



Biomni: A General-Purpose Biomedical AI Agent

<https://doi.org/10.1101/2025.05.30.6>
<https://biomni.stanford.edu>

汪静雅

2025.09.17

研究背景: 随着复杂的实验室实验、大型数据集、分析工具和大量文献的增长, 生物医学研究越来越受到重复和分散的工作流程的限制, 这些人工工作流程减缓了发现速度, 限制了创新, 这突显出需要一种全新的方式来释放研究潜力。

研究重点:

从25个生物医学领域的大量论文中创建了首个统一的Agentic环境:

- 核心工具,
- 数据库,
- 实验步骤

LLM



- 检索增强规划
- 基于代码的执行

动态组合并执行复杂的生物医学 workflows。

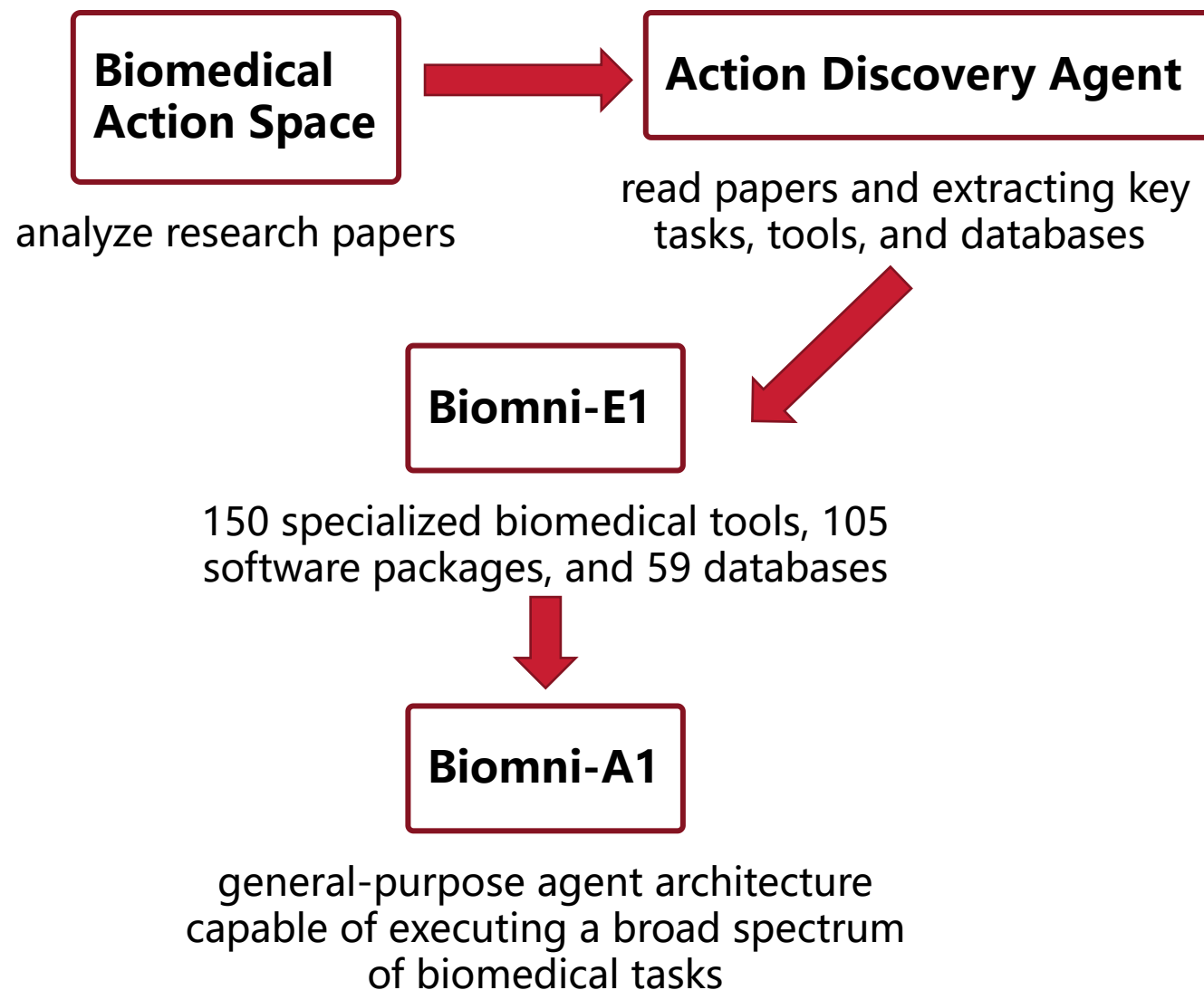
研究成果:

- Benchmark表明: Biomni 无需预设模板或固定流程, 即可在多种异构生物医学任务 (基因因果排序、药物重定位、罕见病诊断、微生物组分析、分子克隆等) 上实现强泛化, 且无需任务级提示调优;
- 真实案例表明: Biomni能解读复杂多模态数据并自主生成可实验验证的方案。

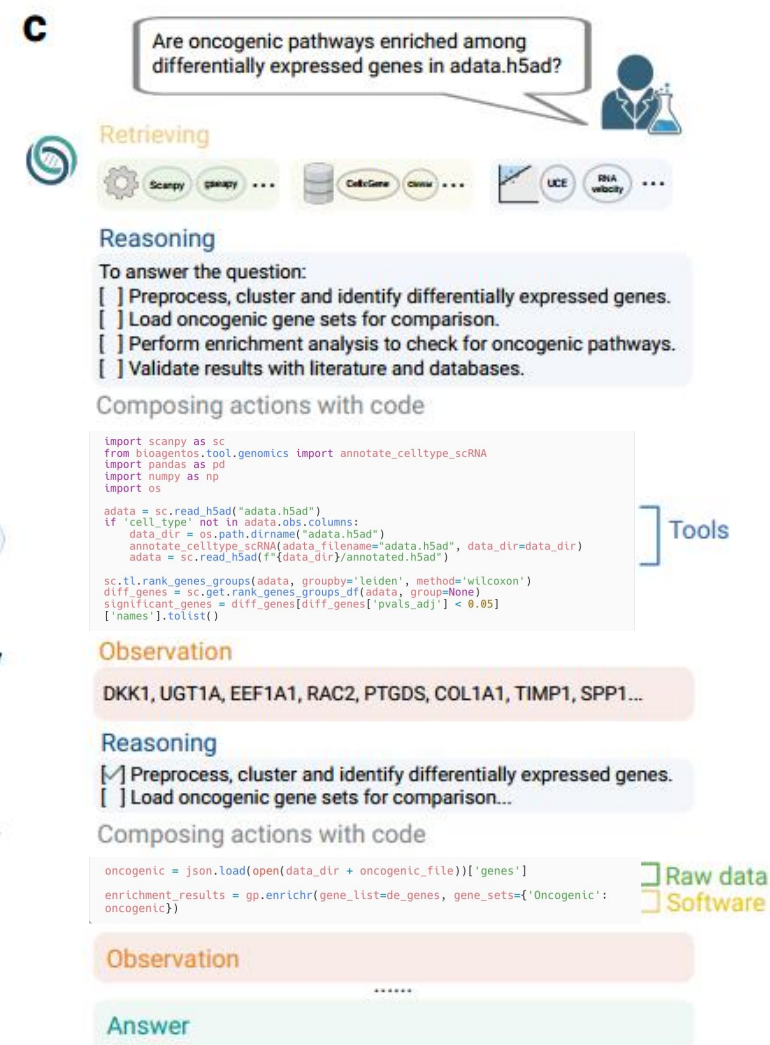
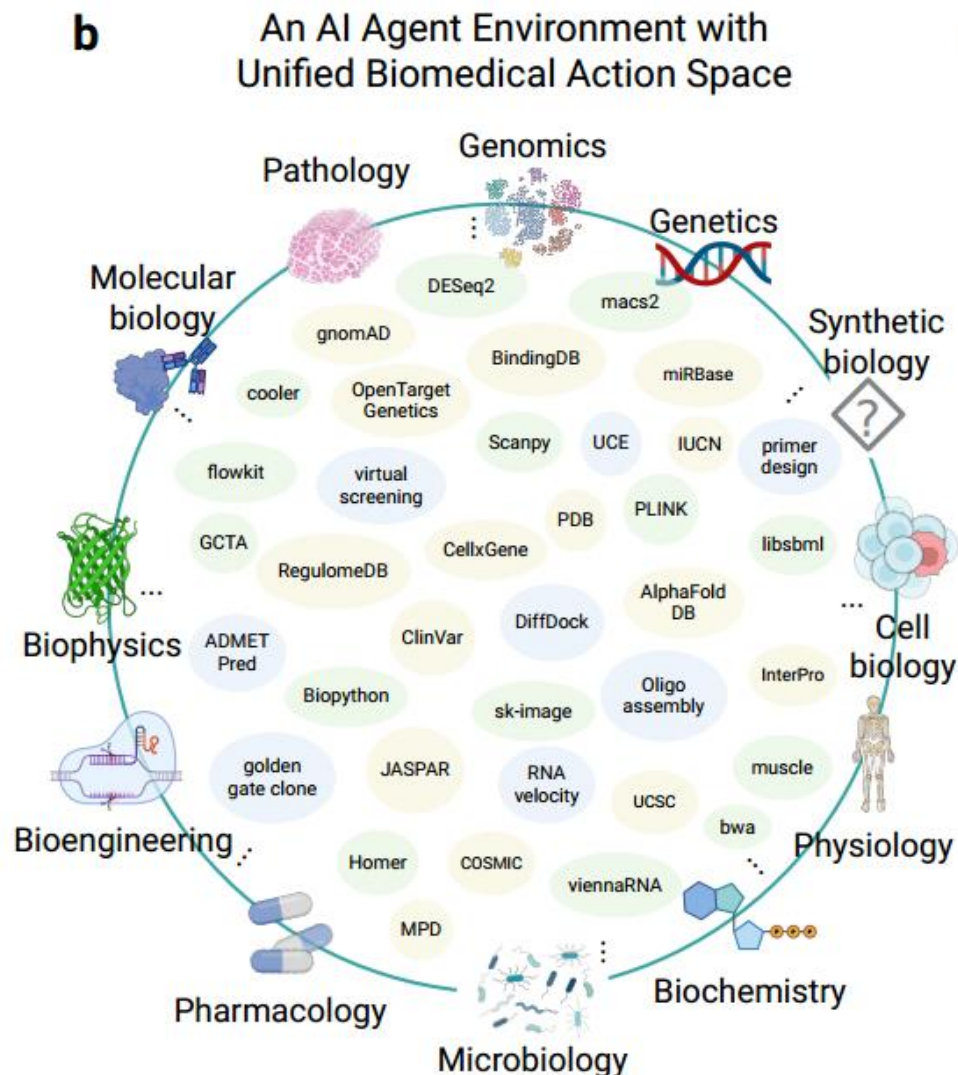
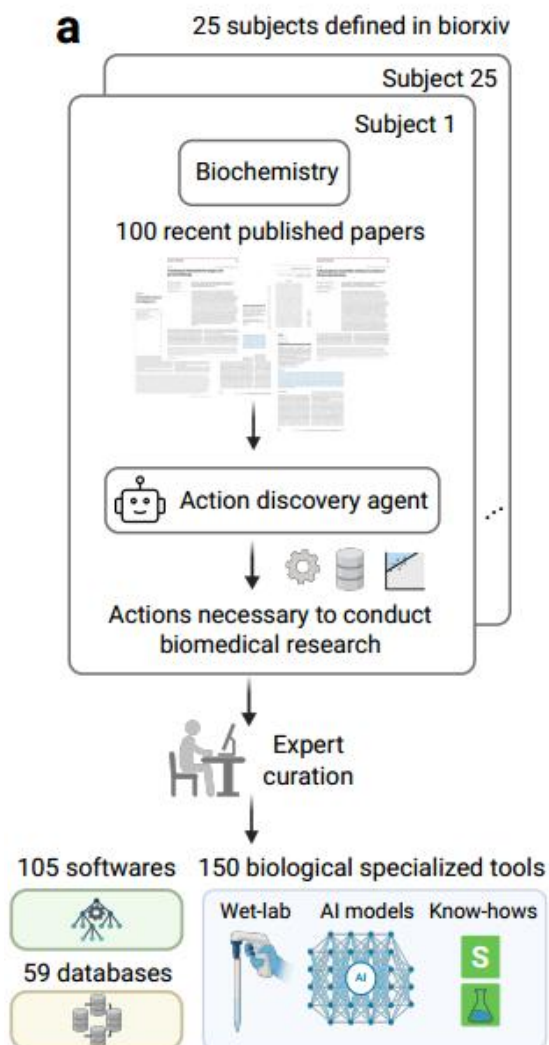
目前研究局限:

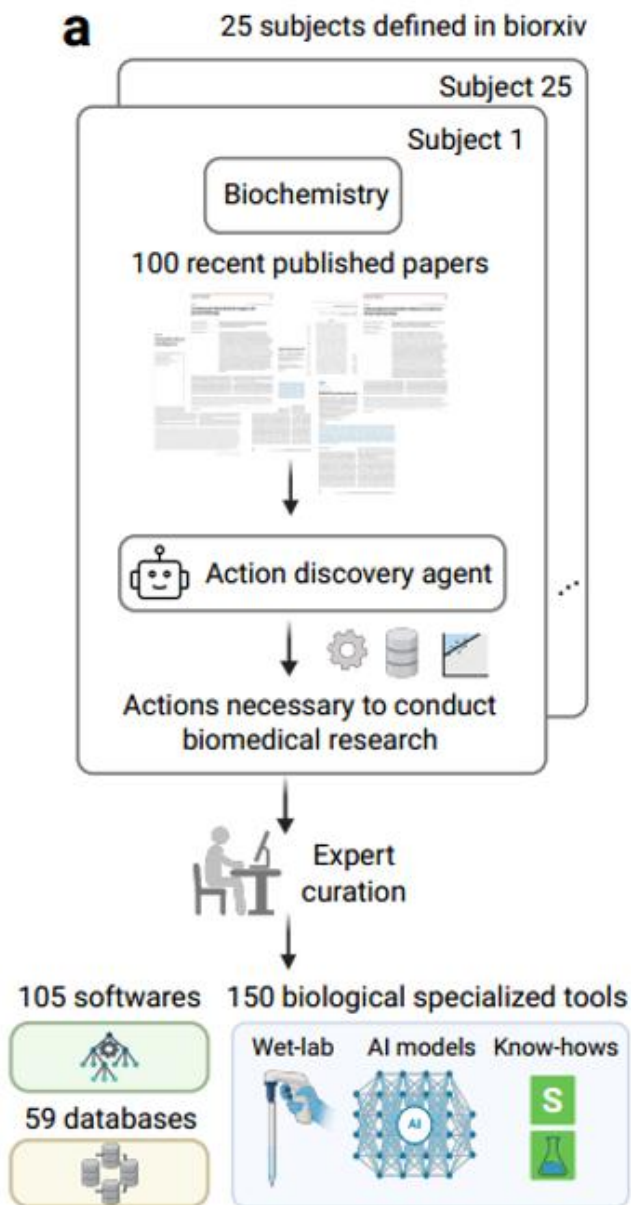
- 处理庞大的生物医学任务会带来重大技术挑战，需要极大时间精力成本；
 - 之前的智能体依赖固定的工作流，导致其只能处理及其受限的任务而无法泛化；
 - 大型语言模型（LLM）的推理能力依赖一个明确定义了“生物医学”的环境。
- 一个真正具备能力的系统需要一种智能体架构，能够原生地与该生物医学环境交互，自主选择 and 组合动作，并运用其推理能力来规划与执行多样化任务，而无需依赖僵化、预定义的工作流。

Biomni架构:



Biomni 的推理与行动组合示例





Action Discovery LLM Agent

1. 以2024年**bioRxiv**的**25个学科**分类为框架，每类选取最新**100**篇预印本；
2. 由**Action Discovery LLM Agent**逐篇提取和解析它们的 PDF 内容。
3. 每篇论文都分块处理，由一个专门的提示引导大型语言模型（LLM）逐块明确识别并提取**任务、软件和数据库**。（对于任务，LLM 被指示突出显示在生物医学研究工作流程中需要专门实现的重复任务）
4. 这些资源共同构成完成大量生物学研究任务所必不可少的 **“动作”集合**。

Action Discovery LLM Agent

```
class PaperTaskExtractor(base_agent):
    """Agent that extracts generalizable tasks/experiments from academic papers.
    It processes papers in chunks and identifies common research tasks that could be shared across papers.
    """

    def __init__(
        self,
        llm="claude-3-7-sonnet-20250219",
        cheap_llm=None,
        tools=None,
        chunk_size=4000,
        chunk_overlap=400,
    ):
        """Initialize the PaperTaskExtractor agent.

        Args:
            llm (str): The LLM model to use
            cheap_llm (str, optional): A cheaper LLM for simpler tasks
            tools (list, optional): Any tools to use (not needed for this agent)
            chunk_size (int): Size of text chunks for processing
            chunk_overlap (int): Overlap between chunks
        """
```

```
def process_paper(self, paper_text: str) -> dict[str, list[dict[str, Any]]]:
    """Process a paper and extract generalizable tasks/experiments, databases, and software.

    Args:
        paper_text (str): The full text of the paper

    Returns:
        Dict[str, List[Dict[str, Any]]]: A dictionary with tasks, databases, and software
    """

    # Split the paper into chunks
    text_splitter = RecursiveCharacterTextSplitter(
        chunk_size=self.chunk_size,
        chunk_overlap=self.chunk_overlap,
        length_function=len,
        separators=["\n\n", "\n", ". ", " ", ""],
    )
    chunks = text_splitter.split_text(paper_text)

    # Process each chunk to extract tasks
    chunk_results = []
    for i, chunk in enumerate(chunks):
        print(f"Processing chunk {i + 1}/{len(chunks)}...")
        chunk_tasks = self._process_chunk(chunk)
        chunk_results.append(chunk_tasks)

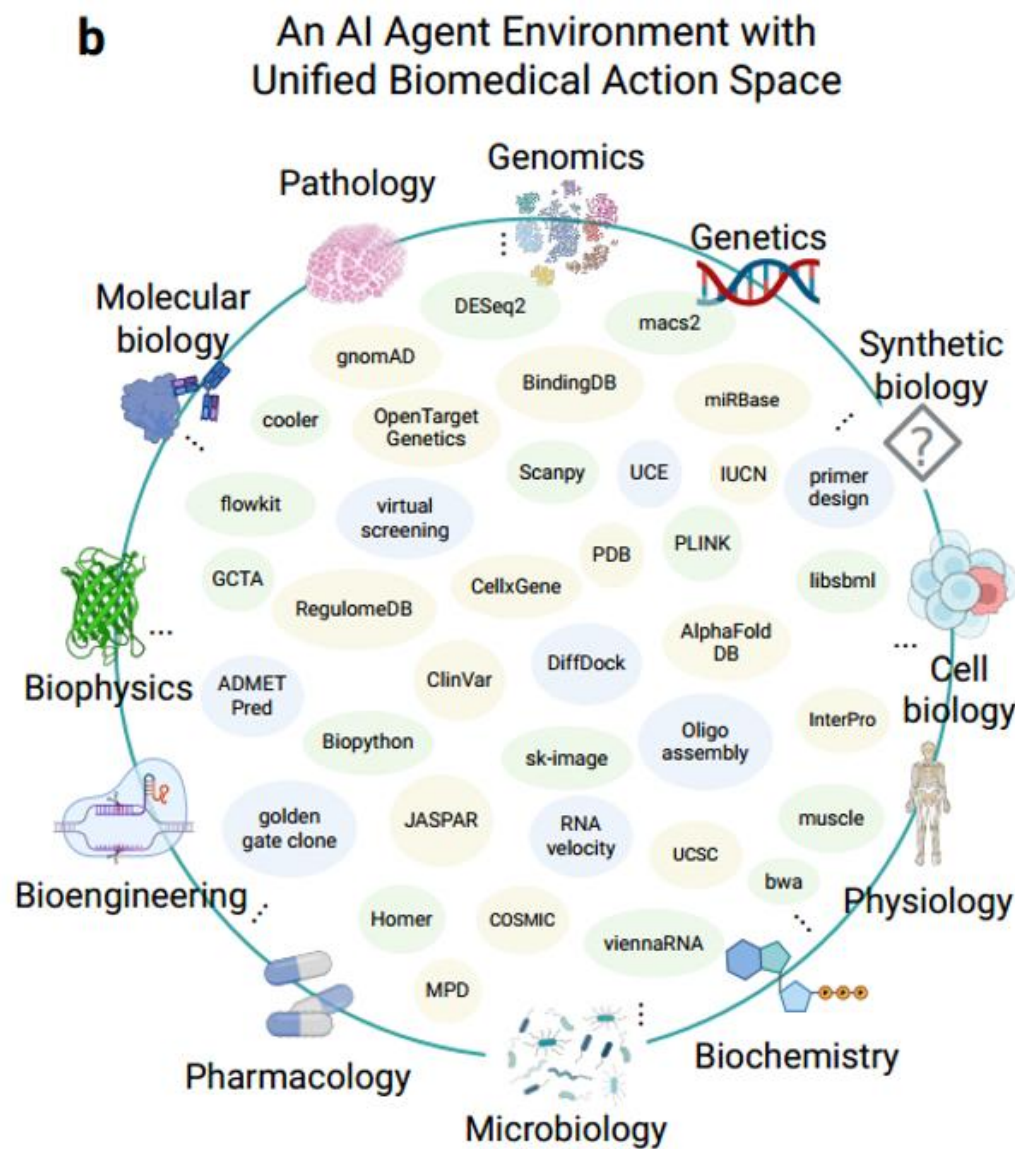
    # Consolidate tasks from all chunks
    consolidated_results = self._consolidate_tasks(chunk_results)
    return consolidated_results
```

两种任务提示词:

- **analyzing paper chunks:** 用于从论文片段中提取专业和常见的任务、数据集、软件
- **consolidating tasks:** 根据任务列表将信息规范成json格式

1. 调用LangChain的 **RecursiveCharacterTextSplitter** 对论文进行分块;
2. 对分块论文进行 **_process_chunk** 处理
3. 对2中得到的任务列表进行 **_consolidate_tasks** 处理

env_collection.py (智能体设置) → **extract_biorxiv_tasks.py** (执行提取任务) → **generate_function.py** (代码生成)

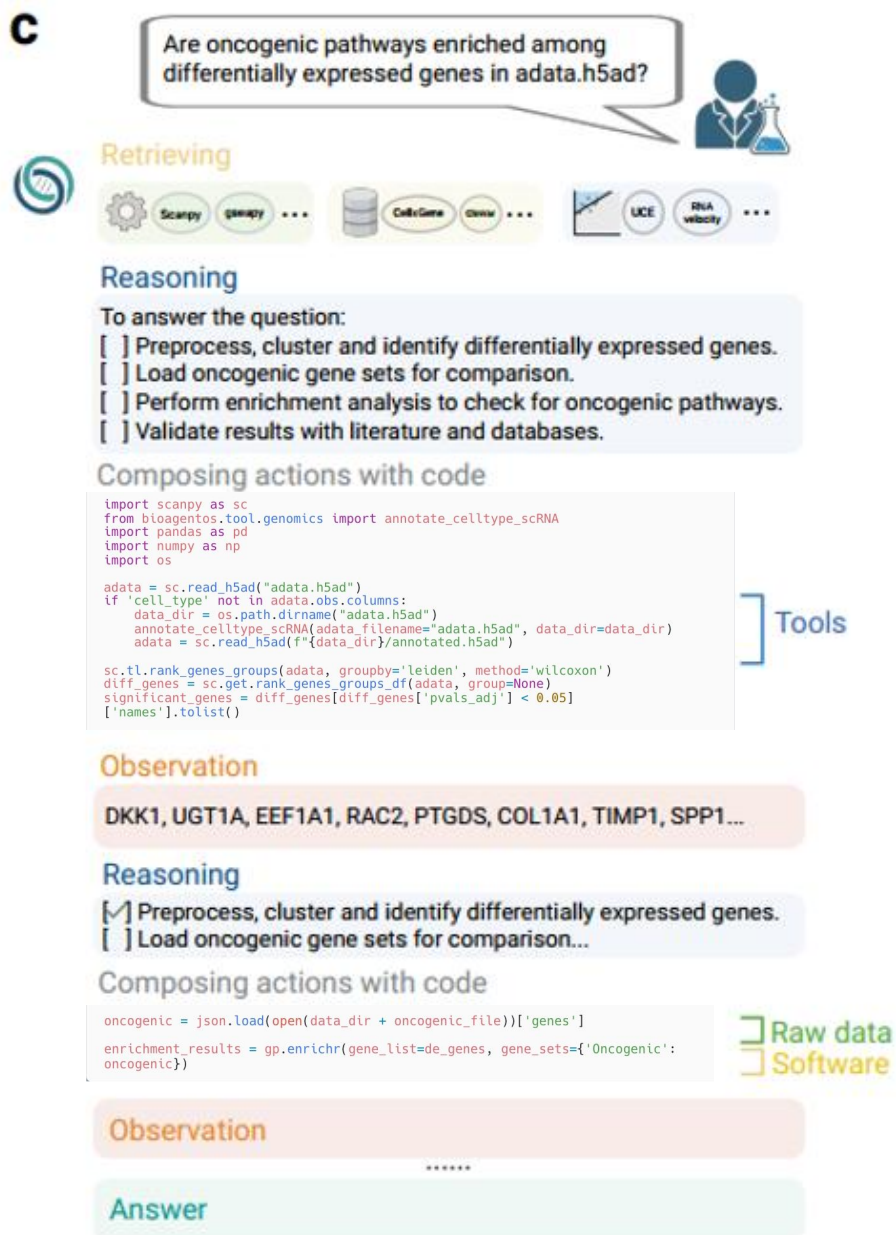


Biomni-E1

- 一个面向生物医学智能体的综合实验环境，包括 **150个专门的生物医学工具、105个软件和59个数据库**。
- 1. **工具**：所有入选工具均经人类专家严格验证，并配有对应测试用例。这些工具的特征包括：非平凡性，复杂代码、领域专有知识或专用 AI 模型。
- 2. **软件**：在环境中预装了 105 个广泛使用的生物软件包，支持 **Python、R 与 Bash** 脚本。
- 3. **数据集**：
 - ① 通过 **API 访问**的开源数据库：为每库实现统一函数；函数接收自然语言查询，内部由大模型解析库表结构并动态生成可执行语句。
 - ② **无 Web 接口**的数据库：下载至本地数据库，预处理为结构化 pandas DataFrame，与智能体无缝集成。

Biomni-E1：工具的具体实现

- 1. 工具筛选：**根据**任务与主要研究兴趣的相关性**（药物发现和临床生物医学领域）进行筛选。随后，这些任务被缩小到大约 1,900 个常见的重复任务。这些任务经过进一步人工审查以消除冗余，强调选择需要**显著领域专业知识的高度专业化任务**。
- 2. 工具实现：**人类科学家与具备网络搜索能力的软件工程Agent合作，以实现每个专业化工具。每个工具都经过严格的验证，需要一个明确定义的测试用例才能成功通过。这一严格的过程最终形成了一个包含 **150 个专业化工具**的精选集合。此外，还包括了 **PubMed** 和 **Google Scholar** 等基本文献检索工具，并为未来的迭代扩展提供了准备。
- 3. 工具规范化：**每个工具都使用一个全面的**检查表**进行严格定义，该检查表要求：
 - ① 清晰且描述性的名称
 - ② 详细的文档
 - ③ 为 LLM 解释优化的详细研究日志格式的输出
 - ④ 包含并成功通过特定测试用例
 - ⑤ 专业化标准：如果一个任务可以通过简短的 LLM 生成代码（例如，简单的数据库查询）轻松实现，则不创建专业化工具。



Biomni-A1

- 一种专门的**智能体架构**，它不再为每项任务硬编码固定流程，动态组合智能动作，与软件、数据及工具交互。
- 1. 引入了**基于 LLM 的工具选择机制**，可针对用户目标动态检索并定制所需的生物医学工具子集，以应对工具的高度复杂与专业性。
- 2. 以**“代码”作为通用动作接口**，可编排并执行包含循环、并行、条件分支等复杂工作流；该方式还能让智能体交错调用软件、工具、数据库及原始数据操作，即使这些资源没有预定义函数签名，也能灵活动态地整合异构资源。
- 3. 智能体采用**自适应规划策略**，先基于生物医学知识制定初始方案，并在执行过程中持续迭代优化，实现响应式、情境感知的行为。

Biomni-A1：具体工作流程

1. 基于 **CodeAct** 框架构建，旨在通过结合大型语言模型（LLMs）与交互式编码环境来系统地解决生物学任务。
2. 给定用户查询，Biomni 首先提示 LLM 生成一个清晰的、编号的**步骤列表**，详细说明解决给定问题所需的步骤，并仔细跟踪进度和调整。
3. 由于工具、软件和数据库空间非常庞大，查询任务可能仅使用这些资源的一小部分。为了避免过长的上下文，利用一个由**单独的 LLM 提供支持的基于提示的检索器**，智能体可以从可用资源中动态选择最相关的功能、数据集和软件库。
4. 在执行过程中，LLM 生成代码，在编码环境（Python、R 或 Bash）中执行，并返回结果观察以指导后续推理。这种**迭代方法**持续进行，直到智能体收敛到一个准确、经过验证的解决方案。

工具导入与选择：

```
read_module2api()  
ToolRegistry(module2api)  
ToolRetriever()
```



提示词设置：

```
A1._generate_system_prompt  
制定计划，并在每一步之后  
更新显示；在回复中可选择  
用<execute>执行代码，或用  
<solution>返回最终回复
```



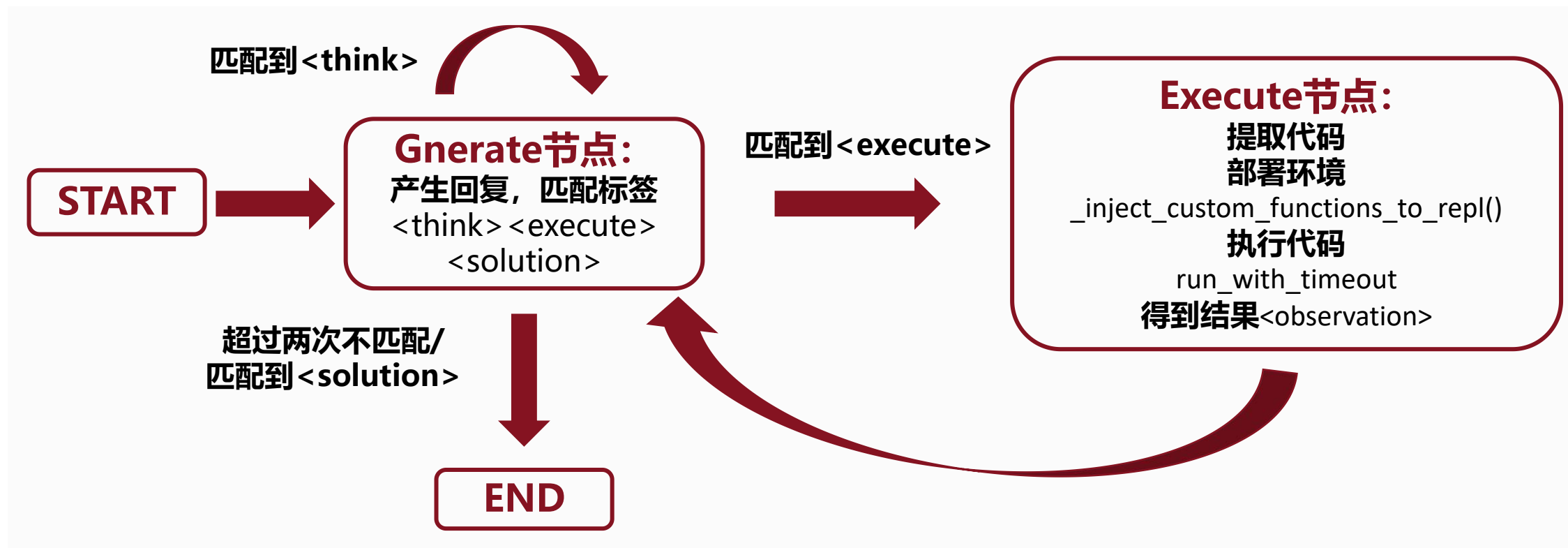
系统设置：

```
A1.configure  
设置数据、工具、软件  
更新系统提示词
```

Biomni-A1: WorkFlow Configure

AgentState

- messages: list[BaseMessage]
- next_step: str | None



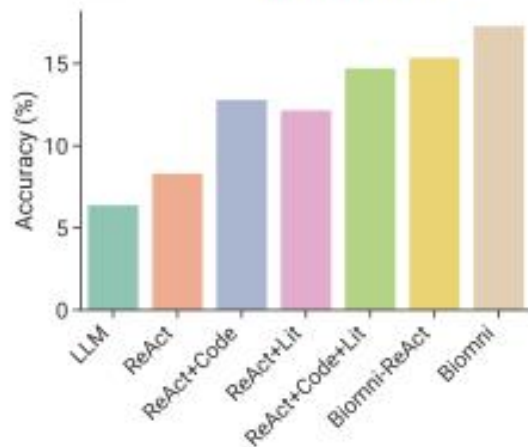
Biomni-A1: go

工具准备

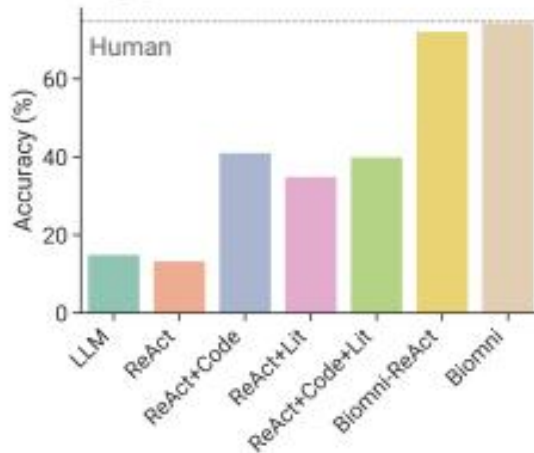
- `_prepare_resources_for_retrieval`
- `update_system_prompt_with_selected_resources`

实际执行

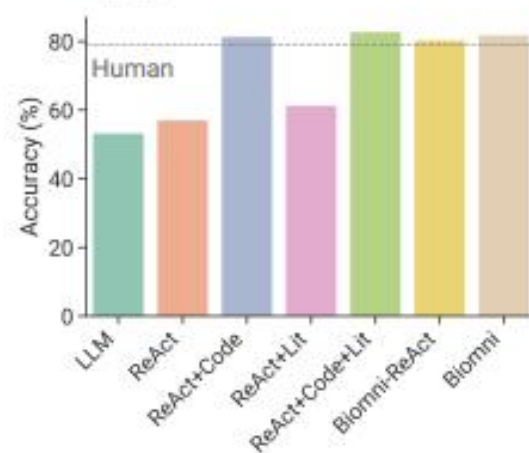
```
for s in self.app.stream(inputs,
stream_mode="values", config=config)
```

a Humanity's Last Exam: Biomedicine

LAB-Bench: DbQA



LAB-Bench: SeqQA



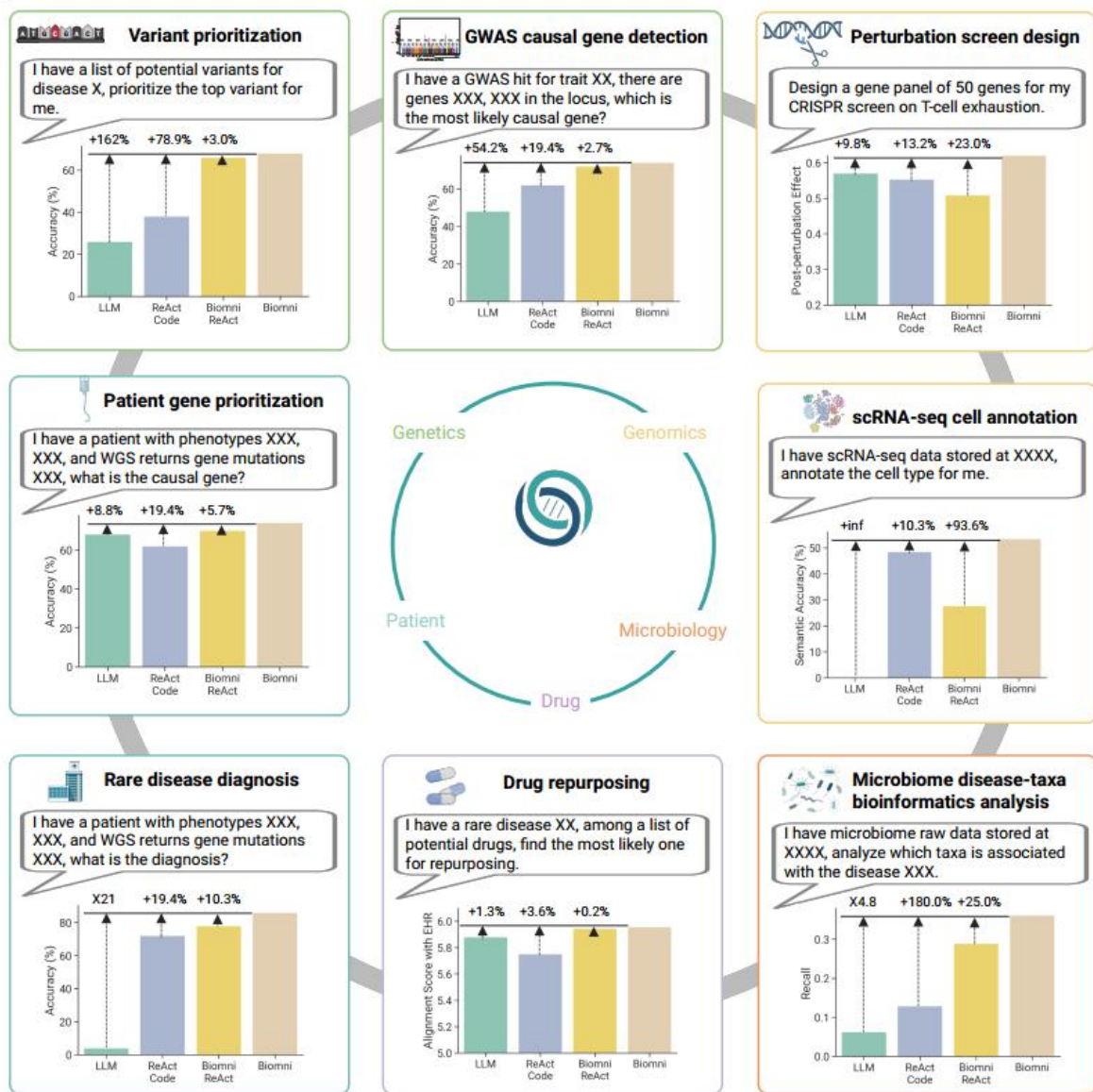
在**三项高难度的多选题基准**上对其评估：**Humanity' s Last Exam** 和 **LAB-Bench**。这些任务涵盖工具使用、符号推理及结构化生物信息检索，是衡量稳健生物医学 AI 智能体的核心能力。

- ✓ **LAB-Bench**: 先用 45 题开发集调优工具/数据库接口，再在 315 题隐藏测试集评估（3 次独立运行平均）。
 - **DbQA**（需对生物数据库做结构化查询）：Biomni 准确率 74.4%，持平专家人类（74.7%），超越所有基线（含 ReAct+Code 编码智能体 40.8%）。
 - **SeqQA**（需对 DNA/蛋白质序列推理）：Biomni 准确率 81.9%，再次超过人类水平（78.8%）。
- ✓ **HLE**: 抽取了涵盖 14 个生物医学子领域的 52 题子集对 Biomni 进行评估。Biomni 取得 17.3 % 的准确率，显著高于基础大模型（6.0 %）、编码智能体（12.8 %）和文献智能体（12.2 %）。

结果表明，Biomni 无需任何任务特定调整即可在陌生、开放的生物医学领域中实现泛化。

03 Results-Generalizes to new, real-world biomedical tasks across diverse subfields

b

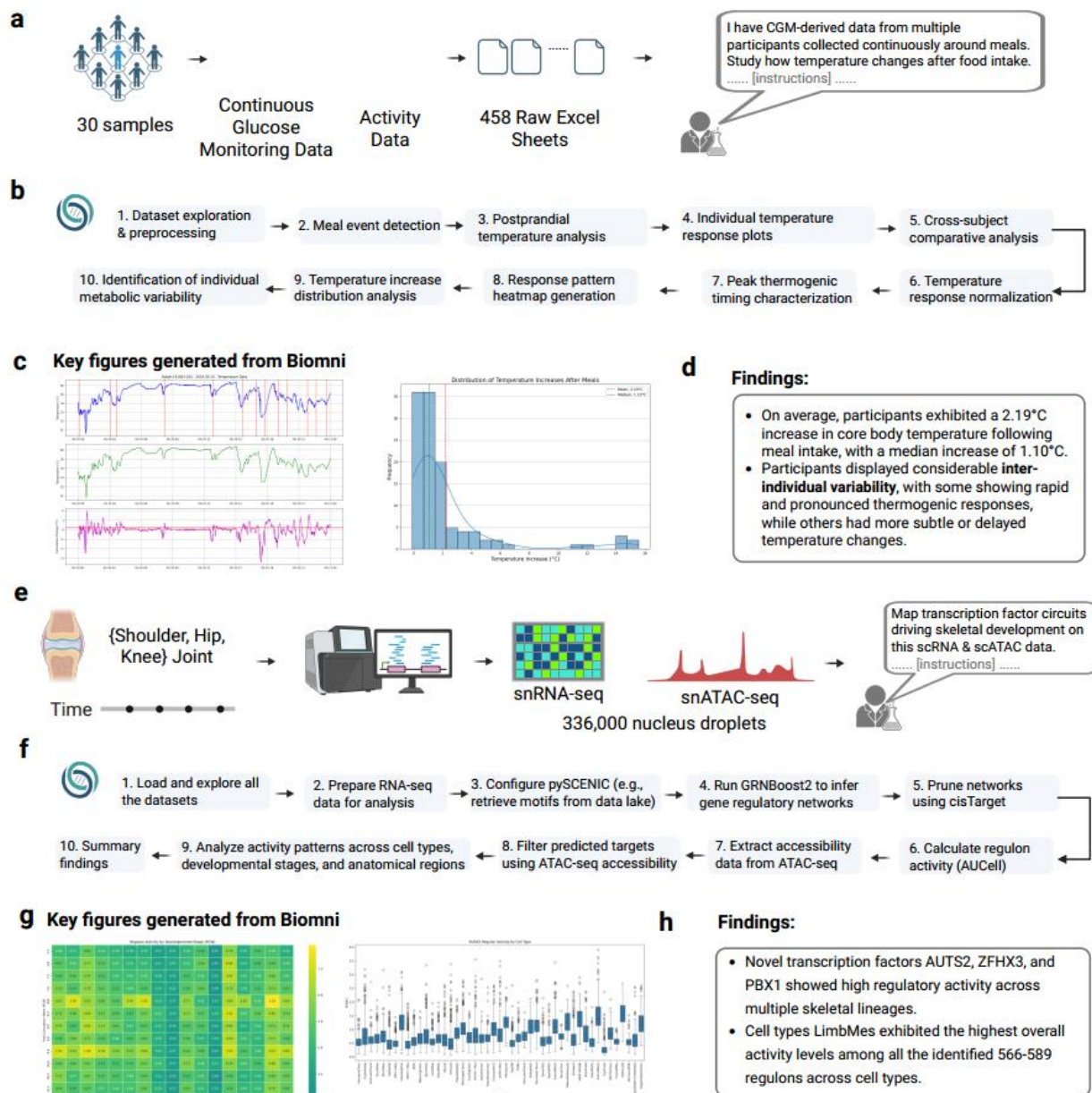


Biomni 可泛化到不同子领域的全新真实生物医学任务。为评估其在现实研究任务中的泛化能力，我们构建了 **8 个全新的生物医学基准**：

1. 变异优先级排序；
2. GWAS 因果基因检测；
3. CRISPR 扰动筛选设计；
4. 罕见病诊断；
5. 药物重定位；
6. 单细胞 RNA-seq 注释；
7. 微生物组疾病-分类群分析；
8. 患者基因优先级排序。

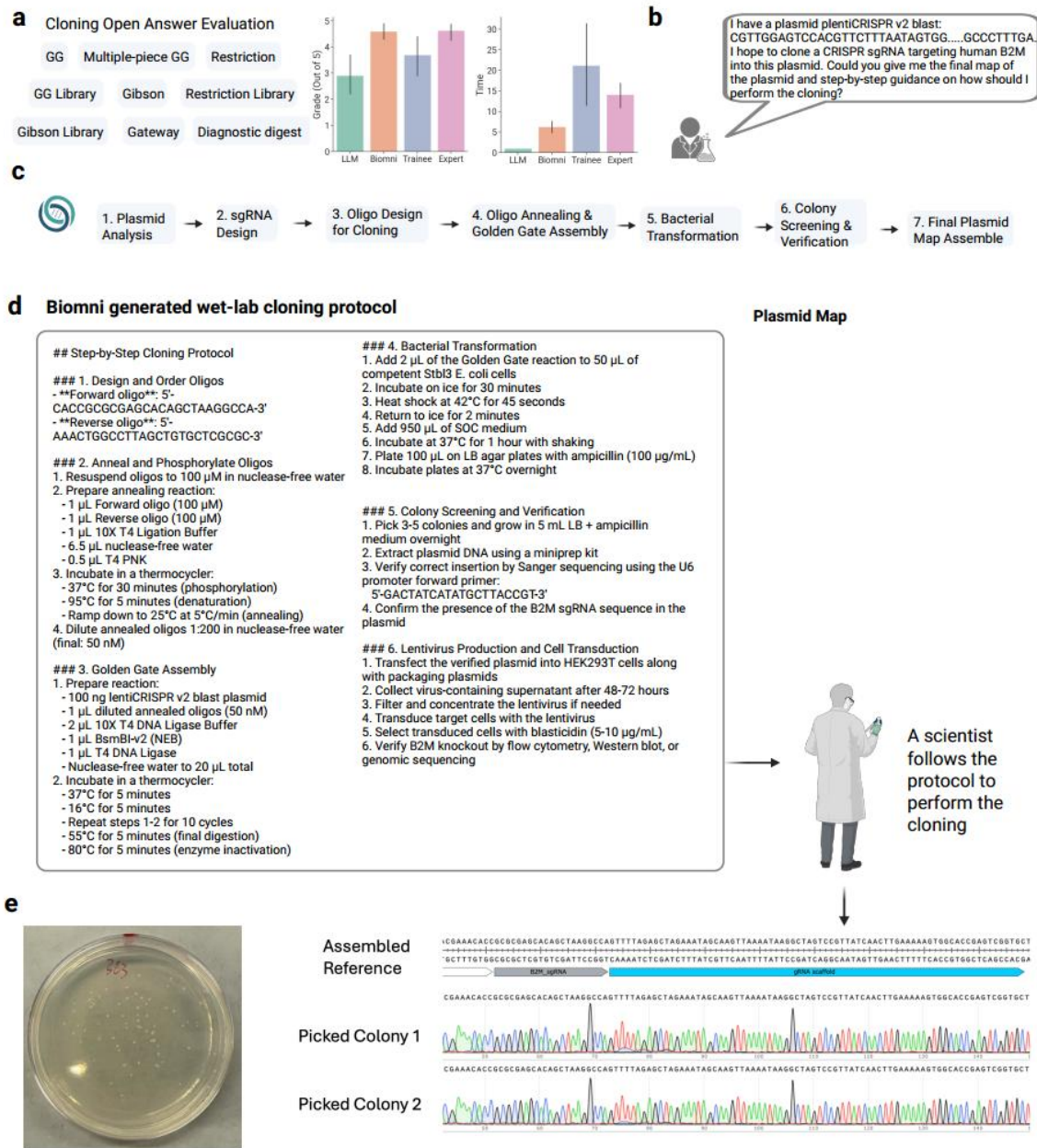
在所有任务中，**Biomni 的平均相对性能比基础大型语言模型（LLM）高出 402.3%，比编码智能体高出 43.0%，比其自身经过简化的变体 Biomni-ReAct 高出 20.4%。**

03 Results-Executes complex multi-modal biomedical analyses to generate hypothesis



- Biomni **快速分析**了来自 30 个个体的 CGM 衍生的热生成反应数据和活动数据。
- 工作流程**: Biomni 自主执行数据预处理、餐食事件检测、餐后温度分析和热生成反应特征化。
- Biomni 自动生成的代表性个体温度-反应图和餐后温度增加分布。
- Biomni 识别的独特生物学发现总结, 包括餐后核心体温显著增加, 以及热生成反应中显著的个体间变异性。
- Biomni 自主分析了大约 336,000 个细胞核液滴的单细胞多组学数据, 涵盖了人类胚胎关节发育。
- 工作流程**: 用于多组学基因调控网络。
- Biomni 生成的两个关键图: 左侧面板显示了发育阶段的调控因子活动热图, 颜色强度表示活动水平。右侧面板显示了 **RUNX2** 调控因子活性的箱线图, 展示了不同细胞群体间的表达变异。
- GRN 分析的关键发现: 1) 主要转录因子在多个骨骼谱系中显示出高调控活性, 尽管之前与骨骼发育没有关联, 2) 在恢复的 566-589 个调控因子中, 肢体间充质细胞显示出最高的平均调控活性得分, 强调了它们在骨骼转录控制中的显著作用。

03 Results-Designs wet-lab experimental protocol



- a. 在 **10 个真实克隆场景**中进行开放式克隆基准测试。将 Biomni 与基础大型语言模型（LLM）、训练水平的人类和专家水平的人类科学家进行了比较。我们发现 **Biomni 的准确性与专家级科学家相似，并且比训练水平的准确性显著更高，同时所用时间更少。**
- b. 用户请求 Biomni 克隆一个针对人类 B2M 基因的 sgRNA 到 lentiCRISPR v2 Blast 质粒中。
- c. Biomni 的自动化步分工作流程，包括质粒分析、sgRNA 设计、寡核苷酸合成、Golden Gate 组装、细菌转化、菌落筛选和最终质粒映射。
- d. Biomni 生成的详细克隆实验步骤，包含逐步说明和全面的质粒图谱，使实验室科学家能够自主执行实验。
- e. 通过在筛选板上成功生长的菌落验证 Biomni 的克隆方案，随后通过 Sanger 测序确认 sgRNA 插入在挑选的菌落中完美对齐，展示了 Biomni 在精确可靠实验设计方面的强有力能力。

实际应用

生物制药的靶点和药物发现

优先排序目标，设计扰动筛选，或重新利用药物

临床应用环境

基因优先排序和罕见病诊断

消费者健康

可穿戴设备数据和多组学分析的整合

局限性

- 评估的任务仅代表某些领域的一个子集，**关键领域仍未被探索。**
- 在归纳工具和数据集时只考虑了近期文献，存在**忽视基础概念和技术**的风险。
- 在需要细致的临床判断、新颖的实验推理、分析发明或深度生物学思维和综合的领域中仍然存在困难。



强化学习

多模态数据

自动发现和整合新工具和数据库

结合更多传统方法



请批评指正！

<https://doi.org/10.1101/2025.05.30.6>
<https://biomni.stanford.edu>

汪静雅

2025.09.17