The scientific paper needs an uncertainty layer

Kaeli Rizzo[1,*], Caroline M. Weisman[2,*,#], Richard Sever[3,4], Peter K. Koo[1,#]

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2] Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
[3] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[4] openRxiv Corp, Davis, CA, USA
* These authors contributed equally
[#] Send correspondence to: cara.weisman@gmail.com and koo@cshl.edu

Summary

AI assistants and scientific agents increasingly shape how biologists navigate a literature no one can keep up with. They can extract results at scale, but judging what those results establish depends on how the evidence was generated, what alternatives remain, and how far the findings generalize. We argue that AI-ready papers should be machine-assessable, not only machine-readable. We propose an uncertainty layer that keeps these qualifications attached as scientific outputs move beyond the paper. By preserving the boundaries of what the evidence can support, this layer could help scientists use literature-scale AI for more reliable scientific judgment.

Introduction

Biology produces more scientific output than any individual can read. Papers, preprints, datasets, models, benchmarks, and tools continue to accumulate[1,2], while the "observable universe" of any individual scientist remains bounded[3]. Researchers can follow a shrinking fraction of the work relevant to their questions, even within relatively narrow specialties. The cost is particularly high for integrative research, where drawing connections across fields depends on finding results, methods, and resources outside a researcher's immediate expertise[3]. Important prior evidence may be missed, useful scientific outputs may go unused, and hypotheses may be evaluated without accounting for results that could change their interpretation. The gap between what the scientific record contains and what researchers can effectively use is widening.

AI may further accelerate the production, processing, and summarization of scientific material[4], but it also offers a way to navigate it. AI could serve as a triage system that scans more of the literature than any individual, structures what it finds, and directs researchers toward work that merits closer examination[5,6]. Here, we focus on assistants based on large language models (LLMs) and on scientific agents that navigate and interpret the scientific literature and support multi-step research workflows, rather than task-specific predictive models. These systems are entering scientific evaluation and research workflows, including manuscript feedback, peer review, hypothesis generation, literature synthesis, experimental planning, and data analysis[5–16].

The scientific value of a paper depends not only on what it reports, but on how its evidence is interpreted. Novice and expert readers can examine the same study yet draw different conclusions. Experts are more likely to recognize which measurements are proxies, which assumptions are contested, which controls address relevant alternatives, how far a result generalizes to other contexts, and what evidence would change the conclusion[17,18]. They are also better positioned to judge when a reported conclusion should be accepted, narrowed, discounted, or left unresolved, and which qualifications matter for the question at hand. At the core of this expert reading is deciding what the evidence establishes and what remains unresolved. That uncertainty may concern statistical variability or measurement error, the limits of what was shown, assumptions required for inference, or alternative interpretations compatible with the evidence. These judgments depend on understanding how a study's framing, data, analyses, and validation constrain its results and claims. The scientific record exposes its contents more readily than the qualifications needed to judge them.

These expert qualifications are rarely recorded in an explicit, reusable form. Each reader must reconstruct them from information distributed across the paper, and AI systems must repeat that process for each assessment. Because LLM outputs are sensitive to prompts and models[19,26], relevant conditions or limitations may be recognized in one assessment and missed in another.

The problem becomes more consequential when individual parts of a study circulate beyond the paper. We use the term *research artifacts*[20–23], or simply *artifacts*, to refer to the data, methods, code, models, benchmarks, results, figures, annotations, and claims associated with a study. Although connected within the study, these artifacts may later be cited, transformed, or reused independently. AI systems may therefore encounter them without the paper's full context. Even when the complete paper is available, the information needed to judge what

an artifact can support remains distributed across methods, captions, supplements, cited resources, and field knowledge. Limitations sections provide only a partial, author-selected account and usually apply to the paper as a whole[24,25]. These qualifications matter when an LLM-based assistant or scientific agent interprets a paper, compares studies, or uses an artifact in a new context. A context-specific result can become a model label, database relation, or apparently general fact once the conditions that delimit its interpretation are left behind. Use of these systems at scale amplifies the consequences of this longstanding loss of context.

The consequences are already visible in current systems. LLM-generated summaries of scientific research can omit details that limit or even undermine a conclusion and may generalize beyond what the original study supports[26]. LLM-generated critiques may also be insufficiently grounded in the source paper[27]. When such outputs enter study comparisons, literature syntheses, extracted claims, experimental plans, or recommendations, errors can propagate faster than expert readers can detect or correct them. As scientific agents increasingly use literature-derived assessments to select analyses and experiments[7-10], a mistaken appraisal could direct substantial time and resources toward a line of inquiry built on evidence that was weaker or narrower than assumed.

Current efforts to make scientific papers more AI-ready address distinct functions, including machine-readable document structure, provenance and reproducibility, documentation for reuse, and explicit representations of scientific claims and evidence[21–23,28–38] (**Box 1**). Booeshaghi and colleagues argue that publication systems should disseminate scientific results in machine-readable form alongside scientific narratives, with claims, evidence, and evaluations made explicit[20]. OpenEval provides a post hoc proof of concept by extracting claims from existing papers, grouping them into results, and representing machine-generated and peer-review-derived evaluations in a structured form[20]. QED Science organizes manuscript assessment around claims and the evidence reported in support of them, using AI agents to identify potential gaps in that support[16]. These approaches make claims, evidence links, and judgments of support more explicit. But assessing whether evidence supports a claim still requires judging the conditions under which that evidence was generated, analyzed, and validated. Those qualifications often remain distributed across the paper and field knowledge. What is missing is a reusable representation of those qualifications, linked to the scientific objects and relationships they constrain.

In this Perspective, we argue that AI-ready papers should be machine-assessable, not only machine-readable. We propose an uncertainty layer that makes this qualified evidence structure explicit. Each record connects a known condition to what it leaves unresolved about the interpretation, scope, or evidentiary support of a specific scientific object or relationship. An artifact map preserves the dependencies among data, analyses, results, and claims, while the records remain linked to the objects and relationships they qualify as those objects are interpreted, transformed, and reused. For a specified scientific question or use, an assessment traces the relevant dependencies, identifies which records matter, and preserves the basis for the resulting judgment. A capable system may reconstruct many of these qualifications from the full paper. The proposal is to retain them as shared records linked to the objects they qualify. Those records could then be inspected, contested, revised, and reused. We also outline an agent-assisted, human-reviewed path for constructing and maintaining these records. The broader aim is to preserve the basis for scientific appraisal so it can be reused and revised rather than rebuilt for each assessment.

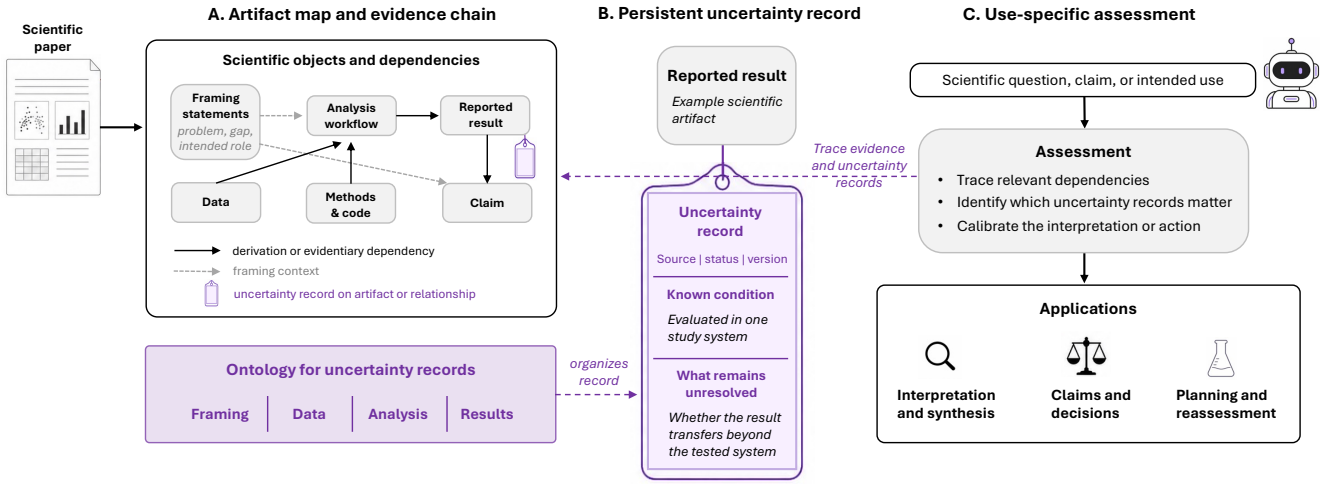


Figure 1. Architecture of the uncertainty layer. (A) A paper is represented as an artifact map of scientific objects and selected dependencies, with framing statements, data, methods and code, analysis workflows, reported results, and claims

shown as examples. Solid arrows indicate derivation or evidentiary dependency; dashed gray arrows indicate framing context. The ontology organizes uncertainty records through one contextual branch, Framing, and three evidence-chain branches, Data, Analysis, and Results. **(B)** A reported result illustrates an artifact linked to a persistent uncertainty record that identifies a known condition, what it leaves unresolved, and the record's source, status, and version. Records remain linked as artifacts are transformed or reused. **(C)** Given a scientific question, claim, or intended use, an LLM-based scientific assistant or agent can use the layer to trace relevant evidence and dependencies, identify which uncertainty records matter for the use, and calibrate its interpretation or action while preserving the basis of the judgment. Illustrative applications include paper interpretation and literature synthesis, claim calibration and suitability assessment, experimental planning, and reassessment as evidence changes.

Box 1. What makes a scientific paper AI-ready?

"AI-ready" has no settled definition. Here, we use the term broadly for efforts to make papers and their associated research records easier for computational systems to access, connect, inspect, interpret, and reuse. Several complementary capabilities contribute to this goal.

Machine readability and structured representation make the contents of a paper computationally recognizable. Titles, abstracts, sections, figures, tables, references, methods, datasets, code, claims, and metadata can be represented through structured reporting, semantic markup, persistent identifiers, and linked entities[22,31]. Nanopublications represent individual scientific assertions as identifiable, machine-interpretable objects connected to provenance and other metadata[23,36].

Claim-evidence representation connects assertions to supporting or challenging evidence, methods, agents, and provenance. Micropublications and SEPIO provide formal structures for these relationships[21,38], while ECO provides a standardized vocabulary for evidence types and assertion methods[37].

Research-object packaging, provenance, and reproducibility preserve the identity, origins, dependencies, and production history of scientific objects. FAIR principles describe findable, accessible, interoperable, and reusable digital research objects; RO-Crate packages research objects with structured metadata; and W3C PROV represents relations among entities, activities, and agents[34,35,42]. Executable workflows, software environments, data, code, and protocols can make an analysis inspectable or rerunnable[32,33]. Reconstructing how an analysis was performed still leaves the interpretation of its results to be assessed.

Documentation frameworks for reuse and evidence synthesis record information needed to select, compare, and evaluate scientific objects. Datasheets describe how datasets were assembled and which uses may be appropriate or problematic[28]. Model cards document intended uses, evaluation conditions, performance, and limitations[29]. GRADE assesses certainty in a body of evidence for a defined outcome and distinguishes certainty of evidence from the strength of a recommendation[30]. These frameworks document conditions and limitations for a designated object or assess certainty for a defined outcome.

Machine-readable results and claim-support assessment make claims, results, and support judgments more explicit. Booeshaghi and colleagues argue that scientific results should be disseminated in machine-readable form alongside scientific narratives, with claims, evidence, and evaluations represented explicitly[20]. OpenEval provides a post hoc proof of concept by extracting atomic claims from existing papers, linking them to source passages and evidence types, grouping them into results, and structuring model-generated and peer-review-derived evaluations[20]. QED Science organizes manuscript assessment around claims and the evidence reported in support of them, identifying gaps in that support[16].

Across these capabilities, the unresolved gap is a shared representation of the qualifications that determine what the linked objects and relationships of a study can support. Without such a representation, computational systems can recover what was reported, how it was produced, and which claims it was used to support, but must still reconstruct whether that support is warranted and how far it extends. A paper can therefore be machine-readable and richly linked, and its analyses reproducible, without its evidence being machine-assessable.

An uncertainty layer for machine-assessable science

Scientific interpretation requires understanding what each aspect of a study supports, the conditions under which that support was established, and what remains unresolved. The proposed uncertainty layer represents this qualified evidence structure through persistent records linked to the scientific objects and relationships that compose a study (**Figure 1**). The layer makes these qualifications available for interpretation, comparison, and summarization while keeping them connected as objects are transformed or reused. Many quantitative frameworks characterize uncertainty in a measurement, estimate, or prediction, using distinctions such as data and model uncertainty, aleatoric and epistemic uncertainty, or reducible and irreducible uncertainty[39–41]. These distinctions are useful, but they capture only part of what remains unresolved in scientific interpretation. A result may have narrow error bars yet depend on a proxy measurement, a restricted study system, an analytical assumption, or validation that does not address the interpretation attached to it. The uncertainty layer therefore records what a known feature or condition leaves unresolved about an artifact's interpretation, scope, or

evidentiary support, including forms that cannot be expressed as a probability or confidence interval. That a result was obtained in the K562 leukemia cell line, for example, is not itself uncertain; what remains uncertain is whether it extends beyond that context. An uncertainty record connects the known condition to what it leaves unresolved, thereby defining a boundary of support for the artifact.

Within this framework, a paper and its associated research record can be represented as a set of artifacts at different levels of granularity, together with the dependencies that connect them. An artifact map identifies scientific objects and the dependencies among them. Together, the artifact map and uncertainty layer represent a paper as a qualified evidence chain in which scientific outputs remain connected to the conditions and dependencies that shape what they support. Assessment follows the relevant dependency paths from a claim, summary, comparison, or proposed use to the supporting evidence and identifies which records matter.

Depending on the field and purpose, the map may distinguish individual datasets, methods, code, models, results, figures, annotations, and claims, or group artifacts according to their stages or roles in a study. A practical starting point is to represent as separate objects those that are independently cited, transformed, or reused. The uncertainty ontology provides a common structure for organizing records linked to those artifacts and relationships. Study-specific records describe uncertainty associated with particular artifacts or relationships. A record might state that a dataset represents a restricted population and that transfer beyond it is untested, or that a derived object depends on a threshold whose stability has not been evaluated.

Uncertainty at one stage of the evidence chain may alter the interpretation of later outputs. A limitation in a measurement may affect the analyses derived from it, while an analytical assumption may alter the interpretation of a reported result or claim. Uncertainty may also arise in the inference connecting one artifact to another, such as whether a measurement captures the intended biological quantity or a result supports a mechanistic claim. The artifact map preserves the relevant dependency path. Most uncertainty records qualify artifacts, while others may qualify the processing or inferential relationship through which one artifact supports another.

Each uncertainty record identifies the artifact or relationship it qualifies, the known feature or condition that grounds the uncertainty, and what remains unresolved. It also records the ontology category, source, status, and record version, together with the artifact and ontology versions to which it applies. These fields allow records to be inspected, contested, updated, and traced over time. Existing proposals for AI-ready publishing provide persistent identifiers, structured metadata, provenance links, and versioning that could keep uncertainty records connected across systems and uses[20,31,34,35,42,43]. The uncertainty layer would record what a known condition leaves unresolved. Records characterize an artifact or relationship independently of any particular assessment. Once a scientific question or use is specified, the relevant dependency paths are traced to determine which records are material. A record is material when accounting for it could change a judgment, narrow that judgment's scope, or alter the conditions under which it holds. Because the same artifact can later serve as a result, model label, experimental prior, or database relation, its records remain linked even as their consequences change across uses.

An ontology for locating uncertainty in scientific interpretation

Reliable interpretation requires locating where uncertainty arises and how it changes what scientific outputs can support. The ontology comprises one contextual branch, *Framing*, and three branches along the evidence chain: *Data*, *Analysis*, and *Results*. Framing qualifies the problem, contribution, or role the evidence is taken to support. Data and Analysis qualify how evidence is generated and transformed; Results qualifies what reported outputs establish. The branches identify where a qualification arises, not the type of artifact being qualified, and a single artifact can carry records from multiple branches. The ontology is a reference structure, not a checklist, and a study need not populate every category. **Supplementary Note S1** and **Table S1** provide the detailed hierarchy, definitions, and examples across biological and computational settings. The ontology is an initial framework for scientific communities to test, revise, and extend. The shared requirement is a lightweight record schema; lower-level categories can remain domain-specific and evolve with scientific practice.

Framing records qualify the scientific problem a paper claims to address, its representation of the field and unresolved gaps, and the roles assigned to its artifacts. They qualify what problem, gap, contribution, or intended role the evidence is taken to support, not whether the underlying result was obtained. A result may therefore remain empirically supported even when the problem, contribution, or intended use is framed more broadly than the evidence justifies. These records do not assign a general score of novelty or importance. How a paper represents prior expectations shapes whether a result is surprising and what it changes about prior knowledge[44]. The stated gap and the role assigned to an artifact likewise determine what contribution or use the evidence is taken to support. Framing records attach to problem, gap, or contribution claims and to artifacts whose intended roles are being asserted.

Data records qualify what was actually studied, how the data were generated and sampled, and what their measurements and labels represent. They also capture restrictions inherited from reused resources and technical or integrity problems such as leakage and batch effects[45,46]. These records distinguish technical precision from representativeness, independence, measurement validity, and source history. A dataset may be technically precise yet poorly represent the population invoked by a paper, while reused data may carry limitations introduced during earlier collection, processing, or labeling.

Analysis records qualify how observations are transformed into reported results and scientific inferences. They capture whether comparisons and controls address confounding and plausible alternatives, how processing choices produce derived objects, which assumptions connect observed data to inferred quantities, and which properties or failure modes were evaluated. The design and conduct of an evaluation belong to Analysis; what that evaluation shows about evidentiary support, robustness, and tested range belongs under Results.

Results records qualify what a reported output establishes. Reported outputs can include effect estimates, trained models, benchmark scores, annotations, structural models, and visualizations. Results records distinguish the evidence supporting an output, its observed stability to the variations examined, the range over which it has been directly tested, and the interpretations that remain compatible with it. The same object may carry Analysis records when it serves as an intermediate and Results records when it is itself a reported output. Which records apply depends on the object's role in the evidence chain.

The branches can overlap because a single issue may enter at one stage and alter what later objects support. Conditions of data generation may delimit a result's tested range; an inadequate control in Analysis may weaken its evidentiary support under Results; and a processing choice may become consequential when the output's stability to that choice has not been established. The map and linked records preserve both where the issue arose and how it affects dependent objects.

The ontology records uncertainty before its importance to any particular question or use is known. Assessment then determines which records are material to the interpretation or use at hand. A model evaluated in one hospital has been tested in that setting; whether its performance transfers to other hospitals remains unresolved. That uncertainty may be central to deployment elsewhere and largely irrelevant to reproducing performance in the original setting. Similarly, an uncertainty in an experimental result may alter a broad claim while leaving a narrower claim intact. For claim assessment, material uncertainty may affect the relation asserted, its scope and strength, and whether the proposed mechanism is established.

A worked example: uncertainty across interpretation and reuse

The uncertainties that matter depend on the question asked of an artifact. The same records may shape interpretation of the source paper and later become material in different ways when the artifact is reused. A clinical model evaluated in one hospital leaves transfer to other care settings unresolved; a protein-variant assay performed in a heterologous system may not establish function in its native cellular context. A regulatory-genomics annotation illustrates more concretely how one artifact can acquire several scientific roles (**Figure 2**). Consider a study that uses CRISPR interference (CRISPRi) in K562 leukemia cells to repress candidate regulatory elements and measure effects on nearby genes[47,48]. CRISPRi targeting candidate element E reproducibly reduces expression of gene G, and the study releases this result as part of a public candidate element–gene annotation resource[49,50]. Within the source paper, the direct observation is that targeting E reduced G expression in K562 cells under the tested conditions. Interpreting the paper requires distinguishing this observation from the stronger inference that E regulates G, while carrying forward the cellular context, perturbation design, assignment procedure, and unresolved directness. These qualifications can inform summaries, comparisons with other studies, and assessment of the authors' conclusions before the annotation is reused.

For interpretation of the source study, the artifact map connects the source data, perturbation design, analysis, reported result, and derived E–G annotation. Linked records state that the effect was observed in K562 cells, leaving transfer across cell states untested, and that the perturbation and expression readout support a context-specific functional relationship without establishing directness or mechanism. Other records address candidate and guide selection, assignment rules, controls, validation, evidentiary support, and tested range. A Framing record distinguishes whether the annotation is assigned the role of a candidate perturbation-associated relation, an established regulatory interaction, or mechanistic evidence. Together, these records show that the evidence supports a context-specific perturbation result while leaving stronger interpretations unresolved.

A later study may use E–G annotations from the resource as labels for training or benchmarking a regulatory-link prediction model. Guide specificity affects whether a positive label can be attributed to the intended element, whereas guide efficacy affects whether a tested pair without a detectable effect can be treated as a negative label[48,51]. Statistical thresholds and assignment rules determine which quantitative observations become positive or negative labels[49,50]. When these observations are converted into labels, uncertainty in guide efficacy becomes

uncertainty about the validity of negative labels, while binarizing a quantitative perturbation effect introduces uncertainty through the selected threshold. Independent guides or orthogonal perturbations may reduce parts of this uncertainty[51,52]. The artifact map preserves the relevant dependencies, while the linked records distinguish uncertainty inherited from the source study, altered by the new role, reduced through validation, or introduced during reuse. The K562 context limits what the labels establish about other cell states, candidate selection shapes the represented distribution of regulatory relationships, and the evidentiary support for each annotation determines whether it should be treated as a perturbation-associated relation or a more strongly validated regulatory assignment. Records that were secondary to interpreting the original result can ultimately become central to label validity, benchmark construction, and the interpretation of apparent prediction errors.

The annotation may instead be used to prioritize candidate perturbations in primary T cells. The key question is whether an effect observed in K562 cells is informative in a different cellular context. Guide properties and the original assignment rule remain relevant, but cellular context and tested range matter more for this use. The annotation may remain useful for prioritization while carrying substantial uncertainty about transfer across cell states and experimental systems.

A third use occurs when the annotation is incorporated into a review, database, or knowledge graph. The source observation may then be compressed into the relation "E regulates G." This compression facilitates retrieval and integration, but it can convert a context-specific perturbation result into an apparently general biological fact. The relation may omit that the evidence came from CRISPRi in K562 cells and depended on a particular expression readout and assignment rule[49,50]. It may also obscure whether the effect was direct, whether G was the direct or primary target[53,54], and whether independent guides or orthogonal perturbations provided additional support[51,52]. Evidentiary support, tested range, and interpretive uncertainty become central to how the relation should be represented and understood.

The E–G annotation first appears as a reported output and later serves as a model label, an input for experimental prioritization, and a relation in a review or knowledge resource. Its role changes across these uses, while its uncertainty records remain connected and traceable. Preserving the broader record makes the same context available when interpreting the source paper and using the annotation in later applications, without reconstructing the study from scratch. Reuse begins with an artifact and asks which uncertainties matter in a new role. Claim assessment begins with a claim and asks how far the supporting evidence reaches.

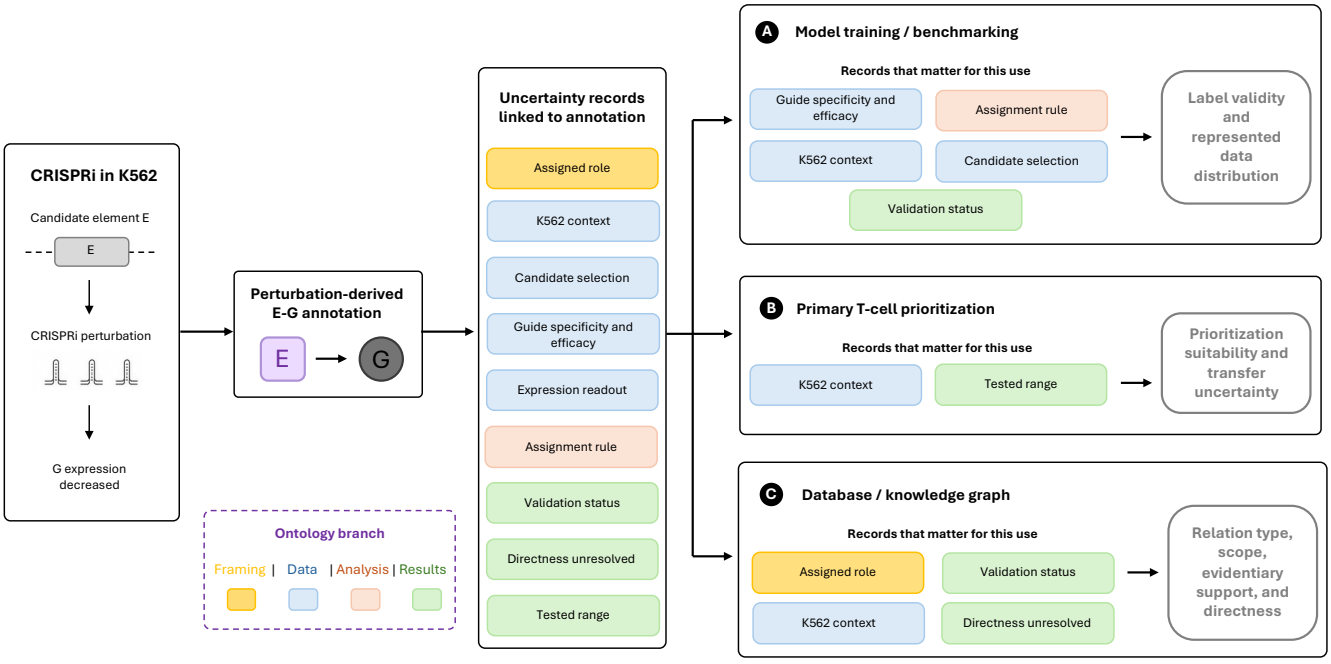


Figure 2. Reuse changes which uncertainties matter. A CRISPRi experiment in K562 cells produces an E–G annotation linked to uncertainty records from the source study. When the annotation is reused as a model label, prioritization input, or database relation, different records matter for label validity, transfer, and relation typing. Colors indicate the ontology branch under which each record is organized; the Framing record "Assigned role" qualifies whether the annotation is treated as a candidate relation, an established regulatory interaction, or mechanistic evidence. Other records remain linked for later uses. Central-column labels abbreviate full records that link the indicated feature or condition to what it leaves unresolved.

From qualified evidence to calibrated claims

Scientific claim extraction and verification systems increasingly identify claims and connect them to supporting evidence[55–58]. A claim may be linked to the correct evidence while expressing a stronger relation, broader scope, greater confidence, or more specific mechanism than the evidence establishes. Claim assessment follows the artifact map from the claim to the results, analyses, data, and dependencies on which it rests. It then asks which uncertainty records affect the claim's relation type, scope, strength, and mechanism. Preserving the evidence chain makes explicit part of the reasoning expert readers use to decide whether a conclusion follows from a study.

Consider the relation extracted in the preceding example: "E regulates G" (**Figure 3**). The trace connects this statement to the perturbation-derived E–G annotation, the observed reduction in G expression following CRISPRi targeting of candidate element E, the analysis used to assign the relation, and the source experiment in K562 cells[49,50]. The observed result is that CRISPRi targeting E reproducibly reduced G expression under the tested conditions. Representing this result as "E regulates G" interprets the observation as evidence of a functional regulatory relationship. Calling E a direct enhancer of G makes stronger claims about enhancer identity, G's status as the direct or primary target, and the underlying mechanism.

CRISPRi provides intervention evidence[47,48,59], but the relation supported by that intervention depends on the assay and analysis. Guide efficacy affects whether the intended perturbation occurred[51,59], whereas guide specificity and the spatial extent of induced repression affect whether the response can be attributed to E[53,54]. The expression readout may include direct and indirect transcriptional effects, while candidate selection, statistical thresholds, and assignment rules shape which element–gene relation is reported. A reproducible effect can support a functional regulatory relationship in the tested context without establishing direct enhancer activity, whether G is the direct or primary target, or the underlying mechanism. Transfer across cellular states also remains unresolved. Independent guides and orthogonal perturbations can address different parts of these uncertainties[51,52].

These records inform how the relation should be represented, including its type, scope, strength, and mechanistic status. Records that have been addressed through controls, sensitivity analyses, or replication remain linked without affecting the current assessment, as do records unrelated to this inference.

A calibrated formulation could be:

> *CRISPRi targeting candidate element E reproducibly reduced G expression in K562 cells, supporting a candidate functional regulatory relationship between E and G in this context. Direct enhancer activity, whether G is the direct or primary target, the underlying mechanism, and transfer across cellular states remain unresolved because the tested perturbation, expression readout, and available validation do not by themselves establish these properties.*

The generic edge "E regulates G" becomes a candidate relation with a defined evidence type, scope, and set of unresolved questions. The observed perturbation effect, the supported biological inference, and the boundary of support remain distinguishable as the relation enters summaries, syntheses, reviews, databases, knowledge graphs, and AI systems. This reduces the risk that an assay-supported candidate interaction will appear as an unqualified biological fact.

The unresolved questions identify what evidence could change the relation's status. Convergent results from orthogonal perturbations or measurements could strengthen support for enhancer identity, direct or primary targeting of G, mechanism, or transfer across contexts. Conflicting evidence could weaken the relation or support a different interpretation. Experimental planning can then focus on the unresolved boundary and prioritize evidence capable of changing the relation's type, scope, strength, or mechanistic status.

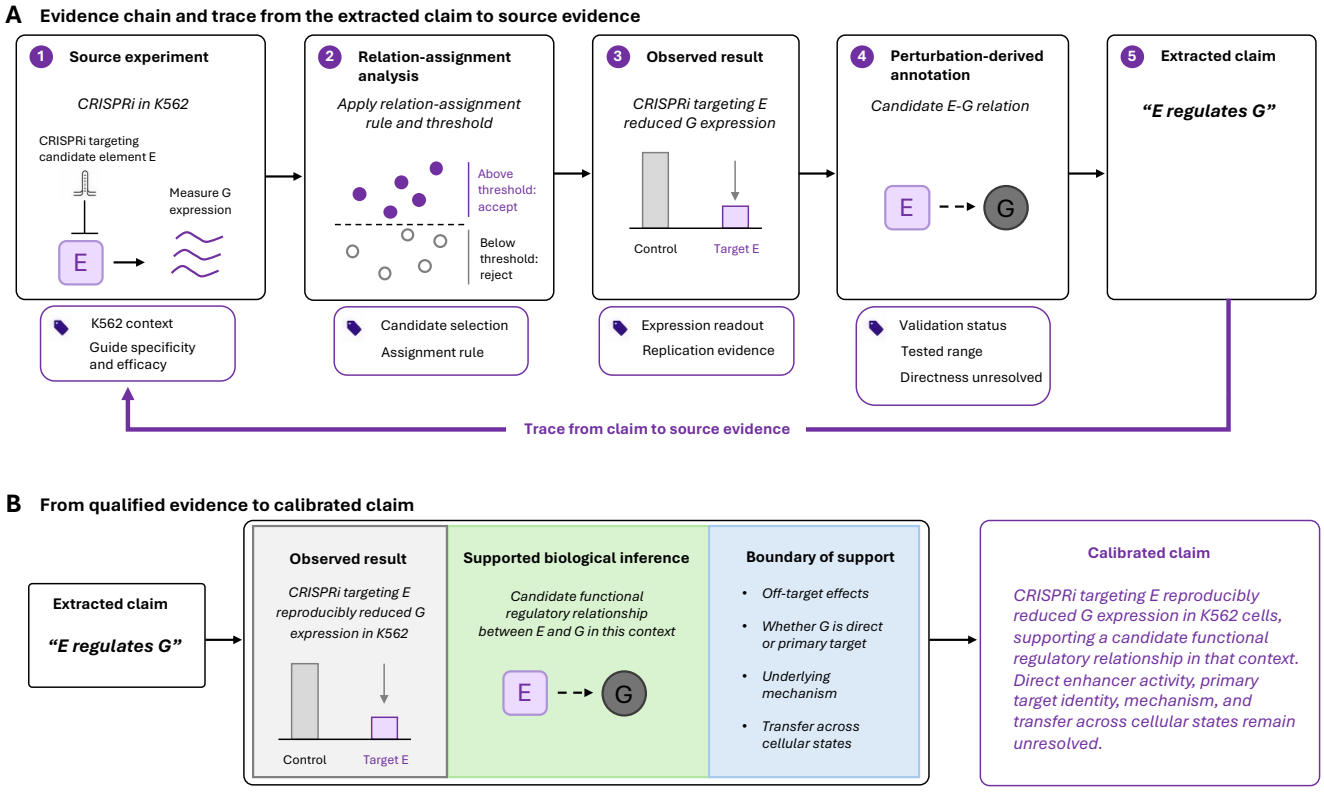


Figure 3. Claim assessment separates observation, inference, and boundary of support. The extracted relation "E regulates G" is traced through the source experiment, relation-assignment analysis, observed result, and perturbation-derived annotation. Relevant uncertainty records delimit the supported inference and yield a calibrated claim.

Discussion: Uncertainty as infrastructure for scientific interpretation

LLM-based assistants and scientific agents will increasingly mediate how scientists navigate, interpret, and synthesize a literature no individual can read. Their usefulness will depend not only on finding relevant papers, but on judging what the evidence supports. The scientific record was built for human readers who reconstruct much of the context needed for that judgment from methods, figures, supplements, prior work, and field knowledge. As AI systems take on more of the work of reading and acting on the literature, repeatedly reconstructing that context for each question becomes a limitation. The problem is compounded when scientific outputs leave the paper in which they first appeared. Datasets, methods, models, results, and annotations are routinely reused independently, while the qualifications that delimit what they support may be left behind. At literature scale, this can make evidence appear to support stronger or broader conclusions than warranted. An uncertainty layer would make those qualifications explicit and keep them linked to the data, analyses, results, and claims they constrain. It would preserve how scientific outputs were produced, the conditions that delimit their interpretation, and how uncertainty in upstream artifacts affects downstream conclusions.

Persistent records would reduce what AI systems must reconstruct. At present, a system interpreting a paper or using one of its artifacts must often recover its assumptions, tested conditions, dependencies, and evidentiary limits from narrative text, captions, supplements, and cited resources. The same reconstruction may be repeated across queries, users, and systems. Given the sensitivity of LLM outputs to prompts and models[19], the recovered context may vary across those reconstructions. The uncertainty layer could reduce this repeated reconstruction by giving systems shared records to draw on during appraisal. The records would not be automatically correct, but they could be inspected, traced to their sources, contested, and revised. Making the basis of AI-mediated appraisal visible would allow scientists to inspect, challenge, and revise such assessments as these systems take on more of the work of reading and synthesis. The structure could also make these judgments more consistent and easier to audit by making clear when an answer should be qualified, when source evidence should be revisited, and when disagreement or expert review is needed.

Uncertainty records could change both scientific navigation and downstream reuse. Conventional search identifies topically relevant papers or artifacts, but scientific use also requires judging whether they were generated

and evaluated under conditions suited to the intended role. As outputs become annotations, training labels, database relations, or review statements, the conditions that delimit them are often lost. Keeping those conditions linked could distinguish objects that address the same topic but support different uses and reduce the risk that provisional, context-bound evidence hardens into general fact.

Linked and versioned records could also make the scientific record more responsive to new evidence. A replication, orthogonal assay, correction, or reanalysis may reduce one uncertainty, introduce another, or change the status of an artifact. If records remain connected to the artifacts and dependencies they qualify, later users could identify which derived labels, models, database relations, or claims require reassessment. The interpretation of a scientific object could then change with the evidence around it without requiring every later user to reconstruct its history independently.

Records would not eliminate disagreement among experts or AI systems. An uncertainty record describes an artifact or relationship independently of any particular assessment and is often grounded in documented features such as the study system, measurement, controls, or tested range. Scientific judgment enters in determining what those features leave unresolved and whether that uncertainty matters for a particular claim or use. Experts or AI systems may therefore agree on the same record yet differ in how it affects a downstream assessment. When disagreement concerns the record itself, multiple records can remain attached to the same artifact, each with its own source and status. The structure makes clear where disagreement enters. Records should indicate whether they were author-stated, inferred from the paper, introduced from field knowledge, or produced through reanalysis. Because no record can enumerate every open question, selecting which boundaries to record and at what granularity is itself a scientific judgment whose basis should remain traceable.

Fields differ in how they generate, validate, report, and reuse evidence. The high-level ontology should therefore provide a shared structure while scientific communities develop and revise lower-level categories suited to their study systems, measurements, methods, and modes of inference. Scientific objects, uncertainty records, and the ontology can change independently and therefore require separate versioning[43,60].

An uncertainty layer can introduce failure modes when its records are treated as complete or authoritative. Records will inevitably be partial, and uncertainties that are easy to formalize may receive more attention than those that depend on tacit or field-specific judgment. Detailed records may create false confidence, while transparent studies may appear weaker than poorly documented work simply because more uncertainty has been made visible. Adoption must avoid rewarding sparse disclosure or treating the number of recorded uncertainties as a measure of study quality. Missing or unassessable information should remain visible, so incomplete reporting does not create the appearance of greater certainty. The layer should remain lightweight and avoid opaque scores or single grades that collapse distinct forms of uncertainty into a universal judgment.

A practical starting point is post hoc, agent-assisted curation with human review. A dedicated scientific agent could construct a provisional artifact map and draft records from relatively explicit information in the paper and associated materials, including the studied system, sample structure, assay or model version, data split, controls, validation setting, tested range, and author-stated limitations[61–64]. Authors, reviewers, curators, and later domain experts could inspect, revise, contest, or append records, with every record and revision retaining provenance and applying to specified artifact and ontology versions. Journals or repositories could preserve these versioned records alongside the relevant artifacts. Preprint and publish-review-curate models increasingly separate dissemination from evaluation and curation, allowing reviews, assessments, and revisions to accumulate after dissemination and across article versions[65,66]. These models point toward a scientific record composed of linked, evolving objects rather than a paper treated as a fixed, stand-alone bundle. Such an ecosystem provides a natural setting for uncertainty records to remain attached to particular artifacts and to be revised as review, curation, or new evidence changes what those artifacts can support. Downstream agents could then retrieve a particular record version for use-specific assessment. Early pilots should test whether links between records and artifacts persist as artifacts are transformed or incorporated into downstream resources. Initial value need not depend on universal coverage; pilots could focus on widely reused or high-consequence artifacts within selected repositories, databases, or journal collections. Who should maintain and fund these records over the long term remains unresolved; early pilots should determine where they create sufficient value to justify sustained curation.

The relevant test is whether structured uncertainty records improve consequential scientific judgments. Evaluation should separate two questions: whether agents and experts can construct accurate, useful records, and whether those records improve downstream assessments. Early pilots should compare the layer with paper-level documentation and with assessments produced directly from the manuscript and associated materials. Record construction should be judged by experts for accuracy and relevance. Downstream value should be judged by whether supplying the same record version improves consistency across prompts and models and leads to better-calibrated claims and judgments about whether an artifact is suitable for a particular use. It should also help systems recognize insufficient or conflicting evidence and identify experiments that could resolve material uncertainty.

Scientific agents will increasingly act on judgments drawn from the literature. Persistent records would allow reviewers, curators, systematic reviewers, database developers, and later users to build on what has already been recorded rather than reconstructing the same assumptions and evidentiary limits. They would make the basis of later judgments easier to inspect, contest, and update. They could also make qualifications routinely recognized by specialists more accessible to early-stage researchers and scientists outside the field. Domain expertise would remain necessary to interpret the records and determine their materiality. Scientific artifacts should carry their contents, provenance, and the boundaries of what they can support.

Acknowledgements

The authors would like to thank Jennifer Doudna for feedback on an early draft of the manuscript, and David McCandlish, Benjamin Cowley, Bruce Stillman, and members of the Koo Lab for discussions that helped shape the uncertainty ontology.

Author contributions

K.R., C.W., and P.K. conceived and wrote the manuscript. R.S. provided conceptual feedback and contributed to manuscript revision.

Declaration of interests

P.K. is a paid consultant for and holds stock options in Inari Agriculture, Inc. P.K. serves on the Scientific Advisory Board of Collage Bio Inc. R.S. serves as Chief Science & Strategy Officer at openRxiv and as an advisor to Sciencecast. The remaining authors declare no competing interests.

Supplemental information

Document S1. Note S1 and Table S1.

References

1. White, K. Publications Output: U.S. Trends and International Comparisons. *National Science Foundation* **NSB-2020-6**, (2019).
2. Bornmann, L., Haunschild, R. & Mutz, R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanit. Soc. Sci. Commun.* **8**, 224 (2021).
3. Jones, B. F. The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder? *Review of Economic Studies* **76**, 283–317 (2009).
4. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
5. Asai, A. *et al.* Synthesizing scientific literature with retrieval-augmented language models. *Nature* **650**, 857–863 (2026).
6. Skarlinski, M. D. *et al.* Language agents achieve superhuman synthesis of scientific knowledge. *arXiv* (2024).
7. Gottweis, J. *et al.* Accelerating scientific discovery with Co-Scientist. *Nature* **655**, 487–496 (2026).
8. Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature* **655**, 497–505 (2026).
9. Lu, C. *et al.* Towards end-to-end automation of AI research. *Nature* **651**, 914–919 (2026).
10. Huang, K. *et al.* Autonomous biomedical research with an artificial intelligence agent. *Science* eadz4351 (2026).
11. Liang, W. *et al.* Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI* **1** (2024).
12. Liang, W. *et al.* Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. in 29575–29620 (Proceedings of the 41st International Conference on Machine Learning, 2024).
13. Thakkar, N. *et al.* A large-scale randomized study of large language model feedback in peer review. *Nature Machine*

Intelligence **8**, 326–336 (2026).

14. Bianchi, F., Queen, O., Thakkar, N., Sun, E. & Zou, J. Exploring the use of AI authors and reviewers at Agents4Science. *Nat. Biotechnol.* **44**, 11–14 (2026).

15. Li, B., Saini, A. K., Hernandez, J. G. & Moore, J. H. Agentic AI and the rise of in silico team science in biomedical research. *Nat. Biotechnol.* **44**, 711–725 (2026).

16. Maor, M. The case for a validity layer in life science research. *QED Science* https://www.qedscience.com/article/life-science-research (2026).

17. Nelms, A. A. & Segura-Totten, M. Expert-novice comparison reveals pedagogical implications for students' analysis of primary literature. *CBE Life Sci. Educ.* **18**, ar56 (2019).

18. Hubbard, K. E. & Dunbar, S. D. Perceptions of scientific research literature and strategies for reading papers depend on academic career stage. *PLoS One* **12**, e0189753 (2017).

19. Zhuo, J. *et al.* ProSA: Assessing and understanding the prompt sensitivity of LLMs. *Findings of the Association for Computational Linguistics: EMNLP* 1950–1976 (2024).

20. Booeshaghi, A. S., Luebbert, L. & Pachter, L. Science should be machine-readable. *bioRxiv* (2026).

21. Clark, T., Ciccarese, P. N. & Goble, C. A. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *J. Biomed. Semantics* **5**, 28 (2014).

22. Jaradeh, M. Y. *et al.* Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19)* **2019**, 243–246.

23. Groth, P., Gibson, A. & Velterop, J. The anatomy of a nanopublication. *Inf. Serv. Use* **30**, 51–56 (2010).

24. Ioannidis, J. P. A. Limitations are not properly acknowledged in the scientific literature. *J. Clin. Epidemiol.* **60**, 324–329 (2007).

25. Smith, J. J., Amershi, S., Barocas, S., Wallach, H. & Wortman Vaughan, J. REAL ML: Recognizing, exploring, and articulating limitations of machine learning research. in *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, New York, NY, USA, 2022).

26. Peters, U. & Chin-Yee, B. Generalization bias in large language model summarization of scientific research. *R. Soc. Open Sci.* **12**, 241776 (2025).

27. Ou, J. *et al.* CLAIMCHECK: How Grounded are LLM Critiques of Scientific Papers? *arXiv* (2025).

28. Gebru, T. *et al.* Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).

29. Mitchell, M. *et al.* Model cards for model reporting. in *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (ACM, New York, NY, USA, 2019).

30. Guyatt, G. H. *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* **336**, 924–926 (2008).

31. Donohoe, P., Sherman, J. & Mistry, A. The Long Road to JATS. JATS-Con Proceedings. *National Center for Biotechnology Information* (2015).

32. National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. *The National Academies Press* (2019) doi:10.17226/25303.

33. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).

34. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

35. World Wide Web Consortium. PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/. (2013).

36. Kuhn, T. *et al.* Nanopublications: A growing resource of provenance-centric scientific linked data. in *2018 IEEE 14th International Conference on e-Science (e-Science)* (IEEE, 2018).

37. Giglio, M. *et al.* ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Res.* **47**, D1186–D1194 (2019).

38. Brush, M. H., Shefchek, K. & Haendel, M. SEPIO: A semantic model for the integration and analysis of scientific evidence. in *CEUR Workshop Proceedings* (2016).

39. Walker, W. E. *et al.* Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integr. Assess.* **4**, 5–17 (2003).

40. Joint Committee for Guides in Metrology. Evaluation of measurement data: Guide to the expression of uncertainty in measurement. *JCGM* **100** (2008).

41. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2021).

42. Soiland-Reyes, S. *et al.* Packaging research artefacts with RO-Crate. *Data Sci.* **5**, 97–138 (2022).

43. Ciccarese, P. *et al.* PAV ontology: provenance, authoring and versioning. *J. Biomed. Semantics* **4**, 37 (2013).

44. Evans, J., Petroff, C. & King, G. Advancing science by designing for surprise. *Science* **393**, eaej4257 (2026).

45. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N. Y.)* **4**, 100804 (2023).

46. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

47. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

48. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

49. Gasperini, M. *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390.e19 (2019).

50. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).

51. Horlbeck, M. A. *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* **5**, e19760 (2016).

52. Pacalin, N. M. *et al.* Bidirectional epigenetic editing reveals hierarchies in gene regulation. *Nat. Biotechnol.* **43**, 355–368 (2025).

53. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).

54. Lensch, S. *et al.* Dynamic spreading of chromatin-mediated gene silencing and reactivation between neighboring genes in single cells. *eLife* **11**, e75115 (2022).

55. Wadden, D. *et al.* Fact or fiction: Verifying scientific claims. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020).

56. Wadden, D. *et al.* SciFact-Open: Towards open-domain scientific claim verification. *Findings of the Association for Computational Linguistics: EMNLP* 4719–4734 (2022).

57. Bekoulis, G., Papagiannopoulou, C. & Deligiannis, N. A review on Fact Extraction and verification. *ACM Comput. Surv.* **55**, 1–35 (2023).

58. Metropolitansky, D. & Larson, J. Towards effective extraction and evaluation of factual claims. *arXiv* (2025).

59. Gilbert, L. A. *et al.* Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).

60. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).

61. Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896 (2008).

62. Rocca-Serra, P. *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation

at the community level. *Bioinformatics* **26**, 2354–2356 (2010).

63. Sansone, S.-A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **37**, 358–367 (2019).
64. Simera, I. *et al.* Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med.* **8**, 24 (2010).
65. Sever, R. Biomedical publishing: Past historic, present continuous, future conditional. *PLoS Biol.* **21**, e3002234 (2023).
66. Stern, B. M. & O'Shea, E. K. A proposal for the future of scientific publishing in the life sciences. *PLoS Biol.* **17**, e3000116 (2019).

Note S1. Detailed uncertainty ontology

The uncertainty ontology organizes uncertainty records linked to scientific artifacts. The artifact map identifies those objects and the dependencies through which they were produced. The ontology provides a common structure for recording uncertainty about the artifacts, including uncertainty that arises from those dependencies. Provenance, status, and version information describe the origin and history of each record. Materiality is assessed once a particular question, claim, or use has been specified.

We use uncertainty for what known features or conditions leave unresolved about an artifact's interpretation, scope, or evidentiary support[39–41]. These records capture sources of actual or potential variability in what an artifact can support, whether or not that variability has been directly measured. Some uncertainty can be quantified directly through measurement error, sampling variability, confidence intervals, or predictive distributions. Other forms concern ambiguity, missing information, threats to validity, dependence on analytical choices, limits of generalization, or alternative interpretations that remain compatible with the evidence. These forms are not equivalent, and their importance cannot be reduced to a common scale.

Uncertainty is often grounded in a known feature of a study. That an experiment was performed in K562 cells is not uncertain. What remains uncertain is whether the result extends beyond that cellular context. A model version, data split, control, threshold, or validation setting may likewise be known. Its uncertainty record states what that feature leaves unresolved about the artifact's interpretation, scope, or evidentiary support. Some uncertainty records describe an unresolved issue directly. Others record evidentiary support, tested conditions, assumptions, or dependencies needed to determine what remains unresolved.

The ontology has four high-level branches: **Framing**, **Data**, **Analysis**, and **Results**. Each branch contains mid-level categories and more specific lower-level issues. Framing provides contextual qualifications; Data, Analysis, and Results locate qualifications along the evidence chain. An uncertainty may arise upstream and have distinct consequences for artifacts downstream. These consequences can be represented through linked records in different branches. A condition in Data may delimit the tested range of a Result. An inadequate control in Analysis may weaken the evidentiary support associated with that Result. A processing choice may become consequential when the stability of a derived output to that choice has not been established.

Study-specific uncertainty records use the ontology to describe particular artifacts and relationships. At minimum, a record should identify the artifact or relationship it qualifies, the ontology category, the known feature or condition that grounds the uncertainty, and what remains unresolved. It should also identify the source passage or evidence supporting the record; whether the record was stated by the authors, inferred from the paper, introduced from field knowledge, or produced through reanalysis; whether it is unresolved, addressed, contested, or unassessable; the record version; and the artifact and ontology versions to which it applies. A single artifact may carry multiple records. When uncertainty arises in a processing or inferential relationship, the record remains linked to that relationship and to the downstream artifact whose interpretation or evidentiary support is affected. Initial implementations could build on existing minimum-information checklists, structured metadata frameworks, standards registries, and reporting-guideline infrastructures[61–64].

Table S1 summarizes the four high-level branches and their mid-level categories. For each category, it states what the category captures and lists representative lower-level issues. The sections below expand this reference hierarchy by clarifying the scope and boundaries of each category, distinguishing related categories, and providing examples of how study-specific uncertainty records may be organized. The hierarchy is intended as an initial framework that scientific communities can test, revise, and extend. The shared requirement is a lightweight record schema; lower-level categories can remain domain-specific and evolve with scientific practice.

Framing

Framing records qualify the scientific problem a paper claims to address, its representation of the field and unresolved gap, and the roles assigned to its artifacts. They qualify what problem, gap, contribution, or intended role the evidence is taken to support, not whether the underlying result was obtained. A result may therefore remain empirically supported even when the problem, contribution, or intended use is framed more broadly than the evidence justifies. These records do not assign a general score of novelty or importance.

Problem formulation, field positioning, and gap identification

This category records uncertainty in how the scientific problem is scoped, why it is considered important, how prior work and contested models are represented, and how the stated gap and contribution are established. It does not provide a general judgment of novelty or importance. It applies when the representation of the problem, field, or gap changes what the study's artifacts or results are taken to support. Lower-level entries can distinguish

problem scope, background assumptions, representation of prior evidence, unresolved disagreement, gap identification, and contribution positioning.

Problem formulation concerns the boundaries and significance assigned to the problem. A study may address a narrow technical limitation while presenting it as a general obstacle to progress. It may define success using an outcome that captures only part of the scientific objective. A paper can also formulate a problem in a way that assumes the relevance of a particular mechanism, population, benchmark, or intervention before that relevance has been established. The resulting experiments may be well designed for the formulated problem even when the broader importance of that problem remains uncertain.

Field positioning concerns how the paper represents existing evidence, prevailing explanations, and areas of disagreement. The record becomes relevant when the representation of the field changes what the study is understood to contribute. A claimed advance may depend on treating a contested model as settled, overlooking evidence that already addresses the stated problem, or comparing against an incomplete account of prior approaches. The reported result may remain valid, but its significance or novelty may be narrower than the framing suggests.

Gap identification records the basis for the unmet need or missing knowledge that the paper claims to address. A gap may concern unavailable data, limited measurement, unresolved mechanism, weak predictive performance, lack of validation, or failure to study a particular system or population. The uncertainty record should preserve what evidence establishes the gap and which assumptions are required for it to be consequential. A paper may establish that a specific benchmark has not been solved without establishing that the benchmark reflects the scientific capability invoked by the paper.

Contribution positioning records how the paper connects its artifacts and results to the stated gap. An incremental improvement may be presented as resolving a general problem. A result that addresses one component of a mechanistic explanation may be presented as establishing the full mechanism. A new dataset may fill a coverage gap without resolving whether its measurements capture the intended biological quantity. These records specify the contribution attributed to the study and the evidence supporting that characterization.

Artifact role or intended use

This category records the function assigned to an artifact and the basis and conditions for treating it as suitable for that role. Lower-level entries can distinguish the assigned role, intended user or application, evidence for suitability, required conditions, and limits on transfer or reuse.

A dataset may be presented as a reference resource, a benchmark as a test of general performance, an annotation as an established biological relation, a trained model as a reusable predictor, or a result as mechanistic evidence. These roles are consequential because they determine how the artifact is interpreted and how later users are likely to reuse it. The role itself may be broader than the evidence supporting it. A dataset collected from a restricted population may be useful as a reference for that population without serving as a general reference. A benchmark may compare models reliably on one task without measuring general biological understanding. A perturbation-derived annotation may support a candidate relation without establishing a direct mechanism.

The record should state why the artifact was assigned that role and under which conditions its suitability was evaluated. These conditions remain relevant when the artifact travels into a new use. A later user may assign the artifact a different role, but the original role and its evidentiary basis remain part of its history.

Data

The Data branch concerns the material on which a study is built and what that material represents. Its categories distinguish the study system and source material, how the data were generated and assembled, the context in which they were produced, the structure of the sample, the history of reused resources, the validity of measurements and labels, and technical or integrity problems affecting the data object. Together, these records describe the represented system, the provenance of the data, and limitations inherited by the analysis.

Study system and source material identify what is represented. Context and conditions identify the state or setting in which it was observed. Data generation and assembly records how the data object was created. Sampling and structure records the composition and dependence structure of the resulting object.

Study system and source material

This category identifies the biological, clinical, observational, computational, or simulated system represented by the data. Lower-level entries can describe system identity, source and origin, population, organism, tissue, cell type, specimen, computational domain, simulation regime, and relevant properties of the source material.

The system determines the domain within which the evidence was generated. A result from an immortalized cell line may strongly support a conclusion about that cell line while leaving transfer to primary tissue unresolved.

A treatment effect estimated in a narrowly recruited cohort may not extend to populations differing in age, ancestry, comorbidity, treatment history, or care setting. An ecological study conducted at one site may not represent other environments or temporal regimes. A benchmark may represent only the computational task and data distribution used in its construction.

The uncertainty record should connect the represented system to the domain invoked by the artifact or claim. It should identify important differences between the source system and the broader system to which the result may later be applied. These differences identify where transfer remains untested.

Data generation and assembly

This category records the procedures and selection decisions through which samples, perturbations, measurements, labels, simulations, benchmarks, and other data objects were generated, collected, combined, curated, or constructed. Lower-level entries can distinguish recruitment, inclusion and exclusion, collection, perturbation delivery, assay execution, label generation, annotation transfer, curation, merging, library design, simulation design, and benchmark construction.

Generation choices determine which observations can appear in the data. Poor perturbation efficiency may mean that a negative readout does not represent the intended intervention. Selective loss to follow-up may leave an observational cohort systematically different from the enrolled population. A protein library may cover only substitutions that pass a synthesis or expression filter. An imaging dataset may include fields selected for image quality or visible phenotype. A benchmark may inherit its difficulty and class balance from the rules used to construct positive and negative examples.

Assembly can introduce additional uncertainty when data from different sources are combined. Differences in assays, collection periods, sites, labeling conventions, preprocessing, or inclusion criteria may create heterogeneity that is not visible in the merged object. The resulting dataset may appear unified even though its components support different comparisons or interpretations.

Context and conditions

This category records the biological, experimental, environmental, clinical, observational, or computational setting in which the data were generated. Lower-level entries can include genotype, developmental stage, cell state, disease severity, dose, timing, treatment, exposure, environmental condition, laboratory setting, clinical setting, assay condition, and computational domain.

Context can determine whether an observed effect appears, its magnitude, and its interpretation. A regulatory interaction measured in one cell state may change after differentiation or stimulation. Protein activity measured in a heterologous expression system may not establish activity in the native cellular environment. A clinical association observed under one treatment regime may not persist under another standard of care. Model performance may depend on acquisition equipment, annotation practice, software environment, or deployment conditions.

Context dependence may be measured through study design, although relevant conditions can remain unmeasured or confounded with the study system. Records should distinguish conditions that were fixed, varied, measured, inferred, or unavailable. What remains unresolved is the extent to which the artifact remains valid outside the recorded context.

Sampling and structure

This category describes what is represented, missing, imbalanced, nested, repeated, or dependent in the data. Lower-level entries can distinguish coverage, exclusions, imbalance, sample size, replicate structure, clustering, nesting, relatedness, repeated observations, missing components, and missing covariates.

A large number of observations may provide little independent evidence when they are nested, repeated, or otherwise dependent. Measurements may be nested within a small number of donors, sites, experimental batches, families, loci, images, or simulation runs. Technical replicates may improve measurement precision without increasing the number of independently sampled biological units. Train and test examples may share subjects, genomic regions, experimental sources, or near-duplicate observations.

Coverage concerns which parts of the relevant space are represented. A dataset may omit rare populations, difficult samples, low-signal conditions, long-range interactions, particular sequence classes, or regions in which the assay performs poorly. Imbalance may cause an aggregate result to be driven by common groups or easy cases. Missing covariates may prevent assessment of whether apparent effects depend on unrecorded differences among samples.

The record should describe the dependence and coverage structure of the sample and identify the comparisons or generalizations it limits.

Source history

This category preserves uncertainty inherited from reused scientific resources. Public datasets, reference genomes, pretrained models, clinical classifications, benchmark releases, curated annotations, and simulation outputs arrive with earlier decisions already embedded in them. Lower-level entries can include prior collection, processing, selection, exclusion, labeling, negative-set construction, split definition, correction, reference version, batch structure, and incomplete source documentation.

Reuse does not reset the history of an artifact. A public dataset may carry labels inferred from another assay, exclusions introduced during curation, or preprocessing that removed information needed for a later analysis. A pretrained model may carry training-data biases, target definitions, filtering rules, and unresolved leakage from its original development. A benchmark may preserve duplicated examples, incorrect labels, or a negative set that encodes an unintended shortcut. A reference annotation may combine evidence of different types and strengths under a single relation type.

Source-history records should remain attached when the artifact is incorporated into a new study. They allow later analyses and derived artifacts to be traced to inherited choices that may not be restated in the paper performing the reuse.

Measurement and label validity

This category concerns the relationship between recorded values, readouts, labels, or annotations and the quantities or categories they are intended to represent. Lower-level entries can distinguish construct validity, directness, proxy measurement, specificity, sensitivity, dynamic range, resolution, measurement error, observer variation, label ambiguity, and transferred labels.

A measurement can be technically precise without capturing the intended construct. A biomarker may correlate with a biological process without uniquely measuring it. A sequencing readout may reflect both the intended molecular event and technical biases in capture or amplification. A structural prediction score may correlate with model confidence without establishing structural correctness. A diagnostic label may combine heterogeneous disease processes or depend on criteria that change across sites.

Labels and annotations can carry additional uncertainty because they often compress evidence into discrete categories. A negative label may reflect absence of the phenomenon, insufficient sensitivity, incomplete testing, or a failed perturbation. A label transferred from one species, cell type, assay, or ontology may not retain the same meaning in the new setting. Consensus labels may obscure disagreement among annotators or uncertainty in the underlying evidence.

The record should distinguish the measured or assigned value from its scientific interpretation and identify unresolved aspects of that interpretation. Assumptions introduced later to infer a parameter, cause, or mechanism belong under Modeling and inference. Ambiguity or mismatch in what a label represents belongs here. Accidental mislabeling, corruption, or loss of sample identity belongs under Technical effects and data integrity.

Technical effects and data integrity

This category records technical effects and integrity problems that may alter the contents, identity, independence, or provenance of a data object. Lower-level entries can include batch effects, contamination, leakage, duplication, sample-identity errors, corruption, failed synchronization, incorrect labels, and untracked version changes.

Technical effects may introduce systematic patterns unrelated to the scientific quantity of interest. Batch effects can align with experimental condition, site, time, instrument, operator, or processing pipeline[46]. Contamination can introduce signal from another sample, species, reagent, or computational source. Duplicated or near-duplicated observations can inflate apparent sample size and performance. Leakage can expose information from validation or test data during feature construction, model selection, preprocessing, or training[45].

Integrity failures are especially consequential because they can invalidate the assumed identity or independence of the artifact. A sample swap, incorrect label, corrupted file, or silent change in a reference resource can produce a precise and reproducible analysis of the wrong object. The uncertainty record should distinguish known technical effects, suspected integrity problems, checks that were performed, and information that cannot be assessed from the available record.

Analysis

The Analysis branch concerns the study design and transformations that connect data to reported results. Its categories distinguish the design, comparisons, and controls used to isolate an effect; the processing steps through which data become intermediate objects; the models and assumptions used for inference; and the tests used to evaluate performance, validity, scope, or failure modes.

Study design, comparisons, and controls

This category records whether the structure of the study and its selected contrasts support the intended inference. Lower-level entries can distinguish allocation, randomization, blinding, temporal structure, comparator choice, confounding, independence, statistical power, positive and negative controls, baselines, null models, and plausible alternative explanations.

A comparison supports an inference only to the extent that relevant alternatives are separated. Confounding may arise during data generation, but it becomes an inferential limitation when the design or analysis cannot distinguish the factor of interest from correlated differences. Unmatched clinical groups may differ in disease severity or treatment history. An experimental perturbation may change several properties simultaneously. A computational method may be compared with weaker baselines, trained on additional information, or evaluated under conditions favorable to its architecture.

Controls establish different properties. A positive control may show that an assay can detect a known effect. A negative control may estimate nonspecific or background responses. An ablation may identify whether a model component contributes to performance. A null model may test whether an apparent pattern exceeds what is expected under a defined alternative. The record should identify which failure mode the control addresses and which alternatives remain unresolved.

Statistical power also belongs here because limited evidence can produce an uninformative result. Power depends on the number of independent units, expected effect size, variability, missingness, multiple testing, and the structure of the comparison. The Analysis branch records the design features that determine power. The Results branch records the precision and evidentiary support achieved.

Data processing and derived objects

This category records the workflows through which source data become analysis-ready inputs or intermediate artifacts. Lower-level entries can include filtering, normalization, alignment, correction, harmonization, aggregation, segmentation, clustering, imputation, feature construction, scalarization, software, parameters, thresholds, and reference versions.

Processing choices alter the information represented in a derived object. Filtering may exclude low-quality observations along with biologically relevant extremes. Normalization may remove technical variation while also altering biological differences. Alignment and correction methods may impose a reference structure on heterogeneous data. Clustering and segmentation convert continuous or ambiguous observations into discrete derived objects. Thresholding can turn quantitative evidence into binary calls.

Derived objects often circulate downstream as if they were direct observations. Counts, inferred abundances, diagnostic labels, clusters, embeddings, peak calls, attribution maps, structural models, and benchmark metrics remain dependent on the procedures that produced them. Processing can also collapse uncertainty. A single cluster label may conceal uncertainty in membership. A consensus peak may conceal variation among replicates. A scalar score may combine several quantities whose individual contributions are no longer visible.

The record should preserve the dependency between the derived object and its source data, software, parameters, and construction choices. It should identify whether alternative reasonable choices were evaluated and what remains unknown about the stability or meaning of the derived object.

Modeling and inference

This category concerns the assumptions, task definitions, and model choices that connect observed data to inferred quantities, mechanisms, or predictions. Lower-level entries can include task and target definition, sampling assumptions, independence, noise, functional form, relevant variables, model class, fitting or training procedure, regularization, calibration, identifiability, and misspecification.

Task and target definitions determine what a model is trained or fitted to represent. A model trained on an assay readout may later be interpreted as representing the underlying biological process. A classifier may reproduce the labeling convention encoded in its training set, while the correspondence of those labels to a stable disease category remains unresolved. A sequence model may accurately predict measurements within the training distribution, while its recovery of causal regulatory rules remains unestablished. These distinctions should be recorded before performance is interpreted.

Model assumptions can remain consequential even when fit or predictive performance is strong. Several kinetic mechanisms may explain the same biochemical time course. Correlated exposures may remain indistinguishable in an observational model. Redundant biological mechanisms may produce similar perturbation responses. A latent representation may support prediction without having a unique biological interpretation.

Identifiability concerns whether the available evidence distinguishes among models, parameters, mechanisms, or explanations. Calibration concerns whether reported probabilities or uncertainty estimates correspond to observed frequencies or errors under the evaluated conditions. Misspecification concerns whether

the model class excludes relevant structure or imposes an incorrect form. The fitting procedure, regularization, initialization, random seed, training curriculum, or stopping rule may further affect the inferred object.

An analytical choice creates uncertainty when reasonable alternatives remain, its effect on the output has not been tested, or the resulting interpretation exceeds what the model identifies.

Testing and validation

This category records what properties, uses, scopes, or failure modes were evaluated, how the evaluation was designed, and under which conditions it was conducted. Lower-level entries can include metrics, data splits, sensitivity analyses, robustness checks, perturbations, independent datasets, external validation, replication, calibration tests, and evaluated failure modes.

An evaluation provides evidence for the property it was designed to test. It may provide little evidence for another property that uses the same metric or terminology. Predictive accuracy does not establish calibration, causal identification, robustness, fairness, interpretability, or practical utility. Reproduction on the same samples or platform provides different evidence from replication in an independent setting.

Data splits define the generalization problem being tested. Randomly held-out observations may evaluate interpolation among similar examples. Holding out donors, sites, time periods, genomic regions, species, perturbation classes, or experimental systems tests different forms of transfer. Strong performance on held-out genomic regions does not establish performance on genetic variants, engineered sequences, or cell states poorly represented in training. Performance in one hospital does not establish transportability to another.

Validation may also address only part of the intended interpretation. Target engagement in an animal model does not establish therapeutic benefit in humans. Agreement between a model prediction and one assay does not establish mechanism. Replication of an association does not remove confounding shared across the original and replication studies.

The evaluation procedure belongs to Analysis; the evidence, stability, and tested range produced by that evaluation belong under Results.

Results

The Results branch records uncertainty associated with reported outputs. A reported output may be an effect estimate, trained model, benchmark score, annotation, ranked list, structural model, or visualization. A dataset or software package can also be qualified under Results when it is itself a primary reported output. The output itself belongs to the artifact map. Results records describe its evidentiary support, observed stability, tested range, and compatible interpretations.

The same type of object may appear under Analysis or Results depending on its role. An embedding used as an intermediate input belongs to the analysis path. An embedding reported as the primary scientific output is also qualified under Results. A fitted model may be an analytical instrument for estimating an effect or a reusable output whose performance and tested range are themselves central results.

Evidentiary support

This category records the quantitative, comparative, or empirical basis supporting a reported output. Lower-level entries can include effect size, uncertainty estimates, statistical support, control outcomes, baseline comparisons, ablations, replication, external validation, convergence among evidence types, and conflicting evidence.

Evidentiary support includes effect magnitude, precision, the number and independence of supporting observations, control performance, and comparison with plausible alternatives. A benchmark improvement may be statistically reliable but too small to change practical conclusions. A large effect may remain difficult to interpret if the comparison is confounded. Replication across sites or modalities may strengthen support, although shared design limitations can persist across repeated studies.

Observational association, intervention, temporal ordering, structural measurement, biochemical binding, predictive performance, and expert annotation provide different forms of support. Convergent evidence can strengthen a relation, although support may differ across mechanistic interpretations. Conflicting evidence should remain in the record with its source, type, and implications.

Stability and robustness

This category records the observed sensitivity of a reported output to the variations that were examined. Lower-level entries can include repeated measurements, donors, sites, batches, perturbations, random seeds, data splits, thresholds, preprocessing choices, model variants, datasets, and analysis pipelines.

A result may be stable under repeated measurement and sensitive to one processing choice. A model may produce similar aggregate performance across seeds while changing predictions for important subgroups or

individual examples. A cluster may recur across parameter settings but change when a different normalization method is used. A biological effect may replicate across guides in one cell line and vary across cellular states.

Robustness records should specify what was varied and what aspect of the output remained stable. Robustness to one class of variation does not establish robustness to untested changes. A sensitivity analysis that varies a threshold over a narrow interval provides different evidence from testing alternative methods or independent datasets.

Stability and robustness describe observed behavior. They should not be inferred solely from the use of a standard method or from the apparent precision of the reported output.

Tested range

This category records the systems, populations, conditions, inputs, and settings over which a reported output was directly evaluated. Lower-level entries can include species, populations, cell types, tissues, assays, doses, time points, environmental conditions, input distributions, perturbation classes, operating ranges, and deployment settings.

Every reported result has a tested range, even when the paper does not state it explicitly. A treatment effect established in one population has been directly tested in that population. A structural conclusion derived under one biochemical condition has been evaluated under that condition. A model validated in one cell type and one perturbation regime has not been directly tested in other biological states or intervention types.

Tested range is an empirical property of the reported output. It differs from the scope assigned in Framing. Scope records where the paper or later user asks the output to apply. Tested range records where the output has been evaluated. A difference between the two establishes an unresolved question about transfer. Whether that difference is material depends on the claim or use being assessed.

Interpretive uncertainty

This category records the interpretations compatible with a reported output and the assumptions, inferential steps, or additional evidence needed to distinguish among them. Lower-level entries can include observation versus inference, directness, causal interpretation, mechanism, target identity, practical meaning, alternative explanations, and unresolved biological significance.

A reported observation can support several interpretations. A biomarker change may reflect altered disease progression, a compensatory response, measurement timing, or another process associated with treatment. A perturbation effect may support a functional relationship without establishing whether the effect is direct or which mechanism produced it. A structural model may fit the available measurements while leaving alternative conformations possible. A benchmark improvement may establish better performance on the benchmark without establishing practical utility or scientific understanding.

Interpretive uncertainty records the supported inference, unresolved stronger interpretations, and evidence that could distinguish among them. Orthogonal measurements, independent interventions, temporal data, rescue experiments, additional controls, or testing in new contexts may address different parts of the uncertainty. No single validation strategy resolves every inferential step. Interpretive uncertainty is attached to the reported output before any particular assessment. Assessment then determines which alternatives and inferential commitments are material to the question or claim being considered.

Cross-cutting uncertainty in dependencies

Uncertainty can arise in the transition from one artifact to another and may generate linked records under multiple branches. When a processing assumption affects a derived result, or an inferential assumption affects a claim, the record remains linked to that relationship and to the downstream artifact whose interpretation or evidentiary support is affected. The artifact map preserves the dependency through which that uncertainty arose.

The conditions under which data were generated belong under Data. The range over which the resulting output has been directly evaluated belongs under Results. An inadequate control belongs under Analysis and changes the evidentiary support recorded under Results. A threshold used to construct a derived object belongs under Data processing and derived objects in Analysis. The observed sensitivity of the final output to that threshold belongs under Stability and robustness.

Uncertainty may change as it passes from an upstream artifact to a dependent artifact. An analytical step can carry an upstream limitation forward, alter its consequence, address it for a particular output, or introduce a separate uncertainty. For example, limited guide efficacy in a perturbation experiment creates uncertainty about whether a negative observation reflects absence of an effect. When observations are converted into binary training labels, that uncertainty becomes uncertainty about the validity of negative labels. Independent validation may reduce it, while the threshold used to binarize the observations may introduce a separate uncertainty.

Linked records preserve uncertainty in the transitions among artifacts. A source measurement and a derived result may each be clearly identified, while the relationship between them depends on a processing assumption. A reported effect and a mechanistic claim may each be explicit, while the inferential relationship connecting them remains uncertain. These records allow later users to trace what an artifact depends on without treating the dependency as a separate artifact.

Persistent records and use-specific materiality

The ontology records uncertainty before its importance to any particular question or use is known. Its records describe what uncertainties exist, where they arise, and which artifacts they qualify. The broader record remains connected as an artifact is interpreted or assigned new roles.

A perturbation-derived annotation may first serve as a reported result, then as a training label, an experimental prior, or a relation in a database. Different records become material in each role. Perturbation efficacy and assignment thresholds may dominate label validity. Cellular context may dominate transfer to another experimental system. Evidence type and directness may dominate representation in a knowledge graph.

Materiality is assessed for a specified question or use. An uncertainty is material when accounting for it could change the judgment being made. It may change how a result is interpreted or summarized, whether an artifact is suitable for a proposed role, or the type, scope, strength, or mechanistic status of a claim.

Whether a record is material is determined during use-specific assessment. A model evaluated in one hospital has a known boundary of tested range and an unresolved question about transfer. That record may be decisive for deployment in another hospital and largely irrelevant for reproducing performance within the original setting. An uncertainty in an experimental result may invalidate a broad mechanistic claim while leaving a narrower observational claim intact.

Extension and revision of the ontology

Fields can extend the high-level ontology with categories specific to their methods and inferential practices. Extensions should preserve the distinction between ontology categories and the study-specific records organized through them. An ontology category identifies a recurring aspect of uncertainty. A study-specific record identifies the artifact, the feature or condition that grounds the uncertainty, and what remains unresolved. Categories should be specific enough to avoid generic limitation statements and broad enough to apply across more than one study or implementation.

New categories may be needed when an existing category repeatedly combines issues that have different consequences for interpretation or propagation. Closely related issues can remain lower-level entries when they share the same role in the evidence chain. When one issue has distinct consequences across branches, linked records should represent those distinct aspects. For example, the conditions under which data were generated and the tested range of the resulting output are related, but they qualify different artifacts or stages of the evidence chain.

The ontology will also require versioning. Scientific methods change, new artifact types emerge, and communities revise what they consider adequate evidence or validation[60]. Ontology versions should preserve the category definitions used when a record was created. Record versions should preserve changes in what is known about an artifact. Artifact versions should identify which scientific object the record qualifies.

Versioned evolution preserves the history of the records organized through the ontology and keeps uncertainty connected to scientific artifacts as they are extracted, transformed, assessed, and reused.

Table S1. High- and mid-level ontology for uncertainty records.

Branch	Mid-level category	What the category captures	Representative lower-level issues
Framing	Problem formulation, field positioning, and gap identification	How the scientific problem is scoped, how the field and prior evidence are represented, and how the stated gap and contribution are established.	Problem scope; background assumptions; representation of prior evidence; unresolved disagreement; gap identification; contribution positioning
Framing	Artifact role or intended use	The role assigned to an artifact, the basis for treating it as suitable for that role, and the conditions and limits of that suitability.	Assigned role; intended user or application; evidence for suitability; required conditions; limits on transfer or reuse
Data	Study system and source material	Which biological, clinical, observational, computational, or simulated system is represented, and how far inferences can extend beyond it.	System identity; source and origin; population or model system; organism; tissue or cell type; specimen; relevant source properties
Data	Data generation and assembly	How samples, perturbations, measurements, labels, simulations, benchmarks, and other data objects were generated, selected, combined, curated, or constructed.	Recruitment or collection; inclusion and exclusion; perturbation delivery; assay execution; label generation; curation; merging; library design; simulation design; benchmark construction
Data	Context and conditions	The biological, experimental, environmental, clinical, observational, or computational setting in which data were generated, and how observations may depend on that setting.	Genotype; developmental stage; cell state; disease severity; dose; timing; treatment or exposure; environmental condition; assay condition; clinical or computational setting
Data	Sampling and structure	What is represented, missing, imbalanced, nested, repeated, or dependent in the data, and how that structure limits comparison and generalization.	Coverage; exclusions; imbalance; sample size; replicate structure; clustering or nesting; relatedness; repeated observations; missing components or covariates
Data	Source history	Uncertainty inherited from the prior collection, processing, selection, labeling, splitting, correction, and versioning of reused resources.	Prior collection; processing history; selection and exclusion; inherited labels; negative-set construction; split definition; correction; reference or resource version; incomplete documentation
Data	Measurement and label validity	The relationship between recorded measurements, readouts, labels, or annotations and the quantities or categories they are intended to represent.	Construct validity; directness; proxy measurement; specificity; sensitivity; dynamic range; resolution; measurement error; label ambiguity; transferred labels
Data	Technical effects and data integrity	Technical effects or integrity problems that may alter the contents, identity, independence, or provenance of a data object.	Batch effects; contamination; leakage; duplication; sample-identity errors; corruption; incorrect labels; untracked version changes
Analysis	Study design, comparisons, and controls	Whether the study design, selected comparisons, and controls isolate the intended effect and distinguish it from plausible alternatives.	Allocation; randomization; blinding; temporal structure; comparator choice; confounding; independence; statistical power; positive and negative controls; baselines; null models; alternative explanations
Analysis	Data processing and derived objects	How transformations and construction choices shape analysis-ready inputs or intermediate artifacts, including their stability and meaning.	Filtering; normalization; alignment; correction; harmonization; aggregation; segmentation; clustering; imputation; feature construction; scalarization; software; parameters; thresholds; reference versions
Analysis	Modeling and inference	The assumptions, task definitions, and model choices that connect observed data to inferred quantities, mechanisms, or predictions.	Task and target definition; sampling assumptions; independence; noise; functional form; relevant variables; model class; fitting or training; regularization; calibration; identifiability; misspecification
Analysis	Testing and validation	What properties, uses, scopes, and failure modes were evaluated, how the evaluation was designed, and whether it addresses the intended interpretation.	Metrics; data splits; sensitivity analyses; robustness checks; perturbations; independent datasets; external validation; replication; calibration tests; evaluated failure modes
Results	Evidentiary support	The quantitative, comparative, or empirical basis supporting a reported output, including the strength, type, convergence, and conflict of evidence.	Effect magnitude; uncertainty estimates; statistical support; control outcomes; baseline comparisons; ablations; replication; external validation; convergence among evidence types; conflicting evidence
Results	Stability and robustness	The observed sensitivity of a reported output to the variations that were examined.	Repeated measurements; donors; sites; batches; perturbations; random seeds; data splits; thresholds; preprocessing choices; model variants; datasets; analysis pipelines
Results	Tested range	The systems, populations, conditions, inputs, and settings over which a reported output was directly evaluated, and what remains untested beyond them.	Species; populations; cell types; tissues; assays; doses; time points; environmental conditions; input distributions; perturbation classes; operating ranges; deployment settings
Results	Interpretive uncertainty	The interpretations compatible with a reported output and the assumptions, inferential steps, or additional evidence needed to distinguish among them.	Observation versus inference; directness; causal interpretation; mechanism; target identity; practical meaning; alternative explanations; unresolved biological significance

Note: Uncertainty in dependencies is cross-cutting rather than a fifth branch or mid-level category and may generate linked records under multiple branches.