人工智能技术辅助对科学实验的评估与反馈

寇亚凤
（长春经济技术开发区世纪小学　吉林长春　130031）

摘要：随着科学研究数据量呈指数级增长，传统实验方法面临效率瓶颈，而人工智能技术为突破这一瓶颈提供了新思路。本研究聚焦于 AI 辅助科学实验的新范式，尤其关注“实验引导的假设排名”方法，以 MOOSE-Chem3 系统为代表。该系统通过功能组分提取、智能聚类、模拟实验执行和迭代总结四步策略，实现对科学假设的动态评估与优化。研究发现，AI 系统能够在模拟环境中准确预测实验趋势（Spearman 相关系数高达 0.96），并将识别最优假设所需的平均实验次数从 32 次降至 15 次。研究表明，AI 辅助实验评估与反馈系统有望显著降低科研成本，缩短研发周期，成为科学发现新范式的关键推动力。

关键词：人工智能；科学实验；大型语言模型；MOOSE-Chem3

引言：近年来，人工智能技术，特别是大型语言模型（LLMs）的快速发展为科学研究提供了新的可能性。传统 AI 在科学研究中的应用主要局限于“实验前假设排名”，即在实验开始前对可能的假设进行预先筛选和排序。虽然这种方法能够在一定程度上提高实验效率，但由于缺乏对实验反馈的动态利用，其效果仍然有限。最近的研究显示，将 AI 深入融入实验过程，形成“实验引导的假设排名”（Experiment-Guided Hypothesis Ranking）的新范式，有望带来科学发现方式的根本性变革。以 MOOSE-Chem3 为代表的系统不再仅仅是假设的“生成器”，而是能够根据实验结果动态调整假设优先级，指导科学家选择最具潜力的下一步实验方向，从而最大限度地减少所需的实验次数和资源投入。

1. AI 在科学研究中的发展历程

1.1 传统科学发现范式及其局限

传统科学发现范式主要依赖科学家基于已有知识提出假设，设计实验验证假设，分析实验数据，并形成结论。这一过程虽然经过几个世纪的发展和完善，但在面对当今复杂科学问题时显现出明显局限性。科学文献的快速增长使得全面掌握某一领域的所有相关知识变得几乎不可能，科学家很难仅凭直觉和有限阅读辨识出最有潜力的研究方向。同时，实验设计和执行过程中的人为因素也可能导致效率低下或结果偏差。

1.2 AI 辅助科学研究的早期尝试

AI 在科学研究中的应用可追溯至 20 世纪 80 年代的专家系统，如 DENDRAL（化学结构推断）和 MYCIN（医疗诊断）。这些系统基于符号推理和规则引擎，能够在特定领域模拟专家思维过程。随着机器学习技术的发展，90 年代和 21 世纪初出现了一批基于数据驱动的科学发现系统，如用于药物发现的 QSAR（定量结构-活性关系）模型。这些系统虽然取得了一定成功，但由于算法和计算能力的限制，其应用范围和效果仍然有限。

1.3 大型语言模型与科学研究的融合

2020 年后，以 GPT、BERT 为代表的大型语言模型技术取得突破性进展，为 AI 辅助科学研究开辟了新天地。这些模型通过预训练获得了大量领域知识，能够理解和生成与人类相似的科学文本，为科学假设生成和实验设计提供支持。

2. 实验引导假设排名的基本原理和技术框架

2.1 实验引导假设排名的理论基础

实验引导假设排名的核心思想是将 AI 深度融入实验过程，通过对实验结果的实时分析和学习，动态调整假设优先级，指导科学家选择最具潜力的下一步实验方向。这一方法基于三个关键假设：

（1）最优解假设（A1）：针对特定科学问题，假设空间中存在一个主导最优解，代表经过实验验证的理想结果。

（2）性能-距离关系假设（A2）：假设与最优解的接近程度与其实验性能正相关，即越接近最优解的假设，实验表现越好。

（3）不完美嵌入假设（A3）：实际中，无论是人类专家还是 AI 对假设空间的理解都存在局限性，导致对“接近度”的感知出现扭曲，使得原本平滑的理想性能曲面变为更复杂的形态。基于这些假设，可以构建一个数学模型来描述假设空间中的性能分布。假设在潜在假设空间 $\mathcal{H}$ 中，每个假设 h 都被表示为一个点，点的坐标反映了假设的不同变体。对于特定科学问题 q，最优假设 h*是使得性能函数 f(h,q)最大化的假设。在理想情况下，性能函数 f(h,q)可以表示为：

$$f(h,q) = f_{\text{ideal}}(h,q) = 1 - \frac{d(h,h^*)}{d_{\max}}$$

其中 d(h,h*)是假设 h 与最优假设 h*之间的距离（如欧几里得距离），d_{max} 是正规化因子。然而，由于不完美嵌入假设（A3），实际观察到的性能函数 $f_{obs}(h,q)$可能与理想情况有所偏差：

$$f_{\text{obs}}(h,q) = f_{\text{ideal}}(h,q) + \epsilon(h,q)$$

其中 $\epsilon(h,q)$ 是系统性修正项，反映了嵌入函数的不完美性。

2.2 实验引导假设排名的技术框架

基于上述理论，实验引导假设排名的技术框架通常包含四个核心组件：

（1）假设生成与表示：利用大型语言模型生成初始假设集合，并将其表示为结构化形式（如化学结构、参数向量等）。

（2）模拟实验环境：构建高保真度的实验模拟器，能够快速评估假设的性能并返回实验结果。

（3）动态排名算法：根据已有实验结果动态调整假设优先级，选择下一个最具潜力的实验方向。

（4）知识积累与总结：分析实验结果，提取关键见解，形成累积知识，指导后续实验。

这一框架的核心是将 AI 从实验前的被动辅助工具转变为实验过程中的主动学习系统，通过与实验环境的持续互动实现对科学问题更深入的理解和更高效的探索。

3. MOOSE-Chem3：AI 辅助实验系统的典型代表

3.1 MOOSE-Chem3 系统概述

MOOSE-Chem3 是一个基于大型语言模型的 AI 辅助化学实验系统，由上海人工智能实验室、中国科学技术大学、南洋理工大学等机构联合研发。该系统实现了“实验引导的假设排名”

范式，能够在化学实验中动态学习和调整，显著提高实验效率。MOOSE-Chem3 的核心优势在于：

（1）实时优化：根据实验结果，动态调整所有假设的优先级

（2）高效决策：帮助科学家选出下一个最具潜力的实验方向

（3）减少试错：最大限度节省实验次数与资源投入

3.2 CSX-smi：高保真实验模拟器

由于真实化学实验成本高昂，难以大规模用于 AI 训练，MOOSE-Chem3 开发了一个名为 CSX-smi 的高保真实验模拟器。该模拟器基于前文提到的三个核心假设（A1、A2、A3），能够精确模拟化学实验的反馈过程。CSX-smi 模拟器的性能在实验验证中表现出色：

（1）趋势预测：在 30 组实验中，CSX-smi 的预测 Spearman 相关系数高达 0.96，其中 26 组实验的预测趋势与真实结果完全一致。

（2）数值准确性：均方根误差仅为 0.213，显示了极高的预测准确性。

这意味着 CSX-smi 能够以极高的精度模拟真实化学实验，为 AI 系统提供可靠的学习环境。

3.3 CSX-Rank：聚类驱动的实验引导假设排名方法

基于 CSX-smi 模拟器，MOOSE-Chem3 开发了名为 CSX-Rank 的聚类驱动假设排名方法。CSX-Rank 采用四步迭代策略：

3.3.1 功能组分提取、分类与聚类

首先，AI 将每个候选假设分解为不同的功能化学组分（即可能对目标反应机制有贡献的独特子结构或基序）。随后，这些组分被分类为：有效、不确定和无效。无效组分被直接排除，剩余组分根据功能相似性进行聚类，每个聚类代表对解决问题的一种独特机制贡献。

这一步骤可以形式化表示为：对于假设集合 $H=\{h_1,h_2,...,h_n\}$，每个假设 h_i 被分解为功能组分集合 $C_i=\{c_{i1},c_{i2},...,c_{im}\}$，然后对所有功能组分进行聚类，形成聚类集合 $G=\{g_1,g_2,...,g_k\}$。

3.3.2 智能聚类与假设选择

在大型语言模型预训练的化学知识引导下，系统识别出最有可能包含与研究问题高度相关组分的聚类。在此基础上，LLM 智能体根据组分相关性和先验知识，在该聚类中选择一个最有前景的假设。这一步骤可以形式化为选择最优聚类 $g^* \in G$ 和最优假设 $h^* \in g^*$，使得期望性能最大化：

$$g^*,h^*=\arg\max_{g\in G,h\in g} E[f(h,q)]$$

3.3.3 模拟实验执行与结果分析

选定的假设被输入到 CSX-smi 模拟器中进行评估，返回标准化性能得分。随后，AI 对模拟实验结果进行深入分析，评估所选聚类的有效性，并验证或更新已有的机制假设。

3.3.4 迭代总结与持续优化

在每次模拟实验评估后，系统进行详细分析，并将获得的分析整合到持续更新的累计总结中。这份总结综合了之前所有分析的见解，突出显示有效的聚类，并为未来的假设和聚类选择提供具体指导。

3.4 MOOSE-Chem3 的实验验证与性能评估

MOOSE-Chem3 在 TOMATO-chem 数据集（包含 1 个“最优假设”和 63 个负样本，共 64 个假设）上进行了测试。实验结果显示，CSX-Rank 将识别最优假设的平均实验次数（N_{trials}）从基线的 32 次降至 15 次，显著提高了实验效率。此外，研究团队还通过在模拟器中引入不同等级的噪声验证了 CSX-Rank 的鲁棒性。结果表明：

（1）随着噪声复杂性的增加，所有方法的性能都逐渐下降。

（2）CSX-Rank 始终优于其消融变体，即使在复杂噪声下也保持了显著的效率优势。

（3）这些结果突显了功能聚类和反馈分析在减轻误导信号和保持搜索效率方面的鲁棒性，验证了 MOOSE-Chem3 在实验引导假设排名中的有效性。

4. AI 辅助实验系统的应用价值与潜在影响

4.1 降低科研成本与缩短研发周期

AI 辅助实验系统如 MOOSE-Chem3 有望显著降低科研成本并缩短研发周期。以新药研发为例，传统药物发现过程中，研究人员可能需要合成和测试数千个化合物，每个化合物的合成和测试成本从数百到数千美元不等。而 AI 辅助系统通过更精准地指导实验方向，可以将所需实验次数减少 50%以上，潜在节省数百万美元的研发成本。

4.2 促进跨学科合作与知识融合

AI 辅助实验系统能够整合来自不同学科的知识和方法，促进跨学科合作。例如，MOOSE-Chem3 系统融合了化学、计算机科学和机器学习的知识，创造出超越单一学科边界的研究工具。这种跨学科融合有望催生新的研究范式和方法论，推动科学发现的范式转变。

4.3 加速科学发现与创新

AI 辅助实验系统最重要的价值在于加速科学发现和创新。通过更高效地探索假设空间，这些系统能够帮助科学家更快地发现新材料、新药物和新机制。例如，DeepMind 的 AlphaFold2 在蛋白质结构预测领域的突破，将原本需要数月甚至数年的结构解析工作缩短至几小时，极大加速了生物医学研究进程。

总结：从传统的“实验前假设排名”到新兴的“实验引导的假设排名”范式，AI 辅助实验系统正在经历根本性变革。以 MOOSE-Chem3 为代表的系统通过功能组分提取、智能聚类、模拟实验执行和迭代总结四步策略，实现对科学假设的动态评估与优化，显著提高了实验效率和科学发现速度。研究表明，这类系统有望大幅降低科研成本，缩短研发周期，促进跨学科合作，加速科学创新。虽然当前系统仍存在模拟器精度限制、领域知识依赖、解释性挑战等局限性，但通过多模态融合、人机协作优化、实验机器人集成等方向的发展，这些挑战有望逐步克服。

参考文献：

[1]李盛阳,刘康,刘云飞,等.数智驱动的空间科学实验研究：AI4S 范式下的新探索[J].中国科学院院刊,2025,40(02):371-379.

[2]邓新凯.基于生成式人工智能的数字代理人实验在社会科学中的应用与方法论探索[J].中国战略新兴产业,2025,(05):56-58.

[3]李钰祥.智能科学与技术专业实验平台的建设[J].电子技术与软件工程,2019,(06):232.DOI:10.20109/j.cnki.etse.2019.06.178.

作者简介：寇亚凤，（1974.05.18），女，汉族，吉林省，长春市，长春经济技术开发区世纪小学，中级教师，大学本科，教育教学方向。