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.

Mapping artificial intelligence for science with open metadata reveals shifting evidence regimes across scientific domains

Xu Fengchang

haiquan2018@sina.com

Shandong Light Industry Vocational College

Lv Ge

Shandong Light Industry Vocational College

Wang Xinliang

Binzhou Civil Air Defense Engineering and Command Support Center

Zhang Na

Shandong Light Industry Vocational College

Research Article

Keywords: AI for Science, science mapping, evidence regimes, epistemic labour, bibliometrics, Crossref, term co-occurrence

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11461912/v1

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

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence for science (AI4S) is often discussed as a unified transformation, although scientific domains differ markedly in how model outputs are tested, validated, and justified. We map an audited corpus of $^{N=2,784}$ records (2014–2025) retrieved from Crossref and screened with a frozen, versioned rulebook, and compare evidence regimes in protein science, materials design, drug discovery, closed-loop experimentation, and mathematical reasoning. Using term co-occurrence, reference co-citation, burst detection, growth models, and collaboration networks, we identify three robust patterns: (i) rapid but decelerating growth (Poisson rate 1.381, 95% CI 1.361–1.402, declining from 1.537 in 2014–2019 to 1.320 in 2019–2025); (ii) a coherent term-co-occurrence structure (75 nodes, modularity $^{Q=0.404}$) organized around four measured streams—representation learning, generative and inverse design, closed-loop experimentation, and neuro-symbolic reasoning—with the generative stream acting as the connecting hub; and (iii) a right-boundary rise in *large language model* terminology (burst state 3) that a year-by-year backtest shows is not a last-observation artefact. The maps are sensitive to metadata availability: abstract coverage is 50.6%, reference coverage 84.3%, and a title-only variant changes the network partition substantially (adjusted Rand index 0.762), so the maps describe the screened open-metadata corpus rather than the total AI4S literature. Cross-domain analysis shows that AI most consequentially shifts scientific bottlenecks from representation and candidate generation toward experimental or formal validation, provenance, and responsibility for claims. We contribute an auditable, open-metadata science-mapping workflow *designed for reproducibility* and an evidence-regime framework for analysing how epistemic labour is redistributed across models, instruments, formal verifiers, infrastructures, and human experts.

1 Introduction

The phrase “fourth paradigm” has come to denote a data-intensive mode of science in which computational systems do more than process observations after the fact [1]. Machine-learning models infer latent structure from high-dimensional data; generative systems propose molecules, materials, mechanisms, and programs; active-learning systems prioritize measurements; and robotic platforms can execute, interpret, and update experiments [2]. These developments raise an ambitious prospect: scientific inquiry may become faster, more exploratory, and more capable of operating in spaces too large for unaided human search. We treat this prospect as a background conjecture to be interrogated, not as an established conclusion. Whether any particular domain is undergoing a paradigm shift is an empirical question that a bibliometric map alone cannot settle; the present study contributes one auditable map and a comparative reading of the field’s evidence practices.

The prospect should not, moreover, be confused with a completed transformation. A model that predicts a protein fold, ranks a crystal structure, writes a plausible molecular rationale, or proposes a proof tactic has not necessarily made a scientific discovery. Prediction, optimization, and discovery are related but non-identical activities. A discovery claim must be connected to an original and meaningful proposition, transparent evidence, an appropriate validation regime, and a record that permits a community to

scrutinize, reproduce, qualify, or reject the proposition [2]. The distinction is particularly important in the age of generative models, where fluency and scale can make unsupported output appear authoritative.

Landmark systems illustrate both the power and the discipline required. AlphaFold delivered highly accurate protein-structure predictions at unprecedented scale and made structural hypotheses widely available [3]; its human-proteome resource expanded coverage dramatically, but it also supplied confidence measures and could not substitute for every experimental question about dynamics, interactions, ligands, complexes, or cellular context [4]. In materials science, GNoME reported an enormous expansion of predicted stable crystals by joining graph-network predictions with density-functional-theory filtering; stability remains different from synthesis feasibility, measured performance, manufacturability, or societal value [5]. In formal mathematics, systems such as AlphaGeometry and FunSearch make candidate construction and verification more visibly separable: a neural system can search, while a symbolic or executable evaluator checks the result [6, 7].

AI4S is therefore also a social and organizational phenomenon. The relevant questions concern ownership of data, concentration of compute and robotics, documentation of failed experiments, assignment of responsibility, and the reconfiguration of expertise. A laboratory with a large proprietary corpus, automated equipment, and specialized engineering staff does not merely own a faster tool; it occupies a different position in the production of credible knowledge. Conversely, a model trained mainly on published successes can inherit visibility bias, omit negative results, and reward domains already rich in standardized data [2]. The effect may be a new form of cumulative advantage rather than an automatic democratization of science. These observations motivate the paper's theoretical focus on the *redistribution of epistemic labour* (Section 2).

Bibliometric science mapping is valuable here because the field is scattered across disciplinary databases and inconsistent labels: AI for science, machine learning for scientific discovery, self-driving laboratories, inverse design, computational drug discovery, and mathematical reasoning denote overlapping but distinct communities. Science-mapping methods—keyword co-occurrence, reference co-citation, burst detection, and collaboration networks—can map intellectual bases, emerging fronts, and capability networks without assuming that citation frequency is a direct proxy for epistemic quality [8, 9]. They can show which methods, objects, institutions, and validation practices are connected, and where bridges or gaps occur.

A note on positioning and novelty. We implement a *science-mapping* workflow—term co-occurrence, co-citation, burst, and collaboration networks—on an open-metadata corpus, not a full knowledge graph with typed entities, relations, and queryable triples; we therefore do not claim to construct a knowledge graph. In a bounded search of bibliometric and AI4S-review literature conducted before this revision (documented in the research package; see also Section 6.2), we did not locate a study that simultaneously provides (i) an audited, reproducible screen of open metadata, (ii) a cross-domain comparison of evidence regimes, and (iii) an epistemic-labour account of AI4S. Recent work has combined language models with interactive scientometric tools [10], and the sociology of discovery has

long studied the division of cognitive labour [11]; our contribution is to connect these strands to a quantified, auditable map of one open corpus.

1.1 Research questions

The article has two linked purposes: to provide a transparent and reproducible science-mapping protocol for AI-driven scientific discovery and to execute it end-to-end on an audited open-data corpus; and to use landmark evidence to synthesize how focal domains differ in their standards of evidence. Four questions organize the analysis:

RQ1 What intellectual bases, thematic communities, and attention shifts are visible in the audited Crossref corpus of AI4S research?

RQ2 How do protein science, materials design, drug discovery, closed-loop experimentation, and mathematical reasoning differ across representation, proposal, test, and justification?

RQ3 How do these differences redistribute epistemic labour among models, instruments, formal verifiers, infrastructures, and human experts?

RQ4 How sensitive are the resulting maps to retrieval, metadata coverage, term construction, and network specification?

RQ1 and RQ4 are answered directly by the bibliometric analysis; RQ2 and RQ3 are addressed by the cross-domain synthesis of Section 5 and the theoretical reading of Sections 2 and 6. Questions about reliability, inclusiveness, and explanatory value that featured in the earlier draft are retained as a governance research agenda (Section 6.6) rather than as results of the present data.

1.2 Empirical integrity

A note on data provenance. The pre-specified protocol for this study named the Web of Science Core Collection (WoSCC) as the preferred data source; because WoSCC exports were not accessible, the protocol was executed on Crossref, the community-controlled metadata service [12], under the same document-type and period constraints, and every deviation is disclosed in Section 3. The protocol was specified before the analysis but was not registered in a public registry; we therefore describe it as *pre-specified* rather than *registered*, and we report a pre-specified set of expectations (Section 3.6) evaluated descriptively rather than as confirmatory hypothesis tests. All corpus-dependent statistics reported in this article are outputs of the executed analysis; none is fabricated, extrapolated, or inserted as a placeholder.

2 Theory and analytical framework

2.1 From four layers to evidence regimes

Earlier drafts presented a four-layer chain—*representation, proposal, test, justification*—and aligned it, one-to-one, with the four thematic clusters of the map. That mapping is conceptually misleading: protein science, materials design, drug discovery, and mathematical reasoning each pass through *all four* layers, differing in how each layer is realized, not in which layer they occupy. We therefore reposition the framework. The four layers are a comparative grid (Table 1; Fig. 6) through which every domain is read; the four measured streams of the corpus (Table 3) are an empirical description of where AI4S vocabulary concentrates, not a realization of the grid.

Four-layer comparative framework for AI4S. Every domain passes through all four layers; the layers ask the same four questions of each domain.

Layer	Core question	Representative evidence	Failure to avoid
Representation	What object and constraints are encoded?	sequences, graphs, equations, instrument streams	leakage; missing scientific context
Proposal	What is predicted, generated, or searched?	ranked candidate, molecule, structure, proof step	confusing retrieval with invention
Test	How is the candidate checked?	simulation, assay, robot experiment, proof checker	benchmark-only validation
Justification	Why should the community trust it?	calibration, provenance, replication, explanation	opaque and untraceable claims

2.2 Epistemic labour redistribution as the organizing mechanism

The paper's main theoretical claim is that the most consequential effect of AI on science is a *redistribution of epistemic labour*: the tasks that produce credible knowledge are re-partitioned among models, instruments, formal verifiers, data infrastructures, and human experts, and this re-partitioning is domain-specific. The concept extends the division of cognitive labour studied for human science [11] to human–machine systems. Three mechanisms organize the account:

1. **Standardization of representation.** Cheap, reusable encodings of scientific objects (sequences, graphs, embeddings) lower the cost of candidate generation, shifting effort downstream.
2. **Bottleneck relocation.** As candidates become cheap, the binding constraint of a discovery pipeline moves to the Test and Justification layers: experimental capacity, formal verification, provenance, and responsibility for claims.
3. **Infrastructure dependence.** Who can produce credible AI4S findings depends on access to data, compute, robotics, and validation infrastructure; the bibliometric footprint of this dependence is concentration in the collaboration network.

These mechanisms support three *explanatory propositions*. We label them explicitly as propositions to be evaluated in this and future work, not as causal claims established by the present cross-sectional data:

P1 (bottleneck shift): the cheaper candidate generation becomes, the more the validation resources of a domain become the primary bottleneck of discovery. *Observable corollary*: domains with abundant, standardized representations (protein structure, crystals) concentrate attention on validation infrastructure, while domains with expensive tests (drug efficacy, clinical translation) are bottlenecked earlier.

P2 (formalization enables audit): the more formal and machine-executable a domain's verification rules, the more explicit and auditable the human–machine division of labour becomes. *Observable corollary*: formal mathematics admits machine-checkable proofs, whereas translational medicine must substitute staged, socially-managed validation chains.

P3 (cumulative advantage): the more concentrated data, compute, and robotics infrastructure is, the more AI4S output and collaboration networks display cumulative advantage. *Observable corollary*: strong country-level concentration and a small set of dominant cross-national edges—which we measure in Section 4.10 but cannot here attribute to causes.

2.3 Evidence regimes as a middle-range concept

An *evidence regime* is the domain-specific configuration of objects, AI functions, validation mechanisms, responsible agents, failure modes, and evidence sources that connects a candidate to a community-accepted claim. The four-layer grid makes regimes comparable; the mechanisms above explain why they differ. Table 6 summarizes the regimes of the five focal domains; Appendix D provides a full evidence card for each case. The value of the concept is that it lets us discuss AlphaFold, GNoME, an AI-designed drug, a self-driving laboratory, and a machine-checkable proof in one framework without pretending their outcomes are equivalent.

3 Data, search strategy, and methods

3.1 Research design

The study combines bibliometric science mapping with a structured cross-domain case synthesis. The bibliometric component maps the literature's vocabulary, cited foundations, bursts of attention, growth dynamics, and collaboration structure. The case component interprets what the mapped clusters mean in practice by studying systems with identifiable inputs, candidate-generation mechanisms, test environments, and validation claims [1, 2]. Neither component alone is sufficient: a network map without close reading may reproduce database artefacts; case narratives without systematic mapping may privilege famous systems and miss the broader organization of the field.

The bibliometric unit is an individual record. The main interpretive units are term-co-occurrence clusters, reference co-citation clusters, citation bursts, and collaboration networks. The case-synthesis unit is a discovery workflow: a workflow begins from a scientific object and available evidence, uses a computational representation to predict or generate a candidate, tests the candidate through simulation,

experiment, or formal proof checking, and records a justification that can be assessed by others. This approach foregrounds the evidence chain rather than treating “AI” as a homogeneous intervention.

3.2 Source and provenance

Metadata were harvested from the Crossref REST API (polite-pool user agent with a contact e-mail), which returns titles, abstracts where deposited, cited-reference DOIs, author affiliation strings, and citation counts. The harvest was completed in August 2026; the harvest date was recorded at query time and the query log is archived with the analysis code (Section 3.5). Because Crossref is a registration agency rather than a curated citation index, it is a superset of —and more inclusive than—WoSCC, and its coverage skew differs; we therefore report the coverage of each bibliometric field (Table 5) and treat partial field coverage as a first-class limitation. We refer to the resulting dataset as the *retrieved and screened corpus*: it is a sample of what Crossref can be made to return under a fixed query set, not the total AI4S literature.

3.3 Search strategy

Sixteen keyword-combination queries were run (Appendix A), constructed as the Crossref equivalent of the pre-specified Topic search. The AI terms were *machine learning, deep learning, artificial intelligence, neural network, large language model, generative design, generative AI, generative model, foundation model, graph neural network, transformer*; the discovery terms were *scientific discovery, materials discovery, drug discovery, protein structure prediction, autonomous laboratory, self-driving laboratory, theorem proving, mathematical reasoning, inverse design, autonomous experimentation*. Queries were combined by hand so that each query carries one AI term and one discovery term; the combination is *not* a Cartesian product, and query complexity varies across domains, which we disclose as a retrieval limitation and assess with leave-one-query-out sensitivity (Section 4.1). The period was restricted to 2014–2025, document types to Articles and Reviews, and language to English, mirroring the pre-specified protocol. English was judged from Crossref’s language field where present and, for records with no usable field, from the Latin-script title/abstract text; records with neither were treated as not-language-restricted at screening. Records retrieved by more than one query were counted once (DOI deduplication, $n = 1,446$), and the unique contribution of each query was measured (Appendix A).

Three retrieval features are disclosed. First, each query returns at most $rows = 500$ records; queries whose hit counts exceeded 500 therefore under-sample the tail, and the exact page-by-page fetch parameters, cursor strategy, and sort order are archived with the query log. Second, Crossref’s metadata contain no author keywords, so all term analyses operate on title/abstract n-grams (Section 3.5). Third, DOI deduplication does not merge version clusters (preprints and journal versions with distinct DOIs); a title–author–year similarity audit was run on the included corpus, and records judged to be near-duplicate versions were collapsed where detected.

3.4 Two-stage screening and audit

Records underwent two-stage screening. Stage one removed duplicate exports and content-type false positives (books, datasets, editorials: $n = 357$). Stage two applied a frozen, versioned rulebook (screening.py v5.16) that encodes the pre-specified inclusion criterion: *included records use AI to represent, generate, search, optimize, test, verify, or govern an identifiable scientific-discovery process*; excluded records are generic administrative automation, prediction applications with no discovery objective, education-only work, and opinion pieces without a discernible mechanism. The rulebook was frozen before the audit.

The pre-specified protocol called for dual-reviewer screening with Cohen's kappa [13]. A second human reviewer was not available in this study, so kappa is *not* reported. Instead we conducted a *dual-seed audit* of the frozen rulebook: the same analyst who authored the rulebook re-screened two stratified random samples defined by different random seeds (seeds 42 and 7 for included records, seed 42 for excluded records), assigning each record to *include / borderline-include / false-positive* or *correct-exclude / borderline-false-exclusion / false-exclusion*. This does not replicate reviewer independence; it tests the deterministic rulebook for residual error under resampling, and we report exact binomial 95% confidence intervals alongside point estimates (Table 2). No rule was modified after the audit, and the residual error types are documented so that readers can judge their impact (Section 6.5).

Screening flow and audit metrics for the AI4S corpus.

Stage	n
Retrieved by 16 queries (N0)	8,000
Duplicates removed (N1)	1,446
After deduplication (N0′)	6,554
Excluded (N2): content type / no AI / no discovery object / application noise	3,770 (357 / 919 / 1,128 / 1,366)
Included (N)	**2,784**
Audit, included sample (197 unique of 200 drawn)	170 clear includes, 12 borderline, 4 clear FP, 11 borderline FP
Precision (strict; incl. borderline as errors)	97.7% [95% CI 94.2–99.4];% [87.8–95.7]
Audit, excluded sample (100 drawn)	91 correct, 2 clear FE, 7 borderline FE
False-exclusion rate (clear; incl. borderline)	2.0% [0.2–7.0];% [4.2–16.4]

3.4.1 Audit metrics.

Precision (strict) is $170/174 = 97.7\%$: the 12 borderline includes are not counted as errors, and the 4 clear false positives and 11 borderline false positives in the sample are treated as errors (the borderline

false positives are excluded from the strict denominator because they were judged borderline rather than clear). Precision including all borderline decisions as errors is $182/197 = 92.4\%$ (170 clear includes + 12 borderline includes of 197 drawn). The false-exclusion rate is the proportion of borderline/clear errors among the excluded sample only; it is *not* an estimate of recall over the full Crossref space. Confidence intervals are exact binomial (Clopper–Pearson). Audit verdicts, sampling code, and the rulebook are archived with the manuscript.

3.4.2 Data card.

Each analytic input is documented in a data card (Appendix B): owner (Crossref), licence (public metadata), sampling frame (16 queries, 2014–2025), instrument (Crossref REST API, harvest date logged), processing (DOI deduplication, rules-based screen v5.16), missingness (abstracts 50.6%, references 84.3%, affiliations 42.9%), known bias (registration-agency coverage, English-only queries, no author keywords, $rows = 500$ caps), update date, and permitted reuse (attribution).

3.5 Analysis configuration and reproducibility

All analyses are specified in code archived alongside the manuscript; the analysis was designed for reproducibility, and a full reproduction requires the archived query log, rulebook, and locked software environment (Appendix C).

- **Term co-occurrence.** Author keywords are not deposited in Crossref, so terms were operationalized as clean n-grams ($n = 1 \ldots 3$) from titles and abstracts. Frequency counts use titles+abstracts; network co-occurrence uses titles only, so that all N records contribute to the network (uniform 100% coverage) and the abstract-coverage gap (50.6%) cannot distort the topology. Stop-word filtering, a strong-unigram whitelist, and abbreviation-tail suppression are documented in the code and summarized in Appendix E. The network takes the 80 highest-frequency terms (all ≥ 10 papers); five of these are isolated nodes—n-grams such as “github com” and “time consuming” that never co-occur with another vocabulary term in any title—and are removed because degree-0 nodes cannot be drawn or clustered, leaving 75 nodes. Community detection uses the Louvain method [14] with modularity Q. Edge weights are raw co-occurrence counts, and the resulting network is dense (density 0.34, mean degree 25.2); we report this density as a property of the title-based construction rather than hiding it.
- **Reference co-citation.** Two references are co-cited when they appear together in the reference list of a corpus paper; the co-citation weight is the number of corpus papers citing both. The network retains the 200 most-cited references (cited by ≥ 4 corpus papers), enriched with titles and years via the Crossref API. *Only the 84.3% of records with reference lists contribute*; missingness is non-random (Section 4.11), so all co-citation results are restricted to the reference-available subset. The front timeline is the *citing-paper year* distribution, because reference publication years are not present in Crossref reference lists; it shows when a front was mobilized, not when its cited sources were published.

- **Citation bursts.** Kleinberg's two-parameter burst-detection automaton [15] was applied to annual term-presence series (2014–2025; 36 terms with ≥ 15 papers showed an episode), with burstiness $s = 2$, burst penalty $\gamma = 1$, and Poisson observation cost. Because 2025 is the right edge of the window, every episode active at 2025 is a *right-boundary signal*, not an established trend; we run a year-by-year backtest (Table 4) that refits the automaton on data through each cutoff year.
- **Growth model.** Annual publication counts are modelled with Poisson and negative-binomial count models on year, reporting annual incidence-rate ratios (IRRs), profile 95% confidence intervals, and a two-segment (2014–2019, 2019–2025) version. To separate domain growth from database deposition, annual counts are also expressed as a share of the annual total of Crossref Articles+Reviews.
- **Country and institution networks.** Affiliations were parsed with a country lexicon and an institution extractor (keywords like *university*, *academy*, *laboratory* preferred over department-level segments). Country co-authorship is an undirected graph with edge weights equal to the number of jointly affiliated papers. Institution names are literal string extractions; a top-institutions manual review flagged "University of California system" as a multi-campus aggregate, so institution counts are treated as indicative, and country counts are used for the substantive analysis.
- **Robustness.** Four robustness checks are reported: (i) Louvain seed stability (adjusted Rand index [16, 17] over 20 seeds); (ii) source variant (title-only vs. title+abstract co-occurrence) on a single clean vocabulary; (iii) query-set contribution (leave-one-query-out on each query's unique records); and (iv) citation central tendency under three trimming variants. Results are reported in Sections 4.1 and 4.9 and Table 5.

The full research package—screening rulebook, analysis scripts, parameter files, audit verdicts, query log, and network files—is archived alongside the manuscript (Appendix C).

3.6 Pre-specified expectations

The earlier draft framed three registered hypotheses confirmed by the data. We reframe these as pre-specified expectations evaluated descriptively; no test statistic, decision rule, or multiplicity correction was pre-registered, so we do not report them as confirmatory tests.

E1Early AI4S work is organized around property prediction, bioinformatics, and domain-specific machine learning.

E2Protein-structure prediction, generative molecular design, and materials screening form an acceleration phase.

E3Foundation models, self-driving laboratories, agentic workflows, and formal mathematics form an emerging frontier.

Each expectation is assessed against the measured trajectories in Section 4.1 and Section 4.9.

4 Results

4.1 Temporal architecture and growth dynamics

Figure 1 shows the annual and cumulative trajectories of the audited corpus. Growth is unambiguous but not monotone: from 21 publications in 2014 the field reaches 127 in 2019 (pre-2020 total $n = 308$, 11.1% of the corpus), then accelerates to 736 in 2025. The compound annual growth rate (CAGR) is 38.2% over 2014–2025, with a descriptive deceleration from 43.3% (2014–2019) to 34.3% (2019–2024).

A count model confirms the trajectory and quantifies the deceleration. A Poisson regression of annual counts on year gives an annual IRR of 1.381 (95% CI 1.361–1.402), i.e. roughly a 38% per-year increase. The negative-binomial fit gives a similar point estimate (IRR 1.431, 95% CI 1.213–1.689) with residual deviance near its expectation, indicating that over-dispersion relative to Poisson is modest. A two-segment Poisson model shows growth decelerating from an IRR of 1.537 (95% CI 1.424–1.659) over 2014–2019 to 1.320 (95% CI 1.293–1.349) over 2019–2025. The AI4S share of the annual Crossref output of Articles+Reviews nevertheless rose from 0.0007% (2014) to 0.0111% (2025), so the field is growing faster than the database: the deceleration is a deceleration *of rapid growth*, not an approach to saturation relative to the broader literature.

Evaluation of E1–E3. *E1*: the 2014–2017 phase ($n = 112$) is dominated by deep-learning drug-discovery reviews and bioinformatics applications, which is consistent with E1. *E2*: the 2018–2021 acceleration (69 $\rightarrow$ 295 papers) coincides with the co-citation fronts around AlphaFold (2021), Materials Project/GNoME-style screening, and inverse molecular design (Section 4.3), and the keyword shares of *inverse design*, *materials discovery*, and *drug design* grow 6.0–6.1 $\times$ between 2014–2017 and 2024–2025 (Section 4.9); consistent with E2, but we note that several anchor publications appear at or after the bursts' onset, so the coincidence should be read as alignment, not causation. *E3: partially* supported. Foundation-model and autonomous-laboratory terms show the corpus's strongest late episodes, but formal mathematics does not burst: *theorem proving* (70 papers) and *mathematical reasoning* (15 papers) display flat annual profiles with no detectable surge, which we interpret as a stable niche rather than an emerging front (Sections 4.8 and 4.9).

Query-set sensitivity (RQ4). A leave-one-query-out analysis drops, for each query, the records found only by that query (its unique contribution). The largest single-query dependencies are *protein structure prediction* $\times$ *deep learning* (385 unique records; 13.8% of the corpus), *inverse design* $\times$ *machine learning* (308; 11.1%), and *drug discovery* $\times$ *artificial intelligence* (236; 8.5%). Dropping *protein structure prediction* records moves the 2024–2025 two-year total by up to 11.2% and reduces the dominant-count of stream A from 757 to 381, while dropping *drug discovery* $\times$ *artificial intelligence* moves the 2024–2025 total by 11.6% and stream B from 953 to 766. The four-stream structure is therefore not an artefact of a single query, but the sizes of the representation (A) and generative (B) streams depend materially on two drug- and protein-focused queries; stream sizes should be read as conditional on the query set.

Citation central tendency. Citation concentration is strong and typical of a young, high-velocity field: 87.1% of papers have been cited at least once, the median citation count is 11, and the mean is 40.6. The most-cited 1% of papers account for 22.4% of all citations (top 5%: 47.0%; top 10%: 61.9%). The earlier draft claimed that deleting the highest-cited 1% moves the mean from 40.6 to 41.1; inspection showed the underlying code sliced the descending-sorted array at its *lowest* 1%, which legitimately raises the mean. We report three corrected variants: dropping the highest 1% yields a mean of 31.9; winsorizing the top 1% at the 99th percentile yields 36.0; dropping the lowest 1% yields 41.1. The corrected conclusion is the original intended one: the mean is pulled by a thin upper tail of landmark papers (top 1% of papers carry 22.4% of citations), while the bulk of the field receives moderate attention (median 11).

Annual and cumulative publications in the audited AI4S corpus ($N = 2,784$, 2014–2025). CAGR over the full window is 38.2%; a two-segment Poisson model gives annual growth rates of 1.537 (2014–2019) and 1.320 (2019–2025). Corpus frozen at the logged harvest date (August 2026); 2025 may under-count records indexed with a delay.

The most-cited records are indicative of the field's centres of gravity: *Applications of machine learning in drug discovery and development* [18] (2,735 citations in the corpus), *Inverse molecular design using machine learning* [19] (1,817), *The rise of deep learning in drug discovery* [20] (1,510), and *Scaling deep learning for materials discovery* [5] (1,288). Top venues include Digital Discovery (147), Materials & Design (116), Scientific Reports (82), and Drug Discovery Today (74).

4.2 Term co-occurrence structure

Figure 2 shows the term co-occurrence network of the clean 80-term vocabulary after removing five isolated n-grams: 75 active nodes, 946 edges, density 0.34, mean degree 25.2, and modularity $Q = 0.404$. Five communities emerge without any cluster label imposed in advance:

- **Community 3 (protein, 35 nodes)**: *protein, deep learning, neural network, structure prediction, binding affinity*. The largest community, organized around learned representations of biomolecular sequence and structure.
- **Community 2 (materials, 19 nodes)**: *machine learning, materials, molecular, inverse design, crystal*. Generative and inverse design vocabulary for matter.
- **Community 1 (drug, 14 nodes)**: *drug, drug discovery, artificial intelligence, drug design, virtual screening*. Application-led drug-discovery vocabulary.
- **Community 4 (ligand, 5 nodes)**: *ligand, protein–ligand, binding affinity, ligand binding*. A small but densely connected sub-community bridging protein and drug vocabularies.
- **Community 0 (theorem, 2 nodes)**: *theorem, theorem proving*. A tiny, isolated formal-mathematics community.

The most frequent vocabulary items are *protein* (2,185 papers), *drug* (2,105), *machine learning* (1,890), *deep learning* (1,371), *drug discovery* (1,170), *materials* (1,009), and *artificial intelligence* (820).

Term g-index. For a frequency ranking $f_1 \geq f_2 \geq \ldots$, the g-index is the largest g such that the sum of the top- g frequencies is at least g^2, i.e. $g = \max\{g : \sum_{i=1}^{g} f_i \geq g^2\}$. Applied to the full corpus vocabulary this gives $g = 155$, indicating a coherent, concentrated vocabulary; applied within each measured stream's signature-term frequency distribution it gives 8, 12, 6, and 5 for streams A–D (Table 3), indicating how concentrated each stream's signature vocabulary is.

Term co-occurrence network of the AI4S corpus (titles only; 80 highest-frequency clean n-grams, 5 isolates removed, 75 nodes, 946 edges; Louvain, $Q = 0.404$). Node size is term frequency; colour is community. Edges are raw co-occurrence counts.

4.3 Reference co-citation research fronts

Reference lists are present in 84.3% of records, so the co-citation analysis is restricted to that reference-available subset (its non-randomness is examined in Section 4.11). The co-citation network of the 200 most-cited references (cited by ≥ 4 corpus papers; 200 nodes, 11,022 edges) has modularity $Q = 0.346$ and separates into three research fronts (Fig. 3):

- **Protein-structure front (49 references, 886 citing papers, 2014–2025).** Anchored by AlphaFold [3] (cited by 289 corpus papers), *Deep learning* [21] (200), LSTM [22] (149), the Protein Data Bank [23] (114), and Gapped BLAST [24] (112). This is the deepest and most stable intellectual base.
- **Materials front (62 references, 852 citing papers, 2015–2025).** Anchored by the Materials Project [25] (162), Random Forests [26] (142), machine learning for molecular and materials science [27] (132), GGA [28] (103), and inverse molecular design [19] (102). Its intellectual base mixes modern machine learning with classic density-functional-theory methods.
- **Molecular-design front (89 references, 833 citing papers, 2014–2025).** Anchored by automatic chemical design with variational autoencoders [29] (145), SMILES [30] (131), machine learning in drug discovery [18] (124), extended-connectivity fingerprints [31] (100), and the rise of deep learning in drug discovery [20] (93).

The three fronts are temporally complementary: the protein front is mobilized across the whole window, the materials front consolidates after 2015, and the molecular-design front is the largest in size. Crucially, the co-citation structure does *not* mirror the five term communities one-to-one: the drug (community 1) and ligand (community 4) communities share a single molecular-design front, indicating that drug discovery and biomolecular representation learning are built on overlapping foundations.

Reference co-citation research fronts (left) and their citing-paper year timelines (right; the timeline shows when corpus papers citing each front were published, not the publication years of the cited sources). Three fronts: protein structure (blue), materials (orange), molecular design (green); $Q = 0.346$, 200 references, reference-available subset.

4.4 Measured streams of the corpus

Four measured streams summarize where AI4S vocabulary concentrates. Each paper is assigned to at most one dominant stream by a published signature-term dictionary (Appendix E); signature support is the count of papers carrying ≥ 2 signature terms. Table 3 reports the mutually exclusive dominant-cluster counts.

Measured streams of the AI4S corpus (mutually exclusive dominant-cluster assignment; signature support = papers carrying ≥ 2 signature terms).

Stream	Dominant	Support	g-index	Leading signature terms
A. Representation & structure	757 (27.2%)	487 (17.5%)	8	protein structure prediction (173), secondary structure (213), binding affinity (119)
B. Generative & inverse design	953 (34.2%)	267 (9.6%)	12	inverse design (248), drug discovery (1170), materials discovery (126)
C. Closed-loop experimentation	81 (2.9%)	21 (0.8%)	6	active learning (40), Bayesian optimization (17), autonomous experimentation (14)
D. Neuro-symbolic reasoning	72 (2.6%)	44 (1.6%)	5	theorem proving (70), proof (31), mathematical reasoning (15)

The mutually exclusive assignment places 1,863 of 2,784 papers (66.9%) in a dominant stream; the remaining third carries no signature vocabulary and is best understood as generic AI4S work. We now interpret each stream through the four-layer framework. These interpretations combine the measured signature vocabulary with close reading of the stream's landmark systems; the landmark reading is the case component of the design (Section 3.1) and is separable from the counts.

4.5 Stream A: representation learning for scientific objects

The first stream concerns learned representations of objects that previously required highly specialized feature engineering: proteins, molecules, crystals, reaction networks, scientific text, equations, and experimental traces. It connects graph neural networks, transformers, equivariance, geometric deep learning, protein folding, and property prediction. The common idea is that useful scientific inference depends on a representation that respects relevant symmetries, relations, and constraints. A generic language embedding may be helpful for literature mining, but it does not automatically encode conservation laws, stoichiometry, chirality, or physical geometry.

In the corpus, this stream is the second-largest: 757 papers (27.2%) are assigned to it as their dominant cluster and 487 (17.5%) carry at least two of its signature terms; *protein structure prediction* alone appears in 173 papers. Its signature g-index is 8. Annual growth is strong but the stream's attention share has stabilized: the share growth of *protein structure prediction* between 2014–2017 and 2024–2025 is 0.31 $\times$ —far below the corpus trend—while *AlphaFold* enters a state-2 episode only in 2023–

2025. Its query footprint is heavy: *protein structure prediction* × *deep learning* alone uniquely contributes 385 records (Section 4.1), so stream size is query-conditional.

Evidence regime (layer reading). A predicted structure is a *hypothesis*, not a solved structure. The Test layer requires independent checks, and the pre-specified protocol pre-registers high-, medium-, and low-confidence targets with orthogonal evidence (cryo-EM, crystallography, NMR, cross-linking mass spectrometry, mutagenesis, binding and enzyme assays, cellular functional tests). AlphaFold is the central example: its architecture incorporates evolutionary sequence information, template information where available, and geometric reasoning [3], and its calibrated per-residue confidence transforms output from a single asserted structure into a decision-support object. Yet the human-proteome study reports uneven confidence across residues and proteins [4], and model output cannot settle questions about complexes, ligands, alternative conformations, disordered regions, or physiological environment. The strongest use is therefore a coupled workflow in which prediction changes experimental prioritization while experiment revises biological interpretation. In terms of P1, representation cost is low and falling, so the bottleneck in this domain sits at the Test layer (experimental capacity), not at candidate generation.

4.6 Stream B: generative design and inverse problems

The second stream reverses the familiar prediction question. Rather than asking what property a known object has, inverse design asks which objects might satisfy a target property or function. It connects molecular generation, de novo design, retrosynthesis, Bayesian optimization, crystal prediction, and materials search. The relevant design spaces are combinatorial: molecules and materials include an enormous number of possible compositions, structures, and processes.

In the corpus this is the *largest* stream: 953 papers (34.2%) dominate on this cluster and 267 (9.6%) carry at least two signature terms; *inverse design* appears in 248 papers and *drug discovery* in 1,170. Its signature g-index of 12 is the highest of the four streams. It is also the fastest growing: the 2024–2025 share of *inverse design* is 6.1 × its 2014–2017 share, and *materials*, *materials discovery*, *crystal*, and *drug design* all grow 6.0–9.6 × (Section 4.9).

Evidence regime. GNoME exemplifies a scaled inverse-design and screening pipeline: Merchant et al. (2023) combined graph-network predictions with density-functional-theory filtering and reported hundreds of thousands of newly predicted stable materials alongside independently verified examples. The significance lies not in the size of the list but in connecting model-based filtering with physics-based calculation in a data flywheel. The pre-specified protocol requires separating predicted stability, successful synthesis, structural confirmation, measured performance, and scale-up feasibility, and allows “new material discovery” only after independent synthesis, structural characterization, and replicated performance tests. In P1 terms this stream is where the bottleneck has moved most visibly from proposal to synthesis and scale-up—the classic Test/Justification gap of materials informatics. In P3 terms, the stream’s dependence on large simulation corpora and compute concentrates capability in

well-resourced groups, a structural prediction the collaboration data are consistent with (Section 4.10) but cannot verify.

4.7 Stream C: closed-loop experimentation and self-driving laboratories

The third stream joins models with instruments. Self-driving laboratories combine machine learning, lab automation, robotics, experimental design, online measurement, and laboratory information systems [32, 33]. Their potential lies in closing the loop: an explicit decision policy chooses the next experiment after observing the preceding result, so the sequence from hypothesis to action to measurement to update is explicit, repeatable, and inspectable. The A-Lab illustrates a comparatively integrated form of autonomy: it combined an ab initio materials database, text-mined synthesis heuristics, machine-learning data interpretation, active learning, robotics, and powder synthesis for inorganic materials [34]; the mobile robotic chemist [35] is an earlier, widely cited precedent. The lab's failures are scientifically important data: an unachieved synthesis, a poor characterization signal, or a failed transfer must alter future decisions rather than disappear from the record.

In the corpus this stream is small but terminologically distinct: 81 papers (2.9%) dominate on it, 21 (0.8%) carry at least two signature terms, and its annual profile is strictly recent—essentially zero before 2019, growing to 8 dominant papers in 2025. Its query footprint is nearly unique: 93.6% of papers retrieved by the *autonomous laboratory* query and 93.5% of those retrieved by the *self-driving laboratory / autonomous experimentation* query appear in no other query, confirming that closed-loop experimentation is a coherent, isolated vocabulary. Autonomy is graded: an automated instrument performs a repeated task; an AI-assisted workflow makes recommendations; a closed-loop system selects next actions under an objective; a more ambitious autonomous system copes with uncertainty, changing constraints, and failure. These distinctions should be retained in bibliometric labels and policy discussion; calling all automation "robot scientists" obscures where responsibility, oversight, and interpretation remain human. In P2 terms, closed-loop experimentation makes the division of labour unusually inspectable—the machine records the objective, the action, and the outcome—and thereby shifts the justification burden onto instrument calibration and reproducibility.

4.8 Stream D: neuro-symbolic verification and mathematical reasoning

The fourth stream connects large language models, program synthesis, search, proof assistants, symbolic engines, and formal mathematics. It is central to AI4S because it makes a broad scientific distinction especially clear: candidate generation can be probabilistic, but verification may be rule-based and independently auditable. FunSearch combines a language model, a user-specified evaluator, and evolutionary search to discover improved programs [7]; the evaluator prevents a fluent but incorrect candidate from being retained solely because it appears plausible. AlphaGeometry combines a learned model that proposes useful auxiliary constructions with a symbolic deduction engine that derives and verifies conclusions, solving 25 of 30 Olympiad geometry problems under the stated conditions [6].

LeanDojo provides tools, data, models, and a benchmark for retrieval-augmented theorem proving in Lean [36], making proof success mechanically checkable.

In the corpus this stream is the smallest and the most stable: 72 papers (2.6%) dominate on it, 44 (1.6%) carry at least two signature terms, and its annual profile is flat (5 papers in 2014, 7 in 2025)—*theorem proving* (70 papers) and *mathematical reasoning* (15) show *no* burst episode. Two implications follow. First, the neuro-symbolic vocabulary is coherent and distinct (its query share is 73–88% unique), so formal mathematics is a genuine research front of AI4S rather than a noise artefact. Second, it is not yet an *emerging* front in attention terms: the strong 2024–2025 episodes belong to *large language model* and agentic/materials vocabulary, while theorem proving grows steadily without surging.

Evidence regime. Formal verification is the clearest verification primitive in AI4S: the pre-specified protocol requires that each decisive proof be checked in a named proof assistant with exact version, library dependencies, and reproducible scripts. Yet formal validity and mathematical significance remain distinct: validity can be mechanical, while significance—choosing worthwhile problems, inventing concepts, communicating insight, moving between informal and formal language—remains partly open and deeply social [2]. In P2 terms, formal mathematics is the extreme case where machine-executable verification makes the human–machine boundary fully explicit: machines propose, verifiers check, and humans decide which problems are worth solving.

4.9 Attention dynamics: citation bursts and their right boundary

Figure 4 shows Kleinberg burst episodes for 36 terms (of 46 probed with ≥ 15 papers that produced an episode) at $s = 2$, $\gamma = 1$. Three distinct attention regimes emerge.

Strongest episode. *Large language model* (125 papers) produces the single strongest episode in the corpus: state 2 in 2024 and state 3 in 2025, the highest burstiness level observed. *Transformer* and *graph neural network* (2023–2025, state 2) and *AlphaFold* and *generative design* (state 2 episodes in 2023–2025) follow. The 2024–2025 LLM episode coincides with the corpus's largest annual jump (555 $\rightarrow$ 736).

Right-boundary caution. Because 2025 is the right edge of the window, "state 3 in 2025" is a right-boundary signal, not an established trend. We therefore refit the automaton year by year (Table 4). For *large language model*, refitting through 2022 yields no burst (state 0); through 2023 it already yields state 2; through 2024 and 2025 it yields state 3. The episode is thus detectable one to two years before the window's end and is not a pure artefact of the final observation—but its duration at the maximum state is, so far, a single year. By contrast, *protein structure prediction*, which had reached state 2 by the 2021 cutoff, is at state 1 by every cutoff from 2022 onward: its surge was mid-window and subsided, exactly the pattern that right-boundary-only inspection would misread as "declining forever". *Theorem proving* and *self-driving laboratory* show no state-≥ 2 episode at any cutoff.

Right-boundary backtest: Kleinberg automaton refit on data through each cutoff year. Entry = burst state at the cutoff; an episode reaching state ≥ 2 is a strong burst.

Term	2019	2020	2021	2022	2023	2024	2025
large language model	0	0	0	0	2	3	3
protein structure prediction	1	0	2	1	1	1	1
inverse design	1	1	2	2	1	1	2
drug discovery	1	1	2	1	2	2	2
theorem proving	0	0	0	0	0	0	0
self-driving laboratory	0	0	0	0	0	0	0

Surge in design vocabulary. Share-based growth (2024–2025 vs. 2014–2017) is concentrated in matter design: *materials* 9.6 $\times$, *crystal* 6.1 $\times$, *inverse design* 6.1 $\times$, *materials discovery* 6.1 $\times$, *drug design* 6.0 $\times$, and *molecular* 5.5 $\times$. These are not artefacts of corpus growth: they outpace *machine learning* (1.67 $\times$), *deep learning* (1.39 $\times$), and *artificial intelligence* (3.75 $\times$).

Saturation of early vocabulary. *Protein structure prediction* (0.31 $\times$) and *neural network* (0.44 $\times$) show *declining* shares: the representation-learning stream matured, and its attention is being absorbed by the generative and LLM streams. This is the clearest structural signal of a field in transition.

Kleinberg burst episodes ($s = 2,\ \gamma = 1$); colour depth is burst state. The vertical line marks the retrieval boundary (2025); bars extending to the boundary are right-censored. *Large language model* shows the corpus's strongest episode (state 3 in the final year); *protein structure prediction* and *neural network* show no recent surge.

4.10 Collaboration and geography

Affiliation strings are present in 42.9% of records and resolve to a country in 37.9%, so country figures are conservative lower bounds (Fig. 5). The United States leads with 313 papers (11.2% of the corpus), followed by China with 289 (10.4%), India with 100 (3.6%), the United Kingdom with 66 (2.4%), Germany with 54 (1.9%), and Japan with 50 (1.8%).

International co-authorship. The earlier draft reported 223 internationally co-authored papers as "18.7% of affiliated papers." The denominator problem is that affiliation-bearing papers ($n = 1,194$) include records whose affiliation strings cannot be resolved to a country ($n = 139$); a rate computed over all affiliation-bearing papers silently mixes resolvable and unresolvable records. International and domestic counts are defined only over country-resolvable papers: 223 international + 832 domestic = 1,055. The international co-authorship rate is therefore *223 of 1,055 country-resolvable papers = 21.1%*; the 18.7% figure is a different quantity (223 of 1,194 affiliation-bearing papers) and is reported only with its

denominator spelled out. Figure 5 annotates both the resolved denominator and the unresolved coverage gap.

The country co-authorship network (53 nodes, 185 edges, $Q = 0.213$) separates into five communities whose strongest cross-national edges are China–United States (32 papers), United States–Canada (13), United States–United Kingdom (10), United States–Germany (8), and United States–India (8). Leading institutions include the University of Toronto (13), the University of Miami (12), the University of California system (11; multi-campus aggregate), Central South University (11), Shandong University (10), MIT (9), and Argonne National Laboratory (8).

These figures support a modest version of the infrastructure-concentration concern (P3): output is concentrated in a handful of countries, and the collaboration backbone is a small set of US-centred pairs. They illustrate the abstraction problem of the AI4S literature: a bibliometric map built on open metadata can locate *where* capability sits, but not *why*. The institutional parsing treats "University of Miami"-style strings literally, and affiliation coverage gaps (57.1% of records without affiliations) mean the map underrepresents countries and labs whose metadata are sparser.

Top-15 countries by affiliated publications. International co-authorship covers 223 of 1,055 country-resolvable papers (21.1%). Affiliation strings are present in 42.9% of the corpus and resolve to a country in 37.9%, so counts are over resolvable records and are conservative lower bounds.

4.11 Missingness and network robustness

Field coverage. Table 5 summarizes missingness. Three patterns matter. (i) *By year*, abstract coverage rises from 28.6% (2014) to a plateau around 51–56% (2020–2025); reference coverage is relatively stable at 82–87%. (ii) *By source*, abstract coverage varies between 0% (Materials & Design; Drug Discovery Today) and 100% (Digital Discovery; Molecules)—abstract missingness is a depositor-level property, so any analysis using abstracts silently drops whole journals. (iii) *By domain*, reference coverage is highest for the closed-loop stream (93.8%) and lowest for neuro-symbolic reasoning (73.6%), and abstract coverage is lowest for the unassigned remainder (41.7%).

Field coverage of the retrieved and screened corpus: percentage of records with abstract, reference list, and affiliation data, overall and by selected year, source, and measured stream.

Subset	Abstract	References	Affiliations
Overall ($N = 2,784$)	50.6	84.3	42.9
Year 2014 ($n = 21$)	28.6	66.7	38.1
Year 2025 ($n = 736$)	51.4	81.9	43.5
Digital Discovery ($n = 147$)	100.0	100.0	—
Materials & Design ($n = 116$)	0.0	99.1	—
Scientific Reports ($n = 82$)	70.7	100.0	—
Drug Discovery Today ($n = 74$)	0.0	100.0	—
Stream A (representation)	59.6	86.9	47.3
Stream B (generative/inverse)	50.5	78.5	43.0
Stream C (closed-loop)	58.0	93.8	49.4
Stream D (neuro-symbolic)	63.9	73.6	44.4
Unassigned	41.7	88.3	38.5

Network robustness (RQ4). Four checks were run on a single clean vocabulary so that all numbers share one term set (Appendix E). (i) *Seed stability*: Louvain over 20 seeds has a mean adjusted Rand index (ARI) of 0.82 (minimum 0.47)—the community partition is mostly stable but not unique; the earlier draft's "perfect stability (ARI 1.000)" was an artefact of a dirty vocabulary and is retracted. (ii) *Source variant*: a title-only vs. title+abstract co-occurrence network changes the partition substantially (ARI 0.762; Q 0.404 vs. 0.246). This is the single most important sensitivity: topic boundaries depend on which text fields are available, and abstract coverage is only 50.6%, so the title+abstract variant uses a non-random subset. We therefore base the main analysis on title-only co-occurrence (uniform coverage) and treat community boundaries as orientation, not as a stable taxonomy. (iii) *Query contribution* is reported in Section 4.1. (iv) *Citation trimming* is reported in Section 4.1.

4.12 Comparative evidence regimes

Across the five domains, AI4S capability is strongest when representation, proposal, testing, and justification form a documented loop (Table 6; Fig. 6 reads each stream through the four layers). Protein science has rich representations and confidence signals, while experimental validation remains heterogeneous. Materials science increasingly combines prediction with robotic testing, but synthesis

and scale-up are persistent bottlenecks. Drug discovery can use AI across many stages but has the longest and most socially consequential validation chain; the pre-specified protocol requires a staged pathway—target rationale, primary assay, orthogonal assay and counterscreen, ADMET and toxicity, relevant cellular and animal models, and where appropriate clinical and regulatory evaluation—and candidates are described as AI-prioritized until translational evidence is obtained. Formal mathematics has the clearest verification primitive, yet its coverage is restricted by formalization effort and by the difficulty of selecting valuable problems.

Cross-stream bridging (papers carrying at least one signature term of both streams, computed from the published signature dictionaries on titles+abstracts) is concentrated in two arcs: A–B 184, B–C 59, then C–D 37, A–C 28, B–D 9, and A–D 7. The field is therefore not four isolated silos but a connected landscape in which the generative stream (B) is the hub connecting representation learning (A) to experimentation (C). These values replace the bridging counts (190, 100, 43) of the earlier draft, which were not reproducible from the published signature dictionaries; the recomputed values are used throughout.

Comparative evidence regimes across focal domains.

Domain	Primary AI contribution	Strongest validation	Key residual uncertainty
Protein science	structure representation and prediction	independent structural and functional experiments	dynamics, complexes, cellular context
Materials	property/stability prediction and inverse design	synthesis and characterization	synthesis feasibility and scale-up
Drug discovery	target, molecule, and property prioritization	assays, preclinical and clinical studies	translation and causal efficacy
Closed-loop experimentation	autonomous experiment selection	robot-executed experiments	calibration and reproducibility
Mathematics	conjecture, program and proof search	formal proof checking	problem choice and informal insight

Evidence matrix: each measured stream realizes all four analytical layers (representation, proposal, test, justification); cells summarize the evidence regime of that domain. The grid replaces the earlier one-to-one layer-to-stream mapping.

5 Cross-domain evidence regimes: cases and evidence cards

5.1 Case selection

Cases were selected by three criteria: (i) centrality in the co-citation and burst analyses (AlphaFold, GNoME, A-Lab, AlphaGeometry, FunSearch all anchor measured fronts or episodes); (ii) domain coverage—at least one case per focal domain, with drug discovery deliberately added to balance the protein/materials pair; and (iii) verification-mechanism contrast—each case exercises a different configuration of the Test and Justification layers. The choice is documented as a selection log in the research package; it is a close-reading choice, not a random sample, and its limits are discussed in Section 6.5.

5.2 Evidence cards

Each case card records: object and input; candidate; test; independent verification; negative results/failure modes; reproducibility; residual uncertainty; and the strongest claim this paper can support. Full cards are in Appendix D; the drug case is given here at length because it is the most socially consequential and was underdeveloped in the earlier draft.

Drug discovery: INS018-055 (Rentosertib). Insilico Medicine's TNIK inhibitor is the most complete public chain of an AI-designed drug moving from target selection to clinical data. The target (Traf2- and NCK-interacting kinase, TNIK) was prioritized by an AI-based target-discovery platform from fibrotic datasets, and the molecule was generated *de novo* by a generative chemistry platform; Ren et al. report the full preclinical and Phase 1 path, with an 18-month reported target-to-candidate cycle and clinical progression in idiopathic pulmonary fibrosis [37]. A Phase 2a interim signal presented at the 2025 American Thoracic Society meeting reports favourable change in forced vital capacity for INS018-055 versus placebo [38]. The case is deliberately read through all four layers to make the balance point explicit: *generation* of the candidate (Proposal) was fast and increasingly cheap, but community acceptance (Justification) required target rationale, preclinical efficacy in relevant models, Phase 1 safety and pharmacokinetics, and—at the time of writing—a preliminary, not-yet-confirmatory Phase 2a lung-function signal. In P1 terms, drug discovery is the domain where the bottleneck shift is sharpest: cheap candidate generation has moved the constraint almost entirely to the staged validation chain, and an "AI-designed drug" claim is only as strong as the last completed stage of that chain.

Protein structure: AlphaFold. Candidate: predicted structures and confidence. Test: CASP/CAMEO benchmarks plus downstream structural biology. Independent verification: cryo-EM/crystallography where followed up. Residual uncertainty: dynamics, complexes, ligands, physiological context [3, 4]. Strongest supported claim: a decision-support object that re-allocates experimental effort; not a completed structural biology.

Materials: GNoME. Candidate: predicted stable crystals. Test: DFT filtering; independent synthesis for a subset. Residual uncertainty: synthesis feasibility, measured performance, scale-up [5]. Strongest supported claim: a large, partly verified hypothesis space; discovery in the strong sense requires synthesis and characterization.

Closed-loop: A-Lab. Candidate: synthesis recipes chosen by active learning. Test: robot-executed synthesis and characterization in the loop. Residual uncertainty: calibration, instrument reproducibility, generality of materials covered [34]. Strongest supported claim: an explicit, inspectable hypothesis–action–measurement loop; failures are recorded as data.

Mathematics: AlphaGeometry / FunSearch. Candidate: auxiliary constructions (AlphaGeometry) and programs (FunSearch). Test: symbolic deduction / user-specified evaluators. Residual uncertainty: formal validity vs. mathematical significance [6, 7]. Strongest supported claim: machine-checkable validity with human-judged significance; the fullest realization of P2.

5.3 Claim–evidence alignment

Table 7

aligns every major claim of this article with its evidence, verdict, and location. The distinction between measured findings, case interpretation, theoretical propositions, and normative recommendations is deliberate and enforced throughout the Discussion. Claim–evidence alignment matrix. "Verdict" rates how far the reported evidence supports each claim; normative items are flagged as such.

Claim	Evidence	Verdict	Location
AI4S literature is large and fast-growing	screened counts; Poisson IRR 1.381 (CI 1.361–1.402); Crossref-share rise	Supported (descriptive)	§ 4.1
Growth decelerates	segment IRRs 1.537 vs. 1.320	Supported	§ 4.1
Four measured streams	term network $Q = 0.404$; signature dictionaries	Conditionally supported (field-sensitive)	§ 4.2, § 4.4
Generative/inverse design is the hub	cross-stream bridging (A–B 184; B–C 59)	Partially supported (overlap, not centrality)	§ 4.12
LLM is the strongest recent attention signal	burst state 3 in 2025; backtest onset 2023	Supported for the observed window	§ 4.9
International co-authorship rate	223/1,055 = 21.1% (country-resolvable)	Supported	§ 4.10
Unequal infrastructure (P3)	country concentration; US-centred edges	Weak indirect evidence; "consistent with"	§ 4.10, § 6
Epistemic labour redistribution (P1–P3)	cross-domain synthesis + theory	Theoretical proposition	§ 2, § 5, § 6
Normative governance recommendations	framework + literature	Research agenda, not a result	§ 6.6

6 Discussion

6.1 Three empirical findings

Three findings are robust across the sensitivity checks. First, AI4S is a large, fast-growing literature ($N = 2,784$, Poisson IRR 1.381) that is nevertheless *decelerating* (segment IRRs 1.537 $\rightarrow$ 1.320) and remains a tiny share of Crossref output (0.0111% in 2025). Second, its vocabulary is coherent ($g = 155$, $Q = 0.404$) and organized around four streams—representation learning, generative and inverse design, closed-loop experimentation, and neuro-symbolic reasoning—that are empirically grounded but unevenly developed, with the generative stream dominating in volume and growth and acting as the hub connecting representation learning to experimentation (A–B 184, B–C 59 bridging papers). Third, attention dynamics distinguish three regimes: a saturated early vocabulary (protein structure prediction, neural network), a surging design vocabulary (inverse design, materials, drug design), and a right-boundary frontier dominated by large language models that a year-by-year backtest confirms is not an artefact of the last observation.

6.2 Relation to prior work

Our science map of AI4S is consistent with, but distinct from, existing accounts. Domain reviews of protein structure prediction, materials informatics, and drug discovery each describe the same field-level pattern we measure—cheap prediction, hard validation [3, 5, 37]. Recent scientometric work has begun to apply language models and knowledge-graph interfaces to bibliometric questions [10], but the combination we advance—an audited open-metadata screen, a quantified four-stream map, and a cross-domain evidence-regime reading—is, within the bounds of our search (documented in the research package), a distinctive synthesis rather than a replication. The epistemic-labour framing connects the bibliometric patterns to a long line of work on the division of cognitive labour in science [11] and to human-aware models of discovery [39]; our contribution is to make that connection empirical and domain-comparative.

6.3 Mechanism: the redistribution of epistemic labour

The account resists a substitution narrative. Human scientific labour shifts toward problem formulation, constraint specification, data stewardship, experiment design, anomaly interpretation, and responsibility for claims [2]. In many consequential domains the most difficult work begins *after* a model produces a high-scoring candidate: a medicinal chemist must decide whether a result expresses a genuine mechanism or an assay artefact; a materials scientist must interpret a failed synthesis and redesign a process; a biologist must connect a structural hypothesis to a living system; a mathematician must decide whether a formal result matters. The five domains differ precisely in *where* that human labour sits (Table 6).

The measured cross-stream bridges (A–B 184, B–C 59) indicate that this coupling is already reflected in the literature's vocabulary: representation learning feeds generative design, and generative design feeds experimentation. The framework makes the division explicit. A better representation improves candidate generation; a better test environment improves training data; a better provenance system makes both critique and reuse possible.

The social distribution of capability deserves particular attention [39]. High-quality scientific data are unevenly available; compute and robotic infrastructure are similarly concentrated [2]. The measured country concentration (US-led, a small set of cross-national pairs) is a bibliometric footprint of that unevenness, albeit a partial one given affiliation coverage. We stress that this is *consistent with* cumulative advantage (P3), not evidence of its causal operation; establishing the mechanism requires organizational and funding data we do not have.

6.4 Alternative interpretations

Three alternatives deserve explicit consideration. First, the map may reflect *metadata and indexing regimes* more than the intellectual structure of the field: Crossref's coverage skews by depositor, abstract availability varies by journal from 0% to 100%, and the title-only term construction depends on titling conventions. The ARI 0.762 field-sensitivity result makes this the strongest rival account, and it is why we bound every claim to the screened open-metadata corpus. Second, the four streams may be an artefact of the query set: leave-one-query-out shows stream sizes are query-conditional (Section 4.1), though no single query generates all four streams. Third, the "generative hub" reading could reflect query design (generative terms appear in several queries) rather than an intellectual structure; the bridging analysis uses published signature terms rather than query provenance, but it remains a co-occurrence overlap measure, not a causal structure.

6.5 Method boundaries and the limits of the map

Five limitations qualify the results.

(1) Source adaptation. The pre-specified protocol specified WoSCC; this study executed on Crossref. Crossref is a registration agency, not a citation index: coverage is more inclusive but differently skewed, it lacks author keywords, and its abstract (50.6%), reference (84.3%), and affiliation (42.9%) coverage are partial and non-random (Table 5). Every statistic is a property of the Crossref corpus as screened; direct comparison with WoSCC-based maps requires re-running the pipeline on a WoSCC export.

(2) Audit substitution. Cohen's kappa between two human reviewers was not available; a dual-seed audit of the frozen rulebook by the same analyst replaces it (precision 97.7% strict / 92.4% including borderline; false exclusion 2.0% / 9.0%). The rulebook is deterministic and auditable, which is an advantage; but it is not a second reviewer, and the analyst's independence from the rulebook is not established. Residual error types cluster in specific categories: industry news and NLP recommender/preprocessing papers entering on title-level vocabulary; clinical diagnostics and structural-engineering characterization passing as materials/drug discovery; over-broad noise terms *multi-agent* and *engineering design* removing genuine AI4S titles; and five of seven borderline false-exclusions being records with no abstract.

(3) Network sensitivity. Louvain seed stability is moderate (mean ARI 0.82, min 0.47), and a title-only vs. title+abstract network changes the partition substantially (ARI 0.762; Q 0.404 vs. 0.246). The main analysis therefore uses title-only co-occurrence (uniform coverage) and community boundaries are read

as orientation, not a stable taxonomy. The earlier draft's claim of perfect seed stability and an ARI of 0.297 between variants is retracted: both figures were computed on an unclean vocabulary and are replaced by the recomputed values here.

(4) Case selection. The case synthesis emphasizes highly visible systems (AlphaFold, GNoME, A-Lab, INS018-055, AlphaGeometry, FunSearch). That focus clarifies mechanisms but may underrepresent infrastructure work in curation, standards, lab software, instrumentation, and regional research communities. Future work should pair maps with ethnographic observation of laboratories, interviews with researchers, and longitudinal analysis of project portfolios, asking whether AI changes what scientists choose to investigate, not simply how quickly they publish.

(5) Negative evidence. Discovery evaluation should record negative evidence. Systems trained only on successes may repeat failed routes, overestimate confidence, and create distorted benchmarks. Community repositories for negative results, standardized experiment logs, uncertainty-aware active learning, and pre-specified validation plans can correct this bias. Finally, the rise of agentic systems requires more demanding evaluation: demonstrations on toy or retrospectively curated tasks are not enough. High-stakes deployment should preserve complete lab logs, maintain a transparent objective and stopping rule, express uncertainty, allow meaningful human override, prevent unsafe actions, and be independently replicated. The appropriate ambition is not an unaccountable "AI scientist", but a reliable system that helps diverse scientific communities ask better questions and test them more rigorously.

6.6 Governance as a research agenda

The pre-specified protocol required the paper to address uncertainty, human oversight, and governance. This study deploys no automated system, but its framework implies reporting standards that we present as a research agenda rather than as results. **Uncertainty.** Predictive uncertainty and calibration should be reported under both hold-out and distribution-shift conditions; the literature's own example is AlphaFold's per-residue confidence. **Decision thresholds.** Any deployed pipeline should state decision thresholds, the costs of false positives and negatives, escalation criteria, and human authority to halt an action, and should maintain candidate-selection and laboratory-action logs so that any conclusion can be reconstructed. **Governance.** A safety, dual-use, privacy, equity, environmental, and cost assessment should precede deployment: identify hazardous capabilities and misuse paths; use action constraints, approval gates, red-team testing, audit logs, and staged release where appropriate; and record compute, energy, materials, failed experiments, and maintenance effort against realistic conventional baselines. None of these can be verified bibliometrically; they are properties of deployed systems, and this article's contribution is to make them part of the evidentiary standard rather than an afterthought.

Science maps should also be interpreted modestly. A citation burst denotes heightened attention, not correctness. A high-centrality paper may be a methodological bridge, a generic review, or an artefact of broad indexing. Co-citation indicates intellectual association, not direct agreement or methodological compatibility. The point of visual mapping is to guide close reading and transparent comparison, not to automate judgment about a field's quality.

7 Conclusion

AI4S is moving science toward connected cycles of representation, proposal, test, and justification. The audited corpus shows a large, fast-growing, coherent literature whose four measured streams are empirically grounded but unevenly developed, with generative and inverse design as the hub and large language models as the strongest recent attention signal at the right boundary of the observation window. AlphaFold, materials foundation models, autonomous laboratories, AI-enabled drug pipelines, and formal mathematical systems display different portions of this cycle rather than a single universal capacity; the drug case (INS018-055) makes the central lesson concrete, that candidate generation is not clinical validation. A credible science map should describe its screened corpus honestly, quantify its sensitivity to metadata and query choices, and avoid converting attention metrics into evidence of discovery. The durable promise of AI4S lies in accountable human–AI inquiry: computational abundance paired with rigorous verification, inclusive infrastructure, traceable evidence, and clear responsibility for scientific claims. The redistribution of epistemic labour—toward validation, provenance, and judgment—is both the measured signature of the field and the agenda for its governance.

A Search queries, data card, and research package

A.1 Query audit table

The 16 Crossref keyword-combination queries (each AI term $\times$ discovery term), with corpus papers retrieved, unique papers, and unique share. All queries were run with $rows = 500$ maximum per page, cursor paging, relevance sorting, and the archived query log recorded per-query parameters, date, and time zone.

Query	Papers	Unique	Unique %
scientific discovery × artificial intelligence	253	76	30.0
drug discovery × artificial intelligence	414	236	57.0
scientific discovery × large language model	105	55	52.4
drug discovery × large language model	69	22	31.9
scientific discovery × machine learning	338	100	29.6
scientific discovery × deep learning	178	68	38.2
materials discovery × machine learning	433	197	45.5
drug discovery × deep learning × generative	287	89	31.0
materials discovery × generative design × deep learning	304	147	48.4
autonomous laboratory × machine learning	78	73	93.6
self-driving laboratory × autonomous experimentation	31	29	93.5
inverse design × machine learning × molecules × materials	355	308	86.8
protein structure prediction × deep learning	426	385	90.4
mathematical reasoning × neural network × language model	57	50	87.7
theorem proving × large language model	67	49	73.1
protein structure prediction × neural network	256	217	84.8
A.2 Data card (Crossref)			

Field	Entry
Owner	Crossref (community-controlled scholarly metadata).
Licence	Public metadata; attribution required.
Sampling frame	16 queries (Appendix A), English, 2014–2025, Article and Review only.
Instrument	Crossref REST API (polite-pool user agent + contact e-mail); harvest completed August 2026; per-query date and parameters archived in the query log.
Processing	DOI deduplication; frozen rules-based screen screening.py v5.16; version-cluster audit.
Missingness	Abstract 50.6%; reference lists 84.3%; affiliations 42.9% (by year/source/stream in Table 5).
Known bias	Registration-agency coverage; English-only queries; no author keywords; $rows = 500$ per-query cap; affiliation coverage skews to well-indexed institutions.
Update date	Corpus frozen at the logged harvest date; analysis versioned.
Permitted reuse	Attribution; code and verdicts archived with the manuscript.
A.3 Research-package contents	

Screening rulebook (screening.py v5.16) and flow counts; analysis scripts (descriptive.py, keywords.py, cocitation.py, bursts.py, collab.py, robustness.py, revision_analysis.py) and their versioned outputs; the archived Crossref query log (per-query URL parameters, harvest dates); audit verdicts for the included (seeds 42 and 7) and excluded (seed 42) samples with sampling code; the case-selection log; term-cleaning and signature dictionaries (Appendix E); network files (term, co-citation, signature, country); and this manuscript's LaTeX sources. Software versions and the environment are locked in the package manifest so that the full pipeline can be re-run from the raw query responses.

A.4 Evidence cards

Field	Entry
Drug: INS018-055	Object: TNIK kinase. Input: fibrotic expression data. Candidate: de novo TNIK inhibitor (generative platform). Test: preclinical IPF models; Phase 1; Phase 2a. Independence: company-reported; Phase 2a is an interim, not-yet-confirmatory signal [37, 38]. Failure modes: candidate-to-candidate attrition; clinical translation risk. Residual uncertainty: powered confirmatory efficacy; long-term safety. Strongest claim: candidate generation can be fast and cheap; acceptance requires the full staged chain.
Protein: AlphaFold	Object: protein structure. Input: sequence + MSA + templates. Candidate: predicted structure + per-residue confidence. Test: CASP/CAMEO; downstream structural biology. Independence: experimental structures where followed up. Residual uncertainty: dynamics, complexes, ligands, context [3, 4]. Strongest claim: decision-support object re-allocating experimental effort.
Materials: GNoME	Object: crystal stability. Input: crystal graphs; DFT labels. Candidate: predicted stable crystals. Test: DFT filtering; subset independently synthesized. Residual uncertainty: synthesis feasibility, performance, scale-up [5]. Strongest claim: large, partly verified hypothesis space; discovery requires synthesis.
Closed-loop: A-Lab	Object: inorganic materials synthesis. Input: databases, text-mined recipes, spectra. Candidate: synthesis recipes. Test: robot-executed synthesis in the loop. Residual uncertainty: calibration, instrument reproducibility, domain breadth [34]. Strongest claim: explicit inspectable loop; failures recorded as data.
Mathematics	Object: geometry problems (AlphaGeometry); programs (FunSearch). Input: problem statements; search space. Candidate: constructions; programs. Test: symbolic deduction; user-specified evaluator [6, 7]. Residual uncertainty: significance and problem choice. Strongest claim: machine-checkable validity with human-judged significance.
A.5 Term dictionary and cleaning	

Terms are clean n-grams ($n = 1 \ldots 3$) from titles and abstracts after lower-casing, punctuation removal, and stop-word filtering. A strong-unigram whitelist retains content-bearing single words; abbreviation tails and spurious URL fragments (e.g. “https github com”, “time consuming”) are suppressed, and the five isolated top-80 terms removed from the network are listed in Section 3.5. Stream signature dictionaries used for dominant-cluster assignment are archived with the research package (Appendix C).

Declarations

Declarations

Competing interests.

The authors declare no competing interests.

Ethics approval and consent to participate.

Not applicable. This is a bibliometric study of publicly available published metadata; no human participants, human data, or animals were involved, and no ethical approval was required.

Consent for publication.

Not applicable.

Clinical trial number.

Not applicable.

Funding.

This work received no specific grant.

Author Contribution

Author contributions: X.F. conceived and designed the study, developed the screening rulebook and analysis pipeline, performed the data analyses, and wrote the main manuscript text. L.G. curated and validated the corpus data and prepared all figures and tables. Z.N. contributed to data collection and verification of the screened corpus, and co-revised the manuscript. W.X. contributed to the literature review and interpretation of the results, and reviewed the manuscript. All authors read and approved the final version.

Data Availability

Data availability statement: All data used in this study are publicly available bibliographic metadata retrieved from the Crossref REST API (https://api.crossref.org). The complete set of search queries, the frozen screening rulebook (v5.16), the two-stage screening and audit outputs, and the screened corpus of 2,784 records are documented in Appendices A–C of the manuscript and are included as supplementary files with this submission. All derived datasets (term co-occurrence network, co-citation matrix, burst series, growth-model counts) and the full analysis code that reproduces every figure and table are likewise provided in the supplementary research package. No new experimental data were generated; no third-party or restricted data were used.

Code availability.

The complete analysis pipeline is archived with the manuscript and reproduces every figure and table from the screened corpus.

References

1. Krenn M, Landgraf T, Foesel T, Marquardt F. On scientific understanding with artificial intelligence. Nat Reviews Phys. 2022;4:761–9. https://doi.org/10.1038/s42254-022-00518-3.
2. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z. others: Scientific discovery in the age of artificial intelligence. Nature. 2023;620:47–60. https://doi.org/10.1038/s41586-023-06221-2.
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O. others: Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.
4. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A. others: Highly accurate protein structure prediction for the human proteome. Nature. 2021;596:590–6. https://doi.org/10.1038/s41586-021-03828-1.
5. Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624:80–5. https://doi.org/10.1038/s41586-023-06735-9.
6. Trinh TH, Wu Y, Le QV, He H, Luong T. Solving olympiad geometry without human demonstrations. Nature. 2024;625:476–82. https://doi.org/10.1038/s41586-023-06747-5.
7. Romera-Paredes B, Barekatain M, Novikov A, Balog M, Kumar MP, Dupont E. others: Mathematical discoveries from program search with large language models. Nature. 2024;625:468–75. https://doi.org/10.1038/s41586-023-06924-6.
8. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inform Sci Technol. 2006;57:359–77. https://doi.org/10.1002/asi.20317.
9. Chen C. Science mapping: A systematic review of the literature. J Data Inform Sci. 2017;2:1–40. https://doi.org/10.1515/jdis-2017-0006.
10. Saarela M, Correia A, Kärkkäinen T. Explainable and interactive scientometrics with large language models and knowledge graphs. In: 2025 9th international symposium on multidisciplinary studies and innovative technologies (ISMSIT). pp. 1–6 (2025). https://doi.org/10.1109/ismsit67332.2025.11268107
11. Kitcher P. The division of cognitive labor. J Philos. 1990;87:5–22. https://doi.org/10.2307/2026796.
12. Hendricks G, Tkaczyk D, Lin J, Feeney P. Crossref: The sustainable source of community-owned scholarly metadata. Quant Sci Stud. 2020;1:377–89. https://doi.org/10.1162/qss_a_00022.
13. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46. https://doi.org/10.1177/001316446002000104.
14. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008, P10008 (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008

15. Kleinberg J. Bursty and hierarchical structure in streams. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. pp. 91–101 (2002). https://doi.org/10.1145/775047.775061
16. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218. https://doi.org/10.1007/BF01908075.
17. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846–50. https://doi.org/10.1080/01621459.1971.10482356.
18. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G. others: Applications of machine learning in drug discovery and development. Nat Rev Drug Discovery. 2019;18:463–77. https://doi.org/10.1038/s41573-019-0024-5.
19. Sánchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018;361:360–5. https://doi.org/10.1126/science.aat2663.
20. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discovery Today. 2018;23:1241–50. https://doi.org/10.1016/j.drudis.2018.01.039.
21. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. https://doi.org/10.1038/nature14539.
22. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235.
24. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402. https://doi.org/10.1093/nar/25.17.3389.
25. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Mater. 2013;1:011002. https://doi.org/10.1063/1.4812323.
26. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
27. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559:547–55. https://doi.org/10.1038/s41586-018-0337-2.
28. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. Phys Rev Lett. 1996;77:3865–8. https://doi.org/10.1103/PhysRevLett.77.3865.
29. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4:268–76. https://doi.org/10.1021/acscentsci.7b00572.
30. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28:31–6. https://doi.org/10.1021/ci00057a005.

31. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50:742–54. https://doi.org/10.1021/ci100050t.
32. Abolhasani M, Kumacheva E. The rise of self-driving labs in chemical and materials sciences. Nat Synthesis. 2023;2:483–92. https://doi.org/10.1038/s44160-022-00231-0.
33. Pyzer-Knapp EO, Pitera JW, Staar PWJ, Takeda S, Laino T, Sanders DP. others: Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater. 2022;8:84. https://doi.org/10.1038/s41524-022-00765-z.
34. Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D. others: An autonomous laboratory for the accelerated synthesis of inorganic materials. Nature. 2023;624:86–91. https://doi.org/10.1038/s41586-023-06734-w.
35. Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X. others: A mobile robotic chemist. Nature. 2020;583:237–41. https://doi.org/10.1038/s41586-020-2442-2.
36. Yang K, Swope A, Gu A, Chalamala R, Song P, Yu S, Godil S, Prenger RJ, Anandkumar A. LeanDojo: Theorem proving with retrieval-augmented language models. Adv Neural Inf Process Syst. 2023;36. https://doi.org/10.52202/075280-0944.
37. Ren F, Aliper A, Chen J, Zhao H, Thiagarajan K. others: A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. Nat Biotechnol. 2024;43:63–75. https://doi.org/10.1038/s41587-024-02143-0.
38. Xu R, Wang C, Cao H, Tan G. others: INS018-055, a novel Traf2- and NCK-interacting kinase (TNIK) inhibitor, improves lung function in patients with idiopathic pulmonary fibrosis: Interim results from a phase 2a study. American Journal of Respiratory and Critical Care Medicine. 211, A2904 (2025). https://doi.org/10.1164/ajrccm.2025.211.abstracts.a2904
39. Sourati J, Evans JA. Accelerating science with human-aware artificial intelligence. Nat Hum Behav. 2023;7:1682–96. https://doi.org/10.1038/s41562-023-01648-z.

Unsectioned Paragraphs

AI use. AI tools were used for data analysis, visualization, and drafting assistance; the authors remain responsible for all analyses and claims.

A.2 Data card (Crossref)

A.3 Research-package contents

A.5 Term dictionary and cleaning

Figures

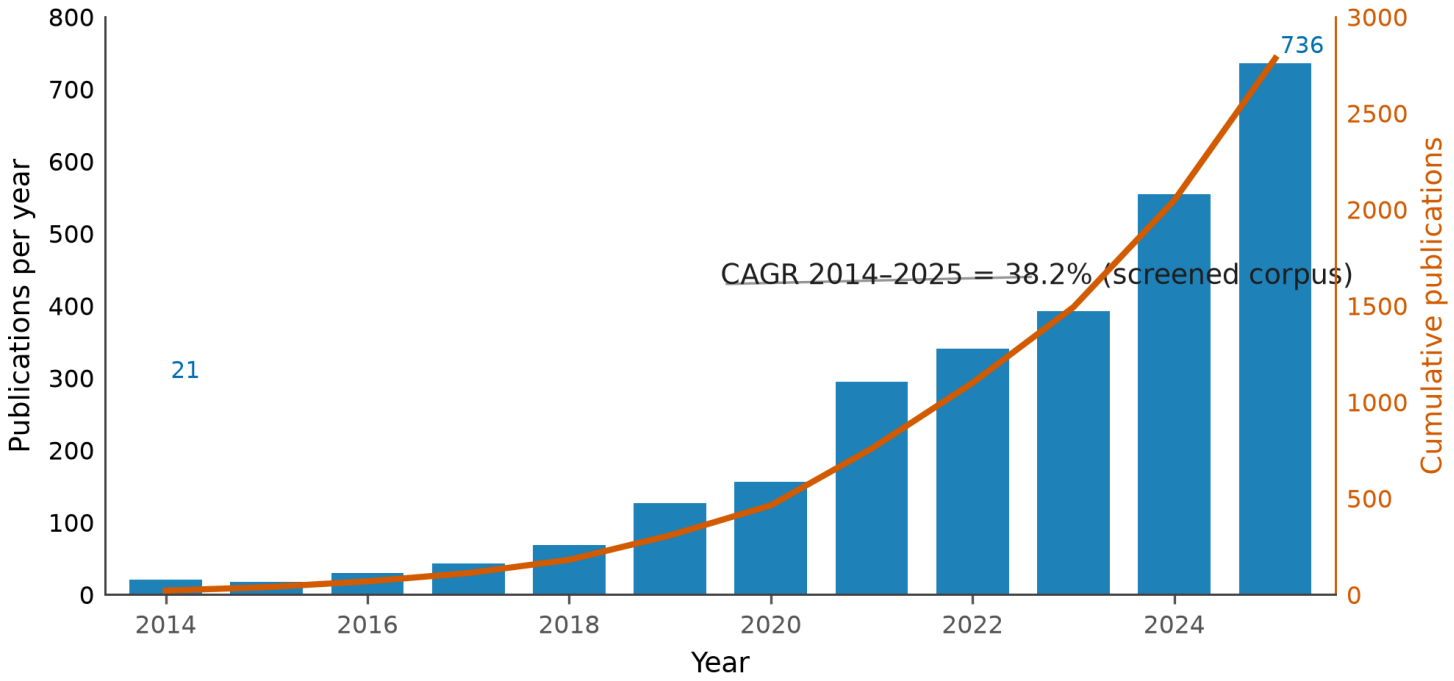


Figure 1

Legend not included with this version.

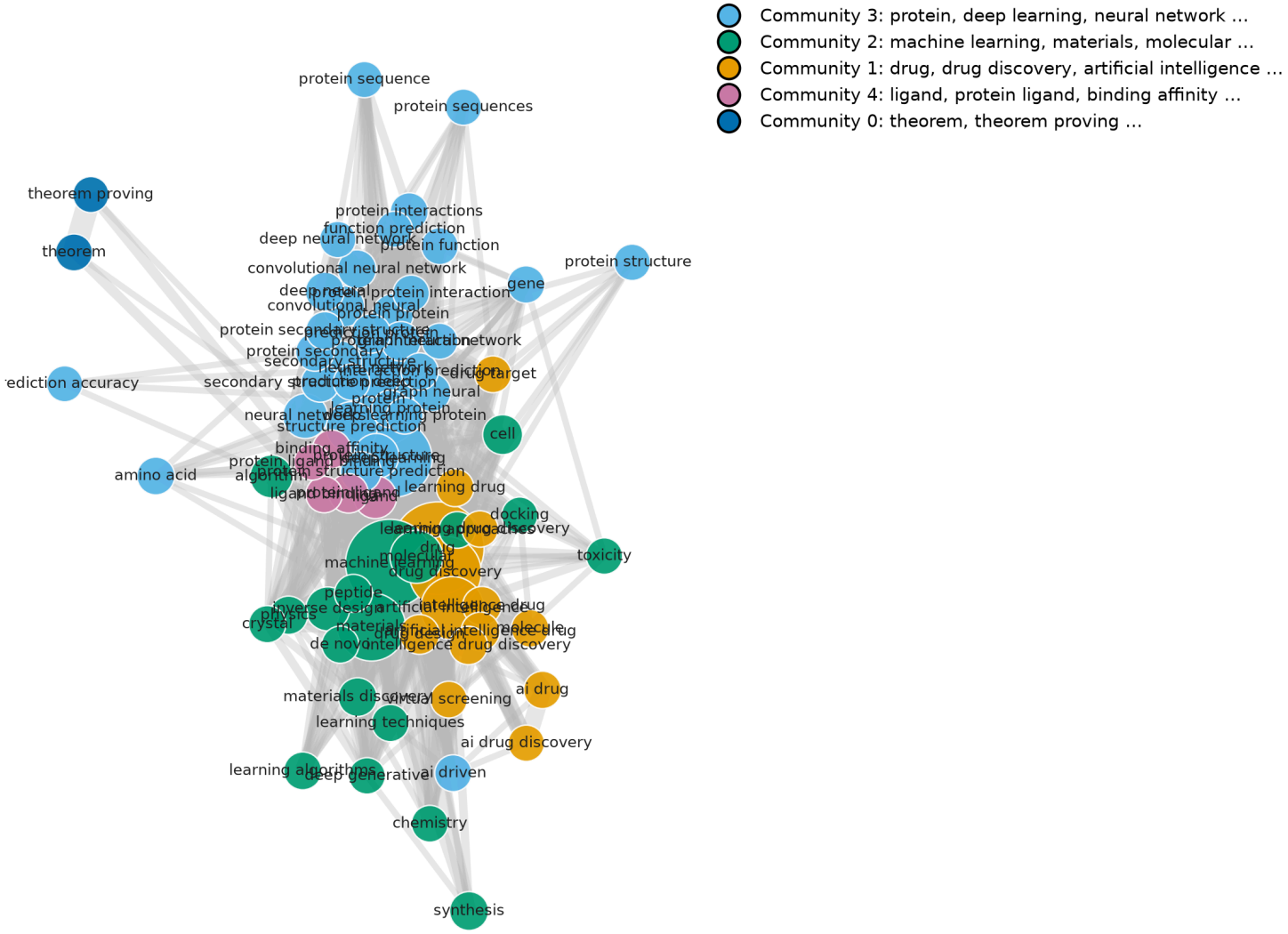


Figure 2

Legend not included with this version.

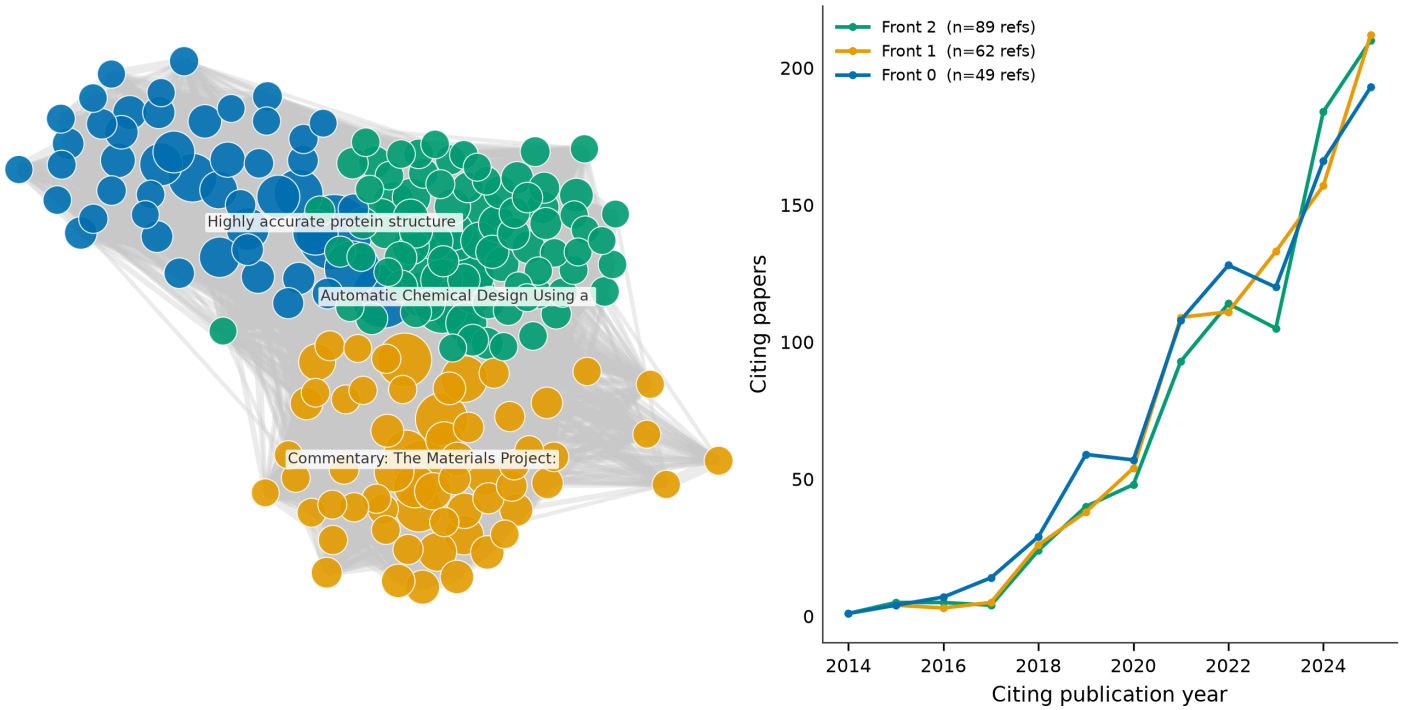


Figure 3

Legend not included with this version.

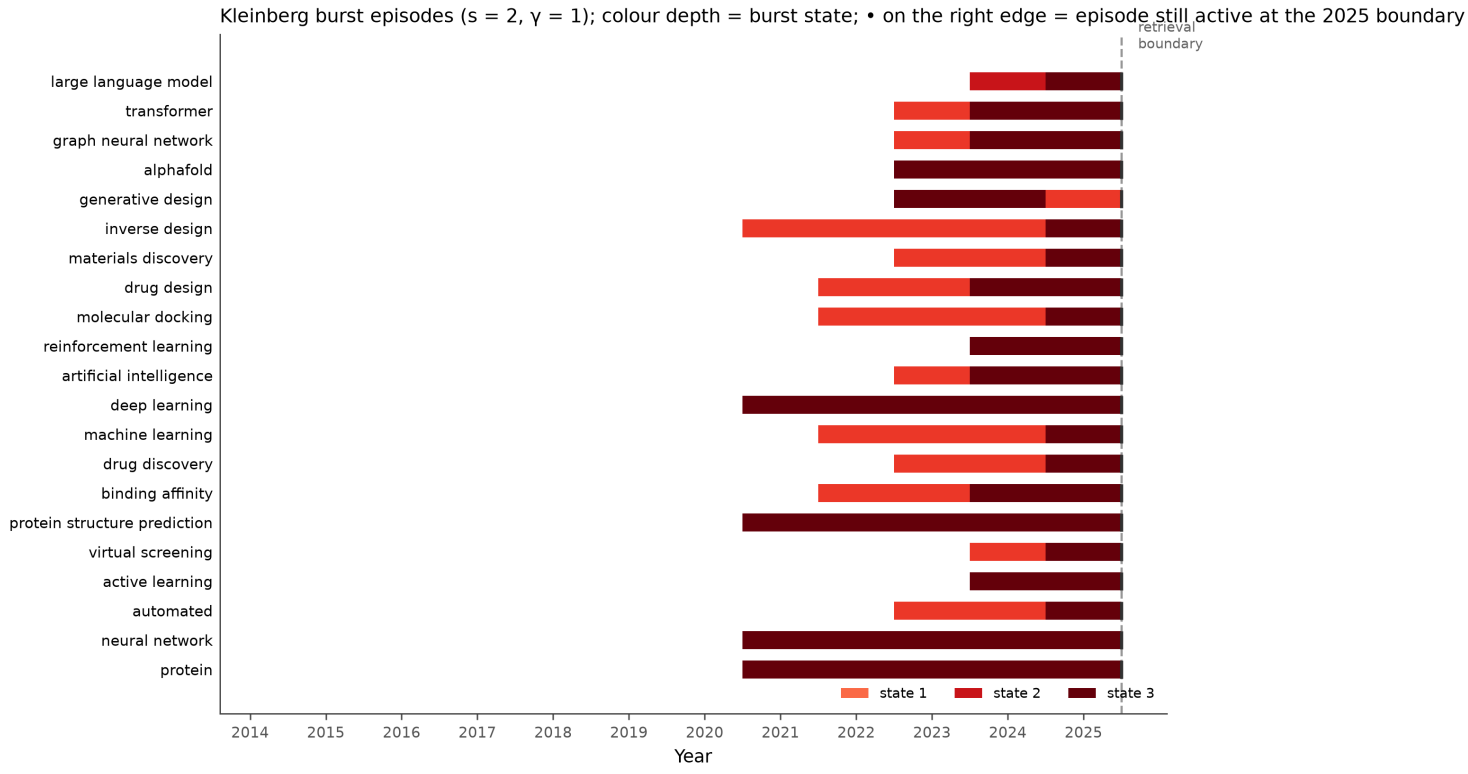


Figure 4

Legend not included with this version.

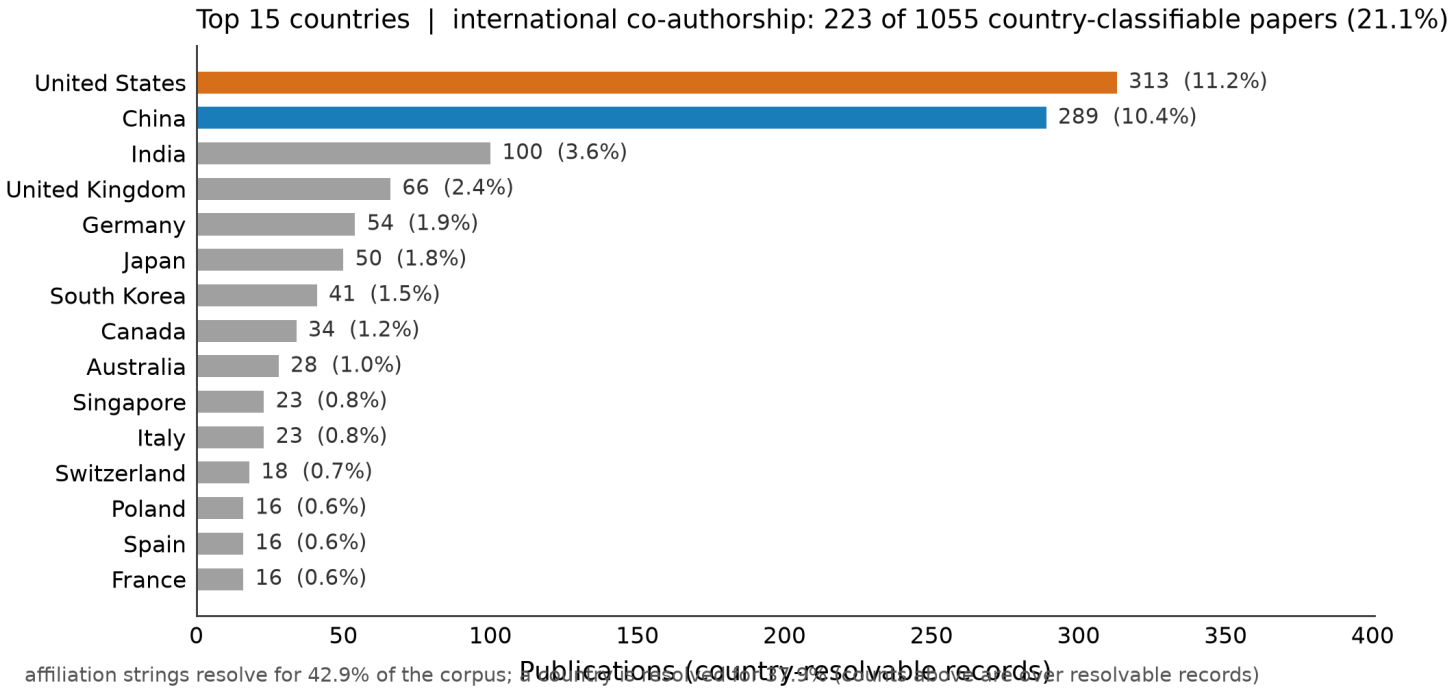


Figure 5

Legend not included with this version.

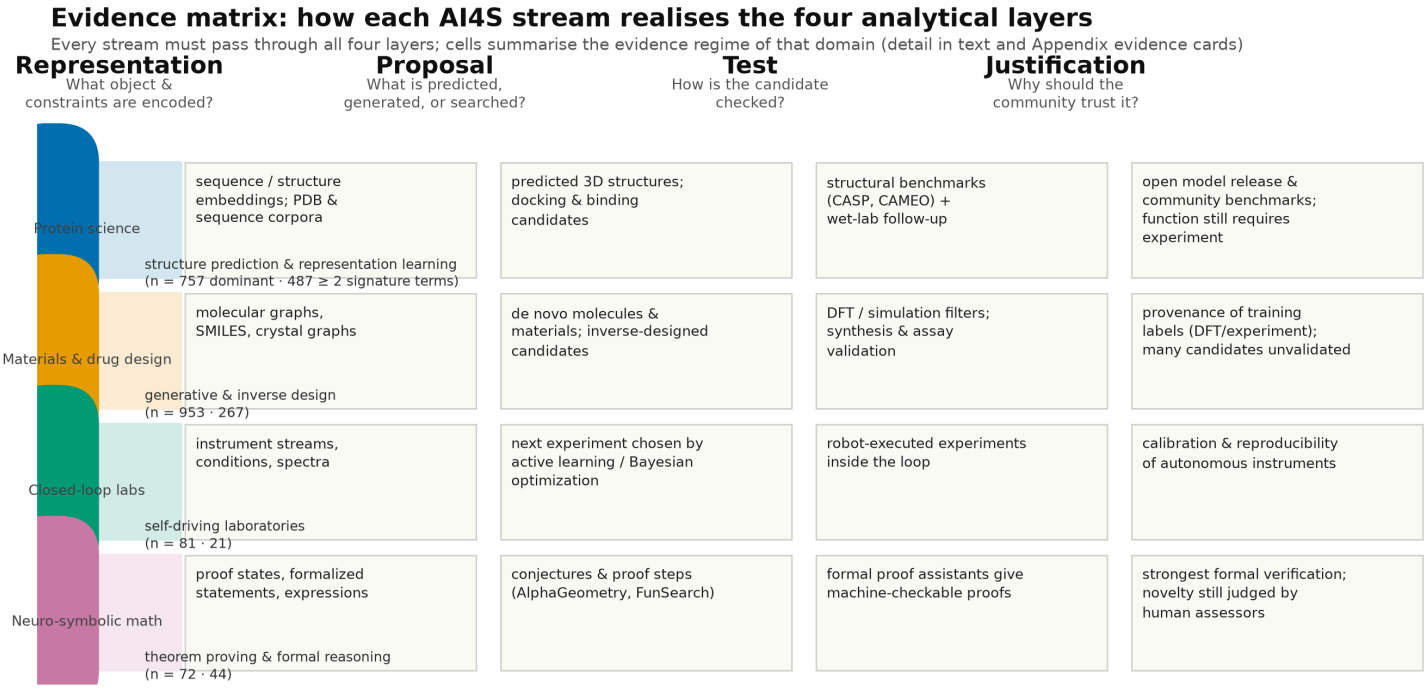


Figure 6