中国科学: 化学 2023 年 第 53 卷 第 5 期: 799 ~ 810
SCIENTIA SINICA Chimica chemcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

专题论述

机器化学家的挑战和机遇

江俊[1], 李淹博[1], 沈祥建[2*], 高飞雪[2*]

1. 中国科学技术大学化学与材料科学学院, 合肥 230026
2. 国家自然科学基金委员会化学部, 北京 100085
*通讯作者, Email: shenxj@nsfc.gov.cn; gaoxf@nsfc.gov.cn

收稿日期: 2023-03-02; 接受日期: 2023-04-14; 网络版发表日期: 2023-04-23

摘要 本文总结了国家自然科学基金委资助下, 中国科学技术大学线上举办的“机器科学家青年论坛”战略研讨会的学术交流内容和研讨成果. 本文主要聚焦在机器科学家领域的一个重要分支——机器化学家. 首先介绍了国内外机器化学家领域的最新研究进展, 即从硬件设施、智能化程度、实验能力等多层面进行了对比; 接着探讨了机器化学家领域的未来发展趋势, 提出5个层级的发展模式; 最后, 说明了机器化学家领域未来面临的一些重要挑战和机遇, 即需构筑精准数据平台, 设计针对自然科学的描述符和AI算法, 以及构建统一的机器指令集、模板和模型库. 此外, 还凝练了机器化学家领域拟解决的重大关键科学问题, 提出了加强学科发展布局的战略性资助建议.

关键词 机器化学家, 人工智能, 合成化学自动化, 材料逆向设计

1 引言

近年来, 由于大规模云计算机计算能力和存储能力的飞速提升、高通量数据的不断积累、先进算法的日趋成熟, 人工智能进入了新一轮的发展热潮, 尤其被广泛应用于计算机视觉、语音识别、自然语言处理、生物信息学、无人驾驶、无人机、智能营销、网络防御、决策管理等高科技前沿领域[1~5]. 以化学这一重要自然基础学科为例, 传统研究范式常常通过试错和变量降维的方式来建立模型, 但由于研究对象复杂化和高维化, 研究效率低和信息丢失的问题变得日益严重. 人工智能擅长对高复杂度、高维度数据进行挖掘和分析, 能够从海量数据中寻找变量之间的“隐藏”关联, 发现物质科学的内在规律[6,7]. 因此, 人工智能被认为继实验、理论、模拟之后的第四大研究范式. 近期, 人工智能在化学科学基础研究中取得了不少突破[8~12]. 例如, 德国明斯特大学Segler等人[13]结合蒙特卡洛树搜索与扩展策略网络, 通过对1200万个反应数据进行训练, 搜索到合适的逆合成路径, 加速了计算机辅助合成路线设计的发展. 复旦大学刘智攀教授[14]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. 美国西北大学Wolverton等人[15,16]搭建了量子材料数据库, 通过数据挖掘算法实现对锂离子电池催化剂和电极材料、镁合金结构和新型三组分材料的预测和设计.

通常而言, 化学科学的实验数据具有碎片化、标准不统一、格式不统一等特点, 而理论数据因大幅采

引用格式: Jiang J, Li Y, Shen X, Gao F. Challenges and opportunities of machine chemists. *Sci Sin Chim*, 2023, 53: 799–810, doi: 10.1360/SSC-2023-0044

用各种近似会引入难以消除的系统误差. 然而, 人工智能对可用数据的数量、质量和多样性均提出了高要求. 因此, 以机器化学家为代表的机器科学家凭借其产生数据可重复性、高效性的特点逐渐受到研究者的重点关注. 当前, 不少世界知名科研院所与企业在机器化学家领域取得了重要进展, 例如瑞典查尔姆斯大学、英国格拉斯哥大学、英国利物浦大学、美国麻省理工学院、美国伊利诺伊大学、美国北卡罗莱州立大学、加拿大英属哥伦比亚大学、德国赫姆霍兹研究所、美国制药企业默克公司和辉瑞公司, 以及北京大学、中国科学技术大学等. 这些团队开发研制的机器化学家虽已经帮助解决了不少科学难点, 但其发展依然存在着一些瓶颈, 如缺乏一个具有化学知识, 能够引入理论和物理模型, 可挖掘整合新旧知识并从中提取新规律并进行理性预测的智慧核心. 具有智慧核心的机器化学家, 能够借助大数据和人工智能深入量子力学底层提炼出构效关系, 为合成逆向预测、材料逆向设计、生物结构功能信息反演提供数据库、智能算法和软件引擎. 这一重要目标的实现, 需要我们开发物质科学研究的数据挖掘工具, 建设精度可控、可扩展的化学、材料和生物数据库; 开发知识图谱, 实现科学家与机器的人机接口; 发展针对化学、材料和生物问题的复杂性进行解耦合处理的人工智能算法; 针对具体问题开发高性能计算平台和模拟软件. 因此, 为了及时凝练机器化学家相关领域的最新研究进展和关键科学问题, 在国家自然科学基金委员会资助下, 中国科学技术大学于2022年12月在线举办了“机器科学家青年论坛”, 邀请了各高校和研究院所数十位从事理论化学、量子计算、生命科学、软件、人工智能等研究方向的专家学者参加. 此次战略研讨会主要聚焦在机器科学家领域的一个重要分支——机器化学家. 讨论了国内外机器化学家的最新研究成果, 分析了机器化学家领域的未来发展趋势, 阐述了机器化学家面临的挑战与机遇. 最后还凝练了机器化学家领域近5~10年的重大关键科学问题和技术问题.

1.1 国内外机器化学家领域研究进展

当前, 机器化学家主要被广泛应用于新型分子合成和功能材料创制两个方面. 国外方面, 世界上第一台机器化学家于2009年由瑞典查尔姆斯大学科学家King团队[17]设计制造并命名为“亚当”(图1a). 它被设计用于功能基因组学研究, 可以自主生成关于酿酒酵母的功能基因组学假设, 并通过实验室自动化对这些假设进行实验验证. 随后该团队于2015年研制出第二台机器化学家“夏娃”[18,19], 用于药物合成领域, 并以高通量和低成本的方式智能筛选治疗疟疾和其他热带疾病的药物. 英国格拉斯哥大学Cronin团队[20~22]在2018年设计制造了一台有机合成机器人, 并在之后不断对其升级改造. 在2022年开发了集文献阅读、实验方案定制、化合物合成和表征于一体的自动化系统Chemputer[21](图1b). 该系统可以将文献中的化学合成操作转化为机器人可读可自动执行的化学描述语言(χDL), 同时这些χDL被存储在内置数据库中. 目前, 该数据库储存了103种化学反应的χDL, 其中53种已经被实验验证. 通过系统配备的核磁共振谱、色谱、质谱对合成的化学物质进行表征, 其产量和纯度与文献中相当. 该团队还研发了便携式自动化学合成平台, 包含合成纯化所需模块[23]. 美国伊利诺伊大学M. Burke团队[24]设计制造了一台化学合成自动化机器(图1c), 实现了复杂3D结构分子的自动化合成流程. 美国麻省理工学院T. F. Jamison和K. F. Jensen团队[25]在2018年设计了即插即用式单元操作的自动化合成平台(如图1d), 并于2019年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[26]. 该系统配备了来自Reaxys反应数据库和美国专利商标局的数百万个反应数据库, 通过神经网络进行训练, 为给定分子提出合成路线, 自动选取简单便宜的起始原料, 给出反应条件, 并根据合成步骤数和预测产量评估最佳路径. 该设备配备了6个合成反应器模块, 包括加热、冷却、光化学反应器、填充床反应器、分离器和旁路管道, 以及3个表征模块包括液相色谱、质谱及振动光谱. 该机器人平台已成功用于15个化学小分子药物的合成路线设计和自动化合成. 此外, 美国两家制药企业默克和辉瑞也研制并搭建了机器化学家平台并实现了自动化高通量化学反应筛选. 辉瑞公司[27]在2018年基于流动化学技术开发了自动化筛选平台, 通过超高效液相色谱–质谱、紫外光谱、质谱技术相结合, 在1天内筛选超过1500个纳摩尔量级的Suzuki–Miyaura偶联反应, 实现了在不同溶剂、温度、压力等条件下进行反应. 同年, 默克公司[28,29]对其3年前的自动化平台进行改造升级, 可实现化合物生物活性测试,

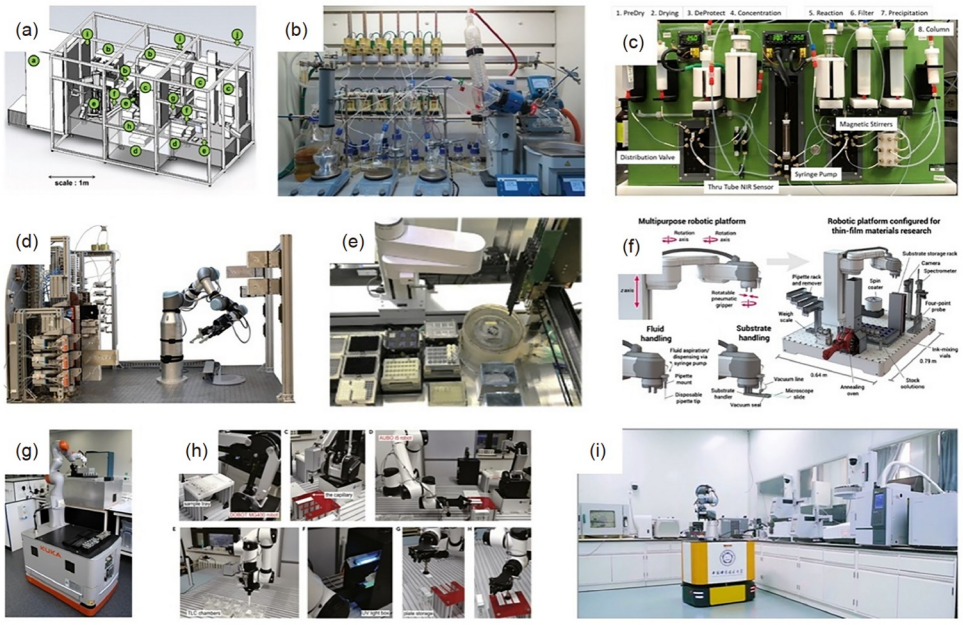

图 1　国内外各研究团队开发的机器化学家. (a) 瑞典查尔姆斯大学团队[17]; (b) 英国格拉斯哥大学团队[22]; (c) 美国伊利诺伊大学团队[24]; (d) 美国麻省理工学院团队[26]; (e) 德国赫姆霍兹研究所团队[31]; (f) 加拿大英属哥伦比亚大学团队[32]; (g) 英国利物浦大学团队[33]; (h) 北京大学团队[34]; (i) 中国科学技术大学团队[36] (网络版彩图)

Figure 1　Robotic AI chemists developed by various research groups worldwide. (a) Chalmers University group from Sweden [17]. (b) University of Glasgow group from the UK [22]. (c) University of Illinois group from the USA [24]. (d) MITgroup from the USA [26]. (e) Helmholtz Research Institute group from Germany [31]. (f) University of British Columbia group from Canada [32]. (g) University of Liverpool group from the UK [33]. (h) Peking University group from China [34]. (i) University of Science and Technology of China group [36] (color online).

使用无需标记的亲和选择–质谱法直接测定反应产物对靶标蛋白的亲和性, 并进行排序, 大大节省常规方法中分离纯化步骤的耗时，并且由此指导后续的合成、纯化、生物活性测试等工作.

在材料合成方面, 美国北卡罗莱那州立大学Abolhasani团队[30]构建了基于机器学习的实验选择和高效自主流动化学的自动化平台. 利用该平台自动合成了无机钙钛矿量子点，同时调整了它们在目标带隙上的量子产率和成分多分散性. 该系统可以通过先前的实验数据，利用贝叶斯优化算法对下一组的实验条件做出自主决定. 德国赫姆霍兹研究所Brabec团队[31]设计了高通量自动化平台用于有机光伏(OPV)材料和器件的效率和光稳定性评估(图1e)，该系统基于光学吸收特征的高斯过程回归(Gaussian Processes Regression, 简称GPR)预测指导优化过程，具有良好的预测精度. 加拿大英属哥伦比亚大学MacLeod团队[32]设计了一种基于模型的优化算法驱动的模块化机器人平台，能够通过修改薄膜成分和加工条件自主优化薄膜材料的光学和电子性能(图1f).

英国利物浦大学Cooper团队[33]于2020年设计制造出可移动机器化学家(图1g)，可以在实验室内自由移动操作不同的工作站，如点胶站、旋盖站、超声站、光解站和色谱站等. 该机器人被用来寻找高活性光催化剂, 用于光解水制氢. 机器人能够在8天内自主运行, 在10个变量的实验空间内执行688个实验. 由贝叶斯优化算法驱动，机器人会根据前一步实验结果进行分析, 确定下一步的实验计划, 最终搜索得到了6倍高活性的光催化剂混合物. 但Cooper指出了该系统存在不足之处: 机器人盲目地执行贝叶斯优化，不能捕获现有的化学知识(包括理论或物理模型)，同时不具备计算大脑, 尚不能自己产生和检验科学假设.

国内方面，北京大学莫凡洋团队[34]开发了薄层色谱(Thin-layer Chromatography，简称TLC)分析的自动化机器人(图1h)，通过高通量实验产生海量标准化的TLC数据，进而应用机器学习对数据进行回归分析, 为纯化条件的选择提供一般指导，加快优质TLC数据的生成和分析. 深圳先进技术研究院赵海涛团队[35]研制的机器化学家，基于自动实验和表征技术得到的实

验数据，使用机器学习模型构建了形貌与配方间的关联，实现了纳米金晶、双钙钛矿纳米晶的逆向设计. 中国科学技术大学江俊团队[36]在2022年研制出数据智能驱动的全流程机器化学家(图1i). 该系统的优势在于：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. 它由三个模块组成，包括机器阅读模块，通过自动读取大量化学文献来获取现有的化学知识，并由此提出科学假设、设计实验方案；移动机器人模块，自动完成多个化学任务的完整实验程序(合成、表征和性能测试共15个工作站)；计算大脑模块，通过进行理论计算来建立理论和实验数据反馈的预测模型，并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. 该套系统内置了5000万个化合物的基本信息、1100条反应路径、8万条分子光谱以及30万分子的量化计算结果, 实验平台包含色谱、紫外、荧光、拉曼等表征仪器. 该系统目前已经被应用于寻找具有聚集诱导发光特性的生物兼容发光团和优化金属氧化物光催化剂的氢掺杂策略等课题研究，显示出在光电催化、发光材料和有机合成等多领域的多功能性.

综上所述, 图2列出了各个研究团队研制的机器化学家的功能对比. 可见, 目前只有英国格拉斯哥和中国科学技术大学团队实现了自动化文献阅读功能，英国利物浦和中国科学技术大学团队实现了机器人可移动性. 最为重要的是, 除中国科学技术大学团队之外均未实现合成、表征、测试、数据处理全流程. 同时, 可进行的研究类型较为单一，仍有一些机器化学家仅处于自动化阶段, 离智能化尚有差距. 因此，一方面说明我国机器化学家发展处于并跑阶段，并在一定程度展现出领跑的苗头和趋势；另一方面，也表明机器科学家领域仍存在很大的发展空间.

1.2 机器化学家的发展趋势

随着国内外机器化学家的蓬勃发展，尤其在分子合成和功能材料自动合成方面取得了显著的进展，充分展示了机器化学家研究领域的无限潜能. 然而，目前大部分自动化平台还停留在单模块形式自动化阶段, 尚未实现化学实验多功能全流程化, 依然需要依靠人类科学家设计实验步骤、分析实验数据、优化实验方案. 因此, 可将该阶段的机器化学家称为机器化学家1.0模式. 展望未来，可根据自动化及智能化的程度、规模和发展趋势，机器化学家大致可发展为5个层级(图3).

1.2.1 机器实验员1.0模式

处于该阶段的机器化学家可以根据指令进行自动

图 2 各个研究团队机器化学家功能对比 (网络版彩图)
Figure 2 Comparison of functionalities for robotic AI chemists developed by different research groups (color online).

图 3　机器化学家发展的5个层级 (网络版彩图)
Figure 3　Five levels of the development for robotic AI chemists (color online).

化实验, 但是功能往往单一, 且无法完成从设计–合成–表征–测试–优化的一体化全流程化学实验. 同时, 实验方案设计依靠人类科学家的经验积累, 过程跟踪也依赖人工解读实验数据, 不能实现机器的主动阅读、方案设计与智能优化等功能, 离实时反馈调节、智能决策优化的目标相距甚远.

1.2.2　机器化学家2.0模式

处于该阶段的机器化学家具备机器人形式智能化, 例如中国科学技术大学研制的机器化学家. 该阶段的机器人初步具备科学大脑, 可以自主阅读文献并设计实验方案. 不仅实现全流程自动化, 还可以分析实验数据, 并且通过优化算法进一步优化. 重要的是, 该阶段的机器化学家还具有物质科学知识图谱和人工智能决策系统.

1.2.3　大规模智慧创制平台

处于该阶段的机器化学家拥有可解的人工智能化学云, 可解决高通量实验的可重复性技术难点, 配备国产自主软硬件设施, 形成机器人操作模板库、实验方案库, 工作站控制系统库和标准化学指令集, 制定机器化学家平台的行业标准. 实现机器化学家在不同实验任务和实验室之间进行迁移学习, 建成标准化智能化学实验室.

1.2.4　复杂高维的智能化学理论

处于该阶段的机器化学家可实现化学洁净与极端条件实验, 已拥有较强的科学思维化学智能, 形成理实交融的科学创造模型. 同时, 在大规模智慧创制平台的基础上, 可完成高维化学关联的智能推演. 最为重要的是, 可以面向需求创制新物质.

1.2.5　具备创造力的智慧科学家

此阶段是该领域未来发展的终极目标, 即可以揭榜挂帅, 自主解决科学问题的智慧科学家. 该阶段的机器化学家可以实现对人类科学家创造力的模仿学习, 掌握物质科学的数字知识体系和针对科学的智能符号语言. 从功能上看, 科研实验自动化整体是从辅助人到替代人的方向演进.

2　机器化学家面临的挑战

随着机器人自动化、集成化、规模化、产业化的未来发展趋势, 机器化学家领域将趋于精准化与智能化. 这一目标的实现, 需要具备智慧的科研“大脑”来驱动, 这也是机器化学家正在面临的重要挑战. 想要打造机器化学家智慧核心, 我们应积极应对以下难题和挑战: 首先, 构筑精准数据平台; 其次, 设计针对自然科学的描述符和AI算法; 最后, 构建统一的机器指令集、模板和模型库.

2.1　精准数据平台的构筑

目前, 我国的科学数据积累处于规模小、检索难、性质不全面的初级阶段, 缺乏能够大规模产生高

精度计算数据的自主量子化学计算软件. 同时, 化学科学数据质量也存在着良莠不齐、碎片化、标准不统一、格式不一致等问题. 这些科学数据问题在一定程度上限制了人工智能在化学科学领域的应用. 为此, 构筑高精度、多维度、融合关联度的物质数据体系是打造机器化学家智慧核心的基础和前提.

近年来, 国内科学家在推动数据平台建设方面取得了一定进展. 中国科学技术大学江俊团队构建了大材库(网址: http://dcaiku.com), 包含了9448万分子基础化合物、40万材料性能数据、90万含磷化合物, 同时又扩展了1120万个化合物分子数据, 初步建立了物理化学及材料化学知识图谱; 中国科学院上海有机化学研究所构建了化学专业数据库(网址: http://www.organchem.csdb.cn/scdb/default.asp), 其中包括化合物结构数据库、质谱数据库、物化性质数据库、晶体结构数据库、核磁谱图数据库、红外谱图数据库等; 中国科学院山西煤炭化学研究所温晓东研究员建立了催化体系数据库及专家系统, 并联合ChemEssen公司开发了面向石油化工、煤化工、药物的基础物性数据库(网址: http://www.imolinstincts.com.cn), 其中包含280万个化合物, 10万亿个数据集; 清华大学程津培院士建立了ibond化学键能数据库(网址: http://ibond.chem.tsinghua.edu.cn/); 国家基础学科公共科学数据中心(网址: https://www.nbsdc.cn/)建立了化学主题数据库和材料学科领域基础科学数据库.

构建统一的化学科学数据标准对于构建精准的数据库平台至关重要. 数据库错误的数据和不统一的化学表示法可能对机器学习模型的预测能力产生重大影响[37]. 目前已有多个国家在力推科学数据标准. 例如, 日本国立材料研究所开发数据与智能软件, 建立了科学数据的分析和机器学习平台; 2023年美国国立卫生研究院启动“数据管理和共享”计划[38], 包括数据格式、标准、工具及代码等. 美国橡树岭国家实验室部署了标准化数据采集通道, 发起了仪器厂商联盟, 从设备硬件底层建立采集管道与数据标准. 英国利物浦大学、格拉斯哥等高校联合, 组建了材料创新工场, 依托大科学装置打造实验数据驱动的机器化学家平台.

机器化学家所需的化学数据不仅要求多样化, 更需要进行标准化和数字化. 这就要求我们需要发展一系列工具来获取、整理、清洗、收纳和表示数据, 不仅包括化学分子的表示, 也包括合成路线中反应底物、催化剂、添加剂和溶剂的结构与性质, 以及对反应温度和时间等因素进行数字化识别, 建立其与催化反应活性和选择性评估数据的匹配关系, 构建以构效关系为关联的合成化学数据库.

2.2 针对自然科学的描述符和AI算法的设计

以自然科学中化学合成和材料设计为例, 化学合成路线优化和材料逆向设计是实现化合物和功能材料精准定制的基础, 然而化学和材料数据库的缺失和尚未明确的构效关系, 严重阻碍其实现的可能性. 描述符和AI算法是打造机器化学家智慧核心的关键. 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, 描述符应尽可能具有物理意义, 以帮助研究人员寻找模型的内涵, 对复杂的构效关系进行解耦, 挖掘描述符与目标性质之间的数学依赖关系, 提出具有物理内涵的解析表达式, 从而实现预测模型的可解释性, 提升机器学习的化学认知.

化学科学常用的描述符分为三类: 结构描述符、电子描述符、反应活性描述符. 结构描述符包括原子半径、原子序数、键长键角、基团数、摩尔体积、晶格常数、配位数、活性位点等几何结构参数, 也可以采用简化分子线性输入规范(Simplified Molecular Input Line Entry System, 简称SMILES)[39]、分子指纹、库伦矩阵[40]等直接表示分子结构. 电子描述符通常从电子结构计算中获得, 涉及d带中心、带隙、电荷/电荷差、价电子、轨道能量等物理参量. 反应活性描述符通常用于描述其接受或失去电子、质子或基团以显示反应活性的能力, 包括吸附能、电负性、电离能、电子亲和能、pK_a等物理化学性能参数. 同时, 针对特定问题, 科学家们还发展了一系列全新的特色描述符. 例如, 中国科学技术大学江俊团队[41]结合量化计算与人工智能技术, 开发了特色的电偶极矩算法和适用于催化剂设计的电(磁)偶极矩描述符; 清华大学罗三中团队[42]开发了将分子指纹和物理有机参数相结合的SPoC描述符, 实现对pK_a的准确预测; 麻省理工学院Shao等人[43]总结出电化学氧化还原电位可以作为析氧反应和氧还原反应的有效描述符.

开发化学原理清晰的AI算法可以快速高效地为机器化学家大脑的开发提供有效的数据支撑. 美国西北大学Agrawal团队[44]开发了一个可以自动获取描述符

的深度神经网络模型ElemNet, 即利用人工智能自动捕捉不同元素之间的物理和化学相互作用以及相似性, 使其能够以更好的精度和速度预测材料性能. Bombarelli等人[45]研发了变分自编码器(Variational Auto Encoder, 简称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. 在此基础上Kadurin等人[46]研发了深度生成对抗自编码器(Adversarial Auto Encoder, 简称AAE)来生成具有预定义抗癌特性的新分子. 其它强化学习方法和生成算法模型也在分子和材料设计中被广泛应用[47~49]. 另外, 当前一个重要的发展方向是基于已有的化学数据库、数字化的化学结构, 针对性地设计AI算法, 开发自动化提取化学描述符的程序, 建立一个基于化学原理的参数集, 结合已知构效关系、经验公式和化学模板提出基本规则, 以线性回归、压缩感知等方式自动化组合各类参数, 运用机器学习技术训练和迭代, 确定有效的构效关系描述符. 基于所得描述符, 形成知识图谱, 为材料和化学的高效精准合成提供数据支撑.

在软件开发方面, 江俊团队开发了光谱人工智能模拟软件(网址: http://dcaiku.com:12880/platform/first), 实现了将蛋白质红外、拉曼、紫外光谱模拟的效率相对量化计算提升3~5个数量级. 北京大学鄂维南院士[50]基于深度神经网络生成的多体势和原子间力, 开发了深度势分子动力学(Deep Potential Molecular Dynamics, 简称DPMD)方法, 计算成本比等效的从头算分子动力学(*Ab Initial* Molecular Dynamics, 简称AIMD)模拟便宜几个数量级.

2.3 机器指令集、模板和模型库的统一构建

构建机器操作指令集、模板库、实验方案库和工作站控制系统库是机器化学家发展的必然趋势. 英国格拉斯哥大学Cronin团队[21]在2020年为其机器化学家开发了化学描述语言χDL, 通过该模块将化学文献转化为机器人指令以供执行. χDL程序、实验结果和相关分析的信息都会永久保存在χDL数据库中. 中国科学技术大学江俊团队[36]开发的机器化学家中也构建了一套自然语言处理模型. 机器人可自动调用该模型将人类利用自然语言书写的论文或专利等文本转化为机器可理解的结构化数据库, 最终生成xml格式的实验操作指令, 由机器人系统和工作站系统进行解析和执行. 针对机器化学家中不同的实验分解操作, 如固/液体进样、离心、紫外光谱表征、X射线衍射表征、光催化测试、电催化测试等, 构建结构化的机器指令、模板, 以便于开展新研究时可以直接进行调用.

未来机器化学家发展的重要趋势是可移动性、多功能化、全流程化, 所以针对类似的研究课题, 应构建统一的机器人操作模板库、实验方案库、模型库和创制模板库. 这有利于实现机器化学家在不同实验任务和实验室之间进行迁移学习, 建成规模化、集成化和标准化的智能化学实验室. 这也有望解决实验数据缺乏, 数据标准不统一的问题, 实现共性规律呈现和预测模型共享, 完成化合物分子与催化材料的精准设计和高效筛选. 另外, 各科研院所、仪器厂商等单位, 应联合制定统一的数据标准和采集管道, 构建面向全球的知识平台包括机器学习模型库、智能科学工具包、科学函数算法集、物质科学数据库, 全力打造物质科学数据的全场景解决方案. 在课题组隐私数据互不交换的前提下, 通过联邦学习算法实现模型共享.

2.4 通用科学智能驱动的机器化学家平台

随着智能技术的快速发展, 未来将出现基于多学科领域、海量专业知识数据的通用科学智能. 尤其是, 面对复杂化、高维化的化学研究问题, 人类研究者往往难以全面、高效地给出解决思路. 但未来或许可借助与通用科学智能模型进行互动交流、探讨问题, 从而直接获得兼具创新且合理的启发性答案. 例如, 2022年11月OpenAI团队推出的ChatGPT语言模型(网址: https://openai.com/blog/chatgpt/), 使用了大量Transformer组件, 构成了大规模语言模型(Large Language Model, 简称LLM)GPT-3, 随后在其基础上通过人类反馈强化学习(Reinforcement Learning from Human Feedback, 简称RLHF), 实现了符合人类逻辑的对话功能. 基于大规模语料数据训练, 基本能理解人类语言对话的意图, 生成具备创造性的答案, 给出相对合理的回应.

当前在物理、化学、材料、生物等物质科学领域, 若能借助GPT等算法工具, 基于海量专业数据训练得到知识模型, 并综合其他领域知识, 便可针对某个探索性科学问题给出具有一定创造力的答案. 进而, 从合理性回答中提取具有启发性的科学假设或研究方案, 便可驱动智能机器科学家平台执行验证性实验及第一

性原理计算等理论手段验证. 装载GPT等算法模型的机器平台将能够覆盖“大胆假设——小心求证——精确预测——解决问题”的科学方法论过程. 未来将在复杂的化学与材料问题中充分利用人类知识库和专家经验, 结合理论预测与实验求证, 实现精确智能预测, 高效解决科学问题.

3 拟解决的关键问题

机器化学家领域集聚众多学科的前沿技术, 极大地帮助人类突破固有思维, 并在众多领域中开启了广泛的应用, 如有机化学的逆向反应预测、功能材料的逆向设计、生物医学精准控制等方面均有所突破. 当前, 在机器化学家领域需要着力解决的关键问题包括发展基于人工智能驱动的化学理论新方法与软件平台、拓展机器化学家的应用领域、覆盖科学方法论全流程、利用机器智能探索复杂体系的化学理论等科学问题, 同时也包括构建完备的科学知识图谱、建立统一且可通用的技术标准、提升机器人的感知、操作和规模化能力等技术问题.

3.1 科学问题

3.1.1 发展基于人工智能驱动的化学理论新方法与软件平台

主要包括: 发展面向智能的电子结构新理论及交换相关泛函; 发展针对大分子和凝聚相体系的低标度且高效电子结构智能算法; 发展人工智能的多尺度模拟算法; 发展多势能面的非绝热过程的智能构建算法; 发展机器学习的增强采样方法; 发展人工智能化的计算模拟软件平台.

3.1.2 拓展机器化学家的应用领域

一方面, 推动人工智能在合成化学、材料逆向设计、生物医药等前沿化学领域中的应用; 另一方面, 加快创制人工智能和机器人在物理、材料、生物等物质科学领域的应用, 最终实现机器科学家.

3.1.3 覆盖科学方法论全流程

针对目前公认的科学方法论全流程的四个步骤: 大胆假设、小心求证、精确预测、解决问题, 实现机器化学家的全覆盖, 开辟解决科学问题的新路径. 结合最新的GPT技术进行大胆假设, 驱动机器人进行高效实验验证和理论模拟, 形成理实交融的预测模型给出最佳方案.

3.1.4 基于机器智能探索复杂体系的化学理论

量子力学等微观理论已建立百年, 但应用于真实体系时却过于复杂难以求解. 其中理实脱节的根源在于实验测量只能给出珍贵稀疏的小数据, 难以找到全局最优解; 而理论模拟的大数据虽能找到全局最优解, 但由于使用大量理想近似而缺失了现实复杂度(图4). 因此, 如何利用机器智能建立复杂高维的化学理论, 构建“理实交融”的模型, 是机器化学家发展的重要方向. 这也是实现数据智能第四范式的必然途径.

如图4所示, 一种可能的解决方法是基于谱学观测量赋予机器人自动构建原子分子模型的能力, 驱动理论模拟产生大数据并形成可解释的预训练智能模型. 接着, 再依托机器人的高质量实测数据做二次训练, 建立面向真实体系的“理实交融”模型, 定位真实全局最优解, 推动数据智能驱动的科学研究新范式. 这种新范式通过智能模型将各学科的底层理论模拟与复杂应用实践结合, 完美释放第二次科学革命的成果, 即量子力学第一性原理在第三次工业革命中未被释放的理性指导能力. 这种方法有望在各个工业领域颠覆低效的试错研究范式, 大幅度加速工业进程.

3.2 技术问题

3.2.1 构建完备的科学知识图谱

在物理、化学、材料和生命等学科领域, 开发自适应大数据挖掘工具、通用的数据清洗与分类手段、高性能的结构和参数检索技术、快速的描述符抽取方法等. 通过结合已知构效关系、经验公式和化学模板提出基本规则, 运用机器学习技术训练和迭代, 确定有效的构效关系描述符, 以描述符为核心准确描述出从化学结构的底层数据到应用性能指标之间的数学映射关系, 形成知识图谱.

3.2.2 建立统一且可通用的技术标准

主要包括数据库标准、机器人性能参数标准、机器人控制软件的参数标准、实验数据格式标准、机器指令集、实验模板, 以及机器化学家平台的智能评估等级和行业标准等.

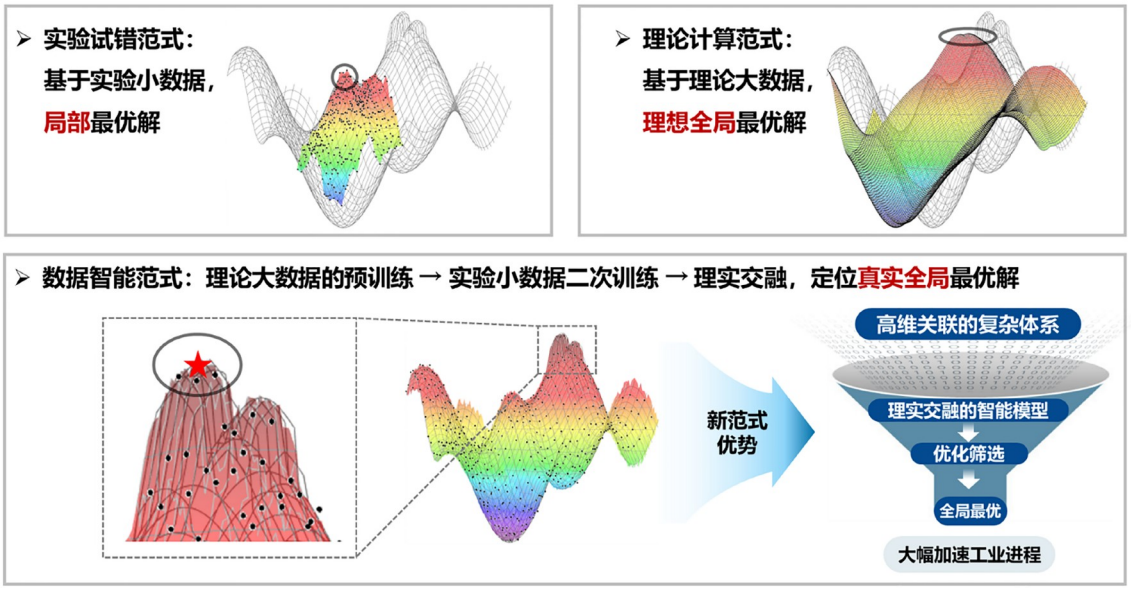

图 4　数据智能研究范式的优势 (网络版彩图)
Figure 4　The advantages of the data intelligence research paradigm (color online).

3.2.3　提升机器人的感知、操作和规模化能力

可通过加入实时谱学观测赋予机器人物质级别的感知能力, 使机器人在光谱的实时反应下能够及时预判问题, 从而具备更加自主的操作能力, 并实现多机协作及大规模智慧创制, 能够覆盖多任务复杂度.

4　强化科学基金资助的发展建议

机器化学家领域开辟了一条探索化学化工、计算机、信息学、软件、生物医学、材料学等众多学科前沿的新研究机遇. 同时, 该领域也面临着不少科学瓶颈, 亟需集聚一批有潜力的年轻科学家联合攻关, 大力推进该领域方向的高质量发展. 为此, 结合当前基础研究的特点和科学基金改革的方向, 可从研究方向的交叉、人才队伍的培养、团队平台的搭建与发展等多方面出发, 以解决上述关键问题为落脚点, 依托国家重大项目组织形式, 着力推动我国在该领域方向的发展.

为此, 我们建议科学基金在面向智能的电子结构新理论与新算法发展、针对大分子和凝聚相体系的高效电子结构智能算法发展、人工智能的多尺度模拟算法发展、人工智能驱动的计算模拟软件平台开发等方面部署重点类项目. 同时, 建议对不同学科领域的数据库及知识图谱的构建、大数据挖掘工具的开发、高性能的结构和参数检索技术的发展等交叉领域进行布局. 此外, 建议相关领域科学家重点关注机器化学家开发及应用过程中涉及到的以下方向: 机器学习, 尤其是深度学习、迁移学习和强化学习在化学反应机理、分子合成、材料及催化剂设计、药物研发等领域的应用; 关注机器化学家在实验过程中的自主控制能力、实验数据采集与处理能力等方面的提升; 推动机器化学家的全流程智能化研究, 探索机器化学家的智能搜索、迭代验证、自主优化能力; 开展机器化学家与人工智能技术的深度融合研究, 如借助GPT等算法工具对机器人进行驱动, 探索机器化学家与人工智能的协同能力, 以及二者在化学领域中的互补作用.

优先发展这些研究方向, 也将大力推动我国研究范式的创新, 有望在形成新理论、发展新方法, 创造新知识体系方面取得突破, 并始终保持在国际前沿领域中抢占先机.

致谢　感谢中国科学技术大学杨金龙院士、罗毅教授、李震宇教授、瞿昆教授、蒋彬教授、胡伟研究员、汪炀教授、李微雪教授、侯中怀教授, 中国科学院计算技术研究所贾伟乐研究员, 中国科学院山西煤炭化学研究所温晓东研

究员、刘星辰研究员, 上海科技大学姜珊研究员、中国科学院深圳先进技术研究院赵海涛研究员等在“机器科学家青年论坛”战略研讨会的精彩报告以及提供的相关素材.

参考文献

1 Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, Bateman A, DePristo MA, Colwell LJ. *Nat Biotechnol*, 2022, 40: 932–937

2 Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. *Neurocomputing*, 2016, 187: 27–48

3 Marcheggiani D, Bastings J, Titov I. Exploiting Semantics in Neural Machine Translation with Graph Convolutional Networks, Preprint at https://arxiv.org/abs/1804.08313, 2018

4 Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645–6649

5 Min S, Lee B, Yoon S. *Brief Bioinform*, 2016, 18: bbw068

6 E W. *Not Amer Math Soc*, 2021, 68: 1

7 Schmidt M, Lipson H. *Science*, 2009, 324: 81–85

8 Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, Zeller M, Friedler SA, Schrier J, Norquist AJ. *Nature*, 2016, 533: 73–76

9 Mikulak-Klucznik B, Gołębiowska P, Bayly AA, Popik O, Klucznik T, Szymkuć S, Gajewska EP, Dittwald P, Staszewska-Krajewska O, Beker W, Badowski T, Scheidt KA, Molga K, Mlynarski J, Mrksich M, Grzybowski BA. *Nature*, 2020, 588: 83–88

10 Wołos A, Koszelewski D, Roszak R, Szymkuć S, Moskal M, Ostaszewski R, Herrera BT, Maier JM, Brezicki G, Samuel J, Lummiss JAM, McQuade DT, Rogers L, Grzybowski BA. *Nature*, 2022, 604: 668–676

11 Sanchez-Lengeling B, Aspuru-Guzik A. *Science*, 2018, 361: 360–365

12 Oliynyk AO, Antono E, Sparks TD, Ghadbeigi L, Gaultois MW, Meredig B, Mar A. *Chem Mater*, 2016, 28: 7324–7331

13 Segler MHS, Preuss M, Waller MP. *Nature*, 2018, 555: 604–610

14 Ma S, Huang SD, Liu ZP. *Nat Catal*, 2019, 2: 671–677

15 Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, Rühl S, Wolverton C. *npj Comput Mater*, 2015, 1: 1–5

16 Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. *JOM*, 2013, 65: 1501–1509

17 King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan KE, Clare A. *Science*, 2009, 324: 85–89

18 Williams K, Bilsland E, Sparkes A, Aubrey W, Young M, Soldatova LN, De Grave K, Ramon J, de Clare M, Sirawaraporn W, Oliver SG, King RD. *J R Soc Interface*, 2015, 12: 20141289

19 Bilsland E, van Vliet L, Williams K, Feltham J, Carrasco MP, Fotoran WL, Cubillos EFG, Wunderlich G, Grøtli M, Hollfelder F, Jackson V, King RD, Oliver SG. *Sci Rep*, 2018, 8: 1038

20 Granda JM, Donina L, Dragone V, Long DL, Cronin L. *Nature*, 2018, 559: 377–381

21 Mehr SHM, Craven M, Leonov AI, Keenan G, Cronin L. *Science*, 2020, 370: 101–108

22 Steiner S, Wolf J, Glatzel S, Andreou A, Granda JM, Keenan G, Hinkley T, Aragon-Camarasa G, Kitson PJ, Angelone D, Cronin L. *Science*, 2019, 363: eaav2211

23 Manzano JS, Hou W, Zalesskiy SS, Frei P, Wang H, Kitson PJ, Cronin L. *Nat Chem*, 2022, 14: 1311–1318

24 Blair DJ, Chitti S, Trobe M, Kostyra DM, Haley HMS, Hansen RL, Ballmer SG, Woods TJ, Wang W, Mubayi V, Schmidt MJ, Pipal RW, Morehouse GF, Palazzolo Ray AME, Gray DL, Gill AL, Burke MD. *Nature*, 2022, 604: 92–97

25 Bédard AC, Adamo A, Aroh KC, Russell MG, Bedermann AA, Torosian J, Yue B, Jensen KF, Jamison TF. *Science*, 2018, 361: 1220–1225

26 Coley CW, Thomas Iii DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, Hart T, Fishman JS, Rogers L, Gao H, Hicklin RW, Plehiers PP, Byington J, Piotti JS, Green WH, Hart AJ, Jamison TF, Jensen KF. *Science*, 2019, 365: eaax1566

27 Perera D, Tucker JW, Brahmbhatt S, Helal CJ, Chong A, Farrell W, Richardson P, Sach NW. *Science*, 2018, 359: 429–434

28 Santanilla AB, Regalado EL, Pereira T, Shevlin M, Bateman K, Campeau LC, Schneeweis J, Berritt S, Shi ZC, Nantermet P, Liu Y, Helmy R, Welch CJ, Vachal P, Davies IW, Cernak T, Dreher SD. *Science*, 2015, 347: 49–53

29 Gesmundo NJ, Sauvagnat B, Curran PJ, Richards MP, Andrews CL, Dandliker PJ, Cernak T. *Nature*, 2018, 557: 228–232

30 Epps RW, Bowen MS, Volk AA, Abdel-Latif K, Han S, Reyes KG, Amassian A, Abolhasani M. *Adv Mater*, 2020, 32: 2001626

31 Du X, Lüer L, Heumueller T, Wagner J, Berger C, Osterrieder T, Wortmann J, Langner S, Vongsaysy U, Bertrand M, Li N, Stubhan T, Hauch J, Brabec CJ. *Joule*, 2021, 5: 495–506

32 MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, Dettelbach KE, Moreira R, Yunker LPE, Rooney MB, Deeth JR, Lai V, Ng GJ, Situ H, Zhang RH, Elliott MS, Haley TH, Dvorak DJ, Aspuru-Guzik A, Hein JE, Berlinguette CP. *Sci Adv*, 2020, 6: eaaz8867

33 Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, Li X, Alston BM, Li B, Clowes R, Rankin N, Harris B, Sprick RS, Cooper AI. *Nature*, 2020, 583: 237–241

34 Xu H, Lin J, Liu Q, Chen Y, Zhang J, Yang Y, Young MC, Xu Y, Zhang D, Mo F. *Chem*, 2022, 8: 3202–3214

35 Zhao H, Chen W, Huang H, Sun Z, Chen Z, Wu L, Zhang B, Lai F, Wang Z, Adam ML, Pang CH, Chu PK, Lu Y, Wu T, Jiang J, Yin Z, Yu XF. *Nat Synth*, 2023, DOI:10.1038/s44160-023-00250-5

36 Zhu Q, Zhang F, Huang Y, Xiao H, Zhao LY, Zhang XC, Song T, Tang XS, Li X, He G, Chong BC, Zhou JY, Zhang YH, Zhang B, Cao JQ, Luo M, Wang S, Ye GL, Zhang WJ, Chen X, Cong S, Zhou D, Li H, Li J, Zou G, Shang WW, Jiang J, Luo Y. *Natl Sci Rev*, 2022, 9: nwac190

37 Young D, Martin T, Venkatapathy R, Harten P. *QSAR Comb Sci*, 2008, 27: 1337–1345

38 Kozlov M. *Nature*, 2022, 602: 558–559

39 Weininger D. *J Chem Inf Model*, 1988, 28: 31–36

40 Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. *Phys Rev Lett*, 2012, 108: 058301

41 Wang X, Ye S, Hu W, Sharman E, Liu R, Liu Y, Luo Y, Jiang J. *J Am Chem Soc*, 2020, 142: 7737–7743

42 Yang Q, Li Y, Yang JD, Liu Y, Zhang L, Luo S, Cheng JP. *Angew Chem Int Ed*, 2020, 59: 19282–19291

43 Kuznetsov DA, Han B, Yu Y, Rao RR, Hwang J, Román-Leshkov Y, Shao-Horn Y. *Joule*, 2018, 2: 225–244

44 Jha D, Ward L, Paul A, Liao W, Choudhary A, Wolverton C, Agrawal A. *Sci Rep*, 2018, 8: 17593

45 Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. *ACS Cent Sci*, 2018, 4: 268–276

46 Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A. *Oncotarget*, 2016, 8: 10883–10890

47 Popova M, Isayev O, Tropsha A. *Sci Adv*, 2018, 4: eaap7885

48 Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. *Mol Pharm*, 2017, 14: 3098–3104

49 Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Zhavoronkov A. *J Chem Inf Model*, 2018, 58: 1194–1204

50 Wang H, Zhang L, Han J, E W. *Comput Phys Commun*, 2018, 228: 178–184

Challenges and opportunities of machine chemists

Jun Jiang[1], Yanbo Li[1], Xiangjian Shen[2*], Feixue Gao[2*]

[1] *School of Chemistry and Materials Science, University of Science and Technology of China, Hefei 230026, China*
[2] *Department of Chemical Sciences, National Science Foundation of China, Beijing 100085, China*
**Corresponding authors (email: shenxj@nsfc.gov.cn; gaoxf@nsfc.gov.cn)*

Abstract: This paper has summarized the main academic communications from the forum titled as "Young Forum of Robotic AI Chemist" which was granted by National Natural Science Foundation of China (NSFC) and held in Hefei in 2022. This strategy seminar mainly focused on an important branch of the field of robotic AI scientists–robotic AI chemists. Firstly, the latest research progress in the field of robotic AI chemists is introduced, which involves hardware facilities, intelligent degree and experimental ability. Secondly, the development trend of robotic AI chemists is discussed, and five possible levels of developing models are proposed. Finally, some important challenges and opportunities faced by this field are proposed, which mainly contains the need to build accurate data platforms, the need to design descriptors and AI algorithms for natural science, and the need to build unified machine instruction sets, templates and model libraries. At the same time, this paper also summarizes the key scientific problems to be solved in this filed, and some strategic suggestions for strengthening the support of NSFC are proposed.

Keywords: robotic AI chemist, artificial intelligence, automation of synthetic chemistry, reverse design of material

doi: 10.1360/SSC-2023-0044