CAVA: A Causal Analysis and Validation Agent for Microscopy-Informed Materials Discovery

Kamyar Barakati[1], Elyar Tourani[1], Vivek Chawla[1], and Mohammad Amin Moradi[2]

[1]University of Tennessee, Knoxville, USA
[2]University of Cincinnati, Cincinnati, USA

December 19, 2025

Summary

Microscopy-informed materials discovery increasingly relies on causal reasoning to interpret composition–processing–microstructure–property relationships. While predictive models accelerate screening, experimental decision-making often requires regime-aware causal interpretation. We introduce CAVA (Causal Analysis and Validation Agent), a two-stage causal framework for microscopy-derived data. In Stage I, CAVA performs data-driven causal analysis using established causal discovery algorithms (PC, FCI, and GRaSP), combined with bootstrap-based regime-wise consensus to identify robust causal relationships. In Stage II, CAVA employs a literature-grounded large language model agent to contextualize and validate the discovered causal relationships by aggregating supporting and conflicting evidence from prior studies. By integrating data-driven causal analysis with literature-based validation, CAVA provides a transparent framework for interpreting and communicating causal reasoning in the absence of ground-truth causal graphs, thereby supporting faster, more reliable decision-making in scientific experimentation.

1 Introduction

Data-driven approaches have become central to modern materials discovery, particularly in microscopy-informed studies of composition–processing–microstructure–property (CPMP) relationships [1]. Advances in high-throughput experimentation, automated characterization, and imaging have enabled the acquisition of large, high-dimensional datasets, motivating the widespread adoption of machine learning models for structure–property prediction, surrogate modeling, and accelerated materials screening [2, 3]. While these approaches have demonstrated substantial success in improving predictive performance, they primarily capture statistical correlations rather than underlying causal relationships.

In experimental materials science, many critical decisions require causal reasoning to guide experimental design and interpret the effects of targeted interventions. Questions such as which compositional variable should be perturbed to induce a targeted microstructural change, which relationships remain stable across processing regimes, or which observed dependencies reflect mechanistic coupling rather than confounding cannot be reliably answered using correlation-based models alone. Addressing these questions requires explicit causal analysis.

Causal discovery methods provide a principled framework for inferring candidate causal relationships directly from observational data. Constraint-based approaches such as the PC algorithm and Fast Causal Inference (FCI) infer causal structure through conditional independence testing, while score-based approaches such as Gradient-based Score Pursuit (GRaSP) optimize global objective functions over directed acyclic graphs [4, 5, 6, 7]. These methods have seen increasing adoption in the physical sciences, including materials systems, where controlled interventions are often limited or expensive [8, 9]. However, causal discovery in experimental settings faces persistent challenges: ground-truth causal graphs are rarely available, inferred structures can be sensitive to noise and finite-sample effects, and causal relationships may vary across experimental regimes or compositional ranges.

At the same time, decades of experimental research in materials science have produced a rich scientific literature containing reported causal hypotheses, intervention outcomes, and mechanistic interpretations. Although individual studies are necessarily limited in scope and experimental context, the collective literature provides an indirect but valuable reference for evaluating newly inferred causal relationships. Systematically integrating this prior knowledge into data-driven causal workflows remains an open challenge.

In this work, we introduce CAVA (Causal Analysis and Validation Agent), an integrated two-stage framework for causal reasoning in microscopy-informed materials discovery. In Stage I, CAVA performs data-driven causal analysis using established causal discovery algorithms (PC, FCI, and GRaSP), combined with bootstrap-based regime-wise consensus analysis to identify robust and regime-dependent causal relationships from microscopy-derived data. In this context, experimental regimes correspond to distinct compositional ranges. In Stage II, CAVA employs a literature-grounded large language model (LLM) agent to contextualize and validate the discovered causal relationships by aggregating supporting and conflicting evidence from prior experimental studies. The objective of this stage is not to infer new causal structure, but to contextualize and evaluate data-inferred causal hypotheses against previously reported experimental evidence.

CAVA explicitly separates causal discovery from causal validation by combining data-driven inference with literature-grounded contextualization, documenting alignment between inferred causal relationships and prior experimental evidence without enforcing consensus.

2 Methodology

2.1 Stage I: Regime-Wise Data-Driven Causal Analysis

Stage I of the CAVA framework performs data-driven causal analysis directly from microscopy-derived experimental data. The objective of this stage is to infer candidate causal relationships, assess their robustness under resampling, and evaluate their consistency across compositional regimes. The material system considered in this study is samarium-doped bismuth ferrite, $Sm_xBi_{1-x}FeO3$, where Sm substitution serves as the primary compositional control parameter. Experimental data were obtained for a total of 14 microscopy snapshots spanning five Sm concentration regimes: 0%, 7%, 10%, 13% and 20%.

Microscopy-derived variables describing composition, lattice geometry, ferroelectric polarization, and structural descriptors were extracted from each snapshot and assembled into tabular form. For each snapshot, the following variables were constructed: total alkali cation content, metal cation content, lattice parameters (a, b), lattice angle, unit-cell volume, and in-plane ferroelectric polarization components (P_x, P_y). Missing or non-finite values were imputed using local interpolation followed by forward–backward filling. All variables were standardized within each compositional regime prior to causal analysis.

The full dataset was partitioned into discrete compositional regimes based on Sm concentration. For each regime, all corresponding snapshots were concatenated to form a regime-specific dataset used for independent causal analysis. Regime-specific cleaned datasets are provided in the project repository for full reproducibility.

Causal relationships were inferred independently within each regime using three causal discovery algorithms: the PC algorithm, FCI, GRaSP. PC and FCI were applied using Fisher's Z conditional independence test, with FCI additionally accounting for latent confounding through partial ancestral graph representations. GRaSP was used as a score-based alternative to constraint-based inference. Directed edges were extracted from PC and GRaSP outputs, while both directed edges and undirected skeleton edges were retained from FCI to assess latent-robust dependencies.

Each causal discovery method was applied under bootstrap resampling. For each regime and method, the dataset was resampled with replacement and causal discovery was repeated over 200 bootstrap iterations across 10 random seeds. Edge frequencies were recorded across bootstrap realizations, enabling estimation of mean occurrence, variability, and minimum confidence across seeds. This procedure filters spurious edges arising from finite-sample effects and isolates stable causal relationships.

For each regime, bootstrap statistics were aggregated to produce method-specific edge confidence tables, including directed causal edges (PC, GRaSP) and latent-robust skeleton edges (FCI). Edges exceeding predefined confidence thresholds were retained as regime-supported relationships. A final consensus table (the "gold table") was constructed by aggregating supported edges across all regimes and recording the presence or absence of each edge per regime. This table provides a compact representation of regime-dependent causal structure and serves as the input to Stage II of the CAVA framework.

2.2 Stage II: Literature-Grounded Causal Validation

Stage II of the CAVA framework performs literature-grounded validation of the candidate causal relationships identified in Stage I. Each candidate causal relationship inferred in Stage I is represented as a directed hypothesis of the form $X \to Y$. These hypotheses are automatically transformed into domain-specific bibliographic search queries using an initial large language model agent (GPT-4o-mini), which optimizes terminology, synonyms, and experimental context. The resulting queries are used to retrieve high-relevance scientific metadata and abstracts from the arXiv repository via its public API.

Retrieved abstracts are segmented into individual sentences and analyzed by a secondary LLM agent trained to identify causal statements. Sentence-level filtering is performed using an explicit intervention-based criterion inspired by the Rubin causal model: a sentence is retained only if it describes an experimental or procedural intervention on variable X that induces a measurable change in variable Y. This filtering step distinguishes causal assertions from purely correlational or descriptive statements.

Validated causal statements are encoded into a directed graph representation using NetworkX. Each directed edge is annotated with its supporting evidence, including the extracted sentence, arXiv identifier, and, when available, associated DOI information. This structure enables traceable linkage between inferred causal relationships and their corresponding literature support.

3 Results

3.1 Stage I: Causal Analysis Results

Causal discovery was performed independently within each Sm concentration regime using the PC, FCI, and GRaSP algorithms. For each regime, directed causal graphs and latent-robust skeleton graphs were inferred and evaluated under bootstrap resampling to assess stability under finite-sample perturbations. For completeness and transparency, the full set of regime-wise causal graphs obtained in Stage I is provided as supplementary visualizations in the project repository (GitHub). These figures illustrate the detailed causal structure inferred within each Sm concentration regime and serve as supporting evidence for the cross-regime analysis summarized below.

Across individual regimes, several causal relationships consistently appeared across bootstrap realizations and across multiple causal discovery methods, indicating stable dependencies that are robust to resampling and algorithmic variation. Other relationships were method-specific, weakly supported, or exhibited reduced stability and were therefore excluded from subsequent consensus analysis.

Overall, the inferred causal structures exhibit clear regime dependence, with both the presence and directionality of candidate causal relationships varying systematically with Sm concentration. While this regime specificity is scientifically meaningful, it also presents a practical limitation for downstream validation: the available experimental literature rarely reports causal relationships at narrowly defined compositional regimes. As a result, direct regime-by-regime literature validation is often infeasible.

To address this mismatch between data-driven regime specificity and literature granularity, we aggregate causal evidence across regimes and focus subsequent validation on cross-regime causal hypotheses that demonstrate sufficient stability. This aggregation enables meaningful comparison with the broader experimental record and forms the basis for Stage II of the CAVA framework.

Cross-Regime Causal Consistency To characterize how causal relationships persist, disappear, or re-emerge across Sm concentration, a consensus table (the "gold table") was constructed by aggregating stable causal relationships across all regimes. This table records the presence or absence of each supported causal relationship within each Sm concentration range and provides a compact representation of cross-regime causal consistency.

Figure 1 presents a heatmap representation of the cross-regime consensus. Several causal relationships are consistently observed across all regimes, including strong couplings between lattice parameters and unit-cell volume, as well as directed relationships linking compositional variables to structural descriptors. Other relationships exhibit pronounced regime dependence, appearing only at intermediate or higher Sm concentrations. In particular, causal interactions involving lattice angle and ferroelectric polarization components display clear sensitivity to composition, indicating progressive structural reorganization with increasing Sm substitution.

3.2 Stage II: Literature-Grounded Validation Results

The directed causal relationships identified in Stage I were subjected to literature-grounded validation using the CAVA agent. Each candidate causal edge was treated as a hypothesis of the form $X \to Y$ and queried against the scientific literature at the abstract level using an LLM-assisted retrieval and filtering pipeline. Supporting and conflicting evidence was extracted using an intervention-based criterion, ensuring that retained statements reflected explicit experimental or procedural manipulations rather than correlational associations.

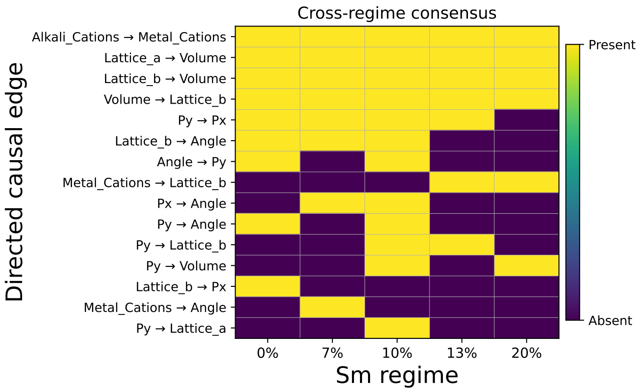

Figure 1: Cross-regime consensus (gold table). Rows correspond to stable directed causal edges and columns to Sm concentration regimes. A filled cell indicates that the edge satisfies the Stage I bootstrap confidence criterion within that regime.

A system of six parameters—alkali cations, metal cations, lattice distance, lattice angle, unit-cell volume, and in-plane ferroelectric polarization—was analyzed, yielding $6 \times 5/2 = 15$ possible pairwise relationships. For each parameter pair, up to 30 relevant publications were queried, resulting in a total of 450 abstract-level evaluations. Each query was treated as a binary assessment of support, in which the literature was examined to determine whether explicit abstract-level evidence indicated that one parameter causally influences another. A directed edge was recorded only when such supporting information was present; otherwise, the relationship was classified as unsupported.

Figure 2 presents the aggregated causal graph constructed from the literature-grounded validation process. Nodes correspond to the same physical variables analyzed in Stage I, while directed edges represent causal relationships supported by intervention-consistent statements extracted from prior studies. Edge presence in this graph reflects literature-level evidence aggregation rather than data-driven inference.

Comparison between the Stage II literature-derived causal graph and the Stage I cross-regime heatmap (Fig. 1) reveals several notable patterns. A subset of causal relationships—particularly lattice–volume couplings and composition-driven structural effects—exhibits strong agreement between data-driven inference and prior experimental reports. These relationships appear both as persistent edges across Sm regimes in Stage I and as well-supported edges in the literature-derived graph, indicating convergence between empirical reporting and physics-based understanding.

Other relationships display partial or context-dependent alignment. In particular, several ferroelectric polarization-mediated interactions inferred in Stage I appear intermittently across Sm concentration regimes and are supported in the literature only under specific experimental conditions or compositional ranges. These cases highlight the importance of contextual interpretation rather than binary validation.

Finally, a small number of inferred causal relationships lack clear precedent in the existing literature. Rather than being dismissed, these edges represent candidate hypotheses for further experimental investigation, especially in compositional regimes that remain underexplored.

For transparency, a tabulated summary of causal edge occurrence across literature sources is provided in the supplementary material (see `causal_edge_occurrence_table.pdf` in the project repository). This table documents the frequency and consistency of literature support for each tested causal relationship.

Consistency between literature-supported relationships and physics-based models reflects convergence between empirical reporting and theoretical understanding, rather than implying that one framework supersedes the other. The resulting graph therefore provides a concise, evidence-based summary of causal relationships among key compositional and structural parameters as articulated in the literature, and serves as a fast, easily accessible reference to support decision-making in scientific experimentation.

Limitations and Scope. CAVA is designed as an interpretive and validation framework rather than a definitive causal oracle. Stage I relies on observational causal discovery methods and is therefore subject to the usual limitations associated with finite sample sizes, latent confounding, and conditional independence testing. While bootstrap-based consensus mitigates instability, inferred causal relationships should be interpreted as candidate hypotheses rather than confirmed mechanisms.

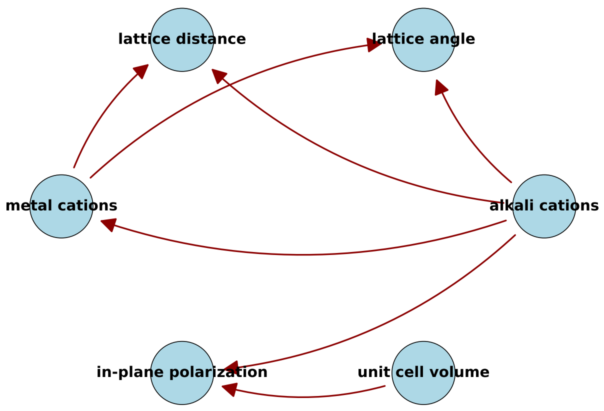

Figure 2: Literature-grounded causal graph produced by the CAVA agent. Directed edges represent causal relationships supported by intervention-consistent statements extracted from prior studies. Node set matches Stage I variables, enabling direct comparison with the cross-regime causal heatmap.

Stage II operates at the abstract level and depends on the availability and clarity of reported experimental interventions in the literature. As such, causal relationships that are underreported, implicit, or domain-specific may be underrepresented. Moreover, literature validation is not performed in a regime-resolved manner, reflecting practical limitations in targeting compositional specificity within existing publications.

Finally, CAVA does not enforce agreement between data-driven and literature-derived causal relationships. Divergence should be interpreted as an opportunity for discovery rather than a failure of inference. Within this scope, CAVA is best viewed as a decision-support tool that augments causal reasoning by combining data-driven structure with literature-grounded context.

4 Conclusion

We introduced CAVA, a two-stage framework for causal reasoning in microscopy-informed materials discovery that explicitly separates data-driven causal analysis from literature-grounded causal validation. In Stage I, CAVA integrates established causal discovery algorithms with bootstrap-based, regime-wise consensus analysis to identify robust and composition-dependent causal relationships directly from experimental data. In Stage II, these candidate relationships are contextualized against the existing scientific literature using an agentic large language model workflow, providing structured evidence without enforcing consensus or mechanistic certainty.

Applied to samarium-doped bismuth ferrite, CAVA reveals both regime-persistent and regime-dependent causal structure linking composition, lattice geometry, and ferroelectric polarization descriptors. By documenting alignment, ambiguity, and divergence between inferred causal relationships and prior experimental evidence, CAVA enables transparent interpretation of causal hypotheses in the absence of ground-truth causal graphs.

More broadly, CAVA provides a reproducible and extensible framework for integrating causal discovery with literature-based contextualization in experimental materials science. The approach is not limited to a specific material system or causal discovery algorithm and can be readily adapted to other microscopy-derived datasets where causal interpretation is essential for guiding experimental design.

Supporting Information

Further details, supporting materials, and source code can be found in the project's GitHub repository.

References

[1] Sergei V Kalinin, Evgheni Strelcov, Alex Belianinov, Suhas Somnath, Rama K Vasudevan, Eric J Lingerfelt, Richard K Archibald, Chaomei Chen, Roger Proksch, Nouamane Laanait, et al. Big, deep, and smart data in scanning probe microscopy, 2016.

[2] Keith T Butler et al. Machine learning for molecular and materials science. *Nature*, 2018.

[3] Krishna Rajan. Materials informatics. *Materials Today*, 2015.

[4] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[5] Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.

[6] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 2008.

[7] Wing Hung Lam et al. Grasp: A greedy relaxation algorithm for learning dags. *Journal of Machine Learning Research*, 2022.

[8] Jakob Runge et al. Inferring causation from time series in earth system sciences. *Nature Communications*, 2019.

[9] Bernhard Schölkopf et al. Toward causal representation learning. *Proceedings of the IEEE*, 2021.