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.

PaperAtlas: an automatically constructed atlas of computational methods and software from 6.4 million open-access articles

Neeti Shah
Bhargava Systems Research Inc
Anagha Vippagunta
Bhargava Systems Research Inc
Yash Bhargava
yash@bhargavaresearch.org

Bhargava Systems Research Inc

Research Article

Keywords: Bioinformatics tool discovery, Literature mining, Software registries, Text mining, Large language models

Posted Date: October 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11143526/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

Computational methods and software are dispersed across a literature that is increasingly difficult to review manually, while curated registries capture only a subset of available resources. We present PaperAtlas, an automatically constructed atlas derived from the PubMed Central open-access corpus. Of 6,446,741 abstracts screened, 1,074,191 were classified as computational, with schema-valid records extracted for 1,074,140. Among these, 267,893 papers describing an algorithm, software package or web server formed 1,438 clusters, with 1,000 retained after restriction to parent-level biomedical categories. In 296 clusters linked to at least five bio.tools entries, the mean concentration of the most frequent EDAM topic was 61.7%, compared with 26.4% under permutation. Of 31,180 distinct software and web-server names, 61.0% lacked a strict match in bio.tools, PyPI, CRAN, Bioconductor or Bioconda. Among 12,207 bio.tools entries with defining papers in the corpus, 56.6% were recovered end to end, increasing to 83.6% among the 8,265 entries whose defining publication entered the atlas. PaperAtlas is fully open source and available as a web server at bhargavaresearch.org/paperatlas.

Introduction

Computational methods are integral to most areas of biological research, and they are published at an increasing rate, which makes them harder to collect and organize (Bornmann et al., 2021). Methods are scattered across subfields, named inconsistently, and often reported in papers whose title presents only a biological result. Retrieval systems, from lexical whole-MEDLINE ranking (Fontaine et al., 2009) to machine-learning recommenders (Allot et al., 2021), rank papers by relevance to a query.

A parallel effort extracts software from the literature, motivated by the finding that only 31–43% of software mentions carry a formal citation (Howison & Bullard, 2016). Softcite supplied a machine-learning-ready collection (Du et al., 2021), SoftwareKG extended extraction to more than 3 million PubMed Central (PMC) articles (Schindler et al., 2022), and the largest effort spans roughly 20 million papers (Istrate et al., 2022). A survey of 713,634 PMC full texts found 70% of resource names mentioned exactly once (Duck et al., 2016). On the curation side, bio.tools is hand-curated (Ison et al., 2019) and FAIRsoft aggregates 43,987 software instances across eleven sources (Martín del Pico et al., 2024). Such extractions are used to identify under-recognized packages (Brown et al., 2026), and related resources annotate the literature at comparable scale (Wei et al., 2024). Registry coverage has been estimated within individual subfields, such as synthetic biology (Cai et al., 2023). In contrast to mention extraction, PaperAtlas is an effort to use large language models (LLMs) to identify papers in which computation is itself a contribution, extract the artifact each contributes and organize these contributions into navigable topics.

We built PaperAtlas to organize the computational contributions of the open-access literature and to ask the following questions (Fig. 1). What computational methods and software are introduced in the literature and how are they distributed across topics? Among contributions of the kind a software

registry can hold, how often do they match registry records, and how many registered tools does an automated LLM pipeline recover?

Materials and methods

Corpus acquisition

We parsed 7,283,594 articles from the PMC bulk Journal Article Tag Suite (JATS) XML distribution (National Information Standards Organization, 2024; Sayers et al., 2026). Each record retained identifiers, bibliographic metadata, sectioned abstract and body text, and figure and table captions. Of these, 6,446,741 had both a title and an abstract and were screened, and the subset excludes methods published only in subscription journals.

Model and encoder selection

Screening and extraction both used Qwen2.5-Instruct checkpoints (Qwen Team, 2024). The model family was chosen as it is openly licensed, released from 3B to 32B parameters and supports the constrained decoding that the screening labels and the extraction schema require. Screening, which reads one abstract and emits one token, used the 32B checkpoint; extraction, which reads a longer window and writes a structured record for every paper in scope, used the 7B checkpoint. Papers were embedded for clustering with bge-base-en-v1.5 (Xiao et al., 2024), a 768-dimensional sentence encoder and smaller than the other two encoders we compared under identical data splits and an identical classification head, bge-large-en-v1.5 and e5-large-v2 (Wang et al., 2022). Table S9 reports both comparisons, across the four model scales and across the three encoders.

Abstract screening

Screening decides which papers enter the atlas, and so we designated a paper as in scope when computation is a central scientific result, including algorithm and model development, software and resource papers, and benchmarking studies. Each abstract was assigned to this computational class or to a primarily experimental class which included experimental, clinical and epidemiological work and reviews in which computation is routine statistical analysis; survival, risk-factor and regression analyses of clinical records, a recurring ambiguous case, have an explicit rule in the prompt. Screening used Qwen2.5-32B-Instruct served with vLLM (Kwon et al., 2023) in bfloat16 at temperature 0, with decoding restricted to the tokens spelling the two labels (Geng et al., 2023; Willard & Louf, 2023). The prompt is reproduced in Table S1.

Structured extraction

Papers classified as computational underwent further extraction with Qwen2.5-7B-Instruct at temperature 0, using guided JSON decoding against a strict schema (Willard & Louf, 2023). Each prompt

included the title, the abstract up to 1,600 characters and the first 2,296 tokens of the body, which was shorter than the full body for 1,014,058 records (94.40%). Each record gives a ten-way artifact type, the artifact name, biological questions, related tools, databases used and a one-sentence summary (Table S2). The prompt frames every paper as biomedical, including those later excluded by the biomedical restriction.

Compiling the atlas & restriction to biomedical topics

To compile the atlas, we clustered all 267,893 papers of the three artifact types that describe a method or software: new algorithms (215,061), software packages (42,455) and web servers (10,377), here called tool-type papers. These include 54,205 papers without a recovered artifact name, since clustering does not require one. We excluded the model class, since only 20.8% of its 290,327 papers name the model they contribute, against 79.8% for the included types, and 43.7% address a clinical prediction question, against 20.4%. Database and pipeline papers, which release data or assemble existing tools, were not clustered; their registry matching is reported separately (Table S12).

Each paper's title, summary and biological questions were embedded with bge-base and L2-normalized, reduced with uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) to five components (15 neighbors, minimum distance 0, cosine, seed 42), and clustered with hierarchical density-based spatial clustering of applications with noise (HDBSCAN) (Campello et al., 2013; McInnes et al., 2017) in scikit-learn (Pedregosa et al., 2011) at minimum cluster size 20, minimum samples 5 and excess-of-mass selection. Papers that HDBSCAN left unassigned were placed in the nearest cluster when their cosine similarity to its centroid reached 0.80, and were left unassigned otherwise. We set that threshold by hand at the eighth percentile of the similarity between clustered papers and their own centroid, so a paper joins a cluster only if it sits about as close to it as the least typical papers already there (Table S13). Cluster keywords were derived by class-based term frequency-inverse document frequency (TF-IDF) (Grootendorst, 2022) and cluster names generated by Qwen2.5-7B-Instruct from the keywords and representative titles. Parent topics were obtained by average-linkage agglomeration of cluster centroids, cut at the prespecified number of parents that maximized centroid silhouette (Table S13).

This corpus included engineering, physics and computer-science articles, because screening did not attempt to constrain the publication's domain. Qwen3-8B (Yang et al., 2025) classified each cluster as biomedical or non-biomedical from its label, keywords and twelve representative titles, and a parent topic was retained when more than half of its papers lay in clusters classified as biomedical.

Validation of the categories

Three properties of the partition were measured: geometric separation, how far apart the clusters lie in the embedding, measured by cosine silhouette (Rousseeuw, 1987) on 12,000 points and by density-

based clustering validation (DBCV) (Moulavi et al., 2014) on 10,000 papers. Semantic coherence, whether a cluster's keywords belong together, was measured by a 200-trial word-intrusion test (Chang et al., 2009) against a chance rate of 0.20. Agreement with an independent labeler was measured against a separate language model's labels for 1,500 papers, by adjusted mutual information (AMI) (Vinh et al., 2010), adjusted Rand index (Hubert & Arabie, 1985) and raw agreement. Stability was assessed by repeating the clustering with other UMAP seeds and settings, other HDBSCAN parameters and without papers lacking a recovered artifact name, comparing each repetition with the reference partition by AMI and by the number of retained clusters overlapping at Jaccard 0.5 or above (Hennig, 2007); each cluster's median overlap across seeds is its released stability score (Table S13).

External validation used curated annotations: for each atlas paper that is the defining publication of a bio.tools entry, identity-resolved as described below, we took the topics and operations of the EDAM ontology of bioinformatics concepts (Ison et al., 2013) that curators assigned to that entry. The primary analysis used entries that are the only entry of their publication, to avoid dependence between entries of one paper; a sensitivity analysis used all linked entries and permuted publications as blocks. For a cluster c with entry set E_c, the concentration of curated annotation is the share of its entries carrying the term t that is most frequent within it,

$$\operatorname{conc}(c) = \frac{\max_t \left|\{\, e \in E_c : t \in T(e) \,\}\right|}{|E_c|},$$

where $T(e)$ is the set of terms curators assigned to entry e. Clusters with at least five entries were evaluated. In each of 1,000 permutations, cluster labels were shuffled among entries and each cluster's most frequent term recomputed; with b permutations reaching the observed value,

$$p = \frac{1+b}{1001}.$$

Qwen2.5-32B-Instruct was given the title, abstract and extracted fields of 300 random papers. Its agreement with the extracted artifact type was 84.3% and with the artifact name 94.9%, a consistency check of the small models and not a measure of accuracy (Table S6).

In a human audit, one annotator labeled stratified random samples of 228 items for four tasks, screening, artifact type and name against the full article, unmatched names and name matches, without seeing the screening class or the extracted type. Estimates weight each stratum by its population, with 95% intervals from 2,000 bootstrap resamples within strata; unclear labels were excluded from the estimate and reported as bounds (Table S16).

Registry comparison

To assess how the extracted software names relate to existing curated resources, we compared them against established software registries and package ecosystems. Extracted names passed two heuristic

filters intended to remove likely non-names: one for names composed entirely of platform, method and container words, and one for multi-token candidates containing neither a capital letter nor a digit (Table S4). Names were also compared with bio.tools (Ison et al., 2019), a curated life-science software registry, and separately with four package ecosystems: PyPI (https://pypi.org), CRAN (https://cran.r-project.org), Bioconductor (Huber et al., 2015) and Bioconda (Grüning et al., 2018). A strict match requires equality after folding case, punctuation and trailing version tokens, or a match of a parenthetical acronym, and all five were retrieved in September 2026 (Table S11).

Strict matching uses registry names alone, whereas an identity-resolved match to a bio.tools entry also uses the entry's publications and links. A match was accepted if the names matched strictly, if the extracted paper was a defining publication for the entry and the names or reported URLs agreed, or if both the URL and name form matched. Name and URL matching rules are defined in Table S12.

Registry matching is reported by artifact type, publication year and parent topic, and papers classified as software packages or web servers form the core software subset, chosen for specificity; the subset is defined by the model's artifact types, so its denominator inherits artifact-classification error. A sensitivity analysis also included all artifact types considered registrable in bio.tools (Table S12). New algorithms are reported in two groups: those whose paper gives a repository or package link, and those whose paper gives none.

Recall against bio.tools

The reference set was derived exclusively from bio.tools, where an entry's defining publications were taken as its Primary publications or, when none were available, its untyped publications, resolved to the corpus by PubMed Central identifier (PMCID), PubMed identifier (PMID) or digital object identifier (DOI) using the PMC ID Converter. Each entry was evaluated using a single defining publication selected a priori: the earliest defining publication represented in the corpus. The principal analysis set, the software subset, comprises entries carrying a tool type other than Database portal and Ontology. An entry counts as recovered when the name extracted from its selected defining publication matches the entry's registry name, so recall over a set R of entries is

$$\operatorname{recall}(R) = \frac{|\{\, r \in R : r \text{ recovered} \,\}|}{|R|}.$$

End-to-end recall takes R to be all 12,207 entries of the software subset, so an entry is a loss whenever its paper was screened as experimental, was extracted as a type that is not clustered, was left unassigned by HDBSCAN, fell outside a retained parent topic, yielded no usable name, or yielded a name that did not match. Recall conditional on entry into the atlas restricts R to entries whose defining publication reached the retained atlas, thereby excluding upstream losses from the denominator and isolating name-matching performance. Recall estimates are reported with Wilson 95% confidence intervals (CI) (Wilson, 1927) and stratified by publication year, tool type, whether the name appears in the

title or abstract, and code-link detection. For extraction-level recall, we included all artifact types designated as registrable in the Table S12 crosswalk (new algorithms, software packages, web servers, databases, pipelines and models), without applying clustering or topic restrictions, and this set was specified before any entry was evaluated.

Retrieval layer and structured store

The retrieval layer, which was not used in the primary analyses, indexes the 1,074,191 papers classified as computational as 30,913,457 passages embedded with Qwen3-Embedding-0.6B (Zhang et al., 2025), fusing BM25 (Robertson & Zaragoza, 2009) and dense retrieval by reciprocal rank fusion (Cormack et al., 2009) with cross-encoder reranking (Chen et al., 2024). In a known-item retrieval benchmark using the curated descriptions of 300 bio.tools tools with names masked, it ranked the defining paper first for 82.7% of queries and in the top ten for 93.7% (Table S15). Alongside the index, a columnar store (Raasveldt & Mühleisen, 2019) holds the papers, 41,890,329 parsed reference entries, dataset accessions detected by 22 identifier patterns, and code links to eight hosts normalized to one repository, project, record or package identity (Tables S8 and S15). The deployed resource also assigns newly indexed bioRxiv (Sever et al., 2019) and arXiv preprints to their nearest cluster above a manually set threshold, holding the rest as emerging; this is not evaluated here (Table S15).

Language models and AI assistance

All models were used as released: Qwen2.5-32B-Instruct for abstract screening and for the extraction consistency check, Qwen2.5-7B-Instruct for structured extraction and cluster naming, Qwen3-8B for classifying clusters as biomedical, bge-base for the clustering embeddings and Qwen3-Embedding-0.6B for the passage index. Every prompt is reproduced in the supplementary material.

During the preparation of this work the authors used OpenAI ChatGPT (GPT-5) to improve the grammar and readability of the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Results

Corpus and screening

Of 7,283,594 parsed articles, 6,446,741 had a title and an abstract, and the screen classified 1,074,191 (16.66%) as computational, 5,371,134 (83.32%) as experimental and 1,416 (0.022%) with no valid label (Table 1, Fig. 1b, 1d). The share classified as computational rose from 13.5% of papers published in 2012 to 20.0% in 2025 (Table S10).

Table 1
Corpus acquisition, screening and extraction

Stage	Count	% of previous
PMC full-text articles parsed	7,283,594	—
Records with title and abstract (screened)	6,446,741	88.51
Screened computational	1,074,191	16.66
No valid screening label	1,416	0.022
Schema-valid extraction	1,074,140	99.995
Named artifact recovered	316,138	29.43
Tool-type papers clustered	267,893	—
Assigned to a cluster	227,771	85.0
Retained after biomedical restriction	165,432	72.6

Structured extraction

Analysis-only studies formed the largest artifact type at 28.19%, followed by models at 27.03% and new algorithms at 20.02% (Fig. S1a, Table S10), and a named artifact was recovered for 316,138 papers (29.43%). In 1,000 randomly drawn papers extracted again with body text up to the model's 32,768-token context limit, the longer window held the whole body for 985. Artifact type agreed between fixed-window and long-context extraction for 73.7% of papers, against 97.3% between two repeated fixed-window extractions, and names agreed for 74.4% of papers named in both. Long-context extraction gave a name for 48.1% of papers against 30.2%, and a tool type for 30.6% against 25.5%. Which window is more accurate was not measured, so these counts are specific to the production window (Table S14).

The atlas and its validation

Clustering the 267,893 tool-type papers yielded 1,438 clusters under 12 parent topics, with 40,122 papers (15.0%) unassigned. The parent-level restriction retained five parents, 1,000 clusters and 165,432 papers, listed with their labels, sizes and keywords in Tables S3 and S5 (Fig. 2a, 2b).

The three measures capture different properties of the partition: the cosine silhouette was − 0.0033 between clusters and 0.118 between parent topics, while DBCV was negative for the reference partition and every variant, indicating that the papers do not form sharply separated geometric clusters under this representation; the clusters partition them for navigation. Word-intrusion accuracy was 0.875, consistent with coherent within-cluster keywords, and agreement with the language-model annotator was modest, at 0.282 AMI, 0.235 adjusted Rand index and 57.5% raw agreement (Fig. 2d).

Repeating UMAP with five additional random seeds produced 1,435-1,486 clusters, with AMI values of 0.78–0.79 relative to the reference partition. Among retained clusters, 72.2% had a stability score of at

least 0.5. Changing the HDBSCAN parameters had a larger effect on the number of clusters, which ranged from 747 to 2,867. We also repeated the clustering after removing papers for which no artifact name was recovered, and 547 retained clusters still overlapped their original counterparts at Jaccard similarity 0.5 or above, showing that excluding these papers changes the resulting partition. Across reassignment thresholds, the proportion of unmatched core names remained nearly unchanged at 60.8–61.4%, whereas overall recall of bio.tools entries varied more substantially, from 30.6% to 65.8%, because this measure includes failures at every stage of the pipeline. Restricted to bio.tools entries whose defining papers entered the atlas, recall rose to 82.8–84.0% (Table S13).

We also compared alternative embedding models, input representations and dimensionality-reduction strategies using 15,000 tool-type papers, of which 7,334 had curated EDAM topic annotations. When UMAP was used for dimensionality reduction, EDAM topic concentration was similar across bge-base, bge-large and e5-large, ranging from 0.61 to 0.66. In contrast, applying HDBSCAN to principal components or directly to the unreduced embeddings produced only 2–7 clusters, indicating that these representations did not yield a useful partition at this scale. We also compared the text used to represent each paper, and a title-and-abstract representation produced more topic-coherent local neighborhoods than the atlas representation based on title, summary and biological questions: 62.9% of nearest linked neighbors shared a curated EDAM topic, compared with 58.2% for the atlas representation (Table S13). The atlas representation had been specified and fixed before these comparisons, and choosing a different representation afterward based on agreement with EDAM topics would effectively optimize the atlas against the same bio.tools annotations later used for validation. We therefore treated these comparisons as sensitivity analyses rather than using them to select the final representation post hoc.

We next asked whether papers grouped into the same atlas clusters also shared independently curated biological annotations. This analysis included 7,261 bio.tools entries, restricted to publications associated with a single bio.tools entry, that mapped to 296 retained clusters containing at least five linked entries each. Within these clusters, the most common EDAM topic was shared by an average of 61.7% of entries, with a per-cluster median of 60.0% and an interquartile range of 48.0–77.0%. Under random permutation, the corresponding concentration was only 26.4% ($p = 0.001$, the smallest attainable value). The most common EDAM operation similarly accounted for 42.4% of entries within clusters, compared with 18.8% under permutation (Fig. 2c). These results were essentially unchanged under stricter linkage criteria or when all 7,584 linked entries were included and publications, rather than individual entries, were permuted. This validation could only be performed for a subset of the atlas: the remaining 704 retained clusters were smaller and more recent, and 69.6% of their papers concerned biomedical image and signal processing, compared with 13.2% in the evaluated clusters. The observed annotation coherence should therefore be interpreted as evidence for the evaluated subset rather than for all retained clusters.

Finally, we tested whether the biomedical filtering strategy affected the atlas. The overall number of retained papers was relatively stable across filtering approaches, although the classification of individual clusters was not. Because the primary rule retains entire parent topics, 13.3% of retained papers

belonged to clusters that the separate cluster-level classifier labeled non-biomedical, and the cluster-level classifications agreed with the parent-level rule for 1,152 of the 1,438 clusters. Despite these differences, downstream name-matching results changed little: the proportion of unmatched extracted name strings ranged from 75.1% to 77.0% across filtering strategies, and strict bio.tools recall was similar, at 56.2% when the cluster-level classifications were used and 57.3% when no biomedical restriction was applied (Table S13). The choice of biomedical filtering rule therefore affected which individual clusters were retained but had relatively little effect on the aggregate name-matching and recall results.

Artifact names contributed by the literature

The retained atlas contains 121,844 distinct artifact name strings identified across 129,169 naming events. Most names were reported only once, with 95.8% contributed by a single paper. By artifact type, 92,110 distinct names were associated with new algorithms, 25,241 with software packages and 6,179 with web servers (Table 2).

Table 2
Registry matches for 121,844 distinct artifact names, by artifact type. Values are row percentages. S, strict name matching; I, identity resolution using publication and URL evidence; Packages, PyPI, CRAN, Bioconductor or Bioconda; None, no registry match. Code links are repository or package URLs reported in the paper

Name strings from	Names	bio.tools S	bio.tools I	Packages	None S	None I
Software packages and web servers (core subset)	31,180	27.7	29.3	23.5	61.0	59.6
Software packages	25,241	25.9	27.2	25.6	61.7	60.6
Web servers	6,179	37.0	39.7	16.2	56.4	53.9
New algorithms, code link detected	14,491	16.5	17.7	23.7	69.3	68.3
New algorithms, link named as the method	6,199	26.2	27.8	30.9	56.8	55.5
New algorithms, no software evidence detected	78,508	6.7	6.8	14.5	83.2	83.1
All tool types	121,844	12.4	13.0	17.0	77.0	76.4

Registry matching in the core software subset

Of 31,180 distinct names extracted from software-package or web-server papers, 61.0% had no strict match in any of the five registries, and 59.6% remained unmatched after identity resolution. bio.tools alone matched 29.3% after identity resolution (Table 2, Fig. 3a). Among new algorithms with a detected code link, 69.3% were unmatched, and this fell to 56.8% when the linked repository or package included

the method name. New algorithms with no software evidence detected were unmatched at 83.2%, and our evidence does not establish their eligibility for a software registry.

The unmatched share increased with publication year in the core subset, from 53.5% for names introduced before 2011 to 66.7% for names introduced from 2022 onward, with values for 2011–2021 ranging from 55.6% to 60.1%. The higher share among the newest names is consistent with registration lag, although a cross-sectional comparison with snapshots from a single month cannot separate lag from cohort and field effects. By parent topic it was 48.1% for biological data analysis and 75.6% for biomedical image and signal processing, where bio.tools matched 12.4% of core names (Table S12).

Across all 121,844 atlas names, 77.0% were unmatched by strict matching and 76.4% after identity resolution. Across the alternative restrictions, name filters and inclusion rules in Table S4, the unmatched share ranged from 75.1% to 81.0%. Nearly two thirds of these names came from algorithms with no software evidence detected, so the all-name figure is not an estimate of registry coverage. Unmatched names should not automatically be interpreted as unregistered software: in a previous large-scale extraction study, only 69.66% of the 10,000 most frequent extracted mentions were confirmed as software by manual curation, despite a cross-validated F1 of 0.92 (Istrate et al., 2022), and other software-mention datasets likewise contain substantial fractions of non-software entities (Druskat et al., 2024) (Table S16).

Of 34,232 bio.tools entries, 16,475 had a defining paper in the screened corpus. Another 11,044 cited papers outside PMC, and 2,738 cited PMC papers outside the open-access subset. Entries whose papers were outside PMC were more often command-line tools (38.4% against 25.8% among included entries) or libraries (24.2% against 15.4%), so the evaluated reference set is not representative of all bio.tools entries and is enriched for web applications and more established software (Table S11).

Strict matching recovered 56.6% of the 12,207 entries in the software subset (95% CI 55.7–57.5), and among the 3,349 entries with a Primary publication recall was 54.3% (95% CI 52.6–55.9; Table 3). Identity resolution increased these estimates to 62.1% and 60.3% respectively, but because the 674 additional matches it introduced were not manually validated, we use strict-match recall as the primary estimate. The largest losses were papers extracted as an artifact type that is not clustered (16.5%), principally databases, models and pipelines, and papers left unassigned by HDBSCAN (14.2%; Fig. 3b), so end-to-end recall reflects both extraction performance and the atlas inclusion criteria. Recall differed most strongly according to whether the registry name appeared in the defining paper’s title or abstract: strict-match recall was 61.6% for the 10,883 entries whose names appeared there, compared with 15.4% for the 1,316 entries whose names did not, and among the latter group 41.7% were lost specifically at the name-matching stage. Differences by code-link detection, tool type and publication year were smaller (Table S11). Of the 8,265 entries whose paper entered the atlas, 83.6% were recovered by strict matching (Table 3).

Table 3

Recall for the 12,207 bio.tools entries in the software subset, defined as entries with a qualifying tool type and a defining paper in the screened corpus. Compatible types are new algorithms, software packages, web servers, databases, pipelines and models. Rows use different denominators: the first row includes the full software subset, whereas entries whose paper entered the atlas are conditional on successful completion of the preceding pipeline stages. Intervals are Wilson 95% confidence intervals. Stage-specific losses are shown in Fig. 3b, with additional analyses in Table S11

Measure	n	Strict	Identity-resolved
Entries recovered	12,207	56.6 (55.7–57.5)	62.1 (61.3–63.0)
Entries with a Primary publication	3,349	54.3 (52.6–55.9)	60.3 (58.7–62.0)
Entries from single-entry publications	11,498	58.7	63.2
Entries whose paper entered the atlas	8,265	83.6	91.8
Compatible types admitted, no clustering	12,207	81.3	88.9
Name in the paper's title or abstract	10,883	61.6	65.7
Name not in the title or abstract	1,316	15.4	33.1

Extracted types and registry tool types agreed where both describe the same kind of object: web servers were registered as web applications in 79% of cases, whereas models diverged most (Table S12).

Human audit

A single annotator manually audited stratified samples from four stages of the pipeline against the full-text articles and relevant registry records (Table S16). Among 50 papers classified as computational, 74.0% (95% CI 62.0–86.0) treated computation as a central result, so the 1,074,191 screened-positive papers should be interpreted as papers classified as computational rather than papers whose primary contribution is computational. After weighting the three sampling strata by their population sizes, the estimated sensitivity of the screening step was 92.9% (95% CI 88.9–96.6). Among 50 name-based registry matches, 55.1% (95% CI 36.4–72.4) referred to the same artifact, including 15 of 23 bio.tools matches, 9 of 25 PyPI matches, and the single audited CRAN and Bioconda matches, so name-based matching to package registries was the least reliable audited stage. All 15 audited unmatched names were judged to represent genuine named methods or software, although for 12 the annotator could not determine whether the same artifact was registered under a different name, and none of the remaining three had an identified registry counterpart. In a separate full-text audit of 13 papers, the extracted artifact type agreed with the annotator in 8 cases, while the extracted artifact name agreed in every paper for which a name was present, and all three papers classified by the pipeline as software packages were judged by the annotator to describe new algorithms. These audits are limited by the small sample sizes and the use of a single annotator, resulting in wide confidence intervals: screening precision was evaluated on 50 papers, extraction on 13 papers and unmatched names on 15 cases. Identity resolution, extraction-window effects, code-link detection, repository and accession extraction, cluster labeling and the biomedical restriction were not independently audited.

Structured resource index

The store holds 499,531 dataset-accession occurrences, 285,615 of them distinct, in 184,701 papers (17.2% of the corpus), and 152,097 code-link occurrences in 97,331 papers, resolving to 77,382 distinct Git repositories and 6,588 package, project and archive records (Table S15). Of 400 sampled repository links, 385 resolved; URL resolution verifies accessibility but not correspondence to the reported artifact.

The citation graph, joined by PMID or DOI, resolves 5,767,772 of the 41,890,329 reference entries to papers in the corpus (Table S15).

Discussion

PaperAtlas is primarily a literature-navigation resource. It organizes computational-methods papers for task-oriented browsing, identifies candidate software names for registry curation, and links papers to reported dataset accessions and code repositories. The registry analyses address a separate question: how often extracted artifact names match entries in five registry snapshots from September 2026.

The results also highlight the limitations of name-based registry matching. Recall fell to 15.4% when the registry name did not appear in the defining paper's title or abstract, indicating that naming differences alone can prevent otherwise related records from being linked. Mention-based resources (Istrate et al., 2022; Schindler et al., 2022), curated registries such as bio.tools (Ison et al., 2019) and FAIRsoft (Martín del Pico et al., 2024), and PaperAtlas therefore capture complementary views of the computational-methods literature.

Conclusion

PaperAtlas organizes 165,432 computational-methods papers from the PubMed Central open-access subset into 1,000 biomedical clusters, providing a reproducible structure for navigating the literature. Where curated bio.tools annotations permit evaluation, these clusters show substantial concentration of shared EDAM topics. Among 31,180 distinct names extracted from software-package and web-server papers, 61.0% had no strict match in the five evaluated registry snapshots. The pipeline recovered 56.6% of bio.tools software entries end to end by strict matching, rising to 83.6% among entries whose defining publication entered the atlas. A small manual audit found the extracted artifact names accurate in the cases examined, while registry name matching was the least reliable audited step, so these results measure concordance between names reported in the literature and names present in the evaluated registry snapshots rather than definitive registry membership. Together, these results show that PaperAtlas can organize the computational-methods literature at scale, while also revealing substantial gaps and naming inconsistencies between published software and existing registries.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

Not applicable: the study analyses published literature and public registry records, with no human participants or animals.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. This work was conducted at Bhargava Systems Research Inc., a registered nonprofit organization.

Author Contribution

N.S. was involved in formal analysis, investigation, data curation, validation, visualization and writing the original draft. A.V. was involved in formal analysis, investigation and writing the original draft. Y.B. was involved in conceptualization, methodology, software, resources, supervision and writing (review and editing). All authors read and approved the final version of the manuscript.

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. This work was conducted at Bhargava Systems Research Inc., a registered nonprofit organization.

Data Availability

The atlas and retrieval service are freely available without registration at https://bhargavaresearch.org/paperatlas, and release files can be downloaded from https://bhargavaresearch.org/paperatlas/download. Pipeline code is available at https://github.com/Bhargava-Systems-Research-Inc/PaperAtlas, and the analysis code, supplementary tables, validation data, atlas, embeddings, registry snapshots and per-paper extracted records are archived at https://doi.org/10.5281/zenodo.22917574. The passage index and parsed corpus are available on request and can be regenerated from the PubMed Central open-access subset using the released code. Verbatim text is distributed only where permitted by the source licence (Online Resource

1, Table S7). This article describes the archived September 2026 release; the live web resource is updated monthly, so current counts may differ.

References

1. Allot, A., Lee, K., Chen, Q., Luo, L., & Lu, Z. (2021). LitSuggest: A web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Research*, *49*(W1), W352–W358. https://doi.org/10.1093/nar/gkab326
2. Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1). https://doi.org/10.1057/s41599-021-00903-w
3. Brown, E. M., Druskat, S., Hébert-Dufresne, L., Howison, J., Mietchen, D., Nesbitt, A., Pimentel, J. F., & Veytsman, B. (2026). Biomedical open source software: Crucial packages and hidden heroes. *PLOS Computational Biology*, *22*(5), e1014260. https://doi.org/10.1371/journal.pcbi.1014260
4. Cai, P., Liu, S., Zhang, D., Xing, H., Han, M., Liu, D., Gong, L., & Hu, Q. N. (2023). SynBioTools: A one-stop facility for searching and selecting synthetic biology tools. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05281-5
5. Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based clustering based on hierarchical density estimates. *Lecture Notes in Computer Science*, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14
6. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22*, 288–296.
7. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity text embeddings through Self-Knowledge distillation. *Findings of the Association for Computational Linguistics ACL 2024*, 2318–2335. https://doi.org/10.18653/v1/2024.findings-acl.137
8. Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 758–759. https://doi.org/10.1145/1571941.1572114
9. Druskat, S., Hong, N. P. C., Buzzard, S., Konovalov, O., & Kornek, P. (2024). Don’t mention it: An approach to assess challenges to using software mentions for citation and discoverability research. *arXiv*. https://doi.org/10.48550/arXiv.2402.14602
10. Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021). Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, *72*(7), 870–884. https://doi.org/10.1002/asi.24454
11. Duck, G., Nenadic, G., Filannino, M., Brass, A., Robertson, D. L., & Stevens, R. (2016). A survey of bioinformatics database and software usage through mining the literature. *PLOS ONE*, *11*(6),

e0157989. https://doi.org/10.1371/journal.pone.0157989

12. Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M., & Andrade-Navarro, M. A. (2009). MedlineRanker: Flexible ranking of biomedical literature. *Nucleic Acids Research*, *37*(suppl_2), W141–W146. https://doi.org/10.1093/nar/gkp353
13. Geng, S., Josifoski, M., Peyrard, M., & West, R. (2023). Grammar-Constrained decoding for structured NLP tasks without finetuning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10932–10952. https://doi.org/10.18653/v1/2023.emnlp-main.674
14. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*. https://doi.org/10.48550/arXiv.2203.05794
15. Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7
16. Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271. https://doi.org/10.1016/j.csda.2006.11.025
17. Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137–2155. https://doi.org/10.1002/asi.23538
18. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252
19. Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075
20. Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H., Kalaš, M., Schwämmle, V., Grüning, B., Beard, N., Lopez, R., Duvaud, S., Stockinger, H., Persson, B., Vařeková, R. S., Raček, T., Vondrášek, J., Peterson, H., Salumets, A., Jonassen, I., … Brunak, S. (2019). The bio.tools registry of software tools and data resources for the life sciences. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1772-6
21. Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., & Rice, P. (2013). EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. https://doi.org/10.1093/bioinformatics/btt113
22. Istrate, A. M., Li, D., Taraborelli, D., Torkar, M., Veytsman, B., & Williams, I. (2022). A large dataset of software mentions in the biomedical literature. *arXiv*. https://doi.org/10.48550/arXiv.2209.00693
23. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. *Proceedings*

of the 29th Symposium on Operating Systems Principles, 611–626. https://doi.org/10.1145/3600006.3613165

24. Martín del Pico, E., Gelpí, J. L., & Capella-Gutierrez, S. (2024). FAIRsoft—a practical implementation of FAIR principles for research software. *Bioinformatics*, *40*(8). https://doi.org/10.1093/bioinformatics/btae464
25. McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205
26. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*. https://doi.org/10.48550/arXiv.1802.03426
27. Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-Based clustering validation. *Proceedings of the 2014 SIAM International Conference on Data Mining*, 839–847. https://doi.org/10.1137/1.9781611973440.96
28. National Information Standards Organization (2024). ANSI/NISO Z39.96-2024, JATS: Journal article tag suite, version 1.4. https://doi.org/10.3789/ansi.niso.z39.96-2024
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
30. Qwen Team (2024). Qwen2.5 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2412.15115
31. Raasveldt, M., & Mühleisen, H. (2019). DuckDB. *Proceedings of the 2019 International Conference on Management of Data*, 1981–1984. https://doi.org/10.1145/3299869.3320212
32. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *4*(1–2), 1–174. https://doi.org/10.1561/1500000019
33. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
34. Sayers, E. W., Bolton, E. E., Fine, A. M., Kelly, C., Kim, S., Landrum, M., Lathrop, S., Malheiro, A., Murphy, T. D., Phan, L., Pujar, S., Trawick, B. W., Schneider, V. A., & Pruitt, K. D. (2026). Database resources of the national center for biotechnology information in 2026. *Nucleic Acids Research*, *54*(D1), D20–D27. https://doi.org/10.1093/nar/gkaf1060
35. Schindler, D., Bensmann, F., Dietze, S., & Krüger, F. (2022). The role of software in science: A knowledge graph-based analysis of software mentions in PubMed central. *PeerJ Computer Science*, *8*, e835. https://doi.org/10.7717/peerj-cs.835
36. Sever, R., Hindle, S., Roeder, T., Fereres, S., Fernández Gayol, O., Ghosh, S., Proietti Onori, M., Croushore, E., Black, K. J., Sussman, L., Argentine, J., Manos, W., Muñoz, M., Sinanan, J., Teal, T. K., & Inglis, J. R. (2019). bioRxiv: The preprint server for biology. https://doi.org/10.1101/833400

37. Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–2854.
38. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2022). Text embeddings by Weakly-Supervised contrastive pre-training. *arXiv*. https://doi.org/10.48550/arXiv.2212.03533
39. Wei, C. H., Allot, A., Lai, P. T., Leaman, R., Tian, S., Luo, L., Jin, Q., Wang, Z., Chen, Q., & Lu, Z. (2024). PubTator 3.0: An AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, *52*(W1), W540–W546. https://doi.org/10.1093/nar/gkae235
40. Willard, B. T., & Louf, R. (2023). Efficient guided generation for large language models. *arXiv*. https://doi.org/10.48550/arXiv.2307.09702
41. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953
42. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., & Nie, J. Y. (2024). C-Pack: Packed resources for general chinese embeddings. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 641–649. https://doi.org/10.1145/3626772.3657878
43. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., … Qiu, Z. (2025). Qwen3 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2505.09388
44. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., & Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2506.05176

Figures

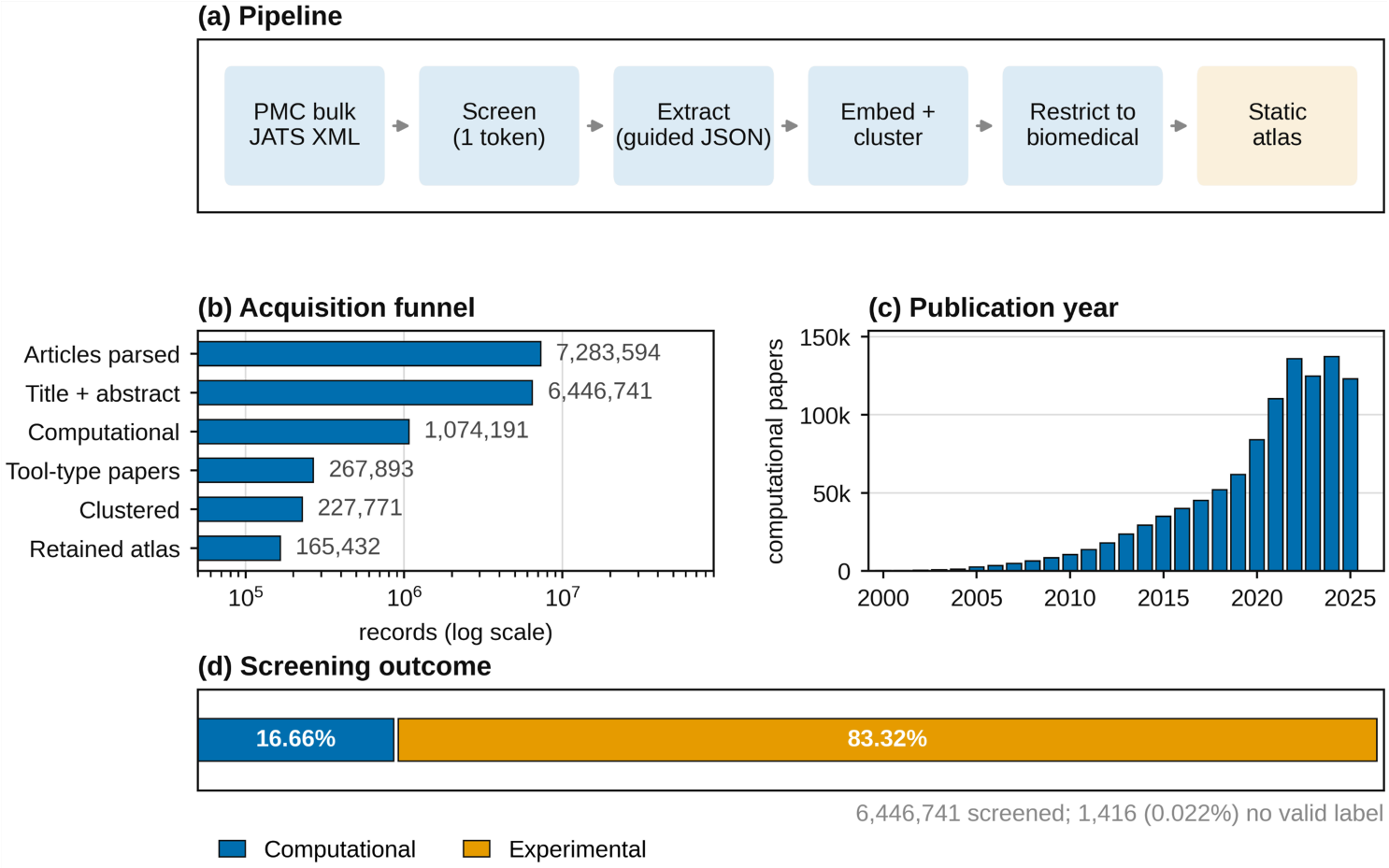


Figure 1

The PaperAtlas pipeline and corpus. (a) End-to-end workflow. (b) Acquisition and screening funnel, 7,283,594 articles parsed to 165,432 retained under the parent-level biomedical rule. (c) Publication year of the computational corpus. (d) Screening outcome.

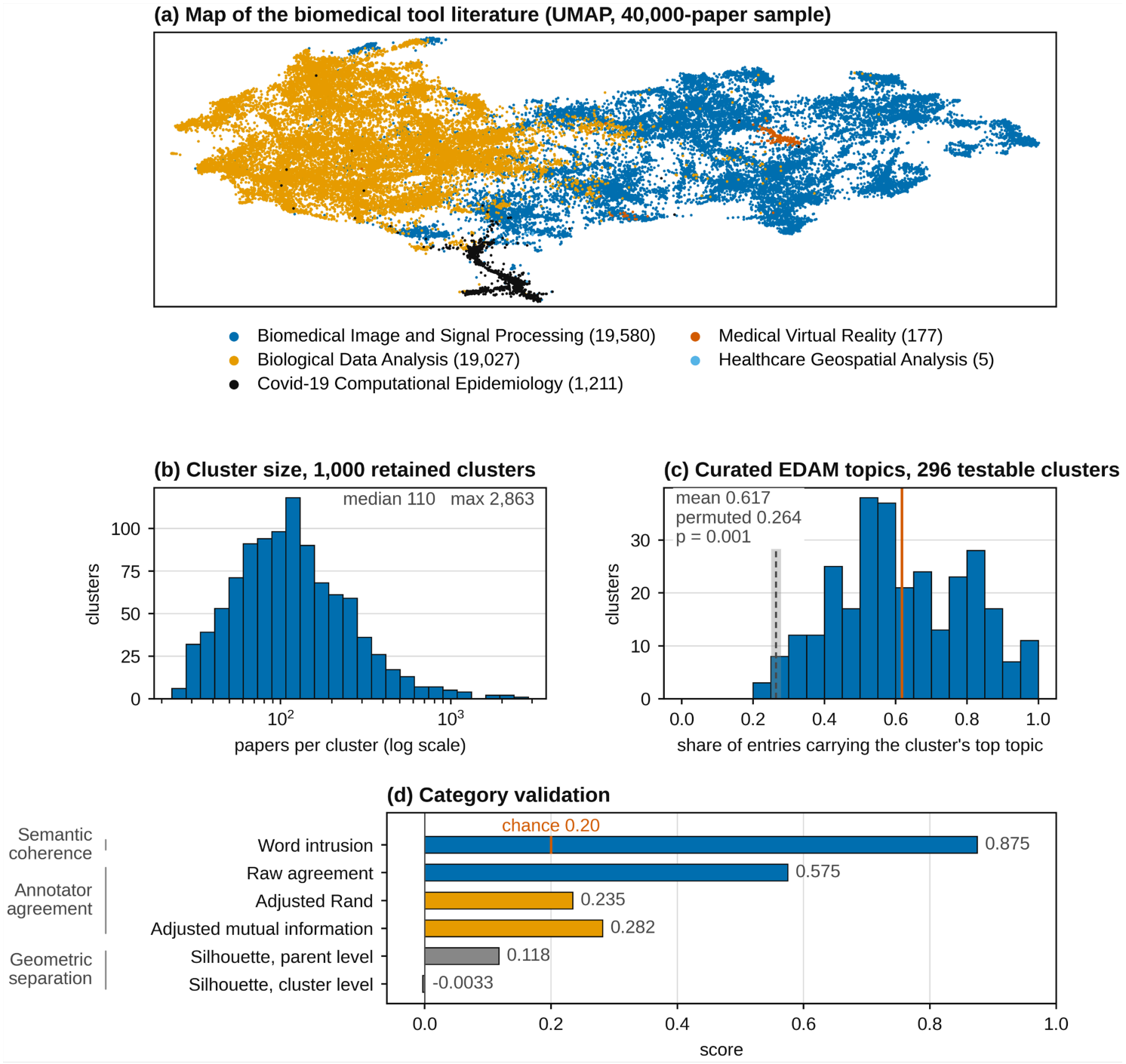


Figure 2

The atlas and its validation. (a) UMAP of the retained tool-type paper embeddings, colored by parent topic, a navigational grouping. (b) Cluster sizes across the 1,000 retained clusters. (c) For the 296 clusters with at least five linked bio.tools entries, the share of entries carrying the cluster's most frequent curated EDAM topic. The solid line is the mean, the dashed line and grey band the mean and range of 1,000 permutations of cluster labels. (d) Validation measures, grouped by the property each tests. Word intrusion is an accuracy with a 0.20 chance floor, marked. The adjusted Rand index and adjusted mutual information are chance-corrected, so their chance value is zero. The silhouette ranges from -1 to 1, and zero denotes overlapping clusters. None of these measures uses human labels.

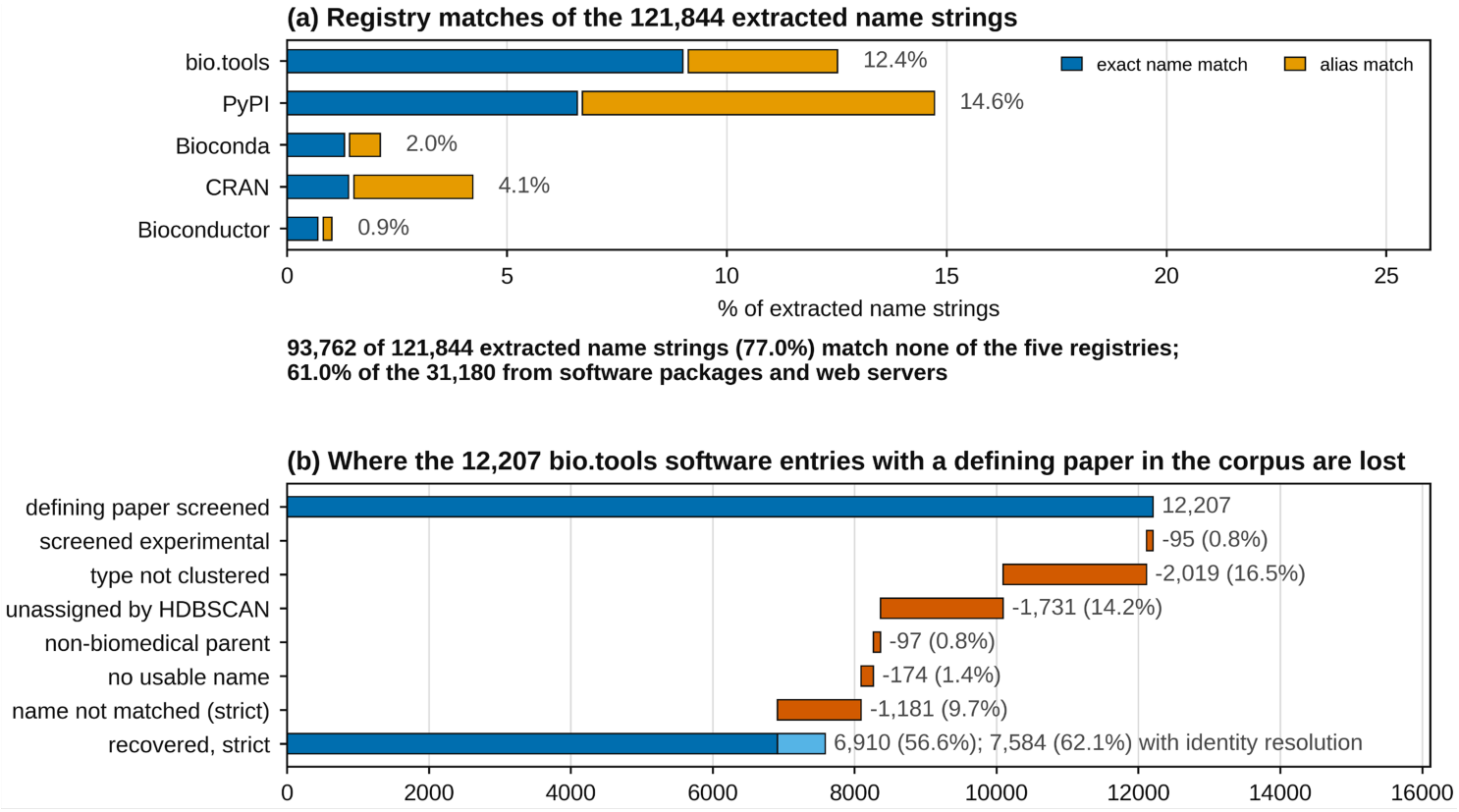


Figure 3

Registry matching and bio.tools recall. (a) Exact and alias match rates of the 121,844 extracted name strings against each of the five evaluated registries, with the share matching none for all name strings and for the core software subset. (b) Where the 12,207 bio.tools software entries with a defining paper in the corpus are lost, stage by stage, ending with the entries recovered by strict matching and, lighter, those added by identity resolution (Table S11).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table3.xlsx
- Table2.xlsx
- Table1.xlsx
- SupplementaryMaterial.pdf