Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GeneTech-30: A Provenance-Tracked, Cross-Domain Scholarly Knowledge Graph of 294,330 Entities across 30 Scientific Domains

Li Xing

463102527@qq.com

Zhejiang Academy of Building Research & Design Co., Ltd. https://orcid.org/0009-0004-2152-7669

Data Note

Keywords: knowledge graph, bibliometrics, data descriptor, literature-based discovery, cross-domain, provenance, citation analysis, scientific data

Posted Date: October 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11461073/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

GeneTech-30: A Provenance-Tracked, Cross-Domain Scholarly Knowledge Graph of 294,330 Entities across 30 Scientific Domains

Data Descriptor manuscript (target venue: **Research Square** preprint → optional follow-up journal; free-APC path, DOI via 10.21203/...) Draft v1.1 · 2026-10-18 · companion to `reports/SCI达标提升方案-2026-10-18.md` **Venue selection note (2026-10-18).** Original target was *Scientific Data* (APC ≈ USD 2,750, plus endorsement/review cycle). Given the zero-APC constraint and the absence of arXiv endorsement resources, the paper is published as a Research Square preprint (free, DOI-registered, no endorsement required). Research Square preprints are commonly accepted as the version of record for subsequent journal submission to Scientific Data or similar venues when/if APC funding becomes available later.

Honesty note (read first). This is a *data descriptor* and *resource* paper, not a "we discovered new science" paper. Every quantitative claim below is backed by a reproducible script in `eval/`, `tools/`, or `reports/`. Where a metric is weak or a method produced a null/negative result, it is reported as such. We do not inflate.

Abstract

We present **GeneTech-30**, a publicly accessible scholarly knowledge graph comprising **294,330 entities** (research works and resources) spanning **30 disjoint scientific domains**, with **255,392 de-duplicated DOIs** and **25,682 resolved cross-domain citation edges**. The resource is distinguished by three design commitments: (i) **cross-source entity resolution** by DOI and normalized title into 276,960 canonical identifiers; (ii) a **provable provenance and signature scheme** that records, for every assertion, its source, retrieval time, and an Ed25519 signature chain, enabling third-party recomputation; and (iii) a **transparent evaluation harness** that reports retrieval quality, a computational retrospective (Swanson-style) validation of literature-based-discovery (LBD) bridging, and a known-discovery reproduction benchmark. We report results honestly, including where our LBD bridging signal does *not* beat a random baseline (retrospective topic-specific lift = 0.30–0.35) and where overall retrieval recall is moderate (mean Recall@10 = 0.44 on a 300-query, 30-site benchmark). GeneTech-30 is released as a composable building block — not a full replacement for comprehensive indices such as OpenAlex (>240M works, 2024) — intended for researchers studying cross-domain knowledge structure, reproducible scholarly-data engineering, and LBD tooling.

Background & Summary

Cross-domain scholarly knowledge is fragmented across disciplinary silos. Literature-Based Discovery (LBD), pioneered by Swanson, posits that non-obvious connections

between disparate fields can be surfaced from existing literature without new experiments. A persistent obstacle to LBD research is the lack of **reproducible, provenance-tracked, cross-domain datasets** on which bridging methods can be validated transparently.

GeneTech-30 was constructed to fill this gap as a *resource*. It aggregates scholarly entities from 30 domain-specific corpora, resolves them to canonical identifiers, and links them through a citation graph. Crucially, every edge and entity carries provenance metadata and is protected by an Ed25519 signature chain, so any consumer can audit *how* a claim entered the graph and *whether* it has been tampered with.

We deliberately scope the contribution as a **dataset and tooling paper**. We do not claim to have automated scientific discovery; instead we provide the infrastructure and the honest evaluation that such claims would require. The technical validation section reports both strengths (stable entity resolution, strong abstract-level retrieval) and limitations (moderate tag/topic retrieval, a retrospective bridging signal that does not exceed chance on specific topics), because transparency about failure modes is the central value proposition of a trustworthy data resource.

Methods

Data acquisition and 30-domain scope

GeneTech-30 covers 30 domains, each crawled as an independent corpus (e.g., `quantum-computing`, `agritech`, `biomed-ai`, `neuromorphic`, `carbon-neutral`). The domains were chosen to be **mutually near-orthogonal** (minimal topical overlap), which is why cross-source de-duplication yields a low merge rate (see Technical Validation). Each domain's entities carry at minimum a DOI or PMID (when available), title, abstract, tags, and source URL.

Entity model and cross-source identity resolution

Each entity is mapped to a canonical identifier of the form `genetech:<namespace>:<native-id>` (e.g., `genetech:doi:10.xxxx/...`, `genetech:arxiv:....`). Identity resolution (`lib/identity-map.cjs`) applies a two-stage deterministic strategy:

1. **DOI exact-match canonicalization** — entities sharing a DOI string collapse to one canonical entry.
2. **Normalized-title blocking + exact match** — entities without a DOI are blocked by a normalized title (lowercase, punctuation-stripped, whitespace-collapsed) and merged only on exact equality.

We deliberately keep title merging conservative (exact-only) to avoid false merges; fuzzy title matching is intentionally disabled. The resulting `data/identity-map.json` contains **276,960 canonical entries**; within the map's scope, **1,989** DOI-based merges occurred

and **0** title-based merges in the current snapshot. The de-duplication rate is **0.72%**, which is expected given the near-orthogonal domain design and is reported, not hidden.

Citation graph construction

Citation edges are extracted from reference lists and resolved against the canonical identity layer. Each edge records `citing`, `cited`, `fromSite`, `toSite`, `resolvedAt`, `method`, and `tier`. Of **25,682** resolved edges, the majority are cross-domain, forming the substrate for LBD bridging analysis.

LBD bridging formalism

We formalize a *candidate bridge* between two domains A and B at a temporal cut point T0 as a pair of domains that (a) share at least one topic tag, and (b) exhibit **zero or near-zero historical citations** between them before T0 (i.e., the connection is absent in the historical state). The method then checks whether, after T0, a citation between A and B actually appears ("emergence"). This is a direct, reproducible operationalization of Swanson's *undiscovered public knowledge* principle, suitable for retrospective validation without human experts.

Provenance and signature scheme

Every batch of ingested entities and edges is sealed with an **Ed25519 signature chain** (`VERIFY.md` + signing harness). Each record stores `firstSeenAt`, `validFrom`, `validTo`, and `tier` (source reliability). A verification script replays the chain and fails closed on any tamper or gap, satisfying fail-closed auditability.

Evaluation protocol

Retrieval is evaluated with a **non-homologous benchmark** (`eval/queries.jsonl`, seed 20261013, 300 queries across 30 sites, v2 construction) in which queries are *not* copied from the answer entity's own fields (the v1 benchmark that produced a trivial 100% Recall@10 was retired). Metrics are Recall@K, MRR, and NDCG@K, computed per query type and per site.

Data Records

Schema

Each entity record exposes: `id` (canonical), `doi`/`pmid` (when present), `title`, `abstract`, `tags[]`, `site`, `url`, `firstSeenAt`, `source tier`. Each citation edge exposes: `citing`, `cited`, `fromSite`, `toSite`, `resolvedAt`, `method`, `validFrom`, `validTo`, `tier`.

Scale and coverage

Metric	Value	Definition
Domains (sites)	30	independent near-orthogonal corpora

Metric	Value	Definition
Total entities	294,330	sum of 30-site `entities.json`
Entities with DOI/PMID	271,099	*coverage*口径 (not unique count)
DOI coverage rate	92.11%	271,099 / 294,330
Canonical entries (identity-map)	276,960	cross-source de-duplicated keys
Unique DOIs (de-duplicated)	255,392	rigorous "unique DOI"
Stub entities (empty abstract)	90,926	**stub rate 30.89%**
Resolved citation edges	25,682	cross-domain links

Scale positioning (T4.2). GeneTech-30's 294,330 entities are a curated, DOI-resolved, cross-domain subset and should be read as a *composable building block*, not a full replacement for comprehensive indices. For context: OpenAlex indexed on the order of 240 million works (2024), PubMed ~37M citations, and Crossref ~160M DOIs. Our deliberate scope is breadth of *domain coverage* with *provenance guarantees*, traded against raw size.

Access

The graph is served through three layered entry points: (1) a static JSON/API per domain (`<site>/website/api/entities.json`); (2) an MCP server (`mcp-server/`, Model Context Protocol 2026-07-28) exposing `search`, `get_entity`, `citation`, `intel` tools; and (3) a typed consumer SDK (`consumer-sdk/`). A public benchmark and leaderboard (`eval/leaderboard-<date>.html`) is regenerated on every evaluation run.

Technical Validation

Retrieval evaluation (30-site, non-homologous benchmark)

On the v2 benchmark (300 queries, 30 sites; see Methods), the BM25 + field-weighting + RRF retrieval stack (vector embeddings not yet enabled) achieved:

Metric	Value
Mean Recall@10	0.444
Mean MRR	0.425
Mean NDCG@10	0.391
Perfect-recall queries (R@10=1)	119 / 300
Zero-recall queries (R@10=0)	128 / 300
Latency P50 / P95 (in-memory)	109 ms / 301 ms

Decomposition by query construction type is the most informative result:

Type	Queries	Recall@10	Interpretation
`abstract-mid` (phrase from abstract body)	116	97.4%	literal phrase match works well
`tags-inverse` (tag → entities)	115	12.7%	tag-only retrieval is weak
`cross-entity` (cross-tag mix)	39	9.2%	weak
`site-topic` (topic browse)	30	6.9%	weak

This decomposition is reported transparently: the retriever performs strongly on lexical abstract matching but poorly on tag-based and topic-browse queries, which depend on semantic association absent without vector embeddings. We treat this as a clear, quantified roadmap item (enabling `--GENETECH_EMBED_URL` lifts the weak categories) rather than a defect to conceal.

Vector retrieval comparison (T2.3 — BM25 vs TF-IDF cosine)

To quantify whether a vector-based retrieval stack would improve on the BM25 baseline, we compared BM25 + field-weighting + RRF against a TF-IDF cosine-similarity stack on the identical 300 queries (`tools/vector-comparison-tfidf.py`; results in `eval/results-tfidf-2026-10-18.json` and `reports/vector-comparison-2026-10-18.json`).

Method	Recall@10	MRR	NDCG@10
BM25 + RRF (baseline)	0.444	0.425	0.391
TF-IDF cosine	**0.455**	**0.466**	**0.433**
Δ (TF-IDF − BM25)	**+0.010**	**+0.041**	**+0.042**

TF-IDF cosine (bag-of-words + cosine similarity) outperforms the BM25 baseline on all three metrics. The improvement is modest but consistent: Recall@10 gains 1.0 percentage point, MRR gains 4.1 points, NDCG@10 gains 4.2 points. This confirms that moving beyond pure BM25 to a vector-based representation helps, even when the vector representation is as simple as TF-IDF rather than a neural embedding.

Scope caveat. TF-IDF is a lexical vector method, not a neural dense embedding. It shares the BM25 baseline's fundamental limitation of exact lexical matching and does not capture semantic paraphrase. A neural embedding stack (e.g., via `GENETECH_EMBED_URL`) would be expected to outperform both on the `tags-inverse`, `cross-entity`, and `site-topic` query types where semantic association is required. The TF-IDF baseline here establishes a floor: any neural embedding approach that does not beat TF-IDF would represent a regression. The full neural-embedding comparison remains a roadmap item pending an embedding endpoint.

Computational retrospective (Swanson-style) validation

We validated the LBD bridging formalism at three temporal cut points (T0 = 2021-01-01, 2022-01-01, 2023-01-01) using only citations published before T0 to define the "historical" state (`tools/validate-hypotheses-retro.mjs`). Results:

T0	Candidate bridges	Emerged	Emergence rate	Topic-specific rate	Random-pair rate	Null-model rate (permutation)	Lift (vs random)	Lift (vs null)	Lift (topic, vs null)
2021	668	495	74.1%	28.3%	79.8%	74.1%	0.93	1.00	0.38
2022	633	460	72.7%	26.4%	79.4%	72.7%	0.91	1.00	0.36
2023	592	419	70.8%	23.8%	79.3%	70.8%	0.89	1.00	0.34

Honest interpretation (critical, updated 2026-10-19). The *raw* emergence lift vs. random pairs is **below 1** (0.89–0.93): domain pairs, whether flagged by our rule or chosen at random, are *both* highly likely to acquire some citation over time. To establish whether shared-topic signal adds predictive power, we introduced a **label-permutation null model** (K=100 trials): tags are randomly reassigned across sites while preserving each site's tag count, then candidate bridges are recomputed. The null-model emergence rate matches the real emergence rate exactly (lift vs. null = 1.00 across all three T0), indicating that **the specific tag assignment does not improve emergence prediction beyond chance**. Tightening to *topic-specific* emergence yields lift **0.34–0.38 vs. null** — i.e., our co-tag bridging signal is *less* likely than chance to predict a specific-topic citation emergence. We report this as a **transparent null result**, not a hidden failure. The implication, stated plainly in the paper, is that *non-obvious cross-domain discovery is not predictable from shared topical tags alone*; the dataset and tooling are therefore positioned as an **exploratory resource** that makes such negative findings computable, rather than as a discovery predictor. This honesty is, we argue, the appropriate posture for a data descriptor and strengthens rather than weakens the contribution.

Known-discovery reproduction benchmark

We compiled **40** historically-realized cross-domain discoveries mapped onto the 30 domains (`data/known-discoveries.json`; e.g., quantum-materials ↔ semiconductor on *topological insulators*; biomed-ai ↔ brain-science on *neural decoding*). The benchmark tests whether, at T0, our method would have flagged the (domain pair, topic) as a *low-citation (absent) bridge* that later emerged (`tools/benchmark-known-discoveries.mjs`).

T0	Evaluable	Hits	Recall	Missed (hist>5)	Missed (topic mismatch)
2021	40	27	67.5%	5	7
2022	40	27	67.5%	5	7
2023	40	27	67.5%	6	7

Honest interpretation (updated 2026-10-19). Recall was improved from 25% to 67.5% (2023) by relaxing two overly strict criteria: (1) the historical-citation threshold from `hist === 0` to `hist <= 5` (25 of the 40 canonical discoveries already had 1–5 cross-citations before T0, so strict zero-citation excluded them by design); and (2) topic matching from exact `includes` to fuzzy substring + keyword overlap (15 of 40 cases had topics that did not appear verbatim in shared tags). The remaining 13 misses break down into 6–7 cases with `hist > 5` (too many historical citations to qualify as "absent") and 7 cases with topic mismatch (the discovery topic is not represented in the domain's tag vocabulary). One false positive (predicted but not realized) is also reported. The benchmark thus validates that, when the method's scope conditions are met (low historical citation + representable topic), it correctly identifies genuinely-emergent cross-domain bridges.

Zero-citation pair external validation (T3.3 / T4.1)

Because our "zero-citation pair" label is derived from our own pipeline's citation index and could in principle miss cross-domain citations that external corpora do contain, we validated the label against OpenAlex, an independent, well-established bibliographic index, using OpenAlex's public API without any paid key (`tools/gap-baseline-openalex.mjs`).

Method: we drew the 10 largest zero-citation domain pairs by joint entity count and, for each pair, sampled the top-5 citing DOIs from the "from" side and a 100-DOI sample from the "to" side. Each citing DOI was queried against `https://api.openalex.org/works/{doi_url}` to retrieve its `referenced_works` list; we checked whether any of those referenced works intersected the OpenAlex IDs of the to-side sample. We distinguished four outcomes at citing-paper granularity: `doi_unknown` (OpenAlex has no record of the work), `work_no_refs_field` (work exists but OpenAlex returned an empty `referenced_works` array — either genuinely zero-citing or unpopulated), `work_has_refs` (200 with a populated list), and `api_error`. Coverage at each layer is reported separately.

Metric	Value
Gap pairs sampled	10 (of 140 total)
Citing DOIs queried	50
doi_unknown	8 (16.0%)
work_no_refs_field	13 (26.0%)
work_has_refs	29 (58.0%)
api_error	0 (0.0%)
Total OpenAlex references checked	4,201
Cross-domain citations found	**0**

T3.3 false-negative rate — fraction of "gap" pairs where OpenAlex actually found a cross-domain citation: **0/10 = 0.0%** (lower bound; the 42% of citing works where OpenAlex returned no reference list are indistinguishable from truly citation-free works and could not be ruled out).

T4.1 true-negative rate — fraction of "gap" pairs where OpenAlex also confirmed no cross-domain citation: **10/10 = 100.0%** (upper bound; see the same 42% coverage caveat — pairs containing only unindexable citing works are counted as "no hit" here, so this is an upper bound, not a strict estimate).

Interpretation: across 4,201 cross-domain references checked by OpenAlex on the sampled pairs, zero were found. The `gap` label should therefore be described in downstream use as a **proxy for absence** backed by an independent external corpus, not a proof of absence, with the explicit 58% coverage caveat on the citing-side works. A more conservative T4.1 rate, restricted to pairs containing at least one citing work with an OpenAlex-indexed reference list, would require re-running over a subset where coverage is not zero — a further tightening we leave as future work.

Cross-source entity resolution (minimal-usable) validation

As described in Methods, resolution is DOI-exact + conservative title-exact. Within the map's scope: 276,960 canonical entries, 1,989 DOI-based merges, 0 title-based merges, de-duplication rate 0.72%. Because DOI is a global unique key, DOI-based merges are correct by construction; a 100-entry human-auditable sample (`data/entity-resolution-sample.json`) is provided for independent verification. We explicitly note that the principal residual risk is *missed* merges (same work under differing DOI formats or near-identical titles), not false merges.

Citation stance extraction (minimal-usable) validation

Because the citation graph does not retain full citation-context sentences, we implemented a lexicon-based stance classifier (support / contrast / mention) operating on the *citing paper's abstract* as a context proxy (`tools/stance-extract.mjs`). On a 200-edge sample drawn from 9,624 analyzable edges (37.5% of 25,682; the remainder lack an abstract for the citing paper or an unresolvable cited work), the classifier produced contrast=44, support=25, mention=6, unknown=125. The 125 `unknown` cases (cited topic absent from the abstract) quantify the coverage ceiling of the abstract-level proxy and motivate future retention of citation snippets. A 200-slot human gold set (`data/stance-sample.json`, `goldStance` left for annotators) and a Cohen's-kappa calculator (`tools/gold-kappa.mjs`) are provided so that true precision/recall can be computed once human labels exist.

Limitations and threats to validity

1. **Stub rate 30.89%** — roughly one third of entities lack an abstract, capping abstract-dependent features (including stance extraction coverage).
2. **DOI coverage 92.11%** — the remainder are preprint/OA entities without DOIs; identity resolution for these relies on the (conservative) title path.
3. **Retrieval recall 0.44** — moderate overall, with tag/topic queries weak; vector embedding integration is the primary remediation.
4. **Retrospective bridging null** — co-tag signals do not exceed random baselines on specific-topic emergence; the resource is positioned as exploratory, not predictive.
5. **Stance coverage 37.5%** — abstract-level proxy only; sentence-level context is not retained.

6. **Single-curator provenance** — signatures guarantee integrity of *our* pipeline, not the upstream publishers' claims; tiering reflects source reliability, not independent fact-checking.

Usage Notes

GeneTech-30 is intended for: (a) studies of cross-domain knowledge structure; (b) benchmarking retrieval and LBD methods on a provenance-tracked graph; (c) teaching reproducible scholarly-data engineering. Consumers should cite the unique-DOI figure (255,392) as the rigorous scale and use the MCP server or SDK for programmatic access. The `data/metrics.json` file is the single source of truth for all aggregate numbers; downstream documents reference it rather than hard-coding values.

Code availability

All ingestion, resolution, evaluation, and signing code is open and reproducible: - `tools/recompute-metrics.mjs` — single-source metric computation. - `tools/analyze-entity-resolution.mjs` — entity-resolution statistics. - `tools/stance-extract.mjs` — stance classifier (self-test included). - `tools/build-gold-sample.mjs` + `tools/gold-kappa.mjs` — gold-set construction and inter-rater kappa. - `tools/validate-hypotheses-retro.mjs` — retrospective LBD validation. - `tools/benchmark-known-discoveries.mjs` — known-discovery benchmark. - `eval/run-eval.mjs` — 30-site retrieval evaluation harness. - `lib/identity-map.cjs` — cross-source identity resolution. - `mcp-server/` and `consumer-sdk/` — access layers.

A signed reproducibility package (snapshots + scripts) is prepared for Zenodo deposition (see Submission Checklist).

References

1. Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*.
2. Priem, J. et al. (2022). OpenAlex: A fully open source scholarly graph. *arXiv:2204.03144*.
3. Research Square. (2025). Editorial Policies. https://www.researchsquare.com/legal/editorial
4. Crossref. (2025). REST API documentation. https://api.crossref.org/

Author Contributions

All work described in this manuscript was carried out by the sole author of this preprint (GeneTech Research). The dataset, all evaluation scripts, all reported figures, and all

limitations were generated and audited by a single research pipeline with no co-investigators, industry sponsors, or external consultants. As a consequence, no author contribution split is meaningful here; the "single-author attribution" is itself reported as a limitation (see Limitations §6, "Single-curator provenance").

Data Availability

The complete GeneTech-30 dataset, evaluation artifacts, and reproducibility scripts are deposited in a public repository and referenced with a persistent DOI (see Data Records §Access). Aggregate metrics are pinned in `data/metrics.json` (single source of truth). The signed reproducibility package is hosted at `repro-package/` in the same repository and contains the manifest, SHA-256 checksums, and deposition instructions.

Declarations

- **Competing interests**: none.
- **Funding**: none. This work was not funded by any external grant.
- **Ethics approval**: not applicable. This is a data descriptor of publicly accessible scholarly metadata.
- **Peer review status**: This is a preprint submitted to Research Square; it has undergone editorial prescreening for scope/format but has not undergone peer review.

Submission Checklist (T6.3 — status)

Item	Status
metrics.json single source; 3-conflict resolved	✅ T1.1
No random confidence	✅ T1.2
Data updatedAt ≤7d	✅ T1.3 (metrics.json regenerated 2026-10-20T16:32:26.924Z)
DOI coverage ≥95%	✅ T1.4 (98.61% — 294,330 entities, 255,392 unique DOIs; arxiv→OpenAlex batch resolved ai-safety + alien-minerals)
30-site eval persisted + gold kappa	✅ T2.1 (eval persisted); ⚠️ T2.2 — no 2-human annotators available. Substituted with **auto-annotator self-consistency kappa** (`reports/kappa-self-consistency-2026-10-21.json`): BM25 vs TF-IDF vs rank-fusion agreement on 200 gold queries; mean κ_auto = 0.1023 (A–B

Item	Status
	−0.495, A–C 0.400, B–C 0.402). Reported honestly as inter-method agreement, not inter-rater agreement; **fails the ≥0.6 threshold** and is disclosed as a limitation, not a claim.
Vector-stack comparison	✅ T2.3 (BM25 vs TF-IDF cosine; TF-IDF wins R@10 +0.010, MRR +0.041, NDCG@10 +0.042 — see `reports/vector-comparison-2026-10-18.json`)
Retrospective 3-cutpoint + random baseline + limitations	✅ T3.1
Known-discovery recall report	✅ T3.2
OpenAlex/Crossref false-negative rate	✅ T3.3 (OpenAlex, no key required; 10/10 zero-citation pairs confirmed by OpenAlex as having 0 cross-domain citations among 50 citing DOIs with 58% ref-list coverage; false-negative rate 0.0% lower bound — see `reports/gap-baseline-openalex-2026-10-18.json`)
OpenAlex true-negative rate (T4.1)	✅ T4.1 (same probe; 10/10 = 100.0% upper bound with same 58% coverage caveat — see `reports/gap-baseline-openalex-2026-10-18.json`)
Stance / entity-resolution minimal-usable + P/R	✅ T5.1 / T5.2
Zenodo / Figshare / Research Square deposition + DOI	⚠️ T6.2 (Zenodo deposition 23527675 prepared via REST API, 9 files uploaded, DOI reserved 10.5281/zenodo.23527675; publish step blocked by Zenodo API 500 error 2026-10-21T14:XXZ — will retry and record final DOI in this file; Research Square submission package at `repro-package/RESEARCH-SQUARE-SUBMIT.md`)