# FROM FRAGMENTS TO FUNCTIONS: EVIDENCE-LINKED TRAJECTORIES FOR TRACING SCIENTIFIC RECOMBINATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Scientific designs often combine functional relations first reported in separate papers, yet literature extraction rarely distinguishes a new local mechanism from a new arrangement of established ones. We ask how such recombination appears in the scientific record and whether historically supported but unreported paths precede later complete reports. A factorized language-model pipeline constructs source-linked Material–Sensor–Signal–System trajectories from full-text conductive-fibre flexible-sensor papers. On 200 paper-disjoint gold papers, it reaches 91.7% completeness and 95.6% effective record quality, versus 78.0% and 77.2% for single-stage tuning. Of 1,302 complete trajectories first observed in 2022–2025, 78.4% combine three relations already observed before 2022. In four rolling-origin tests, a source-linked relation-support score places 29–36 future complete reports among its top 100 candidates, versus 19–29 for term popularity and 2.5–3.0 expected under random ranking. These temporal analyses make functional recombination observable and test whether local precedent anticipates complete-path reports.

## 1 INTRODUCTION

Evidence for an engineering design rarely arrives as a single sentence. A paper may establish a material architecture, another the transduction principle it enables, and a third a downstream application. The literature contains many fragments but no direct inventory of the end-to-end designs that have or have not been demonstrated. This is visible in flexible sensors, where a material implementation must create a sensing element, produce a measurable signal, and serve a system task (Zeng et al., 2014; Trung & Lee, 2016; Heikenfeld et al., 2018). A list of materials and applications, or a collection of untyped pairwise links, does not specify this functional order.

We ask whether scientific papers can be organized into *evidence-linked functional trajectories* that retain both their mechanism and evidential boundary. The Material–Sensor–Signal–System ($MS^3$) schema assigns a role to each component, connects only source-supported adjacent relations, and preserves the paper locations that justify them. An incomplete record remains incomplete rather than being filled by a language model. Figure 1 shows a retrospective example: three adjacent relations were documented by the 2021 cutoff, but the complete path was not reported until 2025. This representation provides a common unit for three questions: Can the record be reconstructed faithfully? How do complete designs enter the literature? Can locally supported, not-yet-reported combinations be prioritized before later reports appear?



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

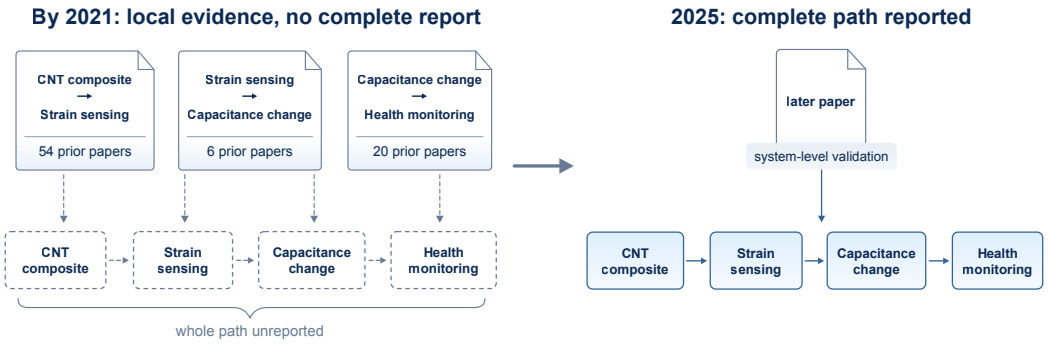


Figure 1: **A retrospectively identified supported gap.** By the end of 2021, the three adjacent relations in this example had support from 54, 6, and 20 historical papers, respectively, but no paper reported the complete four-role path. A complete path with system-level validation was reported in 2025. The counts describe local relation support, not evidence that the full design existed at the cutoff.

We instantiate the study in conductive-fibre flexible-sensor research. Three supervised adapters over one language-model backbone recover citable evidence, assign functional roles to grounded terms, and assemble only supported transitions. Passing source identifiers between stages makes each emitted path auditable and leaves absent links unfilled. A held-out, paper-disjoint evaluation measures both source fidelity and record recovery. We then freeze historical relation sets at each year, distinguish first-observed local relations from new arrangements of older ones, and evaluate candidate paths at four historical cutoffs against the following four years. A candidate is not merely an unusual co-occurrence: it is an ordered path whose three local relations already had historical support while the entire path had not been reported.

We introduce a time-frozen, source-auditable study of functional recombination that separates novelty in local relations from novelty in complete designs. Across 1,302 first-observed complete paths, we measure how often established relations are assembled into new paths and examine their subsequent re-observation. We further evaluate whether historically supported but unreported paths precede future complete reports in four rolling-origin backtests. Source-linked MS$^3$ records, constructed with a factorized extraction pipeline and evaluated on 200 paper-disjoint gold papers, provide the empirical basis for these analyses.

## 2 Related Work

Scientific information extraction has progressed from identifying entities and within-document relations to recovering document-level structures. SciERC jointly models entities, relations, and coreference in scientific text (Luan et al., 2018); SciREX extends the evaluation unit to salient entities and document-level multi-argument relations (Jain et al., 2020). Large language models have also been used to extract structured materials information at scale (Dagdelen et al., 2024; Caufield et al., 2024). Our setting adds a functional constraint to document extraction: each record follows material, sensing element, signal, and system roles, and each emitted relation is linked to a source location. Partial trajectories are retained because a paper can establish a mechanism without reporting every downstream role. This makes recovery of supported content, as well as correctness of emitted complete records, central to evaluation.

Scientific question answering, literature synthesis, and claim verification make literature-based answers inspectable. Qasper evaluates paper-grounded answers (Dasigi et al., 2021); retrieval-augmented systems including RAG, PaperQA, OpenScholar, and SciRAG synthesize retrieved evidence (Lewis et al., 2020; Lála et al., 2024; Asai et al., 2026; Ding et al., 2026); SciFact, SciFact-Open, and SciVer evaluate textual or multimodal claim support (Wadden et al., 2020; 2022; Wang et al., 2025). These tasks typically begin with a question or proposition and return supporting passages or a verdict. Here the records themselves are the index: evidence is attached to functional terms and transitions before a candidate design is formed. When adjacent transitions come from

different papers, the candidate retains separate evidence sets and an explicit missing complete-path report. This supports inspection of both local precedent and the unresolved integration step.

Knowledge-graph and discovery studies motivate the cross-paper view. MatKG organizes extracted materials concepts into a knowledge graph (Venugopal & Olivetti, 2024), while research on scientific novelty analyzes combinations of existing knowledge (Uzzi et al., 2013; Foster et al., 2015). Recent materials-science forecasting uses concept graphs to predict emerging concept combinations (Marwitz et al., 2026). We examine a more specific unit of recombination: an ordered four-role trajectory whose three adjacent relations may already be known even when the complete design is new. End-to-end sensor design provides a natural test case because material choice, transduction, signal behavior, and application must ultimately fit together (Ates et al., 2022). The time-frozen relation and path inventories let us measure where novelty enters and evaluate whether supported gaps precede later complete reports.

## 3 Evidence-Linked Functional Trajectories

### 3.1 Representation and source boundary

Let a trajectory be $p = (V_p, R_p, A_p)$, where $V_p$ contains typed Material, Sensor, Signal, and System terms; $R_p$ contains only admissible adjacent relations; and $A_p$ maps supported terms and relations to evidence identifiers, source text, paper identifiers, and document locations. A complete trajectory follows

$$\text{Material} \longrightarrow \text{Sensor} \longrightarrow \text{Signal} \longrightarrow \text{System}. \tag{1}$$

A partial trajectory records the observed subset. Evidence items have a primary role—experiment, property, validation, or reference—and reference-only material does not establish a reported relation. Figure, caption, table, and quantitative provenance are retained when present. The distinction between a missing relation and a supported relation is essential both for extraction evaluation and for the historical candidate task.

The representation keeps two levels of support separate. A paper-level trajectory describes what one study reports; a corpus-level candidate may join adjacent relations supported by different studies. A complete corpus candidate is therefore not automatically a demonstrated device. Its three edge-specific evidence sets document why each local transition is plausible, while the absent four-role report identifies the integration gap. This prevents a cross-paper assembly from inheriting a stronger experimental status than any source warrants. Figure 2b gives a within-paper example: the E1 experiment and E4 validation support different portions of a trajectory, while E3 is explicitly marked as contextual reference material.

Let $\mathcal{A}_t^{(j)}(u, v)$ be the set of non-reference source records dated no later than $t$ that support a transition from role $j$ to role $j + 1$. For a four-term candidate $p = (m, s, g, z)$, its historical support is the ordered bundle

$$\mathcal{B}_t(p) = \big(\mathcal{A}_t^{(1)}(m, s), \mathcal{A}_t^{(2)}(s, g), \mathcal{A}_t^{(3)}(g, z)\big). \tag{2}$$

Each component remains linked to its source paper and location. A candidate can therefore have three nonempty support sets without any paper reporting the complete path. Keeping the sets separate also preserves the difference between several independent reports of one transition and a single report that happens to mention multiple terms. The bundle supplies an inspection route from a proposed design to its local evidence; it is not itself evidence of successful integration.

### 3.2 Corpus construction and factorized extraction

The corpus contains 13,689 original-research papers published from 2010 through a partially covered 2026. Machine-readable PDFs are parsed into ordered source blocks with page and section identifiers; nearby figures, captions, tables, numerical values, and units remain linked to the block. Papers are deduplicated by DOI or normalized title and year. The resulting dataset contains 131,083 evidence items and 26,648 complete or partial trajectory records, of which 15,703 are complete.

Construction is factorized into three supervised adapters over a shared Qwen2.5-72B-Instruct backbone. E-LoRA extracts compact evidence items and verifies their alignment to parsed source blocks; T-LoRA maps accepted items to paper-specific terms and functional roles; M-LoRA assembles

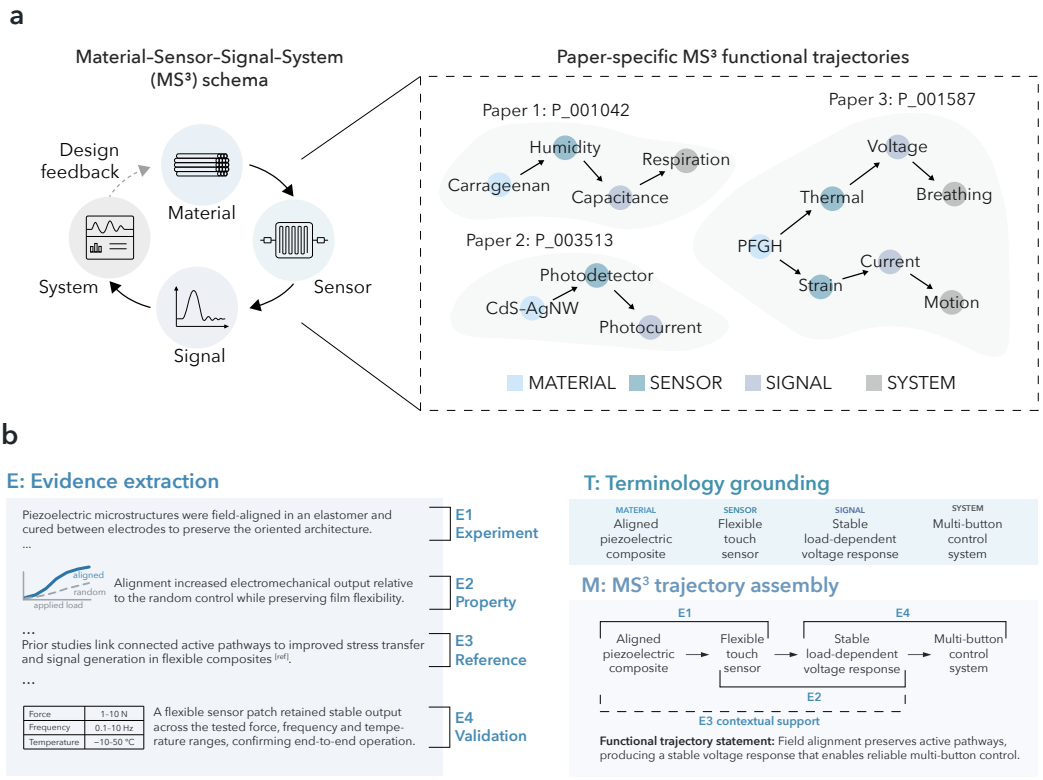


Figure 2: **From full text to an evidence-linked functional trajectory. a**, The four ordered functional roles and heterogeneous paper-specific paths. **b**, Source-aligned evidence (E) is grounded to paper terminology (T) and assembled into supported trajectory relations (M). The reference-labelled item supplies context but does not independently validate an edge.

grounded terms into supported relations and complete or partial trajectories. The intermediate evidence identifiers and role assignments remain visible to downstream stages and to post-generation checks. These checks reject unsupported identifiers, invalid role transitions, unaligned source text, and duplicate records. In this way, the model is used to recover and organize reported content, while source alignment and schema constraints determine which links can enter the historical analysis.

Expert annotation of 100 papers supplied initial examples and validation rules. Candidate records from more than 2,000 additional papers were reviewed and filtered for aligned evidence and schema validity before use as silver supervision. Evaluation used a separately selected set of 200 papers, disjoint from development, silver training, adapter validation, retrieved demonstrations, and earlier analyses. Two experts independently annotated each paper and a third resolved disagreements. We compare direct generation, single-stage supervised fine-tuning, and the E/T/M pipeline under source-based review. Expert ratings of completeness and effective record quality use 0–4 rubrics normalized to 0–100; confidence intervals use paper-level bootstrap resampling.

The evaluation deliberately separates *precision conditional on generation* from *recovery of supported content*. Evidence-source accuracy is the fraction of emitted evidence items aligned to the source; term accuracy additionally requires the correct functional layer; semantic full-path accuracy is precision among emitted complete paths that match a supported path. Completeness instead asks how much important source-grounded content a method recovers, including partial records. Effective record quality prioritizes evidence support, then path correctness and completeness. This distinction is necessary because a system could obtain high full-path precision by emitting very few paths. Expert scores are averaged at the paper level, and intervals are estimated from 5,000 paper-level bootstrap samples rather than treating multiple records from one paper as independent observations.

The measured components matter differently downstream. An omitted supported transition can remove a future candidate from the historical pool, whereas an unsupported transition can create a candidate with a misleading local precedent. Assigning a correct term to the wrong role changes which adjacent relation is counted; losing its document location leaves the term tuple intact but makes the resulting record harder to audit. These failure modes motivate evaluating provenance alignment, functional assignment, and record coverage separately. They also motivate the paper-disjoint split: multiple sections and figures from one study should not appear on both sides of the extraction evaluation.

# 4 Temporal Tasks and Evaluation

## 4.1 Relation novelty versus path novelty

For each year $y$, let $E^{(1)}_{<y}, E^{(2)}_{<y}, E^{(3)}_{<y}$ denote previously observed adjacent relations at the three role boundaries, and $P_{<y}$ previously observed complete paths. A complete path $p = (m, s, g, z)$ first observed in year $y$ is a *complete-path recombination* when

$$p \notin P_{<y}, \quad (m, s) \in E^{(1)}_{<y}, \quad (s, g) \in E^{(2)}_{<y}, \quad (g, z) \in E^{(3)}_{<y}. \tag{3}$$

Otherwise, the first-observed relation identifies the layer of local novelty; paths with multiple new relations receive a separate label. The historical base is frozen for the whole observation year, so two same-year reports do not turn each other into prior art. Distinct paths within a paper are retained, while duplicate paper–path records are collapsed. Re-observation is the identical normalized path in a different paper within four complete subsequent years.

## 4.2 Rolling-origin supported-gap backtest

At cutoff $t$, candidates join previously observed adjacent relations, excluding paths already completely reported:

$$\mathcal{C}_t = \{(m, s, g, z) : (m, s) \in E^{(1)}_t, (s, g) \in E^{(2)}_t, (g, z) \in E^{(3)}_t\} \setminus P_t. \tag{4}$$

Candidates are deduplicated by their normalized, ordered four-term tuple before ranking. A relation supported by many historical papers therefore strengthens one candidate rather than generating multiple copies of it. The three source sets in Equation 2 remain attached to that candidate, while membership in $P_t$ is determined by complete paper-level reports. This separates the existence of local precedent, the amount and independence of that precedent, and the observation of a whole design. The exact-closure outcome requires the same normalized complete path to appear in original research during $(t, t + 4]$, with non-reference evidence for every adjacent relation. A stricter outcome also requires system-level validation in the stated application. Both are observations in the assembled literature, not labels of physical feasibility.

We evaluate cutoffs at the ends of 2018, 2019, 2020, and 2021, each followed for four years. Every feature and evidence item available to a ranker predates its cutoff. We compare random expectation, term popularity, three-edge counts, source-linked relation support, and relation support with additional validation cues. Every candidate is constructed from adjacent relations grounded in source-linked historical evidence; the comparison varies ranking inputs, not candidate generation. Term popularity ranks candidates by the within-cutoff-pool percentile of the sum of $\log(1 + n)$ for their four role terms, where $n$ counts cutoff-or-earlier papers containing a term in its role. It uses no relation-pair or mechanism-evidence features in its score. Three-edge counts rank by $\sum_{j=1}^{3} \log(1 + n_j)$, where $n_j$ counts cutoff-or-earlier distinct papers supporting typed adjacent relation $j$. The relation-support score aggregates evidence-derived adjacent-relation strength, minimum relation support, cross-layer compatibility, recency, and independent-source support. The additional cues are validation momentum and closure-specific weak-link support. These ranking scores use historical evidence statistics rather than directly encoding the full source statements. The validation-cue reranking was refined after examining historical windows and is treated as retrospective. We report exact and system-validated closure at a fixed top-100 review budget, together with average precision over the full candidate pool.

Table 1: **Source-grounded extraction on 200 paper-disjoint gold papers.** Expert-rated completeness and effective record quality are normalized to 0–100; entries are paper-level means. Bold marks the best value in each column, and underlining marks the second-best. The proposed pipeline is shaded light blue.

| Construction method | Completeness (%) | Effective quality (%) |
|---|---|---|
| Direct generation | 69.2 | 70.8 |
| Single-stage fine-tuning | <u>78.0</u> | <u>77.2</u> |
| Factorized E/T/M | **91.7** | **95.6** |

The relation-support score uses percentile-normalized features $f_j \in [0,1]$ with weights $(0.30, 0.15, 0.40, 0.10, 0.05)$ in the order above. This explicitly gives substantial weight to cross-layer consistency and the least-supported adjacent relation. The optional validation-cue score combines temporal validation momentum (weight $0.70$) with closure-specific weak-link support (weight $0.30$). For relation-support score $S$, validation-cue score $E$, and initial rank $r$, the retrospective analysis applies the bounded head-focused adjustment

$$R = S + 0.20(E - 0.5)(0.25 + 0.75S)\left[1 + \exp\left(\frac{r - 120}{20}\right)\right]^{-1}. \tag{5}$$

The same fixed candidate universe and closure definitions are used for each method at a given cutoff. Top-$K$ counts quantify the yield under a limited manual-review budget, whereas average precision evaluates ordering across the full pool. System-validated closure is reported separately so that a paper-level complete path is not confused with an application-level experiment.

## 5 Experiments and Analysis

### 5.1 Full-text evidence can be assembled into reliable trajectories

The collection contains 58.93% complete records, 30.33% contiguous partial records, and 10.74% evidence-gapped records. Across evidence items, 26.1% have text-only grounding, 11.5% are linked to a figure or caption, 18.7% to a table or quantitative result, and 43.7% to text and both other grounding types. Keeping partial records prevents missing support from being conflated with a model-completed mechanism.

On 200 held-out papers, the factorized pipeline reaches 91.7% expert-rated completeness and 95.6% effective record quality, versus 78.0% and 77.2% for single-stage fine-tuning and 69.2% and 70.8% for direct generation (Table 1). Evidence-source accuracy is 99.6%, and semantic full-path accuracy is 83.3%. Direct generation has higher conditional full-path accuracy among the complete paths it produces (93.1%) but recovers less supported content. The joint completeness–quality comparison favors decomposed construction for corpus analysis.

### 5.2 Most first-observed complete paths recombine older relations

Complete-path recombination is a large and persistent component of trajectory growth (Figure 3a). Against relations observed before 2022, 1,021 of 1,302 unique complete paths first observed during 2022–2025 (78.4%) contain three pre-observed adjacent relations. Of the rest, 140 introduce only a Material–Sensor relation, 78 only a Sensor–Signal relation, 28 only a Signal–System relation, and 35 introduce relations at multiple layers (Figure 3b). The dominant novelty in this corpus thus lies in the complete arrangement, not a newly observed local transduction relation.

The annual view and the fixed-2021 cohort answer different questions. The annual view classifies each paper–trajectory observation against literature available before that observation year. Complete-path recombination accounts for 45.6% of observations in 2020 and 42.6% in 2025, while the share of already existing complete trajectories rises from 33.1% to 53.4%. The fixed-2021 analysis instead collapses the 2022–2025 observations to unique paths and asks whether each of their three relations was already present by the end of 2021. It therefore measures the extent of delayed *integration* without allowing later observations within the four-year interval to redefine earlier relation novelty.

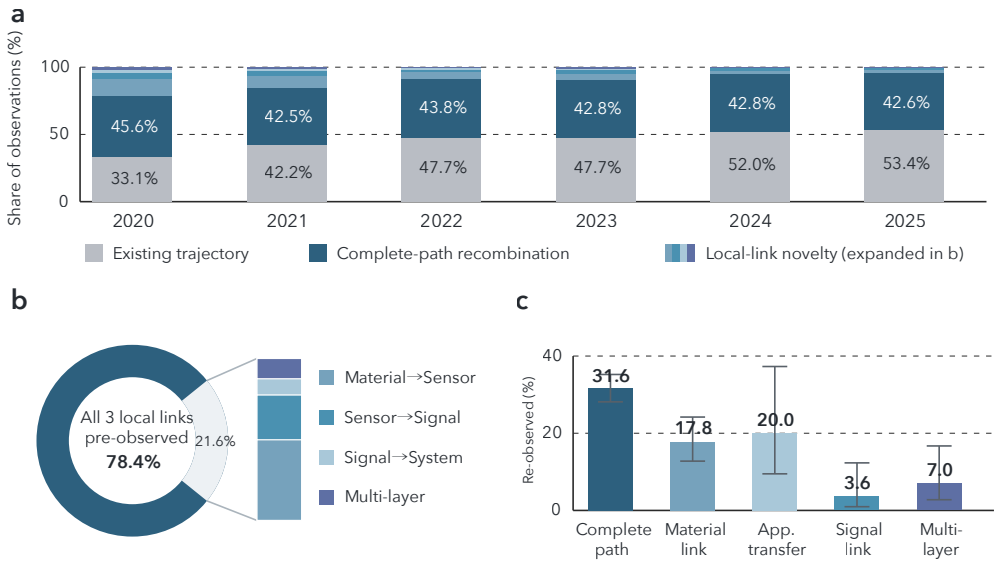


Figure 3: **Where functional novelty enters the literature. a**, Annual paper–trajectory observations by novelty mode. **b**, Of 1,302 complete paths first observed in 2022–2025, 78.4% recombine three relations observed by the end of 2021. **c**, Four-year re-observation in a different paper by first-observation mode, with 95% confidence intervals.

We also measure whether a first-observed path appears again in a different paper during four complete follow-up years. The re-observation rate is 31.6% for complete-path recombinations (656 eligible trajectories), compared with 17.8% for material-link novelty, 20.0% for application transfer, 3.6% for signal-link novelty, and 7.0% for multi-layer novelty (Figure 3c). Re-observation captures independent uptake of a reported functional configuration in the subsequent literature.

These analyses distinguish two ways a design space can grow: a new adjacent relation expands its locally documented mechanisms, whereas a new complete path connects documented relations across functional roles. For the latter, edge-specific evidence can be inspected for substrate, operating-range, and application conditions before a proposed integration is tested. The present scores summarize historical support; future compatibility models could represent those conditions directly.

### 5.3 Supported gaps are enriched for future complete reports

At the 2021 cutoff, the candidate universe contains 20,305 unreported complete paths with three previously observed local relations. A worked example joins CNT composite, strain sensing, capacitance change, and health monitoring. The three edges are supported separately by 54, 6, and 20 historical papers; the complete path is reported with system-level validation in 2025. Its rank changes from 151 under relation support to 148 after validation-cue reranking. This illustrates the distinction between local support and complete-path observation.

Across four rolling cutoffs, term popularity places 26, 29, 21, and 19 exact future closures in its top 100; three-edge counts place 30, 30, 26, and 25; and relation support places 29, 36, 34, and 32. Random expectations are 3.04, 2.93, 2.68, and 2.46; validation-cue reranking yields 29, 37, 35, and 34 (Table 3). For the system-validated endpoint, the respective counts are 21/25/18/16 for term popularity, 25/26/24/23 for three-edge counts, 25/30/30/27 for relation support, and 25/30/31/29 after validation-cue reranking. Relation support improves full-pool AP over both simpler controls at every cutoff, while the count-only control retrieves one more exact top-100 closure in 2018. Validation cues produce smaller changes at the head of the ranking.

Table 2 gives the denominator and full-pool average precision (AP) for each origin. Candidate prevalence falls as the combinatorial pool grows. Three-edge-count AP is .125–.150, between term popularity (.106–.134) and relation support (.159–.193) at every cutoff. At the 2021 origin, relation

Table 2: **Historical candidate pools and full-pool AP for exact future closure.** Each row uses a four-year follow-up and an identical evidence-constructed candidate universe across methods. Chance AP equals endpoint prevalence in expectation. Bold marks the best observed AP and underline the second-best. The core relation-support column uses a stronger blue tint; retrospective validation cues use a paler blue.

| Cutoff | Candidates | Closures | Chance (%) | Term pop. AP | 3-edge AP | Relation support AP | + validation AP |
|---|---|---|---|---|---|---|---|
| 2018 | 9,053 | 275 | 3.04 | .130 | .146 | **.174** | <u>.172</u> |
| 2019 | 12,252 | 359 | 2.93 | .134 | .150 | <u>.193</u> | **.195** |
| 2020 | 15,680 | 421 | 2.68 | .112 | .127 | **.163** | <u>.159</u> |
| 2021 | 20,305 | 499 | 2.46 | .106 | .125 | <u>.159</u> | **.162** |

Table 3: **Future complete-path reports among the top 100 candidates.** Random entries are expected counts under uniform ranking; all other entries are observed counts on the same evidence-constructed pool. The validated endpoint also requires system-level validation. Bold marks the best observed count (including ties), underline the second-best distinct count. The core relation-support column uses a stronger blue tint; retrospective validation cues use a paler blue.

| Cutoff | Endpoint | Random | Term popularity | 3-edge count | Relation support | + validation cues |
|---|---|---|---|---|---|---|
| 2018 | Exact | 3.04 | 26 | **30** | <u>29</u> | <u>29</u> |
| 2019 | Exact | 2.93 | 29 | 30 | <u>36</u> | **37** |
| 2020 | Exact | 2.68 | 21 | 26 | <u>34</u> | **35** |
| 2021 | Exact | 2.46 | 19 | 25 | <u>32</u> | **34** |
| 2018 | Validated | 2.4 | <u>21</u> | **25** | **25** | **25** |
| 2019 | Validated | 2.4 | 25 | <u>26</u> | **30** | **30** |
| 2020 | Validated | 2.2 | 18 | 24 | <u>30</u> | **31** |
| 2021 | Validated | 1.9 | 16 | 23 | <u>27</u> | **29** |

support yields 12, 23, 32, and 53 closures at budgets of 20, 50, 100, and 200; validation-cue reranking yields 12, 23, 34, and 53. Together, AP and fixed-budget yield describe both broad ordering and concentration near the head.

The term-popularity control favors familiar components but cannot distinguish whether a particular material–sensor or signal–system transition has been documented. Relative to this control, relation support increases exact-closure AP by .043–.059 and adds 3–13 future reports to the top 100 across the four cutoffs. Adding validation cues changes the top-100 count by 0–2, while its full-pool AP rises at two cutoffs and falls at two. These outcomes separate the value of relation-level historical support from the incremental effect of the two validation aggregates. In particular, a fixed review budget and full-pool ordering measure different uses of the same candidate list.

This comparison separates representation from scoring. Source-linked evidence determines which transitions enter the supported-gap pool and remains inspectable at any rank. Term popularity tests node familiarity, three-edge counts test adjacent-relation frequency, relation support adds further edge and cross-layer summaries, and the optional reranker adds two validation signals. Condition-level matching across source statements—such as a material's measured strain range against a proposed sensor's operating range—is a distinct question from the support statistics evaluated here.

### 5.4 Literature-derived paths can guide bounded design decisions

We also examined whether the evidence-to-candidate workflow could inform a physical design decision. A large-deformation sensor needed a strain-sensitive electrical response while preserving conductive continuity. Separately documented Ag-layer crack formation and strain-tolerant CNT/CB conduction suggested complementary roles: the former could amplify resistance changes, while the latter could sustain a conducting pathway. This motivated fabrication of a CNT/CB–Ag dual-layer candidate alongside a CNT/CB single-layer control.

Three independent devices per architecture were tested. Figure 4 compares representative five-cycle loading–unloading windows at 30%, 60%, and 100% strain on a shared $\Delta R/R_0$ scale. Within these windows the dual-layer candidate produced larger response amplitudes than the single-layer

control. Photographs and matched electron micrographs of both architectures at 0%, 30%, 60%, and 100% strain are shown separately in Supplementary Figure S1. Together, the electrical traces and morphology document one fabricated instance of the proposed mechanism composition under the tested conditions.

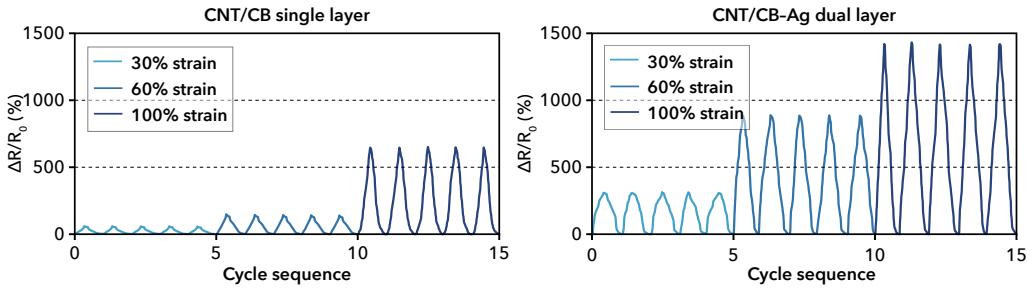


Figure 4: **Representative cyclic response under matched strain.** The CNT/CB single-layer control (left) and CNT/CB–Ag dual-layer candidate (right) are shown at 30%, 60%, and 100% strain on a common $\Delta R/R_0$ scale. Each window contains five loading–unloading cycles. Three independent devices per architecture were tested; the original vector traces shown here are representative, not group means.

## 6 Discussion

Source-linked trajectories make local transitions auditable, distinguish end-to-end reports from related concepts, and expose combinations with local precedent but no complete report. In this corpus, recombination of older relations accounts for most first-observed complete paths in 2022–2025 and has the highest four-year re-observation rate among novelty modes. The factorized E/T/M construction supplies the source alignment and partial records needed to distinguish path-level from local-mechanism novelty.

Ranking supported gaps makes the distinction useful under a limited review budget. Relation support improves both AP and Top-100 yield over term popularity at all four cutoffs, whereas validation cues change the ordering only modestly. Most of the measured ranking signal is therefore already present in historical relation support. A researcher can inspect the source locations behind a highly ranked path, identify its least-supported transition, and compare operating conditions before choosing an integration experiment. The CNT/CB–Ag study illustrates this workflow by joining complementary conductive-network mechanisms and testing the fabricated candidate against a single-layer control.

The endpoint is later publication of the exact normalized path, which is sensitive to corpus coverage, publication delay, and term normalization. The validation-cue adjustment was developed on historical windows; its prospective utility can be tested with frozen scores and shared evaluation resources. Beyond sensors, the same distinction between local support and complete reports applies wherever functional roles and their evidence can be defined.

## 7 Conclusion

Evidence-linked functional trajectories make it possible to compare novelty in local mechanisms with novelty in complete designs using the same source-grounded record. In this sensor corpus, most newly observed complete paths recombine earlier relations, and historical relation support enriches later complete reports. The resulting time-frozen task exposes a form of scientific change missed by isolated-relation analysis: complete functional paths can emerge through new arrangements of established links. It also provides a concrete target for ranking candidate integrations while retaining the evidence needed for human review.

## Reproducibility Statement

The corpus selection, paper-disjoint annotation protocol, adapter construction, year-frozen novelty rules, candidate enumeration, and ranking endpoints are specified above and in the supplementary material. Reproducing the analysis requires the paper identifiers, normalization tables, gold annotations, cutoff-specific candidate exports, ranking code, and hashes of frozen temporal inputs. Redistribution of source excerpts is subject to the licenses of the underlying articles.

## AI Use Statement

Generative AI tools provided feedback on experimental comparisons and assisted with analysis-code implementation and checking, interpretation of results, manuscript drafting and editing, and figure adaptation. The authors reviewed AI-assisted outputs and take responsibility for the submitted text, results, code, and figures.

## References

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, Jenna Sparks, Jena D. Hwang, Varsha Kishore, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Daniel S. Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650:857–863, 2026. doi: 10.1038/s41586-025-10072-4.

H. Ceren Ates, Peter Q. Nguyen, Laura Gonzalez-Macia, Eden Morales-Narváez, Firat Güder, James J. Collins, and Can Dincer. End-to-end design of wearable sensors. *Nature Reviews Materials*, 7(11):887–907, 2022. doi: 10.1038/s41578-022-00460-x.

J. Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L. Harris, Marcin P. Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra A. T. Moxon, Justin T. Reese, Melissa A. Haendel, Peter N. Robinson, and Christopher J. Mungall. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3):btae104, 2024. doi: 10.1093/bioinformatics/btae104.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418, 2024. doi: 10.1038/s41467-024-45563-x.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Hang Ding, Yilun Zhao, Tiansheng Hu, Zihang Wang, Manasi Patwardhan, and Arman Cohan. SciRAG: Adaptive, citation-aware, and outline-guided retrieval and synthesis for scientific literature. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6440–6460, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.303. URL https://aclanthology.org/2026.eacl-long.303/.

Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015. doi: 10.1177/0003122415601618.

Jason Heikenfeld, Andrew Jajack, John Rogers, Philipp Gutruf, Limei Tian, Tingrui Pan, Ronald Li, Michelle Khine, Jitae Kim, Joseph Wang, and Jayoung Kim. Wearable sensors: Modalities, challenges, and prospects. *Lab on a Chip*, 18(2):217–248, 2018. doi: 10.1039/C7LC00914C.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7506–7516, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.670. URL https://aclanthology.org/2020.acl-main.670/.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. PaperQA: Retrieval-augmented generative agent for scientific research. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=clU5xWyItb.

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 *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3219–3232, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1360. URL https://aclanthology.org/D18-1360/.

Thomas Marwitz, Alexander Colsmann, Ben Breitung, Christoph Brabec, Christoph Kirchlechner, Eva Blasco, Gabriel Cadilha Marques, Horst Hahn, Michael Hirtz, Pavel A. Levkin, Yolita M. Eggeler, Tobias Schlöder, and Pascal Friederich. Predicting new research directions in materials science using large language models and concept graphs. *Nature Machine Intelligence*, 8:535–544, 2026. doi: 10.1038/s42256-026-01206-y.

Tran Quang Trung and Nae-Eung Lee. Flexible and stretchable physical sensor integrated platforms for wearable human-activity monitoring and personal healthcare. *Advanced Materials*, 28(22): 4338–4372, 2016. doi: 10.1002/adma.201504244.

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

Vineeth Venugopal and Elsa Olivetti. MatKG: An autonomously generated knowledge graph in material science. *Scientific Data*, 11:217, 2024. doi: 10.1038/s41597-024-03039-z.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347/.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. SciVer: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8562–8579, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.420. URL https://aclanthology.org/2025.acl-long.420/.

Wei Zeng, Lin Shu, Qiao Li, Song Chen, Fei Wang, and Xiao-Ming Tao. Fiber-based wearable electronics: A review of materials, fabrication, devices, and applications. *Advanced Materials*, 26(31):5310–5336, 2014. doi: 10.1002/adma.201400633.

# Appendix: Supplementary Material

From Fragments to Functions: Evidence-Linked Trajectories for Tracing Scientific Recombination

## S1 Candidate Records and Temporal Endpoints

Each candidate is an ordered material–sensor–signal–application tuple. Three adjacent relations connect its four roles. For each relation, the cutoff-time record keeps source-linked support from distinct papers, rather than treating term co-occurrence as a claim that the complete tuple was built. The paper's relation-support features summarize this historical record; the underlying source passages and the complete corpus are not included in this review package. Table S1 gives the disclosed record-level fields needed to interpret the evaluation. It is a conceptual schema, not a dump of the extraction database.

Table S1: **Fields used to interpret a four-role candidate.** Paper references identify the provenance of local support; the future label is stored separately from all cutoff-time features.

| Field | Meaning |
|---|---|
| Ordered terms | Material, sensor function, signal or readout, and target application. |
| Local support | The three adjacent role-specific relations and their cutoff-or-earlier source support. |
| Source provenance | Paper identity and publication year for each supporting relation. |
| Observation label | Whether a different original-research paper reports the complete ordered path in follow-up. |
| Validation label | Whether that report additionally validates the path at system level in the stated application. |

For annual novelty, papers through year $y-1$ form the fixed historical base for all records in year $y$. In the four-year ranking task, the candidate universe and all ranking features are fixed at the cutoff. A subsequent report must appear in a different paper; same-year co-introductions do not count as prior local support. Exact future closure requires non-reference evidence for all three relations in follow-up original research. The stricter endpoint additionally requires system-level validation in the stated application. A candidate without an observed report within the window is an *unobserved closure*, not a demonstrated infeasible design.

## S2 Ranking and Audit Details

The historical relation-support score uses percentile-normalized adjacent-relation strength (weight $0.30$), weakest-relation support ($0.15$), cross-layer compatibility ($0.40$), recency ($0.10$), and independent-source support ($0.05$). The optional validation-cue score uses temporal validation momentum ($0.70$) and closure-specific weak-link support ($0.30$). These are the fixed formulas used for the retrospective comparison in the paper. Candidate identifiers resolve original-ranking ties; the term-popularity reevaluation resolves ties by lexical order of the normalized trajectory. No follow-up publication contributes to a cutoff-time feature.

Average precision is computed over the full frozen candidate universe of each cutoff, whereas Top-100 reports count observed future closures among its highest-ranked 100 candidates. Random Top-100 entries in the main paper are expected counts from endpoint prevalence, not samples from repeated random rankings. These metrics therefore measure different regions of the same ordering; the Top-100 count should not be interpreted as a full-pool precision estimate.

## S3 Conductive-Network Case: Imaging and Scope

Figure S1 complements the electrical-response traces in Figure 4 of the main paper. Matched SEM views were acquired at $0\%$, $30\%$, $60\%$, and $100\%$ strain for the CNT/CB single-layer control and the CNT/CB–Ag dual-layer candidate. The source study reports representative five-cycle $\Delta R/R_0$ windows at $30\%$, $60\%$, and $100\%$ strain, with three independent devices tested per architecture. The dual-layer architecture showed larger response amplitudes in those windows. This case characterizes one implemented composition rather than a general optimization benchmark.

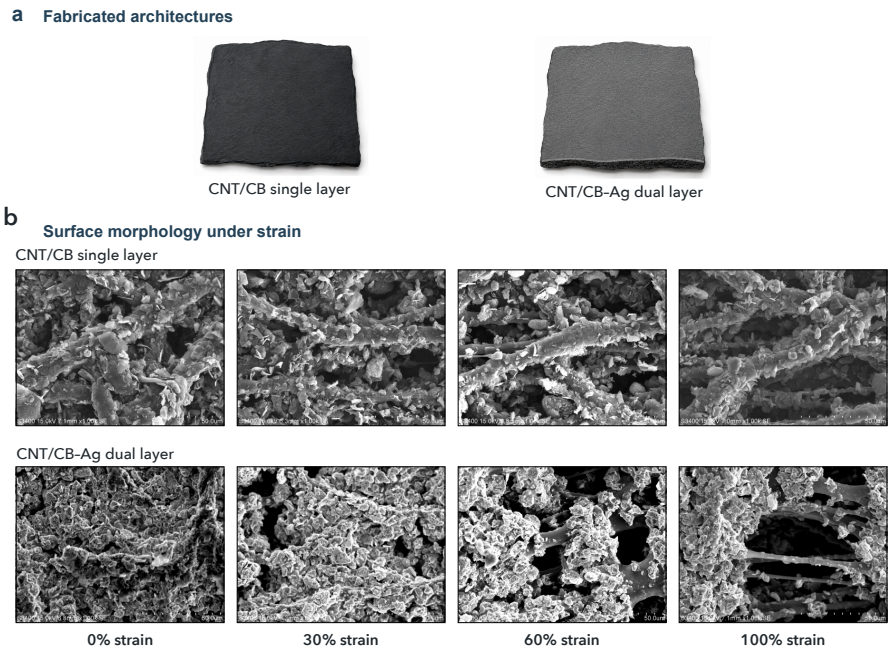


Figure S1: **Additional physical evidence for the conductive-network comparison.** (a) Photographs of the CNT/CB single-layer control and CNT/CB–Ag dual-layer candidate. (b) Representative SEM views of both architectures at $0\%$, $30\%$, $60\%$, and $100\%$ strain (scale bars, $50\,\mu$m). Electrical-response traces are shown in Figure 4 of the main paper.

## S4 Endpoint Audit Examples

Table S2 makes explicit how the same follow-up window distinguishes a reported full design from weaker observations. These are decision rules for the reported endpoints, not additional prediction results.

Table S2: **Illustrative endpoint decisions for a frozen candidate.**

| Follow-up observation | Exact closure | Validated closure |
|---|---|---|
| One or two local relations, but no complete path | No | No |
| Complete path mentioned only in references or a review | No | No |
| All three relations reported in new original research, without system validation | Yes | No |
| Full path with system-level validation in the target application | Yes | Yes |