Causal Subgraphs for Reconciling Experiment and First-Principles Prediction

Frank Rogers, Raymond Reed, Jack Cook*

* Corresponding author: Jack Cook

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

ABSTRACT

Which linked observations justify preferring a thermodynamic interpretation over a metastable one? This review answers by treating model-experiment reconciliation as a property of a sociotechnical workflow rather than a feature that can be read from average accuracy. The focal setting is high-pressure crystallography, where a plausible conclusion may depend on an unrecorded stress state, preprocessing choice, or alternative structural interpretation. Evidence from the assigned publications is synthesized with foundational studies of calibration, distribution shift, causal structure, and responsible deployment. Four requirements follow: preserve the lineage of instrument traces, diffraction and spectroscopy products, simulation outputs, laboratory records, and analysis repositories; measure stability across relevant perturbations; connect confidence to a specific action; and maintain a route for human challenge and correction. The framework distinguishes descriptive performance from decision utility and separates uncertainty about the world from uncertainty created by the model and its evaluator. It also shows why faster inference or richer reasoning is valuable only when it improves a defined decision under a transparent resource budget. The article is a literature review and research agenda, not a report of a newly completed trial.

KEYWORDS

causal subgraphs; reconciling experiment; first-principles prediction; causal; interpretation; decision; uncertainty

Introduction

Which linked observations justify preferring a thermodynamic interpretation over a metastable one? The problem resists a model-only answer; once deployed, a predictor becomes part of a wider decision process. Its behavior reflects a chain of design decisions about evidence, representation, alternatives, resource use, and response. In high-pressure crystallography, these choices interact with institutions and physical processes that the estimator 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. Accordingly, this review treats model-experiment reconciliation as an evidence problem. The task is therefore to identify the evidence that makes a dependability claim testable. In high-pressure crystallography, this distinction should be tested before it changes an operational decision.

This review follows the inference process from source to action. Its unit is the linked history of sample preparation, pressure-temperature path, instrument output, code version, fit, and scientific claim. Such a unit of analysis guards against one common mistake: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. It also gives the target studies a common basis for comparison. Although their application settings differ, they each make some part of an inference chain more explicit - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The key test is whether that explicit 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 explicit preservation of competing hypotheses instead of merely producing another metric. This point narrows the research task for high-pressure crystallography: compare alternatives under the same information and resource constraints.

What the Core Studies Establish

MuSK: Multi-Scale Knowledge Learning for Provenance-Graph Anomaly Detection asks how stealthy multi-stage attacks can be detected in heterogeneous system-provenance graphs without abundant attack labels. It uses self-supervised recovery of masked node attributes, meta-path edges, and positional structure, followed by meta-path-guided subgraph verification. The relevant evidence is that 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. In the present review, this work matters because model-experiment reconciliation cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. Performance may depend on audit completeness, benign workload diversity, attack coverage, and the stability of hand-selected meta-path semantics. (Ma et al. 2026).

A useful test case is BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models. Rather than treating its output as an isolated score, the study frames how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit. Its central mechanism 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, and its evidential basis is 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. For high-pressure crystallography, the transferable lesson is methodological: a system should preserve the path from signal to decision and expose the conditions under which that path may fail. The boundary is equally important. Benchmark success does not by itself establish maintainability, specification fidelity, or robustness to incomplete and flaky tests. (Li et al. 2026).

Evidence for model-experiment reconciliation becomes concrete in Crystal Structure of CaSiO3 Perovskite at 28-62 GPa and 300 K under Quasi-Hydrostatic Stress Conditions. The paper addresses which CaSiO3 perovskite structure is thermodynamically stable under lower-mantle pressure when deviatoric stress is minimized through synchrotron X-ray diffraction and Rietveld refinement of thermally annealed samples in a neon pressure medium from 28 to 62 GPa. Its evaluation is informative because 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. 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, the 300 k protocol isolates stress effects but does not reproduce the full temperature path of the lower mantle. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment conditions (Chen et al. 2018).

Conceptual Foundations

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 rather than presenting confidence as a decorative number. This requirement gives practical content to the article's central question: Which linked observations justify preferring a thermodynamic interpretation over a metastable one?

The analysis distinguishes evidence, representation, inference, and decision as separate layers. 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 explicit 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 model-experiment reconciliation testable because each claim can be assigned to a component and a

measurable transition. This requirement gives practical content to the article's central question: Which linked observations justify preferring a thermodynamic interpretation over a metastable one?

A Traceable System Architecture

Operational design in high-pressure crystallography 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 maintain 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 point narrows the research task for high-pressure crystallography: compare alternatives under the same information and resource constraints.

A uniform inference budget is rarely the best operational policy. 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 deployment 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 model-experiment reconciliation, the issue is not merely technical; it determines which claims remain open to challenge.

Relevant Foundations

Several established literatures clarify the design space. Backtracking work established causal system history as a practical resource for intrusion investigation (King and Chen 2003). 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). Together these findings imply that model-experiment reconciliation should be evaluated against explicit alternatives and failure costs. In Causal Subgraphs for Reconciling Experiment and First-Principles Prediction, this literature supplies a specific test of model-experiment reconciliation within high-pressure crystallography.

The evidence base offers complementary tests rather than one universal metric. Relational graph convolution provides a foundation for treating typed edges as distinct evidence channels (Schlichtkrull et al. 2018). Inductive graph representation learning is essential when new entities arrive after training (Hamilton, Ying, and Leskovec 2017). Graph attention makes neighborhood weighting learnable but does not by itself guarantee causal relevance (Velickovic et al. 2018). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In Causal Subgraphs for Reconciling Experiment and First-Principles Prediction, this literature supplies a specific test of model-experiment reconciliation within high-pressure crystallography.

A credible assurance case can be built from findings that operate at different levels. 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). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In Causal Subgraphs for Reconciling Experiment and First-Principles Prediction, this literature supplies a specific test of model-experiment reconciliation within high-pressure crystallography.

Testing Claims Rather Than Scores

A defensible protocol starts from explicit 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. This requirement gives practical content to the article's central question: Which linked observations justify preferring a thermodynamic interpretation over a metastable one?

The action and its costs should determine the reported metrics. 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 model-experiment reconciliation, 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. This point narrows the research task for high-pressure crystallography: compare alternatives under the same information and resource constraints.

Experiments the Field Still Needs

The first research priority is failure-mechanism sampling in place of another convenient random split. 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. Seen through model-experiment reconciliation, the issue is not merely technical; it determines which claims remain open to challenge.

A complementary research program should improve how evidence accumulates over time. Benchmarks should retain 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. This requirement gives practical content to the article's central question: Which linked observations justify preferring a thermodynamic interpretation over a metastable one?

Deployment Controls

Measurements become useful only when they alter deployment limits. A deployment specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the operational tool. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed process, or new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the model and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. This requirement gives practical content to the article's central question: Which linked observations justify preferring a thermodynamic interpretation over a metastable one?

Oversight requires its own capacity, interface, and testing plan. 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 high-pressure crystallography, 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 estimator. This point narrows the research task for high-pressure crystallography: compare alternatives under the same information and resource constraints.

Conclusion

Model-experiment reconciliation is credible only when it is attached to an clearly stated evidence chain and decision policy. Across the reviewed studies, several mechanisms can strengthen that chain: 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 high-pressure crystallography, the practical standard should be a traceable pipeline that measures shift and instability, supports replication, reanalysis, anomaly review, and clearly stated 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 clearly stated deferral and independent verifiability. This point narrows the research task for high-pressure crystallography: compare alternatives under the same information and resource constraints.

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

5. 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.

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

7. 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.

8. Schlichtkrull, Michael, et al. "Modeling Relational Data with Graph Convolutional Networks." The Semantic Web, Springer, 2018, pp. 593-607.

9. Hamilton, William L., Rex Ying, and Jure Leskovec. "Inductive Representation Learning on Large Graphs." Advances in Neural Information Processing Systems, vol. 30, 2017.

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

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

12. 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.

13. 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.