科学数据湖：通过基于嵌入的本体对齐，整合八个学术来源中 2.93 亿篇论文的统一开放基础设施

Jonas Wilinski*汉堡工业大学（*TUHH*），德国汉堡
jonas.wilinski@tuhh.de ORCID：0009-0005-4672-7197

arXiv:2603.03126v1 [cs.DL] 3 Mar 2026

2026年3月

摘要

学术数据大多分散在彼此孤立的数据库中，这些数据库的元数据各不相同，且彼此之间缺少关联。我们介绍科学数据湖：一种可在本地部署的基础设施，基于 DuckDB 和简单的 Parquet 文件构建，通过 DOI 规范化整合八个开放来源，同时保留数据源级模式。这一资源包含约 960 GB 的 Parquet 文件，涵盖 22 种模式和 153 个 SQL 视图，其中有 2.93 亿篇可唯一识别的论文。我们采用基于 BGE-large 句子嵌入的本体对齐，将 4,516 个 OpenAlex 主题映射至 13 个科学本体（130 万个术语），生成 16,150 个映射，覆盖 99.8% 的主题（≥ 0.65 阈值）；在推荐的 ≥ 0.85 工作点上达到 $F1 = 0.77$，在包含 300 对样本的金标准评估中优于 TF-IDF、BM25 和 Jaro–Winkler 基线。我们通过 10 项自动化检查、跨来源引文一致性分析（成对皮尔逊 $r = 0.76$–0。87）以及分层人工标注进行了验证。四个示例展示了无法通过任何单一数据库实现的跨来源分析。该资源采用开源形式，可部署在单个硬盘上，也可通过 HuggingFace 远程查询，并附有结构化文档，适用于基于大语言模型（LLM）的研究智能体。

1 背景与摘要

追踪科学运作的大规模数据集的出现，催生了一个迅速发展的“科学学”领域，并为其建立了自身的数据基础设施、指标和分析框架 [1]。然而，支撑这一领域的数据库仍然彼此割裂：Semantic Scholar 为数百万篇论文提供重要的引文标记、开放获取元数据和全文覆盖 [2]；迄今最全面的开放索引 OpenAlex 提供领域加权引文影响力 (FWCI)、包含 4,516 个叶级主题的四级主题分类体系、121 K 家机构的地理编码机构隶属关系，以及 1,170 万项附有金额的资助奖项，且均采用 CC0 许可 [3]；SciSciNet 提供了一套具有里程碑意义的预计算科学学指标，包括颠覆性指数、非典型性评分和专利关联 [4]；Papers with Code 将论文与可复现代码关联起来，但该平台已停止积极运营，因此其存档快照是不可再生的资源 [5]；Retraction Watch 追踪科研诚信事件 [6]；Reliance on Science 则绘制专利至论文的引文关系图 [7]。这些来源都为开放科学记录作出了重大贡献；然而，没有任何单一来源能够涵盖所有

*通讯作者。

表 1：与现有学术数据集成系统的比较。

系统	源	多个	模式	开放	关键局限
SciSciNet [4]	1+	×	—	✓	单一文献计量学基础（MAG/OpenAlex）；不独立——附属跨数据源比较
PubGraph [10]	3	✓	已合并	✓	折叠为统一的Wikidata模式；数据源级别溯源信息丢失
SemOpenAlex [11]	1	×	—	✓	单一来源（仅 OpenAlex）；需要 SPARQL 专业知识-tise
Dimensions [12]	1	×	—	~	商业平台；程序化访问受限；无法完全复-现
ORKG [13]	3+	✓	KG	✓	聚焦于经过整理的研究贡献，而非文献计量-ric记录级比较
科学数据湖	**8**	✓	**已保留**	✓	需要 ~1太字节本地存储；基于快照

方面。希望将这些来源结合起来的研究人员——例如，研究颠覆性论文是否更有可能发布代码，或撤稿论文在不同数据库中是否呈现异常引文模式——必须编写临时性的集成脚本，而这些脚本很少被共享或复现。

对59个学术数据库的系统评估发现，后向与前向引文覆盖率存在显著差异 [8]。对书目数据源的大规模两两比较揭示了元数据、文献类型和引用次数方面的非平凡差异 [9]，但跨两个以上来源的记录级连接仍然少见。缺乏共享基础设施迫使每个研究团队重复相同的数据整理步骤，既浪费精力，又引入不一致性。

若干系统已开始弥补这一空白（表1）。SciSciNet [4] 提供了一个基于 Microsoft Academic Graph（现为 OpenAlex）构建的丰富“数据湖”，包含预计算的科学学指标以及与专利、资助项目和临床试验的关联，但其文献计量学基础来自单一索引，且不保留独立的数据源级模式以进行跨数据源比较。PubGraph [10] 使用 Wikidata本体将 Wikidata、OpenAlex 和 Semantic Scholar 合并为一个统一的知识图谱，但折叠了数据源级模式，从而牺牲了比较不同来源如何描述同一篇论文的能力。SemOpenAlex [11] 将 OpenAlex 重新编码为260亿条资源描述框架（RDF）三元组，提供语义网互操作性，但仍是单一数据源资源。Dimensions [12] 提供对综合性商业数据库的可通过SQL查询的访问，但其专有性质限制了可复现性。开放研究知识图谱（ORKG） [13]采取互补方法，侧重于对研究贡献（方法、结果、数据集）的结构化描述，将其作为经策展的 RDF 三元组，而非文献计量元数据；它汇集来自 Crossref、Semantic Scholar 和 OpenAIRE 的元数据，但不提供记录级引用次数、科学计量指标，也不对同一篇论文在独立数据库之间进行跨数据源比较。

科学数据湖并不试图取代这些各自优秀的资源，而是为科学学研究提供开放的基础设施基础——这类似于 SciSciNet [4] 为单一数据源指标所做的贡献，但将其扩展到了八个独立的数据来源。该资源有三项贡献。**首先**，它采用多源保留架构，将八个开放学术数据库整合为一个资源，同时保留各数据源的原生模式，从而能够在记录级别直接进行跨数据源比较。由于上游数据源（尤其是 OpenAlex）会在不同快照分区中调整其模式，转换流水线采用动态 Python脚本自动发现实体类型和列结构，而非依赖硬编码模式，从而确保基础设施能够随数据源的变化保持更新。**其次**，它采用基于嵌入的本体对齐 方法，利用 BGE-large 句子嵌入 [14],将 OpenAlex 的扁平主题分类体系与 13 个正式科学本体对接；在金标准评估中，该方法在推荐工作点达到 $F1 = 0.77$，表现优于 TF-IDF、BM25 和 Jaro–Winkler 基线。**第**三，它提供跨数据源、记录级

表2：集成数据源概览。记录数反映了本次发布所使用的快照版本。

来源	记录数	许可协议	主要内容
Semantic Scholar（S2AG）	231M	ODC-BY	引文、有影响力引文、开放获取
OpenAlex	479M	CC0	领域加权引文影响力、主题、类型、语言、资助奖项
SciSciNet	250M	知识共享署名许可	颠覆性、非典型性、团队规模
Papers with Code	513K	知识共享署名-相同方式共享许可	代码仓库、任务、数据集
Retraction Watch	69K	Open	撤稿原因和日期
Reliance on Science (RoS)	47.8M	知识共享署名—非商业性使用	专利—论文引文对
预印本到正式发表 (P2P)	146K	Open	bioRxiv/medRxiv DOI 至正式发表论文的 DOI
Crossref	—[a]	Open	DOI 元数据、参考文献列表

[a]Crossref 作为元数据增强和 DOI 验证层，而非独立的记录来源；不为其单独统计记录数。

比较层 (`unified_papers`，293M 行、29 列），支持跨所有数据源同时进行查询，从而实现多数据库引文可靠性评估等分析，而这些分析无法通过单一数据源或两两比较完成。除传统研究工作流外，原生 SQL 架构和结构化文档也使数据湖尤其适用于新兴的基于大语言模型的研究智能体 [15]：机器可读的模式参考旨在帮助用户浏览由 153 个视图和 22 个模式构成的结构。

2 方法

2.1 数据来源

Science DataLake 集成了八个开放的学术数据源，每个数据源都提供不同的元数据维度（表 2）。

Semantic Scholar 学术图谱 (S2AG) [2] 提供约 2.31 亿篇论文的文献计量元数据，包括引用次数、重要引用次数（基于引用语境分析）以及开放获取状态标记。

OpenAlex [3] 是一个开放目录，收录了 4.79 亿篇学术作品，包含领域加权引用影响力 (FWCI)、四级主题分类体系（领域、研究领域、子领域和主题，共有 4,516 个叶级主题）、文献类型和语言标注。

SciSciNet [4] 为 OpenAlex 记录补充了预先计算的科学学指标，包括颠覆性指数 CD_5 [16, 17], 期刊非典型性 z评分、团队规模指标，以及 2.5 亿篇论文的专利引用次数。

Papers with Code [5] 将 51.3 万篇机器学习论文与其相关代码仓库、基准任务和数据集关联起来。

Retraction Watch [6] 收录约 6.9 万篇撤稿或更正的出版物，并提供结构化的撤稿原因和日期。

Reliance on Science (RoS) [7] 提供了来自全球专利局的 4780 万条专利与论文之间的引用关系，并附有置信度分数和引用位置元数据。

预印本到正式发表 (P2P) 提供了约 14.6 万条映射，将 bioRxiv 和 medRxiv 预印本 DOI 与其对应的正式发表版本 DOI 关联起来。

Crossref 提供 DOI 元数据和参考文献列表，用于 DOI 验证和补充关联。

图1 展示了各数据源的时间覆盖范围，揭示出结构上的差异：Ope- nAlex 的历史可追溯至数个世纪前；S2AG 集中于近几十年，侧重计算机科学；SciSciNet 的指标则止于 2022年左右。

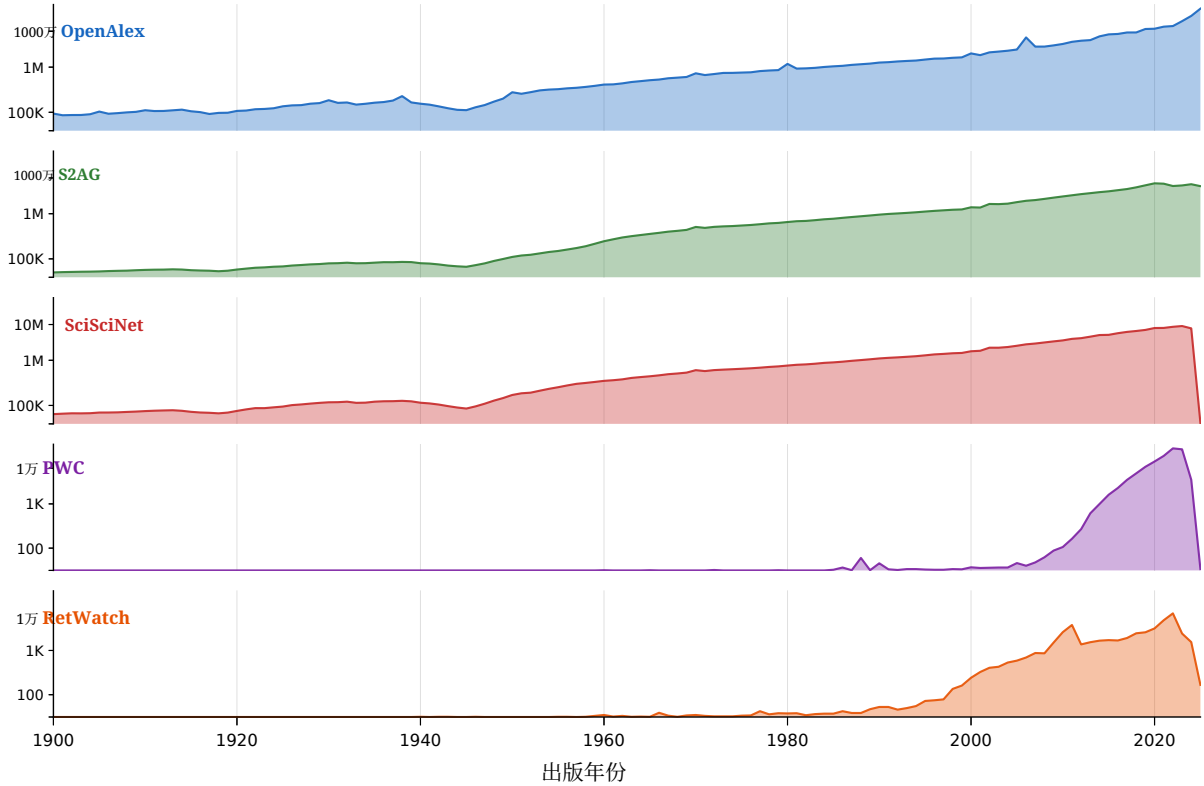

图1：各数据源的时间覆盖范围（对称对数刻度）。统一索引中带 DOI 的记录的出版年份分布（完整的数据源规模见表 2 ）。对称对数纵轴在零附近呈线性，在更高数值处呈对数变化，既展现了历史深度，也揭示了近期增长。OpenAlex、S2AG 和 SciSciNet 的数据均可追溯至 1900 年，每年涵盖数万篇论文；这三个数据源在 ~1960 之后均呈现出明显的增长加速。SciSciNet 在 ~2022 之后出现明显截止，此时其指标计算已结束。专业数据源（PWC、Retraction Watch）的时间覆盖范围较窄，主要集中在过去二十年。

2.2 架构

科学数据湖采用基于 *Parquet* 文件的视图架构，并使用 DuckDB [19]（图2）。每个数据源首先从其原生格式（JSON Lines、CSV、N-Triples）转换为列式 Apache Parquet 文件，磁盘占用总计约 960 GB。一个轻量级 DuckDB 数据库（~270KB）定义了 153 个 SQL 视图，分布在 22 个模式中，并引用这些 Parquet 文件，而不复制数据。

模式设计遵循两项原则。**源级保留**：每个数据源都在专属命名空间中保留其原生模式（例如 `s2ag.papers`、 `openalex.works`、`sciscinet.paper_metrics`），便于直接检查不同数据库如何表示同一篇论文。**通过 `xref` 模式进行交叉引用**：三个物化视图实现了跨数据源关联——`doi_map`（DOI 规范化）、 `unified_papers` （含 2.93 亿行的连接表）和 `topic_ontology_map` （本体对齐）。

我们选择 DuckDB，而非 PostgreSQL 等传统的服务器型系统，原因有三：它采用可嵌入式、无服务器的设计，无需安装或管理；其列式引擎针对文献计量研究常见的在线分析处理（OLAP）查询模式进行了优化；它还能通过 SQL 视图直接读取 Parquet 文件，无需导入数据，使数据库文件保持在 300 KB 以下，同时可访问近 1 太字节的数据。

系统支持双模式访问：可部署在大容量硬盘上，以进行全速分析查询；也可通过 HuggingFace 托管的 Parquet 文件远程访问，供没有本地存储的用户使用。

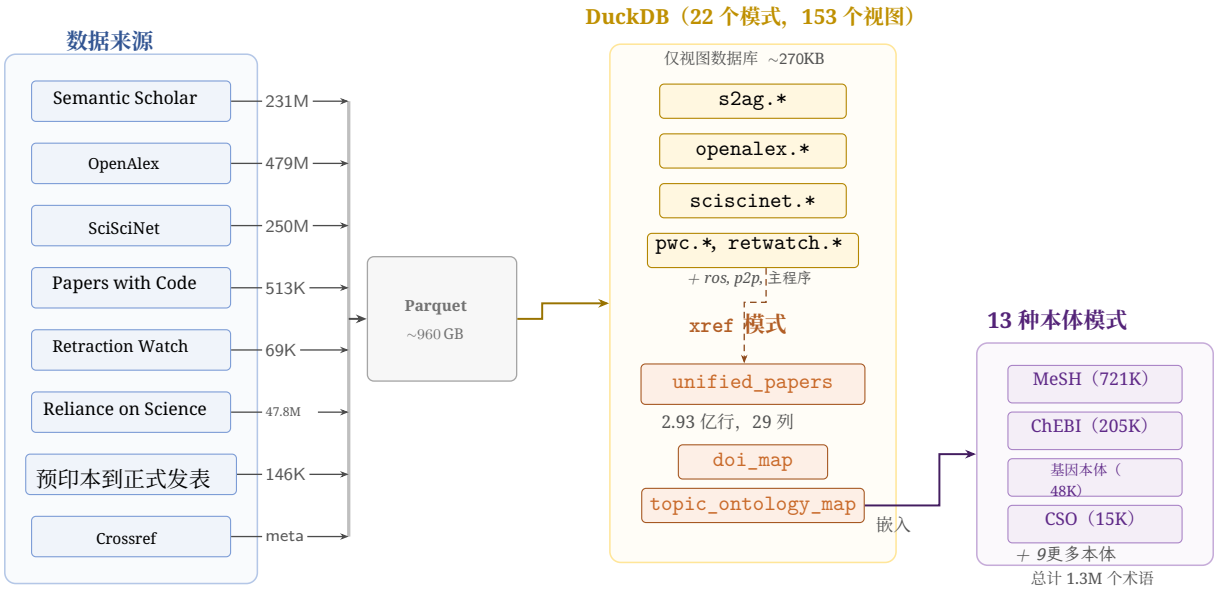

图2：科学数据湖的架构。八个开放学术数据源（左）被转换为 Apache Parquet 格式（~960 GB），并通过轻量级 DuckDB 数据库（中）以 SQL 视图的形式提供。每个数据源都保留其原生模式，以确保源级保真度。交叉引用 `xref` 模式（橙色）通过 DOI 规范化（`unified_papers`，2.93 亿行）关联记录，并通过基于 BGE-large 嵌入的对齐，将 OpenAlex 主题与 13 个科学本体连接起来（右）。

表 3：不同数据源之间的 DOI 格式差异及所采用的规范化处理。

来源	原始 DOI 格式	规范化
S2AG	小写，无前缀（`10.1038/...`）	规范形式（无）
OpenAlex	小写，`https://doi.org/`前缀	去除前缀
SciSciNet	小写，`https://doi.org/` 前缀	去除前缀
Papers with Code	小写，无前缀	None
Retraction Watch	小写，无前缀	None
Crossref	大小写混合	小写

由主命令行界面脚本（`datalake_cli.py`）编排的可复现流水线可自动完成整个工作流：下载数据源快照、转换为 Parquet、创建 DuckDB 视图、物化交叉引用表，并构建本体关联。

2.3 DOI 规范化与记录关联

数字对象标识符 (DOI) 是跨来源记录关联的主键，但各来源以不兼容的格式存储这些标识符（见表 3）。

所有 DOI 均已规范为标准小写无前缀格式。 `xref.doi_map`视图通过一组针对各个来源的子查询来实现这种规范化，每个子查询都会应用相应的转换。

最终生成的 `xref.unified_papers` 表包含 293,123,121 个唯一 DOI，共 29 列，数据来自所有来源，其中包括六个布尔覆盖标志，用于指示每篇论文出现在哪些来源中。表4总结了成对覆盖率。

OpenAlex 的覆盖范围最广，涵盖了所有 DOI 的 99.67%，这与其作为综合性开放索引的定位一致。SciSciNet 和 S2AG 覆盖了约一半的 DOI 范围，反映出它们侧重于收录引文数据足以用于指标计算的论文。该

表 4：2.93亿篇统一论文的跨来源覆盖率。每个单元格显示列来源中的论文同时也存在于行来源中的百分比。

	OpenAlex	SciSciNet	S2AG	PWC	RetWatch	RoS
覆盖率（%）	99.67	54.08	45.52	0.048	0.020	0.19

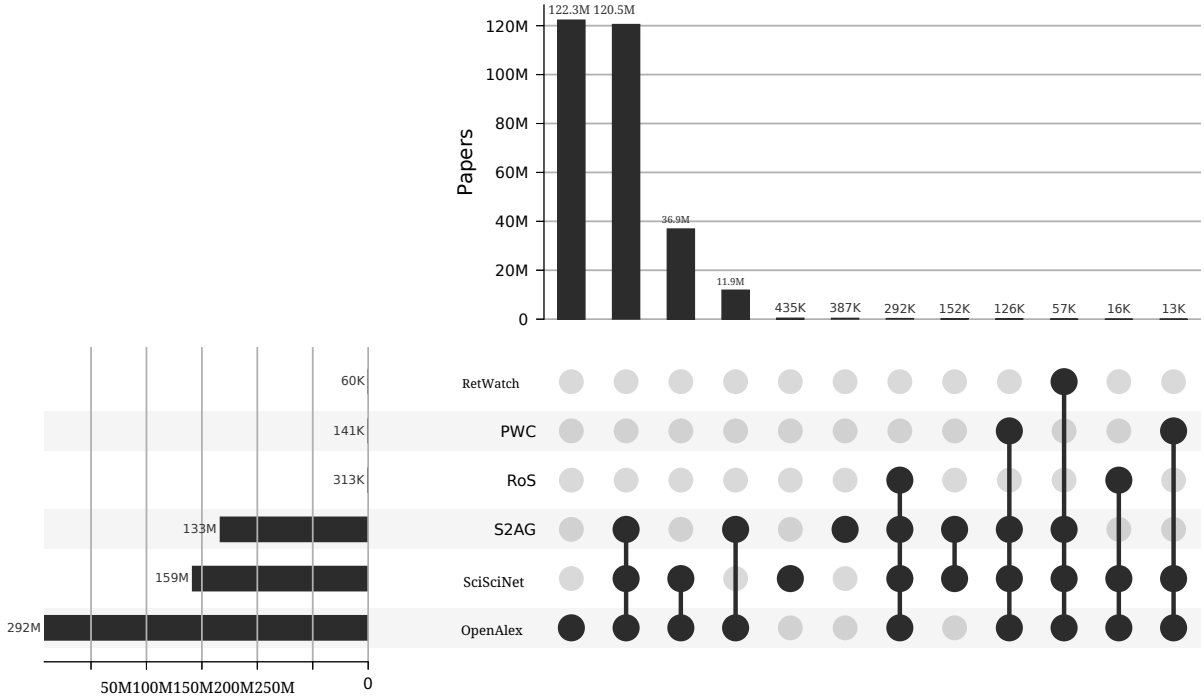

图3：展示六个数据来源交集结构的 UpSet图。柱形表示各来源组合中的论文数量。在观察到的34种组合中，OpenAlex、SciSciNet 和 S2AG 三者的重叠构成了最大的多来源交集。

专业数据源（Papers with Code、Retraction Watch、Reliance on Science）贡献了规模较小但独特的记录集，无法从三个大型数据库中获取。

图 3 展示了六个来源之间重叠情况的UpSet图，揭示了34种已观察到的来源组合。占主导地位的组合仅包含OpenAlex（45.0%），其次是OpenAlex、SciSciNet和S2AG三者重叠（38.2%）。

2.4 基于嵌入的本体对齐

OpenAlex将论文归入由4,516个主题构成的扁平主题分类体系，该体系分为四个层级（252个子领域、26个领域、4个学科域），但这些主题尚未映射到能够编码领域专属知识的正式科学本体。为弥合这一差距，我们开发了一种基于嵌入的对齐方法，将OpenAlex主题映射到13个科学本体，这些本体总共包含130万个术语。

这13个本体涵盖了多样的科学学科域：医学主题词表 (MeSH；721K个术语) [20], 生物学兴趣化学实体 (ChEBI；205K) [21], NCI 同义词库 (NCIT；204K) [22], 基因本体 (GO；48K) [23], AGROVOC (42K) [24], 计算机科学本体 (CSO；15K) [25], 疾病本体 (DOID) [26], 人类表型本体 (HPO) [27], EDAM生物信息学本体 [28],联合国教科文组织叙词表[1], 标准经济学叙词表 (STW)[2],

[1] https://vocabularies.unesco.org/browser/thesaurus/en/
[2] https://zbw.eu/stw/

表 5：本体对齐质量等级。每个等级均包含相似度达到或超过相似度阈值的所有映射。

质量等级	相似度	映射	涵盖主题
完全匹配	≥ 0.95	85	71 (1.6%)
高质量	≥ 0.85	2,527	1,647 (36.5%)
All	≥ 0.65	16,150	4,509 (99.84%)

表6：本体对齐方法比较。所有方法使用相同的主题和本体术语集。映射和主题数量均采用各方法的运行阈值。精确率 (P)、召回率 (R) 和 F1 均依据分层的300对金标准标注集进行评估（见第 4 节）；只有当标注员将映射标注为 正确 （严格标准；部分相关将计为假阳性）。F1 根据未四舍五入的 P 和 R 计算；对显示值进行四舍五入可能导致结果略有差异。

方法	映射	主题	P	R	F1
Jaro–Winkler (≥ 0.90)	937	675	0.80	0.52	0.63
TF-IDF余弦相似度	4,861	2,620	0.61	0.84	0.71
BM25	14,929	4,513	0.45	0.92	0.61
BGE-large（我们的方法）	16,150	4,509	0.38	1.00	0.55
BGE-large (≥ 0.85)	2,527	1,647	0.67	0.89	0.77

物理学主题词 (PhySH)[3], 和数学主题分类 (MSC2020)[4]。每个本体都从其原生格式（开放生物医学本体 (OBO)、简单知识组织系统 (SKOS)/资源描述框架 (RDF)、N-Triples、CSV）转换为统一的 Parquet 表示形式，转换时使用了特定格式的解析器；同时，这些本体也被加载到 Oxigraph RDF三元组存储中，以便执行 SPARQL 查询。

我们采用了混合对齐策略。对于 10 个较小的本体（包含同义词在内共 291K 个术语），我们使用智源通用嵌入（BGE）large-en-v1.5 模型 [14] （3.35 亿个参数，1024 维）计算稠密嵌入，并通过 Facebook人工智能相似性搜索（FAISS）索引 [29] 执行最近邻搜索。对于三个最大的本体（MeSH、ChEBI、NCIT），它们合计包含 110 万个术语，若采用嵌入将主导嵌入空间，因此我们使用精确字符串匹配来确保精确率。

表5 汇总了各质量等级的对齐质量，并列举了代表性示例。在精确匹配层级（≥ 0.95），主题与几乎完全相同的本体术语相对应；例如，OpenAlex 的“机器学习”主题映射到 CSO 的“machine learning”（相似度为 0.98）。在高质量层级（≥ 0.85），则匹配语义相关但名称不同的概念，例如，“医学人工智能”映射到 EDAM 的“医学信息学”（0.87）。放宽至 ≥ 0.65 后，共得到 16,150 个映射，覆盖 4,516 个主题中的 4,509 个（99.84%），其中包括更宽泛的关联，例如，“土壤化学”映射到 AGROVOC 的“土壤化学物理性质”（0.68）。

为将嵌入方法置于背景中，我们将其与三种标准基线方法进行了比较（表6）：字符级 TF-IDF 余弦相似度（2 至 4 元字符组， `char_wb` 分析器）、BM25 [30] （对按空白字符分词的标签计算 Okapi BM25，进行最小-最大归一化，每个主题取前 20 项），以及阈值为 0.90 的 Jaro–Winkler 字符串相似度。所有方法均针对相同的 4,516 个主题和 291K 个本体术语；为确保一致性，对于大型本体（MeSH、ChEBI、NCIT），所有情况下均采用完全匹配。

嵌入方法能够捕捉词汇方法无法识别的语义相似性：例如，OpenAlex 的“医学人工智能”主题映射到 EDAM 的“医学

[3] https://physh.org/
[4] https://msc2020.org/

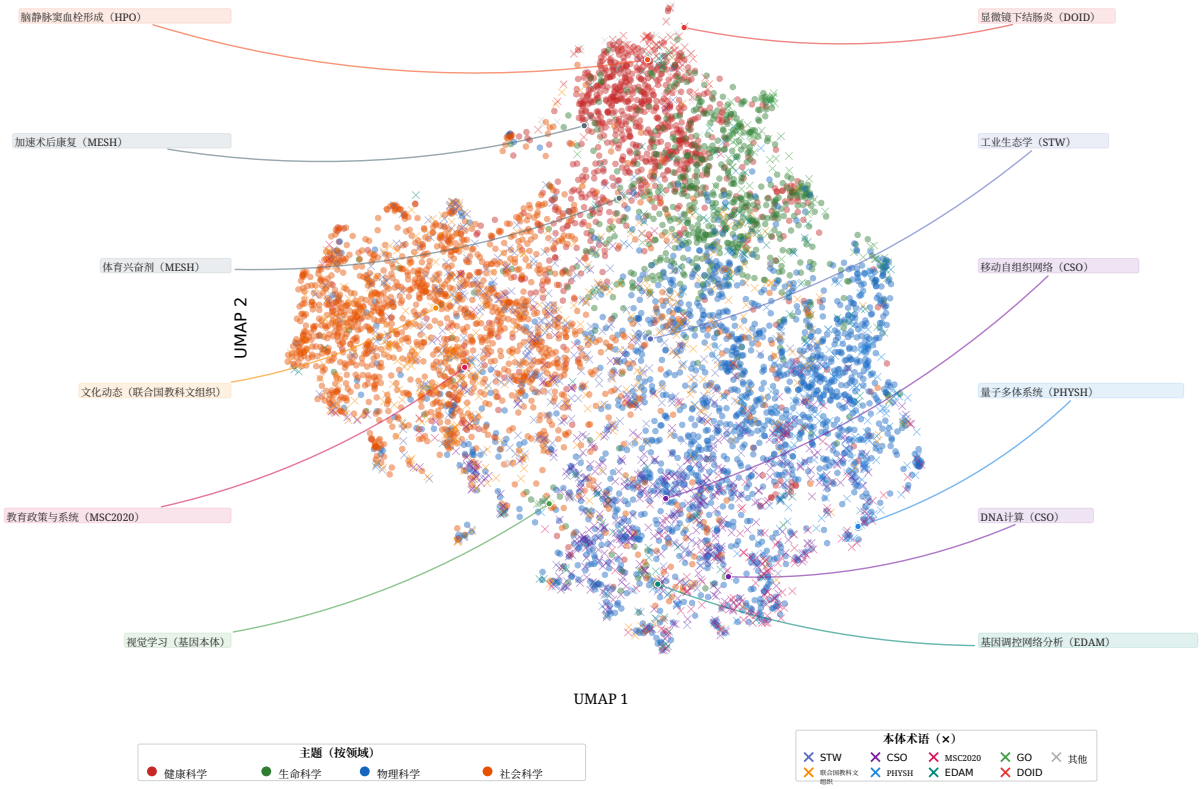

图 4：OpenAlex 主题（点）和匹配的本体术语（叉号）的 BGE-large 嵌入 UMAP 投影，按 OpenAlex领域着色。语义簇自然形成，特定领域的本体术语与相应主题聚集在一起。

信息学”（余弦相似度为 0.87）和 NCIT 的“生物医学信息学”（0.85），而这三种基于字符串的基线方法均无法发现这种对应关系。

图4 使用 UMAP 可视化联合嵌入空间 [31]，，展示 OpenAlex 主题如何按领域聚类，并与领域本体中的术语对齐。图5展示了本体按领域覆盖热图，证实不同本体专注于不同的科学领域：MeSH 在健康科学领域占主导地位，CSO 覆盖计算机科学，GO 涵盖分子生物学，而 AGROVOC 则连接农业与环境科学。

3 数据记录

科学数据湖托管于 HuggingFace Datasets（https://huggingface.co/datasets/JOnasW/science-datalake，DOI：10.57967/hf/7850），该平台提供用于引用的持久 DataCite DOI。远程用户无需下载完整数据集，即可通过 DuckDB 的 `hf://` 协议直接查询 Parquet 文件。由于许可限制，托管在 HuggingFace 上的版本不包含 Semantic Scholar（根据 ODC-BY 许可分发，并附加服务条款，仅限非商业研究用途）或 Reliance on Science（CC BY-NC 4.0）。可使用所提供的下载脚本和免费的 Semantic Scholar API 密钥在本地获取这两个来源；下载后，流水线会自动将其整合到统一模式中。

完整数据集包含约 960 GB 的压缩 Apache Parquet 文件，组织为 22 个模式目录，每个目录包含一个或多个与该模式中各表相对应的 Parquet 文件。轻量级 DuckDB 数据库文件（~270KB）定义了 153 个 SQL 视图，这些视图

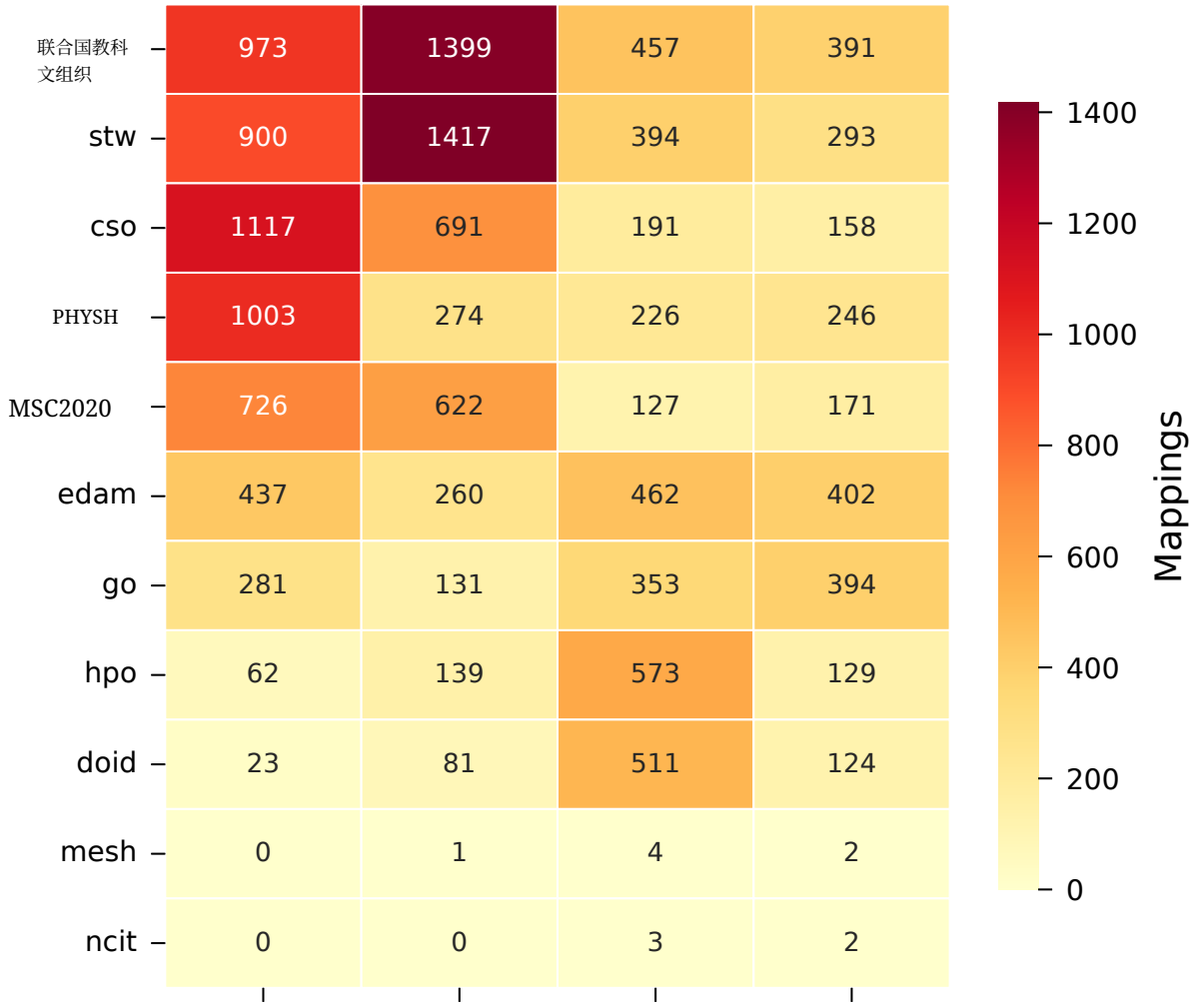

图 5：本体覆盖热力图，显示各本体与各 OpenAlex领域之间高质量映射的数量（相似度 ≥ 0.85）。多本体设计确保覆盖所有科学领域。

引用这些 Parquet 文件，并可使用所提供的流水线脚本从源重新生成。

表 7：自动化合理性检查摘要。全部数据集中进行的 10 项检查均通过，未发现违规。

#	检查项	结果	详情
1	DOI 格式（无前缀、小写）	PASS	0 项违规 / 2.93 亿
2	覆盖标记与数据是否存在相符	PASS	0 处不匹配（OA、S2AG、SSN）
3	主键唯一性（无重复 DOI）	PASS	293,123,121 个唯一 = 总计
4	OpenAlex ID 格式与可连接性	PASS	0 处格式违规；69% 主题匹配
5	本体映射：没有孤立的主题 ID	PASS	0 个孤立主题_ID
6	RoS 与 OpenAlex 的关联（1 万条样本）	PASS	匹配率为 86%
7	引文跨源相关性	PASS	$r = 0.76$–0.87 两两相关
8	年份分布（NULL/无效）	PASS	NULL：0.53%，无效：0.001%
9	抽查已知论文	PASS	Wakefield 撤稿标记正确
10	案例计数可复现性	PASS	4 项计数全部完全匹配

主要数据模式及其内容如下：

- `s2ag`：来自 Semantic Scholar 的论文（2.31 亿篇）、摘要、引文、作者和出版场所。
- `openalex`：来自 OpenAlex 的作品（4.79 亿篇）、作者、机构、来源、主题、概念、出版商和资助机构。
- `sciscinet`：来自 SciSciNet 的论文指标（2.5 亿条）、颠覆性指数、非典型性评分和团队规模指标。
- `pwc`：来自 Papers with Code 的论文（51.3 万篇）、代码链接、任务、数据集和方法。
- `retwatch`：来自 Retraction Watch 的撤稿论文（6.9 万篇），包含撤稿原因和日期。
- `ros`：来自 Reliance on Science 的专利—论文引文对（4,780 万）。
- `p2p`：来自 bioRxiv/medRxiv 的预印本至已发表论文 DOI 映射（14.6 万）。
- `xref`：跨来源关联表——`unified_papers`（2.93 亿，29 列）、`doi_map`、`topic_ontology_map`（16,150 个映射），以及时间范围校验视图（`source_temporal_coverage`、`paper_temporal_flags`）。
- 13本体模式（例如 `mesh`、`go`、`cso`），每个模式都包含 `*_terms`、`*_hierarchy`，并可选包含 `*_xrefs`表。

每个来源均保留其原始许可协议：CC0（OpenAlex）、ODC-BY（S2AG）、CC BY 4.0（SciSciNet）、CC BY-SA 4.0（Papers with Code）、CC BY-NC 4.0（Reliance on Science），其余来源则采用开放／公有领域许可。用户应遵守其查询来源中适用的最严格许可协议。

4 技术验证

我们通过 10 项自动化合理性检查（表 7）、跨来源引文相关性分析以及人工检查本体映射，对 Science DataLake 进行了验证。

4.1 DOI 与模式完整性

检查 1–5 用于验证交叉引用层的结构完整性。DOI 格式检查（检查 1）确认，`unified_papers` 中的 2.93 亿条记录均采用标准小写无前缀格式，违规数为零。覆盖标记（检查 2）是布尔列，用于指示每篇论文是否出现在 OpenAlex、S2AG 和 SciSciNet 中；所有标记都正确反映了相应来源表中数据的有无。主键唯一性（检查 3）确认每个 DOI 只出现一次。OpenAlex ID 格式检查（检查 4）验证所有 ID 均符合预期模式，且 69% 的论文成功关联到

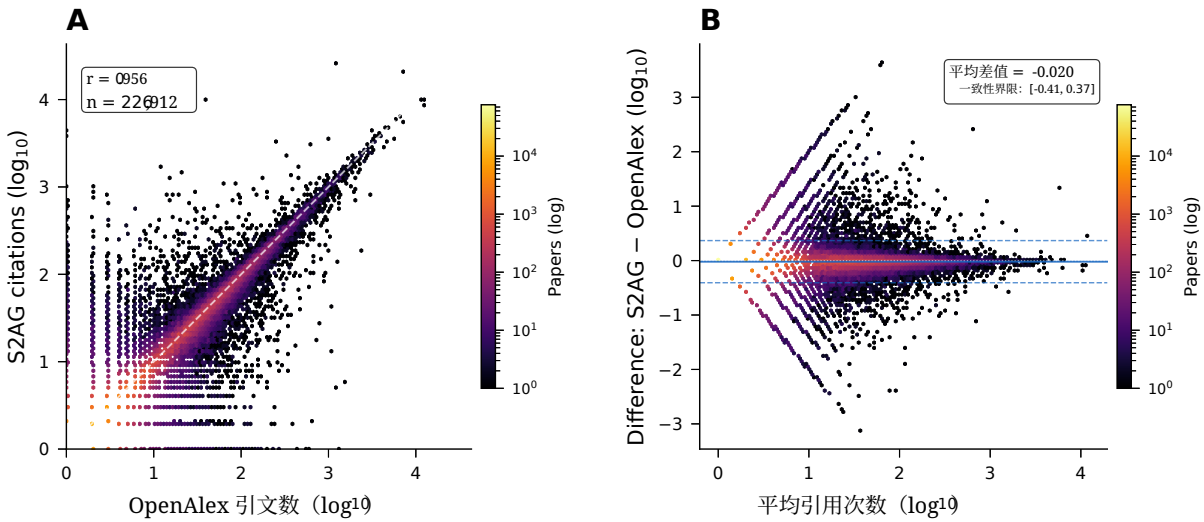

图 6：Semantic Scholar（S2AG）与 OpenAlex 之间引用次数一致性的 Bland–Altman 图。每个点代表一篇论文；x轴显示两个来源的平均引用次数，y轴显示差值（S2AG − OpenAlex）。虚线表示平均差值和 95% 一致性界限。

`works_topics` 表（其余论文在 OpenAlex 中没有主题分配）。检查 5 验证本体映射表中的每个主题 ID 均存在于 OpenAlex 主题分类体系中，未发现孤立项。

4.2 跨来源引文一致性

为评估不同数据库之间引用次数的一致性，我们计算了同时收录于三个大型来源（S2AG、OpenAlex、SciSciNet；$n \approx 121$M）的论文的成对皮尔逊相关系数。相关系数分别为：S2AG–OpenAlex $r = 0.76$、S2AG–SciSciNet $r = 0.87$，以及 OpenAlex–SciSciNet $r = 0.86$。平均绝对差分别为 4.14（S2AG–OA）、3.84（S2AG–SSN）和 2.30（OA–SSN）条引文。

图 6 展示了 S2AG 与 OpenAlex 之间引文一致性的 Bland–Altman 分析。该图显示，分歧随引用量增大而增加，并识别出系统性离群值——最显著的是一篇论文在 S2AG 中有 257,887 条引文，而在 OpenAlex 中为零，这可归因于引用计数方法与覆盖范围的差异。

超过 $r = 0.8$ 的三者相关性证实，这三个来源提供了大体一致的引文信息，而不可忽视的分歧（尤其是 $r = 0.76$处的 S2AG–OA）凸显了保留全部三项计数以进行敏感性分析的价值。

我们有意保留全部三项独立的引用次数，而不是强加单一的消解值。六个布尔覆盖标志（`has_s2ag`, `has_openalex`, `has_sciscinet`,`has_pwc`, `has_retraction`, `has_patent`）充当隐式冲突检测指标：出现在较少来源中的论文，在比较依赖引文的指标时需要格外谨慎。实施自动冲突解决将需要对来源可靠性作出编辑判断，而这种判断可能因研究领域、时代和研究问题而异——我们认为这些选择应留给分析者。

4.3 本体对齐验证

为通过人工检查之外的方式量化对齐质量，我们从总计 16,150 个映射中抽取了 300 个作为分层样本：精确层级抽取 50 个（≥ 0.95），高质量层级（0.85–0.95）抽取 100 个，中等范围（0.75–0.85）抽取 100 个，临界层级（0.65–0.75）抽取 50 个，并按比例

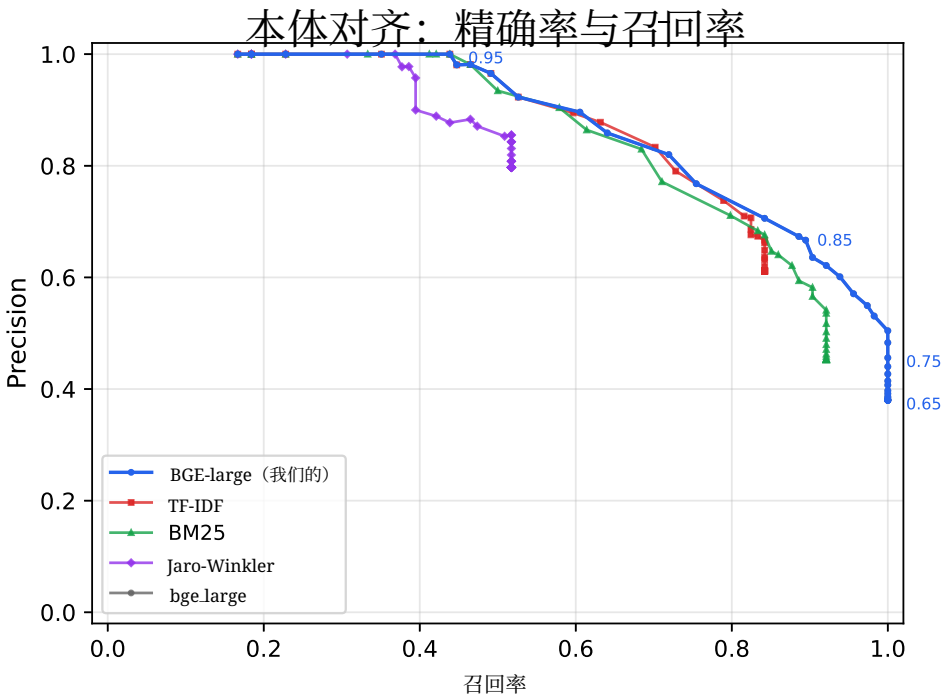

图 7：四种本体对齐方法的精确率—召回率曲线，展示了余弦相似度阈值从 0.60 到 1.00 的变化。标注点表示关键阈值。

表示每个层中的全部 13 个本体。每对映射均由独立人员标注为正确（语义等价）、部分相关（存在有意义的关系，但并不等价）或错误（无关）。

在严格评估下（仅将正确的 = 真正例计入），各层的精确率为：精确层级 1.00（50/50）、高质量层级 0.51（51/100）、中等范围 0.13（13/100）以及边缘层 0.00（0/50）。值得注意的是，在精确层级和高质量层级中均未发现错误映射——其中所有被标记为不正确的标签都属于部分相关（即有意义但并非等价的关系）。在推荐的 ≥ 0.85工作点，BGE-large 达到 $P = 0.67$， $R = 0.89$，$F1 = 0.77$。我们使用同一金标准集评估了三种词汇基线（TF-IDF、BM25、Jaro–Winkler；见表 6）。图7 展示了四种方法在不同相似度阈值下的精确率—召回率权衡。

高质量层级中的典型边缘匹配包括合理的跨领域关联，例如“生物电子学”（OpenAlex）→ “生物传感器”（EDAM，相似度为 0.856），这反映了有意义的语义邻近性，而非错误。

4.4 时间覆盖范围防护措施

各个来源的时间覆盖范围会给纵向分析带来注意事项。为便于通过程序获取这些信息， `xref` 模式包含两个辅助视图。`xref.source_temporal_coverage` 记录每个来源的年份范围和覆盖类型（例如，SciSciNet 指标截至约 2022年，RoS 专利引用在 2023年末之前最为完整）。`xref.paper_temporal_flags` 提供逐篇论文布尔标志： `sciscinet_metrics_stale`（当论文年份晚于 2022年且该论文有 SciSciNet 记录时为 true）、 `ros_coverage_incomplete`（当年份晚于 2023且该论文有专利引文时为 true），以及 `year_missing` （当出版年份为 NULL 时为 true）。用户可以将这些标志联接到任意分析查询中，以筛选出时间上可靠的子集——例如，在计算领域层面的颠覆性趋势之前，过滤掉颠覆性分数已过时的论文。

4.5 已知局限

用户在根据近期趋势得出结论之前，应核实每个来源的时间覆盖范围。OpenAlex 主题分类体系可能会随快照而变化，从而可能影响本体映射的稳定性。

5 使用说明

5.1 设置

克隆代码库，通过`datalake_cli.py` 运行流水线（该流水线会下载、转换并关联所有来源），然后连接到生成的 DuckDB 数据库，即可部署科学数据湖。对于没有本地存储的用户，可通过 DuckDB 的 `httpfs` 扩展直接查询托管于 HuggingFace 的 Parquet 文件。以下所有查询均使用标准 SQL，并在 DuckDB 中执行。

5.2 小节1：颠覆性、代码采用与本体格局 e

本小节考察发布代码的论文是否呈现出不同于未发布代码论文的颠覆性特征，并将这一模式映射到按本体划分的学科域中。

结合 `sciscinet.paper_metrics` （颠覆性指数 CD_5）、 `xref.unified_papers` （来自 Papers with Code 的代码可用性标记）和 `xref.topic_ontology_map` （本体桥接），我们识别出 139,873 篇关联有代码仓库的论文（占统一表格的 0.048%）。有代码的论文平均 CD_5 为 -0.0005，相比之下，没有代码的论文为 $+0.0026$ ，这表明发布代码的论文往往略微更具整合性（基于现有研究），而非颠覆性。本体映射显示，这一模式因学科域而异：计算机科学主题（通过 CSO 映射）表现出最高的代码采用率，而生物医学主题（通过基因本体和 MeSH 映射）的代码比例较低，但颠覆性差异更大。

这项分析只有在能够同时访问颠覆性分数（SciSciNet）、代码标记（Papers with Code）、主题分配（OpenAlex）和本体桥接（我们的关联数据）时才有可能——这四种资源不存在于任何单一数据库中。

5.3 小插曲 2：撤稿特征与本体富集

本小插曲根据撤稿前影响力指标刻画撤稿论文，并识别出撤稿率异常的本体领域。

通过 объедин合并 `retwatch.retracted_papers`、 `sciscinet.paper_metrics` 和 `xref.unified_papers`，我们获得了 58,775 篇具有相关 SciSciNet 指标的撤稿论文。撤稿论文的平均颠覆性为 0.0035，而非撤稿论文为 0.0026；引用次数最多的撤稿论文在撤稿前累计获得了 8,062 次引文。本体层面的富集分析揭示了撤稿热点：映射至“人工智能应用”的主题显示出 $394\times$ 富集，而“先进技术”主题相较于基准撤稿率显示出 $338\times$ 富集。

该分析需要撤稿标记（Retraction Watch）、颠覆性分数（SciSciNet）、引用次数（OpenAlex）以及本体映射——这些数据在任何单一来源中都无法同时获得。

5.4 小插图 3：专利影响与多本体足迹

本小插图按本体领域细分，量化了被专利引用的论文的引用及影响特征。

结合 `ros.patent_paper_pairs`、 `xref.unified_papers`和 `xref.topic_ontology_map`，我们识别出312,929篇被专利引用的论文（占统一表的0.107%）。这些论文具有

显著更高的影响力：其平均引用次数为94.3，而未被专利引用的论文为16.1（5.8×）；领域加权引文影响力平均值为4.7，而后者为1.5（3.1×）。多本体足迹显示，被专利引用的论文集中于应用型学科域：由MeSH映射的健康科学主题、由CSO映射的计算机科学主题，以及由ChEBI映射的化学主题占据主导地位。

时间覆盖范围局限仍然适用：由于专利处理延迟，RoS 专利引用在 2023 年末之前最为全面，因此最新发表的论文可能尚未完整关联专利。

5.5 示例 4：跨来源引文可靠性

本示例展示了如何跨三个独立数据库进行记录级引文比较。

我们将范围限定为三个大型来源（S2AG、OpenAlex、SciSciNet）均收录的 1.21 亿篇论文，并计算两两引文相关性：S2AG–OpenAlex $r = 0.76$、S2AG– SciSciNet $r = 0.87$ 和 OpenAlex–SciSciNet $r = 0.86$。平均绝对差介于 2.3 至 4.1 条引文之间。相对不一致性在低引文量论文中最为明显（引文数为 <10 的论文平均相对差异为 $\sim$20%），而在高引文量论文中则有所减弱。最极端的离群值是一篇在 S2AG 中有 257,887 条引文、在 OpenAlex 中却为零的论文，这说明覆盖范围和方法差异可能导致显著的记录级差异。

第 4 节利用这些相关性验证数据完整性，而本示例展示的是分析工作流程本身——说明研究人员如何诊断来源特定偏差，并根据研究设计选择最合适的引用次数。三方比较要求来自独立来源的平行引用次数共存于单个可查询表中——在这一规模下，通过成对的基于 API 的比较来构建这种配置并不切实际。

图 8 总结了全部四个示例的结果。

5.6 人工智能辅助查询

科学数据湖包含一份结构化模式参考（`SCHEMA.md`，约 1,200 行），旨在为基于大语言模型（LLM）的编码代理提供上下文。该文档列出了每张表的名称、列、数据类型、行数和大小级别，以及九种跨数据集连接策略和常见查询范式；其格式针对大语言模型的读取进行了优化，而非用于叙述性阅读。例如，诸如“找出计算机科学领域中最具颠覆性的、具有开源代码的论文，并检查其撤稿状态”这样的查询，需要连接四种模式（`sciscinet`、`xref`、`pwc`、`retwatch`），并进行适当的 DOI 规范化；`SCHEMA.md` 汇总了构造此类连接所需的元数据。

我们没有构建定制的自然语言界面（这需要维护，也会限制查询范围），而是提供结构化文档，供任何通用大语言模型智能体使用。随着大语言模型能力不断发展，同一份模式参考文档无需修改即可继续发挥作用。对不同大语言模型提供商的文本到 SQL 性能进行系统评估，超出了本数据描述的范围。

5.7 局限性与可扩展性

科学数据湖继承了其组成来源的局限性。时间覆盖范围各不相同：SciSciNet 的指标数据截至约 2022年，RoS 存在专利处理延迟，而 OpenAlex 快照日期可能比实时数据晚数周至数月。没有 DOI 的论文（根据研究领域和时代不同，估计占 5–15%）无法进行跨源关联。本体映射依赖当前的 OpenAlex 主题分类体系，而该体系可能会随快照更新而演变。

该架构旨在便于扩展：添加新的数据源需要编写 Parquet转换器，并在流水线配置中注册模式。我们欢迎社区贡献更多来源、本体或交叉引用方法。

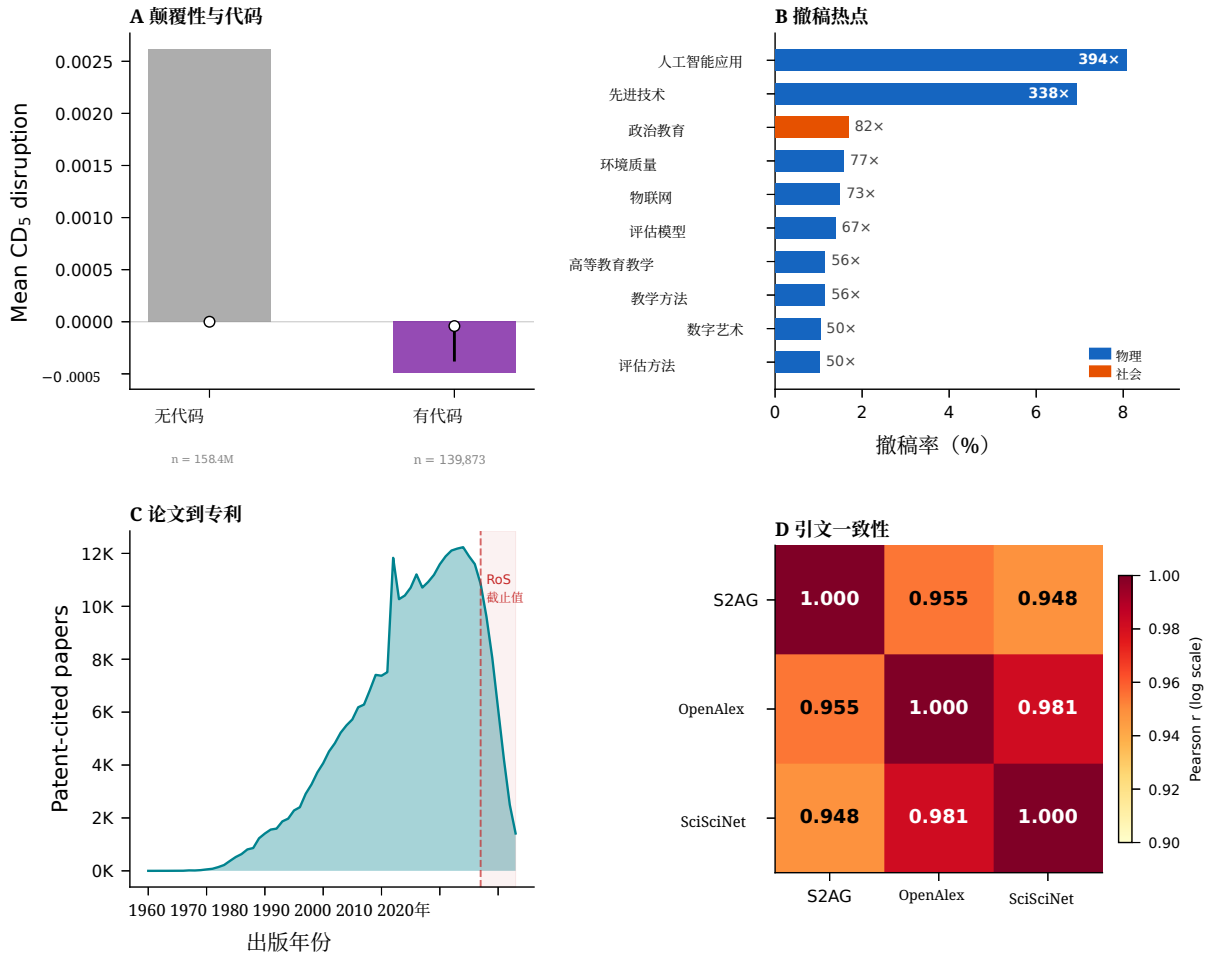

图 8：综合案例结果（2×2 个面板）。左上：有代码与无代码论文的颠覆性分布（示例 1）。右上：按本体领域划分的撤稿富集（示例 2）。左下：被专利引用与未被专利引用论文的引文分布（示例 3）。右下：三个来源之间的成对引文一致性（示例 4）。

6 代码可用性

用于构建和查询科学数据湖的所有代码均可在公开的 GitHub 仓库中获取（https://github.com/JOnasW/science-datalake）。该流水线使用 Python 3.12 实现，主要依赖项如下：DuckDB 1.4。2 [19], PyArrow 22.0 和 sentence-transformers 5.2.2（用于本体对齐）。

该流水线包含五个阶段，每个阶段均作为主命令行界面脚本的一个子命令实现：

1. **下载** (`datalake_cli.py download`)：从官方分发渠道（S3 存储桶、API、直接下载）获取数据源快照。2. **转换** (`datalake_cli.py convert`)：将各数据源从原生格式（JSON Lines、CSV、N-Triples）转换为列式 Apache Parquet 文件。3. **创建视图** (`create_unified_db.py`)：生成包含 22 个模式、153 个 SQL 视图的 DuckDB 数据库。4. **物化** (`materialize_unified_papers.py`)：通过 DOI 规范化和多源记录链接构建 `xref.unified_papers`连接表。5. **构建链接**(`build_embedding_linkage.py`)：计算本体术语和 OpenAlex 主题的 BGE-large 嵌入，构建 FAISS 索引，并生成本体对齐表 `xref.topic_ontology_map`。

其他脚本包括 `convert_ontologies.py`（用于 13 个本体的五个特定格式解析器）和 `ontology_registry.py`（本体 URL 与格式的声明式注册表）。该代码库还包含结构化模式参考（`SCHEMA.md`），记录全部 153 个视图的列、类型、行数和跨数据集连接策略，旨在供基于大语言模型的编程智能体和人类开发者使用。

致谢

作者感谢汉堡工业大学（TUHH）提供机构支持。计算资源由 TUHH 高性能计算集群提供。作者衷心感谢 OpenAlex 团队通过 Amazon Web Services S3 提供开放、可免费访问的数据快照，也感谢 Semantic Scholar、SciSciNet、Papers with Code、Retraction Watch、Reliance on Science 以及 PreprintToPaper 项目背后的团队公开提供其数据集。

作者贡献

J.W.：概念构思、方法设计、软件、数据整理、验证、形式分析、调查、可视化、撰写初稿、审阅与编辑。

利益冲突

作者声明不存在利益冲突。

参考文献

[1] Lu Liu、Benjamin F. Jones、Brian Uzzi 和 Dashun Wang。科学学中的数据、测量与实证方法。*Nature Human Behaviour*，7:1046–1058，2023。doi: 10.1038/s41562-023-01562-4。[2] Rodney Kinney、Chloe Anastasiades、Russell Authur、Iz Beltagy、Jonathan Bragg、Alexandra Buraczynski、Isabel Cachola、Stefan Candra、Yoganand Chandrasekhar、Arman Cohan 等。Semantic Scholar 开放数据平台。*arXiv* 预印本 *arXiv:2301.10140*，2023。[3] Jason Priem、Heather Piwowar 和 Richard Orr。OpenAlex：学术作品、作者、出版场所、机构和概念的完全开放索引。*arXiv* 预印本 *arXiv:2205.01833*，2022。[4] Zihang Lin、Yian Yin、刘璐和 Dashun Wang。SciSciNet：面向科学学研究的大规模开放数据湖。*Scientific Data*，10:315，2023。doi: 10.1038/ s41597-023-02198-9。[5] Papers with Code。Papers with Code：面向机器学习的免费开放资源。https://paperswithcode.com，2024。访问于 2024 年 10 月。[6] 科学诚信中心。撤稿观察数据库。 https://retractionwatch.com，2024。访问于 2024 年 9 月。[7] Matt Marx 和 Aaron Fuegi。对科学的依赖：全球专利首页对科学文章的引文。*Strategic Management Journal*，41(9):1572–1594，2020。doi: 10.1002/smj. 3145。

[8] Michael Gusenbauer。超越 Google Scholar、Scopus 和 Web of Science：对 59 个数据库引文索引的后向与前向引文覆盖率进行评估。*ResearchSynthesis Methods*，15(5):802–817，2024。doi: 10.1002/jrsm.1729。[9] Martijn Visser、Nees Jan van Eck 和 Ludo Waltman。文献计量数据源的大规模比较：Scopus、Web of Science、Dimensions、Crossref 和 Microsoft Academic。*Quantitative Science Studies*，2(1):20–41，2021。doi: 10.1162/qss_a_00112。[10] Kian Ahrabian、Xinwei Du、Richard Delwin Myloth、Arun Baalaaji Sankar Ananthan 和 Jay Pujara。PubGraph：大规模科学知识图谱。*arXiv* 预印本 *arXiv:2302.02231*，2023。[11] Michael Färber、David Lamprecht、Johan Krause、Linn Aung 和 Peter Haase。SemOpenAlex：260 亿个 RDF 三元组中的科学图景。载于语义网 – *ISWC 2023*，第 94–112 页。Springer，2023。doi: 10.1007/978-3-031-47243-5_6。[12] Daniel W. Hook、Simon J. Porter 和 Christian Herzog。Dimensions：构建搜索与评估的背景信息。*Frontiers in Research Metrics and Analytics*，3:23，2018。doi: 10.3389/frma.2018.00023。[13] Mohamad Yaser Jaradeh、Allard Oelen、Manuel Prinz、Markus Stocker 和 Sören Auer。开放研究知识图谱：系统演示。载于开放知识数字图书馆 – *TPDL 2019*，第 11799 卷计算机科学讲义，第 348–351 页。Springer，2019。doi: 10.1007/978-3-030-30760-8_31。[14] Shitao Xiao、Zheng Liu、Peitian Zhang、Niklas Muennighoff、Defu Lian 和 Jian-Yun Nie。C-pack：通用中文嵌入的资源集合。载于第 *47* 届 *ACM SIGIR* 信息检索研究与发展大会论文集，第 2436–2446 页，2024。doi: 10.1145/3626772.3657878。[15] Hanchen Wang、Tianfan Fu、Yuanqi Du、Wenhao Gao、Kexin Huang、Ziming Liu、Payal Chandak、Shengchao Liu、Peter Van Katwyk、Andreea Deac 等。人工智能时代的科学发现。*Nature*，620:47–60，2023。doi: 10.1038/s41586-023-06221-2。[16] Russell J. Funk 和 Jason Owen-Smith。技术变革的动态网络指标。*Management Science*，63(3):791–817，2017。doi: 10.1287/mnsc.2015.2366。[17] Lingfei Wu、Dashun Wang 和 James A. Evans。大团队推动科学与技术发展，小团队则带来颠覆。*Nature*，566:378–382，2019。doi: 10.1038/s41586-019-0941-9。[18] Brian Uzzi、Satyam Mukherjee、Michael Stringer 和 Ben Jones。非典型组合与科学影响力。*Science*，342(6157):468–472，2013。doi: 10.1126/science.1240474。[19] Mark Raasveldt 和 Hannes Mühleisen。DuckDB：可嵌入式分析型数据库。载于*ACM SIGMOD* 数据管理国际会议论文集，第 1981–1984 页，2019。doi: 10.1145/3299869.3320212。[20] Carolyn E. Lipscomb。医学主题词表（MeSH）。*Bulletin of the Medical LibraryAssociation*，88(3):265–266，2000。[21] Janna Hastings、Gareth Owen、Adriano Dekker、Marcus Ennis、Namrata Kale、Venkatesh Muthukrishnan、Steve Turner、Neil Swainston 和 Christoph Steinbeck。2016 年的 ChEBI：服务改进与代谢物集合扩展。*Nucleic Acids Research*，44

(D1):D1214–D1219，2016。doi：10.1093 /nar/gkv1031.

[22] Nicholas Sioutos、Sherri de Coronado、Margaret W. Haber、Frank W. Hartel、Wen-Ling Shaiu 和 Lawrence W. Wright。NCI 同义词库：整合癌症相关临床与分子信息的语义模型。《生物医学信息学杂志》，40(1):30–43，2007年。doi: 10.1016/j.jbi.2006.02.013。[23] 基因本体联盟。基因本体资源：丰富一座“金矿”。*NucleicAcids Research*，49(D1):D325–D334，2021年。doi: 10.1093/nar/gkaa1113。[24] Caterina Caracciolo、Armando Stellato、Ahsan Morshed、Gudrun Ferreri、Aimee Johannsen、Sachit Rajbhandari、Yves Jaques、Johannes Keizer 和 Rajan Shankar。AGROVOC 链接数据集。语义网，4(3):341–348，2013。doi: 10.3233/SW-130106。[25] Angelo A. Salatino、Thiviyan Thanapalasingam、Andrea Mannocci、Aliaksandr Birukou、Francesco Osborne 和 Enrico Motta。计算机科学本体：一个全面的自动生成研究领域本体。《数据智能》，2(3):379–416，2020年。doi: 10.1162/dint_a_00055。[26] Lynn M. Schriml、James B. Munro、Mike Schor、Dustin Olley、Cynthia McCracken、Victor Felix、J. Allen Baron、Rebecca Jackson、Susan M. Bello、Claudia Bearer 等。人类疾病本体 2022 年更新。*Nucleic Acids Research*，50(D1):D1255–D1261，2022年。doi: 10.1093/nar/gkab1063。[27] Sebastian Köhler、Michael Gargano、Nicolas Matentzoglu、Leigh C. Carmody、David Lewis- Smith、Nicole A. Vasilevsky、Daniel Danis、Ganna Balagura、Gareth Baynam、Amy M. Brower 等。2021 年的人类表型本体。*Nucleic Acids Research*，49(D1): D1207–D1217，2021年。doi: 10.1093/nar/gkaa1043。[28] Jon Ison、Matúš Kalaš、Inge Jonassen、Dan Bolser、Mahmut Uludag、Hamish McWilliam、James Malone、Rodrigo Lopez、Steve Pettifer 和 Peter Rice。EDAM：生物信息学操作、数据类型、标识符、主题和格式的本体。《生物信息学》，29(10):1325–1332，2013。doi: 10.1093/bioinformatics/btt113。[29] Jeff Johnson、Matthijs Douze 和 Hervé Jégou。利用图形处理器进行十亿级相似性搜索。《*IEEE*大数据汇刊》，7(3):535–547，2021年。doi: 10.1109/TBDATA.2019.2921572。[30] Stephen Robertson 和 Hugo Zaragoza。概率相关性框架：BM25 及其后续发展。《信息检索基础与趋势》，3(4):333–389，2009年。doi: 10.1561/1500000019。[31] Leland McInnes、John Healy 和 James Melville。UMAP：用于降维的均匀流形近似与投影。*arXiv* 预印本 *arXiv:1802.03426*，2018。