From Diffraction Peaks to Auditable Scientific Conclusions

Aaron Bailey, Jose Rivera, Adam Cooper*

* Corresponding author: Adam Cooper

Review Article | Frontiers in Integrative Science | PaperSpread Research Publishing | Published 23 September 2026

ABSTRACT

This evidence synthesis examines auditability in Rietveld refinement and competing structural models. It argues that the appropriate object of evaluation is the linked history of sample preparation, pressure-temperature path, instrument output, code version, fit, and scientific claim, not a model score in isolation. The review brings together the assigned studies with established work on uncertainty, robustness, provenance, and governance. Across these literatures, a common problem emerges: a plausible conclusion may depend on an unrecorded stress state, preprocessing choice, or alternative structural interpretation. The proposed framework separates evidence quality, model behavior, decision policy, and operational monitoring, then asks how each layer changes under distribution shift, adversarial pressure, or incomplete information. It recommends evaluation by slices and repeated trials, explicit reject and escalation policies, preservation of data and reasoning lineage, and prospective monitoring tied to defined actions. The result is a research agenda for systems that are efficient enough to use but also bounded enough to audit. No new experiment is claimed; the article develops a comparative conceptual model and identifies tests that would make future empirical claims more credible.

KEYWORDS

diffraction peaks; auditable scientific conclusions; scientific; structural; evaluation; monitoring; enough

Introduction

How can an evidence graph maintain alternative fits and discarded hypotheses? A satisfactory answer must look beyond the predictor, because operational use joins several technical and institutional stages. Before an output appears, designers have already decided what counts as evidence, which representations are admissible, how alternatives are compared, and how uncertainty changes action. In Rietveld refinement and competing structural models, these choices interact with institutions and physical processes that the predictive component only partially represents. A high score can coexist with fragile inputs, an evaluator that rewards the wrong surface feature, or an action rule that turns modest uncertainty into a costly decision. For this reason, the synthesis examines auditability as an evidence problem. The inquiry focuses on the minimum evidence and comparison set required for a credible claim. This requirement gives practical content to the article's central question: How can an evidence graph preserve alternative fits and discarded hypotheses?

A workflow view organizes the comparison. Its unit is the linked history of sample preparation, pressure-temperature path, instrument output, code version, fit, and scientific claim. This scope makes it harder to commit a familiar error: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. The same view makes differences among the focal studies easier to interpret. Although their application settings differ, they each make some part of an inference chain more documented - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The comparison examines whether that documented component remains meaningful when instrument traces, diffraction and spectroscopy products, simulation outputs, laboratory records, and analysis repositories change, and whether the proposed safeguards lead to replication, reanalysis, anomaly review, and documented preservation of competing hypotheses rather than merely producing another metric. Seen through auditability, the issue is not merely technical; it determines which claims remain open to challenge.

Comparative Reading of the Target Literature

Evidence for auditability becomes concrete in MuSK: Multi-Scale Knowledge Learning for Provenance-Graph Anomaly Detection. The paper addresses how stealthy multi-stage attacks can be detected in heterogeneous system-provenance graphs without abundant attack labels through self-supervised recovery of masked node attributes, meta-path edges, and positional structure, followed by meta-path-guided subgraph verification. Its evaluation is informative because evaluation spans nine provenance settings, including DARPA Transparent Computing hosts, and reports high precision, complete recall on those hosts, and efficiency gains from cached incremental expansion. The design joins local attributes, heterogeneous causal paths, and global position while moving decisions from isolated nodes to suspicious subgraphs. This does not make the method a universal component; it identifies a design hypothesis that must be re-tested in the article's focal setting. In particular, performance may depend on audit completeness, benign workload diversity, attack coverage, and the stability of hand-selected meta-path semantics. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment conditions (Ma et al. 2026).

The strongest contribution of BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models is not a generic claim that learning improves performance. It is the more specific proposition that how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit. The proposed route is a three-stage pipeline combining execution-verified supervised traces, sequence- and line-level reward models, and PPO with reward redistribution to critical edit regions. Support comes from the fact that evaluation spans SWE-bench Verified, Defects4J transfer, HumanEval-Java, and QuixBugs, with reported gains in both repository repair and cross-language generalization. Line-level credit assignment makes test execution informative at the granularity where patches actually change behavior. Read through the lens of auditability, the study also illustrates why a reported average must remain connected to the workflow that produced it. Benchmark success does not by itself establish maintainability, specification fidelity, or robustness to incomplete and flaky tests. (Li et al. 2026).

Crystal Structure of CaSiO3 Perovskite at 28-62 GPa and 300 K under Quasi-Hydrostatic Stress Conditions asks which CaSiO3 perovskite structure is thermodynamically stable under lower-mantle pressure when deviatoric stress is minimized. It uses synchrotron X-ray diffraction and Rietveld refinement of thermally annealed samples in a neon pressure medium from 28 to 62 GPa. The relevant evidence is that a tetragonal model with octahedral rotation better fits split diffraction lines, while the fitted bulk modulus agrees with first-principles calculations. The study reconciles earlier experimental disagreements with computation by identifying stress state as a source of metastable structures. In the present review, this work matters because auditability cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. The 300 K protocol isolates stress effects but does not reproduce the full temperature path of the lower mantle. (Chen et al. 2018).

Connections to the Wider Literature

Several established literatures clarify the design space. Graph attention makes neighborhood weighting learnable but does not by itself guarantee causal relevance (Velickovic et al. 2018). Anomaly-detection research distinguishes point, contextual, and collective anomalies, a distinction that maps naturally to provenance subgraphs (Chandola, Banerjee, and Kumar 2009). Open-world security analysis warns that laboratory labels and stationary traffic assumptions rarely survive deployment (Sommer and Paxson 2010). Automated threat-intelligence extraction shows both the reach and the noise of public indicator sources (Liao et al. 2016). Together these findings imply that auditability should be evaluated against explicit alternatives and failure costs. In From Diffraction Peaks to Auditable Scientific Conclusions, this literature supplies a specific test of auditability within Rietveld refinement and competing structural models.

The evidence base offers complementary tests rather than one universal metric. ATT&CK provides a shared behavioral vocabulary, but its categories remain an evolving analyst model rather than ground truth (Strom et al. 2018). FAIR principles make provenance, identifiers, and reusable metadata part of scientific evidence quality (Wilkinson et al. 2016). Backtracking work established causal system history as a practical resource for intrusion investigation (King and Chen 2003). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In From Diffraction Peaks to Auditable Scientific Conclusions, this literature supplies a specific test of auditability within Rietveld refinement and competing structural models.

A credible assurance case can be built from findings that operate at different levels. HOLMES demonstrates the value of correlating low-level events into higher-level attack stages (Milajerdi et al. 2019). UNICORN uses provenance structure to

learn long-running system behavior and surface persistent threats (Han et al. 2020). Whole-system provenance research exposes the engineering tension among capture completeness, query utility, and runtime overhead (Pasquier et al. 2017). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In From Diffraction Peaks to Auditable Scientific Conclusions, this literature supplies a specific test of auditability within Rietveld refinement and competing structural models.

What the System Must Make Visible

Another important separation concerns random variation, limited knowledge, and feedback from strategic actors. Aleatory variation concerns irreducible differences among cases. Epistemic uncertainty concerns missing knowledge, limited samples, or unsupported regions of the input space. Strategic adaptation occurs when users, adversaries, markets, or institutions respond to the application itself. These sources demand different controls. Repeated measurement can characterize variation; new evidence and conservative extrapolation can reduce epistemic uncertainty; red teaming and feedback-aware assessment address adaptation. Combining them into one confidence score hides which intervention is appropriate. A defensible system therefore reports uncertainty in a form tied to an available response instead of presenting confidence as a decorative number. For the present focus, auditability therefore becomes a criterion for deciding which evidence deserves further scrutiny.

A four-layer model exposes where assurance claims can fail. The evidence layer records observations and their provenance. The representation layer turns those observations into features, graphs, tokens, or simulated states. The inference layer produces predictions, explanations, translations, anomaly scores, or candidate actions. The decision layer maps those outputs to replication, reanalysis, anomaly review, and clearly stated preservation of competing hypotheses. Reliability can fail at any boundary. Better inference cannot repair evidence that omitted the relevant condition, and a calibrated probability cannot rescue an action rule whose costs were never specified. Conversely, a conservative decision policy may reduce harm even when the underlying model remains imperfect. This layered view makes auditability testable because each claim can be assigned to a component and a measurable transition. For the present focus, auditability therefore becomes a criterion for deciding which evidence deserves further scrutiny.

Evaluation That Matches the Decision

A defensible protocol starts from documented claims instead of available metrics. Each intended claim - such as improved robustness, safer abstention, faster updating, or better structural localization - needs a target population, comparator, outcome, and boundary condition. Random splits are insufficient when related samples share a patient, repository, instrument run, season, organization, or simulated design family. Grouped and temporal splits better approximate the novelty encountered after deployment. Stress tests should vary one factor at a time when attribution is important, then combine factors to measure interaction. Repeated runs are necessary whenever decoding, optimization, retrieval, or numerical kernels can vary. Reporting only the best seed or a pooled mean makes operational instability disappear. Seen through auditability, the issue is not merely technical; it determines which claims remain open to challenge.

Metric choice must be derived from the decision. Discrimination measures whether cases are ranked correctly; calibration asks whether confidence corresponds to observed frequency; selective risk asks what error remains after deferral; robustness measures performance across defined perturbations; and utility assigns costs to actions. These are not interchangeable. For auditability, the minimum report should include overall performance, slice-level results, uncertainty intervals, coverage-risk curves, stability across runs, and failure examples linked back to instrument traces, diffraction and spectroscopy products, simulation outputs, laboratory records, and analysis repositories. If the system updates incrementally, the testing must also test forgetting, stale evidence, and rollback. If an automatic judge supplies labels, a sample needs independent human review and content-preserving perturbations that reveal whether the judge is grading substance. For the present focus, auditability therefore becomes a criterion for deciding which evidence deserves further scrutiny.

Operational Governance and Human Review

A governance layer translates validation results into permissions and constraints. A deployment specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the deployed pipeline. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed pipeline, or new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the estimator and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. Seen through auditability, the issue is not merely technical; it determines which claims remain open to challenge.

Human intervention works only when the review task is feasible and consequential. Reviewers need enough context to challenge the output, but not so much generated rationale that weak evidence is obscured by volume. Escalation queues require prioritization and workload limits. A reject option that sends nearly everything to experts is safe only on paper; a reject option that rarely fires may be equally ineffective. For Rietveld refinement and competing structural models, oversight should therefore be evaluated with realistic staffing, time pressure, and disagreement. The system should record the final action and its reason without turning that record into surveillance unrelated to the stated purpose. Contestability, privacy, and security are properties of this institutional process as much as of the learned component. For the present focus, auditability therefore becomes a criterion for deciding which evidence deserves further scrutiny.

Designing the Operational Workflow

Operational design in Rietveld refinement and competing structural models begins with typed evidence. Each record should identify its source, time, collection conditions, transformations, and relation to other records. Graphs are useful when relations carry meaning, but graph construction is itself a hypothesis. An edge may denote temporal adjacency, physical causation, code dependency, shared infrastructure, or analyst assertion; treating these as interchangeable creates false certainty. The architecture should therefore preserve edge types and confidence, retain alternative links when evidence is ambiguous, and version both the data schema and the rules that construct the graph. A model may aggregate this structure, but the original evidence must remain recoverable for audit. This requirement gives practical content to the article's central question: How can an evidence graph preserve alternative fits and discarded hypotheses?

The next design choice concerns where additional computation is worth its cost. Easy, well-supported cases can take a fast path. Shifted, ambiguous, or high-cost cases should activate additional precision, retrieval, alternative drafts, simulation, or expert review. This conditional design is preferable to applying maximum computation everywhere because resource cost is part of implementation quality. It is also preferable to an unconditional fast path because efficiency can amplify a systematic error at scale. The trigger must be evaluated as carefully as the expensive model: coverage, false reassurance, subgroup effects, latency, and the consequences of abstention all matter. The output is not simply a prediction but a package containing evidence links, uncertainty, the action taken, and the conditions that caused escalation. Seen through auditability, the issue is not merely technical; it determines which claims remain open to challenge.

Research Agenda

Research should begin with cases selected for what they can reveal about failure, not merely for availability. Cases should represent unsupported regions, ambiguous evidence, conflicting modalities, rare but costly conditions, and realistic adversarial transformations. Each case needs provenance and a reason for inclusion. The second priority is intervention testing: when uncertainty is detected, compare additional computation, retrieval, alternative reasoning paths, model ensembles, and human escalation under the same resource budget. This reveals whether a proposed safeguard actually changes the decision or only changes the explanation. The third priority is transfer. A method should be re-evaluated across sites, devices, languages, organizations, or physical regimes before broad claims are made. In Rietveld refinement and competing structural models, this distinction should be tested before it changes an operational decision.

A complementary research program should improve how evidence accumulates over time. Benchmarks should maintain negative results, failed branches, and changed definitions so later work does not learn only from successful demonstrations. Shared reporting should include data lineage, versioned assessment code, confidence intervals, and enough error material to reproduce major conclusions. Prospective studies should predefine monitoring and stopping rules,

then measure how people adapt to the application. Finally, researchers should compare simple baselines with complex architectures under equivalent information. Complexity is justified when it adds a measurable capability - for example, structural localization, calibrated deferral, or incremental updating - not merely when it creates a richer narrative around the same errors. For the present focus, auditability therefore becomes a criterion for deciding which evidence deserves further scrutiny.

Conclusion

Auditability is credible only when it is attached to an explicit evidence chain and decision policy. The studies considered here make the evidence chain more explicit through structured exploration, typed graphs, conditional revision, adaptive computation, physical constraints, and repeated assessment. A reported benchmark gain still leaves the boundary of safe transfer to be established. For Rietveld refinement and competing structural models, the practical standard should be a traceable workflow that measures shift and instability, supports replication, reanalysis, anomaly review, and explicit preservation of competing hypotheses, and preserves enough evidence for an independent reviewer to reconstruct failure. Future work should compare safeguards under realistic resource and staffing limits, publish negative and subgroup results, and treat monitors and automated judges as components requiring their own validation. The final standard is not uninterrupted automation but evidence-linked action with explicit deferral and independent verifiability. Seen through auditability, the issue is not merely technical; it determines which claims remain open to challenge.

References

1. Ma, Mingjun, et al. "MuSK: Multi-Scale Knowledge Learning for Provenance-Graph Anomaly Detection." Computer Networks 289 (2026): 112728.

2. Li, Yuanhao, et al. "BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models." arXiv preprint arXiv:2605.09134 (2026).

3. Chen, Huawei, et al. "Crystal Structure of CaSiO3 Perovskite at 28-62 GPa and 300 K under Quasi-Hydrostatic Stress Conditions." American Mineralogist 103.3 (2018): 462-468.

4. Velickovic, Petar, et al. "Graph Attention Networks." International Conference on Learning Representations, 2018.

5. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey." ACM Computing Surveys, vol. 41, no. 3, 2009, article 15.

6. Sommer, Robin, and Vern Paxson. "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection." 2010 IEEE Symposium on Security and Privacy, 2010, pp. 305-316.

7. Liao, Xiaojing, et al. "Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 755-766.

8. Strom, Blake E., et al. MITRE ATT&CK: Design and Philosophy. MITRE Corporation, 2018.

9. Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data, vol. 3, 2016, article 160018.

10. King, Samuel T., and Peter M. Chen. "Backtracking Intrusions." Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, 2003, pp. 223-236.

11. Milajerdi, Sadegh M., et al. "HOLMES: Real-Time APT Detection through Correlation of Suspicious Information Flows." 2019 IEEE Symposium on Security and Privacy, 2019, pp. 1137-1152.

12. Han, Xueyuan, et al. "UNICORN: Runtime Provenance-Based Detector for Advanced Persistent Threats." Network and Distributed System Security Symposium, 2020.

13. Pasquier, Thomas, et al. "Runtime Analysis of Whole-System Provenance." Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1601-1614.