Stress State, Metastability, and the Reproducibility of Crystal Claims

Gregory Edwards, Alexander Stewart, Patrick Morris*

* Corresponding author: Patrick Morris

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

ABSTRACT

This evidence synthesis examines experimental boundary conditions in quasi-hydrostatic structure determination. 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

stress state; the reproducibility of crystal claims; stress; state; claims; evaluation; monitoring

Introduction

How should stress history enter the evidential record for a phase assignment? The problem resists a predictive component-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 quasi-hydrostatic structure determination, these choices interact with institutions and physical processes that the model 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. This makes it useful to treat experimental boundary conditions as an evidence problem. The task is therefore to identify the evidence that makes a assurance claim testable. This requirement gives practical content to the article's central question: How should stress history enter the evidential record for a phase assignment?

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. The choice corrects a frequent analytical shortcut: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. A process-level unit also supports a more precise comparison of the assigned studies. Although their application settings differ, they each make some part of an inference chain more clearly stated - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. Evaluation must establish whether that clearly stated 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 clearly stated preservation of competing hypotheses in place of merely producing another metric. The article's emphasis on experimental boundary conditions makes that boundary observable and, in principle, auditable.

Separating Evidence, Inference, and Action

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 experimental boundary conditions testable because each claim can be assigned to a component and a measurable transition. This point narrows the research task for quasi-hydrostatic structure determination: compare alternatives under the same information and resource constraints.

Three kinds of uncertainty require different responses: aleatory variation, epistemic limits, and strategic adaptation. 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 operational tool 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 evaluation 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. The article's emphasis on experimental boundary conditions makes that boundary observable and, in principle, auditable.

Methods and Boundaries in the Assigned Studies

The strongest contribution of MuSK: Multi-Scale Knowledge Learning for Provenance-Graph Anomaly Detection is not a generic claim that learning improves performance. It is the more specific proposition that how stealthy multi-stage attacks can be detected in heterogeneous system-provenance graphs without abundant attack labels. The proposed route is self-supervised recovery of masked node attributes, meta-path edges, and positional structure, followed by meta-path-guided subgraph verification. Support comes from the fact 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. Read through the lens of experimental boundary conditions, the study also illustrates why a reported average must remain connected to the workflow that produced it. Performance may depend on audit completeness, benign workload diversity, attack coverage, and the stability of hand-selected meta-path semantics. (Ma et al. 2026).

BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models asks how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit. It uses a three-stage pipeline combining execution-verified supervised traces, sequence- and line-level reward models, and PPO with reward redistribution to critical edit regions. The relevant evidence 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. In the present review, this work matters because experimental boundary conditions cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. Benchmark success does not by itself establish maintainability, specification fidelity, or robustness to incomplete and flaky tests. (Li et al. 2026).

A useful test case is Possible H_2O Storage in the Crystal Structure of $CaSiO_3$ Perovskite. Rather than treating its output as an isolated score, the study frames whether $CaSiO_3$ perovskite can host hydrogen at lower-mantle pressures and temperatures. Its central mechanism is laser-heated diamond-anvil synthesis from starting materials with different water contents, combined with in-situ diffraction and infrared observations, and its evidential basis is that non-cubic peak splitting, a possible OH vibration, and systematically smaller unit-cell volumes in hydrous syntheses provide convergent but indirect evidence of water storage. The work expands the lower-mantle water budget beyond bridgmanite and ferropericlase. For quasi-hydrostatic structure determination, 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. The authors emphasize that exact water quantification and exclusion of signals from coexisting hydrous phases remain difficult under pressure. (Chen et al. 2020).

Established Tests and Design Principles

Several established literatures clarify the design space. Automated threat-intelligence extraction shows both the reach and the noise of public indicator sources (Liao et al. 2016). 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). Together these findings imply that experimental boundary conditions should be evaluated against explicit alternatives and failure costs. In Stress State, Metastability, and the Reproducibility of Crystal Claims, this literature supplies a specific test of experimental boundary conditions within quasi-hydrostatic structure determination.

The evidence base offers complementary tests rather than one universal metric. 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). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In Stress State, Metastability, and the Reproducibility of Crystal Claims, this literature supplies a specific test of experimental boundary conditions within quasi-hydrostatic structure determination.

A credible assurance case can be built from findings that operate at different levels. 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). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In Stress State, Metastability, and the Reproducibility of Crystal Claims, this literature supplies a specific test of experimental boundary conditions within quasi-hydrostatic structure determination.

Architecture for Bounded Automation

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 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. In quasi-hydrostatic structure determination, this distinction should be tested before it changes an operational decision.

A bounded automation architecture for quasi-hydrostatic structure determination 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 retain 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 quasi-hydrostatic structure determination: compare alternatives under the same information and resource constraints.

Measurement Under Shift and Repetition

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 experimental boundary conditions, 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 deployed operational sequence updates incrementally, the assessment 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 requirement gives practical content to the article's central question: How should stress history enter the evidential record for a phase assignment?

Evaluation begins by writing each claim in testable form. 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 experimental boundary conditions, the issue is not merely technical; it determines which claims remain open to challenge.

Limits, Monitoring, and Contestability

Human review is an engineered operational sequence, not an automatic safeguard. 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 quasi-hydrostatic structure determination, 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 requirement gives practical content to the article's central question: How should stress history enter the evidential record for a phase assignment?

Governance turns empirical findings into operating boundaries. A deployment specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the deployed workflow. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed workflow, 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 experimental boundary conditions, the issue is not merely technical; it determines which claims remain open to challenge.

Priorities for Empirical Work

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 evaluation 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 deployed process. 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. The implication is especially consequential in quasi-hydrostatic structure determination, where a small modeling choice can redirect attention and resources.

The first research priority is failure-mechanism sampling rather than 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. This point narrows the research task for quasi-hydrostatic structure determination: compare alternatives under the same information and resource constraints.

Conclusion

Experimental boundary conditions is credible only when it is attached to an clearly stated evidence chain and decision policy. Useful mechanisms recur across the literature: structured exploration, typed graphs, conditional revision, adaptive computation, physical constraints, and repeated evaluation. None of these results makes the original benchmark a complete map of safe use. For quasi-hydrostatic structure determination, the practical standard should be a traceable operational sequence 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. A system earns trust by limiting its claims, acting on uncertainty, and making both its evidence and its decision reconstructable. In quasi-hydrostatic structure determination, this distinction should be tested before it changes an operational decision.

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. "Possible H2O Storage in the Crystal Structure of CaSiO3 Perovskite." Physics of the Earth and Planetary Interiors 299 (2020): 106412.

4. 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.

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

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

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

8. 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.

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

10. 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.

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

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

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