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.

An Automated Multi-Source Framework for Constructing Domain-Specific Literature Corpora for AI-Driven Knowledge Extraction

DIBAKAR DATTA

dibakar_datta@alumni.brown.edu

New Jersey Institute of Technology

Zinia Sultana Joti

New Jersey Institute of Technology

Research Article

Keywords: Literature corpus construction; retrieval-augmented generation; scholarly metadata; open-access acquisition; relevance ranking; nanoparticle drug delivery; electro-optic materials; AI-ready databases

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11836049/v1

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

Additional Declarations: The authors declare no competing interests.

An Automated Multi-Source Framework for Constructing Domain-Specific Literature Corpora for AI-Driven Knowledge Extraction

Zinia Sultana Joti, Dibakar Datta *

Department of Mechanical and Industrial Engineering
New Jersey Institute of Technology (NJIT), Newark, NJ 07102, USA

* Corresponding author: ddlab@njit.edu

October 26, 2026

ABSTRACT

Literature-grounded AI systems, from retrieval-augmented generation to automated database curation, can use only the papers that their acquisition stage collects, yet that stage is rarely specified or measured. We present an open, modular pipeline that builds a domain-specific full-text corpus and reports a measurable quantity at every stage. Anchored queries from ten template families are issued to Crossref, arXiv, PubMed and OpenAlex; the records are deduplicated, scored by a transparent domain-weighted lexical function, and passed to a cascading resolver that tries five kinds of open-access location until a valid PDF is obtained. Applied to nanoparticle–membrane interaction in cancer drug delivery with 500 queries per source, the pipeline reduced 42 918 raw records to 5863 relevant records and 1229 validated PDFs. The sources were complementary: 95 % of the relevant records came from one source only, and no source supplied more than 41 % of them. The acquisition yield of 21.0 % factorises into URL coverage (53.2 %) and download success (39.4 %); outside arXiv the download success was close to 30 % for every source, so coverage set the differences between sources. The yield fell with relevance, from 23 % in the weakest score band to 14 % in the strongest. Of 100 hand-labelled records at the maximum score, 93 were relevant (95 % Wilson interval 0.86–0.97). A cross-check on electro-optic materials, with only the vocabulary pack replaced, raised the yield to 37.3 %, mainly because more of that literature is on arXiv. The domain enters only through this pack, so the approach is domain agnostic and can be applied to any field, within science and beyond it.

Keywords Literature corpus construction; retrieval-augmented generation; scholarly metadata; open-access acquisition; relevance ranking; nanoparticle drug delivery; electro-optic materials; AI-ready databases

HIGHLIGHTS

- A four-source, resumable pipeline turns a domain vocabulary into a ranked, deduplicated and provenance-logged PDF corpus.
- The sources are complementary: 95 % of relevant records come from one source only, and no source supplies more than 41 %.
- Acquisition yield factorises as $\eta = \phi\,\psi$; outside arXiv, download success is near 30 %, so URL coverage decides which source contributes full text.
- The most relevant records are the least likely to be acquired as full text.

PLAIN-LANGUAGE SUMMARY

AI tools that read the scientific literature can only learn from the papers they are given. We built software that searches four scholarly databases, removes duplicates, ranks what it finds by topic relevance, and downloads the papers that are freely available. Tested on nanoparticle drug-delivery research, each database found papers that the others missed, and the main bottleneck was obtaining the full text of relevant papers, most of which are not freely downloadable. A second test on electro-optic materials needed only a new list of topic words, so the same software can serve other fields, in science and beyond.

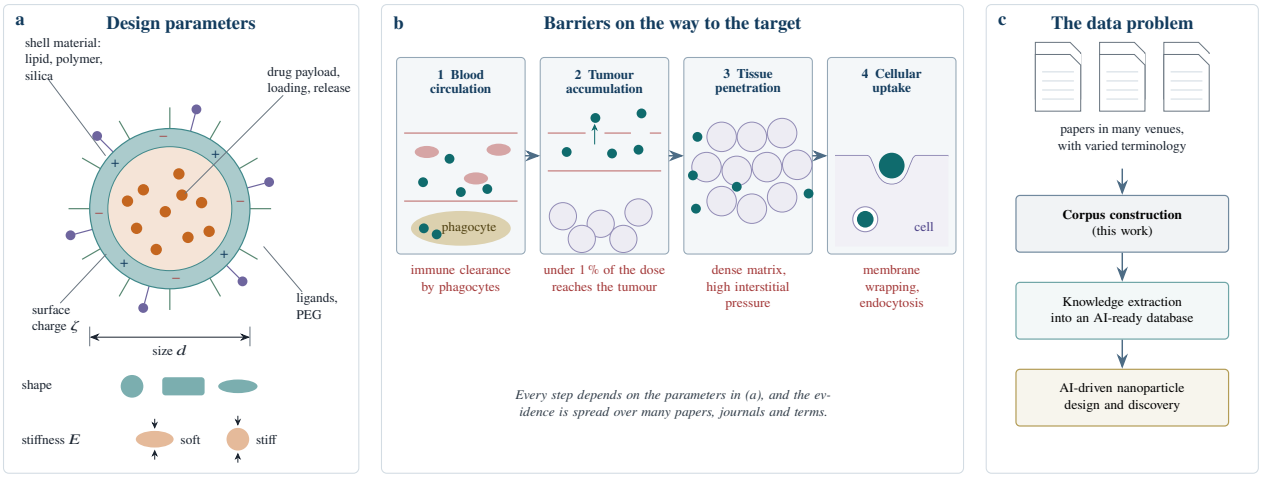


Figure 1. Why the nanoparticle drug-delivery literature needs automated mining. (a) Design parameters of a drug-delivery nanoparticle: size d, shape, stiffness E (Young's modulus), surface charge ζ (zeta potential), surface ligands and poly(ethylene glycol) (PEG) chains, shell material and drug payload [4, 5, 8]. (b) The barriers a particle meets between injection and the target cell; a median of 0.7 % of the injected dose reaches a solid tumour [6]. (c) The data that link (a) to (b) are spread over a large and varied literature. This paper addresses the first step of the chain: building the corpus from which a database, and then AI-driven design, can be derived.

1. Introduction

Artificial intelligence now enters every stage of research, from machine-learned pattern recognition and property prediction to large language models (LLMs) that automate multi-step workflows. Machine-learning models, however, depend on large and structured datasets. Materials science has curated resources such as the Materials Project [1] and the Inorganic Crystal Structure Database [2], but many applied domains have nothing comparable. The scientific literature itself doubles roughly every 17 years [3] and is scattered across papers that differ in terminology and reporting conventions, which makes manual curation slow and incomplete.

Nanoparticles for cancer drug delivery show the problem clearly (Fig. 1). A nanoparticle carrier is defined by many coupled design parameters: its size, shape, surface charge and surface chemistry, the material of its core and shell, its drug loading and release, and its mechanical stiffness [4, 5]. Each parameter changes how the particle behaves at each biological barrier: clearance from the blood by phagocytes, accumulation in the tumour, penetration into the tissue and uptake by the target cells. The barriers are severe. A meta-analysis of published studies found that a median of only 0.7 % of the injected dose reaches a solid tumour [6]. Mechanical properties such as elasticity and Young's modulus affect circulation, tumour penetration and cellular uptake, yet they are much less studied than size and shape [7, 8]. The measurements that would guide the design of a better carrier therefore exist, but they are spread over thousands of papers in medicine, chemistry, materials science and mechanics. Curated nanomaterial databases help, but they remain small: Yan et al. [9] assembled 705 nanomaterials with up to six physicochemical properties each. Data on nanoparticle mechanics, membrane interaction and uptake efficiency remain locked in unstructured text.

Datta [10] proposed an autonomous framework that converts the literature into an AI-ready knowledge base in three steps: (i) literature acquisition, (ii) knowledge extraction, and (iii) data fusion and harmonisation. The first step fixes what every later step can see. A paper that is not acquired cannot be summarised, extracted, or cited, and no downstream model can recover it.

LLMs handle long inputs and summarise large documents well [11], and they have been used to delegate relevance screening and study selection in systematic reviews under human supervision [12]. They also hallucinate references and answer from static training data [13]. Retrieval-augmented generation (RAG) [14] addresses this by grounding generation in retrieved documents, and a series of scientific RAG systems now answer questions, summarise and recommend literature [15–18]. These systems are evaluated on corpora that were fixed in advance, collected by hand, or retrieved with one query against one database [15, 16, 19, 20]. AutoLit, for example, derives five to ten keywords from a seed paper with an LLM and searches arXiv alone [19]. Combining LLM-generated keyword search with embedding

search raised recall by about 30 % over either method alone [21], which shows that the query, and not only the ranker, decides what reaches the corpus. Coverage also differs widely between academic search systems [22], so a single source sees only part of a field.

The problem is sharpest in interdisciplinary domains such as nanoparticle mechanics in cancer drug delivery: *stiffness* and *elasticity* retrieve work on composites and coatings, while drug-delivery terms retrieve studies that never measure a mechanical property. One ambiguous query reaches part of the field; one database indexes part of the literature. HySemRAG [23] is the closest multi-source design: an eight-stage extract–transform–load pipeline that harvests and deduplicates metadata from PubMed, OpenAlex and Scopus, retrieves open-access PDFs and supports hybrid retrieval, citation verification and gap analysis. It does not report the precision of its acquisition stage, the yield of its PDF retrieval, or how the acquisition behaves in a second domain.

> **Key Idea:** Corpus construction is a measurable stage with its own failure modes, and it should be reported like one. We treat it as a pipeline of explicit operators (query generation, retrieval, deduplication, relevance thresholding and PDF resolution) and attach a quantity to each: record retention, source overlap, URL coverage, download success, resolver success rates, and the precision of the highest-scored records.

This paper implements and evaluates step (i) of the framework of Datta [10]. Our contributions are:

1. An open, resumable pipeline (Fig. 2) that generates anchored queries from ten template families, retrieves metadata from Crossref, arXiv, PubMed and OpenAlex, merges and scores the records, and acquires open-access PDFs through a five-step resolver cascade with per-attempt provenance.
2. A complete, code-faithful specification of the relevance function (Eqs. (7) to (10)), including its gate, saturation and quantisation, so that the ranking can be reproduced and criticised.
3. A decomposition of acquisition yield into URL coverage and download success (Eq. (11)) that locates the bottleneck by source and by resolver.
4. An evaluation on nanoparticle–membrane interaction in cancer drug delivery (500 queries per source) that measures how much each source contributes that no other source returns, where relevant records are lost, and how acquisition varies with relevance, together with manually labelled precision for the highest-scored records.
5. A cross-check on a second, unrelated domain, electro-optic materials, in which only the vocabulary pack changes. It shows that the pipeline is domain agnostic: the domain enters only through a replaceable set of terms and weights.

Section 2 specifies the pipeline, Sec. 3 the runs and metrics, Sec. 4 the results, and Sec. 5 what they show, what they leave open, and how the approach extends to other fields.

2. Framework

2.1 Architecture and orchestration

The pipeline is a set of ten Python modules (M1–M10, Table 1) coordinated by a pipeline manager (M9). It runs in four stages (Fig. 2): *plan* the queries, *retrieve* metadata, *curate* the records, and *acquire and enrich* the full text. Every intermediate result is written to a single SQLite database (M1): query checkpoints, papers, abstracts, keywords, every PDF download attempt, and the final labels. A run interrupted by a network failure or an exhausted API quota resumes from its checkpoints under the same run name; completed queries are served from the checkpoint store and existing valid PDFs are reused. Algorithm 1 states the order of operations, and Table 1 lists what each module does; the module labels are used throughout the paper.

2.2 Anchored query generation

A single keyword list either misses the vocabulary of a subfield or admits records from unrelated fields that share a term. The query engine (M2) therefore builds queries from two kinds of vocabulary (Table 2 and Fig. 3). *Anchor families* define the domain: nanoparticle ($\mathcal{A}_N$, 33 terms), membrane and uptake ($\mathcal{A}_M$, 29) and drug delivery ($\mathcal{A}_D$, 17). *Context vocabularies* add the aspect of interest: mechanics, interaction, design, simulation, experiment, and AI/ML (11–23 terms each). A context term is never used alone. Every query must contain at least one anchor term. With $F(q)$ the set

Table 1. The ten modules of the pipeline, the stage of Fig. 2 in which each acts, its responsibilities, and what it writes. All persistent state lives in one SQLite database per run.

Module	Stage	Responsibility	Writes
M1	all	SQLite schema, normalised DOI and title keys, query checkpoints, download-attempt log, exports	`papers`, `abstracts`, `keywords`, `query_checkpoints`, `download_attempts`, CSV
M2	plan	Anchor and context vocabularies, ten template families, anchor gate, balanced sampling, source adaptation	generated query catalogue
M3	retrieve	Per-source minimum interval, retry with back-off, arXiv cool-down, daily-quota deferral	failed-query log
M3A–D	retrieve	Crossref, arXiv, PubMed and OpenAlex clients returning one record schema	records, PDF hints, OA locations
M4	curate	Concept groups, weights, bonuses, penalty and normalisation of Eqs. (7) to (10)	relevance score
M5	acquire	Candidate list, 403 retry, PDF validation, file naming, per-attempt provenance	PDFs, attempt records
M6	enrich	Optional embeddings and cosine-similarity search	`embeddings`
M7	enrich	DOI content negotiation with local fallback, unique citation keys	BibTeX, `.bib` file
M8	enrich	Keyword themes and optional k-means clusters	`paper_topics`
M9	all	Orchestration, year validation, merging, threshold, resume	run report
M10	output	Optional dashboard over the database	none

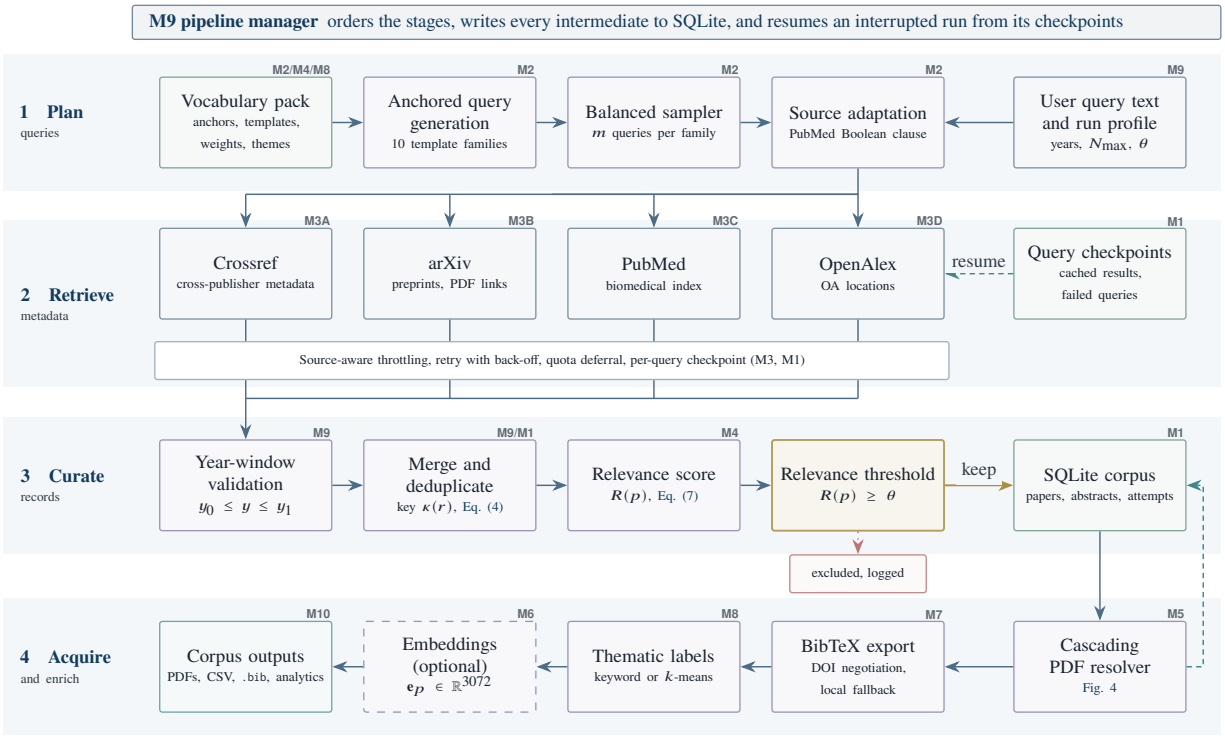


Figure 2. Architecture of the corpus-construction pipeline. Each stage is a lane; module numbers sit above the top right corner of each node and are described in Table 1. Node colour follows the house vocabulary: green for configuration and stores, indigo for methods, navy for structure and sources, gold for the decision point, teal for outputs, red for rejection. Solid edges carry records forward; dashed edges return state (cached query results on resume, download outcomes to the corpus). The same code serves another domain when the vocabulary pack (top left) is replaced.

Algorithm 1. Corpus construction with resume (M9)

Input: Vocabulary pack $\mathcal{V}$; user query q_0; year window $[y_0, y_1]$; queries per source Q; records per query $N_{\max}$; relevance threshold θ
Output: SQLite corpus $\mathcal{D}$, validated PDFs, BibTeX file, theme labels

Plan
$\{\mathcal{Q}_c\}_{c=1}^{10} \coloneqq \textsc{Generate}(\mathcal{V})$; prepend q_0 to the mixed family *// Eqs. (1) and (2)*
$\mathcal{B} \coloneqq \textsc{Sample}(\{\mathcal{Q}_c\}, m = Q/10)$ *// Eq. (3)*

Retrieve
for all source $s \in$ (Crossref, arXiv, PubMed, OpenAlex) **do**
  for all $q \in \mathcal{B}_s$ **do**
    if a completed checkpoint exists for $(s, q, y_0, y_1, N_{\max})$ **then**
      load its cached records *// resume*
    else
      query s under its rate limit; store records or mark q failed
    end if
  end for
end for

Curate
keep records with a parseable year in $[y_0, y_1]$
merge records sharing the key $\kappa(r)$ *// Eq. (4)*
for all unique record p **do**
  compute $R(p)$ *// Eq. (10)*
  if $R(p) \geq \theta$ **then**
    insert or update p in $\mathcal{D}$
  end if
end for

Acquire and enrich
resolve PDFs for every retained record *// Fig. 4*
optional embeddings (M6); BibTeX (M7); theme labels (M8); exports

of anchor families present in a query q,

$$F(q) = \left\{ f \in \{N, M, D\} : \exists\, a \in \mathcal{A}_f,\ a \sqsubseteq q \right\}, \qquad q \text{ is kept if and only if } F(q) \neq \varnothing, \tag{1}$$

where $a \sqsubseteq q$ means that a occurs in q as a case-insensitive substring.

The vocabularies are crossed through ten template families $\mathcal{Q}_c$ (Table 3). A family is the set of strings produced by one template τ_c over its vocabulary slots,

$$\mathcal{Q}_c = \left\{ \tau_c(t_1, \ldots, t_{r_c}) : t_i \in \mathcal{V}_{c,i} \right\}, \qquad |\mathcal{Q}_c| \leq \prod_{i=1}^{r_c} |\mathcal{V}_{c,i}|, \tag{2}$$

where r_c is the number of vocabulary slots of template τ_c, $\mathcal{V}_{c,i}$ is the vocabulary that fills slot i (an anchor family, a context vocabulary or a list of suffixes), and t_i is a term drawn from it. Equality holds unless case-insensitive duplicates are removed. For example, the mechanics template $a_N\ v_{\text{mech}}\ a_M\ s$ with the suffix $s =$ *cellular uptake* produces *nanoparticle stiffness lipid bilayer cellular uptake*. In the standard mode the families hold 151 393 anchored queries, from 64 in the experiment family to 88 044 in the mechanics family. The user's own query text is placed first in the mixed family, which then holds 121 queries.

The families differ in size by three orders of magnitude, so drawing queries uniformly from their union would return almost only mechanics queries. The sampler instead takes m queries from every family at evenly spaced positions.

Table 2. Anchor families and context vocabularies for the nanoparticle–membrane domain, with term counts and examples. The full lists are in module M2.

Type	Vocabulary	Terms	Example terms
Anchor	Nanoparticle $\mathcal{A}_N$	33	nanoparticle, nanocarrier, liposome, lipid nanoparticle, polymer–lipid hybrid nanoparticle
	Membrane, uptake $\mathcal{A}_M$	29	cell membrane, lipid bilayer, endocytosis, membrane wrapping, cellular uptake
	Drug delivery $\mathcal{A}_D$	17	drug delivery, tumor targeting, tumor penetration, controlled drug release, biodistribution
Context	Mechanics	23	mechanical properties, elasticity, stiffness, Young's modulus, bending rigidity
	Interaction	21	binding affinity, dissociation constant, adhesion energy, potential of mean force
	Design	16	particle size, aspect ratio, zeta potential, PEGylation, surface functionalization
	Simulation	17	molecular dynamics, dissipative particle dynamics, Martini force field, umbrella sampling
	Experiment	16	atomic force microscopy, nanoindentation, dynamic light scattering, flow cytometry
	AI and ML	11	machine learning, graph neural network, Bayesian optimization, formulation optimization

Table 3. Template families and their sizes in the standard generation mode. s denotes a family-specific suffix drawn from a short fixed list.

c	Family	Template τ_c	$\lvert Q_c \rvert$
1	NP–membrane	$a_N\ a_M$	957
2	NP–drug delivery	$a_N\ a_D$	561
3	NP–membrane–delivery	$a_N\ a_M\ a_D$	16 269
4	Interaction	$a_N\ a_M\ v_{\text{int}}$	20 097
5	Mechanics	$a_N\ v_{\text{mech}}\ a_M\ s$, 4 suffixes	88 044
6	Design by simulation	$v_{\text{sim}}\ a_N\ v_{\text{des}}\ s$, 2 suffixes	17 952
7	Experiment	v_{exp} *nanoparticle* s, 4 suffixes	64
8	AI and ML	$v_{\text{AI}}\ a_N\ s$ or $v_{\text{AI}}\ a_N\ v_{\text{des}}\ s$	7260
9	Uptake pathway	20 curated, plus 10 pathways × 5 contexts	69
10	Mixed	capped crosses (plus the user query)	120
	Total		151 393

Writing $Q_c[k]$ for the k-th query of family c in generation order ($k = 0, 1, \ldots$),

$$\mathcal{B}_c = \left\{ Q_c\left[\lfloor i\,|Q_c|/m \rfloor\right] : i = 0, \ldots, m-1 \right\}, \qquad \mathcal{B} = \bigcup_{c=1}^{10} \mathcal{B}_c, \tag{3}$$

which is deterministic, spans each family from its first to its last combination, and gives $|\mathcal{B}| = 10m$ whenever every family holds at least m queries. A family smaller than m contributes all of its queries. The number of queries per source is therefore $Q = 10m$.

Same batch, four sources: In the balanced runs of this paper ($Q = 500$ and $Q = 60$ queries per source), Crossref, arXiv and OpenAlex receive the identical batch $\mathcal{B}$. PubMed receives every query of $\mathcal{B}$ joined to the clause *AND (drug delivery OR cancer OR tumor OR uptake OR membrane OR nanoparticle)*. The engine can also append simulation- and AI-specific queries for arXiv, but only in the unbounded mode, which was not used here. Differences between sources in Sec. 4 therefore come from the sources, not from different query sets.

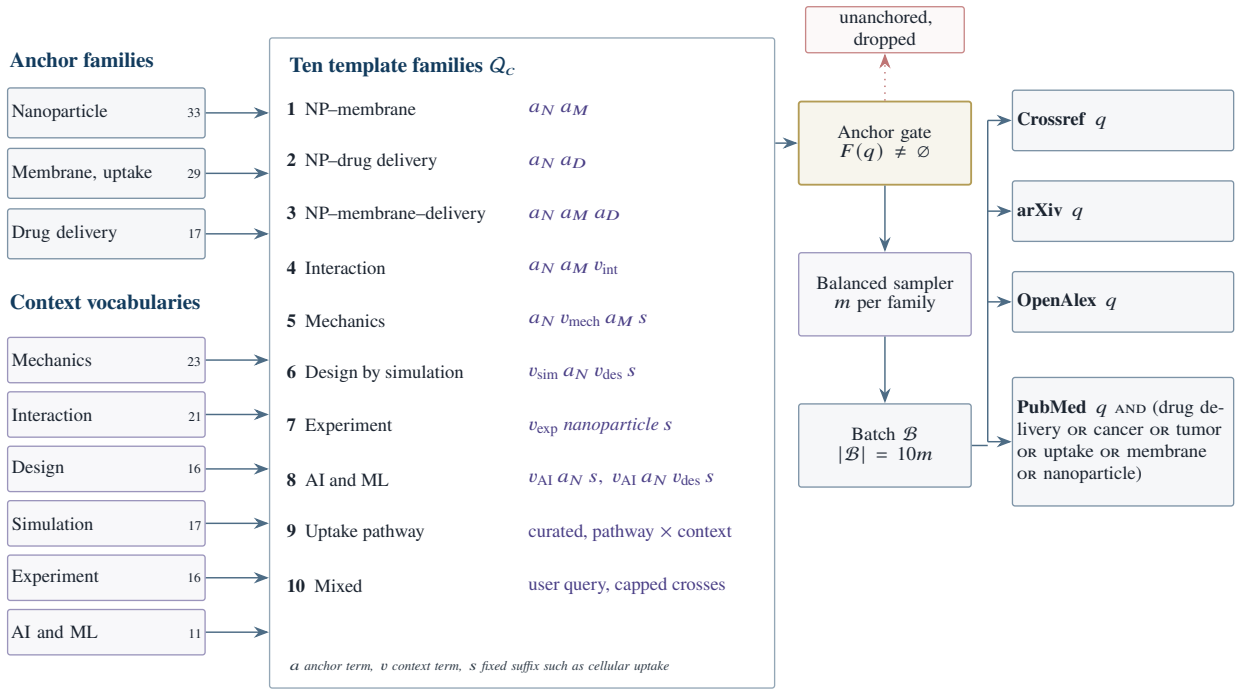


Figure 3. Anchored query generation. Anchor and context vocabularies are crossed through ten template families; an anchor gate (Eq. (1)) drops any string without an anchor term; a deterministic sampler draws m queries from each family (Eq. (3)). In the balanced runs reported here every source receives the same batch $\mathcal{B}$; PubMed receives each query with a Boolean clause that restricts it to biomedical context.

2.3 Metadata retrieval

Four retrievers (M3A–M3D) query the public APIs of Crossref [24], arXiv, PubMed through NCBI E-utilities, and OpenAlex [25], each returning at most $N_{\max}$ records per query. Crossref contributes cross-publisher metadata and publisher full-text links; arXiv contributes preprints with direct PDF links; PubMed contributes biomedical records with PMID and DOI identifiers but no PDF link; OpenAlex contributes multidisciplinary metadata, reconstructed abstracts, and every open-access location it knows, labelled as best, primary or other. Crossref and OpenAlex apply the year window at the source; arXiv and PubMed do not, so the window is enforced again centrally (Sec. 2.4).

A shared rate-limit layer (M3) enforces a minimum interval per source (for example 0.11 s for Crossref and at least 5 s for arXiv), retries transient failures (HTTP 408, 425, 429, 500, 502, 503 and 504, and 403 for OpenAlex) with exponential back-off, uses a longer cool-down for repeated arXiv 429 responses, and defers all remaining OpenAlex queries once its daily quota is exhausted. Each query is checkpointed as started, completed with its records, or failed with its error, and failed queries are exported for a later retry.

2.4 Year validation, merging and deduplication

Retrieved records first pass a year check: a record is kept only if a four-digit year can be parsed from it and lies in $[y_0, y_1]$; the supported window starts in 2000. Records are then merged on a canonical key,

$$\kappa(r) = \begin{cases} (\text{doi},\ \nu(d_r)) & \text{if record } r \text{ has a DOI } d_r, \\ (\text{title},\ \nu(t_r)) & \text{otherwise,} \end{cases} \tag{4}$$

where ν lower-cases and trims a string, and for a title also collapses internal white space. When two records share a key, the first one is kept and its empty fields (abstract, PMID, PDF URL, journal, OpenAlex identifier and others) are filled from the second; the OpenAlex PDF locations of both are pooled. The database applies a second check on insertion, matching a normalised DOI or a title stripped of punctuation, so a record already stored by an earlier run is updated, not

Table 4. Concept groups of the relevance function, with pattern counts, weights w_g and example patterns. The out-of-scope group subtracts a penalty (Eq. (9)).

Concept group	Symbol	$\lvert\Pi_g\rvert$	w_g	Example patterns
Mechanical properties	mech	14	4.5	stiffness, Young's modulus, deformability, AFM
Nanoparticle types	NP	21	4.0	nanoparticle, liposome, PLGA, dendrimer, exosome
Membrane, uptake and delivery	mem	20	4.0	lipid bilayer, endocytosis, drug delivery, EPR effect
Membrane interaction	int	11	3.0	wrapping energy, adhesion energy, Helfrich
Simulation and free energy	sim	16	2.5	molecular dynamics, coarse-grained, umbrella sampling
Targeting and binding	tgt	11	2.0	ligand–receptor, multivalency, PEGylation, transferrin
AI for nanoparticle design	ai	7	2.0	machine learning nanoparticle, QSAR
Experimental characterisation	exp	12	1.5	dynamic light scattering, TEM, flow cytometry
Out of scope	oos	22	–	electro-optic, battery, solar cell, photocatalysis

duplicated, and its relevance score is replaced by the latest one. This check falls back to the title even when both records carry different DOIs.

Source attribution is order dependent: A merged record keeps the source label of the first record seen. Sources are queried in the fixed order Crossref, arXiv, PubMed, OpenAlex, so a paper returned by all four is credited to Crossref. The per-source counts in Sec. 4.3 therefore measure *first* attribution, not unique contribution. The merged list of contributing sources is not stored with the record, but every source's raw response to every query is kept in the checkpoint store, and Sec. 4.2 reconstructs the overlap between sources from it.

2.5 Domain-weighted relevance scoring

The ranker (M4) scores each unique record from a single lower-cased text $x(p)$ that concatenates its title, abstract, keywords, journal name, source label, DOI and URL; in practice keywords are empty and the journal name is supplied only by OpenAlex. Eight positive concept groups g and one out-of-scope group are each defined by a list of regular expressions Π_g (7 to 22 patterns per group; Table 4). We write $[S]$ for the Iverson bracket, which equals 1 when the statement S is true and 0 otherwise. The signal of a group is the number of *distinct* patterns $\pi \in \Pi_g$ that occur at least once in $x(p)$,

$$n_g(p) = \sum_{\pi \in \Pi_g} \left[\, \pi \text{ matches } x(p) \,\right], \tag{5}$$

so a term repeated many times counts once, and a group can contribute through several different terms.

A record is scored only if it carries a nanoparticle, membrane or interaction signal; otherwise its score is zero:

$$G(p) = \left[\, n_{\mathrm{NP}} > 0 \;\vee\; n_{\mathrm{mem}} > 0 \;\vee\; n_{\mathrm{int}} > 0 \,\right]. \tag{6}$$

For a gated record, the raw score sums capped group signals, adds bonuses for co-occurring concepts and subtracts an out-of-scope penalty.

Table 5. Heuristic interpretation of score bands.

Score band	Reading
$0.80 \le R \le 1.00$	Core paper
$0.60 \le R < 0.80$	Strongly relevant
$0.40 \le R < 0.60$	Useful background
$0.15 \le R < 0.40$	Weak
$R < 0.15$	Excluded

Domain-weighted relevance score M4

$$R_{\text{raw}}(p) = \sum_g w_g \, \min\big(n_g, 4\big) \; + \; B(p) \; - \; P(p) \tag{7}$$

$$\begin{aligned} B(p) = {} & 5\,[\text{NP} \wedge \text{mech}] + 5\,[\text{NP} \wedge \text{mem}] + 4\,[\text{mech} \wedge \text{mem}] + 6\,[\text{NP} \wedge \text{mech} \wedge \text{mem}] \\ & + 3\,[\text{sim} \wedge (\text{NP} \vee \text{mem})] + 2\,[\text{tgt} \wedge \text{mem}] \end{aligned} \tag{8}$$

$$P(p) = 5 \, \min\big(n_{\text{oos}}, 4\big) \times \begin{cases} \frac{1}{2} & \text{if } n_{\text{NP}} > 0 \text{ and } n_{\text{mem}} > 0, \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

$$R(p) = G(p) \cdot \min\Big\{1, \; \max\Big(0, \; \frac{R_{\text{raw}}(p)}{40}\Big)\Big\} \tag{10}$$

The sum in Eq. (7) runs over the eight positive groups of Table 4, w_g is the weight of group g, and each signal is capped at 4 so that no single group dominates. $B(p)$ is the co-occurrence bonus, $P(p)$ the out-of-scope penalty, and n_{oos} the signal of the out-of-scope group. Inside the Iverson brackets of Eq. (8) a group symbol stands for the statement that its signal is non-zero, so $[\text{NP} \wedge \text{mech}] = 1$ when $n_{\text{NP}} > 0$ and $n_{\text{mech}} > 0$. The divisor 40 in Eq. (10) sets the raw score at which R saturates, and the outer min and max keep R in $[0, 1]$. The bonuses encode the target of this study, which is mechanics-informed nanoparticle design for delivery rather than nanomedicine in general: the strongest bonus rewards the joint presence of nanoparticle, mechanical and membrane terms. The penalty is halved for records that carry both a nanoparticle and a membrane or delivery signal, because such papers may mention an out-of-scope term in passing. Scores are rounded to three decimals, and records with $R(p) < \theta$ are discarded before download. We use $\theta = 0.15$. Table 5 gives a heuristic reading of the score bands, used only for reporting.

Saturation and quantisation of R: Two properties of Eq. (10) shape every distribution reported later. *Saturation:* the largest attainable raw score is $4 \sum_g w_g + 25 = 94 + 25 = 119$, but R reaches 1 at $R_{\text{raw}} = 40$, so all records above 40 tie at $R = 1$. A title and a one-sentence abstract that name two nanoparticle terms, two mechanical terms and four delivery terms already score $R_{\text{raw}} = 33 + 20 = 53$ (Appendix A). *Quantisation:* all weights, bonuses and penalties are multiples of 0.5, so R lies on a lattice of step $0.5/40 = 0.0125$; and because the dominant groups add 4 raw points per matched pattern, typical scores cluster at steps of $4/40 = 0.1$; the acquired PDFs of the main run concentrate in the bins starting at 0.20, 0.30, 0.40 and 0.50 (Fig. 9). Histograms of R therefore show a regular comb that reflects the weights, not structure in the literature.

2.6 Cascading PDF acquisition

Metadata rarely guarantee full text: links may be missing, point to a landing page, or be refused. The downloader (M5) therefore builds, for each retained record p, an ordered list of candidate URLs $C(p)$ and tries them in turn until one returns a valid PDF (Fig. 4). The list is built in five steps: (1) the direct PDF link supplied by the retriever; (2) every OpenAlex location that lists a PDF URL, best open-access location first; (3) Crossref full-text links whose declared content type is PDF; (4) for any record whose URL points to an arXiv abstract page, that URL rewritten to the PDF

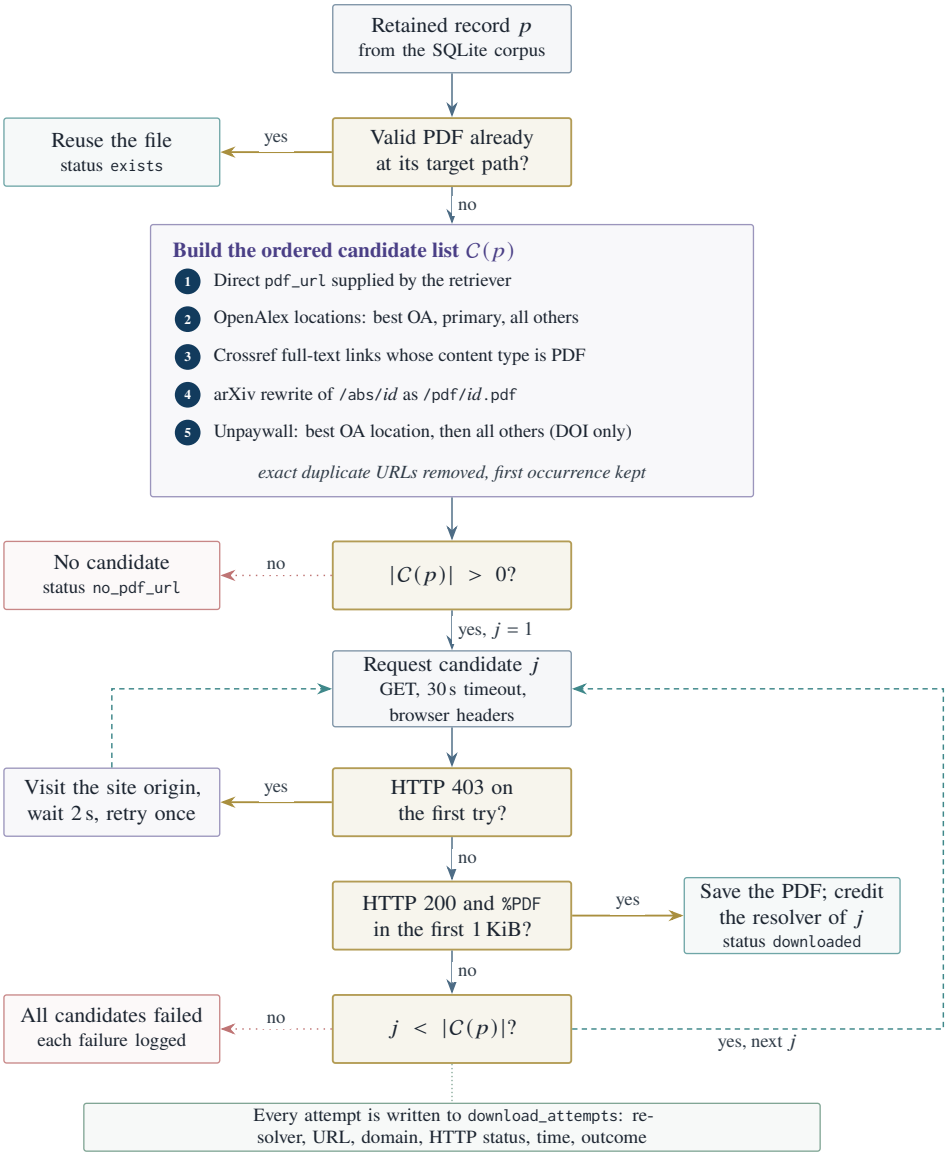


Figure 4. The cascading PDF resolver (M5), drawn from the code path. A file counts as a PDF only if the response is HTTP 200 and the bytes `%PDF` occur in its first kilobyte. A first HTTP 403 triggers one retry after the site origin is visited. The resolver of the first successful candidate is credited with the download.

endpoint (`arxiv.org/abs/<id>` becomes `arxiv.org/pdf/<id>.pdf`); and (5) when a DOI exists, the best and then every other open-access location returned by Unpaywall [26]. Exact duplicate URLs are removed, keeping the first occurrence.

Requests use browser-like headers with a contact address, a 30 s timeout and redirects. A first HTTP 403 triggers a visit to the site origin, a 2 s wait and one retry. A response is accepted only if its status is 200 and the PDF signature `%PDF` occurs in its first 1024 bytes; HTML pages and other content are rejected and logged with their domain and status. Files are named from the sanitised title, truncated to 80 characters (or from the DOI when the title is empty), and a valid file already at that path is reused. Each attempt is written to the database with its resolver label, candidate rank, URL, domain, HTTP status, elapsed time and outcome. The module uses only openly accessible locations and does not bypass publisher access controls.

For reporting, the resolver labels are collapsed into four provenance groups: Crossref, arXiv, OpenAlex and Unpaywall. Step (1) has no group of its own, because a direct link carries the OpenAlex location label when a merged OpenAlex record supplied one and the label of the first-attributed source otherwise. Since merging fills a missing PDF link and its label independently, a direct link can be credited to a group other than the host it points to (an arXiv URL labelled as an

OpenAlex location, for example); the group shares reported in Sec. 4.4 are therefore approximate. The acquisition of a set of N_{ret} retained records, of which N_{url} have at least one candidate and N_{pdf} yield a validated PDF, factorises as

Acquisition yield coverage times success

$$\eta = \frac{N_{\text{pdf}}}{N_{\text{ret}}} = \underbrace{\frac{N_{\text{url}}}{N_{\text{ret}}}}_{\phi\ \text{(URL coverage)}} \times \underbrace{\frac{N_{\text{pdf}}}{N_{\text{url}}}}_{\psi\ \text{(download success)}} . \tag{11}$$

ϕ is a property of the metadata: whether any index knows an open location. ψ is a property of the web: whether that location actually serves a PDF to a script.

2.7 Enrichment

Citations (M7). Each retained record receives a BibTeX entry. When a DOI exists, the entry is obtained by DOI content negotiation; otherwise, or if negotiation fails, a local entry is built from the stored title, authors, year and URL, typed `@article` for records from Crossref, PubMed or OpenAlex and `@misc` otherwise. Entries carry unique citation keys and are exported as one `.bib` file.

Thematic labels (M8). The keyword classifier counts occurrences of the terms of eleven predefined themes (for example *drug delivery and cancer*, *cellular uptake*, *mechanics and elasticity*) in the title and abstract, and assigns the theme with the largest count; a record with no matching term is labelled *unclassified*. An optional semantic mode clusters the stored embeddings with k-means and names each cluster by its dominant keyword theme. The labels support browsing of the corpus; they are not evaluated in this paper.

Embeddings (M6, optional). Title and abstract can be encoded with `text-embedding-3-large` into $\mathbf{e}_p \in \mathbb{R}^{3072}$ and stored per paper. A natural-language query $\mathbf{q}$ is embedded with the same model, and records are ranked by cosine similarity $\text{sim}(\mathbf{q}, \mathbf{e}_p) = \mathbf{q} \cdot \mathbf{e}_p / (\|\mathbf{q}\|\, \|\mathbf{e}_p\|)$. The same similarity supports related-paper search, and the optional thematic clustering of M8 runs k-means on the embeddings. Embeddings are not used by the ranking of Eq. (10).

2.8 Adapting the pipeline to another domain

Orchestration, retrieval, rate control, storage, deduplication, PDF acquisition, citation export and analytics are domain independent. A new domain replaces the *vocabulary pack*, which is at present a set of constants edited in the module source: the anchor families, context vocabularies and templates of M2; the concept groups, weights, bonuses and out-of-scope list of M4; and the themes of M8. The out-of-scope list matters: the nanoparticle configuration penalises *electro-optic* and *lithium niobate*, so a move to electro-optic materials must replace M4 and not only the query anchors.

3. Experimental design

3.1 Runs

We report three runs (Table 6). The main run, Q500–NM, targets nanoparticle–membrane interaction in cancer drug delivery with $m = 50$ queries from each of the ten families, that is 500 queries per source and 2000 source-specific queries, each returning at most 25 records. Two smaller runs, Q60–NM and Q60–EO, use 60 queries per source ($m = 6$) and differ only in the vocabulary pack. Q60–EO is a cross-check on a domain unrelated to drug delivery: Q60–EO replaces the query vocabularies and templates (M2), the concept groups and out-of-scope list of the ranker (M4) and the themes (M8) with an electro-optic materials configuration, and leaves the pipeline logic unchanged. In Q60–EO, 55 of the 60 PubMed queries completed. All runs use the publication window 2000–2026 and the threshold $\theta = 0.15$.

3.2 Metrics

Stage retention. For each stage we report the number of records that leave it and the fraction of its input that this represents.

Table 6. Run configurations and raw retrieval. NM: nanoparticle–membrane; EO: electro-optic materials. Per-source entries are in the order Crossref / arXiv / PubMed / OpenAlex. PubMed returns fewer records than the other sources because its queries carry the Boolean clause of Sec. 2.2.

Setting	Q500–NM	Q60–NM	Q60–EO
Domain	NP–membrane	NP–membrane	Electro-optic
Queries per family, m	50	6	6
Completed queries per source	500/500/500/500	60/60/60/60	60/60/55/60
Records per query, $N_{\max}$	25	25	25
Raw records per source	12 500/12 500/5591/12 327	1500/1500/771/1488	1500/1500/441/1483
Raw records, total	42 918	5259	4924
Retained records, $R \geq \theta$	5863	1360	1349
Publication window	2000–2026	2000–2026	2000–2026
Relevance threshold, θ	0.15	0.15	0.15

Source overlap. The merged record keeps only its first source (Sec. 2.4), so overlap is reconstructed from the checkpoint store, which holds every source's raw response to every query. A retained record is counted as returned by a source when that source returned a year-valid record with the same normalised DOI or the same normalised title. For each source we report the retained records it returned, the number returned by no other source, and the share of the final set covered by the best combination of $n = 1, 2, 3$ sources.

Acquisition. URL coverage ϕ, download success ψ and yield η follow Eq. (11), overall, per first-attributed source and per relevance band (Table 5). For a resolver group j with A_j download attempts and S_j successes we report the conditional success rate and the share of acquired PDFs,

$$s_j = \frac{S_j}{A_j}, \qquad \sigma_j = \frac{S_j}{\sum_{j'} S_{j'}}. \tag{12}$$

Because the cascade stops at the first success, a resolver late in the order sees only the records that the earlier ones failed to resolve. s_j is therefore conditional on the cascade order and is not a head-to-head comparison of resolvers.

Precision of the highest-scored records. The 100 highest-scored retained records of Q500–NM were labelled by hand. A record was labelled relevant ($y_i = 1$) when it studies nanoparticles or nanocarriers in relation to membrane interaction, cellular uptake, cancer drug delivery, targeting, ligand–receptor binding, biodistribution, or the effect of mechanics or design on these; a record about nanoparticles without a meaningful link to membrane interaction or drug delivery was labelled irrelevant ($y_i = 0$). With $n = 100$ labelled records, the precision $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is reported with the 95 % Wilson score interval [27, 28],

$$\frac{\hat{p} + \dfrac{z^2}{2n} \pm z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n} + \dfrac{z^2}{4n^2}}}{1 + \dfrac{z^2}{n}}, \qquad z = 1.96, \tag{13}$$

where z is the 0.975 quantile of the standard normal distribution. This interval, unlike the normal approximation, stays inside $[0, 1]$ when $\hat{p}$ is close to 1.

4. Results

4.1 From raw records to a PDF corpus

The 2000 queries of Q500–NM could return at most 50 000 records; they returned 42 918 (85.8 % of the ceiling). Year validation removed 1149 records without a parseable year and 252 records outside 2000–2026, leaving 41 517. Merging on the key of Eq. (4) reduced these to 14 213 unique records. Almost all of this reduction is repetition within a source: 500 related queries sent to the same source return the same papers many times, whereas only 315 of the 14 213 unique records (2.2 %) were returned by more than one source. The relevance threshold retained 5863 records (41.3 % of the

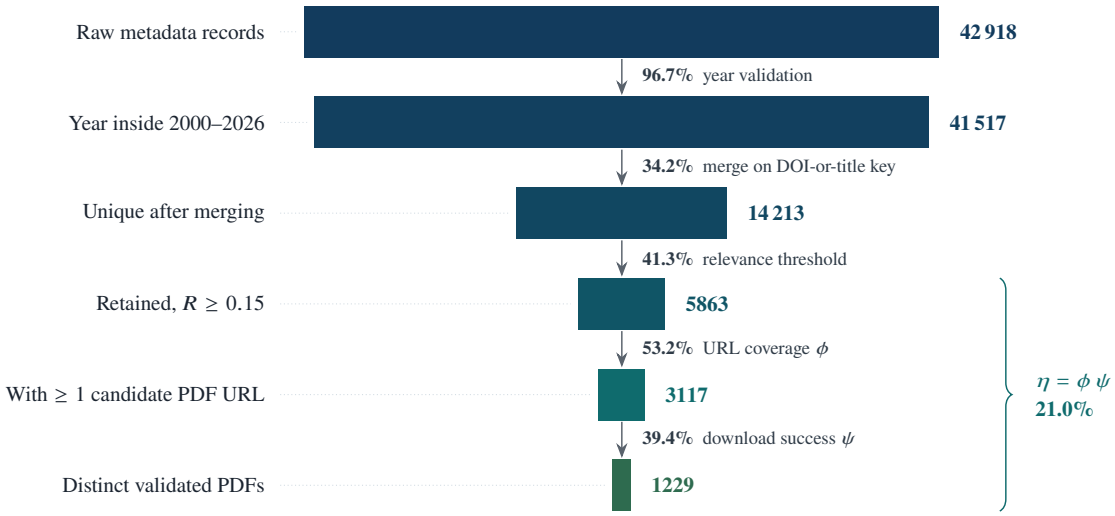


Figure 5. Record counts through the Q500–NM run, with the fraction of each stage's input that survives it. Bar width is proportional to the count. The last two transitions are the URL coverage ϕ and the download success ψ of Eq. (11); their product is the acquisition yield $\eta = 21.0\,\%$ of retained records (2.9 % of raw records).

unique records), 3117 of which had at least one candidate PDF URL, and 1229 distinct validated PDFs were acquired (Fig. 5).

> **Where the relevant records are lost:** The threshold removes more records in number (8350), but acquisition loses the largest fraction, and it loses records already judged relevant. Of the 5863 retained records, 4634 (79.0 %) produced no PDF: 2746 had no candidate URL in any index, and 1888 had candidates that all failed. Open-access availability, not the search, decided how much of the relevant set became full text.

4.2 Complementarity of the four sources

The four sources return largely disjoint sets of relevant records (Fig. 6). Of the 5863 retained records, 5581 (95.2 %) were returned by one source only, 265 by two and 17 by three; none was returned by all four. Every source contributed records that no other source returned: PubMed 2215 of the 2387 retained records it returned, OpenAlex 1777 of 2017, Crossref 1223 of 1371 and arXiv 366 of 387. The best single source, PubMed, covers 40.7 % of the retained set; the best pair (PubMed and OpenAlex) covers 72.8 % and the best three (Crossref, PubMed and OpenAlex) 93.8 %. For the acquired PDFs the best single source is OpenAlex (40.0 %), and arXiv belongs to the best pair (70.2 %) and the best three (85.6 %), because almost every arXiv record becomes full text.

> **What the overlap measures:** The overlap is measured at a depth of 25 records per query. Each source ranks its own matches, so the same query draws different top-25 lists from different sources, and deeper retrieval would raise the overlap. A preprint and its journal version also carry different identifiers and often different titles, and are then counted as distinct records. The result therefore shows that, at this depth, no single source returns more than 41 % of the retained set; it does not estimate the overlap of the underlying indices.

4.3 Coverage and success by source

Table 7 and Fig. 7a break the retained records down by first-attributed source. PubMed supplied the most retained records (2335, 39.8 %), followed by OpenAlex (1777), Crossref (1369) and arXiv (382, 6.5 %). At acquisition the order changes: arXiv rises from last to second and PubMed falls from first to third. Almost every arXiv record carries a PDF link ($\phi = 0.997$) and almost every link works ($\psi = 0.997$), so arXiv delivers 30.9 % of the PDFs from 6.5 % of the

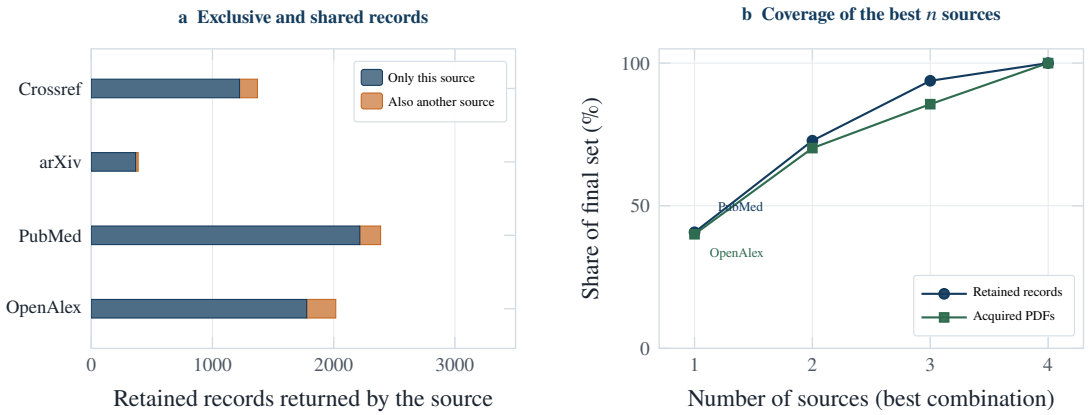


Figure 6. Complementarity of the sources in Q500–NM. (a) Retained records returned by each source, split into records that no other source returned and records that at least one other source also returned. A record counts as returned by a source when that source returned a year-valid record with the same normalised DOI or title. (b) Share of the 5863 retained records and of the 1229 acquired PDFs covered by the best combination of n sources; the labels name the best single source.

Table 7. Retained records, URL coverage, download success and yield by first source in Q500–NM (Eq. (11)). A record returned by several sources is credited to the first in the order Crossref, arXiv, PubMed, OpenAlex.

Source	Retained	With URL	PDFs	ϕ	ψ	η
Crossref	1369	627	195	0.458	0.311	0.142
arXiv	382	381	380	0.997	0.997	0.995
OpenAlex	1777	1401	443	0.788	0.316	0.249
PubMed	2335	708	211	0.303	0.298	0.090
All	5863	3117	1229	0.532	0.394	0.210

records. PubMed records carry no PDF link of their own, and only 30.3 % gain one from merged OpenAlex locations or from Unpaywall.

Why coverage, not retrieval, sets the yield: For Crossref, OpenAlex and PubMed records in Q500–NM the download success lies within two percentage points (ψ = 0.298 to 0.316), while their coverage varies from 0.30 to 0.79 (Fig. 7b). Of the 3270 download attempts outside the arXiv group, 2410 (73.7 %) failed: 70.0 % of these failures were HTTP 403 refusals, 19.8 % HTML pages served with status 200 in place of a PDF, 5.9 % HTTP 404 responses, and 4.3 % other errors. The losses therefore sit with the hosts behind the URLs rather than with the metadata. Among these sources, yield tracks coverage: OpenAlex leads because more of its records carry at least one location (ϕ = 0.79), and PubMed trails because its records arrive without any. In the smaller Q60–NM run the same three sources span a wider range (ψ = 0.30 to 0.40), so the equality is approximate. arXiv sits apart because it serves a stable, script-friendly PDF endpoint for every record it indexes.

4.4 Resolver behaviour

The cascade made 3641 download attempts (Table 8, Fig. 8). OpenAlex locations produced the largest share of PDFs (490, 39.9 %) at a conditional rate of 26.9 %. The arXiv group resolved 369 of 371 attempts; both failures were HTTP 404 responses. Unpaywall, the last step, received only records whose earlier candidates had all failed or were absent, and still supplied 310 PDFs (25.2 %); without it the run would have ended at 919 PDFs and a yield of 15.7 %. Attempts in the Crossref group, which includes direct Crossref links as well as full-text links, succeeded in 18.0 % of cases. Direct links do not appear as a group of their own, and because their labels follow the merged metadata rather than the URL host, the group shares are approximate (Sec. 2.6).

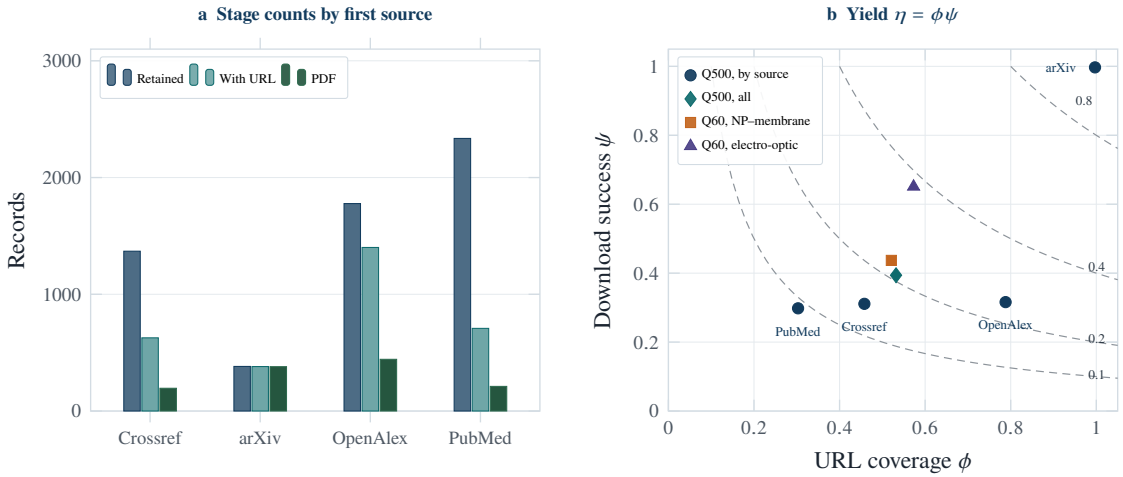


Figure 7. Acquisition by source. (a) Retained records, records with at least one candidate URL, and acquired PDFs, by first-attributed source (Q500–NM). (b) The same sources in the plane of URL coverage ϕ and download success ψ; dashed curves are lines of constant yield $\eta = \phi\psi$. In Q500–NM the three non-preprint sources share $\psi \approx 0.30$ and differ mainly in ϕ. The run totals of Q500–NM and both Q60 runs are shown for comparison.

Table 8. Download attempts and successes by resolver group in Q500–NM (Eq. (12)). Rates are conditional on the cascade order.

Resolver group	Attempts A_j	Successes S_j	s_j (%)	σ_j (%)
Crossref	334	60	18.0	4.9
arXiv	371	369	99.5	30.0
OpenAlex	1819	490	26.9	39.9
Unpaywall	1117	310	27.8	25.2
Total	3641	1229	33.8	100.0

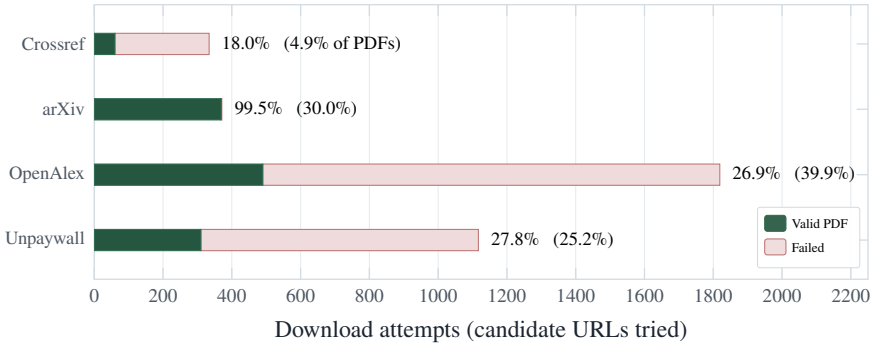


Figure 8. Download attempts by resolver group, split into validated PDFs and failures (Q500–NM). Labels give the conditional success rate s_j and, in parentheses, the share σ_j of all acquired PDFs.

4.5 Relevance and acquisition

Precision of the highest-scored records. Because the score saturates (Sec. 2.5), 169 retained records share the maximum $R = 1$, 116 of them first attributed to PubMed. The 100 labelled records all have $R = 1$: they are 100 of these 169 tied records, and their order within the tie carries no information. Of the 100, 93 were relevant, a precision of 0.93 with a 95 % Wilson interval of $[0.86, 0.97]$ (Eq. (13)). We report one value rather than precision at several rank cut-offs, because a cut-off inside a tie is arbitrary.

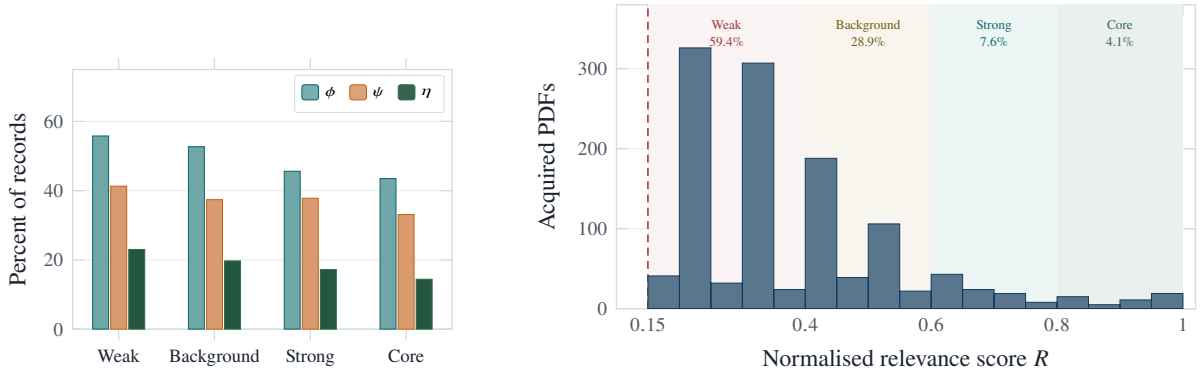


Figure 9. Relevance and acquisition in Q500–NM. (a) URL coverage ϕ, download success ψ and yield η of the retained records in each score band of Table 5; band sizes are 3169, 1801, 546 and 347 records. (b) Relevance scores of the 1229 acquired PDFs in bins of 0.05, shaded by band; the dashed line is the threshold θ.

Scores of the acquired corpus. The rest of the corpus scores much lower (Fig. 9b). Of the 1229 acquired PDFs, 59.4 % fall in the weak band $0.15 \le R < 0.40$, 28.9 % in the background band, 7.6 % in the strong band and 4.1 % in the core band. The comb in the histogram reflects the 0.1 step of the dominant weights (Sec. 2.5), not a property of the literature.

Yield by relevance band. Acquisition favours the weaker records (Fig. 9a). The yield falls from 23.0 % in the weak band to 19.7 %, 17.2 % and 14.4 % in the background, strong and core bands, and both factors fall with it: coverage from 0.56 to 0.44 and download success from 0.41 to 0.33. Part of the trend is composition, since arXiv supplies 8.4 % of the weak records but only 3.7 % of the core records. Excluding arXiv-attributed records does not remove it: the yield of the remaining records still falls from 16.0 % in the weak band to 11.4 % in the core band, mainly through coverage (0.52 to 0.41).

> **Evidence:** **Determined:** 93 of 100 records at the maximum score are relevant (Wilson lower bound 0.86), and the records with the highest scores are the least likely to be acquired as full text. **Not determined:** the precision of the corpus as a whole. Most acquired PDFs score in the weak band, which the labelled sample does not reach, and recall (the fraction of relevant literature that the pipeline finds) was not measured. **Implication:** the measured precision describes the records a user would read first, and the acquired corpus is weighted towards the less relevant end of the retained set.

4.6 Citations and publication years

Citation export succeeded for all retained records: 5666 entries (96.6 %) came from DOI content negotiation and 197 were built locally from the stored metadata. The acquired PDFs span 2000–2026 and rise from fewer than ten per year through 2008 to 117 in 2020 (Fig. 10). The 154 PDFs from 2026, a partial year, are not a trend. Each query returns at most 25 records per source, and this cut-off is not neutral in time: PubMed, the largest source, is queried without an explicit sort order, and the order of the E-utilities default can differ from the relevance ranking of web PubMed. Setting the sort order explicitly removes this ambiguity.

4.7 Cross-check on electro-optic materials

To test whether the pipeline depends on its first domain, we ran it on a field with a different vocabulary, different journals and a different preprint culture. Electro-optic materials change their refractive index in an applied electric field (the Pockels effect) and are the basis of the modulators that encode electrical signals onto light in optical communication (Fig. 11) [29, 30]. Their literature reports material constants, such as the Pockels coefficient r_{33} and the extraordinary refractive index n_e, and device figures of merit, such as the half-wave voltage V_π, the voltage–length product $V_\pi L$, the modulation bandwidth and the optical loss. None of these terms overlaps with the drug-delivery vocabulary, and

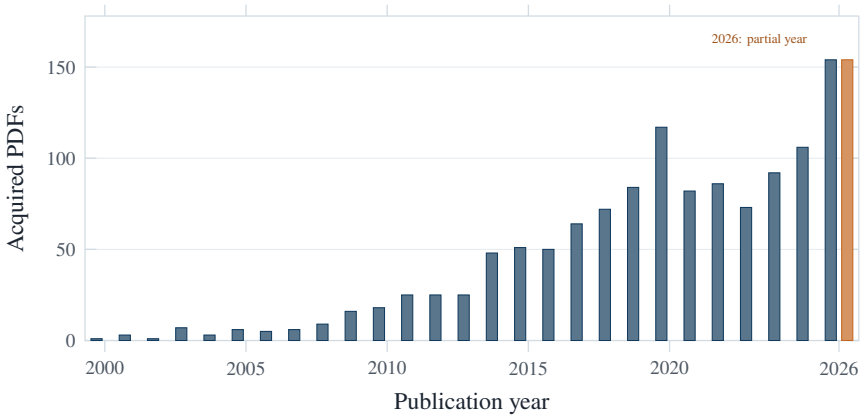


Figure 10. Publication year of the 1229 PDFs acquired in Q500–NM; 2026 is a partial year.

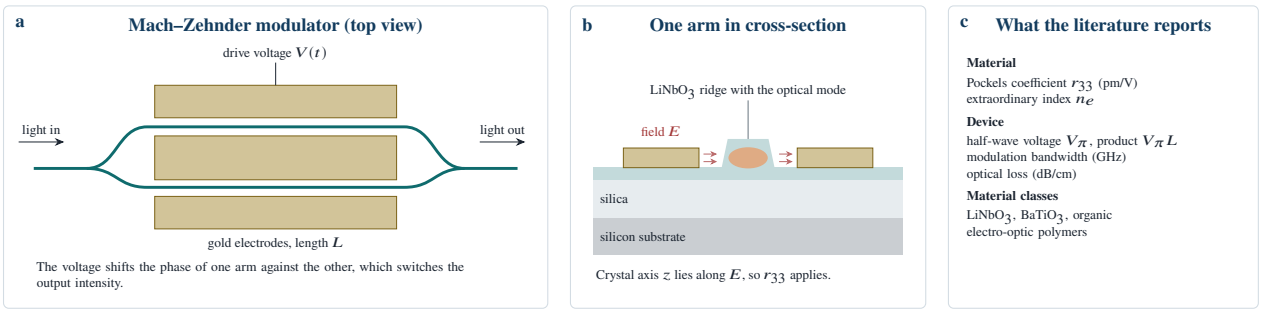


Figure 11. The electro-optic cross-check domain. (a) A Mach–Zehnder modulator: light is split into two waveguide arms, a drive voltage $V(t)$ applied over electrodes of length L shifts the optical phase, and recombination turns the phase difference into an intensity change. (b) Cross-section of one arm on thin-film lithium niobate ($LiNbO_3$); the in-plane field E acts on the optical mode through the Pockels coefficient r_{33}. (c) Quantities that papers in this domain report: n_e is the extraordinary refractive index and V_π the voltage that shifts the phase by π. Adapted in simplified form from the device geometry of refs. [29, 30].

electro-optic is on the out-of-scope list of the nanoparticle ranker (Table 4), so the cross-check requires a new vocabulary pack for M2, M4 and M8 and nothing else.

Q60–NM and Q60–EO ran the same pipeline logic with different vocabulary packs (Table 6) and retained similar numbers of records (1360 and 1349). The yield rose from 22.8 % for nanoparticle–membrane interaction to 37.3 % for electro-optic materials (Fig. 12a). Decomposing it with Eq. (11) shows where the gain came from: coverage rose only from 0.52 to 0.57, while download success rose from 0.44 to 0.65. Most of the difference is composition. arXiv supplies 24.4 % of the retained electro-optic records against 8.5 % of the nanoparticle records, and almost all of its records become PDFs: the electro-optic run obtained 293 of its 503 PDFs (58.3 %) through the arXiv group, against 112 of 310 (36.1 %) for the nanoparticle run (Fig. 12b). Download success outside arXiv-attributed records also rose, from 0.33 to 0.40. The smaller nanoparticle run agrees with the main run in coverage (0.52 against 0.53) and lies four points higher in download success (0.44 against 0.39).

What the cross-check shows: Replacing the vocabulary pack was sufficient to run the pipeline in a second domain without any change to its logic. The higher yield in that domain comes from download success, largely because a larger share of the literature is on arXiv, and not from better metadata coverage. The relevance of the electro-optic corpus was not labelled, so the cross-check shows that the acquisition stage carries over; whether the ranking carries over equally well remains to be measured.

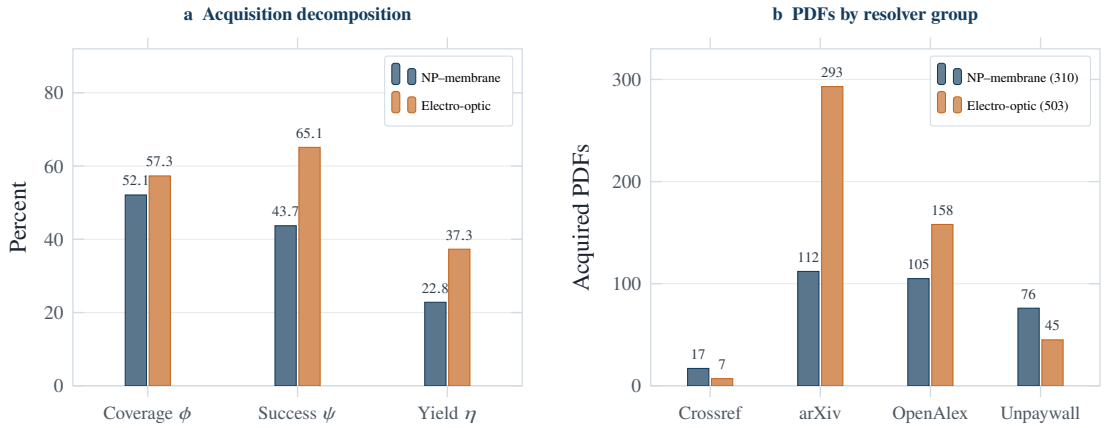


Figure 12. Cross-check on electro-optic materials, with 60 queries per source. (a) URL coverage ϕ, download success ψ and yield η for the nanoparticle–membrane and electro-optic runs. (b) Acquired PDFs by resolver group.

5. Discussion

5.1 What the measurements show

The pipeline turns a domain vocabulary into a ranked, deduplicated, citable and provenance-logged PDF corpus without manual filtering, and each stage carries a number that can be compared across runs, domains and pipelines. This is the point of reporting corpus construction as a stage in its own right: the corpus bounds the recall of every retrieval and extraction step built on it.

Three results follow. First, the sources are complementary at the retrieval depth used: 95 % of the retained records came from one source only, and no single source covered more than 41 % of the retained set or of the acquired PDFs (Fig. 6). A single-source design would have lost more than half of the corpus. Second, acquisition, not the search, loses the largest share of the relevant records (Fig. 5), and the factorisation of Eq. (11) separates its two causes. In the main run, download success outside arXiv is close to 0.30 whatever the source, and coverage decides the yield (Fig. 7b); the electro-optic cross-check moves the other factor, because more of that literature is on arXiv. Coverage improves with more open-location indices; success improves with better handling of the hosts that those indices point to. Third, acquisition is not neutral with respect to relevance: the yield falls from 23 % in the weakest score band to 14 % in the strongest (Fig. 9a), so a full-text corpus built from open-access copies under-represents the records that the ranker judges most relevant. A downstream extractor that reads only the acquired PDFs inherits this skew.

5.2 Limitations

Limitations and Scope: **Evaluation scope.** Precision was measured on 100 of the 169 records that share the maximum score. Precision within each score band, corpus-level precision and recall were not measured, and the electro-optic corpus was not labelled.

Lexical scoring. The relevance function is hand-weighted and lexical. It reads the DOI and URL and, for OpenAlex records, the journal name as well as the title and abstract, so a venue such as a drug-delivery journal can add signal by itself. Records without an abstract are scored from their title and identifiers alone, and PubMed records keep only the first section of a structured abstract and carry no author list.

Open-access bias. Only openly accessible PDFs are acquired, so the corpus over-represents preprints and open-access venues relative to the literature retrieved, and, as shown above, under-represents the highest-scored records.

Overlap at fixed depth. Source overlap was measured with 25 records per query and with matching on DOI or normalised title, so it is a property of this retrieval depth, and preprint and journal versions of one paper count as two records (Sec. 4.2).

Non-article records. Crossref also indexes peer-review documents, meeting abstracts and notices. Among the 5863 retained records of the main run, 81 (1.4 %) are of this kind: 35 decision letters, author responses or editor's evaluations, 30 meeting abstracts and 16 corrections or retractions. Three of them reached the PDF corpus.

Deduplication by title. Records without a DOI are merged on their normalised title, and the database check falls back to the title even when two records carry different DOIs. Two different papers with identical titles are therefore merged into one.

File-name collisions. PDF file names are the first 80 characters of the sanitised title, and an existing valid file is reused. Papers whose titles differ only after the 80th character therefore share one file. In each nanoparticle run this affected one pair, Parts I and II of the same study series [31, 32]; all counts of PDFs in this paper are of distinct files.

5.3 Extensions within reach

Most of these limitations have short fixes that need no change to the design. A stratified sample of labelled records from each score band would give corpus-level precision and a calibrated threshold. A work-type filter on Crossref (`type:journal-article` and preprint types) removes review documents and meeting abstracts. Appending a DOI hash to each file name removes the collision, labelling each download by the host of its URL makes the resolver shares exact, and storing the list of contributing sources with each record makes overlap a direct output of the run. The two acquisition levers suggest two additions to the cascade: a PubMed Central and Europe PMC step keyed on the PMCID would raise coverage where it is lowest, and reading the `citation_pdf_url` tag of a returned landing page could raise download success for hosts that serve an HTML page in place of the PDF. Finally, the hand-tuned weights of Eq. (7) can be replaced by a ranker trained on labelled records, with the lexical score kept as an interpretable baseline.

Open Question: What recall does the pipeline achieve? Answering it requires an independent reference set, for example the included studies of a published systematic review in the domain, against which the retrieved and the acquired records can both be checked.

5.4 A domain-agnostic approach and outlook

Only the vocabulary pack is specific to a domain (Sec. 2.8). The query engine, the four retrievers, rate control, deduplication, the relevance function's structure, the PDF cascade and the provenance log are reused unchanged, as the electro-optic cross-check showed (Sec. 4.7). The same core can therefore be pointed at other fields of science and engineering, such as electrode and electrolyte materials for energy storage, and at literatures outside science, such as law, policy or economics, wherever Crossref, OpenAlex or a comparable index covers the field (Fig. 13). PubMed and arXiv contribute little outside biomedicine and the physical sciences, so for such fields the set of sources is part of the configuration as well.

Corpus construction is the first step of a longer chain. Once the PDFs are acquired, AI tools can be built on them for data mining: language models that extract property values, synthesis and processing conditions, and measurement methods from the full text, together with the source of each value. Harmonised into consistent units and linked back to the source papers, these values form an AI-ready database, the missing resource identified in Sec. 1 [10]. For nanoparticle drug delivery, such a database would relate size, shape, stiffness and surface chemistry to circulation, tumour accumulation and uptake (Fig. 1); for energy and electro-optic materials, it would relate composition and structure to the figures of merit of devices. Machine-learning models trained on these databases can then support AI-driven materials discovery, in the way that curated databases such as the Materials Project already support inorganic materials design [1]. The measurements of this paper, and in particular the open-access bias against the most relevant records, determine how complete such a database can be.

6. Conclusion

We built and evaluated an open, resumable pipeline that constructs a domain-specific literature corpus from four scholarly sources. Anchored template families give broad but controlled query coverage, and the four sources return largely

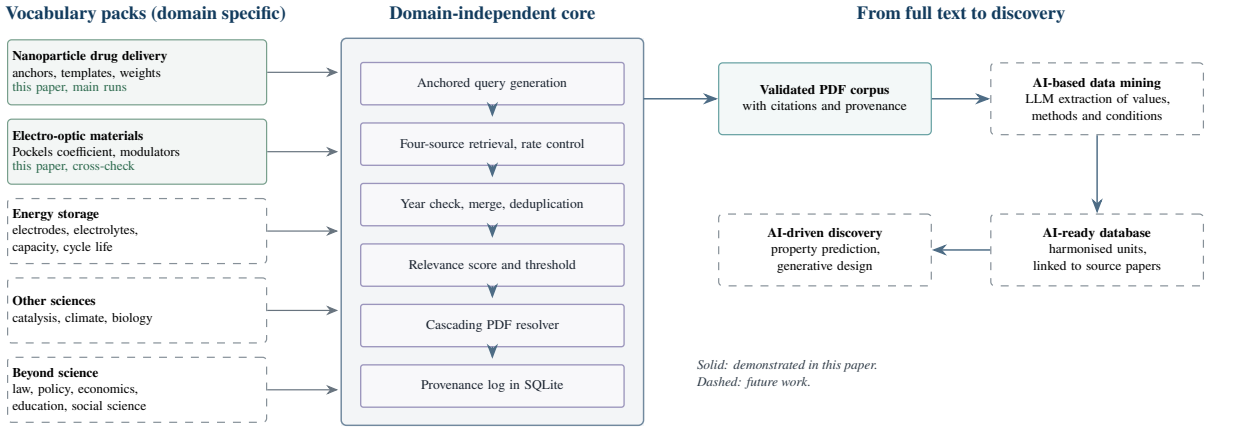


Figure 13. A domain-agnostic pipeline. Left: each domain is described by a vocabulary pack of anchor terms, templates, weights and themes; the nanoparticle drug-delivery and electro-optic packs are used in this paper, and the others illustrate possible applications. Centre: the domain-independent core, reused unchanged. Right: the acquired PDF corpus is the input to AI-based data mining, an AI-ready database and AI-driven discovery, which are future work.

disjoint sets of relevant records, so that no single source supplies more than 41 % of the corpus. A transparent weighted lexical score selects records of which 93 % are relevant at the top of the ranking, and a five-step resolver cascade turns 21 % of the retained records into validated full text. Factorising that yield shows that, outside arXiv, the supply of open-access locations, not the success of retrieving them, decides how much full text a source contributes; the same factorisation explains the higher yield of the electro-optic cross-check, and shows that the most relevant records are the least likely to be acquired. Because the domain enters only through the vocabulary pack, the pipeline is domain agnostic and can be applied to other fields, within science and beyond it. Corpus construction can thus be reported and compared as a stage with measurable properties that bound every downstream step of an AI-driven knowledge-generation pipeline, from data mining to AI-ready databases and materials discovery.

> **Reproducibility:** Each run is stored under `RUNDIR/<run_name>/` with its SQLite database, record table, generated query catalogue, failed-query log, BibTeX file, PDFs and analytics. Every number in Sec. 4 is computed from the three run databases, and every figure is drawn by `pgfplots` from the tables in the `Data/` folder of the manuscript source. Rerunning a configuration requires the vocabulary pack, the run profile of Table 6, and API credentials for OpenAlex and, optionally, NCBI; results will differ slightly from those reported because the indices change over time.

Author contributions

Z.S.J.: methodology, software, investigation, data curation, formal analysis, visualisation, writing (original draft). D.D.: conceptualisation, methodology, supervision, writing (review and editing), funding acquisition.

Funding

This work was supported by the NJIT Gary Thomas Fellowship.

Data availability

The three run databases, exported record tables, generated query catalogues, failed-query logs and the manual relevance labels will be deposited with an archived DOI at submission. The tables from which every figure in this paper is drawn are included with the manuscript source.

Code availability

The pipeline is available at https://github.com/ZiniaJoti/corpus-construction-pipeline. An archived release with a DOI will be cited at submission.

Competing interests

The authors declare no competing interests.

Acknowledgments

The authors acknowledge the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) for computational resources (allocations DMR180013 and MAT250009).

References

[1] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi:10.1063/1.4812323.

[2] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown. The inorganic crystal structure data base. *Journal of Chemical Information and Computer Sciences*, 23(2):66–69, 1983. doi:10.1021/ci00038a003.

[3] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 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):224, October 2021. ISSN 2662-9992. doi:10.1057/s41599-021-00903-w. URL https://doi.org/10.1057/s41599-021-00903-w.

[4] Alexandre Albanese, Peter S. Tang, and Warren C. W. Chan. The effect of nanoparticle size, shape, and surface chemistry on biological systems. *Annual Review of Biomedical Engineering*, 14:1–16, 2012. doi:10.1146/annurev-bioeng-071811-150124.

[5] Elvin Blanco, Haifa Shen, and Mauro Ferrari. Principles of nanoparticle design for overcoming biological barriers to drug delivery. *Nature Biotechnology*, 33(9):941–951, 2015. doi:10.1038/nbt.3330.

[6] Stefan Wilhelm, Anthony J. Tavares, Qin Dai, Seiichi Ohta, Julie Audet, Harold F. Dvorak, and Warren C. W. Chan. Analysis of nanoparticle delivery to tumours. *Nature Reviews Materials*, 1(5):16014, 2016. doi:10.1038/natrevmats.2016.14.

[7] Aaron C. Anselmo and Samir Mitragotri. Impact of particle elasticity on particle-based drug delivery systems. *Advanced Drug Delivery Reviews*, 108:51–67, 2017. doi:10.1016/j.addr.2016.01.007.

[8] Yue Hui, Xin Yi, Fei Hou, David Wibowo, Fan Zhang, Dongyuan Zhao, Huajian Gao, and Chun-Xia Zhao. Role of nanoparticle mechanical properties in cancer drug delivery. *ACS Nano*, 13(7):7410–7424, 2019. doi:10.1021/acsnano.9b03924.

[9] Xiang Yan, Alexander Sedykh, Wenyi Wang, et al. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nature Communications*, 11(1):2519, May 2020. doi:10.1038/s41467-020-16413-3. URL https://doi.org/10.1038/s41467-020-16413-3.

[10] Dibakar Datta. An autonomous scientific knowledge generation framework for AI-driven scientific discovery. *arXiv preprint arXiv:2607.09806*, July 2026. doi:10.48550/arXiv.2607.09806. URL https://arxiv.org/abs/2607.09806.

[11] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Yiwen Hu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models. *Frontiers of Computer Science*, 20(12): 2012627, May 2026. ISSN 2095-2236. doi:10.1007/s11704-026-60308-3. URL https://doi.org/10.1007/s11704-026-60308-3.

[12] Constantin Brîncoveanu, K. Valerie Carl, Aaron Witzki, and Oliver Hinz. Augmenting Systematic Literature Reviews: A Human-AI Collaborative Framework. In Tanya Braun, Benjamin Paaßen, and Frieder Stolzenburg, editors, *KI 2025: Advances in Artificial Intelligence*, volume 15956 of *Lecture Notes in Computer Science*, pages 3–17, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-02813-6. doi:10.1007/978-3-032-02813-6_1.

[13] Xuemei Tang, Xufeng Duan, and Zhenguang G. Cai. Large Language Models for Automated Literature Review: An Evaluation of Reference Generation, Abstract Writing, and Review Composition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1617, Suzhou, China, 2025. Association for Computational Linguistics. doi:10.18653/v1/2025.emnlp-main.83.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[15] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research, December 2023. URL http://arxiv.org/abs/2312.07559. arXiv:2312.07559 [cs.CL].

[16] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao

Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs, November 2024. URL http://arxiv.org/abs/2411.14199. arXiv:2411.14199 [cs.CL].

[17] Yiming Li, Jeff Zhao, Manqi Li, Yifang Dang, Evan Yu, Jianfu Li, Zenan Sun, Usama Hussein, Jianguo Wen, Ahmed M Abdelhameed, Junhua Mai, Shenduo Li, Yue Yu, Xinyue Hu, Daowei Yang, Jingna Feng, Zehan Li, Jianping He, Wei Tao, Tiehang Duan, Yanyan Lou, Fang Li, and Cui Tao. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. *Journal of the American Medical Informatics Association*, 31(9):2030–2039, September 2024. ISSN 1067-5027, 1527-974X. doi:10.1093/jamia/ocae129. URL https://academic.oup.com/jamia/article/31/9/2030/7690757.

[18] Aditya Nagori, Ricardo Accorsi Casonatto, Ayush Gautam, Abhinav Manikantha Sai Cheruvu, and Rishikesan Kamaleswaran. Open-Source Agentic Hybrid RAG Framework for Scientific Literature Review, July 2025. URL http://arxiv.org/abs/2508.05660. arXiv:2508.05660 [cs.IR].

[19] Abhiyan Dhakal, Kausik Paudel, and Sanjog Sigdel. An Artificial Intelligence Driven Semantic Similarity-Based Pipeline for Rapid Literature Review, September 2025. URL http://arxiv.org/abs/2509.15292. arXiv:2509.15292 [cs.AI].

[20] Mahmoud Amiri, Jamile Mohammad Jafari, Sara Mostafapour, and Thomas Bocklitz. Lit2Vec: A Reproducible Workflow for Building a Legally Screened Chemistry Corpus from S2ORC for Downstream Retrieval and Text Mining, April 2026. URL http://arxiv.org/abs/2604.12498. arXiv:2604.12498 [cs.DB].

[21] Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. LitLLMs, LLMs for Literature Review: Are we there yet? *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2412.15249. arXiv:2412.15249.

[22] Michael Gusenbauer and Neal R. Haddaway. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11 (2):181–217, 2020. doi:10.1002/jrsm.1378.

[23] Alejandro Godinez. HySemRAG: A Hybrid Semantic Retrieval-Augmented Generation Framework for Automated Literature Synthesis and Methodological Gap Analysis, August 2025. URL http://arxiv.org/abs/2508.05666. arXiv:2508.05666 [cs.IR].

[24] Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427, 2020. doi:10.1162/qss_a_00022.

[25] Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. URL https://arxiv.org/abs/2205.01833.

[26] Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. The state of OA: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, 6:e4375, 2018. doi:10.7717/peerj.4375.

[27] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi:10.1080/01621459.1927.10502953.

[28] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

[29] Cheng Wang, Mian Zhang, Xi Chen, Maxime Bertrand, Amirhassan Shams-Ansari, Sethumadhavan Chandrasekhar, Peter Winzer, and Marko Lončar. Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. *Nature*, 562(7725):101–104, 2018. doi:10.1038/s41586-018-0551-y.

[30] Di Zhu, Linbo Shao, Mengjie Yu, Rebecca Cheng, Boris Desiatov, C. J. Xin, Yaowen Hu, Jeffrey Holzgrafe, Soumya Ghosh, Amirhassan Shams-Ansari, Eric Puma, Neil Sinclair, Christian Reimer, Mian Zhang, and Marko Lončar. Integrated photonics on thin-film lithium niobate. *Advances in Optics and Photonics*, 13(2):242–352, 2021. doi:10.1364/AOP.411024.

[31] S. Jenkins, S. R. Kirk, M. Persson, J. Carlen, and Z. Abbas. Molecular dynamics simulation of nanocolloidal amorphous silica particles: Part I. *The Journal of Chemical Physics*, 127(22):224711, December 2007. ISSN 0021-9606. doi:10.1063/1.2803897. URL https://doi.org/10.1063/1.2803897.

[32] S. Jenkins, S. R. Kirk, M. Persson, J. Carlen, and Z. Abbas. Molecular dynamics simulation of nanocolloidal amorphous silica particles: Part II, 2007. URL https://arxiv.org/abs/0708.2531.

Appendix A. Worked scoring example

The record below is the first test case shipped with M4.

Title. Nanoparticle elasticity directs tumor uptake.
Abstract. Soft polymer–lipid nanoparticles with low Young's modulus show enhanced tumor penetration, cellular uptake, and improved cancer drug delivery.

Distinct patterns matched: two nanoparticle patterns (*nanoparticle*, *nanoparticles*), two mechanical patterns (*elasticity*, *Young's modulus*), and four membrane, uptake and delivery patterns (*tumor penetration*, *cellular uptake*, *drug delivery*, *cancer drug delivery*); no other group matches. Then

$$\sum_g w_g \min(n_g, 4) = 4.0 \times 2 + 4.5 \times 2 + 4.0 \times 4 = 33,$$
$$B = 5 + 5 + 4 + 6 = 20, \qquad P = 0,$$
$$R_{\text{raw}} = 53, \qquad R = \min(1,\ 53/40) = 1.$$

A title and a one-sentence abstract therefore saturate the score. By contrast, a record titled *Green synthesis of silver nanoparticles for wastewater treatment* matches one nanoparticle pattern only, scores $R = 4/40 = 0.10 < \theta$, and is excluded.