Math. Appl. 13(2)2024, 171–185

An Autonomous Scientific Discovery Platform Using AI Agents, Knowledge Graphs, and Reproducible Data Pipelines

Chaitanya Nagabhyru[1]

Abstract: *The volume of scientific literature and data is growing faster than any researcher can read or connect, and the same growth makes irreproducible results costly. Large language model agents can search, summarise and write code, but they invent citations, state unsupported claims with confidence and leave no auditable trail. This paper presents a platform for AI-assisted scientific discovery that addresses these weaknesses by design. A knowledge graph, built from literature and structured datasets with evidence spans and provenance attached to every edge, serves as the shared memory of a team of specialised agents. The agents retrieve evidence, propose testable hypotheses, rank them by a transparent score that combines plausibility, evidence support, novelty and expected information gain, and run computational experiments inside containerised, version-controlled pipelines whose every step is logged. An independent verification layer, separate from the generating agents, checks each claim against its cited source, recomputes reported numbers, re-executes the pipeline, and flags unsupported statements. Humans approve every consequential step. The paper specifies the architecture, the formal definitions of the main scores and metrics, and a pre-specified evaluation protocol with four measures: reproducibility, hypothesis-ranking quality, citation accuracy and researcher time saved. The protocol uses time-sliced historical evaluation, in which hypotheses are ranked from a graph frozen at a past date and tested against discoveries made afterwards, together with a controlled within-subject user study. This manuscript presents the design, formal definitions and protocol. It does not report experimental results: the result tables in Section VI are templates to be completed from actual runs, and no performance claim is made.*

Introduction

Science is limited by attention as much as by instruments. Biomedical literature adds more than a million articles a year, and materials and genomic repositories hold millions of records. The facts needed to form a good hypothesis are scattered across papers, supplementary tables and datasets in different formats and vocabularies, and a researcher must search, read, reconcile and compute, much of it repetitive [1].

The reliability of the published record is also a concern. Surveys and replication projects have found that a large share of results cannot be reproduced from the description in the paper, because code, data versions, parameters or environments are missing [2], [3]. A platform that speeds up analysis without improving traceability will only increase the volume of unverifiable results.

[1]12-1006, Rahmath Nagar 2nd Line, Chilakaluripet, Guntur Pin: 522616, Andhra Pradesh, India
Email:chaitanyanagabhyru@gmail.com

Keywords:AI for Science, Knowledge Graphs, AI Agents, Reproducibility, Hypothesis Generation, Provenance, Retrieval-Augmented Generation.

Language-model agents can read papers, call tools, write and run analysis code, and propose ideas in fluent prose. They also fabricate plausible references, attribute to a source a statement it does not make, vary between runs, and leave reasoning that another scientist cannot audit [4]. In science, a fluent but ungrounded answer is worse than none.

National research programmes have begun to treat AI-enabled science as infrastructure. The United States federal Genesis Mission, established by Executive Order 14363 in November 2025, emphasises the use of AI to accelerate scientific discovery and the integration of scientific datasets and computing resources across national laboratories and research institutions [5]. A platform in this setting must be trustworthy enough to support work that will be published, funded and built upon.

This paper proposes a platform designed around that requirement. Its central principle is that a generated statement is acceptable only if it is tied to evidence that can be checked, and a computed result is acceptable only if it can be reproduced from a recorded pipeline. The research questions are as follows. RQ1 Does grounding agents in an evidence-linked knowledge graph reduce unsupported claims and citation errors compared with retrieval alone and with an ungrounded model? RQ2 Do the platform's hypothesis scores rank candidate hypotheses in agreement with later discoveries better than baselines? RQ3 Do automatically tracked pipelines raise the rate at which analyses can be reproduced? RQ4 Does the platform reduce the time that researchers need for evidence synthesis and analysis tasks, without lowering the quality of the results? RQ5 Does an independent verification layer catch errors that the generating agents miss? The contributions are: an architecture that unites a provenance-rich knowledge graph, specialised agents, tracked pipelines and independent verification; formal definitions of the hypothesis score and of the verification checks; a human-oversight model with risk tiers; and a pre-specified evaluation protocol with success criteria and an explicit statement of data limitations. Section II reviews the background, Section III related work, Section IV the methodology, Section V the datasets and setup, Section VI the results structure, Section VII the discussion and Section VIII concludes.

LITERATURE REVIEW

A. Scientific Knowledge Graphs

A knowledge graph represents entities (genes, compounds, diseases, materials, methods, papers) as nodes and relations (inhibits, is associated with, has property, cites) as edges. In biomedicine, large resources such as curated pathway databases, drug–target collections and ontologies have been integrated into graphs that support queries across sources and have been used for drug repurposing and mechanism discovery [6]. Such graphs are incomplete, contain extraction errors, and often present edges without supporting evidence, so a user cannot tell an established fact from a single uncertain mention. Attaching provenance, meaning the source document, the text span, the method of extraction and a confidence, to every edge is therefore essential for scientific use.

B. Link Prediction and Hypothesis Generation

Predicting missing edges in a knowledge graph is a form of hypothesis generation: a predicted link between a drug and a disease is a candidate repurposing hypothesis. Translation-based embedding models score a triple by the distance between the translated head and the tail, and complex-valued models capture asymmetric relations [7]. Literature-based discovery, going back to Swanson's work on fish oil and Raynaud's syndrome, connects concepts in separate literatures through shared intermediates. Evaluation of such methods is commonly done by hiding a share of known edges, which tests recovery of old facts but not prediction of new ones. A stricter design freezes the graph at a past date and tests predictions against edges that appeared afterwards.

C. Retrieval-Augmented Generation and Citation Accuracy

Retrieval-augmented generation supplies a language model with retrieved passages so that its answer is based on documents and not only on its parameters. Retrieval combines sparse lexical matching with dense vector similarity. The gain in factuality is incomplete: models can ignore the retrieved text or cite a document that does not contain the claim. Evaluations of generative search systems have found that a considerable share of statements are not fully supported by the sources cited next to them [8]. Citation accuracy has two parts: whether the cited source exists, and whether it entails the statement. A verifier that tests entailment between the source passage and the claim is the natural check, and claim-verification datasets provide a standard to test such verifiers [9].

D. LLM Agents and Automated Science

Agents extend a language model with tools, memory and the ability to plan and iterate. Systems have been demonstrated that read literature, plan syntheses, control laboratory equipment and draft manuscripts [10]. Reported weaknesses are error accumulation over many steps, difficulty in judging which ideas are worth testing, and results that appear valid but come from flawed code or leaked information. For the physical sciences, the dual-use issue is serious: a system able to design and plan experiments in chemistry and biology could be misused, and so safeguards and human oversight are part of the requirements and not an optional feature [10].

E. Reproducibility and Provenance

Reproducibility means that an independent run of the same analysis on the same data gives the same results within a stated tolerance; replicability means that a new study reaches a consistent conclusion. Computational reproducibility fails when the data version, code version, parameters, random seeds or software environment is not recorded. The FAIR principles ask that data be findable, accessible, interoperable and reusable [11], and the W3C PROV model gives a standard vocabulary for recording which entities, activities and agents produced which results. Workflow engines, container images and experiment trackers provide the means; recording every agent-initiated run automatically makes the record complete by construction.

RELATED WORK

Knowledge-graph-based drug repurposing has been demonstrated by projects that integrated many biomedical resources and used embeddings to rank candidates, including retrospective recovery of known indications [6]. Lewis et al. [12] introduced retrieval-augmented generation; Liu et al. [8] found that only part of the statements of generative search engines were fully supported by their citations; and Wadden et al. [9] released a scientific claim-verification benchmark. A comparison of idea generation by language models and by expert researchers found model ideas rated more novel but less feasible [13], which is why ranking by feasibility and evidence, and not novelty alone, is needed. Baker's survey [2] reported that a majority of researchers had failed to reproduce another scientist's experiments, and studies of computational notebooks found that only a fraction run without error [3], [14].

The gaps are as follows. Knowledge graphs, agents, workflow tracking and verification are studied separately. Hypothesis generators are evaluated by recovery of old edges, not by prediction of later discoveries under a frozen graph. Citation accuracy is seldom reported for a full chain of claims, and time saved is rarely measured in controlled studies. The present work is built around these gaps.

METHODOLOGY

A. Platform Architecture

The platform has six layers: ingestion (open literature, datasets and licences), knowledge (entity and relation extraction into a provenance graph), agent (specialised agents reading and writing the graph), execution (sandboxed, containerised experiments under a workflow engine), verification (independent audit of claims and results) and oversight (proposals, approval gates and audit log). All layers write to a common provenance store, so any statement or number in a report can be traced back to its sources and data.

B. Knowledge Graph Construction

Documents are processed by a pipeline that recognises entities, links them to standard identifiers in ontologies and databases, extracts relations with their direction and polarity (for example "increases" or "does not affect"), and records for each relation the document, the text span, the section type (abstract, results, discussion, citation of earlier work), the method and a confidence. An edge is stored as a triple of head entity, relation and tail entity together with an evidence set. The support of an edge is aggregated over its evidence, with weights for source type, extraction confidence and study design, and a penalty for evidence that merely cites other work, so a statement repeated by many papers citing one original is not counted as many confirmations. Contradictory evidence is stored as separate edges with opposite polarity and is not discarded, and the confidence of an edge falls when strong contradiction exists.

Let an edge e have an evidence set $\mathcal{E}_e = \{(s_i, c_i, w_i)\}$, where $c_i \in [0,1]$ is the extraction confidence of evidence item i, $w_i \in [0,1]$ is its source weight and s_i its independence group (the original study to which the evidence traces).

The aggregated support combines evidence within a group by taking the maximum and across independent groups by a noisy-or:

$$\mathrm{Sup}(e) = 1 - \prod_{g \in \mathcal{G}_e} \left(1 - \max_{i \in g} c_i \, w_i\right) \qquad (1)$$

so that repeated citations of one study count once, while independent studies strengthen the support with diminishing returns.

C. Retrieval

Retrieval combines lexical and dense methods. The lexical score of a document d for a query q is the BM25 score:

$$\mathrm{BM25}(q, d) = \sum_{t \in q} \mathrm{IDF}(t) \, \frac{f(t, d)\,(k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b\,|d| / \overline{|d|}\right)} \qquad (2)$$

where $f(t, d)$ is the frequency of term t in d, $|d|$ the document length, $\overline{|d|}$ the mean length, and k_1 and b are parameters. The dense score is the cosine similarity of embeddings of the query and the passage:

$$s_{\mathrm{dense}}(q, p) = \frac{e_q^{\top} e_p}{\| e_q \| \; \| e_p \|} \qquad (3)$$

and the two rankings are merged by reciprocal rank fusion, in which a passage receives the sum over rankers r of $1/\left(k + \mathrm{rank}_r(p)\right)$ with $k = 60$. A graph-aware step then expands the query entities to their neighbours in the knowledge graph and retrieves the passages that are evidence for edges on short paths between the entities, so that the retrieved context contains the connecting evidence and not only the passages that match the query words.

D. Agent Roles

Five agents with defined permissions share the graph. The retrieval agent answers questions by gathering passages and graph paths and returns them with identifiers; it cannot write to the graph without evidence. The reading agent extracts structured findings from documents and proposes new edges with their evidence spans. The hypothesis agent proposes testable statements in a fixed template (an entity, a proposed relation, a predicted observable outcome and a stated way in which it could be refuted). The experiment agent translates an approved hypothesis into a computational plan: data to use, analysis steps, controls, success criteria set before running, and the resources required. The analysis agent runs the plan in the execution layer and reports results. Each agent produces output in a structured form, and every sentence of free text in a report is linked to evidence items or to run identifiers, or is marked as the agent's own inference. This convention is what makes automatic verification possible.

E. Hypothesis Scoring and Ranking

Candidate hypotheses are scored on five criteria. Plausibility is the score of a link predictor trained on the graph at the current date. For a triple (h, r, t),

a translation-based embedding model scores plausibility by

$$\phi(h, r, t) = -\| h + r - t\|_2 \qquad (4)$$

and is trained with a margin loss over true triples and corrupted ones:

$$\mathcal{L}_{KG} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t')\in\mathcal{T}'} \max\left(0,\ \gamma + \| h + r - t\|_2 - \| h' + r - t'\|_2\right) \qquad (5)$$

where γ is the margin, $\mathcal{T}$ the set of observed triples and $\mathcal{T}'$ the set of corrupted triples. The raw score is converted to a probability by calibration on held-out edges. Evidence support is the aggregate support of the existing edges on the paths that connect the entities of the hypothesis, computed by Eq. (1), so that a hypothesis resting on strong, independent evidence ranks above one that rests on a single mention. Novelty measures how little is already known about the direct relation: it is high when no direct edge exists and when the entities are far apart in the literature, and it is penalised to zero when the hypothesis restates an existing edge. Testability and cost is the estimated cost, in compute time or laboratory effort, of a decisive test, together with the availability of data and methods. Expected information gain measures how much a test is expected to reduce uncertainty about the hypothesis. With a prior probability p that the hypothesis is true, and a test whose true-positive rate is α and false-positive rate is β, the expected information gain is the mutual information between the hypothesis and the test outcome:

$$\mathrm{EIG} = H(p) - \left[P(+)\, H\big(P(\mathrm{true}\mid +)\big) + P(-)\, H\big(P(\mathrm{true}\mid -)\big)\right] \qquad (6)$$

where $H(x) = -x\log x - (1-x)\log(1-x)$ is the binary entropy, $P(+) = \alpha p + \beta(1-p)$, and the posterior probabilities follow from Bayes' rule. The posterior after a positive test, for example, is

$$P(\mathrm{true}\mid +) = \frac{\alpha\, p}{\alpha\, p + \beta\,(1-p)} \qquad (7)$$

The overall ranking score is a weighted combination, with weights set by the research team and stated in the report:

$$S(\mathcal{H}) = w_1\,\mathrm{Plaus} + w_2\,\mathrm{Sup} + w_3\,\mathrm{Nov} + w_4\,\frac{\mathrm{EIG}}{\mathrm{Cost}} \qquad (8)$$

with $\sum_k w_k = 1$ and the cost term normalised to the unit interval. Because the weights encode priorities that differ between a pharmaceutical company, which may prefer low-risk repurposing, and a national laboratory exploring new materials, they are exposed as parameters, and the ranking is reported together with the component scores so that the researcher can see why a hypothesis is placed where it is.

F. Reproducible Execution

Every experiment runs in a container whose image is identified by a content hash, with dependencies pinned. Inputs are referenced by the version

identifier of the dataset, with checksums, and the code by the commit identifier. Random seeds are fixed and recorded. The workflow engine records each step as an activity in the provenance store, with inputs, outputs, parameters, environment, start and end times, and resource use. Each record contains the hash of its predecessors, so the records form a chain in which a change to any earlier step changes all the later hashes:

$$\mathrm{id}_k = \mathrm{H}(\mathrm{id}_{k-1} \,\|\, \mathrm{input}_k \,\|\, \mathrm{code}_k \,\|\, \mathrm{params}_k \,\|\, \mathrm{env}_k) \qquad (9)$$

where H is a cryptographic hash and $\|$ denotes concatenation. This makes tampering or silent changes detectable, and it makes the claim "this figure was produced by this pipeline" checkable. For stochastic analyses, reproducibility is defined by agreement within a tolerance. A re-run reproduces a numerical result x if

$$\frac{|x_{\mathrm{rerun}} - x_{\mathrm{orig}}|}{\max(|x_{\mathrm{orig}}|, \epsilon)} \leq \tau \qquad (10)$$

for a tolerance τ that is stated in advance and depends on the type of result (for example 10^{-6} for deterministic computations, and a bound derived from the run-to-run standard deviation for stochastic ones). The reproducibility rate of a set of N analyses is the share for which all declared key results satisfy this condition.

G. Independent Verification

The verification layer is separate from the generating agents: it uses a different model instance, different prompts and, where possible, a different model family, and it has read-only access to sources. It performs four checks on every report. The citation existence check confirms that each cited identifier resolves to a real document and that the quoted span occurs in it. The entailment check tests whether the cited passage supports the sentence it is attached to. For a claim c and a passage p, the verifier outputs probabilities of the three labels supported, refuted and not enough information, and the claim is accepted only if the supported probability exceeds a threshold chosen on validation data:

$$\mathrm{Accept}(c, p) = \mathbb{1}[P(\mathrm{supported}|c, p) \geq \theta] \qquad (11)$$

The numerical recomputation check recomputes each reported number from the recorded data and code, independently of the analysis agent, and compares it by Eq. (10). The statistical audit checks the analysis against a checklist: whether the test used suits the data, whether multiple comparisons were corrected, whether the success criteria were fixed before the run, and whether information from the test data could have leaked into training. A claim that fails any check is flagged and removed from the summary, or shown with a visible warning, and it cannot be promoted into the knowledge graph. Where the same error could be made by both the generator and the verifier because they share training data, disagreement between two verifier models is treated as grounds for human review.

H. Human Oversight and Risk Tiers

Actions are classified into tiers. Tier 1 actions are read-only, such as search and summarisation, and may run automatically. Tier 2 actions are computational experiments within an approved plan, with bounded compute and no external side effects, and run automatically after the plan has been approved by a researcher. Tier 3 actions have real-world consequences or are costly, such as ordering materials, instructing laboratory automation or sending data outside the institution, and always need explicit approval by a named researcher. Tier 4 actions are prohibited for autonomous operation: any step related to the design or production of hazardous biological, chemical or radiological agents is blocked and reported, and hypotheses in these areas are screened by a dual-use filter and reviewed by a human before display. Every approval, edit and rejection is logged. The platform proposes and records; researchers decide.

I. Evaluation Measures

Four groups of measures follow the brief. Reproducibility is the reproducibility rate of Eq. (10) over a sample of analyses, reported separately for analyses re-run from the platform's own records and for analyses taken from published papers or public notebooks, and also the share of the pipeline steps that carry a complete provenance record. Hypothesis-ranking quality uses ranking metrics against later discoveries. If the relevant hypotheses are at ranks $r_1, \ldots, r_m$ in a list, the mean reciprocal rank is $\mathrm{MRR} = \frac{1}{m}\sum_j 1/r_j$ taken over queries, and the normalised discounted cumulative gain at depth K is

$$\mathrm{NDCG@K} = \frac{1}{\mathrm{IDCG@K}} \sum_{i=1}^{K} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)} \qquad (12)$$

where rel_i is the relevance of the item at position i and $\mathrm{IDCG@K}$ is the value for an ideal ordering. The precision of the top K and the area under the precision–recall curve are also reported, and agreement between the platform's ranking and the ranking of an expert panel is measured by Spearman's rank correlation

$$\rho = 1 - \frac{6\sum_i d_i^2}{n\left(n^2 - 1\right)} \qquad (13)$$

with d_i the difference in ranks of item i and n the number of items. Citation accuracy is measured at two levels. Citation precision is the share of cited sources that both exist and support the sentence to which they are attached, and citation recall is the share of sentences that need a citation and have a supporting one:

$$P_{\mathrm{cite}} = \frac{\#\{\text{citations that exist and support the claim}\}}{\#\{\text{citations given}\}}, \qquad R_{\mathrm{cite}} = \frac{\#\{\text{claims with a supporting citation}\}}{\#\{\text{claims needing a citation}\}} \qquad (14)$$

The assessment is made by trained annotators who read the source, and the agreement between annotators is measured by Cohen's kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad (15)$$

Researcher time saved is measured in a within-subject study: each participating researcher performs comparable tasks with and without the platform, in an order balanced across participants, and the relative saving is

$$\Delta T = \frac{\bar{T}_{ctrl} - \bar{T}_{plat}}{\bar{T}_{ctrl}} \qquad (16)$$

reported together with the quality of the output as judged by blinded graders, since time saved is of no value if quality falls. The effect on time is reported with a 95% confidence interval obtained by bootstrap resampling over participants, and the paired difference is tested with the Wilcoxon signed-rank test, because task times are skewed.

DATASETS AND EVALUATION SETUP

A. Data Sources

Table 1 lists the candidate sources by domain. Final selection depends on access and licensing. All sources are public or openly licensed, and the licence of each is recorded in the provenance store.

TABLE 1. Candidate Data Sources for the Knowledge Graph and Evaluation

Domain	Literature	Structured data	Use in evaluation
Biotechnology and pharmaceuticals	PubMed abstracts and open-access full texts	Drug–target and bioactivity databases; protein and pathway resources; gene-expression repositories; clinical-trial registry	Time-sliced drug repurposing and target-disease association; citation accuracy
Materials science	Open-access materials papers and preprints	Computed and experimental materials property databases	Time-sliced prediction of later-reported compounds or properties
Cross-domain	Open scholarly graph of papers, authors and citations	Ontologies and identifier systems	Entity linking; independence groups for evidence
Claim verification	Scientific claim–abstract pairs with labels	—	Calibration and testing of the entailment verifier
Reproducibility	Public computational notebooks and code deposited with papers	Archived datasets	Re-execution study

B. Time-Sliced Hypothesis Evaluation

The principal test of the ranking is retrospective and time-sliced. The knowledge graph is built using only information published before a cut-off date T_0. The platform then ranks candidate hypotheses of a stated type, for example "compound X has activity against target Y" or "material M has property P". The candidates are checked against the sources published between T_0 and a later date T_1, and a candidate counts as confirmed if it appears there in a curated database or a peer-reviewed paper. Negative examples are drawn in two ways, at random and by sampling pairs that are

close to the positives in graph structure, because performance on random negatives is known to be too optimistic. Several cut-off dates are used, to see whether performance is stable. Because a confirmed hypothesis is only a hypothesis found true by later work, and the absence of confirmation is not proof of falsity, the evaluation measures the ranking of later-confirmed items and never claims that other items are false.

C. Contamination Control

A language model trained on text published after T_0 may have seen the confirmations. Three safeguards are used: models whose training data end before T_0 are used where such models exist; for other models a canary test checks whether post-T_0 facts can be reproduced and the contamination risk is stated beside every result; and the knowledge-graph link predictor, trained only on the graph as it stood at T_0, is evaluated separately from the language-model components. Where contamination cannot be excluded, the claim is limited to components that can be assessed.

D. Reproducibility and Citation Studies

For the reproducibility study, a sample of published computational analyses with public code and data is drawn, stratified by domain, and each is re-run twice: once by the platform, which builds the environment and executes the analysis, and once by a human expert, to give a ceiling. The outcome is reproduction of declared key results within the tolerance of Eq. (10); failures are classified by cause. For the citation study, a sample of platform-generated reports and of baseline outputs is annotated claim by claim by trained annotators who do not know which system produced the text.

E. User Study Design

Researchers from biotechnology, pharmaceuticals and materials science perform matched tasks (evidence synthesis, data integration, hypothesis generation and analysis) with condition order counterbalanced and a time limit. Graders blinded to the condition score outputs against a rubric. The study is pre-registered, with outcome, analysis and exclusions fixed in advance, and the sample size is set by a power calculation for the paired comparison.

F. Baselines

The baselines separate the contributions of the components. They are keyword search and manual reading; a language model without retrieval; retrieval-augmented generation without the knowledge graph; the knowledge graph with link prediction and no agents; the full platform without the verification layer; and the full platform without tracked pipelines (agents run code in an ordinary, unrecorded environment). For ranking, a popularity baseline (rank by the number of papers on the entities), a nearest-neighbour baseline in embedding space, and an expert panel are also included.

.

RESULTS: REPORTING STRUCTURE AND SUCCESS CRITERIA

Important note. No experiments have been run for this version of the manuscript, and **no measured result is reported here.** The tables below define how results will be reported once the experiments in Section V are completed. All numerical cells are intentionally empty. They must be filled only with values from the pre-specified protocol, and the abstract, discussion and conclusion must be revised to match them. A result that does not meet the criteria below should be reported as it is.

A. Pre-Specified Success Criteria

The criteria are fixed in advance. For RQ1, the proportion of claims that are fully supported by their citations in the full platform should exceed that of retrieval-augmented generation without the graph, with a confidence interval of the difference that excludes zero, and citation precision should be reported with the inter-annotator kappa. For RQ2, the NDCG at the chosen depth and the MRR of the platform's ranking should exceed those of the popularity and the embedding-neighbour baselines at every cut-off date tested, under both random and hard negatives. For RQ3, the reproducibility rate for analyses run through the platform should exceed that of the unrecorded-agent baseline, and the failures should be classified. For RQ4, the time saved should have a confidence interval excluding zero while the quality score is not lower than in the control condition by more than a pre-set margin. For RQ5, the verification layer should catch a stated share of seeded errors (errors inserted deliberately into reports in a test set) and the false-flag rate should be reported. If a criterion is not met, the corresponding claim is withdrawn.

B. Citation Accuracy and Grounding

Table 2 is the template for the citation study.

System	Citation precision	Citation recall	Claims fully supported (%)	Fabricated citations (%)	Annotator kappa
Language model without retrieval	[ ]	[ ]	[ ]	[ ]	[ ]
Retrieval-augmented generation, no graph	[ ]	[ ]	[ ]	[ ]	[ ]
Platform without verification layer	[ ]	[ ]	[ ]	[ ]	[ ]
Platform (full)	[ ]	[ ]	[ ]	[ ]	[ ]

TABLE 2. Citation Accuracy and Claim Support (to be completed)

C. Hypothesis-Ranking Quality

Table 3 is the template for the time-sliced ranking results, with one block for each domain and cut-off date.

TABLE 3. Hypothesis Ranking under Time-Sliced Evaluation (to be completed; state cut-off dates and negative sampling)

Method	Negatives	MRR	NDCG@20	Precision@20	AUPRC	Spearman with expert panel
Popularity baseline	Random / hard	[ ]	[ ]	[ ]	[ ]	[ ]
Embedding nearest neighbour	Random / hard	[ ]	[ ]	[ ]	[ ]	[ ]
Link predictor only	Random / hard	[ ]	[ ]	[ ]	[ ]	[ ]
Platform score, Eq. (8)	Random / hard	[ ]	[ ]	[ ]	[ ]	[ ]
Expert panel	—	[ ]	[ ]	[ ]	—	—

D. Reproducibility

Table 4 is the template for the re-execution study.

TABLE 4. Reproducibility of Computational Analyses (to be completed)

Condition	Analyses (N)	Reproduced within tolerance (%)	Complete provenance record (%)	Main failure causes
Agent without tracking	[ ]	[ ]	[ ]	[ ]
Platform with tracked pipelines	[ ]	[ ]	[ ]	[ ]
Human expert (ceiling)	[ ]	[ ]	—	[ ]

E. Researcher Time Saved

Table 5 is the template for the user study. Time and quality must be reported together.

TABLE 5. Time and Quality in the Within-Subject Study (to be completed)

Task	Median time without platform (min)	Median time with platform (min)	Relative time saved, 95% CI	Quality score without	Quality score with
Evidence synthesis	[ ]	[ ]	[ ]	[ ]	[ ]
Data integration	[ ]	[ ]	[ ]	[ ]	[ ]
Hypothesis generation and analysis	[ ]	[ ]	[ ]	[ ]	[ ]

F. Verification Layer Effectiveness

Table 6 is the template for the test of the verifier on seeded errors.

TABLE 6. Detection of Seeded Errors by the Verification Layer (to be completed)

Error type	Seeded errors (N)	Caught by verifier (%)	Missed (%)	False flags per 100 correct claims
Fabricated citation	[ ]	[ ]	[ ]	[ ]
Real source, unsupported claim	[ ]	[ ]	[ ]	[ ]
Altered number	[ ]	[ ]	[ ]	[ ]
Statistical misuse	[ ]	[ ]	[ ]	[ ]
Data leakage	[ ]	[ ]	[ ]	[ ]

G. Ablations

Each component (graph grounding, evidence-support term, information-gain term, independence correction of Eq. (1), verification layer, pipeline tracking) will be removed in turn, and the effect on the four primary measures, with compute cost and latency, will be reported.

DISCUSSION

Three principles stand out. First, evidence should be attached to every statement: linking each sentence to evidence, or marking it as inference, with an independent check, turns a vague worry about hallucination into a measurable rate. Second, a result is not a result unless it can be re-run; automatic recording removes a failure that depends on researcher discipline, and the hash chain makes the record verifiable. Third, evaluation of discovery must look forward: recovering hidden edges shows a model can reconstruct what is known, whereas the time-sliced design asks whether it ranks what was later found.

The criteria of Eq. (8) pull in different directions: plausible hypotheses are often unsurprising, and novel ones poorly supported, which is why the platform exposes the weights and components. The independence correction in Eq. (1) addresses a known problem: a claim repeated by many papers that cite the same early study looks well established but has been tested once.

Human oversight keeps the speed of automation where it is harmless (reading, searching, bounded computation) and brings in a person where consequences are real. Autonomous action in dual-use areas is prohibited, and a named researcher approves anything beyond computation.

Programmes that aim to connect scientific datasets and computing resources need components that make heterogeneous data usable by AI while keeping results traceable; the provenance store, identifiers and licence records are intended as such a component. The paper makes no claim of national benefit, which would need evidence from deployment at several institutions over time.

The data limitations bound any result. Publication bias means the literature over-represents positive findings, so a graph built from it overstates support and contains few reliable negatives. Automatic extraction errors propagate into the graph and should be measured on an annotated sample. The literature is not independent, which independence groups only partly correct. Later publications used as truth are themselves incomplete, so low

scores may reflect lack of testing and not a wrong hypothesis. Language-model contamination (Section V-C) may inflate results. Reproducible studies are not representative of all science, licensing limits redistribution, verification roughly doubles compute cost, and results from one field cannot be assumed to transfer to another.

The platform works with what has been written down and measured, so it is best at connecting and prioritising and weaker at creative leaps that rest on tacit knowledge. A computational test is no substitute for a laboratory experiment where one is needed, and many automated analyses raise the danger of false discovery through multiple testing, which the statistical audit and pre-set criteria reduce but cannot remove.

CONCLUSION

This paper proposed an autonomous scientific discovery platform in which a provenance-rich knowledge graph, specialised evidence-grounded agents, automatically tracked pipelines and an independent verification layer work together under human oversight. It defined the main scores and checks formally, including evidence support with independence correction, hypothesis ranking with expected information gain, hash-chained provenance, tolerance-based reproducibility and entailment-based citation checks, and it set out a pre-specified evaluation protocol using time-sliced historical tests, contamination controls, a re-execution study, a citation audit and a within-subject user study.

The manuscript does not yet contain experimental results. The next steps are to build the knowledge graph for at least two domains from the sources of Table 1, to run the experiments under the stated protocol, to fill the tables of Section VI with measured values, to revise the abstract, discussion and conclusion accordingly, and to report the limitations observed. If the results meet the criteria, the contribution will be evidence that grounding and verification reduce unsupported claims, that tracked pipelines raise reproducibility, and that researchers save time without losing quality. If they do not, the analysis will show which component fails and why, which is also useful for those who plan to build or adopt such systems.

References

[1] L. Bornmann, R. Haunschild, and R. Mutz, “Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases,” *Humanities and Social Sciences Communications*, vol. 8, p. 224, 2021.

[2] M. Baker, “1,500 scientists lift the lid on reproducibility,” *Nature*, vol. 533, no. 7604, pp. 452–454, 2016.

[3] J. F. Pimentel, L. Murta, V. Braganholo, and J. Freire, “A large-scale study about quality and reproducibility of Jupyter notebooks,” in *Proc. IEEE/ACM International Conference on Mining Software Repositories (MSR)*, 2019, pp. 507–517.

[4] Z. Ji et al., “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[5] Executive Office of the President, “Launching the Genesis Mission,” Executive Order 14363, *Federal Register*, vol. 90, no. 227, Nov. 28, 2025.

[6] D. S. Himmelstein et al., “Systematic integration of biomedical knowledge prioritizes drugs for repurposing,” *eLife*, vol. 6, p. e26726, 2017.

[7] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 26, 2013, pp. 2787–2795.
[8] N. F. Liu, T. Zhang, and P. Liang, "Evaluating verifiability in generative search engines," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025.
[9] D. Wadden et al., "Fact or fiction: Verifying scientific claims," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7534–7550.
[10] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous chemical research with large language models," *Nature*, vol. 624, no. 7992, pp. 570–578, 2023.
[11] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016.
[12] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.
[13] C. Si, D. Yang, and T. Hashimoto, "Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers," in *Proc. International Conference on Learning Representations (ICLR)*, 2025.
[14] S. Samuel and D. Mietchen, "Computational reproducibility of Jupyter notebooks from biomedical publications," *GigaScience*, vol. 13, p. giad113, 2024.
[15] L. Moreau and P. Missier, Eds., "PROV-DM: The PROV Data Model," W3C Recommendation, 2013.