)是一种用于有机化合物的综合光谱数据库系统, 其中包括电子轰击电离质谱(electron ionization-mass spectroscopy, EI-MS)、傅里叶变换红外光谱(Fourier-transform infrared spectroscopy, FT-IR)、^{1}H核磁共振谱(^{1}H nuclear magnetic resonance spectroscopy, ^{1}H NMR)、^{13}C NMR谱、激光拉曼光谱和电子自旋共振谱(electron spin resonance spectroscopy, ESR). SMPDB(http://smpdb.ca/)是一个交互式可视化数据库, 包含超过3万个人类小分子路径的信息. 其专门设计用于发展代谢组学、转录组学、蛋白质组学和系统生物学中的路径解释与发现, 支持全文、序列和化学结构搜索. 由公开数据集中获取的数据可能存在来自实验的测量误差或来自不同计算级别的误差的影响, 导致数据不统一或者包含缺失值、异常值、重复值和无用值. 数据库错误的数据和化学结构的不正确处理可能对机器学习模型预测能力产生重大影响[14]. 识别和消除这些错误对避免机器学习算法被误导至关重要. 因此, 在训练模型之前, 应确保所有数据经过清理和处理, 这是确保模型可重复的重要一步. 此外, 在模型运行的初始阶段, 可能会发现一些额外有问题的数据样本, 它们会对模型性能产生不利的影响. 在这种情况下, 应考虑在最终确定模型之前执行另一轮数据清理.

2 分子描述符

分子描述符是建立机器学习与化学科学联系的关键步骤, 是指分子在某一方面性质的度量, 既可以是通过各种算法推导出的表示分子结构的数值指标, 也可以是分子的物理化学性质. 分子描述符选取得越合适, 算法映射到输出数据的精确度就越高. 同时, 描述符应尽可能地具有物理意义, 以帮助我们寻找模型的内涵, 实现预测模型可解释. 分子描述符的选取是当前一个开放的研究问题, 虽然目前有许多种分子表示方法, 但是似乎没有一种表示适用于所有情况.

2.1 结构描述符

传统上, 我们用含有原子和键的结构图来表示分子. 在数据驱动的新范式中, 分子需要被表示成计算机可读可解释的格式. 图1展示了几种常见的分子结构表示方法[15]. 其中, 分子指纹是一种分子的向量化表示, 可以捕捉分子内部原子构型的精确细节. 常见的分子指纹有如下几种: (1) 基于子结构的分子指纹, 如MACCS[16]和PubChem指纹; (2) 基于拓扑和路径的分子指纹, 如Daylight指纹和Tree指纹; (3) 圆形分子指纹, 如Morgan指纹[17]、MolPrint2D[18]、ECFP指纹[19]、FCFP[19]等; (4) 混合指纹, 如UNITY 2D、MP-MFP[20]等.

简化分子线性输入规范(simplified molecular input line entry specification, SMILES)[21]是一种被广泛接受的分子简写方式. 在SMILES中, 原子符号用来表示原子, 脂肪族和芳香族原子分别用大写和小写符号表示, 双键和三键分别写成“=”和“#”, 支链用圆括号“()”表示. 通常氢原子不会被明确地显示出来. 通过给两个相连的环原子分配数字来表示环. 同一分子如果沿着不同的顺序可能会生成不同的SMILES字符串. 为了解决这个问题, 国际化学标识符(international chemical Identifier, InChI)[22]的提出为每个分子提供了唯一的标识符. SMARTS(SMILES arbitrary target specification)是在SMILES基础之上的改进版[23], SMARTS允许使用通配符表示原子和化学键, 因此被广泛用于化学数据库中的结构搜索.

分子图本质上是将分子中的原子和键分别映射到图的节点和边. 分子结构是典型的图结构, 分子图表示可以自然地描述具有丰富结构和空间信息的分子. 基于分子图表示的图神经网络模型(graph neural network, GNN)近几年发展得很快[24~27]. 库仑矩阵(Coulomb ma-

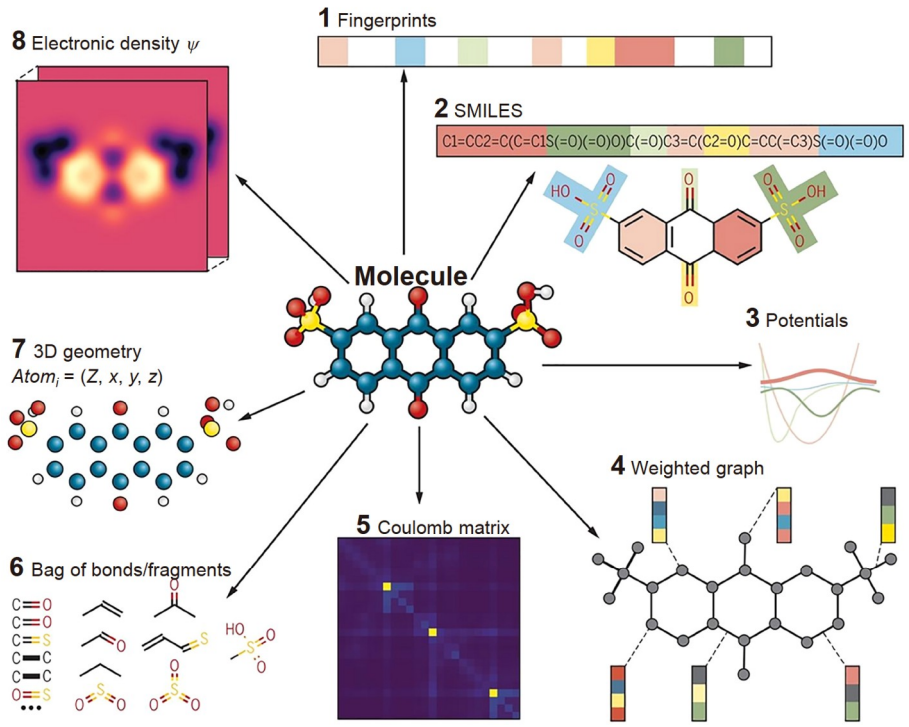


图 1 (网络版彩色)同一分子不同类型的分子表示[15]
Figure 1 (Color online) Different types of molecular representations applied to one molecule[15]

trix)[28]是一种特殊的分子图, 是用来表示分子内笛卡尔坐标系集合和核电荷的矩阵. 此外, 还有BOB(bag of bonds)[29]、BA(bonds-angles)[30]、BIM(bonds in molecule)[31]、BAND(bonds-angles-nonbonded-dihedrals)[32]等分子结构描述符.

2.2 理化性质描述符

上节提到的分子表示均是基于分子结构的直接表示方法, 在进行一些特定问题的机器学习时, 我们可以利用已有的专家知识, 选用具有物理意义的描述符, 从而降低机器学习的难度. 一个好的分子描述符应该至少满足以下3个标准: (1) 分子的独特描述; (2) 对目标属性敏感; (3) 容易获得. 一些基于量子力学的性质如电子能量、最高占据分子轨道(highest occupied molecular orbital, HOMO)、最低未占据分子轨道(lowest unoccupied molecular orbital, LUMO)的轨道能量等也是一种可用的描述符. 同样, 一些实验可观测值如电离能、亲核性等也是可供选择的描述符. 中国科学技术大学江俊团队[33]创新地提出偶极矩描述符, 成功预测CO在金属表面的吸附性质, 揭示了复杂体系的构效关系. 基于结构描述符建立了偶极矩预测模型, 高效模拟分子光谱, 并提出了谱学描述符, 建立关键性质的定量预测关系. 江俊团队[34]还将NMR化学位移和键能作为描述符, 运用到有机逆合成反应路径预测中.

分子描述符可以从前文提到的数据库中直接获取, 也可以利用一些专门的描述符计算软件如DRAGON[35]、Mordred[36]、PaDEL-Descriptor[37]、BlueDesc、ChemoPy[38]、PyDPI[39]、Rcpi[40]、Cinfony[41]、ChemDes[42]等获取. 或者通过一些量子化学软件如Gaussian等计算得到.

3 智能算法

当数据被收集并用恰当的描述符进行表示之后, 就需要选择一个模型来从中学习. 根据不同的数据类型和研究问题, 可以应用一系列不同的机器学习算法(图2). 从学习方式上分, 可以分为监督学习、无监督学习和强化学习. 后文将对分子科学研究中常用的监督学习和无监督学习算法进行介绍.

3.1 监督学习

监督学习是分子科学研究中使用较为广泛的一类算法, 比如用于分子性质的预测. 监督学习是基于已标

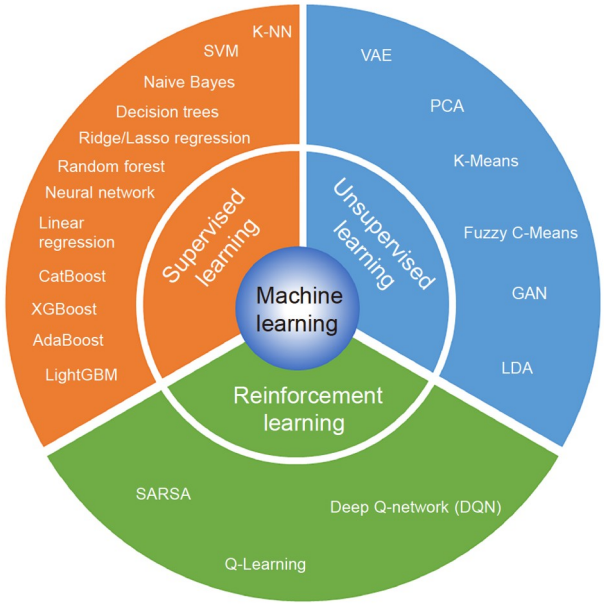


图 2 (网络版彩色)各种机器学习算法
Figure 2 (Color online) Various machine learning approaches

记的输入-输出对, 学习将输入映射到输出的函数. 该算法的目标是最小化学习过程中输入-输出值的差别. 监督学习模型可以在离散集(如药理活性或非活性分类)或连续集(如分子键能)内预测输出值. 前者建立的模型称为分类模型, 后者称为回归模型. 接下来简单介绍常见的监督学习模型.

朴素贝叶斯分类器[43]是一类假设各特征之间相互独立、运用贝叶斯定理进行分类的监督学习算法. 其优点是所需数据较少, 但是对数据的输入方式较为敏感, 只适用于分类问题. k-近邻算法[44]采用测量不同特征值之间的距离进行分类. 预测的输出值依赖于数据中k个"最近邻"的值, 其中k为整数. 该方法具有精度高、对异常值不敏感、无数据输入假定等优点. 它可以用在分类及回归问题中, 但是具有较高的计算复杂度和空间复杂度. 逻辑回归算法通过最优化算法寻找分类边界的最佳拟合参数, 其计算代价不高, 易于理解和实现, 但容易欠拟合, 分类精度不高. 支持向量机(support vector machine, SVM)[45]是一种二分类模型, 其基本思想是, 求解能够正确划分训练数据集并且几何间隔最大的分离超平面. 除了进行线性分类, SVM还可以使用核技巧(kernel trick)处理非线性分类, 通过将低维空间中的非线性问题转换成高维空间下的线性问题进行求解. 其计算量小、泛化错误率低、结果易解释, 但是对参数调节和核函数的选择敏感.

决策树[46]是用于确定行动过程或结果的流程图. 树的每个分支代表一个可能的决策、事件或反应. 这棵树的结构是为了显示一个选择如何以及为什么会导致下一个选择, 不同分支表明每个选择相互排斥. 决策树由根节点、叶节点和分支组成. 根节点是树的起点. 根节点和叶节点都包含要处理的问题或标准. 分支是连接节点的箭头, 显示从问题到答案的流程. 其计算复杂度不高, 输出结果易于理解, 对中间值的缺失不敏感, 可以处理不相关特征数据. 但是通常来说, 一棵简单的树并不能产生有效的结果, 随机森林算法应运而生. 随机森林是由许多决策树随机组合而成的森林, 将单个决策树的输出整合起来生成最后的输出结果.

神经网络是近年来用途最为广泛并且发展迅速的一类算法[47]. 并产生了许多分支如全连接神经网络、卷积神经网络、循环神经网络等. 神经网络分为输入层、隐藏层以及输出层三部分. 输入层单元接受外部输入的信号与数据, 隐藏层连接输入和输出层, 主要用于调整神经元之间的权重实现特征提取, 并将处理结果传递到输出层. 输出层单元被激活函数激活后, 对处理结果进行输出, 激活函数需要针对不同的研究问题进行选择. 神经网络技术已经越来越被认为是化学研究中强有力的工具.

3.2 无监督学习

虽然监督学习功能强大、应用广泛, 但其在分子科学领域的应用仍面临重大的挑战. 监督学习模型需要用大量的带标签的数据集来训练, 但是可直接获取的高质量带标签的数据较少, 这就使得可供学习的样本很少, 从而影响其预测效果. 无监督学习与之不同, 它所需要的输入数据集不需要带标签(输出值), 所以很好地克服了这个问题. 但是无监督学习算法简单, 功能有限, 因而现在只应用于分子科学的少数领域, 如聚类、特征提取等. 接下来, 对最常见的几种无监督学习算法进行概述.

(1) 变分自编码器(variational auto-encoder, VAE)[48]是一种随机生成模型, 其目的是尽可能地重建输入数据. 它由编码器、解码器和预测器三部分组成. 编码器可以将分子离散的字符串表示转换为连续的向量表示. 使用这种新的连续向量值表示, 可以通过随机向量解码、扰乱已知化学结构或在分子间插入等方法产生新的分子, 并通过优化算法产生特定性质的分子. 该方法在分子设计领域得到了广泛的应用.

(2) 主成分分析(principal component analysis, PCA)[49]是一种常见的对数据进行降维的无监督学习算法. 其目标是从输入数据中提取重要信息, 将其表示为一组称为主成分的线性不相关变量, 并将观察结果和变量的相似模式显示为映射中的点. PCA模型常用于查看分子的特征分布或特征数据的选取, 以加速或提高其他机器学习算法的预测效果.

(3) 生成对抗网络(generative adversarial network, GAN)[50]是一个快速发展的研究领域. GAN会同时训练两个模型: 生成器模型和鉴别器模型. 生成器模型用于学习真实样本, 并不断生成"假"数据; 鉴别器模型的目的是将这些"假"数据从真实样本中尽可能分辨出来; 而生成网络则要尽可能地"欺骗"鉴别网络. 两个网络相互对抗、不断调整参数, 最终目的是使鉴别网络无法判断生成网络的输出结果是否真实. 训练好的GAN具备很好的数据生成能力, 因而可以用于生成有特定功能的新分子, 在分子生成中被广泛使用.

4 数据智能在分子科学中的应用

4.1 分子性质预测

在整个化合物空间中准确而高效地预测分子性质是化学和制药工业中合理设计化合物的关键问题. 近几年, 由于深度学习的发展, 分子性质预测领域也在迅速发展, 包括分子键能、光谱、生物活性、药物的吸收、分布、代谢、排泄、毒性(absorption, distribution, metabolism, excretion and toxicity, ADMET)等性能以及与毒性靶标的相互作用等.

Hansen等人[29]提出了BOB模型, 实现对有机分子的原子化能、极化率以及分子前线轨道能量的预测. 该模型对平衡态分子有较好的预测效果, 同样也可以扩展到非平衡态分子. 江俊团队[51]提出了一种高效预测SARS-CoV-2红外光谱的机器学习模型, 并可以表征与蛋白质二级结构变化相关的红外光谱的细微变化, 为监测SARS-CoV-2与人类ACE2之间的实时相互作用提供了一种具有成本效益的工具. 江俊团队[52]还提出了基于二维光谱描述符的蛋白质二级结构机器学习识别方法, 驱动机器自动解读光谱信号并反演识别相应结构, 对同源和非同源蛋白片段的识别准确率分别达到了97%和91%. Paton团队[53]开发了一个GNN模型, 能够在几分之一秒内预测有机分子的键能, 其预测结果相对于密度泛函理论(density functional theory, DFT)结果的平均绝对误差(mean absolute error, MAE)仅为0.58 kcal/mol. Saini等人[54]采用数据驱动的方法, 从不同的亲核试剂和相关溶剂中提取27个量子力学分子和热力学描述符, 实现对亲核性值的预测. Lee等人[55]也利用机器学习, 开展了亲电性和亲核性的研究: 先用传统的描述符包括硬度、电子亲和能、电离能、电负性等进行预测, 发现效果不尽人意, 在添加分子指纹描述符之后, 预测效果得到改善.

4.2 分子设计

分子设计旨在虚拟地创建和分析具有相关优化性能的分子, 如可合成性、药物ADMET等. VAE、GAN或强化学习模型在分子设计特别是药物分子的反向设计领域被广泛地应用. Gómez-Bombarelli等人[48]研发了VAE, 通过优化潜伏空间中编码的分子向量, 成功生成了具有用户所需属性的分子. 随后, Kadurin等人[56]研发了深度生成对抗自编码器(generative adversarial autoencoders, AAE)来生成具有预定义抗癌特性的新分子, 该模型相较于VAE具备在生成分子指纹可调性方面的优势和处理大分子数据集的能力. 一些研究已经采用强化学习模型来生成类药物分子. Popova等人[57]在深度和强化学习方法的基础上设计了结构进化强化学习(reinforcement learning for structural evolution, ReLeaSE). 该模型集成了生成和预测两种深度神经网络, 首先, 生成模型和预测模型分别训练与监督学习算法; 然后, 这两种模型将与强化学习方法联合训练, 以倾向于生成那些具有所需物理或生物特性的新化学结构. 生成算法的组合模型在分子设计中也被广泛应用, 以充分利用每种算法的优点. 例如, druGAN[58]采用对抗性自编码器网络, RANC[59]采用强化学习和对抗性网络.

4.3 化学反应

4.3.1 正向预测

机器学习技术可以用于开发与实验直接相关的可观测数据的综合模型, 包括反应势垒预测[60~62]、反应速率常数预测[63,64]、量子反应速率的预测[65]以及化学反应产率预测[66]. 例如, Jorner等人[67]训练了一个高斯过程回归模型对亲核芳香取代反应的反应活化能进行训练, 并使用该模型来预测外部测试集, 预测结果的MAE值为0.77 kcal/mol. 麻省理工学院Green团队[60,62,68~77]在机器学习应用于小分子化学反应方向做

了许多杰出的工作. 例如, 构建了RMG数据库[75], 其能够预测反应的热力学、动力学、溶剂化效应和输运性质, 并自动生成反应机理; 构建了一个有向信息传递神经网络对24000个反应进行训练, 仅使用来自反应物和生成物的信息, 对反应势垒的预测效果相对于耦合簇方法计算势垒的MAE为2.6 kcal/mol[62]; 开发了开源统计力学软件Arkane[77]进行自动反应动力学和网络探索, 有助于计算化学物种的热力学性质、高压极限反应速率系数, 以及多势阱分子势能面上的压力相关速率系数, 包括碰撞能量转移对反应速率常数的影响等.

4.3.2 逆合成

实现逆向合成一直以来都是有机化学家的研究目标[78]. 在过去的几年里, 我们见证了机器学习在该领域的巨大进展[79~81]. Segler等人[79]将蒙特卡罗树搜索、扩展策略网络以及过滤器网络相结合用于搜索逆合成路线. 他们对1200万反应数据进行训练, 结果表明, 该模型比基于提取规则和手工设计启发式的传统计算机辅助搜索方法快30倍, 并且盲测的化学工作者很难将机器学习设计以及人类设计的逆合成路线区分开来. Wolos课题组[81]使用正向合成Allchemy平台, 从大约200种商业规模回收的废弃化学品中生成大规模合成网络, 从中检索出数万条通向大约300种重要药物和农药的路线. 他们还根据可持续化学的公认指标对这些合成进行算法排序, 并通过实验验证了其中几种路线的可行性. 江俊团队[34]利用化学信息分子图(chemistry-informed molecular graph, CIMG)作为描述符, 设计了逆合成机器学习模型. 以NMR化学位移作为顶点特征, 键解离能作为边特征, 溶剂和催化剂信息作为全局特征, 引入到分子图中来表示分子和反应, 构建了逆合成规划模型. 该模型利用基于分子图的神经网络和蒙特卡罗树搜索开发, 使用140万反应数据集进行训练, 在反应模板选择方面前50的反应模板准确率达到94%, 在催化剂预测方面达到了前10位的准确率达到93%, 在溶剂预测方面前10位的准确率达到89%.

4.4 计算化学

使用计算手段研究分子科学已经被大家广为接受, 并且占据越来越重要的地位. 但是随着体系规模的增加, 计算的复杂度迅速增长. 因此, 许多研究致力于对求解薛定谔方程做各种简化和近似, 使计算保持可接受的准确性和合理的计算时间. 机器学习的应用使得高精度和高复杂度的同时实现成为可能. 2017年, Müller团队[82]用机器学习训练密度和能量数据获得了丙二醛分子的密度泛函模型, 随后结合分子动力学模拟成功描述了丙二醛分子内质子转移过程, 这验证了利用机器学习构造泛函并绕开传统Kohn-Sham方程自洽场迭代求解电子结构的可行性. 中国科学院大连化学物理研究所张东辉团队、复旦大学徐昕团队[83]在高通量量子化学计算数据的基础上, 成功应用人工智能技术预测了分子势能面. 复旦大学刘智攀团队[84]结合表面随机行走算法和神经网络技术训练分子和材料的势能面数据, 大幅度加速了材料结构演化和化学反应的预测.

4.5 自动化合成

除了前文介绍的针对特定化学问题的机器学习研究案例之外, 科学家基于大数据和人工智能深入量子力学底层提炼构效关系, 研制了能够自主探索化学反应性的机器化学家. 真正实现从文献数据抓取, 到实验方案设计, 再到自主完成合成、表征、测试以及数据分析迭代的全流程自动化操作. 自动化合成的实现为合成化学的逆向反应预测、功能材料的逆向设计、生物大分子的结构自动化反演提供高质量数据库、智能算法和软件引擎. 2019年, 美国麻省理工学院Jamison和Jensen团队[85]将逆向合成预测算法与机器人可重构流动装置配对, 开发了基于人工智能规划的有机化合物流动合成机器人平台, 实现了自主化学合成的里程碑. 2020年, 英国格拉斯哥大学Cronin团队[86]研发了自主化学机器人, 通过探索超过109种可能反应的化学空间来发现超分子结构. 同年, 英国利物浦大学Cooper团队[87]设计了第一台可移动机器化学家, 它能够比人类更快地执行实验, 并能够使用贝叶斯优化自主选择光催化剂. 机器科学家在生物大分子结构研究、自动化药物发现方面也初露锋芒. 2021年, 瑞典查尔姆斯理工大学[88]研制机器科学家“夏娃”用于识别和测试针对新型冠状病毒的药物. 瑞士苏黎世联邦理工学院[89]开发人工智能虚拟化学家, 筛选与天然产物具有相似功能但易于合成的分子, 并设计新分子的合成路线, 实现药物发现分子层面的逆向工程.

2022年, 中国科学技术大学科研团队[90]研制出数据智能驱动的全流程机器化学家——小来. 小来集成了移动机器人、化学工作站、智能操作系统、科学大数据库于一身, 其优势在于: 装载有物理模型、全流程自主实验、数据分析理实交融、实验结果自主优化迭

代, 可以快速地实现全局搜索、提炼出理实交融的机器学习模型. 这个模型融合了底层物理规律和化学实验演化, 具有指导实践的预测能力. 它的工作流程如下(图3)[90]: 首先, 针对提出的科学问题通过自然语言处理系统, 从云数据库中读取文献或专利, 整理成机器可以理解的结构化数据库, 并提出假设和设计试验方案; 然后, 根据提出的方案, 在室内或在线控制移动机器人, 可以自动调用计算和实验测试平台, 进行自主的量子化学模拟并执行实验合成、表征、测试等14个工作站的完整实验过程, 采集标准化实验数据; 最后, 利用其计算大脑通过机器学习和贝叶斯优化同时分析实验数据, 为下一次迭代提出新的假设, 实现理论与实验数据的交融. 小来目前已经被应用于寻找具有聚集诱导发射(aggregation-induced emission, AIE)特性的生物兼容发光团. 在这个实验中, 小来的机器阅读模块首先被激活, 从15979篇论文中发现了4865个分子与“AIE”、“aggregation-induced emission”或者“aggregation induced emission”关键词有关. 在基于专业知识规则的数据清理后, 确定了306个分子为商业上可用的AIE荧光团. 其中, Berberine chloride(BBR)是唯一一种发射波长位于可见光区域的化合物, 因此选择该化合物进行进一步研究. 小来紧接着进行了一系列自动化实验, 以合成不同的BBR溶液并测量其荧光. 先通过固体点胶工作站称量出适量的BBR固体样品, 然后转移到液体点胶工作站溶解成溶液, 并利用光致发光光谱和紫外光谱测量了溶液的光学性质. 通过对比不同浓度和溶剂的BBR溶液, 确定了BBR的最佳浓度为20 mmol/L. 除此之外, 小来还被应用于优化金属氧化物光催化剂的氢掺杂策略等课题研究, 在实验中显示出了较好的准确性和重复性, 显著缩短了实验耗时, 展现出在电催化剂、光催化剂和发光材料领域的多功能性.

5 总结与展望

综上所述, 数据驱动的机器学习为规则清晰而演

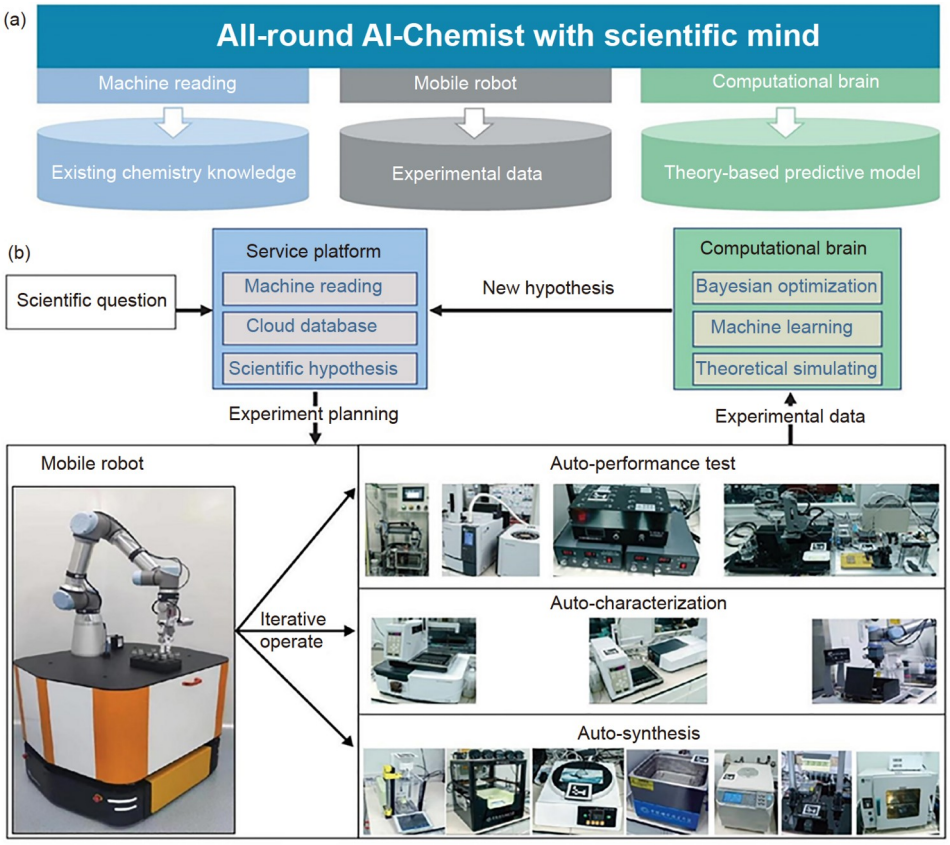


图 3 (网络版彩色)具有科学头脑的全方位人工智能化学家设计[90]. (a) AI-Chemist的3个模块: 机器阅读模块、移动机器人模块和计算大脑模块. (b) 人工智能化学家的工作流程和各个模块的功能

Figure 3 (Color online) Design of the all-round AI-Chemist with a scientific mind[90]. (a) Three modules of the AI-Chemist: A machine-reading module, a mobile robot module and a computational brain module. (b) The workflow of the AI-Chemist and the functions of each module

化复杂的分子科学研究提供了新工具，在分子性质预测、分子设计、有机反应预测及逆合成、量子化学计算、自动化合成等领域获得了广泛应用. 对于分子科学中特定的机器学习研究，想要获得准确的机器学习模型，有几点至关重要: 构建高质量的化学数据集; 将化学语言转化成机器可读的数字语言，最好是蕴含物理化学信息的描述符; 搭配合适的机器学习算法. 但是由于机器学习自身的特点，在分子科学领域的发展也受到了一定的限制，同时也说明该领域还存在许多进一步突破的机会. 本文提出几点该领域的潜在挑战与机遇:

首先, 最大的挑战就是缺乏高质量数据集. 对高质量的机器学习模型来说，许多现有的实验数据库要么太小，要么太不统一(如不同的测量技术或实验条件). 大量的实验数据只能在期刊上发表，不方便直接下载使用. 为应对这一难题，在未来可以从以下3个方面入手: (1) 数据及模型的开放共享. 现在许多机器学习所用数据是通过高通量计算得到的，研究者应尽量将数据和模型公开, 以便于建立更多的标准化学数据库, 有利于进行基准测试并促进机器学习算法的开发. 在这一过程中，应尽量做到化学及计算机语言标准化. (2) 电子实验记录本的普及. 由于从文献或者数据库中下载的实验数据往往是实验成功的案例，这样会损失掉很多不成功案例中的有效信息，导致数据带有一定的偏见. 另外, 不同课题组或者不同化学工作者对同一类实验的记录可能存在维度不统一、数据缺失的问题. 电子实验记录本的普及可以保证数据最大程度地保留, 以及维度的统一，并便于公开和共享. (3) 构建可从小数据集中充分挖掘信息的机器学习模型. 例如, 可以通过理论数据产生可解释的预训练模型，再依托实验小数据做迁移学习. 另一个有希望的方案是元学习[91], 即首先让机器进行学习，使其具备分析和解决问题的能力, 再使机器学习模型可以更好地泛化到新领域中, 从而实现“学会学习”. 神经图灵机[92]和模仿学习[93]等新概念的提出使这一过程得以实现. 此外, 文献[94]表明, 贝叶斯框架可以在数据有限的情况下，在一次性分类学习问题上实现与人类水平相当的性能，这将对数据稀疏、通常昂贵且获取缓慢的分子科学有一定帮助.

其次, 构建可解释的机器学习模型. 机器学习的过程是一个黑盒子，内部进行的数学运算过程我们无从知晓，这导致利用机器学习建立的模型是否具有实际的物理化学意义有待考量，也限制了进一步对其进行调试和改进. 上海大学Ouyang等人[95]提出基于压缩感知的SISSO算法, 用于在大量候选描述符中找出最佳的低维描述符，并且可以帮助获得机器学习模型更好的可解释性. 如何构建具有清晰规则的机器学习模型是未来一个重要的发展方向，通过化学家已有的化学知识构建可解释的机器化学模型，可解释模型具有较好的鲁棒性、容错性和较好的迁移能力，可以再反向重构我们对化学的认识.

随着越来越多的科学家利用数据驱动的机器学习模型开展分子科学领域的相关研究，该领域的底层原理越来越清晰. 基于数据智能的研究范式已经开始改变我们发现分子和研究分子科学的方式，在更大的数据和模型开放共享的支持下，可能会有颠覆性的新发现.

参考文献

1 Frisch M J, Trucks G W, Schlegel H B, et al. Gaussian 09. Wallingford: Gaussian, Inc., 2009
2 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
3 Bileschi M L, Belanger D, Bryant D H, et al. Using deep learning to annotate the protein universe. Nat Biotechnol, 2022, 40: 932–937
4 Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: A review. Neurocomputing, 2016, 187: 27–48
5 Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform, 2017, 18: 851–869
6 Zhang Z, Cui P, Zhu W. Deep learning on graphs: A survey. IEEE Trans Knowl Data Eng, 2022, 34: 249–270
7 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016. 770–778
8 Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag, 2012, 29: 82–97
9 Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks. In: International Conference on Acoustics Speech and Signal Processing ICASSP. New York: IEEE, 2013. 6645–6649
10 Bastings J, Titov I, Aziz W, et al. Graph convolutional encoders for syntax-aware neural machine translation. 2017, arXiv: 1704.04675

11 Marcheggiani D, Bastings J, Titov I. Exploiting semantics in neural machine translation with graph convolutional networks. 2018, arXiv: 1804.08313
12 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550: 354–359
13 Hiller S A, Golender V E, Rosenblit A B, et al. Cybernetic methods of drug design. I. Statement of the problem—The perceptron approach. Comput BioMed Res, 1973, 6: 411–421
14 Young D, Martin T, Venkatapathy R, et al. Are the chemical structures in your QSAR correct? QSAR Comb Sci, 2008, 27: 1337–1345
15 Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science, 2018, 361: 360–365
16 Durant J L, Leland B A, Henry D R, et al. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci, 2002, 42: 1273–1280
17 Morgan H L. The generation of a unique machine description for chemical structures—A technique developed at chemical abstracts service. J Chem Doc, 1965, 5: 107–113
18 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 Chem Inf Comput Sci, 2004, 44: 1708–1718
19 Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model, 2010, 50: 742–754
20 Xue L, Godden J W, Stahura F L, et al. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. J Chem Inf Comput Sci, 2003, 43: 1151–1157
21 Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Model, 1988, 28: 31–36
22 Heller S, McNaught A, Stein S, et al. InChI—The worldwide chemical structure identifier standard. J Cheminform, 2013, 5: 7
23 Jeliazkova N, Kochev N. AMBIT-SMARTS: Efficient searching of chemical structures and fragments. Mol Inf, 2011, 30: 707–720
24 Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst, 2021, 32: 4–24
25 Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model, 2019, 59: 3370–3388
26 Wu Z, Ramsundar B, Feinberg E N, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci, 2018, 9: 513–530
27 Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). Ann Statist, 2000, 28: 337–407
28 Rupp M, Tkatchenko A, Müller K R, et al. Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett, 2012, 108: 058301
29 Hansen K, Biegler F, Ramakrishnan R, et al. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. J Phys Chem Lett, 2015, 6: 2326–2331
30 Huang B, von Lilienfeld O A. Understanding molecular representations in machine learning: The role of uniqueness and target similarity. J Chem Phys, 2016, 145: 161102
31 Yao K, Herr J E, Brown S N, et al. Intrinsic bond energies from a bonds-in-molecules neural network. J Phys Chem Lett, 2017, 8: 2689–2694
32 Laghuvarapu S, Pathak Y, Priyakumar U D. BAND NN: A deep learning framework for energy prediction and geometry optimization of organic small molecules. J Comput Chem, 2020, 41: 790–799
33 Wang X, Ye S, Hu W, et al. Electric dipole descriptor for machine learning prediction of catalyst surface-molecular adsorbate interactions. J Am Chem Soc, 2020, 142: 7737–7743
34 Zhang B, Zhang X, Du W, et al. Chemistry-informed molecular graph as reaction descriptor for machine-learned retrosynthesis planning. Proc Natl Acad Sci USA, 2022, 119: e2212711119
35 Mauri A, Consonni V, Pavan M, et al. DRAGON software: An easy approach to molecular descriptor calculations. MATCH Commun Math Comput Chem, 2006, 56: 237–248
36 Moriwaki H, Tian Y S, Kawashita N, et al. Mordred: A molecular descriptor calculator. J Cheminform, 2018, 10: 4
37 Yap C W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem, 2011, 32: 1466–1474
38 Cao D S, Xu Q S, Hu Q N, et al. ChemoPy: Freely available python package for computational biology and chemoinformatics. Bioinformatics, 2013, 29: 1092–1094
39 Cao D S, Liang Y Z, Yan J, et al. PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. J Chem Inf Model, 2013, 53: 3086–3096
40 Cao D S, Xiao N, Xu Q S, et al. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. Bioinformatics, 2015, 31: 279–281
41 O'Boyle N M, Hutchison G R. Cinfony-combining open source cheminformatics toolkits behind a common interface. Chem Cent J, 2008, 2: 24
42 Dong J, Cao D S, Miao H Y, et al. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform, 2015, 7: 60

43 Hand D J, Yu K. Idiot's Bayes—Not so stupid after all? Int Stat Rev, 2001, 69: 385–398
44 Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theor, 1967, 13: 21–27
45 Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995, 20: 273–297
46 Quinlan J R. Induction of decision trees. Mach Learn, 1986, 1: 81–106
47 Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw, 2015, 61: 85–117
48 Gómez-Bombarelli R, Wei J N, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci, 2018, 4: 268–276
49 Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics Intell Lab Syst, 1987, 2: 37–52
50 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM, 2020, 63: 139–144
51 Ye S, Zhang G, Jiang J. AI-based spectroscopic monitoring of real-time interactions between SARS-CoV-2 and human ACE2. Proc Natl Acad Sci USA, 2021, 118: e2025879118
52 Ren H, Zhang Q, Wang Z, et al. Machine learning recognition of protein secondary structures based on two-dimensional spectroscopic descriptors. Proc Natl Acad Sci USA, 2022, 119: e2202713119
53 St John P C, Guan Y, Kim Y, et al. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. Nat Commun, 2020, 11: 2328
54 Saini V, Sharma A, Nivatia D. A machine learning approach for predicting the nucleophilicity of organic molecules. Phys Chem Chem Phys, 2022, 24: 1821–1829
55 Lee B, Yoo J, Kang K. Predicting the chemical reactivity of organic materials using a machine-learning approach. Chem Sci, 2020, 11: 7813–7822
56 Kadurin A, Aliper A, Kazennov A, et al. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget, 2016, 8: 10883–10890
57 Popova M, Isayev O, Tropsha A. Deep reinforcement learning for *de novo* drug design. Sci Adv, 2018, 4: eaap7885
58 Kadurin A, Nikolenko S, Khrabrov K, et al. druGAN: An advanced generative adversarial autoencoder model for *de novo* generation of new molecules with desired molecular properties *in silico*. Mol Pharm, 2017, 14: 3098–3104
59 Putin E, Asadulaev A, Ivanenkov Y, et al. Reinforced adversarial neural computer for *de novo* molecular design. J Chem Inf Model, 2018, 58: 1194–1204
60 Grambow C A, Pattanaik L, Green W H. Deep learning of activation energies. J Phys Chem Lett, 2020, 11: 2992–2997
61 Lewis-Atwell T, Townsend P A, Grayson M N. Machine learning activation energies of chemical reactions. WIREs Comput Mol Sci, 2022, 12: e1593
62 Spiekermann K A, Pattanaik L, Green W H. Fast predictions of reaction barrier heights: Toward coupled-cluster accuracy. J Phys Chem A, 2022, 126: 3976–3986
63 Houston P L, Nandi A, Bowman J M. A machine learning approach for prediction of rate constants. J Phys Chem Lett, 2019, 10: 5250–5258
64 Nandi A, Bowman J M, Houston P. A machine learning approach for rate constants. II. Clustering, training, and predictions for the $O(^3P) + HCl \rightarrow OH + Cl$ reaction. J Phys Chem A, 2020, 124: 5746–5755
65 Komp E, Valleau S. Machine learning quantum reaction rate constants. J Phys Chem A, 2020, 124: 8607–8613
66 Schwaller P, Vaucher A C, Laino T, et al. Prediction of chemical reaction yields using deep learning. Mach Learn-Sci Technol, 2021, 2: 015016
67 Jorner K, Brinck T, Norrby P O, et al. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. Chem Sci, 2021, 12: 1163–1175
68 Coley C W, Barzilay R, Jaakkola T S, et al. Prediction of organic reaction outcomes using machine learning. ACS Cent Sci, 2017, 3: 434–443
69 Gao H, Struble T J, Coley C W, et al. Using machine learning to predict suitable conditions for organic reactions. ACS Cent Sci, 2018, 4: 1465–1476
70 Coley C W, Jin W, Rogers L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci, 2019, 10: 370–377
71 Ahuja K, Green W H, Li Y P. Learning to optimize molecular geometries using reinforcement learning. J Chem Theor Comput, 2021, 17: 818–825
72 Chung Y, Vermeire F H, Wu H, et al. Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy. J Chem Inf Model, 2022, 62: 433–446
73 Coley C W, Green W H, Jensen K F. Machine learning in computer-aided synthesis planning. Acc Chem Res, 2018, 51: 1281–1289
74 Grambow C A, Li Y P, Green W H. Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. J Phys Chem A, 2019, 123: 5826–5835
75 Johnson M S, Dong X, Grinberg Dana A, et al. RMG database for chemical property prediction. J Chem Inf Model, 2022, 62: 4906–4915
76 Spiekermann K, Pattanaik L, Green W H. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. Sci Data, 2022, 9: 417

77 Dana A G, Johnson M, Allen J, et al. Automated reaction kinetics and network exploration (Arkane): A statistical mechanics, thermodynamics, transition state theory, and master equation software. ChemRxiv, 2022, doi: 10.26434/chemrxiv-2022-4klsm
78 Corey E J. The logic of chemical synthesis: Multistep synthesis of complex carbogenic molecules (Nobel Lecture). Angew Chem Int Edit, 1991, 30: 455–465
79 Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 2018, 555: 604–610
80 Mikulak-Klucznik B, Gołębiowska P, Bayly A A, et al. Computational planning of the synthesis of complex natural products. Nature, 2020, 588: 83–88
81 Wołos A, Koszelewski D, Roszak R, et al. Computer-designed repurposing of chemical wastes into drugs. Nature, 2022, 604: 668–676
82 Brockherde F, Vogt L, Li L, et al. Bypassing the Kohn-Sham equations with machine learning. Nat Commun, 2017, 8: 872
83 Su N Q, Chen J, Xu X, et al. Quantum reaction dynamics based on a new generation density functional and neural network potential energy surfaces. Acta Phys Chim Sin, 2016, 32: 119–130
84 Ma S, Huang S D, Liu Z P. Dynamic coordination of cations and catalytic selectivity on zinc-chromium oxide alloys during syngas conversion. Nat Catal, 2019, 2: 671–677
85 Coley C W, Thomas III D A, Lummiss J A M, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science, 2019, 365: eaax1566
86 Porwol L, Kowalski D J, Henson A, et al. An autonomous chemical robot discovers the rules of inorganic coordination chemistry without prior knowledge. Angew Chem Int Edit, 2020, 59: 11256–11261
87 Burger B, Maffettone P M, Gusev V V, et al. A mobile robotic chemist. Nature, 2020, 583: 237–241
88 Alalam H, Sigurdardóttir S, Bourgard C, et al. A genetic trap in yeast for inhibitors of SARS-CoV-2 main protease. Msystems, 2021, 6: e01087-21
89 Friedrich L, Cingolani G, Ko Y H, et al. Learning from nature: From a marine natural product to synthetic cyclooxygenase-1 inhibitors by automated *de novo* design. Adv Sci, 2021, 8: e2100832
90 Zhu Q, Zhang F, Huang Y, et al. An all-round AI-Chemist with a scientific mind. Natl Sci Rev, 2022, 9: nwac190
91 Jankowski N, Duch W, Grabczewski K. Meta-learning in Computational Intelligence: Vol. 358. Berlin, Heidelberg: Springer, 2011
92 Graves A, Wayne G, Danihelka I. Neural turing machines. 2014, arXiv: 1410.5401
93 Duan Y, Andrychowicz M, Stadie B, et al. One-shot imitation learning. In: Guyon I, Luxburg U V, Bengio S, eds. Advances in Neural Information Processing Systems. California: Neural Information Processing Systems (NIPS), 2017, 30: 1087–1098
94 Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. Science, 2015, 350: 1332–1338
95 Ouyang R, Curtarolo S, Ahmetcik E, et al. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Phys Rev Mater, 2018, 2: 083802

补充材料

表S1 现有分子数据库汇总表

本文以上补充材料见网络版csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

Summary for “面向分子科学的数据智能”

Data intelligence for molecular science

Yanbo Li[1,2], Jun Jiang[2,3*] & Yi Luo[2,3*]

[1] GuSu Laboratory of Materials, Suzhou 215123, China;
[2] School of Chemistry and Materials Science, University of Science and Technology of China, Hefei 230026, China;
[3] Hefei National Research Center for Physical Sciences at the Microscale, Hefei 230026, China
* Corresponding authors, E-mail: jiangj1@ustc.edu.cn; yiluo@ustc.edu.cn

Molecular science is the core of chemistry, but also the basis of biology, materials, pharmacy and other disciplines. Traditional molecular science research was carried out by experimental or theoretical means, which is costly, time-consuming and hard to solve the high complexity system. With the advent of the era of big data, data-driven innovation of artificial intelligence (AI) has become the fourth research paradigm after experiment, theory and simulation. With its fast and efficient data processing capabilities, data-driven machine learning has shown great potential in molecular science research.

A standard machine learning workflow usually includes three steps: data set construction, molecular descriptor selection, and model building. Firstly, the quality, quantity, and diversity of available data impose an upper limit on the accuracy and generality of the model. In Chapter 1, we classify publicly available data sources according to the research categories of molecular science, including molecular basic information, combustion, atmosphere and interstellar, biological, protein, pharmaceutical and organic. The contents, characteristics and obtaining methods of each database are introduced in detail. Secondly, descriptors are the key step to connect machine learning and chemical science. A good molecular descriptor should satisfy at least three criteria: a unique description of a molecule, sensitive to target attributes and easy to obtain. At the same time, descriptors should have physical significance to help find the connotation of the model and realize the interpretation of the model. In Chapter 2, we introduce some widely used structural descriptors and physical or chemical properties descriptors. Then, once the data has been collected and represented with the appropriate descriptor, a model needs to be selected according to the data type and research question. In Chapter 3, supervised and unsupervised learning algorithms are introduced which are widely used in molecular science.

After introducing three important steps of machine learning, Chapter 4 presents its application in several subarea of molecular science research. For example, in molecular property prediction, machine learning has been used to predict the atomization energy, polarizability, electron affinity energy, ionization energy, electronegativity, etc. In molecular design, variational auto-encoder (VAE), generative adversarial network (GAN) and reinforcement learning model are widely used, especially in reverse design of drug molecules. In chemical reaction, machine learning has been used to predict reaction barrier, reaction rate constant, quantum reaction rate and chemical reaction yield, and it has also made great strides in backward synthesis. In theoretical chemistry, by training the potential energy surface of molecules and materials, the structural evolution of materials and the prediction of chemical reactions are greatly accelerated. Furthermore, based on big data and AI, robotic chemists which can explore chemical reactivity and achieve automatic synthesis were developed. It is worth mentioning that Chinese scientists have made breakthroughs in this field.

In summary, data-driven machine learning provides new tools for molecular science with clear rules and complex evolution. However, it also faces some challenges and opportunities. Firstly, the biggest challenge is the lack of high-quality data sets. Here we propose three possible solutions: open sharing of data and models, popularization of electronic experimental record books, and construction of models that can fully mine information from small data sets. Secondly, how to build explicable machine model which has good fault tolerance and good transferable ability, and can reverse reconstruct our understanding of chemistry. As more and more scientists use data-driven machine learning models in molecular science, the underlying principles are clearer. The new research paradigm is changing the way we discover molecules and study molecular science, which could lead to disruptive new discoveries.

molecular science, data intelligence, molecular properties prediction, chemical reaction prediction, automatic synthesis

doi: 10.1360/TB-2022-1152