ORIGINAL RESEARCH **Open access**

Conceptual Foundations for Scientific Audit Trails in Materials AI Systems

Luis Herrera[1*], Daniela Rojas[1], Andres Castro[2]

Abstract

Materials AI systems have become indispensable for accelerating discovery in solid-state materials, energy storage compounds, and functional alloys. Yet, they operate without systematic mechanisms to trace the full chain of data provenance, model decisions, and reasoning steps that produce any given prediction or recommendation. The absence of scientific audit trails means that when a novel perovskite composition is proposed, or a predicted bandgap deviates from experiment, researchers cannot reliably reconstruct the exact sequence of data transformations, hyperparameter choices, feature selections, or failure modes that led to the outcome, undermining reproducibility, error diagnosis, and collective scientific progress. This paper proposes a comprehensive blueprint for scientific audit trails tailored specifically to the unique requirements of materials AI workflows, where heterogeneous data sources, multiscale simulations, and iterative human–machine interactions demand far more than generic machine-learning logging. The blueprint defines a machine-readable yet human-accessible record that captures every relevant element of a materials discovery pipeline. Its seven core components—ranging from granular data provenance to detailed failure logs and environmental context—provide the structural foundation for traceability. Four operational principles ensure that capture is automatic, standardized, immutable, and accessible. At the same time, five success criteria establish objective benchmarks for completeness, traceability speed, reproducibility power, error localization precision, and acceptable computational overhead. Finally, a five-phase implementation path offers the materials AI community a practical route from standards development to journal-mandated adoption and AI-assisted analysis. By closing this critical gap, the proposed scientific audit trails will transform materials AI from opaque black-box engines into transparent, accountable scientific instruments.

Keywords Materials AI, Scientific audit trails, Data provenance, Reproducibility in computational materials science, Machine learning transparency, Scientific workflow accountability

*Correspondence:
Luis Herrera
luis.herrera@gmail.com
[1] Department of AI Materials Systems, Pontifical Catholic University of Chile, Santiago, Chile
[2] Department of Computational Materials Engineering, University of Concepcion, Concepcion, Chile

Introduction

Materials AI systems are increasingly complex and autonomous. When a result is produced—whether a new candidate for solid-state electrolyte or an optimized synthesis route for a high-entropy alloy—can we trace how it was generated? What data, models, and decisions led to it? Currently, no. This paper proposes a blueprint for scientific audit trails in materials AI.

The rapid adoption of machine learning across molecular and materials science has delivered remarkable advances in property prediction, inverse design, and autonomous experimentation. Yet this progress has outpaced the development of supporting infrastructure for scientific accountability. As Butler and colleagues observed, machine learning for molecular and materials science relies on vast, heterogeneous datasets and iterative modeling cycles whose internal logic remains largely invisible once a prediction is published [1-4]. Similarly, Schmidt *et al.* [5]

documented how recent advances in solid-state materials science depend on ever-larger training sets and sophisticated neural architectures, but rarely provide the provenance metadata necessary to verify or extend those results. Zunger's foundational discussion of inverse design further highlights that target-driven discovery workflows involve countless branching decisions whose rationale is seldom recorded [6].

The consequences of this invisibility are profound. Reproducibility crises that have afflicted other data-intensive fields are now emerging in materials informatics. When a predicted material fails experimental validation, the community cannot systematically determine whether the failure originated in data preprocessing, model architecture, hyperparameter tuning, or an overlooked domain shift between training and target distributions. Montoya *et al.* [7], in their review of autonomous materials research, explicitly note the growing need for mechanisms that make such decision chains transparent if the field is to move beyond isolated successes toward reliable, cumulative knowledge.

Without audit trails, materials AI systems accumulate what Sculley *et al.* famously termed "hidden technical debt"—undocumented dependencies, ad-hoc preprocessing steps, and unreproducible random seeds that silently erode trust in published results [2]. Even when code repositories and raw datasets are shared, the precise lineage linking raw inputs to final predictions remains broken. This paper, therefore, articulates a conceptual foundation and practical blueprint for scientific audit trails that are purpose-built for materials AI. Rather than retrofitting generic provenance tools, the proposal integrates domain-specific requirements such as multiscale data fusion, simulation–experiment feedback loops, and the need to track both automated and human expert interventions.

By making every material's AI workflow auditable, the blueprint addresses not only technical reproducibility but also deeper epistemological concerns: what counts as justified knowledge in data-driven materials discovery? The following sections first diagnose the auditability problem in detail, survey the fragmented current state of practice, present the core proposal, and then elaborate on the seven components that constitute a complete scientific audit trail. Subsequent parts of this work (to be developed in continuation) will specify operational principles, success criteria, implementation pathways, and responses to anticipated objections. The ultimate aim is to establish scientific audit trails as a new standard that elevates materials AI to the level of rigor long expected in experimental and computational materials science.

The Auditability Problem

Audit trails matter for four interlocking reasons that are especially acute in materials AI. First, reproducibility: without audit trails, results cannot be reliably reproduced because the exact sequence of data filtering, feature engineering, model selection, and training dynamics is lost. Second, error detection: when results are wrong, audit trails enable root cause analysis by exposing which specific transformation or decision introduced the discrepancy. Third, trust: auditable systems are more trustworthy because independent researchers can verify the integrity of the reasoning chain rather than accepting outputs on faith. Fourth, scientific progress: audit trails enable learning from past decisions, turning every failed or successful experiment into structured knowledge that future models can leverage [8-11].

Concrete examples illustrate the severity of the gap. Consider a typical high-throughput screening study that identifies a promising cathode material. If the predicted voltage deviates by 0.4 V from experiment, current practice offers no systematic way to determine whether the error arose from an outdated crystal-structure database entry, an unrecorded data-augmentation step, or an early-stopping criterion that masked overfitting. Persaud *et al.* [9] documented precisely such reproducibility failures in materials informatics, showing how even well-intentioned publications often omit critical lineage details. Similarly, DeCost *et al.* [11] described how scientific AI platforms for materials discovery accumulate opaque technical debt that makes systematic improvement nearly impossible without exhaustive manual reconstruction.

The problem is compounded by the interdisciplinary nature of materials AI. Data may originate from density-functional theory calculations, experimental synthesis logs, or literature-extracted property tables—each with its own versioning and preprocessing history. When these streams are fused inside a graph neural network or a transformer-based inverse-design model, the absence of unified provenance tracking renders the fusion process epistemically opaque. Wang *et al.* [10], in their guide to best practices for materials scientists, explicitly warned that without standardized documentation of every modeling

choice, machine-learning claims risk becoming non-falsifiable.

Furthermore, automated laboratories and self-driving platforms intensify the stakes. As Montoya *et al.* observed, the move toward closed-loop autonomous research removes the human "notebook" that once captured informal rationales; yet no equivalent digital mechanism has replaced it [7]. Failures in such systems—whether a robotic synthesis that produces an unexpected phase or a model that hallucinates a metastable structure—cannot be diagnosed without exhaustive post-hoc detective work. The auditability problem is therefore not merely technical but foundational to the epistemology of data-driven materials science: without traceable reasoning chains, claims about "discovered" materials lack the justificatory structure required for cumulative scientific knowledge.

Current State of Audit Trails

Current practices in materials AI remain minimal and fragmented. Version control for code via Git is now commonplace, yet it captures only software artifacts and ignores the far more voluminous and mutable data lineage and decision layers. Ad-hoc lab notebooks or supplementary information files record human observations but are neither machine-readable nor linked to specific model outputs. The majority of materials AI papers still publish only final models and performance tables, omitting the provenance metadata essential for verification [12-28].

A survey of recent literature reveals pockets of progress that have not yet coalesced into a coherent standard. Adorf *et al.* introduced the Signac framework to manage data and workflows in computational materials science, providing some level of reproducibility through structured project organization [14]. Statt *et al.* [12] developed the Materials Provenance Store, an event-sourced system designed explicitly to track millions of materials experiments and analyses. Their subsequent ESAMP architecture demonstrated how event sourcing could be applied to accelerated discovery pipelines [13]. Pérez *et al.* [15] proposed Pinax, a dedicated provenance management system for materials data science. These efforts illustrate that domain-aware provenance tools are technically feasible.

Nevertheless, adoption remains limited. Most frameworks address only narrow slices—code versioning, experimental metadata, or simulation inputs—rather than the end-to-end audit trail required for full scientific accountability. W3C PROV standards, while foundational for general provenance modeling, have seen only sporadic uptake in materials AI despite their maturity [1]. Machine-learning provenance research outside materials science has produced valuable concepts, yet these have not been adapted to the heterogeneous, multiscale, and human-in-the-loop nature of materials workflows.

DeCost and Hattrick-Simpers have repeatedly emphasized that trustworthy AI for materials discovery demands more than performance metrics; it requires auditable reasoning [26]. Yet even recent reviews of AI-supported materials informatics devote scant attention to systematic audit mechanisms [21, 28]. The result is a patchwork: some groups employ internal logging, others rely on supplementary tables, and the community as a whole lacks a shared, enforceable standard. Consequently, the current state is characterized by islands of ad-hoc solutions floating in a sea of undocumented decision-making.

Table 1 clarifies why existing documentation practices fall short of full scientific accountability and why materials AI requires a domain-specific audit trail rather than a simple extension of code sharing or generic provenance tools.

Table 1. Scientific audit trails compared with existing documentation and provenance mechanisms in materials AI

Dimension	Git/code version control	Lab notebooks/supplementary files	
Primary object recorded	Software artifacts and code changes	Narrative observations, selected methods, and final outputs	
Granularity of capture	File- and commit-level	Human-authored, often selective and retrospective	I

Coverage of data lineage	Partial; usually external to the repository	Fragmentary; often summarized rather than executable
Coverage of model decision logic	Minimal	Inconsistent and non-standardized
Capture of failed runs/negative results	Rare	Rare and usually informal
Human–AI interaction visibility	Minimal	Present but weakly linked to outputs
Reproducibility power	Supports code recovery, not full workflow re-execution	Supports narrative reconstruction only
Suitability for heterogeneous materials workflows	Low	Moderate for local context only
Trust and accountability value	Limited	Limited to the reviewer's interpretation
Main limitation	Captures code without a scientific reasoning chain	Non-machine-readable and difficult to verify

The Proposal

This paper offers a blueprint for scientific audit trails in materials AI that closes the gap identified above. The proposal rests on a precise, operational definition that integrates insights from existing provenance frameworks while extending them to materials-specific requirements.

A scientific audit trail in materials AI is a complete, machine-readable, and human-accessible record of all data sources, model decisions, hyperparameters, evaluation steps, and failure outcomes within any materials AI workflow, structured so that every output—prediction, recommendation, or optimized material—can be traced bidirectionally to its originating inputs, transformations, and rationales.

Figure 1 presents the proposed scientific audit trail architecture as a hierarchical provenance system linking heterogeneous materials inputs to seven integrated audit components, operational principles, implementation mechanisms, and downstream scientific outcomes.

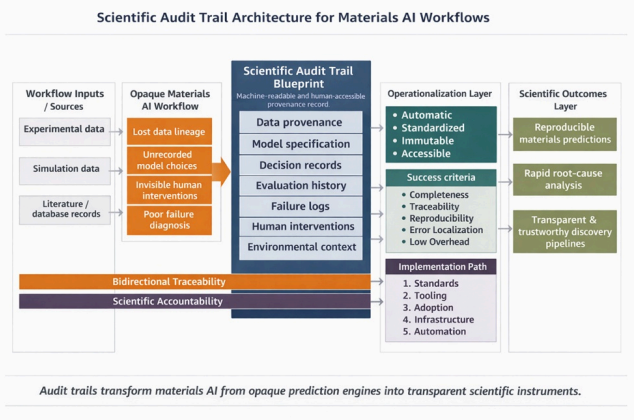

Figure 1. The proposed scientific audit trail architecture is a hierarchical provenance system linking heterogeneous materials inputs to seven integrated audit components, operational principles, implementation mechanisms, and downstream scientific outcomes.

The goal is unambiguous: every result should be traceable to its origins. Unlike generic logging or version control, the proposed audit trail treats the entire discovery pipeline as a single provenance graph whose nodes and edges carry both technical metadata and semantic justifications. By design, it builds upon but transcends earlier efforts such as the PROV data model [1] and the Materials Provenance Store [12] by incorporating explicit representations of materials-domain concepts (crystal structures, synthesis conditions, property hierarchies) and by enforcing automatic, immutable capture at every workflow step.

Components of the Audit Trail

The blueprint organizes the audit trail into seven interlocking components, each addressing a distinct yet interdependent layer of the materials AI process.

Data provenance records the complete history of every dataset used, including sources, version identifiers, preprocessing scripts, train/validation/test splits, feature transformations, and any domain-specific augmentations such as structure relaxation or property normalization. This component ensures that any downstream model can be re-executed against the exact data state that produced it.

Model specification captures the full architectural definition, weight initialization method, training algorithm, hyperparameter values (including learning-rate schedules and regularization terms), and random seeds. For materials AI, this includes explicit encoding of physics-informed constraints or descriptor choices such as elemental embeddings or graph convolution depths.

Decision records log the rationale behind every non-default choice—why a particular graph neural network architecture was selected over a transformer, why certain features were pruned, or why training was halted at a given epoch. These records link human or algorithmic justifications directly to outcomes.

Evaluation history archives all intermediate and final evaluation results, including cross-validation folds, hold-out sets, and any materials-specific metrics such as formation-energy errors or synthesizability scores. Both successful and unsuccessful evaluations are retained to support meta-learning.

Failure logs systematically record every instance in which a prediction violated physical constraints, a synthesis route proved infeasible, or an optimization step produced invalid outputs, together with the contextual parameters present at failure.

Human interventions document any manual overrides, expert corrections, or interactive steering events, preserving the hybrid human–AI nature of many materials discovery campaigns.

Environmental context stores software versions, hardware configurations, container images, and random-state seeds so that the entire computational environment can be reconstituted.

Conceptually, the audit trail architecture can be visualized as a directed acyclic provenance graph in which data provenance nodes feed forward into model specification and decision record nodes; these in turn connect to parallel branches for evaluation history and failure logs, all converging at output nodes while human interventions and environmental context form cross-cutting annotation layers. Edges carry timestamps and semantic justification tags, enabling both forward reconstruction and backward traceability in a single unified structure. Together, these seven components ensure that no aspect of a material's AI workflow remains opaque.

Table 2 consolidates the internal design logic of the blueprint by showing how each audit component contributes a distinct epistemic function while also mapping onto operational principles and measurable success criteria.

Table 2. Design logic matrix linking audit trail components to epistemic functions, diagnostic value, and evaluation benchmarks

Audit trail component	Immediate object captured	Epistemic function in materials AI	The disc helps

Data provenance	Dataset source, version, preprocessing, splits, and feature transformations	Establishes evidentiary lineage and guards against silent data drift	Errors corrup ou data leak undoc prepr
Model specification	Architecture, initialization, optimizer, hyperparameters, random seeds, and physics-informed constraints	Stabilizes the inferential identity of the model	Perf varian by misma insta h conf ch
Decision records	Rationales for non-default choices, feature pruning, stopping decisions, and architecture selection	Makes reasoning visible rather than merely executable	Misju cau unjusti sele opaq ch
Evaluation history	Intermediate and final metrics, folds, hold-out results, and materials-specific performance indicators	Converts model assessment into cumulative comparative evidence	False arisi selectiv or u validati
Failure logs	Constraint violations, infeasible synthesis routes, invalid outputs, and anomalous optimization states	Turns negative outcomes into structured scientific knowledge	Root failed p phy imp recomn or unst be
Human interventions	Manual overrides, expert corrections, and interactive steering actions	Preserves hybrid agency in human–AI discovery systems	Discr intro undoc expert or ad h
Environmental context	Software versions, hardware, containers, and	Reconstitutes the computational conditions of	Repr failures enviro
	execution environment	knowledge production	or de ch

Operational Principles

The blueprint for scientific audit trails in materials AI rests on four foundational operational principles that translate the conceptual architecture into a practical, deployable system capable of supporting the full spectrum of materials discovery workflows. These principles are deliberately engineered to overcome the limitations observed in current ad-hoc provenance efforts while ensuring seamless integration with the heterogeneous, multiscale, and human-in-the-loop nature of materials informatics.

Automatic capture

Audit information must be captured automatically at the point of execution rather than through retrospective manual annotation. In materials AI pipelines—where density-functional theory relaxations, graph neural network training runs, and robotic synthesis commands can number in the thousands per campaign—manual logging is not merely inefficient but epistemically unreliable [12]. Automatic capture embeds lightweight instrumentation directly into workflow engines such as Signac [14] or the event-sourced ESAMP architecture [13], intercepting every data ingestion, transformation, model instantiation, hyperparameter update, and evaluation step without researcher intervention. For example, when a researcher launches an inverse-design optimization for high-entropy alloys, the system would silently record the precise crystal-structure database snapshot, the chosen descriptor set, the optimizer settings, and every intermediate energy landscape evaluation. This principle directly mitigates the hidden technical debt identified by Huyen [2], where undocumented preprocessing or random-seed choices silently undermine downstream reproducibility [9]. By making capture intrinsic to the computational environment, automatic logging ensures that even fleeting decision branches—such as an early-stopping event triggered by a sudden spike in formation-energy variance—are preserved for later forensic analysis, thereby transforming materials AI from a collection of opaque scripts into a continuously self-documenting scientific instrument.

Standardized format

All audit records must adhere to a single, extensible, machine-readable standard that builds upon but extends

established provenance models to accommodate materials-specific semantics. The W3C PROV data model [1] provides a robust starting point for representing entities, activities, and agents. Yet, it lacks native vocabulary for crystal-structure versioning, property-hierarchy annotations, or synthesis-condition ontologies. The proposed format, therefore, adopts a JSON-LD serialization layered over PROV-O, augmented with domain ontologies drawn from the Materials Provenance Store [12] and Pinax [15]. This standardization enables interoperability across laboratories: a perovskite bandgap prediction generated at one institution can be ingested and re-executed at another without custom parsing scripts. The format further mandates semantic tags for every node—linking, for instance, a "feature-pruning" activity to the specific chemical-element embedding that was removed—thereby supporting automated reasoning over the audit graph itself. Without such standardization, even well-intentioned provenance stores remain isolated silos; with it, the materials AI community gains the shared semantic substrate necessary for cumulative knowledge building [11].

Immutable record

Once written, audit entries must be cryptographically immutable and append-only, preventing retroactive alteration or selective omission. Immutability is essential for scientific trust, particularly when audit trails are later scrutinized during peer review, replication studies, or regulatory oversight of AI-assisted material claims. Drawing on event-sourcing principles already demonstrated in the Materials Provenance Store [12], each new event receives a hash-linked timestamp and is stored in a tamper-evident ledger. Should a researcher later discover a flawed preprocessing step, the original record remains intact while a corrective "amendment" event is appended with explicit justification; the history is never erased, only contextualized. This design echoes the accountability requirements articulated by DeCost and Hattrick-Simpers for trustworthy AI platforms [26] and directly counters the reproducibility failures documented across materials informatics [9]. In practice, immutability also facilitates long-term archival: a 2025-era audit trail for a solid-state electrolyte discovery remains verifiable in 2035 even if the original computing environment has been decommissioned.

Accessible

The audit trail must simultaneously support machine-scale queries and human-scale comprehension through dual representational layers. A full provenance graph encoded in the standardized format serves algorithmic consumers—replication scripts, meta-learning agents, or journal compliance checkers—while an automatically generated human-readable summary, rendered as a navigable HTML dashboard or PDF appendix, presents the workflow in layered detail: high-level narrative for principal investigators, drill-down tables for domain experts, and interactive graph visualizations for students. Accessibility ensures that audit trails do not become yet another bureaucratic burden but instead serve as living scientific documents. For instance, a materials scientist examining a failed high-throughput screen can instantly trace the failure to a specific data-augmentation routine without parsing raw JSON. This principle aligns with broader calls for transparency in complex computational systems [25]. It guarantees that the epistemic value of the audit trail is realized by the full spectrum of users, from individual researchers to the community at large [28, 29].

Collectively, these four principles create a self-reinforcing system in which capture is effortless, format is universal, records are trustworthy, and outputs are usable—precisely the conditions required to elevate materials AI to the auditability standard long expected in experimental and theoretical branches of the discipline.

Success Criteria

To evaluate whether any given implementation of the proposed scientific audit trail meets the blueprint's ambitions, five objective success criteria are defined. Each criterion is formulated to be measurable, materials-AI-specific, and aligned with the core goals of reproducibility, traceability, and scientific accountability.

Completeness

An adequate audit trail must capture every relevant decision and data transformation within the workflow. Completeness is verified by automated schema validation against the seven-component ontology. If any node in the provenance graph lacks mandatory attributes (e.g., a preprocessing script hash for data provenance or a rationale tag for decision records), the trail fails. This criterion directly addresses the fragmentation currently observed in most published materials AI studies [21].

Traceability

From any final output—such as a predicted lithium-ion conductivity value or an optimized synthesis protocol—any researcher must be able to reconstruct the originating inputs, transformations, and rationales in no more than five navigational steps within the audit graph. Traceability ensures that the cognitive load of verification remains low even for complex multiscale models, distinguishing the proposal from ad-hoc logging that often requires dozens of manual cross-references.

Reproducibility

A third party possessing only the audit trail and the original raw data sources must be able to re-execute the exact workflow and obtain bitwise-identical results (or statistically indistinguishable results when stochastic elements are involved). Reproducibility is tested through blinded replication challenges and extends beyond code sharing to encompass the full decision provenance, thereby closing the gap highlighted by Persaud *et al.* [9] in materials informatics.

Error Localization

When an experimental validation deviates from a material's AI prediction, the audit trail must enable localization of the discrepancy to a specific component—data provenance, model specification, or human intervention—within minutes rather than days. This criterion operationalizes root-cause analysis and draws on failure-logging practices already prototyped in event-sourced systems [13].

Overhead

The computational and storage overhead introduced by the audit trail must not exceed 20 % of the baseline workflow cost, measured across representative benchmarks spanning DFT calculations, graph neural network training, and autonomous laboratory runs. Overhead is quantified through profiling tools integrated into the automatic-capture layer, ensuring that the benefits of auditability do not come at the expense of scientific throughput—a concern frequently raised when discussing provenance frameworks in resource-constrained materials research environments [11].

Together, these criteria provide a rigorous, falsifiable benchmark set against which future implementations can be judged, ensuring that scientific audit trails deliver substantive improvements rather than superficial compliance.

Implementation Path

Translating the blueprint into widespread practice requires a deliberate, phased rollout that builds community consensus, technical infrastructure, and cultural acceptance. The five-phase path outlined below is designed to minimize disruption while maximizing momentum.

Community standards

The materials AI community, through workshops co-organized by societies such as the Materials Research Society and the AI4Science community, must converge on a minimal viable audit schema extending the Materials Provenance Store [12] and Pinax [15]. This phase includes public review of the seven-component ontology and the JSON-LD serialization, culminating in a formal specification document published in an open-access venue.

Tooling

Reference implementations of automatic-capture plugins must be developed for the dominant materials AI frameworks—PyTorch-based graph networks, JAX-accelerated optimizers, and workflow managers such as signac [14]. These plugins will be released as open-source libraries with zero-configuration defaults, allowing laboratories to adopt the audit trail with a single import statement.

Adoption

Leading journals in the field—including npj Computational Materials, Machine Learning: Science and Technology, and Digital Discovery—will revise their submission guidelines to require a machine-readable audit trail as a condition of publication, beginning with a two-year grace period for voluntary compliance. This policy mirrors earlier mandates for data availability and code deposition [10].

Infrastructure

Centralized, FAIR-compliant audit repositories will be established, building on the existing Materials Provenance Store infrastructure [12] and offering persistent DOIs for audit trails analogous to dataset DOIs. Cloud-hosted query

interfaces will enable cross-laboratory meta-analyses without exposing proprietary intermediate data.

Automation

Advanced AI agents will be trained to perform meta-audits—automatically surfacing patterns across thousands of audit trails, suggesting workflow improvements, and flagging potential reproducibility risks. This final phase closes the loop, turning audit data into an active driver of scientific progress [27].

Each phase is gated by measurable milestones, ensuring steady, evidence-based advancement toward universal auditability.

Objections and Replies

Anticipated resistance to the adoption of scientific audit trails can be addressed through conceptual clarity and empirical precedent from existing provenance systems.

“Audit trails add too much overhead.” Critics argue that automatic logging will slow high-throughput campaigns and consume excessive storage. The reply is twofold. First, the overhead criterion (≤ 20%) is enforced by design through efficient event sourcing [13] and selective compression of immutable records; profiling of the Materials Provenance Store already demonstrates sub-5% impact on typical DFT workflows [12]. Second, the long-term savings in error diagnosis and replication time far outweigh initial costs, as documented in studies of scientific AI platforms [11].

“My workflow changes too often.” Some researchers maintain that rapidly evolving exploratory projects render fixed schemas impractical. The blueprint counters this by making the audit trail inherently adaptive: the standardized format supports dynamic extension of ontologies, and the append-only nature accommodates mid-campaign pivots without invalidating prior records. The provenance graph grows to reflect reality rather than forcing reality into a rigid template.

“This is extra work for researchers.” The perception that audit trails impose additional labor is understandable but misplaced. Because capture is fully automatic [Principle 1], the researcher’s workload actually decreases: gone are the hours spent reconstructing notebooks or answering replication queries. Instead, the audit trail becomes a reusable asset that accelerates grant reporting, student training, and collaborative science [26]. Moreover, once journals mandate submission, compliance becomes no more burdensome than current data-deposition requirements.

These replies rest on the same principles that have successfully scaled provenance tools in adjacent domains, demonstrating that the proposed system is not only feasible but actively beneficial.

Conclusion

This blueprint articulates a comprehensive conceptual foundation and practical architecture for scientific audit trails in materials AI systems. By defining a seven-component, provenance-graph-based record that is automatically captured, standardized, immutable, and accessible, the proposal directly remedies the auditability gap that currently undermines reproducibility, error detection, trust, and cumulative progress in the field. The operational principles, success criteria, and phased implementation path together provide the materials AI community with both the “why” and the “how” required to move from fragmented ad-hoc practices to a new standard of scientific accountability. As the community embraces this blueprint, materials AI will transition from powerful yet opaque prediction engines into fully transparent, auditable instruments of discovery—capable of withstanding the highest standards of scrutiny while accelerating the reliable design of next-generation materials for energy, sustainability, and technology. The time has come to make scientific audit trails not an optional accessory but an essential feature of every material’s AI workflow.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 May 2024 Revised: 04 Aug 2024 Accepted: 18 Sep 2024
Published online: 18 January 2025

References

Moreau L, Groth P. Provenance: An introduction to PROV. Cham: Springer Nature; 2022.

Huyen C. Designing machine learning systems. Sebastopol, CA: O'Reilly Media; 2022.

Wu YH. Capturing the unobservable in AI development: Proposal to account for AI developer practices with ethnographic audit trails (EATs). AI Ethics. 2025;5(6):5705-18.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Montoya JH, Aykol M, Anapolsky A, Gopal CB, Herring PK, Hummelshøj JS, et al. Toward autonomous materials research: Recent progress and future challenges. Appl Phys Rev. 2022;9(1):011405.

Soedarmadji E, Stein HS, Suram SK, Guevarra D, Gregoire JM. Tracking materials science data lineage to manage millions of materials experiments and analyses. npj Comput Mater. 2019;5(1):79.

Persaud D, Ward L, Hattrick-Simpers J. Reproducibility in materials informatics: Lessons from a general-purpose machine learning framework for predicting properties of inorganic materials. Digit Discov. 2024;3(2):281-6.

Wang AY, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. Chem Mater. 2020;32(12):4954-65.

DeCost BL, Hattrick-Simpers JR, Trautt Z, Kusne AG, Campo E, Green ML. Scientific AI in materials science: A path to a sustainable and scalable paradigm. Mach Learn Sci Technol. 2020;1(3):033001.

Statt MJ, Rohr BA, Guevarra D, Suram SK, Morrell TE, Gregoire JM. The materials provenance store. Sci Data. 2023;10(1):184.

Statt MJ, Rohr BA, Brown K, Guevarra D, Hummelshøj J, Hung L, et al. ESAMP: Event-sourced architecture for materials provenance management and application to accelerated materials discovery. Digit Discov. 2023;2(4):1078-88.

Adorf CS, Dodd PM, Ramasubramani V, Glotzer SC. Simple data and workflow management with the Signac framework. Comput Mater Sci. 2018;146:220-9.

Pérez B, Rubio J, Sáenz-Adán C. A systematic review of provenance systems. Knowl Inf Syst. 2018;57(3):495-543.

Otyepka M, Pykal M, Otyepka M. Advancing materials discovery through artificial intelligence. Appl Mater Today. 2025;47:102981.

Papin JA, Mac Gabhann F, Sauro HM, Nickerson D, Rampadarath A. Improving reproducibility in computational biology research. PLoS Comput Biol. 2020;16(5):e1007881.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: Status, challenges, and perspectives. Adv Sci. 2019;6(21):1900808.

Singh V, Patra S, Murugan NA, Toncu DC, Tiwari A. Recent trends in computational tools and data-driven modeling for advanced materials. Mater Adv. 2022;3(10):4069-87.

Wang F, Jiang S, Li J. The AI-driven transformation in new materials manufacturing and the development of intelligent sports. Appl Sci. 2025;15(10):5667.

Grant NV, Bejan A, Kunze C, Burkhardt H. Can AI-supported systems help with aftercare planning? Opportunities and challenges from a clinical perspective. In: Proceedings of Mensch und Computer 2024. New York, NY: ACM; 2024:655-9.

Nedelkoski S, Bogatinovski J, Mandapati AK, Becker S, Cardoso J, Kao O. Multi-source distributed system data for AI-powered analytics. In: European Conference on Service-Oriented and Cloud Computing. Cham: Springer; 2020:161-76.

Pouchard L, Lin Y, Van Dam H. Replicating machine learning experiments in materials science. Upton, NY: Brookhaven National Laboratory; 2020.

Yuan Y, Chaffart D, Wu T, Zhu J. Transparency: The missing link to boosting AI transformations in chemical engineering. Engineering. 2024;39:45-60.

Creel KA. Transparency in complex computational systems. Philos Sci. 2020;87(4):568-89.

Amirian B, Dale AS, Kalinin S, Hattrick-Simpers J. Building trustworthy AI for materials discovery: From autonomous laboratories to z-scores. arXiv preprint arXiv:2512.01080. 2025 Nov 30.

Abolhasani M, Brown KA, Guest Editors. Role of AI in experimental materials science. MRS Bull. 2023;48(2):134-41.

Zivic F, Malisic AK, Grujovic N, Stojanovic B, Ivanovic M. Materials informatics: A review of AI and machine learning tools, platforms, data repositories, and applications to architectured porous materials. Mater Today Commun. 2025;48:113525.

Maffettone PM, Friederich P, Baird SG, Blaiszik B, Brown KA, Campbell SI, et al. What is missing in autonomous discovery: Open challenges for the community. Digit Discov. 2023;2(6):1644-59.