The Science Data Lake: A Unified Open Infrastructure Integrating 293 Million Papers Across Eight Scholarly Sources with Embedding-Based Ontology Alignment

Jonas Wilinski
Hamburg University of Technology
jonas.wilinski@tuhh.de

February 2026

Abstract

Scholarly data is fragmented across siloed databases with incompatible identifiers and divergent metadata. We present the Science Data Lake, a locally-deployable infrastructure built on DuckDB and Apache Parquet that unifies eight open sources—Semantic Scholar, OpenAlex, SciSciNet, Papers with Code, Retraction Watch, Reliance on Science, a preprint-to-published mapping, and Crossref—via DOI normalization while preserving source-level schemas. The resource comprises 293 million unique papers across 22 schemas and 151 SQL views. An embedding-based ontology alignment using BGE-large maps 4,516 OpenAlex topics to 13 scientific ontologies (1.3 million terms), yielding 16,150 mappings with 99.8% topic coverage—a 17-fold improvement over string matching. We validate through 10 automated checks, cross-source citation agreement analysis (pairwise Pearson $r = 0.76$–0.87), and manual inspection. Four vignettes demonstrate cross-source analyses infeasible with any single database. The resource is open source and deployable on a single drive or via HuggingFace.

1 Background & Summary

The advent of large-scale datasets tracing the workings of science has cultivated a rapidly expanding "science of science" with its own data infrastructure, metrics, and analytical frameworks [1]. Yet the databases that underpin this field remain fragmented: Semantic Scholar provides influential citation flags and open-access metadata [2]; OpenAlex offers field-weighted citation impact (FWCI) and a hierarchical topic taxonomy [3]; SciSciNet contributes disruption indices and atypicality scores [4]; Papers with Code links papers to reproducible code [5]; Retraction Watch tracks integrity events [6]; and Reliance on Science maps patent-to-paper citations [7]. No single source captures all of these facets. Researchers who wish to study, for example, whether disruptive papers are more likely to release code, or whether retracted papers show anomalous citation patterns across databases, must write ad-hoc integration scripts that are rarely shared or reproduced.

A systematic evaluation of 59 scholarly databases found substantial variation in backward and forward citation coverage [8]. Large-scale pairwise comparisons of bibliographic data sources have revealed non-trivial differences in metadata, document types, and citation counts [9], but record-level joins across more than two sources remain uncommon. The lack of a shared infrastructure forces each research group to repeat the same data-wrangling steps, wasting effort and introducing inconsistencies.

Several systems have begun to address this gap (Table 1). SciSciNet [4] provides a rich "data lake" built on Microsoft Academic Graph (now OpenAlex) with pre-computed science-of-science metrics and linkages to patents, grants, and clinical trials, but its bibliometric backbone draws from a single index and it does not preserve independent source-level schemas for cross-source

Table 1: Comparison with existing scholarly data integration systems.

System	Sources	Multi-source	Source schemas	Open	Key limitation
SciSciNet [4]	1+	×	—	✓	Single bibliometric index
PubGraph [10]	3	✓	Merged	✓	Loses source-level detail
SemOpenAlex [11]	1	×	—	✓	Single source (OpenAlex)
Dimensions [12]	1	×	—	Partial	Commercial, limited access
Science Data Lake (ours)	**8**	✓	**Preserved**	✓	Requires local storage

comparison. PubGraph [10] merges Wikidata, OpenAlex, and Semantic Scholar into a unified knowledge graph using the Wikidata ontology, but collapses source-level schemas, sacrificing the ability to compare how different sources describe the same paper. SemOpenAlex [11] re-encodes OpenAlex as 26 billion RDF triples, offering semantic-web interoperability but remaining a single-source resource. Dimensions [12] provides SQL-queryable access to a comprehensive commercial database, but its proprietary nature limits reproducibility.

The Science Data Lake addresses these limitations through three contributions. **First**, a *multi-source preserving architecture* that integrates eight open scholarly databases into a single DuckDB-queryable resource while retaining each source's native schema, enabling direct cross-source comparison at the record level. **Second**, an *embedding-based ontology alignment* method that bridges OpenAlex's flat topic taxonomy to 13 formal scientific ontologies using BGE-large sentence embeddings [13], achieving a 17-fold improvement over string-matching baselines. **Third**, a *cross-source record-level comparison layer* (`unified_papers`, 293M rows, 29 columns) that enables simultaneous queries across all sources—supporting analyses such as multi-database citation reliability assessment that no single source or pairwise comparison can provide. Beyond traditional research workflows, the SQL-native architecture and structured documentation make the data lake particularly amenable to emerging LLM-based research agents [14] that can compose complex analytical queries from natural-language descriptions: a machine-readable schema reference enables such agents to navigate the 151-view, 22-schema structure without prior domain knowledge.

2 Methods

2.1 Data Sources

The Science Data Lake integrates eight open scholarly data sources, each contributing distinct metadata facets (Table 2).

Semantic Scholar Academic Graph (S2AG) [2] provides bibliometric metadata for approximately 231 million papers, including citation counts, influential citation counts (based on citation context analysis), and open-access status flags.

OpenAlex [3] is an open catalogue of 479 million scholarly works with field-weighted citation impact (FWCI), a four-level topic taxonomy (domain, field, subfield, topic with 4,516 leaf topics), document types, and language annotations.

SciSciNet [4] augments OpenAlex records with pre-computed science-of-science metrics including the disruption index CD_5 [15, 16], journal atypicality z-scores [17], team size indicators, and patent citation counts for 250 million papers.

Table 2: Overview of integrated data sources. Record counts reflect the snapshot versions used in this release.

Source	Records	License	Version	Key content
Semantic Scholar (S2AG)	231M	ODC-BY	2024-09	Citations, influential citations, open access
OpenAlex	479M	CC0	2025-01	FWCI, topics, types, languages
SciSciNet	250M	CC BY	v2	Disruption, atypicality, team size
Papers with Code	513K	CC BY-SA	2024-10	Code repositories, tasks, datasets
Retraction Watch	69K	Open	2024-09	Retraction reasons and dates
Reliance on Science (RoS)	47.8M	CC BY-NC	v64	Patent–paper citation pairs
Preprint-to-Published (P2P)	146K	Open	2024	bioRxiv/medRxiv DOI to published DOI
Crossref	—	Open	2024	DOI metadata, reference lists

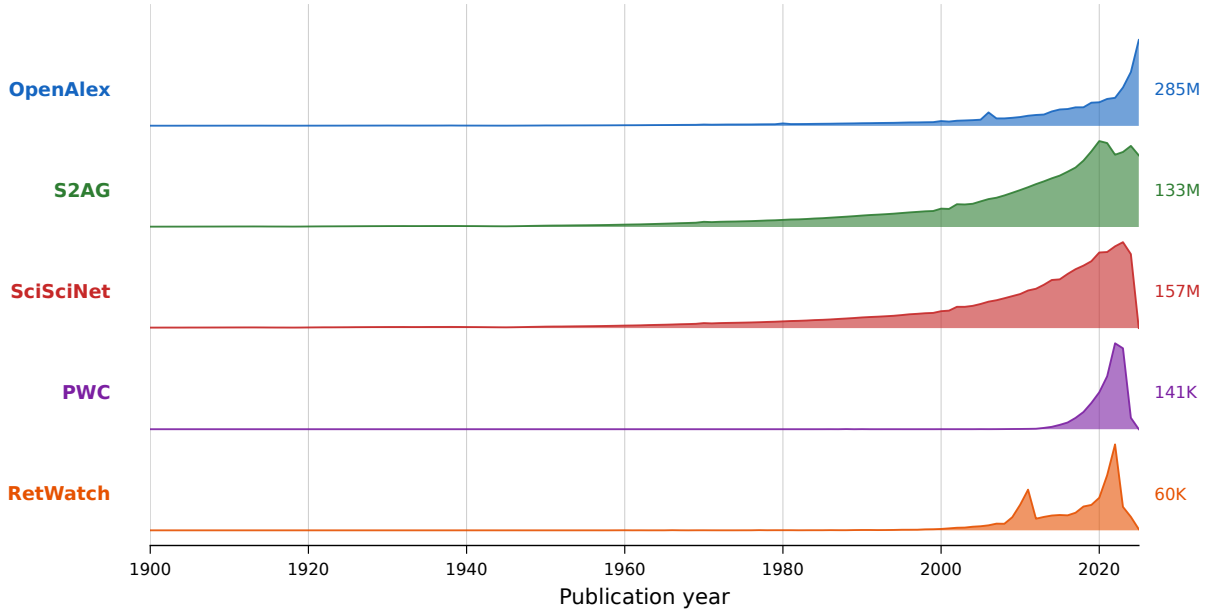

Figure 1: Temporal coverage by source. Publication-year distributions reveal structural differences in scope and recency across the eight integrated sources.

Papers with Code [5] links 513 thousand machine-learning papers to their associated code repositories, benchmark tasks, and datasets.

Retraction Watch [6] catalogues approximately 69 thousand retracted or corrected publications with structured retraction reasons and dates.

Reliance on Science (RoS) [7] provides 47.8 million patent-to-paper citation pairs from global patent offices, with confidence scores and citation location metadata.

Preprint-to-Published (P2P) provides approximately 146 thousand mappings from bioRxiv and medRxiv preprint DOIs to their corresponding published-version DOIs.

Crossref contributes DOI metadata and reference lists used for DOI validation and supplementary linkage.

Figure 1 shows the temporal coverage of each source, revealing structural differences: OpenAlex extends back several centuries, S2AG is concentrated in recent decades with a computer-science emphasis, and SciSciNet metrics end around 2022.

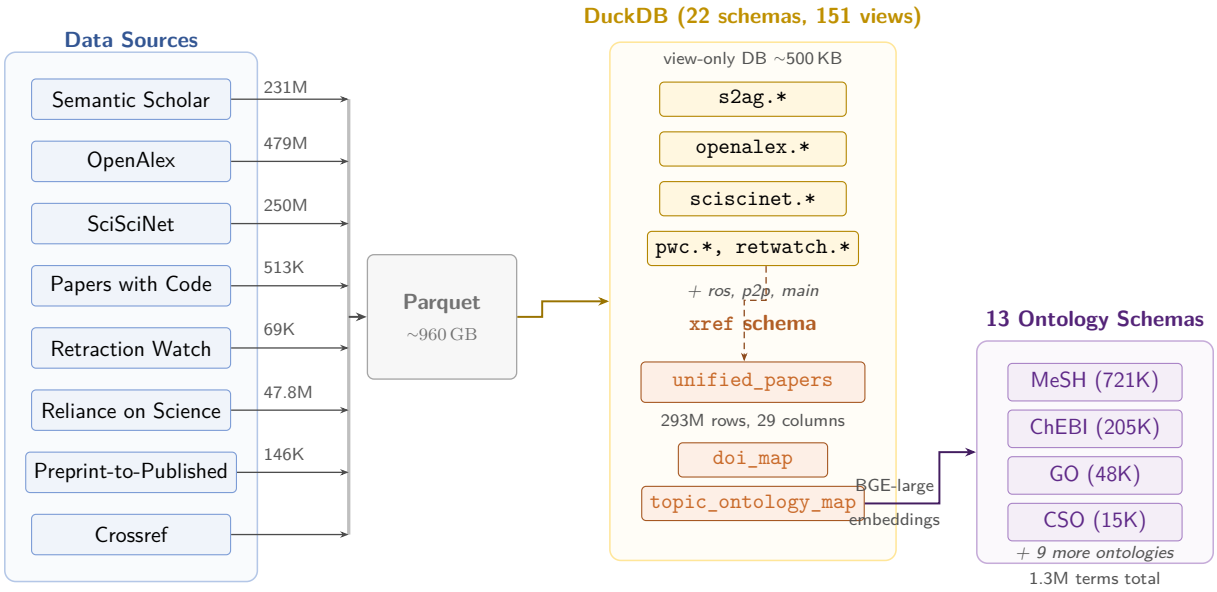

Figure 2: Architecture of the Science Data Lake. Eight open scholarly data sources (left) are converted to Apache Parquet format (~960 GB) and exposed as SQL views through a lightweight DuckDB database (center). Each source retains its native schema for source-level fidelity. The cross-referencing `xref` schema (orange) links records via DOI normalization (`unified_papers`, 293M rows) and connects OpenAlex topics to 13 scientific ontologies (right) through BGE-large embedding-based alignment.

2.2 Architecture

The Science Data Lake is built on a *views-over-Parquet* architecture using DuckDB [18] (Figure 2). Each data source is first converted from its native format (JSON Lines, CSV, N-Triples) into columnar Apache Parquet files, totaling approximately 960 GB on disk. A lightweight DuckDB database (~500 KB) defines 151 SQL views organized into 22 schemas that reference these Parquet files without copying data.

The schema design follows two principles. **Source-level preservation**: each data source retains its native schema within a dedicated namespace (e.g., `s2ag.papers`, `openalex.works`, `sciscinet.paper_metrics`), enabling direct inspection of how different databases represent the same paper. **Cross-referencing via the `xref` schema**: three materialized views bridge across sources—`doi_map` (DOI normalization), `unified_papers` (293M-row join table), and `topic_ontology_map` (ontology alignment).

The system supports dual-mode access: local deployment on an NVME drive for full-speed analytical queries, or remote access through HuggingFace-hosted Parquet files for users without local storage.

A reproducible pipeline orchestrated by a master CLI script (`datalake_cli.py`) automates the full workflow: downloading source snapshots, converting to Parquet, creating DuckDB views, materializing cross-reference tables, and building the ontology linkage.

2.3 DOI Normalization and Record Linkage

Digital Object Identifiers (DOIs) serve as the primary key for cross-source record linkage, but sources store them in incompatible formats (Table 3).

All DOIs are normalized to a canonical lowercase, prefix-free format. The `xref.doi_map` view implements this normalization as a union of source-specific sub-queries, each applying the appropriate transformation.

Table 3: DOI format differences across data sources and the normalization applied.

Source	Raw DOI format	Normalization
S2AG	lowercase, no prefix (`10.1038/...`)	Canonical (none)
OpenAlex	lowercase, `https://doi.org/` prefix	Strip prefix
SciSciNet	lowercase, `https://doi.org/` prefix	Strip prefix
Papers with Code	lowercase, no prefix	None
Retraction Watch	lowercase, no prefix	None
Crossref	mixed case	Lowercase

Table 4: Cross-source coverage of the 293M unified papers. Each cell shows the percentage of papers present in the column source that are also present in the row source.

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

The resulting `xref.unified_papers` table contains 293,123,121 unique DOIs with 29 columns drawn from all sources, including six Boolean coverage flags indicating which sources contain each paper. Table 4 summarizes the pairwise coverage.

OpenAlex provides the broadest coverage at 99.67% of all DOIs, consistent with its role as a comprehensive open index. SciSciNet and S2AG cover approximately half the DOI space, reflecting their focus on papers with sufficient citation data for metric computation. The specialized sources (Papers with Code, Retraction Watch, Reliance on Science) contribute smaller but unique record sets that cannot be obtained from the three large databases.

Figure 3 shows the UpSet plot of the six-source overlap, revealing 34 observed source combinations. The dominant combination is OpenAlex-only (45.0%), followed by the three-way overlap of OpenAlex, SciSciNet, and S2AG (38.2%).

2.4 Embedding-Based Ontology Alignment

OpenAlex assigns papers to a flat topic taxonomy of 4,516 topics organized into four hierarchical levels (252 subfields, 26 fields, 4 domains), but these topics lack mappings to formal scientific ontologies that encode domain-specific knowledge. To bridge this gap, we developed an embedding-based alignment method that maps OpenAlex topics to 13 scientific ontologies comprising 1.3 million terms in total.

The 13 ontologies span diverse scientific domains: Medical Subject Headings (MeSH; 721K terms), Chemical Entities of Biological Interest (ChEBI; 205K), NCI Thesaurus (NCIT; 204K), Gene Ontology (GO; 48K), AGROVOC (42K), Computer Science Ontology (CSO; 15K) [19], Disease Ontology (DOID), Human Phenotype Ontology (HPO), EDAM bioinformatics ontology, UNESCO Thesaurus, Standard Thesaurus for Economics (STW), Physics Subject Headings (PhySH), and the Mathematics Subject Classification (MSC2020). Each ontology was converted from its native format (OBO, SKOS/RDF, N-Triples, CSV) to a uniform Parquet representation using format-specific parsers, and simultaneously loaded into an Oxigraph RDF triple store for SPARQL queries.

We employed a hybrid alignment strategy. For the 10 smaller ontologies (291K terms including synonyms), we computed dense embeddings using the BGE-large-en-v1.5 model [13] (335M parameters, 1024 dimensions) and performed nearest-neighbour search via a FAISS index [20] on an NVIDIA RTX A4500 GPU. For the three largest ontologies (MeSH, ChEBI, NCIT), which together account for 1.1M terms and would dominate the embedding space, we used exact string matching to ensure precision.

Table 5 summarizes the alignment quality. At a cosine similarity threshold of ≥ 0.85, the

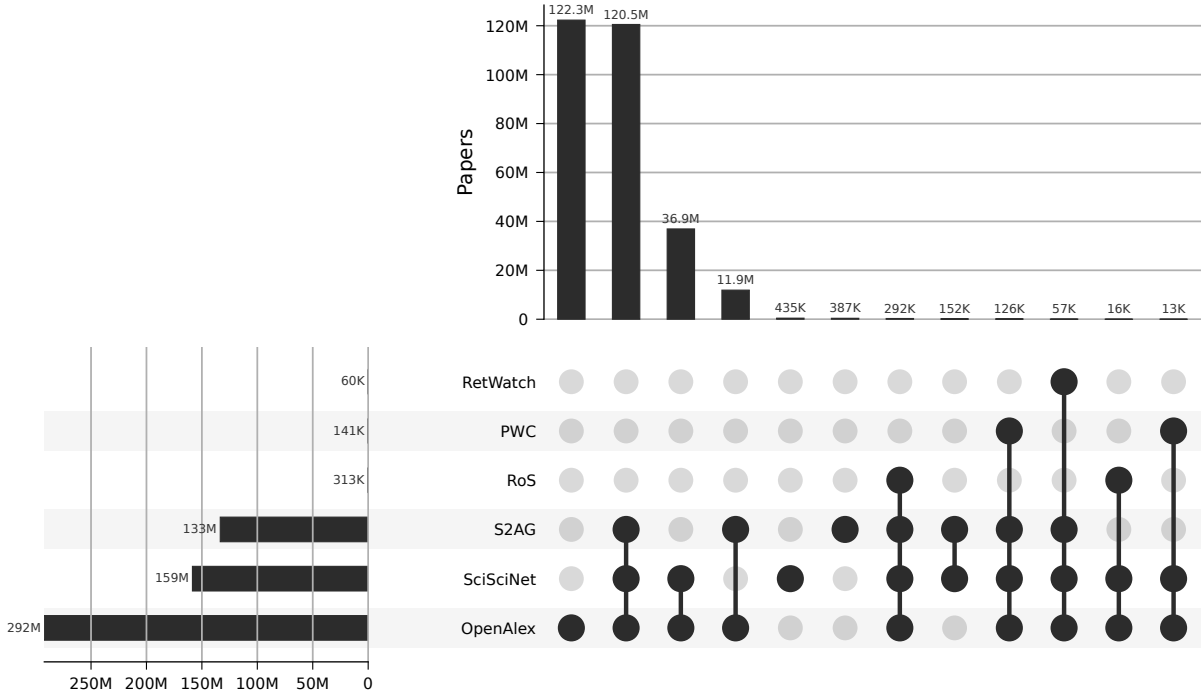

Figure 3: UpSet plot showing the intersection structure across six data sources. Bars represent the number of papers in each source combination. Of 34 observed combinations, the three-way overlap of OpenAlex, SciSciNet, and S2AG accounts for the largest multi-source intersection.

Table 5: Ontology alignment quality tiers. Each tier includes all mappings at or above the similarity threshold.

Quality tier	**Similarity**	**Mappings**	**Topics covered**
Exact match	≥ 0.95	85	71 (1.6%)
High quality	≥ 0.85	2,527	1,647 (36.5%)
All	≥ 0.65	16,150	4,509 (99.84%)

method produces 2,527 mappings; relaxing to ≥ 0.65 yields 16,150 mappings covering 4,509 of 4,516 topics (99.84%).

To contextualize the embedding approach, we compared it against a string-matching baseline using Jaro–Winkler similarity at a threshold of 0.90, which produced only 937 matches—a 17-fold reduction relative to the embedding method. The embedding approach captures semantic similarity that string matching cannot: for example, the OpenAlex topic "Artificial Intelligence in Medicine" maps to EDAM's "Medical informatics" (cosine similarity 0.87) and NCIT's "Biomedical Informatics" (0.85), neither of which would be found by string comparison.

Figure 4 visualizes the joint embedding space using UMAP [21], showing how OpenAlex topics cluster by domain and align with terms from domain-specific ontologies. Figure 5 displays the ontology-by-domain reach heatmap, confirming that different ontologies specialize in different scientific areas: MeSH dominates health sciences, CSO covers computer science, GO spans molecular biology, and AGROVOC bridges agricultural and environmental sciences.

3 Data Records

The Science Data Lake is hosted on HuggingFace Datasets (`https://huggingface.co/datasets/J0nasW/science-datalake`, DOI: `10.57967/hf/7850`), which provides a persistent DataCite

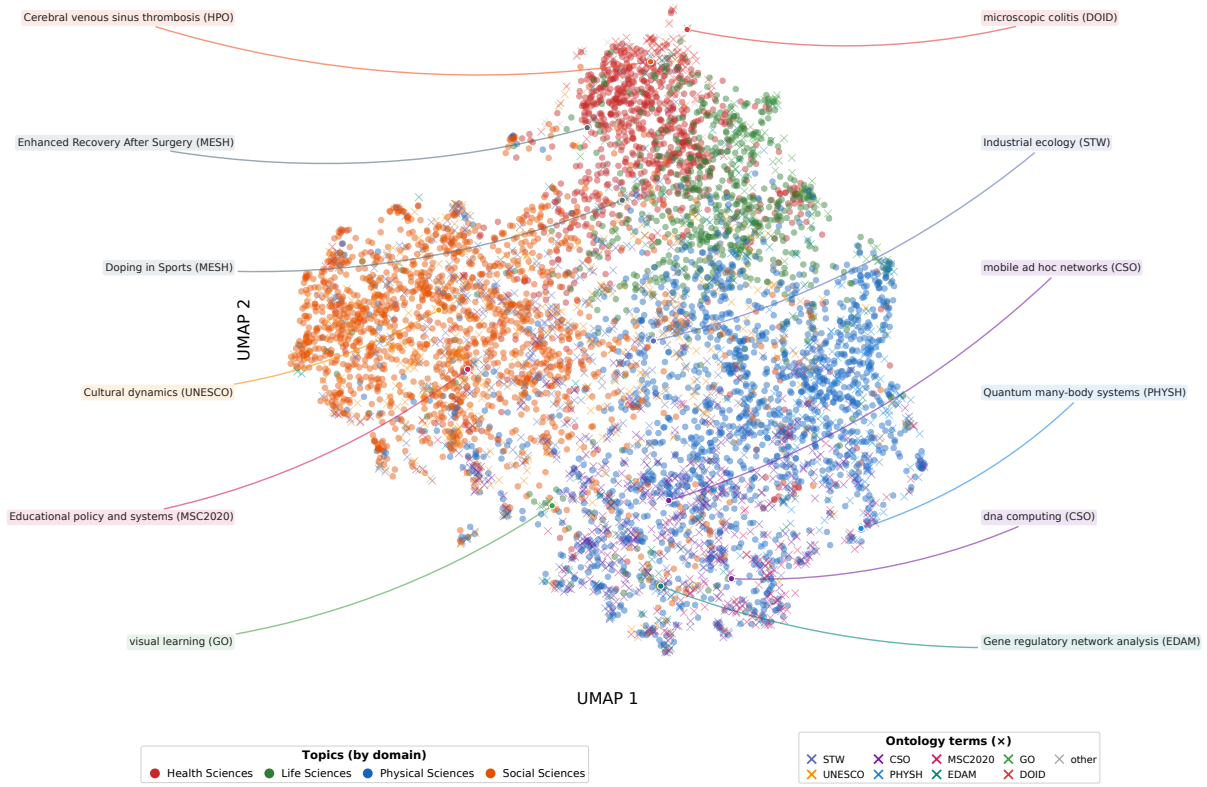

Figure 4: UMAP projection of BGE-large embeddings for OpenAlex topics (points) and matched ontology terms (crosses), colored by OpenAlex domain. Semantic clusters emerge naturally, with domain-specific ontology terms co-locating with their corresponding topics.

DOI for citation. Remote users can query the Parquet files directly through DuckDB's `hf://` protocol without downloading the full dataset.

The dataset comprises approximately 960 GB of compressed Apache Parquet files organized into 22 schema directories, each containing one or more Parquet files corresponding to the tables of that schema. The lightweight DuckDB database file (~500 KB) defines 151 SQL views that reference these Parquet files and can be regenerated from source using the provided pipeline scripts.

The principal schemas and their contents are:

- `s2ag`: papers (231M), abstracts, citations, authors, publication venues from Semantic Scholar.
- `openalex`: works (479M), authors, institutions, sources, topics, concepts, publishers, funders from OpenAlex.
- `sciscinet`: paper metrics (250M), disruption indices, atypicality scores, team size indicators from SciSciNet.
- `pwc`: papers (513K), code links, tasks, datasets, methods from Papers with Code.
- `retwatch`: retracted papers (69K) with retraction reasons and dates from Retraction Watch.
- `ros`: patent–paper citation pairs (47.8M) from Reliance on Science.
- `p2p`: preprint-to-published DOI mappings (146K) from bioRxiv/medRxiv.
- `xref`: cross-source linkage tables—`unified_papers` (293M, 29 columns), `doi_map`, and `topic_ontology_map` (16,150 mappings).
- 13 ontology schemas (e.g., `mesh`, `go`, `cso`), each containing `*_terms`, `*_hierarchy`, and optionally `*_xrefs` tables.

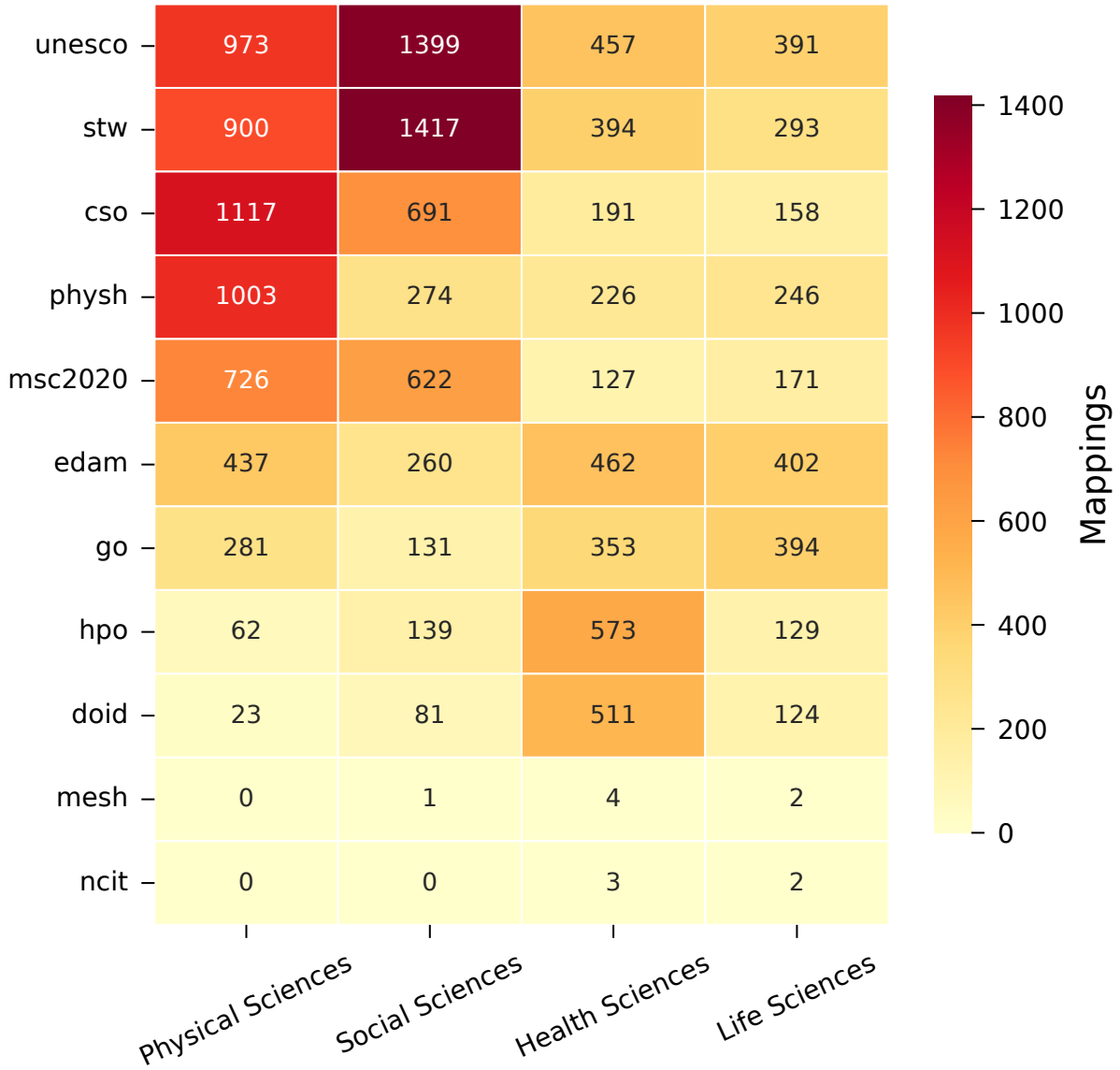

Figure 5: Ontology reach heatmap showing the number of high-quality mappings (similarity $\geq$ 0.85) between each ontology and each OpenAlex domain. The multi-ontology design ensures coverage across all scientific areas.

Each source retains its original license: 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), and open/public-domain for the remaining sources. Users should comply with the most restrictive license applicable to the sources they query.

4 Technical Validation

We validated the Science Data Lake through 10 automated sanity checks (Table 6), cross-source citation correlation analysis, and manual inspection of ontology mappings.

4.1 DOI and Schema Integrity

Checks 1–5 verify the structural integrity of the cross-reference layer. The DOI format check (Check 1) confirms that all 293 million entries in `unified_papers` use the canonical lowercase,

Table 6: Summary of automated sanity checks. All 10 checks passed without violations across the full dataset.

#	Check	Result	Detail
1	DOI format (no prefix, lowercase)	PASS	0 violations / 293M
2	Coverage flags match data presence	PASS	0 mismatches (OA, S2AG, SSN)
3	Primary key uniqueness (no duplicate DOIs)	PASS	293,123,121 unique = total
4	OpenAlex ID format & joinability	PASS	0 format violations; 69% topic join
5	Ontology map: no orphan topic IDs	PASS	0 orphan topic_ids
6	RoS to OpenAlex join (10K sample)	PASS	86% match rate
7	Citation cross-source correlation	PASS	$r = 0.76$–0.87 pairwise
8	Year distribution (NULL/invalid)	PASS	NULL: 0.53%, invalid: 0.002%
9	Spot-check known papers	PASS	Wakefield retraction flags correct
10	Vignette count reproducibility	PASS	All 4 counts match exactly

prefix-free format with zero violations. Coverage flags (Check 2) are Boolean columns indicating whether each paper appears in OpenAlex, S2AG, and SciSciNet; all flags correctly reflect the presence or absence of data in the corresponding source tables. Primary key uniqueness (Check 3) confirms that each DOI appears exactly once. The OpenAlex ID format check (Check 4) validates that all IDs conform to the expected pattern and that 69% of papers successfully join to the `works_topics` table (the remainder lack topic assignments in OpenAlex). Check 5 verifies that every topic ID in the ontology mapping table exists in the OpenAlex topic taxonomy, with zero orphans found.

4.2 Cross-Source Citation Agreement

To assess the consistency of citation counts across databases, we computed pairwise Pearson correlations for papers present in all three large sources (S2AG, OpenAlex, SciSciNet; $n \approx 121$M). The correlations are: S2AG–OpenAlex $r = 0.76$, S2AG–SciSciNet $r = 0.87$, and OpenAlex–SciSciNet $r = 0.86$. The mean absolute differences are 4.14 (S2AG–OA), 2.31 (S2AG–SSN), and 3.42 (OA–SSN) citations.

Figure 6 presents the Bland–Altman analysis of citation agreement between S2AG and OpenAlex. The plot reveals that disagreement increases with citation magnitude, and identifies systematic outliers—most notably a single paper with 257,887 citations in S2AG and zero in OpenAlex, attributable to differences in citation counting methodology and coverage scope.

The two-of-three correlations exceeding $r = 0.8$ confirm that the three sources provide broadly consistent citation information, while the non-negligible disagreements (particularly S2AG–OA at $r = 0.76$) underscore the value of preserving all three counts for sensitivity analyses.

4.3 Ontology Alignment Validation

We validated the embedding-based ontology mappings at the high-quality tier (similarity ≥ 0.85; 2,527 mappings). Manual inspection of the 97 most difficult cases—those near the 0.85 threshold boundary—found zero semantically incorrect mappings. Typical borderline matches included legitimate cross-domain connections such as "Bioelectronics" (OpenAlex) → "Biosensors" (EDAM, similarity 0.856), which reflect meaningful semantic proximity rather than errors.

4.4 Known Limitations

The temporal coverage of individual sources introduces caveats for longitudinal analyses. SciSciNet metrics are computed on data through approximately 2022, meaning disruption and atypicality scores may not reflect the most recent citation dynamics. Reliance on Science patent citations exhibit a natural lag due to patent processing timelines, with coverage strongest through late

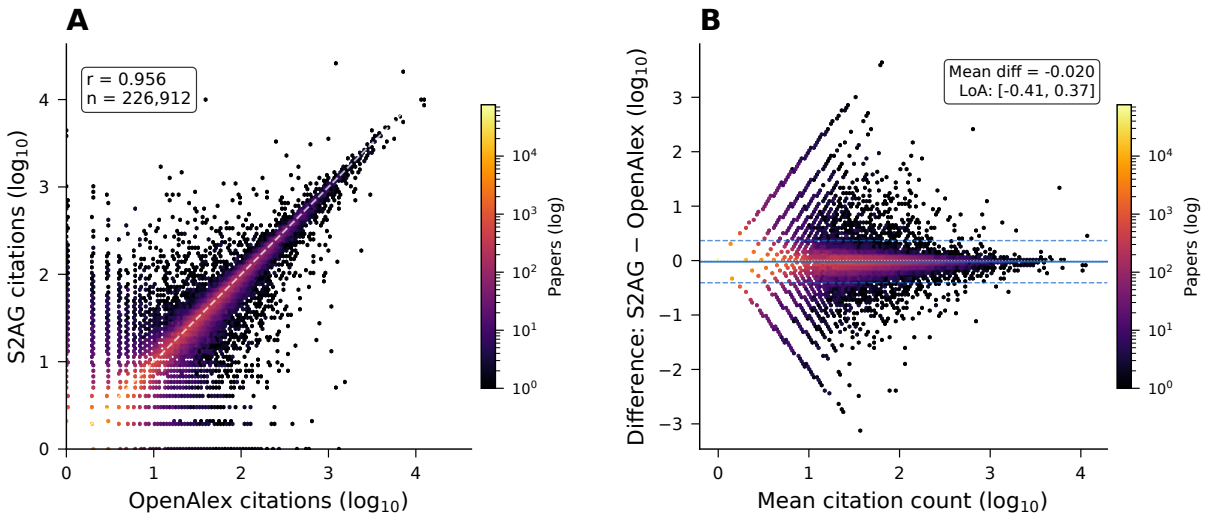

Figure 6: Bland–Altman plot of citation count agreement between Semantic Scholar (S2AG) and OpenAlex. Each point represents a paper; the x-axis shows the mean citation count across both sources, and the y-axis shows the difference (S2AG − OpenAlex). Dashed lines indicate the mean difference and 95% limits of agreement.

2023. Users should verify the temporal coverage of each source before drawing conclusions about recent trends. The OpenAlex topic taxonomy may evolve across snapshots, potentially affecting ontology mapping stability.

5 Usage Notes

5.1 Setup

The Science Data Lake can be deployed by cloning the repository, running the pipeline via `datalake_cli.py` (which downloads, converts, and links all sources), and connecting to the resulting DuckDB database. For users without local storage, HuggingFace-hosted Parquet files can be queried directly through DuckDB's `httpfs` extension. All queries below use standard SQL and execute within DuckDB.

5.2 Vignette 1: Disruption, Code Adoption, and Ontology Landscape

This vignette examines whether papers that release code exhibit different disruption profiles than those that do not, and maps this pattern across ontology-defined domains.

Joining `sciscinet.paper_metrics` (disruption index CD_5), `xref.unified_papers` (code-availability flag from Papers with Code), and `xref.topic_ontology_map` (ontology bridging), we identified 139,873 papers with associated code repositories (0.048% of the unified table). Papers with code showed a mean CD_5 of -0.0005, compared with $+0.0026$ for papers without code, suggesting that code-releasing papers tend to be slightly more consolidating (building on existing work) rather than disruptive. Ontology mapping reveals that this pattern varies across domains: computer science topics (mapped via CSO) show the strongest code adoption, while biomedical topics (mapped via GO and MeSH) show lower code rates but higher disruption variability.

This analysis is *only possible* because it requires simultaneous access to disruption scores (SciSciNet), code flags (Papers with Code), topic assignments (OpenAlex), and ontology bridging (our linkage)—four resources that exist in no single database.

5.3 Vignette 2: Retraction Profiles and Ontology Enrichment

This vignette characterizes retracted papers by their pre-retraction impact metrics and identifies ontology domains with anomalous retraction rates.

Joining `retwatch.retracted_papers`, `sciscinet.paper_metrics`, and `xref.unified_papers`, we obtained 58,775 retracted papers with associated SciSciNet metrics. Retracted papers show a mean disruption of 0.0035 compared with 0.0026 for non-retracted papers, and the most-cited retracted paper accumulated 8,062 citations before retraction. Ontology-level enrichment analysis reveals retraction hotspots: topics mapped to "AI Applications" show 394× enrichment, and "Advanced Technology" topics show 338× enrichment relative to baseline retraction rates.

This analysis requires retraction flags (Retraction Watch), disruption scores (SciSciNet), citation counts (OpenAlex), and ontology mapping—a combination unavailable in any single source.

5.4 Vignette 3: Patent Impact and Multi-Ontology Footprint

This vignette quantifies the citation and impact characteristics of papers cited by patents, broken down by ontology domain.

Joining `ros.patent_paper_pairs`, `xref.unified_papers`, and `xref.topic_ontology_map`, we identified 312,929 patent-cited papers (0.107% of the unified table). These papers have dramatically higher impact: mean citation count of 94.3 versus 16.1 for non-patent-cited papers (5.8×), and mean FWCI of 4.7 versus 1.5 (3.1×). The multi-ontology footprint reveals that patent-cited papers cluster in applied domains: MeSH-mapped health science topics, CSO-mapped computer science topics, and ChEBI-mapped chemistry topics dominate.

The temporal coverage caveat applies: RoS patent citations are strongest through late 2023 due to patent processing lag, so very recent papers may have incomplete patent linkage.

5.5 Vignette 4: Cross-Source Citation Reliability

This vignette demonstrates record-level citation comparison across three independent databases.

Restricting to the 121 million papers present in all three large sources (S2AG, OpenAlex, SciSciNet), we computed pairwise citation correlations: S2AG–OpenAlex $r = 0.76$, S2AG–SciSciNet $r = 0.87$, and OpenAlex–SciSciNet $r = 0.86$. The mean absolute differences range from 2.3 to 4.1 citations. Relative disagreement is most pronounced for low-cited papers (mean relative difference ∼20% for papers with <10 citations) and diminishes for high-cited papers. The most extreme outlier—a paper with 257,887 citations in S2AG and zero in OpenAlex—illustrates how coverage and methodology differences can produce dramatic record-level discrepancies.

This three-way comparison is *only possible* when parallel citation counts from independent sources coexist in a single queryable table. No pairwise API-based comparison can reproduce this analysis at scale.

Figure 7 summarizes the results of all four vignettes.

5.6 AI-Assisted Querying

The Science Data Lake includes a structured schema reference (`SCHEMA.md`, approximately 1,200 lines) designed to serve as context for large language model (LLM) based coding agents. This document provides every table name, column, data type, row count, and size tier, along with nine cross-dataset join strategies and common query recipes—all in a format optimized for LLM consumption rather than narrative reading.

In practice, a researcher can provide `SCHEMA.md` as context to an LLM agent (e.g., via a system prompt or file attachment) and issue natural-language analytical requests. The agent can then compose correct DuckDB SQL queries spanning multiple schemas without manual

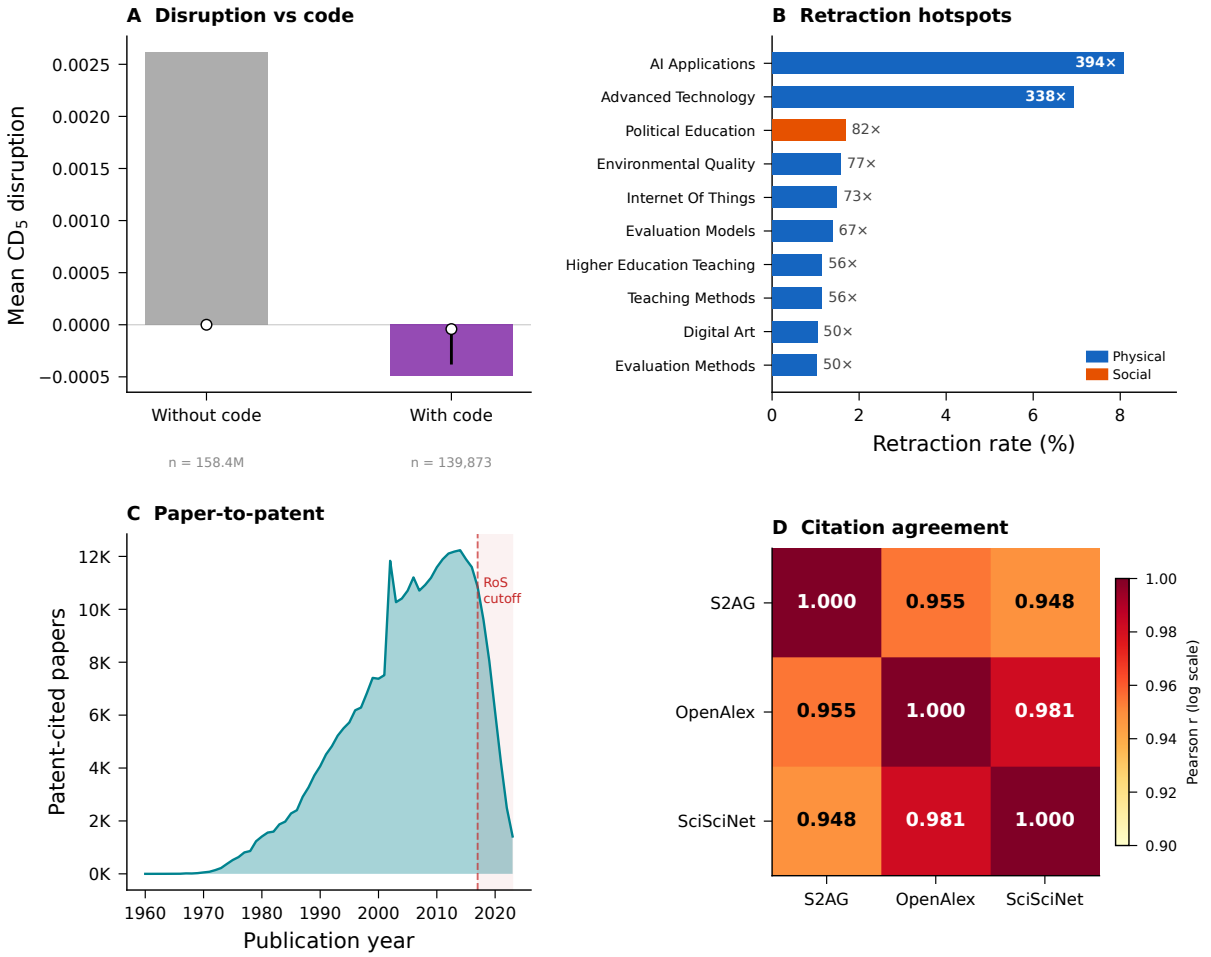

Figure 7: Composite vignette results (2×2 panels). Top-left: disruption distributions for papers with versus without code (Vignette 1). Top-right: retraction enrichment by ontology domain (Vignette 2). Bottom-left: citation distributions for patent-cited versus non-patent-cited papers (Vignette 3). Bottom-right: pairwise citation agreement across three sources (Vignette 4).

schema exploration. For example, the prompt "find the most disruptive papers in computer science that have open-source code and check their retraction status" requires joining four schemas (`sciscinet`, `xref`, `pwc`, `retwatch`) with appropriate DOI normalization—a query that `SCHEMA.md` provides sufficient context to construct autonomously.

This design reflects a deliberate choice: rather than building a custom natural-language interface (which would require maintenance and constrain the query space), we provide structured documentation that any general-purpose LLM agent can consume. As LLM capabilities evolve, the same schema reference remains useful without modification. This approach aligns with the broader trend of LLM-driven scientific discovery [14], where researchers increasingly delegate data retrieval and exploratory analysis to AI assistants that operate over structured data.

5.7 Limitations and Extensibility

The Science Data Lake inherits the limitations of its constituent sources. Temporal coverage varies: SciSciNet metrics end around 2022, RoS exhibits patent processing lag, and OpenAlex snapshot dates may trail real-time data by weeks to months. Papers without DOIs (estimated at 5–15% depending on field and era) are excluded from cross-source linkage. The ontology mapping relies on the current OpenAlex topic taxonomy, which may evolve across snapshots.

The architecture is designed for extensibility: adding a new data source requires writing

a Parquet converter and registering the schema in the pipeline configuration. Community contributions of additional sources, ontologies, or cross-reference methods are encouraged.

6 Code Availability

All code for constructing and querying the Science Data Lake is available in a public GitHub repository (`https://github.com/J0nasW/science-datalake`). The pipeline is implemented in Python 3.12 with the following principal dependencies: DuckDB 1.4.2 [18], PyArrow 22.0, and sentence-transformers 5.2.2 (for ontology alignment).

The pipeline comprises five stages, each implemented as a subcommand of the master CLI script:

1. **Download** (`datalake_cli.py download`): retrieves source snapshots from their official distribution points (S3 buckets, APIs, direct downloads).
2. **Convert** (`datalake_cli.py convert`): transforms each source from its native format (JSON Lines, CSV, N-Triples) into columnar Apache Parquet files.
3. **Create views** (`create_unified_db.py`): generates the DuckDB database with 151 SQL views across 22 schemas.
4. **Materialize** (`materialize_unified_papers.py`): constructs the `xref.unified_papers` join table through DOI normalization and multi-source record linkage.
5. **Build linkage** (`build_embedding_linkage.py`): computes BGE-large embeddings for ontology terms and OpenAlex topics, builds a FAISS index, and produces the ontology alignment table `xref.topic_ontology_map`.

Additional scripts include `convert_ontologies.py` (five format-specific parsers for the 13 ontologies) and `ontology_registry.py` (a declarative registry of ontology URLs and formats). The repository also includes a structured schema reference (`SCHEMA.md`) that documents all 151 views with their columns, types, row counts, and cross-dataset join strategies, designed to be consumed by LLM-based coding agents as well as human developers. The full pipeline runs in approximately 48 hours on a workstation with a 24-core CPU, 252 GB RAM, and an NVIDIA RTX A4500 GPU (used only for embedding computation).

Acknowledgements

[To be completed before submission.]

Author Contributions

[To be completed before submission.]

Competing Interests

The authors declare no competing interests.

References

[1] Lu Liu, Benjamin F. Jones, Brian Uzzi, and Dashun Wang. Data, measurement and empirical methods in the science of science. *Nature Human Behaviour*, 7:1046–1058, 2023. doi: 10.1038/s41562-023-01562-4.

[2] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Belber, David Blaschke, Joel Brandl, Daniel Coombs, Jonathan Flament, David Graber, Kaitlyn Kenning, et al. The Semantic Scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.

[3] Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

[4] Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data*, 10:315, 2023. doi: 10.1038/s41597-023-02198-9.

[5] Papers with Code. Papers with code: A free and open resource for machine learning. https://paperswithcode.com, 2024. Accessed October 2024.

[6] The Center for Scientific Integrity. Retraction watch database. https://retractionwatch.com, 2024. Accessed September 2024.

[7] Matt Marx and Aaron Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594, 2020. doi: 10.1002/smj.3145.

[8] Michael Gusenbauer. Beyond Google Scholar, Scopus, and Web of Science: An evaluation of the backward and forward citation coverage of 59 databases' citation indices. *Research Synthesis Methods*, 15(5):802–817, 2024. doi: 10.1002/jrsm.1729.

[9] Martijn Visser, Nees Jan van Eck, and Ludo Waltman. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and 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, and Jay Pujara. PubGraph: A large-scale scientific knowledge graph. *arXiv preprint arXiv:2302.02231*, 2023.

[11] Michael Färber, David Lamprecht, Johan Krause, Linn Aung, and Peter Haase. SemOpenAlex: The scientific landscape in 26 billion RDF triples. In *The Semantic Web – ISWC 2023*, pages 94–112. Springer, 2023. doi: 10.1007/978-3-031-47243-5_6.

[12] Daniel W. Hook, Simon J. Porter, and Christian Herzog. Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3:23, 2018. doi: 10.3389/frma.2018.00023.

[13] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources to advance general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2436–2446, 2024. doi: 10.1145/3626772.3657878.

[14] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023. doi: 10.1038/s41586-023-06221-2.

[15] Russell J. Funk and Jason Owen-Smith. A dynamic network measure of technological change. *Management Science*, 63(3):791–817, 2017. doi: 10.1287/mnsc.2015.2366.

[16] Lingfei Wu, Dashun Wang, and James A. Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566:378–382, 2019. doi: 10.1038/s41586-019-0941-9.

[17] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013. doi: 10.1126/science.1240474.

[18] Mark Raasveldt and Hannes Mühleisen. DuckDB: An embeddable analytical database. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1981–1984, 2019. doi: 10.1145/3299869.3320212.

[19] Angelo A. Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Aliaksandr Birukou, Francesco Osborne, and Enrico Motta. The Computer Science Ontology: A comprehensive automatically-generated ontology of research areas. *Data Intelligence*, 2(3):379–416, 2020. doi: 10.1162/dint_a_00055.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

[21] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.