

Artificial Intelligence (AI) in Chemistry: A Review of Methods, Applications, Challenges and Prospects

1. 背景与意义 / Background & Motivation

传统化学研究（合成、筛选、性质测试、反应预测等）高度依赖人工经验、试错 (trial-and-error) 和专家直觉 (chemical intuition); 这不仅费时、耗力，而且在面对庞大化学空间 (chemical space) 时效率低下。

随着计算能力、数据积累 (databases, high-throughput experiments) 与算法 (machine learning, deep learning) 的发展，AI 为化学家提供了一种基于数据 + 算法 + 计算资源的新范式 (paradigm)，使得“预测 – 设计 – 生成 – 优化”成为可能。

ResearchGate

+2

web.cas.org

+2

对于制药 (drug discovery)、材料化学 (materials chemistry)、有机合成 (organic synthesis)、环境化学 (environmental chemistry) 等多个子领域，AI 有潜力极大地加速发现周期 (time-to-discovery)、降低成本 (cost)、提升成功率 (hit rate / yield / selectivity) 。

ScienceDirect

+2

ScienceDirect

+2

因此，将 AI 与化学结合 — “智能化学 (AI-chemistry / digital chemistry)” — 被认为是

化学研究范式 (research paradigm) 的重大变革。最近已有综述总结这一“交汇时刻”。

ResearchGate

+2

ResearchGate

+2

2. 方法基础与技术路线 / Methods and Key Techniques

在 AI 化学 (AI-chemistry) 中, 核心要素通常包括: 化学数据 (chemical data) + 分子 / 反应表示 (molecular/reaction representation) + 机器学习 / 深度学习模型 (ML / DL models) + 高质量数据集 (datasets)。下面是几个关键构件与技术路线。

2.1 分子与反应的表示 (Molecular / Reaction Representation)

传统 cheminformatics 多用指纹 (fingerprint)、SMILES 字符串 (SMILES strings)、分子图 (molecular graph) 等作表示。

SpringerLink

+1

近年深度学习 (deep learning) 推动了 分子表征学习 (Molecular Representation Learning, MRL) — 通过 neural networks 从原始分子结构 (graph / coordinates) 学习 latent vector (嵌入空间) 表示。这样, AI 可以“理解”分子结构特征, 而不完全依赖人工定义的描述符 (descriptors / fingerprints) 。

RSC 出版

+2

arXiv

+2

对于反应 (reaction) 本身，也有专门表示法 (reaction representation)，比如图神经网络 (graph-based neural networks) 直接把反应前后分子 graph 当输入，从而学习反应模式 (reaction patterns)。一个例子是 GraphRXN 模型，用于有机反应预测 (reaction prediction)。

PMC

2.2 机器学习 / 深度学习模型 (ML / DL Models)

常见模型和技术包括：

传统 ML：随机森林 (Random Forests)、支持向量机 (SVM)、kernel methods 等。适用于分子性质 (properties) 预测等。

PMC

+2

PMC

+2

深度学习 / 图神经网络 (Graph Neural Networks, GNNs)、几何深度学习 (Geometric Deep Learning, GDL) —— 特别适合处理图 (graph) 或结构化分子/反应数据。

arXiv

+2

RSC 出版

+2

生成模型 (Generative Models)：通过编码器-解码器 (encoder-decoder) 构造 latent space，使 AI 能“生成 (generate)”新分子 (molecule generation) 或预测潜在反应路径。经典工作如 Automatic chemical design using a data-driven continuous representation of molecules。

arXiv

+1

混合模型 (hybrid models): 将 ML / DL 与传统计算化学 (如分子动力学, 量子化学) 结合, 在效率与物理准确性之间寻求平衡。近年来这方面发展迅速。

ScienceDirect

+2

SpringerLink

+2

2.3 数据集 (Datasets) 与数据质量 (Data Quality)

AI 模型的表现高度依赖于训练数据 (data)——包括结构 (geometry)、电子/量子化学特性 (energies, orbitals 等)、实验性质 (reactivity, yields, kinetics, ADMET 性质) 等。

SpringerLink

+2

PMC

+2

常用量子化学数据集 (quantum chemistry datasets) 有如 QM9、ANI-1 等, 它们包含小分子的几何、能量、电子结构等信息, 可用于训练 AI 模型预测分子性质或反应特性。

SpringerLink

+1

对于反应与合成 (synthesis / retrosynthesis) 任务, 还需大规模、准确标注 (annotated) 的反应数据库 (包括反应条件、产率、选择性、局部结构变化等)。数据稀疏性 (data sparsity)、偏倚 (bias)、不一致 (inconsistency) 是目前的重要瓶颈。

ResearchGate

+2

ScienceDirect

+2

3. 主要应用领域 / Key Application Domains

AI 在化学中的应用非常广泛。以下是几个最具代表性和前沿的方向。

3.1 分子性质预测 (Molecular Property Prediction)

预测分子的物理 / 化学性质 (melting point, solubility, HOMO–LUMO 能级, 极化率等), 以及 ADMET (吸收、分布、代谢、毒性) 性质。这样可以在合成之前筛选出潜力化合物。

chemai.io

+2

SpringerLink

+2

使用 representation learning + ML 模型 (如 GNNs) 对未知 (novel) 分子进行预测, 比传统经验 / 经验式关系 (empirical rules) 更具泛化能力 (generalizability)。

RSC 出版

+2

ResearchGate

+2

3.2 分子生成与新分子设计 (Molecule Generation & Design)

生成模型 (generative models) 允许通过 latent space 探索 (interpolate / extrapolate)

来发现“化学空间 (chemical space)”中的新分子。例如自动编码器 (autoencoder)、变分自动编码器 (VAE)、生成对抗网络 (GAN) 或其他生成机制。经典工作包括 上文提到的 Automatic chemical design。

arXiv

+1

这些方法可用于药物先导化合物 (lead compounds) 的发现，也可用于材料 (功能分子) 的设计，如光电材料、催化剂、功能性高分子等。

ScienceDirect

+2

ResearchGate

+2

3.3 合成规划 (Synthesis Planning) 与反应预测 (Reaction Prediction / Retrosynthesis)

传统合成分析 (synthetic planning) 长期依赖专家经验。现在，深度学习 + 图模型 + 搜索算法 (e.g. Monte Carlo Tree Search, MCTS) 已被用于 retrosynthesis (逆合成) 规划，以自动建议从原料到目标分子的合成路线。

arXiv

+2

ResearchGate

+2

向前 (forward) 反应预测 (给定反应物 + 条件 \rightarrow 产物) 也取得了显著进展。比如一项最近工作 (2025 年) 使用新生成模型 (generative AI) 能够更准确地预测反应产物，同时保持物理 (electron / mass) 守恒 (conservation) 约束。

MIT 新闻

+2

国家科学院院刊

+2

用于自动化与机器人 (robotics) 实验平台 — 将 AI 用于合成规划 + 实验执行 + 反馈 (闭环, closed-loop), 即自动化实验室 (autonomous labs)。这种 “数字化 / 自动化化学 (digital chemistry / autonomous chemistry)” 被认为是未来趋势。

ResearchGate

+2

SpringerLink

+2

3.4 材料设计 (Materials Discovery)

AI 应用于新材料 (功能材料) 的发现 — 包括催化剂、光电材料、储能材料等。通过 ML / DL 模型对材料组成 / 结构 / 性能之间关系建模 (structure–property mapping)。

ScienceDirect

+2

SpringerLink

+2

特别是对于复杂 / 高维 / 多成分体系 (如合金、掺杂材料、聚合物等), AI 提供比传统经验或试错更系统、高效的筛选与优化手段。

itginsight.com

+2

SpringerLink

+2

3.5 化学生物学与药物发现 (Chemical Biology & Drug Discovery)

AI 辅助药物设计 (AI-driven drug discovery): 从分子生成、性质预测、构象 / 靶点结合预测 (docking / binding affinity prediction)、ADMET 预估, 到合成路径规划, 全流程都有 AI 的身影。

ScienceDirect

+2

SpringerLink

+2

对大分子 (例如蛋白质)、生物大分子结构 / 功能 / 相互作用 (structure–function, binding) 的预测也借助 AI。尽管这一领域还面对挑战 (尤其是功能性构象, 动态性, 生理环境)。

ScienceDirect

+1

3.6 分析化学、光谱、性质测定自动化 (Analytical Chemistry, Spectroscopy, High-Throughput Prediction)

AI 可以用于预测 / 解析分子光谱 (IR, NMR, UV/Vis 等), 加速谱学分析 (spectroscopic analysis)。最近有综述指出: 通过 ML, 从分子结构 / 化学式直接预测光谱 (spectrum) 的方法正在发展。

arXiv

+1

对于高通量筛选 (high-throughput screening, HTS) + 实验 + 数据分析 + ML 模型 — 未来可实现 “从合成、测定、分析、反馈, 再到优化 / 设计” 的闭环 (closed-loop) 自动化流程 (autonomous lab)。

ResearchGate

+2

SpringerLink

+2

4. 最新进展与前沿成果 / Recent Advances & Breakthroughs

2025 年综述 — The converging moment of chemistry and artificial intelligence: a review and outlook (2025) 系统总结了 AI 在化学数据提取、反应预测与优化、合成路径规划与“化学智能体 (chemical agents)” 的最新进展，指出 AI 正在推动化学研究迈向智能化与自动化。

ResearchGate

最新 generative AI — 反应预测 — FlowER 系统 (2025, 出自 MIT) 将电子 (electrons) 的重新分布 (electron redistribution) 纳入模型，确保反应预测时满足质量守恒 (mass conservation) 和电子守恒 (electron conservation)，大大提升预测可靠性与现实可操作性。

MIT 新闻

分子表征学习 (MRL) 的突破 — 最近在分子性质预测上的研究，如 Molecular representation learning for the prediction of chemical properties (2024)，展示了 MRL 如何通过 deep-learning 对传统任务 (如分子性质预测) 提高准确性与泛化能力。

RSC 出版

AI + 材料化学 与 自动化 / 自主实验室 (Autonomous Labs) — 最新综述指出，AI + 自动化将从单纯“预测”转向“自动发现 (discovery)”与“自动实验 (autonomous experimentation)” — 例如高通量 + ML + 自动合成 / 测试平台。

ScienceDirect

+2

SpringerLink

+2

药物发现领域的新进展 — 最新 ML 方法已经显著提升结构基础 (structure-based)

药物设计 (drug design) 的效率和精度, 同时增强对毒性 (toxicity)、ADMET 性质的预测能力。2025 年的综述 (special issue) 总结这些进步, 并指出 AI 正成为现代药物研发不可或缺的工具。

SpringerLink

+1

5. 优势、挑战与局限 / Strengths, Challenges & Limitations

5.1 优势 (Strengths / Why AI helps)

效率与规模 (Efficiency & Scalability): AI 能够处理大量数据 (数据库 + 高通量实验数据), 短时间内筛选、预测、设计数千到数百万个化合物 / 反应 / 材料。

发现隐藏规律 (Hidden Patterns): 深度学习 / 图神经网络能捕捉传统方法难以识别的非线性、多维特征 — 帮助发现传统规则 (rules) 难以总结的新规律 (new patterns) 。

节省时间和资源 (Cost & Time Saving): 大幅减少合成 / 筛选 / 测试所需的物理实验次数, 降低失败代价 (failed experiments); 加快新药 / 新材料 / 新反应的开发周期。

跨领域 (Multidisciplinary) 和灵活性 (Flexibility): AI 可应用于有机合成、药物、材料、分析、环境、能源等多个子领域, 增强研究广度和融合性。

5.2 挑战与局限 (Challenges / Limitations)

数据问题 (Data Quality & Availability): 很多化学反应或材料体系缺乏足够高质量、标准化、可公开的大规模数据; 数据偏倚、标签不一致 (inconsistent labeling)、报告选择性 (reporting bias) 常见。

ResearchGate

+2

SpringerLink

+2

模型可解释性 (Interpretability): 深度学习模型 (尤其是 graph-based / generative models) 往往像 “黑盒 (black box)” —— 虽然预测/生成能力强, 但很难解释其“为什么这样预测/生成 (why)”。这对于理解反应机制 (mechanism)、机制合理性 (chemical validity) 很不利。

SpringerLink

+2

ResearchGate

+2

物理 / 量子力学准确性 (Physical / QM Accuracy): AI 模型多数是数据驱动 (data-driven), 可能忽略量子力学、热力学、动力学等物理约束 (constraints) — 导致一些预测可能不符合真实物理 / 化学规律。尽管已有混合模型和生成模型尝试解决, 但仍是挑战。

SpringerLink

+2

中国科学院院刊

+2

通用性 (Generalizability) 与偏差 (Bias): AI 模型通常对其训练集 (training distribution) 表现好, 但对于训练集中未覆盖的 “化学空间 (chemical space)” 部分 (novel scaffolds, 未知反应类型) 的预测/生成能力可能较差。

自动化实验与闭环 (Autonomous Lab) 的局限: 虽然自动化 + AI 是趋势, 但现实中实验条件复杂 (溶剂、温度、混合、纯化…) —— AI 很难完全取代化学家的专业判断。数据共享、标准化、实验 reproducibility (可重复性) 也是难题。

ResearchGate

+2

SpringerLink

+2

6. 当前热点 & 研究趋势 / Hot Topics & Future Directions

端到端 (end-to-end) 自动化 / 自主实验室 (Autonomous Chemistry): 将 AI + 高通量实验 + 机器人 / 自动合成/分析 — 构建闭环自动发现系统 (closed-loop discovery pipelines), 未来可能大幅加速新分子 / 新材料 / 新反应发现。

ResearchGate

+2

国家科学院院刊

+2

混合模型 (Hybrid AI + Physics / QM) 的发展: 为了兼顾物理 / 化学准确性与效率, 越来越多研究尝试把量子化学 (quantum chemistry) + ML 结合 — 在高准确性 (energy, reactivity) 与高效率之间取得平衡。

SpringerLink

+2

ScienceDirect

+2

更好的分子 / 反应表示 (Representation): 例如更先进的图 / 几何 / 对称 (symmetry-aware) 表示 (geometry-aware GNNs, GDL) —— 更好地捕捉分子几何、电子结构、对称性等。

arXiv

+2

RSC 出版

+2

可解释性 AI (Explainable AI, XAI) 在化学中的应用： 越来越多人认识到，仅仅“预测 / 生成”是不够的，必须理解“为什么 (why)”。因此，将来在 reaction mechanism, catalyst design, toxicity origin 等方面，对可解释性模型 (解释其内部决定过程) 的需求将大大增加。

ScienceDirect

+2

ResearchGate

+2

多模态 / 混合数据 + 多任务学习 (multimodal / multitask learning): 将结构、光谱 (spectra)、动力学 (kinetics)、物性 (properties)、合成条件 (conditions) 等多种数据集合起来，一起训练模型，使 AI 具备跨领域、跨任务 (multi-property, multi-purpose) 能力。

ScienceDirect

+2

arXiv

+2

数据共享、标准化与社区建设： 要实现真正有效、可靠的 AI 化学研究，需要更多共享、高质量、标准化的数据库 (public, curated datasets)，以及良好的社区规范 (standards, reporting guidelines)。这是未来能否大规模推广 AI 化学 (especially autonomous labs) 的基础。

ResearchGate

+2

SpringerLink

+2

7. 对你个人研究 / 兴趣 (结合你背景) 的启示

鉴于你目前:

正在学习有机化学 (reaction mechanisms, 合成路径)、物理、量子、计算化学理论基础,

对有机合成、有机机制 (organic synthesis, mechanism) 有兴趣,

并且希望未来可能申请物理 / 化学研究生项目 (比较不同学校),

那么你可以考虑将 AI 化学 作为一个有前景、跨学科相关方向。具体而言:

如果你对有机合成 / 反应机制有兴趣, 可以关注 AI 反应预测 (reaction prediction / retrosynthesis planning) 的研究 — 这有助于自动化合成路线设计 + 优化。

如果你对材料 / 功能分子 (functional molecules) 感兴趣, AI 分子生成 + 性质预测 + 材料筛选 是一个非常有潜力的方向。

你的量子力学 / 物理背景 + 对计算 / 编程可能有兴趣, 结合 AI + 计算化学 (hybrid QM+ML) 会是很好的研究基础。

可以尝试找一些公开数据库 (如 QM9, ANI-1) + 学习分子表征 (molecular representation)、Graph Neural Network (GNN)、Representation Learning 等, 为将来做 AI 化学方向打基础。

8. 总结 / Conclusion

总的来说, AI 正在深刻改变化学研究的范式 — 从传统经验与试错 (trial-and-error) 向数据驱动、算法辅助、自动化与智能化 (data-driven, algorithm-assisted,

autonomous) 转变。分子 / 反应 / 材料设计、性质预测、合成规划都受益良多。尽管目前仍面临数据、可解释性、物理准确性、通用性等挑战，但混合模型、自动化实验室、可解释 AI、多模态学习等研究趋势正在迅速推进这个领域。对你这样兼具化学基础 + 物理 / 量子背景 + 学术兴趣的人来说，AI 化学极具吸引力，可能成为你未来研究 / 学习路线中的重要方向