SciDataCopilot：面向通用人工智能驱动科学发现的智能体数据准备框架

Jiyong Rao*、Yicheng Qiu*、Jiahui Zhang*、Juntao Deng*、Shangquan Sun*、Fenghua Ling、Hao Chen、Nanqing Dong、Zhangyang Gao、Siqi Sun、Yuqiang Li、Dongzhan Zhou、Guangyu Wang、Lijun Wu、Conghui He、Xuhong Wang、Jing Shao、Xiang Liu、Yu Zhu、Mianxin Liu、Qihao Zheng、Yinghui Zhang、Jiamin Wu、Xiaosong Wang、Shixiang Tang、Wenlong Zhang、Bo Zhang、Wanli Ouyang、Runkai Zhao ✉、ChunfengSong ✉、LeiBai ✉、Chi Zhang ✉ †

上海人工智能实验室 * 共同第一作者 ✉ 通讯作者

†项目负责人✉zhangchi2@pjlab.org.cn 项目主页

arXiv:2602.09132v1 [cs.DB] 9 Feb 2026

当前科学人工智能（AI4S）主要依托大规模文本语料库，生成式人工智能系统在假设生成、文献检索和多模态推理方面表现出色。然而，要加速闭环科学发现，原始实验数据的利用仍是一个关键瓶颈。原始数据具有极端异构性、高度特异性，且需要深厚的领域专业知识；它们既无法与语言表征直接实现语义对齐，也不具备适用于统一嵌入空间的结构同质性。这种脱节使新兴的科学通用人工智能（AGI4S）无法有效地与实验所处的物理现实交互。在这项工作中，我们将以文本为中心的人工智能就绪概念扩展至科学人工智能就绪数据范式，明确规范科学数据在计算工作流中的定义、结构化和构成方式。为将这一理念付诸实践，我们提出*SciDataCopilot*，这是一种自主智能体框架，旨在以端到端的方式处理数据摄取、科学意图解析和多模态集成。该框架将数据就绪度定位为核心操作原语，为可复用、可迁移的系统奠定了有原则的基础，从而推动向实验驱动的科学通用智能转型。在三个异构科学领域开展的大量评估表明，与人工流水线相比，SciDataCopilot提升了效率、可扩展性和一致性，数据准备速度最高可提升 30× 倍。

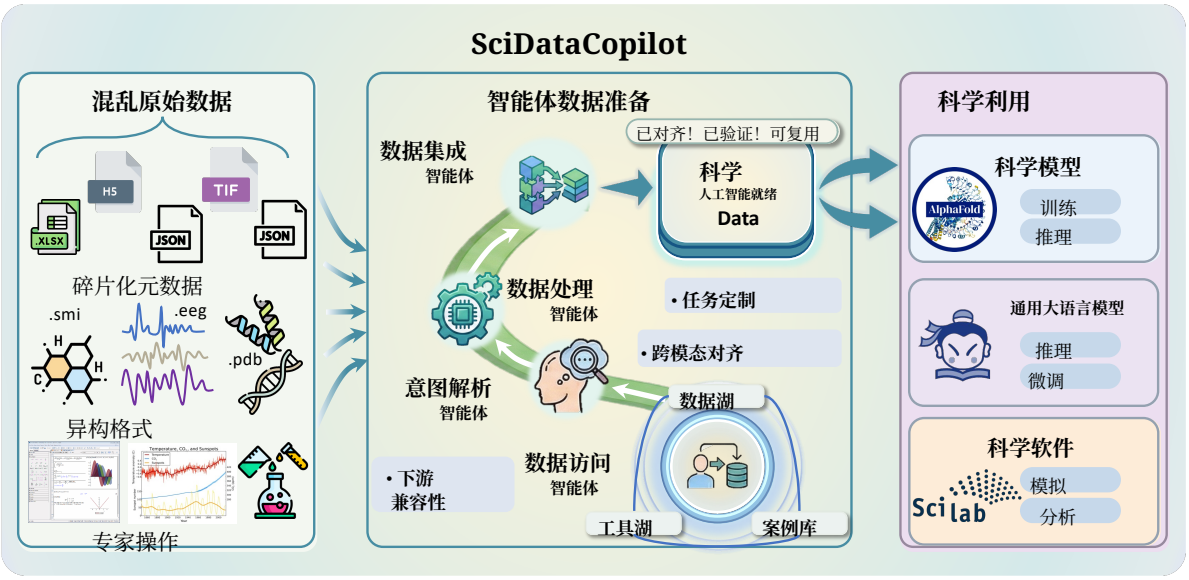


图 1 | 科学AI就绪范式将原始数据规范化为任务条件化、跨模态对齐的数据，以支持自主科学发现。提出的 **SciDataCopilot** 通过一个智能体框架将异构原始数据转化为科学人工智能就绪数据，从而实现这一范式。

目录

1 引言 3

2 相关工作 52.1 通用自动化数据准备框架 52.2 领域专用自动化数据准备框架 52.3 自动化数据准备的研究前沿 6

3 面向科学发现，重新思考数据智能体3.1 概念阐述 **7**
3.2 设计启示 8
8

4 SciDataCopilot 框架 84.1 数据访问智能体 94.1.1 科学数据库构建 104.1.2 数据访问智能体方法 114.2 意图解析智能体 124.2.1 科学数据需求分析 134.2.2 案例检索与适配 134.2.3 科学数据处理计划的生成与验证 144.3 数据处理智能体 144.3.1 处理流水线 154.3.2 自修复执行循环 174.3.3 可复现性与可追溯性 174.4 数据集成智能体 174.4.1 数据集成策略分析 184.4.2 生成数据集成流水线 18

5 用例5.1 案例 1：酶催化数据的自动化收集 5.1.1 问题陈述 **19**
5.1.2 实验设置 5.1.3 实验结果 5.1.4 总结 5.2 案例 2：端到端神 19
经科学数据分析 5.2.1 问题陈述 5.2.2 实验设置 5.2.3 实验结果 19
5.2.4 总结 5.3 案例3：跨学科地球数据准备 5.3.1 问题陈述 19
5.3.2 实验设置 5.3.3 实验结果 5.3.4 总结 21
21
22
22
23
23
25
26
26
27
27
27

6 结论 **27**

参考文献 **29**

1. 引言

生成式人工智能 [1, 2, 3, 4, 5] 正在推动科学发现领域的范式转变，使传统研究实践迈入科学人工智能（AI4S）的新时代。其影响贯穿研究生命周期的多个阶段，包括假设构建 [6, 7], 深度文献检索 [8, 9], 干实验/湿实验设计 [10, 11], 以及证据解读 [12, 13]。Intern-S1 [14] 和 ScienceOne-AI [15] 等代表性成果提供了实证证据：以论文、专利和技术报告的综合文本档案为条件时，科学智能便能大规模涌现。多智能体系统利用这些庞大的科学语料库作为基础数据源，例如 InternAgent [7], AI Co-Scientist [16, 17] 和 SciAgents [18], 能够自动化端到端的研究工作流。这些智能体协同编排复杂的发现流程，执行从构思到验证的各项任务，绕过传统以人为中心的方法。

然而，传统的人工智能就绪范式会将结构化科学数据线性化，迫使本质上复杂的表示形式转化为序列格式，以适配以大语言模型为中心的对齐学习。 [12,19, 20] 实践中，如图1所示，科学数据形式多样，既包括受化学与生物学有效性约束的结构化分子序列，也包括受依赖实验方案的预处理支配的高维神经记录，以及需要显式时空对齐的多源观测数据。这种复杂性揭示了一个根本性缺口：目前缺乏一种标准化的数据范式，能够系统地协调这些不同模态，使其转化为可供下游科学分析直接使用的格式[21]。

为弥合这一缺口，我们提出了**科学人工智能就绪数据范式**，这是一种统一的表述方式，以第一性原理为基石，开展基于人工智能的科学建模与推理。与以文本为中心的格式不同，它将原始异构数据转化为科学目标、领域特定约束条件和程序化实验逻辑的结构化表示。由此，为从任务条件化的 AI4S 迈向面向科学的通用人工智能（AGI4S）奠定必要基础，使人工智能成为能够与科学家直觉协同的合作伙伴，从而加速科学突破。

与优先考虑数据清洁度和以大语言模型为中心的拟合的传统人工智能就绪数据范式相比，科学人工智能就绪数据范式重新定义了科学数据的规范、组织和使用方式，以最大限度地提升其科学可用性。该范式的特点如下：

- **任务条件化原则：** 科学人工智能就绪数据范式将**科学任务作为首要组织原则**，把科学意图转化为所需的数据单元、变量和约束条件。这使任务条件化数据规范摆脱低效的手动收集，转而采用自动化的可复用工作流。
- **下游兼容性：** 科学人工智能就绪数据范式优先考虑与**下游科学分析的直接兼容性**，确保准备好的数据满足模型特定输入约束，并支持独立推理之外可组合、可执行的工作流。
- **跨域整合能力：** 科学人工智能就绪数据范式强调**有原则的跨模态与跨学科对齐**，从而能够在异构科学领域间系统地进行数据关联、检索、对齐和组合。

因此，上述区别体现了科学数据根本目标的范式转变：从用于模型拟合的数据，转向以科学任务为条件、约束一致且可供科学分析直接使用的数据。

将科学人工智能就绪数据范式付诸实施，在广度和深度上都带来了重大挑战。首先，科学数据准备本质上是任务条件化的，并源自多样化来源。这会造成多样化目标，进而导致碎片化的数据资产与处理

逻辑分散于不同实验与领域之中。此外，科学数据具有高度的结构与语义复杂性，并取决于任务意图、数据属性以及领域知识。在不同模态、尺度和实验情境之间对齐并组合数据的要求，进一步加剧了这些困难。当数据必须在异构模态、尺度和实验情境之间进行关联、对齐与组合时，这些挑战会被进一步放大。因此，这些特性使得单体式工作流难以胜任，因为它们缺乏灵活性，无法使处理逻辑适应不断变化的任务需求与数据条件。相反，它们促使采用一种分阶段但具有适应性的工作流，其中数据访问、意图解析、处理与集成通过协作智能体协同完成，每个智能体在整体运作的同时做出情境感知的决策。

基于这一观察，我们提出了**SciDataCopilot**，这是一个统一的智能体框架，通过分阶段工作流实现科学人工智能就绪数据准备。该框架的概览如图2所示，由四个协同工作的智能体构成：负责建立结构化数据关联的数据访问智能体 、负责检索并规划任务相关数据的意图解析智能体 、负责在明确约束条件下执行领域感知型转换的数据处理代理 ，以及负责跨模态和实验情境对齐并组合已处理数据单元的数据集成智能体 。这些智能体共同构建端到端流水线，涵盖从数据关联和检索到处理与集成的全过程，生成可直接供下游科学模型和工作流使用的数据集。更多详细说明见第4节。

在异构科学领域开展的大量实验表明，该框架具有良好的效果和泛化能力。在生命科学领域，SciDataCopilot能够根据一条自然语言指令，全自动准备大规模酶催化数据集，在数小时内生成214K条科学人工智能就绪的酶-反应记录，其规模几乎提升了 20×倍，超过此前收集的数据库。在神经科学领域，SciDataCopilot将脑电图/脑磁图分析标准化为四项子任务，包括Alpha提取、眼电回归、独立成分分析分解和大规模脑电图预处理。其定量表现可媲美人类专家，同时端到端执行速度提升3–5× 倍，并生成结构化执行轨迹。在地球科学领域，SciDataCopilot能够在严格的时间约束下执行跨学科气象数据准备，与基于电子表格的人工工作流相比，效率提升超过30× 。这些结果共同表明，SciDataCopilot是一个可扩展且不依赖特定领域的框架，能够将异构原始科学数据转化为科学人工智能就绪数据，弥合人类意图与自动化、质量感知型数据准备之间的鸿沟，适用于不同学科。我们正积极在更广泛的科学领域验证该框架。总体而言，我们的主要贡献概括如下：

- **任务条件化科学人工智能就绪数据范式。** 我们将科学人工智能就绪数据作为一种新型数据范式提出，它扩展了传统的人工智能就绪格式。该范式明确规定了任务条件化数据规范、基于科学依据的数据结构化以及跨领域数据组合，使复杂的科学数据能够与下游科学发现系统地对齐。

- **科学数据准备统一智能体框架。** 我们提出了*SciDataCopilot*，这是一个统一的智能体框架，通过数据访问、意图解析、数据处理和数据集成来分解科学数据准备流程。该框架支持跨异构数据源进行任务驱动的数据定制。

- **科学数据的跨领域泛化。** 我们在多个不同科学领域（包括生命科学、神经科学和地球科学）中验证了所提出的数据范式和框架。这些结果表明，科学人工智能就绪数据范式能够支持在数据模态、结构和实验约束截然不同的领域中进行一致的数据准备。

2. 相关工作

科学发现领域的数据基础设施长期以来围绕资源聚合、元数据注册、搜索与下载以及共享治理这一主线构建，代表性工作包括 Intern-Discovery [22], OpenDataLab [23], ScienceDataBank [24], 和 SciHorizon [25]。这些工作提供了大规模目录和基础元数据服务，用于回答诸如“数据在哪里”和“如何获取数据”等问题。然而，它们提供的成果往往是以文本为中心的输出（例如论文、报告或结构化文本摘要），或对传统多模态数据（音频、图像/视频和文本）进行轻度处理后得到的衍生数据，而非面向分析的科学数据集。因此，许多更广泛的科学数据模态往往缺失，包括时间序列测量数据、地理空间和遥感栅格数据，以及分子/结构表示。要解决具体的研究问题，仍需开展大量下游工作，包括清理、结构化、对齐、标注和准备AI 就绪输入。此外，通用和特定领域的数据准备方法仍存在适应性与可移植性有限、互操作成本高等问题，难以支持跨学科、多模态科学数据的可扩展处理与一致集成。

2.1. 通用自动化数据准备框架

通用领域的数据处理发展迅速，其中一个代表性方向是算子化、基于流水线的数据生产系统。DataFlow [26] 将大语言模型置于数据处理生态系统的核心，并提供可复用算子和通用流水线，用于清洗、筛选、格式化和合成高质量训练数据。Data- Juicer [27, 28, 29] , 、Curator [30], 和 DataTrove [31] 等系统也具备丰富的功能，支持数据剖析、去重、筛选、格式转换和多模态处理，并高度重视工程可扩展性和可扩展的执行后端。Deequ [32] 等开源数据准备框架提供数据质量检查、剖析和验证。Apache Spark [33]和 Apache Flink [34] 等云原生大数据计算基础设施提供可扩展的执行引擎，以支持大规模数据准备，适用于批处理和流处理场景；但在科学数据的语义理解、实验背景以及与领域工具链的集成方面，仍需开展更多工作。

这些系统显著降低了构建通用语料库和多模态训练数据的门槛，并推动了以数据配方和可复用流水线为基础的工业化方法。然而，将这些框架应用于科学数据时，往往会面临若干困难。科学数据通常以高度异质的实验数据包形式存储在文件系统中，而非扁平样本集。处理链依赖领域软件和模型，通用算子无法全面覆盖这些需求。科学数据准备不仅用于训练集构建，还用于提供符合研究意图定制的数据集、支持可复现实验流程，以及促进跨模态、跨学科的对齐与集成。

因此，除了通用数据准备范式之外，科学发现还需要结构化的元数据理解、任务驱动规划，以及与领域工具和模型建立有原则的关联。

2.2. 领域专用自动化数据准备框架

在特定领域的社群中，生物信息学（例如 AlphaFold 生态系统） [35], 地球科学（例如 eo4ai、earthlink） [36, 37], 和天体物理学等学科已建立起成熟的海量多模态数据处理流水线。NanoPyx [38],Astropy [39],和 MDAnalysis [40] 等领域工具包提供了成熟的解析、处理和分析功能，并形成了稳定的用户社群。近年来，更高层次的智能系统和行业级

数据处理平台也相继涌现。例如，EarthLink [37]基于 Intern-S1 基础模型构建，系统地整合了气候领域数据、知识、工具以及通用大模型能力。它通过人在回路的智能规划、自演化实验平台和多场景科学分析，降低了编程和软件使用门槛，并支持由自然语言指令驱动的研究工作流。在大规模生物医学队列场景中，与英国生物银行相关的内部数据处理工具已具备可扩展的样本与表型管理、基因组数据清洗与质量控制、特征构建及批处理能力 [41] [42] [43]。此外，许多学科仍高度依赖专用工具链。典型示例包括用于神经科学脑电图和脑磁图数据的 MNE-Python [44] 、用于地球与气候科学中网格数据和再分析数据的 xarray [45]和 NCO [46]、用于化学信息学中分子和反应数据的 RDKit [47] ，以及用于生物图像分析数据的 Fiji [48] 和 ImageJ [49] 。

尽管这些工具和平台相对成熟，但通常需要借助领域知识库访问和基于脚本的流水线来使用。它们普遍面临诸多局限，例如工具链复杂且紧密耦合、跨学科互操作性差，以及高度依赖专家经验。因此，实际研究工作流仍持续面临挑战。首先，缺乏对高度异构文件内容的机器可计算理解，难以将处理步骤规范化为可复现、可审计的标准化流程。其次，当研究目标、数据源或工具版本发生变化时，现有工作流难以快速迁移和复用，导致工作仍持续依赖专家进行手动调试和基于经验的决策。

2.3. 自动化数据准备的研究前沿

为弥合数据与知识之间的鸿沟，全栈智能体解决方案正逐渐成为一种发展趋势。例如，Intern-S1 [14] 和 SciReasoner [13] 通过多模态训练和改进推理能力，增强了基础科学能力。在此基础上，Intern-Agent [7] 构建了面向科学研究的闭环多智能体工作流，涵盖构思、实验和反思，从而实现自主计划，并对数据处理和实验流水线进行迭代优化。Kosmos [50]和 AI-Scientist [51]也探索了实现更自主、端到端科学发现的相关路径。与此同时，面向数据操作与分析的智能体范式正在快速发展。Data-Copilot [52] 将高层自然语言意图分解为可执行的工具链操作；ChatDB [53]则专注于数据库交互，并维护一个可操作的记忆流 ，以保留多轮一致性并支持自我修正，从而降低复杂查询和管理的门槛。此外，AutoGen [54] ，LangChain [55],和 LlamaIndex [56]等通用智能体框架，支持多智能体协作、工具调用和工作流编排，非常适合用于构建跨数据源分析、ETL及报告生成流水线。最后，SGI-Bench [9] 和 SciEvalKit [57] 等科学智能体评测套件提供了系统化基准测试，用于评估真实研究环境中的对齐性和可靠性。总而言之，从基础模型到智能体再到评测的这一分层技术栈，表明科学数据处理正从辅助工具转向日益自主的科学发现。然而，现有数据智能体仍主要面向通用工具使用，往往缺乏针对科学数据的系统化语义锚定、可复用且资产化的工作流组件，以及具备强可追溯性的版本化闭环流水线。在整合多模态和跨学科数据时，它们对约束驱动对齐的支持也有限。

为缓解这些问题，案例增强的工作流生成与智能体闭环编程提供了一条更具可迁移性的路径。通过检索先前成功的解决方案、将其适配到新需求，并纳入执行反馈以实现自动代码修复，这些方法

系统	多领域	多模态	Data 语义	Case 可复用	封闭 Loop	Data 融合
DataFlow [26]	✓	×	×	△	△	×
Data-Juicer [27, 28, 29]	✓	✓	×	×	×	×
Curator [30]	✓	✓	×	×	×	×
DataTrove [31]	✓	×	×	×	×	×
SciDataCopilot	✓	✓	✓	✓	✓	✓

表 1 | 代表性通用数据准备框架与我们的 SciDataCopilot 之间的能力对比。各列概括系统是否支持跨领域与多模态数据准备、科学数据语义（机器可计算元数据与本体对齐）、案例级工作流复用性、智能体闭环迭代，以及用于跨模态集成的约束驱动的数据融合。 ✓：支持，△：部分支持， ×：不支持。

降低幻觉并提升复杂任务的可运行性。在实践中，此类系统通常被实例化为由计划、验证、生成、执行、收集反馈和迭代构成的闭环管线。我们的工作与这一思路一致，但进一步将其系统化为面向科学数据的统一架构，如表1。

- **数据感知：**借助可计算的科学元数据与本体进行数据访问，使系统能够摄取异构数据集，并自动推导出机器可执行的解释和数据集描述。
- **案例复用：**案例库存储可复用的工作流资产。每个案例捕获数据单元、工具链和集成策略，支持检索与适配，以降低规划不确定性。
- **需求闭环：**一种多智能体闭环工作流，将用户需求转化为可执行管线和版本化制品，并具备溯源跟踪、故障可追溯性和迭代修复能力。
- **数据集成：**面向跨模态和跨学科对齐的约束驱动型集成，通过显式约束指导协调与融合，生成契合用户科学目标的高价值数据集。

总体而言，这项工作将算子与管线驱动的数据生产范式进一步拓展至科学领域。我们开发了一套系统化实现，将科学数据流层与多智能体编排以及基于本体与案例的知识库相结合，从而推动科学数据平台从单纯提供资源，迈向生产科学人工智能就绪数据集。

3. 重新思考面向科学发现的数据智能体

本节重新审视科学数据智能体，将其视为端到端工作流中连接科学问题、多样化数据资产、领域算子和下游科学模型的底层基础。我们认为，核心瓶颈不只是数据量或存储，而是缺乏一种系统化接口，无法让实验数据在不同模态和学科间具备科学有效性、符合约束且可组合进工作流。因此，本节系统阐述科学数据智能体的核心需求，分析现有数据准备系统的功能缺陷，并提出一个支持智能体科学 [58]的演进框架。

3.1. 概念构建

什么是科学数据智能体？科学数据智能体体现了人工智能系统设计的一种转变：从执行孤立预测的模型，转向能够基于科学数据和工作流进行自主推理与执行的系统。这类智能体通常构建于科学大语言模型之上 [59]，，旨在贯穿发现生命周期进行计划、执行和迭代，以实现高层次研究目标。与传统人工智能助手不同，科学智能体能够自主将复杂目标（例如识别候选药物）分解为结构化且可执行的子任务。它们收集并整合多模态线索，开展实验，并综合结果，形成连贯的科学发现。

关键在于，科学数据智能体处理的是*AI*就绪数据。在当前的 AI4S 系统中，这通常指经过清洗、标准化并格式化、可供 AI 模型（尤其是大语言模型）直接使用的数据。这类数据强调语法一致性和模式规范化，有助于大规模模型拟合和基于文献的推理。然而，AI就绪数据并未明确编码科学任务意图、实验背景或领域约束，因此不足以支持自主的端到端科学工作流。通过遵循体现科学方法的结构化假设驱动工作流，这些智能体纳入领域特定约束和验证协议，以确保自主发现具有可重复性并符合科学严谨性。

3.2. 设计启示

科学数据智能体有哪些不足？当前的科学数据智能体存在显著的操作和认知瓶颈，主要表现为过度依赖文本模态而忽视一手实证数据，以及训练语料库中缺少失败结果。这些系统在长期研究任务中持续表现出脆弱性，常因底层模型的静态特性和缺乏评估数据质量与实验背景所需的整合元数据而面临知识过时。因此，在编排复杂工具时，这些智能体往往难以取得较高的成功率；面对快速演进的研究发现前沿时，也难以保持科学严谨性。

需要怎样的架构？要实现稳健的智能体科学，必须建立一种以三阶段演进框架为核心的分层架构：构建多模态数据基础、促进科学知识涌现，并推动自主发现前沿的发展。该架构必须纳入操作系统级交互协议，例如模型上下文协议 [60]和科学上下文协议 [61]，，以统一异构环境（如计算模拟器和机器人实验室）中的工具发现、授权与执行。此外，该架构还需要自动化的持续集成流水线，以实时摄取数据，并建立不可篡改的可追溯链，从而确保整个自主研究生命周期的透明性、可审计性和可复现性。

综合来看，上述要求意味着科学数据准备需要采用分阶段架构，我们将在第4节中将其实现为SciDataCopilot。

4. SciDataCopilot 框架

SciDataCopilot 是一个统一的分阶段多智能体框架，通过将科学意图映射到涵盖数据探查、计划、转换和集成的可组合数据准备工作流，使科学 AI 就绪范式得以落地。给定一个自然语言科学查询q，该框架基于结构化知识库 $\{D,\ T,\ C\}$运行。其中，D 表示经过重新组织的科学数据集及其描述，T 表示可用的处理工具

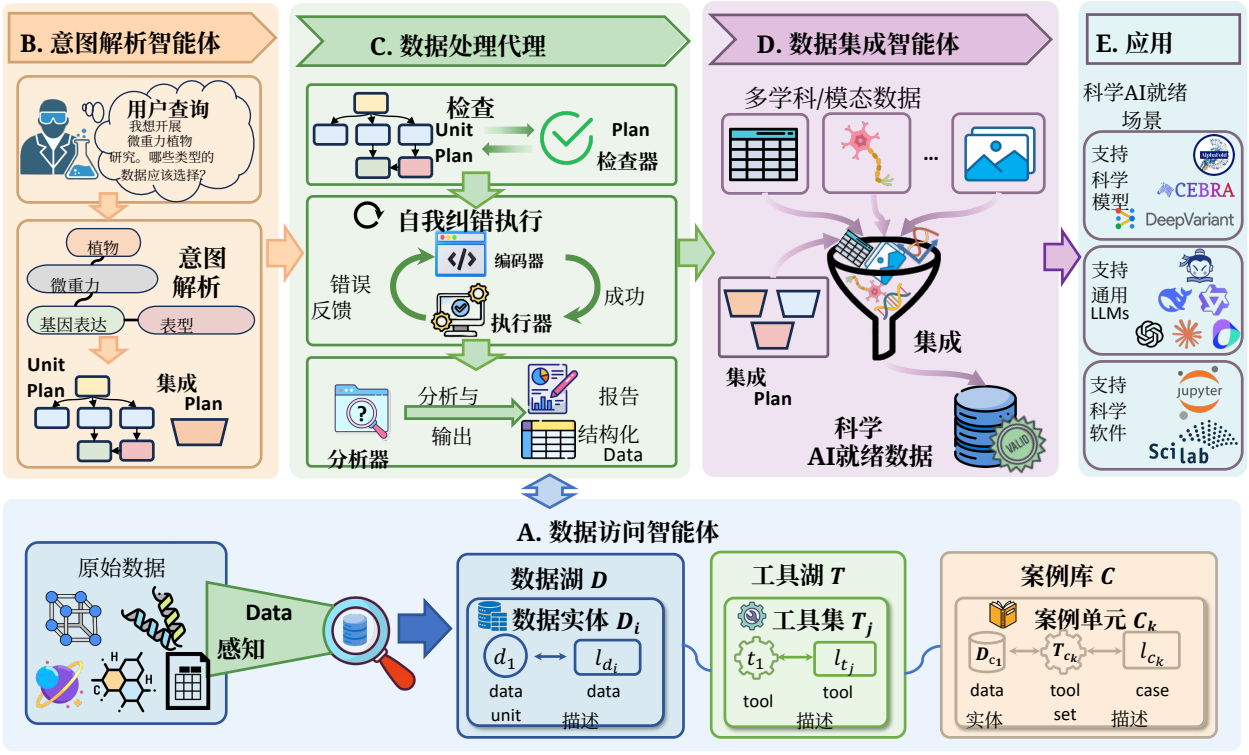


图 2 | **SciDataCopilot 的架构**。该框架集成了四个协作智能体（数据访问、意图解析、数据处理和数据集成），以自主地将用户意图与复杂的数据资源相匹配。最终，SciDataCopilot 将异构科学数据与特定模型连接起来，重新定义了任务引导的数据定制和跨学科集成，从而支持多样化的科学研究任务。

及其规范， C 则记录了关联科学问题、数据实体和处理流水线的历史案例，包括已准备的案例以及从外部数据库检索到的案例。

如图2所示，整体框架分为四个阶段。首先，SciDataCopilot 通过将异构原始数据、工具和历史先例组织到统一的知识库中，执行数据访问和知识构建。随后，在意图解析和计划制定阶段，系统会解析查询 q ，以筛选相关数据资产并制定特定任务的处理策略。在这些计划的指导下，SciDataCopilot 执行数据处理流程，通过领域感知操作和迭代调试，将选定的数据单元转换为与任务相匹配的输出。最后，根据明确的集成策略集成已处理的数据输出，生成连贯的科学人工智能就绪数据集，以满足原始科学意图。各智能体的详细设计和功能将在以下章节中介绍。

4.1. 数据访问智能体

科学数据很少以统一、标准化表格的形式呈现。实际中，数据集通常以多源、多模态、多版本制品的形式积累，分散存储在本地文件系统、共享存储或导出的实验包中。即使属于同一任务领域，不同团队或项目生成的数据也往往存在命名约定不一致、目录层级结构随意、字段定义不兼容，以及缺失值和异常值模式差异很大等问题，使得一次性编写的脚本十分脆弱，也阻碍了可扩展自动化。为此，我们将数据感知概念化为一种协作系统，而非简单的文件读取例程，用于路径探索、

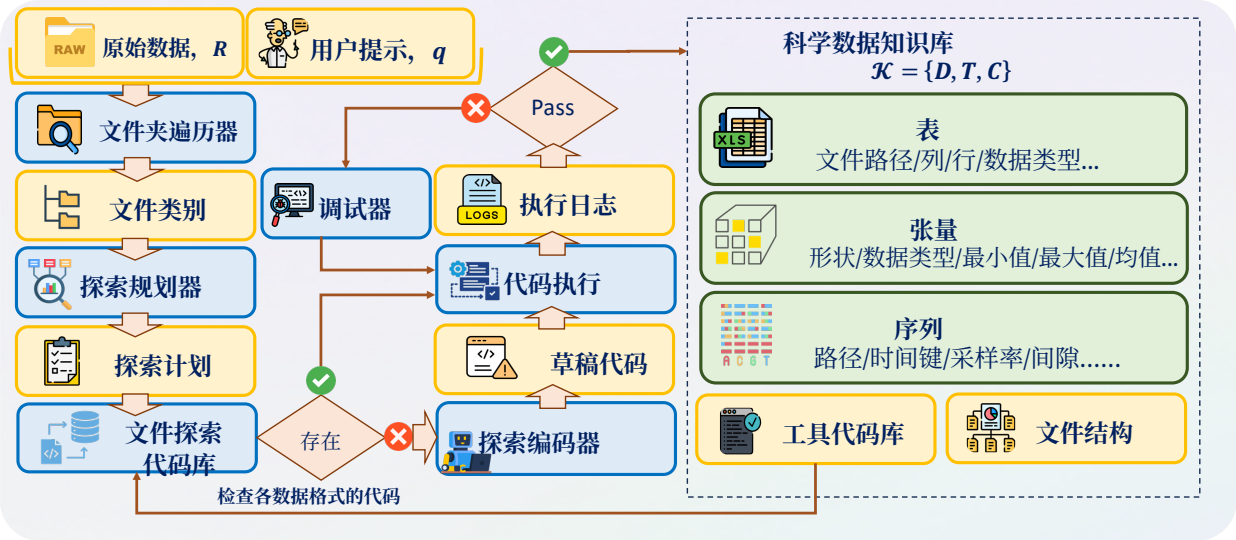


图 3 | **数据访问智能体**。输入：用户查询 q 和数据集根目录 R。输出：科学数据知识库 $\mathbf{K} = \{D, T, C\}$，即规范化数据单元和描述符，它们作为共享输入供下游意图解析、处理和集成使用。

模式推断和元数据规范化。基于这一观点，我们开发了一种**数据访问智能体**，能够自主探索数据目录，并以自组织的方式处理数据集，持续提取并规范化元数据，使其成为下游自动计划模块可使用的形式，从而初步构建科学数据本体。

4.1.1. 科学数据库构建

如图 3所示，数据访问智能体以用户查询 q 和数据集根目录 R为输入，并输出科学数据知识库

$$(q, R) \longrightarrow \mathcal{K} = \{D, T, C\}, \tag{1}$$

其中， D 表示数据湖， T 表示工具湖， C 表示案例库。生成的知识库 $\mathbf{K}$ 将由后续智能体（意图解析智能体、数据处理智能体和数据集成智能体）使用，作为用于计划和执行的统一数据集表示。

数据湖存储规范化的数据集实体及其描述：

$$D = \{[D_i, l_{D_i}]\}, \qquad D_i = \{[d_j, l_{d_j}]\}, \tag{2}$$

其中，每个数据集 D_i 都表示为一组数据单元 d_j，并配有单元级描述符l_{d_j}，而 l_{D_i} 则提供数据集级描述。数据单元被刻意限定为单一模态、单一时空范围和单一目标对象（或感兴趣实体）。这一限定使数据单元具备原子性和可组合性，并为索引、检索、验证和流水线合成等下游操作提供了稳定的抽象。描述符 l_{D_i} 和 l_{d_j}构成本体的规范化元数据层，以跨领域一致的格式记录模态标签、结构模式、可用时的时空索引、溯源、质量指标和汇总统计。

工具湖以同样的规范化方式对处理生态系统进行建模：

$$T = \{[T_i, l_{T_i}]\}, \qquad T_i = \{[t_j, l_{t_j}]\}, \tag{3}$$

其中，工具 t_j 被定义为一个功能完备模块，而 l_{t_j} 则说明其能力、输入–输出契约、依赖关系、运行时约束及适用模态。通过将工具作为具有明确描述符的一级实体进行存储，SciDataCopilot 可以根据任务需求和数据单元属性检索并组合工具，而不必依赖硬编码脚本或非正式约定。集合级描述符 l_{T_i} 进一步概括一个工具集，以支持规划期间更高层级的路由决策和兼容性检查。

案例库捕获了连接科学意图、数据与工具的可复用解决方案 ls:

$$C = \{C_i\}, \qquad C_i = \left\{l_{c_i}, \left[d_j, l_{d_j}, (t_{j_1}, t_{j_2}, \ldots, t_{j_k})\right], l_G, (t_{G_1}, t_{G_2}, \ldots, t_{G_m})\right\}, \tag{4}$$

其中，案例 C_i 包含案例描述 l_{c_i}，记录科学问题陈述及相应的处理策略。它还包含一组数据单元，其中每个单元都关联一个工具序列（t_{j_1}，……，t_{j_k}），构成单元级处理流水线。最后，案例定义了集成描述 l_G 和集成工具序列（t_{G_1}，……，t_{G_m}），说明如何将多个已处理单元整合为连贯输出。按照这一表述，案例库充当数据表示与可执行工具之间的显式接口：每个案例都将过程性知识编码为由带类型的数据单元和带类型的工具构成的组合图，支持跨任务的案例检索、流水线构建和可追溯复用。

数据库设计 K $=\{D, T, C\}$提供了一个统一且全面的信息平台，支持数据关联、集成与共享。通过以规范化描述符为所有实体奠定基础，并对数据集、原子数据单元、工具和可复用案例之间的关联进行显式建模，SciDataCopilot 建立了一个本体导向底层架构，使下游智能体规划与执行能够在异构科学领域中可靠运行。

4.1.2. 数据访问智能体方法

递归探索。 给定指定的数据集根目录后，数据访问智能体会递归扫描，生成完整的文件清单，以及记录数据集组织结构的相对路径索引。此步骤提供数据集的初步结构视图，包括目录模式、文件分组和基于扩展名的模态划分。

类型识别、脚本复用与执行–反思循环。 对于发现的每个资源，智能体首先执行文件类型识别，例如，张量数据格式包括 pkl、npy 和 pt，分类表格格式包括 csv 和 xlsx，序列文件包括 json 和 fasta。随后，系统会优先复用先前针对相同文件类型或相似结构模式验证过的探索脚本。如果复用失败或不存在可复用的逻辑，智能体便会调用大语言模型来生成新的探索脚本。至关重要的是，不能假定探索代码首次尝试就能运行。相反，智能体会执行脚本、捕获运行时错误，并进入自动的执行–错误–反思循环，在有限的迭代预算内不断运行，直至得到可运行的解析器和摘要生成器。这种闭环执行使数据集感知在实际操作中更加稳健，并减少对手动调试的依赖。

可复现的数据单元摘要和可追溯工件。 数据访问成功后，智能体会将结果保存到磁盘，包括数据单元标识符 d_j（例如文件路径、资产键、模态标签）及对应的描述 l_{d_j}。描述 l_{d_j} 包含重要的统计和结构摘要，例如张量形状、字段列表、列列表、数据类型、取值范围、缺失模式和代表性样本。此外，智能体还会保存用于生成摘要的脚本，以便复现和验证。

从探索输出到科学数据本体原型。 探索数据目录后，系统会将发现的内容转化为后续工作流阶段可直接使用的结构化信息。它从许多不同的文件和格式中收集元数据，然后进行清理和标准化

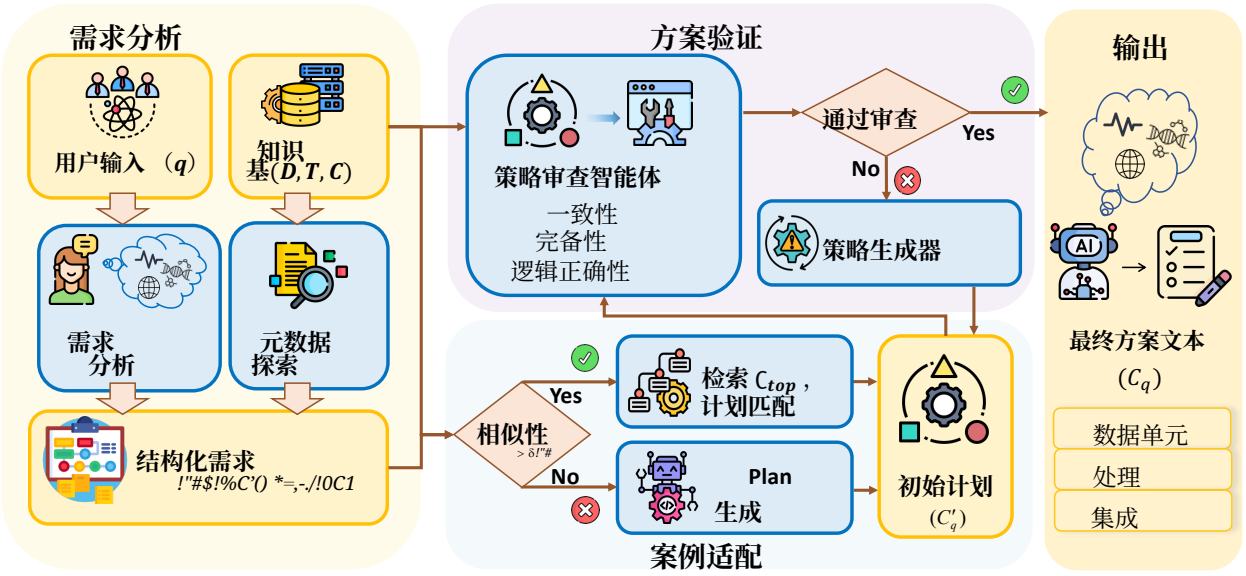


图 4 | **意图解析智能体**。借助知识库 K ={D, T, C}，该智能体通过需求分析、案例适配或生成以及迭代式策略审查来处理用户需求 q 并生成可执行计划 C_q。

并将其整合为单一的数据集级描述，记为 l_{D_i}。该描述概括了数据集的关键信息——其中包含哪些类型的数据、字段如何组织、基本统计信息如何、是否存在合理的时间或空间索引、数据可能如何划分为训练集和评估集，以及可从文件夹结构和文件名中推断出哪些语义线索。最终的总结存储在 l_{D_i} 中，作为下一步（例如意图解析和流水线规划）的机器可读上下文。

面向下游智能体的接口。 总而言之，数据访问智能体通过生成可复用、可检查的知识库 K，为整个系统提供基础支撑。在下游，意图解析智能体查询 K，以选择相关数据单元并生成意图级计划；处理智能体在所选数据单元上执行这些计划；集成智能体则将中间输出整合为最终的数据集级交付成果。

4.2. 意图解析智能体

基于面向本体的知识库 $\mathcal{K} = \{D, T, C\}$ 第 4.1 节中构建的知识库，本节的意图解析模块使用数据单元、工具和案例的规范化描述符作为语义锚点，将自然语言查询 q 转换为结构化规划集 C_q，其中包含可验证的数据单元选择和可执行的处理策略。整体工作流程如图 4 所示。具体而言，给定用户提供的需求描述或科学问题 q 以及知识库 K ={D, T, C}，目标是将模糊意图转换为结构化的数据单元及相应的处理策略。系统会自动生成一个包含 n数据单元的规划集 C_q，其形式化表达如下：

$$q \xrightarrow{\{D,T,C\}} C_q. \tag{5}$$

其中， d_{q_k} 表示与用户查询 q相关的第 k个数据单元； l_{q_k} 表示该数据单元的处理策略；而$l_{q,\ G}$ 表示用于整合多个数据单元的联合策略。最终输出是一个可预执行的文本计划，随后可由执行器将其落实为具体的脚本

和工作流。为了将模糊需求系统地转化为具体计划，我们的方法采用了一个认知框架，包含三个连续阶段：需求分析、案例检索与适配，以及策略生成与验证。

4.2.1. 科学数据需求分析

数据分析的主要挑战在于建立自然语言与知识库之间的语义映射。本模块通过智能体解析用户输入，提取核心研究变量和约束条件，并形成结构化需求表示 $R = \{Obj、Var、Con、Task\}$，分别对应研究对象、关键变量、约束条件（例如时间粒度和质量阈值）以及分析任务类型。在此过程中，系统会对 D 进行轻量级元数据扫描，以确保提取的变量在物理上可用或能够构建。例如，针对用户输入“探索微重力环境下植物的生长机制”，系统得出：

$$R = \begin{cases} Obj = \text{“plant”} \\ Var = \text{“growth performance, growth mechanism”} \\ Con = \text{“microgravity”} \\ Task = \text{“exploratory study, mechanism analysis”} \end{cases} \tag{6}$$

结构化需求表示 R 为后续在案例库中进行检索提供了语义锚点。

4.2.2. 案例检索与适配

为减少自动规划引发的幻觉，并提高规划质量和可复现性，我们引入案例库 C 作为先验解决方案。案例库由历史科学问题及其对应的数据处理方案构成，形式化表示为：

$$C = \{C_i\}, \quad C_i = \{l_{c_i}, [d_j, l_{d_j}, (t_{j_1}, t_{j_2}, \ldots, t_{j_k})], l_G, (t_{G_1}, t_{G_2}, \ldots, t_{G_k})\}. \tag{7}$$

在此形式中，C_i 表示存储库 C中的第 i个案例，l_{c_i} 表示案例 C_i的语义描述，涵盖自然语言描述的科学问题及其对应的解决方案。元组$[d_j, l_{d_j}, (t_{j_1}, t_{j_2}, \ldots\ldots, t_{j_k})]$记录了案例中使用的数据单元 d_j、应用于该数据单元的处理策略 l_{d_j}以及相关工具序列（t_{j_1}，t_{j_2}，……，t_{j_k}）。此外，l_G与（t_{G_1}，t_{G_2}，……，t_{G_k}）共同指定了跨单元集成策略及相应的集成流水线。

基于这一结构，我们提出一种混合推理机制，包含检索、适配和生成阶段：

- **案例检索：**我们计算用户需求 R 与案例库中语义描述 l_{c_i} 之间的相似性 $Sim(R, l_{c_i})$。如果某个高置信度案例 C_{top} 满足 $Sim > \delta$，则将其检索为“种子方案”。
- **计划适配：**如果 C_{top} 成功检索到，系统将执行差异分析。由于新的用户问题通常涉及变量或条件的细微变化（例如，将“小鼠”替换为“植物”，或将“区域环境”替换为“生长机制”），系统使用智能体将 C_{top} 中的原始研究对象映射到 R中的新对象，从而修订处理策略 l_{d_j} 和集成策略 l_G，以生成适配后的计划。
- **计划生成：**如果存储库中不存在相关案例（$Sim < \delta$），系统便会转而采用由多步推理驱动的生成模式。它根据 R 中的意图逻辑推断所需的数据单元，并通过智能体调用大模型，综合领域知识，进而推导出相应的处理和集成策略。

4.2.3. 科学数据处理计划的生成与验证

完成意图解析和案例适配后，系统进入计划合成的最后阶段。为确保生成的计划在逻辑上合理且在实际中可执行，我们建立了一个由生成、审查和修订步骤构成的闭环机制。该迭代过程由计划生成器和计划审查代理共同协调。

初始计划生成。 在此阶段，系统利用第 4.2.2 节中得到的适配案例，实例化一份针对用户具体查询定制的自然语言计划。该计划的形式定义如下：

$$C_q = \{(d_{q_1}, l_{q_1}), (d_{q_2}, l_{q_2}), \ldots, (d_{q_n}, l_{q_n}), l_{q,G}\}. \tag{8}$$

该计划可分解为三个内容层级： d_{q_i} 表示提取的数据单元，明确指定解决目标问题所需的 n 个独立数据单元； l_{q_i} 表示各单元的处理策略，用于定义要应用于相应数据单元的清洗和预处理逻辑；$l_{q,G}$ 表示集成策略，用于形式化数据单元之间的关系及相关集成逻辑。

计划审查与优化。 鉴于初始计划可能存在逻辑疏漏或资源不可用的情况，我们部署策略审查代理，利用结构化需求 R、数据湖 D和工具库 T 作为参考，对生成的计划 C_q进行多维度评估。此验证过程涵盖三个关键维度：首先是需求对齐，用于验证生成的计划是否符合用户的输入需求 q ，并检测是否存在偏题、概念替换或范围偏移（例如，若用户要求分析“植物生长机制 ”，但计划转而聚焦于“ 根系生长机制 ”或仅讨论单个子模块，则判定为不匹配）；其次是覆盖完整性，用于验证策略是否全面涵盖 R 中定义的关键变量和约束条件（例如，若用户要求提供“ 植物生长机制”相关数据，但计划仅提供“ 根生长机制 ”相关数据，则判定计划不完整）；第三是逻辑正确性，用于评估集成策略$l_{q,G}$ 的合理性，以避免结构性错误，例如错误合并时间粒度不匹配的数据。

迭代修订与计划定稿。 如果策略审查代理检测到潜在问题，就会生成自然语言反馈，明确具体的错误原因。随后，计划生成器利用大模型的上下文学习能力进行局部修订 C_q。这一过程可能迭代多轮，直至策略审查代理批准该计划或达到最大迭代次数限制。最终输出是一份经过严格验证且可执行的规划文本。

4.3. 数据处理代理

如图 5所示，数据处理代理由五个核心模块组成：(1) **计划细化**在数据与资源约束条件下，将意图解析阶段生成的初始计划修订为可行、可执行的流水线规范，重点关注要做什么，而非编写冗长的代码。(2)**计划检查** 对精炼后的流水线进行技术审查，以检查其正确性和可行性；审查通过则予以批准，否则提供具体的修订建议，并在必要时将工作流程退回细化阶段。(3) **代码合成** 基于已批准的精炼计划和当前错误上下文，生成完整、可运行的程序，并为每次修订保留版本，以便进行事后分析并确保可复现性。(4) **执行** 运行生成的程序，捕获运行时输出和错误跟踪，将日志写入实验工作区，并将失败信息反馈至共享状态以触发自动修复。(5) **分析** 在执行成功时生成科学解读报告；否则，在多次失败后生成可审计的终止报告，总结失败原因和迭代轨迹。



图 5 | **数据处理代理**。给定用户查询 q、选定的原始数据单元以及意图级处理计划，该代理会执行可复现的流水线，生成已处理数据、可视化结果和分析摘要，以供下游集成。

4.3.1. 处理流水线

在图 5中，给定由用户查询 q （以及可选的全局集成策略 $l_{q,\,G}$）得出的数据单元计划 $\{(d_{q_i},\ l_{q_i})\}_{i=1}^{n}$，数据处理代理会针对相应的数据单元 d_{q_i} 执行每项策略 l_{q_i}，并生成可供下游整合的中间单元级输出。

科学数据处理高度依赖领域知识和工具生态系统。不同学科在预处理惯例、统计假设、特征构建和评估指标方面存在显著差异。此外，真实世界数据常常存在缺失值、异常值、分布偏移和抽样偏差，使固定模板脚本随时间推移变得脆弱。因此，该系统强调两项原则，即可控领域推理和可执行闭环。可控领域推理原则是利用大模型推理进行流程编排，同时使过程可审计并受验证约束。相比之下，可执行闭环是将代码生成视为可执行、可验证且可自我修复的循环，而非一次性文本输出。

形式上，智能体以用户请求 q、经意图解析的处理计划，以及在数据访问阶段构建的共享知识库 $\mathrm{K} = \{D, T, C\}$ 作为输入：

$$\left[q,\ \mathcal{P} = \{(d_{q_i}, l_{q_i})\}_{i=1}^{n},\ \mathcal{K}\right] \longrightarrow [O,\ \mathcal{V},\ \mathcal{A}]\,. \tag{9}$$

此处，$\mathrm{O} = \{d_{o_i}\}_{i=1}^{n}$ 表示已处理数据单元，V 表示生成的可视化集合（用于诊断和报告的绘图/表），A 表示简明分析摘要（例如关键统计量、合理性检查结果和指标报告）。这些输出构成中间交付成果，将由下游数据集成代理根据 $l_{q,G}$ 进行组装。

依据上游规划阶段生成的数据处理规范，我们的系统旨在自动选择适当的数据处理工具、自动完善可复现处理流程、自动生成可执行的数据处理代码，并通过迭代调试和验证自动优化执行效率。此外，系统还考虑了科学数据处理的实际特点，包括单元级异质性、依赖感知调度，以及对相互独立的数据单元进行并行执行，以提升

在计算资源和输入/输出约束条件有限的情况下的吞吐量。

让意图解析智能体提取所需数据单元的集合

$$\mathcal{D}_s = \{d_{s_1}, d_{s_2}, \ldots, d_{s_n}\}, \tag{10}$$

并为每个单元制定处理策略

$$\mathcal{L}_s = \{l_{s_1}, l_{s_2}, \ldots, l_{s_n}\}. \tag{11}$$

意图解析智能体还会为每个单元提供初始处理计划，作为下游优化的起点，

$$\mathcal{P}_s^{(0)} = \left\{p_{s_1}^{(0)}, p_{s_2}^{(0)}, \ldots, p_{s_n}^{(0)}\right\}. \tag{12}$$

给定一个包含以往流水线、工具用法和最佳实践的案例库的领域知识库，数据处理代理会优化每个初始计划 $p^{(0)}_{s_i}$，将其转化为可执行的单元级流水线规范 p_{s_i}，并针对每种策略 l_{s_i} 编排一条工具序列作为流水线，其形式为

$$\mathcal{T}_{s_i} = \left(t_{s_i,j_1}, t_{s_i,j_2}, \ldots, t_{s_i,j_k}\right), \tag{13}$$

并执行这些流水线，将提取的数据单元转换为处理后的输出

$$\mathcal{D}_o = \{d_{o_1}, d_{o_2}, \ldots, d_{o_n}\}. \tag{14}$$

除 D_o 外，智能体还会生成可视化产物（例如诊断图和汇总表）以及简明的分析摘要，以支持质量控制和技术报告。

我们将科学数据处理智能体实现为一个显式的状态机工作流，其中每个步骤都有明确的职责，系统可以根据中间结果决定接下来运行哪个步骤。端到端工作流分为五个阶段。首先，智能体从知识库中检索相关的既有案例，并将意图解析阶段得到的初始计划作为输入。其次，智能体进行计划细化，通过确定工具选择、确保输入输出兼容性，并纳入数据量、内存预算和运行时间限制等可行性约束，将这些初始计划修改为可执行流水线规范。第三，在精炼计划获批后，智能体会生成一个完整、可运行的程序，用于实现所提出的流水线。第四，系统进入调试阶段，通过修复程序来应对警告和运行时错误。最后，系统执行程序、收集执行反馈，并持续迭代，直至运行成功或流程终止。在整个过程中，智能体维护一个共享的运行时状态，用于记录不断演进的计划、生成的代码、执行结果和中间产物。

针对每个候选处理计划，智能体会搜索既往解决方案存储库，寻找相似案例，并据此组合合适的工具与操作序列。在生成任何代码之前，都会先检查生成的流水线是否可行。如果计划不可行，智能体会请求基于模型的评议，并利用反馈精炼计划。计划被判定为有效后，代码生成组件会生成端到端实现，并将其保存为专用实验工作区中的带版本的产物，以便后续检查和复用。为体现实际数据处理的特点，智能体会识别独立单元，并在依赖关系允许时安排并行执行，同时通过并发限制和分阶段执行，将共享资源（如磁盘带宽和 GPU 显存）的使用控制在一定范围内。

4.3.2. 自修复执行循环

我们并不假设生成的程序首次运行就能正确执行。相反，程序执行与程序合成构成一个闭环反馈。执行器运行当前程序并捕获完整的运行时输出，包括标准输出和错误跟踪。这些信号会写回共享状态，随后用于指导修复步骤：代码生成组件会结合预期的精炼计划和最新的运行时失败，重写一个完整、可运行的程序。实际上，系统不再将代码视为静态文本来生成，而是把代码生成作为一个由真实执行反馈驱动的迭代工程过程，从而提升其在实际环境中的稳健性。

执行成功时，智能体会生成适用于技术报告的分析，其中包括对结果的科学解释、对计算指标的说明以及简洁的结论摘要。如果多次尝试仍未能收敛，系统不会悄无声息地失败，而是会生成可追溯的失败报告，记录运行终止的原因并保留迭代历史，使结果可供审计且便于采取后续行动，而非晦涩不明。

4.3.3. 可复现性与可追溯性

该系统旨在支持可复现的工程交付。每次运行都会创建一个带时间戳的实验工作区，其中保存精炼后的流水线计划、每个程序修订版本、执行日志和最终报告的快照。通过将数据集摄取输出（K 中的规范化数据集元数据）与处理输出（精炼计划、实现、可视化结果和报告）关联起来，智能体形成了一个数据驱动自动化闭环：数据集理解为计划制定提供依据，计划推动工具选择与实现，而执行则验证其正确性。此设计支持在异构科学数据集上实现可扩展自动化，同时保留可追溯性，以便调试和审查。

运行时由统一智能体状态管理，其中存储了协调工作流所需的所有信息。它包括用户的请求和任务类型、数据集标识符及规范化数据集描述、意图解析阶段生成的初始计划和当前精炼计划、程序实现、细化与执行过程中的迭代计数器和记录的错误，以及最终分析输出和成功标志。该状态还会跟踪当前实验工作区，以便将运行期间创建的所有工件关联到一个具体且可检查的目录。

4.4. 数据集成智能体

前一数据处理阶段会生成一组已处理数据单元 $\{d_{o_1}, d_{o_2}, \ldots, d_{o_n}\}$。所提出的**数据集成智能体** 旨在将这些分散且异构的数据单元整合为一个具有一致全局语义的数据对象 Duni:

$$\{d_{o_1}, d_{o_2}, \ldots, d_{o_n}\} \xrightarrow{\{(t_{o,G_1}, t_{o,G_2}, \ldots, t_{o,G_k})\}} D_{\text{uni}}. \tag{15}$$

在此，集成流程 $\{(t_{o,\,G_1},\ t_{o,\,G_2},\ \ldots\ldots,\ t_{o,\,G_k})\}$ 由集成策略 $l_{o,\,G}$在上游的意图解析与规划阶段生成。重要的是，$l_{o,\,G}$ 并非操作指令序列，而是一种联合的关系—结构约束规范（图6）。它编码了数据单元之间应如何关联、对齐和聚合，以及最终数据集所需的全局组织方式。每个集成算子 $t_{o,\,G_i}$ 均选自工具知识库（工具湖，$T = \{[T_i, l_{T_i}]\}$），其中 T_i 表示可用的集成工具，l_{T_i} 则规定其功能语义和约束条件。

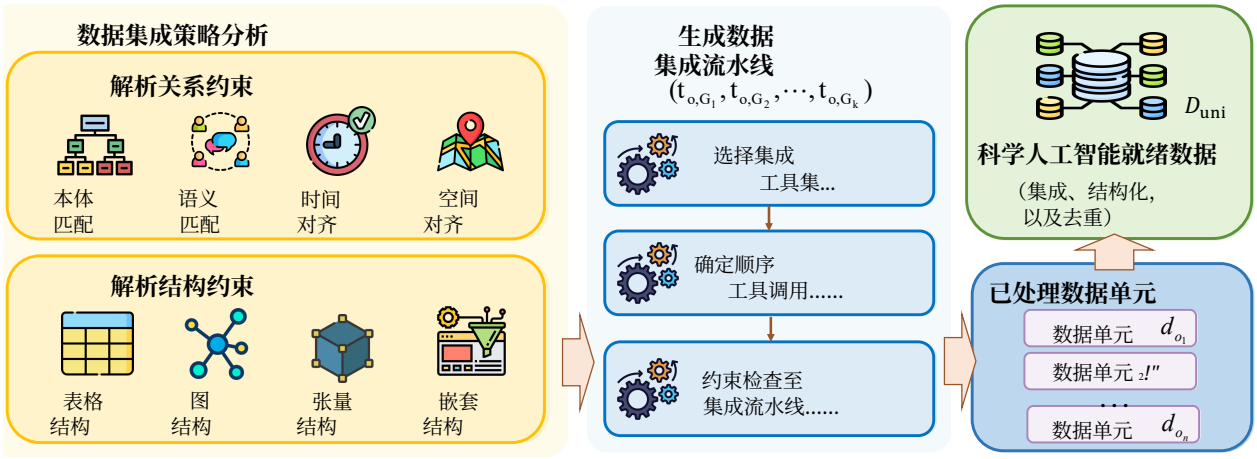


图 6 | **数据集成智能体**。该系统通过强制执行结构约束、选择经过验证的工具序列，并将分布式数据单元整合为一个连贯的科学人工智能就绪数据集，从而落实集成策略。

4.4.1. 数据集成策略分析

给定集成策略 $l_{o,G}$，此阶段的目标是将高层集成意图——通常以自然语言或抽象语义描述表达——转化为一组形式化、明确且可验证的集成约束条件：

$$l_{o,G} \longrightarrow \Gamma_G = \{\gamma_1, \gamma_2, \ldots, \gamma_m\}. \tag{16}$$

每项集成约束 γ_k 定义如下：

$$\gamma_k = (R_k, S_k), \tag{17}$$

其中，R_k 指定数据单元 $\{d_{o_1}, d_{o_2}, \ldots\ldots, d_{o_n}\}$ 之间必须满足的关系类型，而 S_k 指定集成输出所需的结构形式。

关系组件 R_k 可能包括但不限于本体对齐、语义对应、时间同步以及变量或模式映射。结构组件 S_k 定义集成数据集的预期表示形式，例如表格结构、基于图的结构或高阶张量表示。大语言模型根据原始用户查询 q 和全局科学意图推断 R_k 和 S_k。具体而言，大语言模型会确定哪些数据单元必须建立关联、它们之间关系的性质（例如语义、时间或变量层面的关系），以及最终数据集必须满足哪些结构要求，才能达到科学AI就绪。

4.4.2. 生成数据集成流水线

给定集成约束集合 $\Gamma_G = \{\gamma_1, \gamma_2, \ldots\ldots, \gamma_m\}$ 和工具湖 $T = \{[T_i, l_{T_i}]\}$，系统会生成一个集成工具序列 $\{(t_{o,G_1}, t_{o,G_2}, \ldots\ldots, t_{o,G_k})\}$。

每个选定的集成工具 t_{o,G_i} 必须共同满足 Γ_G 中定义的所有约束条件，包括所需的关系类型 R_k 和目标结构形式 S_k。生成集成流程需要大语言模型对以下方面进行推理：

1. **工具-约束匹配**：识别哪些工具能够实现特定关系

约束条件 R_k，以及哪些工具能够生成符合结构要求的输出S_k；2. **基于工具的序列推理**：确定有效的工具执行顺序，例如，在变量映射或聚合之前先进行时间对齐；3. **感知失败的回溯**：如果所选工具未能满足某项约束条件 $\gamma_k = (R_k,\ S_k)$，系统就必须通过重新选择备选工具或重构集成序列来修订其决策。

这一过程确保生成的集成流程在语义上合理，并且在指定约束条件下具备实际可行性。最后，按照指定顺序，依次在已处理数据单元$\{d_{o_1}, d_{o_2}, \ldots, d_{o_n}\}$上执行生成的有序集成工具序列 $\{(t_{o,\ G_1},\ t_{o,\ G_2},\ \ldots,\ t_{o,\ G_k})\}$，以生成最终的集成数据集：

$$D_{\text{uni}} = \{[d_i, l_{d_i}]\}\,. \tag{18}$$

此处，每个 d_i 表示集成数据集的一个组成部分，而 l_{d_i} 则记录其相关的语义注释和结构元数据。生成的 D_{uni}构成一个连贯、全局对齐且人工智能就绪的数据集，可直接满足查询中表达的原始科学意图 q。

5. 用例

为评估 SciDataCopilot 在真实科学工作流中的实际效果，我们在涵盖多个学科和数据模态的三个用例中对该框架进行了评估。每个案例均依据自然语言指令，在一致的实验设置下端到端执行，并将结果与专家的人工工作流进行比较。所有实验均在配备英特尔至强铂金 8558 处理器的服务器上进行，框架使用的大语言模型为 GPT-5。2 [62]评估重点关注效率、输出标准化，以及在真实数据和工具约束条件下的稳健性。

5.1. 案例 1：酶催化数据的自动化收集

5.1.1. 问题陈述

在许多生命科学领域，新兴研究问题缺乏标准化、可直接用于下游建模和分析的任务就绪数据集。尤其是酶催化研究，往往需要同时具备酶序列、催化反应以及化学结构明确的底物–产物表示，但这些内容很少以统一且机器可读的形式提供。

因此，数据准备通常依赖多阶段且耗费大量人工的工作流，包括文献检索、查询多个异构数据库，以及大量的数据清洗和规范化。这些碎片化流程限制了可扩展性和可复现性，也为数据驱动方法在酶催化研究中的快速应用造成了重大障碍。

本节将酶催化数据获取作为一个代表性用例进行考察，评估能否将高层次的自然语言规范可靠地转化为结构化、机器可读且可大规模构建的酶催化数据集。

5.1.2. 实验设置

在本案例研究中，我们将 SciDataCopilot 应用于大型蛋白质知识库，以评估其能否根据自然语言

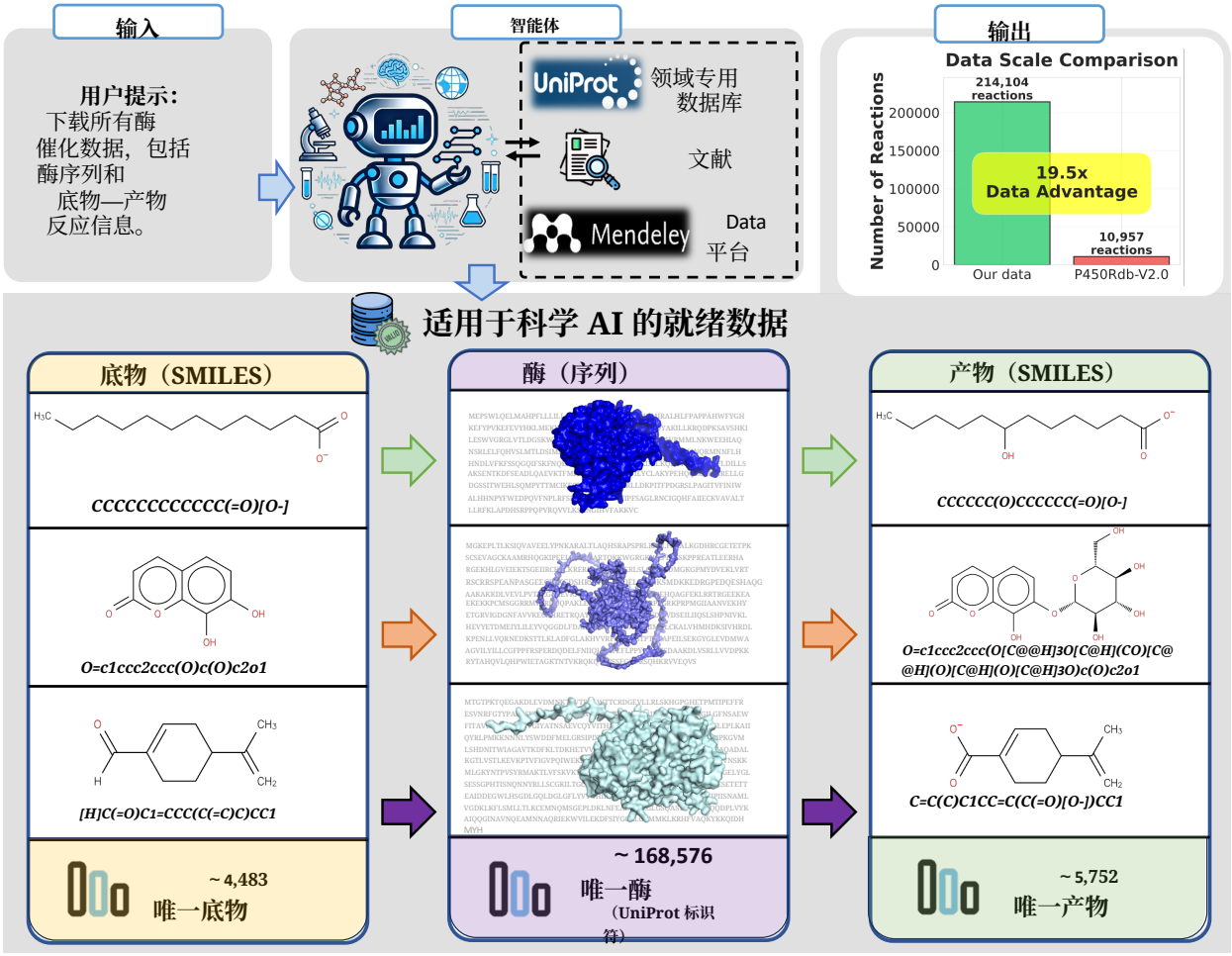


图 7 | 由 SciDataCopilot 准备的自动整理酶催化数据库案例研究示意图。

规范自动获取并预处理酶催化数据。实验重点在于评估能否将高层次科学意图转化为结构化、可用于机器学习的数据集，而无需用户手动与数据库模式或查询接口交互。

- **用户提示**：下载所有酶催化数据，包括酶序列和底物–产物反应信息。

我们将酶催化视为一项具有代表性的任务，因为它依赖于异构信息，包括蛋白质序列、功能注释以及化学明确的反应描述。实验输入是一条自然语言查询，要求获取全面的酶催化数据，包括酶序列及相应的底物–产物反应。在最少用户干预的情况下，该设置使我们能够评估 SciDataCopilot 的端到端数据获取能力。

其他查询示例，例如蛋白质家族筛选或特定生物数据收集，也由同一框架支持，但并非本次评估的主要重点。这些提示完全根据用户需求制定，以下实例对此进行了示例说明。

- 用户提示 1：收集人类中具有催化活性的蛋白质。

- 用户提示 2：搜索小鼠中的 RNA结合蛋白。
- 用户提示 3：提取蛋白激酶序列和底物信息。
- 用户提示 4：检索DNA结合蛋白的结构和功能数据。
- 用户提示 5：收集氧化还原酶及其催化反应。

5.1.3. 实验结果

我们评估了 SciDataCopilot 在大规模酶催化数据获取与整理任务中的表现。给定自然语言规范后，该系统会自动从蛋白质知识库中收集并预处理与酶相关的信息，生成可用于机器学习的酶—反应记录，其中明确表示底物和产物。

利用该流水线，我们构建了一个大规模酶催化数据集，其中包含214,104条结构化反应记录（图7）。每条记录均包括酶序列、规范化的催化反应描述，以及相应的底物—产物SMILES对。所得数据集支持多种下游应用，包括酶特异性预测、催化反应预测以及科学基础模型。

图 7 总结了整理后数据集的整体组成，展示了酶家族多样性、反应覆盖范围以及底物–产物统计信息。该数据集涵盖数千种不同的底物和产物，并涉及广泛的酶类别，体现出显著的化学与功能多样性。值得注意的是，SciDataCopilot 可在约五小时内自主完成整个数据收集与预处理流程，最终结果验证仅需极少量的人工检查。

我们将整理后的数据集与 P450Rdb-v2.0 [63]，（一个专注于细胞色素 P450 酶的专业数据库）进行比较。我们的数据集包含 214,104 个酶促反应，涉及 168,576 种唯一酶、4,483 种底物和 5,752 种产物。与包含 10,957 个经整理的 P450 特异性反应的 P450Rdb-v2.0 相比，我们的数据集总体反应规模约增加了 19.5× 。

这一规模优势体现了 SciDataCopilot 所实现的广泛酶类覆盖，其中包括甲基转移酶、脱氢酶、激酶、氧化还原酶和转移酶等多种酶家族；该数据集非常适合数据驱动建模，因为此类建模受益于规模庞大且异质性强的训练语料库。目前，由于现阶段仅依赖源自 UniProt 的数据来源，细胞色素 P450 反应的数量少于 P450Rdb-v2.0；不过，随着计划整合源自文献的数据和公共生物化学数据库，我们预计 P450 相关数据的规模和质量都将超越现有的领域专用资源。

总体而言，这些结果表明，SciDataCopilot 能够快速、大规模地构建酶催化数据集，为持续扩展生物化学覆盖范围、使其更加丰富全面奠定了可扩展的基础。

5.1.4. 总结

酶催化案例研究表明，所提出的数据整理范式本质上并不局限于单一数据库、模式或生物学领域。同一种交互模式——通过自然语言说明高层次科学意图，并获取结构化、可直接用于任务的数据——可以应用于其他涉及异构且标准化程度较低的科学资源的数据获取场景。

除了酶催化之外，这一范式还适用于一系列相关任务，例如收集蛋白质–配体相互作用数据、汇总不同数据库中的结构和功能注释，或整理生物体特异性分子数据集。更广泛地说，该案例研究表明，

自动化数据整理可以作为统一接口，连接蛋白质数据库、生化资源以及其他结构化科学资料库。

这些观察结果表明，SciDataCopilot 提供了一种可泛化的机制，能够以极少的人工工作实现可扩展的数据准备，并支持在不同科学领域中可复现的数据集构建。

5.2. 案例 2：端到端神经科学数据分析

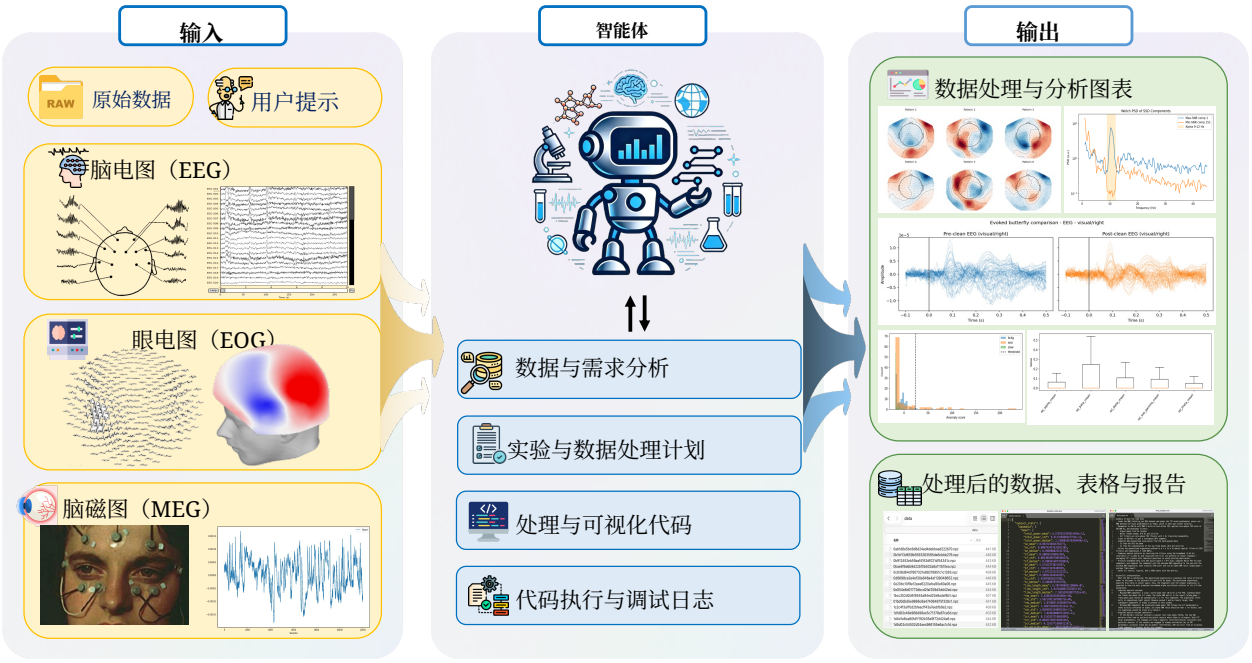


图 8 | SciDataCopilot 在脑电图（EEG）、眼电图（EOG）和脑磁图（MEG）中的案例研究示意图。

本用例聚焦于电生理记录的自动化分析，包括脑电图（EEG）、眼电图（EOG）和脑磁图（MEG）。在神经科学实验室中，这类数据价值很高，但众所周知，难以大规模处理。研究团队常常面临工具链零散、参数选择不一致以及可复现性差等问题。我们的目标是将这一工作流程转变为标准化、模块化且可复现的流水线，使其能够在不同数据集、研究团队和实验阶段中得到一致执行。

5.2.1. 问题陈述

脑电图和脑磁图分析是现代神经科学的一项基础技术，但现实中的实践仍面临若干长期存在的障碍。首先，数据具有高度异质性：不同的采集系统会产生不同的文件格式、元数据规范、通道定义和采样率，因此很难构建能够在不同数据集上可靠运行的统一流水线。其次，工作流往往分散在多个软件包和脚本中，依赖手工编写的衔接代码；这些代码脆弱、难以维护，并且数据集一旦变化就容易出错。第三，滤波器设置、伪迹剔除阈值和独立成分分析配置等关键预处理与分析选项，常常依赖经验进行调整，这会引入操作人员偏倚并降低可复现性。第四，验证往往不完整或不一致：研究人员可能会检查图表，但通常缺少一个能够系统检查统计质量的统一框架，

生理学合理性和可解释性。最后，交付成果通常散落在不同的文件夹和工具输出中，导致协作、审计和重新运行实验都耗费不必要的时间。为解决这些问题，我们旨在将依赖专家经验的“逐步式”电生理处理转变为标准化、模块化且可复现的工作流，将原始记录与经过验证、可审计的科学产出连接起来。

5.2.2. 实验设置

SciDataCopilot 将专家驱动的工作流转化为自动化端到端流水线，从用户的研究意图出发，最终生成一套结构化的科学产出。该框架能够理解分析请求并将其映射为一致的配置，从不同设备和文件格式中加载并规范化原始记录，执行标准化预处理，运行特定任务的分析模块，开展多视角验证，并将已处理数据、图表、指标和报告导出到可追溯目录结构中。我们使用常见的公共数据集和成熟的分析实践，在具有代表性的电生理场景中对其进行评估。该智能体处理来自多个采集生态系统的记录，涵盖广泛使用的电生理格式，例如CTF数据集、MNE格式文件以及大规模临床脑电片段集合。具体而言，我们测试神经科学工作流中常见的四项子任务：(i) 从脑磁图中提取α频段活动，并进行频谱和空间合理性检查（任务 1）；(ii) 对脑电图/脑磁图进行眼动伪迹校正，并开展前后验证（任务 2）；(iii) 执行独立成分分析分解，并检查数值有效性和可解释性（任务 3）；(iv) 对大规模临床脑电进行预处理和特征提取，并开展癫痫发作相关异常筛查（任务 4）。为兼顾实用性和科学有效性，我们从两个方面进行评估：**(i) 效率：**从提出分析请求到完成整套交付成果（已处理数据、图表、指标和报告）的端到端完成时间；**(ii) 有效性：**结合定性诊断证据评估特定任务的结果指标。对于任务 1，我们报告α频段与侧频带频谱信噪比，并检查空间模式和功率谱密度比较结果。对于任务 2–3，在可用时报告定量模型拟合和稳定性指标（例如解释方差、运行时间、收敛/异常统计），并提供可审计的诊断材料（前后对比图、空间权重图、成分地形图和日志）。对于任务 4，我们报告数据集级完整性指标（有效片段率、采样率和通道一致性）、预处理操作（伪迹剔除和独立成分分析剔除统计），以及通过混淆计数和分类异常率衡量的异常筛查表现，同时导出逐文件预处理日志以供审计。对于所有任务，智能体都会记录配置、中间产物、日志和最终报告，使整个运行过程能够在自包含工作空间中复现。为提供切实可行的基线，我们还招募了一名具有电生理处理经验的人工测试人员，使用标准工具链执行相同的四项任务，并交付可比的输出。我们从特定任务的定量指标和端到端完成时间两个方面，对SciDataCopilot与人工基线进行并列比较（表 2）。

5.2.3. 实验结果

在各项代表性任务中，该智能体均能产出完整的端到端交付成果，并在异构输入下表现可靠。

任务 1（脑磁图阿尔法波提取）。使用公开的 FieldTrip CMC 数据集（CTF MEG），智能体加载 SubjectCMC.ds，将数据裁剪至 $[50, 110]$ 秒，重采样至 250 赫兹，并处理 151 个 CTF 轴向梯度计。它采用 SSD，信号频带为 9–12 赫兹，邻近噪声频带为 8–9 赫兹和 12–13 赫兹（OAS 正则化协方差），同时结合实用的预处理流程（50/100 赫兹陷波滤波器和 1–45 赫兹零相位 FIR 带通滤波器），以稳定估计结果。提取的成分阿尔法波与邻频带功率比为 23.53（13.72 分贝），而信噪比最低的成分为 0.53（-2.78 分贝），表明最大值与最小值之比高达 44.6×。从定性结果来看，导出的功率谱密度

比较结果和地形图模式显示，所选成分的阿尔法峰更为尖锐；对于 CTF 三联传感器，则采用位置感知去重，并记录该处理以确保可解释性。作为比较，人工测试人员取得了相近的信噪比（最大值/最小值）：23.71（14.11 分贝）/ 0.52（-2.59 分贝），但端到端耗时更长，为 31.9 分钟（智能体耗时 9.2 分钟）。

任务 2（眼电回归）。在 MNE 示例数据集上，智能体识别刺激事件（共计 320 个；{visual/left:3, visual/right:4}），应用 0.3 赫兹高通滤波器，并通过短时滞设计（ ±100毫秒以内）与岭正则化对 EOG 061 进行回归，以去除眼动伪迹。定量而言，该回归解释了相当大比例的眼动驱动方差（均值解释方差 0.8638，中位数 1.0000）。定性而言，交付成果工作区包含前后诱发响应对比与回归权重地形图，可支持合理性检验（与额部／眶周传感器的耦合更强），并表明诱发响应形态得以保留、眨眼污染有所减轻。需要外部资源的可选步骤（例如麦克斯韦滤波）被视为尽力而为模块：缺失文件会被记录，同时流水线继续采用已验证的替代方案。相比之下，人工基线达到的均值／中位数解释方差为 0.8614/1.0000，而端到端时间增至 40.5 分钟（智能体为 11.3 分钟）。

任务 3（独立成分分析分解）。智能体对 0–60 s 时间段内仅 MEG 信号进行 ICA 拟合，采用 1–30 赫兹零相位 FIR 带通滤波器，将信号标准化为 20 个成分，并记录运行时间和稳定性指标。FastICA 经 62 次迭代后收敛（0.43 s），Infomax 和扩展 Infomax 均成功完成（分别为 0.69 s 和 0.75 s；两者均达到 500 次迭代）。由于缺少可选依赖项，Picard 被自动跳过；系统检测到此情况并记录日志，而非导致执行中断。系统导出了成分图和摘要，以便检查可解释性并查看类似伪迹的模式。在相同设置下，人工运行的 ICA 后端耗时更长（FastICA 为 0.7 s；Infomax 为 2.1 s；Ext-Infomax 为 3.2 s；Picard 为 0.9 s，迭代上限均相同），端到端完成时间为 35.4 分钟（智能体为 13.8 分钟）。

任务 4（TUSL 大规模脑电图预处理与异常筛查）。智能体在统一配置下处理了 300 个 TUSL脑电图片段（带通滤波 0.5–40 赫兹；目标采样率 250 赫兹），最终得到 296 个有效片段，另有 4 个因不符合伪迹标准而被拒绝。采样率完全一致（唯一值为 250 赫兹），通道数则略有差异（最小值 21，中位数 23，最大值 23）；逐文件预处理日志记录了丢弃的通道、移除的独立成分分析索引以及坏样本比例，确保可审计性。特征（标准频带上的功率谱密度和微分熵）采用 2.0 秒时间窗和 0.5 的重叠率计算，并使用全局异常阈值（分数 23.19）进行癫痫相关筛查。在这一工作点上，各类别的异常率分别为 5.05%（类别 0）、14.0%（类别 1）和 0%（类别 2）；混淆计数为 TP = 14、FP = 5、FN = 86 和 TN = 191。这表明其筛查表现较为保守（FP 较低，但 FN 较高）；智能体通过导出数据集级汇总和逐文件审计记录，明确呈现这一权衡，以便进行阈值调优和临床审查。相比之下，人工测试人员成功处理了 298/300 个文件（2 个被拒绝），得到的混淆计数为 TP = 20、FP = 11、FN = 62、TN = 203，但端到端耗时明显更长，为 51.9 分钟（智能体为 15.6 分钟）。

从专业评估的角度来看，这些结果表明，该智能体实施的是质量感知执行，而不仅仅是自动化脚本。它将定量检查（频谱信噪比、解释方差、运行时间/迭代、产出/一致性统计和混淆计数）与可解释诊断（地形图模式、功率谱密度、诱发前后图、ICA成分图和逐文件审计日志）相结合，降低了默默接受伪影或数值不稳定输出的风险。我们还记录了边界情况和局限性，以加强可审计性：当振荡结构较弱时，频谱比可能趋于平坦，提取模式也会变得分散，从而触发警告并保留中间产物；缺少可选资源（例如 Maxwell 滤波文件或可选 ICA 后端）时，则通过尽力执行并生成结构化日志来处理；对于大规模临床脑电图，通道变异性和类别不平衡可能导致特征分布发生偏移，因此有必要

表 2 | 四项电生理学任务的精简定量总结，比较 SciDataCopilot 与人类研究人员在核心指标、运行时间和可审计成果方面的表现。

Task	数据集 / 设置	SciDataCopilot	人类研究人员	审计工件
任务 1：Alpha 提取（SSD）	FieldTrip CMC（CTF 脑磁图）；250 赫兹；151 通道	**指标：** 信噪比（最大值/最小值）23.53（13.72 分贝）/ 0.53（-2.78 分贝） **时间：** 9.2 分钟	**指标：** 信噪比（最大值/最小值）23.71 (14.11 分贝) / 0.52 (-2.59 分贝) **时间：** 31.9 分钟	功率谱密度比较；地形图；报告
任务 2：眼电图回归-sion	MNE样本；0.3 赫兹高通；回归器 EOG 061	**指标：** 解释方差 均值 0.8638，中位数 1.0000 **时间：** 11.3 分钟	**指标：** 解释方差 均值 0.8614，中位数 1.0000 **时间：** 40.5 分钟	诱发前后；权重地形图；模型
任务 3：ICA 分解位置	MNE样本；仅MEG；0–60 秒；1–30 Hz频带通过；20 个成分	**运行时间（秒 / 迭代次数）：** FastICA 0.43/62；Picard 0.83/500 Infomax 0.69/500；Ext-Infomax 0.75/500 **时间：** 13.8 分钟	**运行时间（秒/迭代次数）：** FastICA 0.7/62；Picard 0.9/500 Infomax 2.1/500；Ext-Infomax 3.2/500 **时间：** 35.4 分钟	成分图；运行时日志；状态
任务 4：TUSL脑电图处理	300 个文件；0.5–40 赫兹带通滤波；采样率 250 赫兹；时间窗 2.0 秒（0.5 重叠）	**有效性：** 296 / 4（有效 / 已拒绝）；通道 21–23 **混淆：** TP/F-P/FN/TN = 14/5/86/191 **时间：** 15.6 分钟	**有效性：** 298 / 2（有效 / 已拒绝）；通道 21–23 **混淆：** TP/F-P/FN/TN = 20/11/62/203 **时间：** 51.9 分钟	清单；预-处理日志；特征

受试者/会话级校准或基于 ROC/PR 的阈值分析，智能体可通过导出受试者级特征统计数据和预测轨迹来支持这些分析。

5.2.4. 总结

除了这些已评估的设置外，同一智能体设计还支持广泛的电生理场景，且无需改变整体工作流结构。这得益于流水线的模块化设计：标准化数据接入、参数归一化、预处理、日志记录和交付均可复用，而分析和验证阶段则可根据研究问题进行替换或扩展。例如，通过调整事件对齐规则、分段时间窗和成分提取目标，同时沿用相同的数据协调和报告主干，系统即可适用于事件相关电位研究。通过启用对所选脑区划分进行连接性估计并计算网络指标，系统也可支持脑网络建模。通过添加基于振荡的分期启发式方法和周期检测，系统可扩展至睡眠分析；通过集成特征优化和分类器训练模块，系统还可支持脑机接口工作流。对于癫痫样活动筛查等临床脑电图用例，可启用带有领域特定阈值和报告模板的异常检测组件。重要的是，这些扩展无需从头重写流水线：智能体会复用相同的数据接入、预处理、审计追踪和交付成果打包机制，只需更改特定任务的分析逻辑和验证规则。图8 展示了这一端到端流程，总结了智能体如何将用户请求转换为归一化配置，探索并协调异构数据集，执行标准化预处理流程，路由至场景特定分析模块，应用多视角验证，并导出包含已处理数据和科学报告的完整交付包。



图 9 | SciDataCopilot 在海洋学、海洋化学和极地气象学中的案例研究示意图。

5.3. 案例3：跨学科地球数据准备

5.3.1. 问题陈述

长期地球科学分析日益依赖于将涵盖物理海洋学、海洋化学、沉积学、冰冻圈过程和大气环境的大规模异构观测数据集成起来，形成统一且可用于分析的数据集。然而，传统的数据整理仍受学科边界和格式差异的限制，往往需要多个部门和领域专家开展跨组协作。在实际工作中，仅对单个年份的数据进行清洗和标准化，就可能需要数周持续的人工劳动。

历史地球档案还具有多个维度系统性错位的特征，包括数据格式、时间分辨率和语义定义。这些不一致性使异构观测源无法进行联合分析，也无法直接供数据驱动和物理信息人工智能模型使用。

一个核心障碍是，许多遗留观测数据并未以便于下游集成的方式组织。观测值通常存储在结构较弱的文件中，而语义表头和时间字段则被分别编码，或以不一致的惯例记录。因此，构建列含义正确且时间索引标准化的连贯表格，往往需要反复人工检查和非正式转换。

从以数据为中心的人工智能视角来看，这一对齐步骤是必要的前提：必须先将地球观测数据转换为语义连贯且时间标准统一的表示形式，才能进行跨来源和跨模态集成。若不进行这种转换，就难以构建能够支持可复现分析和建模的可靠科学人工智能就绪数据集。

5.3.2. 实验设置

在本案例研究中，我们评估了一项具有代表性的对齐任务：观测值和语义表头分别存储在不同的旧版文件中，且时间组织方式与下游分析需求不符。具体而言，本次评估使用了2005—2006年在Marble Point采集的气象观测数据，共5840行[64]，，用户指令为“处理极地表格数据：合并表头和记录，根据小时值计算日平均值，然后按月拆分输出”。

选择这项任务，是因为它能够单独检验一项基础能力，而这项能力是实现更广泛地球数据集成所必需的。它评估能否将一条自然语言规范转化为确定、可审计的处理流水线，以重建列语义、规范化时间字段、执行时间聚合，并导出标准化输出，作为多源和多模态对齐的上游输入。

5.3.3. 实验结果

我们在 Marble Point 数据集上开展了对比研究，以评估 SciDataCopilot 的效率和有效性。该任务涉及处理 5840 行气象数据，并严格遵循时间划分和平均值计算要求。智能体成功理解了自然语言指令，并仅用 3.5 分钟就执行完完整的数据处理流水线。相比之下，一名熟练的人工操作员使用带有自动补全功能的 Excel 完成相同任务，需要 75 分钟。这些结果表明，与传统人工工作流程相比，智能体的效率提升超过 $20\times$ 。重要的是，智能体不仅加快了处理速度，还确保生成高质量的标准化输出，可立即用于下游分析，从而有效消除了人工数据处理所带来的差异和潜在错误。

5.3.4. 总结

所展示的工作流遵循 SciDataCopilot 的智能体数据准备设计，通过协调智能体角色将零散的原始文件转换为科学人工智能就绪数据。意图解析组件将用户请求转化为包含明确操作和约束条件的可执行计划；数据访问组件则识别相关文件，提取观测矩阵和元数据工件，并将它们规范化为规范化内部表示。随后，数据处理组件在迭代检查与执行过程中进行语义重建和时间重组，确保中间产物可追溯，并通过结构化诊断而非静默损坏来处理故障。

尽管评估使用的是气象表格，但相同的交互模式也适用于模态和存储惯例各异的异构地球观测数据。相同的流水线级结构可摄取并规范化航次测量数据、系泊仪器记录和沉积物采样输出，同时复用一致的模式重建、时间标准化、日志记录和交付成果打包机制。从这个意义上说，该案例应被视为跨学科集成的前置模块，可支持后续步骤，例如跨源变量匹配，以及海洋、冰、大气和生物学领域之间的多模态融合。

6. 结论

总而言之，我们介绍了 **SciDataCopilot**，这是一个统一的智能体框架，旨在弥合异构实验数据与人工智能驱动的科学发现所需严格要求之间的差距。通过将数据准备范式从人工智能就绪提升至科学AI就绪，该系统明确

应对任务依赖、数据碎片化和领域特定处理约束等挑战。该框架协调四个协作智能体，自动完成数据准备的整个生命周期，将原始、杂乱的数据转化为科学AI就绪资源。

在多个不同科学领域的大规模部署证实了系统的有效性和泛化能力。在生命科学、神经科学和地球科学的三个用例中，SciDataCopilot 通过在真实的科学约束条件下，可靠地将异构原始数据转化为标准化、任务对齐的资源，验证了科学人工智能就绪数据范式。

展望未来，仍有若干方向值得探索。首先，我们计划将 SciDataCopilot 扩展到更广泛的科学领域和数据模态，包括大规模仿真输出和实时实验流，以进一步检验科学AI就绪范式的普适性。其次，尽管当前系统依赖经过筛选的领域工具和案例，未来工作将探索更具适应性的工具发现、验证机制，以及从执行轨迹中学习的机制，从而在尽量减少人工干预的情况下持续改进。最后，我们旨在制定针对科学AI就绪数据的评估协议，评估维度不止效率和正确性，还将纳入科学有效性、可复现性，以及对科学建模和科学发现的下游影响。

参考文献

[1] Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、Ahmed El-Kishky 等人。《OpenAI o1 系统卡》。发表于：2024年。arXiv：2412.16720 [cs.AI]。网址：https://arxiv.org/abs/ 2412.16720。[2] Marah Abdin、Jyoti Aneja、Harkirat Behl、Sébastien Bubeck、Ronen Eldan 等人。《Phi-4 技术报告》。发表于：2024年。arXiv：2412.08905 [cs.CL]。网址：https://arxiv.org/abs/2412. 08905。[3] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Peiyi Wang 等人。《DeepSeek-R1 通过强化学习激励大语言模型中的推理》。发表于：第645卷，第8081期。2025年，第633–638页。网址：http://dx.doi.org/10.1038/s41586-025-09422-z。[4] Google DeepMind。《Gemini 3 模型说明卡》。发表于：2025年。网址：https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf。[5] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui 等人。《Qwen3 技术报告》。发表于：2025年。arXiv：2505.09388 [cs.CL]。网址：https://arxiv.org/abs/2505.09388。[6] Zonglin Yang、Wanhao Liu、Ben Gao、Yujie Liu、Wei Li 等人。《MOOSE-Chem2：通过分层搜索探索大语言模型在细粒度科学假设发现方面的局限》。发表于：2025年。arXiv：2505.19209 [cs.AI]。网址：https://arxiv.org/abs/2512.16969。[7] InternAgent 团队、Bo Zhang、Shiyang Feng、Xiangchao Yan、Jiakang Yuan 等人。《InternAgent：当智能体成为科学家——构建从假设到验证的闭环系统》。发表于：2025年。arXiv：2505.16938 [cs.AI]。网址：https://arxiv.org/abs/2505.16938。[8] Mingxuan Du、Benfeng Xu、Chiwei Zhu、Xiaorui Wang 和 Zhendong Mao。《DeepResearch Bench：深度研究智能体综合基准测试》。发表于：2025年。arXiv：2506.11763[cs.CL]。网址：https://arxiv.org/abs/2506.11763。[9] Wanghan Xu、Yuhao Zhou、Yifan Zhou、Qinglong Cao、Shuo Li 等人。《通过科学家对齐工作流探测大语言模型的科学通用智能》。发表于：2025年。arXiv：2512.16969[cs.AI]。网址：https://arxiv.org/abs/2512.16969。[10] Joseph L Watson、David Juergens、Nathaniel R Bennett、Brian L Trippe、Jason Yim 等人。《使用 RFdiffusion 从头设计蛋白质结构与功能》。发表于：第620卷，第7976期。2023年，第1089–1100页。网址：https://doi.org/10.1038/s41586-023-06415-8。[11] J. Dauparas、I. Anishchenko、N. Bennett、H. Bai、R. J. Ragotte 等人。《使用 ProteinMPNN 进行基于深度学习的稳健蛋白质序列设计》。发表于：第378卷，第6615期。2022年，第49–56页。网址：https://www.science.org/doi/abs/10.1126/science.add2187。[12] Yuhao Zhou、Yiheng Wang、Xuming He、Ao Shen、Ruoyao Xiao 等人。《科学家的首次考试：通过感知、理解和推理探测多模态大语言模型的认知能力》。发表于：2025年。arXiv：2506.10521 [cs.AI]。网址：https://arxiv.org/abs/2506.10521。[13] Yizhou Wang、Chen Tang、Han Deng、Jiabei Xiao、Jiaqi Liu 等人。《SciReasoner：奠定跨学科科学推理的基础》。发表于：2025年。arXiv：2509.21320 [cs.CL]。网址：https://arxiv.org/abs/2509.21320。[14] Lei Bai、Zhongrui Cai、Yuhang Cao、Maosong Cao、Weihan Cao 等人。《Intern-S1：科学多模态基础模型》。发表于：2025年。arXiv：2508.15763 [cs.LG]。网址：https:// arxiv.org/abs/2508.15763。[15] ScienceOne 团队。《S1-DeepResearch：高性能深度研究智能体》。2025年。网址：https://github.com/ScienceOne-AI/S1-DeepResearch。

[16] Juraj Gottweis、Wei-Hung Weng、Alexander Daryin、Tao Tu、Anil Palepu 等。“迈向 AI co-scientist”。发表于：2025年。arXiv：2502.18864 [cs.AI]。网址：https://arxiv.org/abs/ 2502.18864。[17] The Swarm Corporation。“AI-CoScientist：用于协作科学研究的多智能体框架”。2024年。网址：https : / / github . com / The- Swarm-Corporation / AI - CoScientist。[18] Alireza Ghafarollahi 和 Markus J Buehler。“SciAgents：通过仿生多智能体智能图推理自动化科学发现”。发表于：第 37 卷第 22 期。2025年，第 2413523 页。网址：https : / / advanced . onlinelibrary . wiley . com / doi / abs / 10 . 1002 / adma . 202413523。[19] Mourad Gridach、Jay Nanavati、Khaldoun Zine El Abidine、Lenon Mendes 和 Christina Mack。“用于科学发现的智能体人工智能：进展、挑战与未来方向综述”。发表于：2025年。arXiv：2503.08979 [cs.CL]。网址：https://arxiv.org/abs/2503.08979。[20] Bowen Zhou、Ning Ding、Lei Bai 和 Hao Zhou。“推进科学人工智能：从工具革命到革命性工具”。发表于：第 6 卷。2025年，第 323–328 页。网址：https://www.sciencedirect.com/science/inproceedings/pii/S266665102500021X。[21] John Jumper、Richard Evans、Alexander Pritzel、Tim Green、Michael Figurnov 等。“利用 AlphaFold 实现高精度蛋白质结构预测”。发表于：第 596 卷第 7873 期。2021年，第 583–589 页。网址：https://doi.org/10.1038/s41586-021-03819-2。[22] 上海人工智能实验室。“Intern-Discovery”。发表于：2025年。网址：https:// discovery.intern-ai.org.cn/home。[23] Conghui He、Wei Li、Zhenjiang Jin、Chao Xu、Bin Wang 等。“Opendatalab：利用开放数据集赋能通用人工智能”。发表于：2024年。arXiv：2407.13773 [cs.DL]。网址：https://arxiv.org/abs/2407.13773。[24] Chengzan Li、Yanfei Hou、Jianhui Li 和 Lili Zhang。“ScienceDB：面向电子科学的公共多学科研究数据存储库”。发表于：IEEE电子科学国际会议。2017年，第 248–255 页。网址：https://doi.ieeecomputersociety.org/10.1109/eScience. 2017.38。[25] Chuan Qin、Xin Chen、Chengrui Wang、Pengmin Wu、Xi Chen 等。“SciHorizon：评测从科学数据到大语言模型的科学人工智能就绪度”。发表于：ACM SIGKDD知识发现与数据挖掘会议论文集。2025年，第 5754–5765 页。网址：https://doi.org/10.1145/3711896.3737403。[26] Hao Liang、Xiaochen Ma、Zhou Liu、Zhen Hao Wong、Zhengyang Zhao 等。“DataFlow：以数据为中心的人工智能时代中用于统一数据准备和工作流自动化的大语言模型驱动框架”。发表于：2025年。arXiv：2512.16676 [cs.LG]。网址：https://arxiv.org/ abs/2512.16676。[27] Daoyuan Chen、Yilun Huang、Zhijian Ma、Hesen Chen、Xuchen Pan 等。“Data-Juicer：面向大语言模型的一站式数据处理系统”。发表于：国际数据管理会议。2024年，第 120–134 页。网址：https://doi.org/10.1145/3626246. 3653385。[28] Daoyuan Chen、Haibin Wang、Yilun Huang、Ce Ge、Yaliang Li 等。“Data-Juicer Sandbox：用于多模态数据-模型协同开发的反馈驱动套件”。发表于：国际机器学习会议。2025年。网址：https : / / openreview . net / forum ? id =zIGIvysR1H。

[29] Daoyuan Chen、Yilun Huang、Xuchen Pan、Jiang Nana、Haibin Wang 等。“Data-Juicer 2.0：面向基础模型并与基础模型协同的云规模自适应数据处理”。发表于：神经信息处理系统数据集与基准赛道年会。2025年。网址：https: //openreview.net/forum?id=DHA9uoeMQx。[30] NVIDIA NeMo。“NVIDIA-NeMo/Curator：面向大语言模型的可扩展数据预处理与整理工具包”。发表于：2025年。网址：https://github.com/NVIDIA-NeMo/Curator。[31] Guilherme Penedo、Hynek Kydlíček、Alessandro Cappelli、Mario Sasko 和 Thomas Wolf。“DataTrove：大规模数据处理”。发表于：2024年。网址：https://github.com/huggingface/ datatrove。[32] Sebastian Schelter、Dustin Lange、Philipp Schmidt、Meltem Celikel、Felix Biessmann 等。“自动化大规模数据质量验证”。发表于：第 11 卷第 12 期。2018年，第 1781–1794 页。网址：https://doi.org/10.14778/3229863.3229867。[33] Apache Spark。“Apache Spark”。2018年。网址：https://spark.apache.org。[34] Paris Carbone、Asterios Katsifodimos、Stephan Ewen、Volker Markl、Seif Haridi 等。“Apache Flink：在单一引擎中实现流处理与批处理”。发表于：第 38 卷第 4 期。2015年。网址：https: // api.semanticscholar.org/CorpusID:263802939。[35] Richard Evans、Michael O’Neill、Alexander Pritzel、Natasha Antropova、Andrew Senior 等。“利用 AlphaFold-Multimer 预测蛋白质复合物”。发表于：2021年。网址：https://www. biorxiv.org/content/early/2021/10/04/2021.10.04.463034。[36] ESAPhiLab 团队。“Eo4ai：面向人工智能的地球观测”。2022年。网址：https://github.com/ESA- PhiLab/eo4ai。[37] Z. Guo、J. Wang、X. Yue 等。“EarthLink：面向气候科学的自演化人工智能体”。发表于：2025年。arXiv：2507.17311 [cs.LG]。网址：https://arxiv.org/abs/2507.17311。[38] Bruno M Saraiva、Inês Cunha、António D Brito、Gautier Follain、Raquel Portela 等。“利用 Liquid Engine 驱动的 Python 框架 NanoPyx 高效加速生物图像分析”。发表于：第 22 卷第 2 期。2025年，第 283–286 页。网址：https://doi.org/10.1038/s41592-024-02562- 6。[39] Thomas P Robitaille、Erik J Tollerud、Perry Greenfield、Michael Droettboom、Erik Bray 等。“Astropy：面向天文学的社区 Python 软件包”。发表于：第 558 卷。2013年，A33。网址：https://ui.adsabs.harvard.edu/abs/2013A&A...558A..33A。[40] Naveen Michaud-Agrawal、Elizabeth J Denning、Thomas B Woolf 和 Oliver Beckstein。“MD-Analysis：分子动力学模拟分析工具包”。发表于：第 32 卷第 10 期。2011年，第 2319–2327 页。网址：https://doi.org/10.1002/jcc.21787。[41] Ken Hanscombe。“Ukbtools：用于导入和处理英国生物银行数据的工具”。2017年。网址：https://github.com/kenhanscombe/ukbtools。[42] J. Mbatchou、L. Barnard、J. Backman 等。“用于定量性状和二元性状的高效全基因组回归计算方法”。发表于：第 53 卷。2021年，第 1097–1103 页。网址：https://doi. org/10.1038/s41588-021-00870-7。[43] Wenjia Bai。“ukbb_cardiac”。发表于：2025年。网址：https://github.com/baiwenjia/ukbb_cardiac。[44] Alexandre Gramfort、Martin Luessi、Eric Larson、Denis A. Engemann、Daniel Strohmeier 等。“使用 MNE-Python 分析脑磁图和脑电图数据”。发表于：第 7 卷。2013年，第 267 页。网址：https: // www.frontiersin.org/journals/neuroscience/inproceedingss/10.3389/ fnins.2013.00267。

[45] Stephan Hoyer 和 Joe Hamman。《xarray：Python 中的 N维标记数组和数据集》。发表于第 5 卷，第 1 期。2017年，第 10 页。网址：https://openresearchsoftware.metajnl.com/inproceedingss/jors.148。[46] Pedro Vicente、Charlie Zender、Henry Butowsky。《NCO：NetCDF 运算符》。2025年。网址：https://nco.sourceforge.net/。[47] Greg Landrum。《Rdkit 文档》。发表于第 1 卷。第 1–79 页。2013年，第 4 页。网址：https://media.readthedocs.org/pdf/rdkit/latest/rdkit.pdf。[48] Johannes Schindelin、Ignacio Arganda-Carreras、Erwin Frise、Verena Kaynig、Mark Longair 等。《Fiji：用于生物图像分析的开源平台》。发表于第 9 卷，第 7 期。2012年，第 676–682 页。网址：https://doi.org/10.1038/nmeth.2019。[49] Tony J. Collins。《用于显微镜学的 ImageJ》。发表于第 43 卷，增刊 1。2007年，第 S25–S30 页。网址：https://doi.org/10.2144/000112517。[50] Ludovico Mitchener、Angela Yiu、Benjamin Chang、Mathieu Bourdenx、Tyler Nadolski 等。《Kosmos：用于自主发现的人工智能科学家》。发表于 2025年。arXiv：2511.02824 [cs.AI]。网址：https://arxiv.org/abs/2511.02824。[51] Chris Lu、Cong Lu、Robert Tjarko Lange、Jakob Foerster、Jeff Clune 等。《The AI Scientist：迈向全自动开放式科学发现》。发表于 2024年。arXiv：2408.06292[cs.AI]。网址：https://arxiv.org/abs/2408.06292。[52] Wenqi Zhang、Yongliang Shen、Weiming Lu 和 Yueting Zhuang。《Data-Copilot：通过自主工作流连接数十亿数据与人类》。发表于 2023年。arXiv：2306.07209[cs.CL]。网址：https://arxiv.org/abs/2306.07209。[53] Chenxu Hu、Jie Fu、Chenzhuang Du、Simian Luo、Junbo Zhao 等。《ChatDB：以数据库作为符号记忆增强大语言模型》。发表于 2023年。arXiv：2306.03901 [cs.AI]。网址：https://arxiv.org/abs/2306.03901。[54] Qingyun Wu、Gagan Bansal、Jieyu Zhang、Yiran Wu、Beibin Li 等。《AutoGen：通过多智能体对话实现新一代大语言模型应用》。发表于 COLM 2024。2024年。网址：
h
t
t
p
s://
w
w
w.
m
i
c
r
o
s
o
ft.
com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/。[55] Vasilios Mavroudis。《LangChain》。2024年。网址：https://github.com/langchain-ai/langchain。[56] Jerry Liu。《LlamaIndex》。2022年。网址：https://github.com/jerryjliu/llama_index。[57] Yiheng Wang、Yixin Chen、Shuo Li、Yifan Zhou、Bo Liu 等。《SciEvalKit：面向科学通用智能的开源评估工具包》。发表于 2025年。arXiv：2512.22334 [cs.AI]。网址：https://arxiv.org/abs/2512.22334。[58] Jiaqi Wei、Yuejin Yang、Xiang Zhang、Yuhan Chen、Xiang Zhuang 等。《从科学人工智能到智能体科学：自主科学发现综述》。发表于 2025年。arXiv：2508.14111[cs.LG]。网址：https://arxiv.org/abs/2508.14111。[59] Ming Hu、Chenglong Ma、Wei Li、Wanghan Xu、Jiamin Wu 等。《科学大语言模型综述：从数据基础到智能体前沿》。发表于 2025年。arXiv：2508.21148[cs.CL]。网址：https://arxiv.org/abs/2508.21148。[60] Anthropic。模型上下文协议。2024年。网址：https://github.com/modelcontextprotocol。[61] Yankai Jiang、Wenjie Lou、Lilong Wang、Zhenyu Tang、Shiyang Feng 等。《SCP：借助自主科学智能体全球网络加速科学发现》。发表于 2025年。arXiv：2512.24189[cs.AI]。网址：https://arxiv.org/abs/2512.24189。

[62] OpenAI。《GPT-5 系统卡更新：GPT-5.2》。发表于 2025年。 网址：https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf。[63] Yang Zhang、Xianrun Pan、Tianyu Shi、Zhifeng Gu、Zhaochang Yang 等。《P450Rdb：人工整理的细胞色素 P450 酶催化反应数据库》。发表于第 63 卷。2024年，第 35–42 页。 网址：https://www.sciencedirect.com/science/inproceedings/pii/S2090123223003168。[64] Yetang Wang、Xueying Zhang、Wentao Ning、Matthew A Lazzara、Minghu Ding 等。《AntAWS 数据集：南极洲自动气象站观测数据汇编》。发表于第 15 卷，第 1 期。2023年，第 411–429 页。 网址：https://essd.copernicus.org/articles/15/411/2023/。