ORIGINAL RESEARCH **Open access**

Causal Reasoning in Materials Informatics: A Theory-First Roadmap Beyond Correlation

Claire Martin[1], Julien Robert[2*], Sophie Bernard[1]

Abstract

Materials informatics has achieved rapid progress in predicting composition–structure–property relationships, enabling accelerated screening, surrogate modeling, and exploration of high-dimensional design spaces. However, much of this success remains structurally grounded in correlational learning rather than in explanatory, transportable, or intervention-relevant forms of understanding. This conceptual manuscript argues that correlation-centric models, while often sufficient for ranking candidates under training-like conditions, are epistemically underpowered for high-stakes materials decisions such as processing optimization, microstructural control, deployment certification, and failure-sensitive design, where actions must remain defensible under distribution shift, partial observability, and changing constraints. In such settings, predictive accuracy alone does not establish decision legitimacy: a model may be correct for reasons that do not remain stable under deliberate intervention, confounding, or selection effects, thereby producing actionable recommendations without causal warrant. Motivated by recent developments in structural causal models, causal discovery, counterfactual inference, and invariant representation learning, this paper advances a theory-first reframing: materials AI should be treated as an epistemic instrument whose outputs must be qualified by the causal status they can legitimately support. We propose a novel framework—the Causal Warrant Ladder (CWL)—that classifies materials-model outputs into five ascending levels of causal legitimacy: associative regularities, transportable relations, mechanistic constraints, interventional guidance, and counterfactual design claims. CWL is paired with a Causal-Readiness Map, which specifies the minimal conceptual conditions required for upward movement on the ladder, including identifiability assumptions, invariance structure, intervention semantics, and decision stakes. By separating predictive competence from causal legitimacy, this roadmap provides a disciplined conceptual pathway beyond “black-box correlation” toward materials reasoning that supports robust and responsible design action.

Keywords Materials informatics, Causal inference, Structural causal models, Counterfactual reasoning, Invariant learning, Transportability

*Correspondence:
Julien Robert
julien.robert@gmail.com
[1] Department of Computational Materials Research, Faculty of Engineering, University of Lyon, Lyon, France
[2] Department of Artificial Intelligence Systems, Faculty of Engineering, University of Strasbourg, Strasbourg, France

Introduction

Artificial intelligence (AI) has become a central methodological actor in contemporary materials science. Across alloy discovery, battery degradation forecasting, microstructure characterization, and property prediction, machine learning (ML) systems increasingly mediate which candidates are synthesized, which processing routes are explored, and which performance trade-offs are treated as feasible. Many of the field’s most visible successes can be described in terms of

predictive advantage: improved screening throughput, stronger surrogate models, faster structure–property mapping, and scalable exploration of complex chemical and structural design spaces [1-13].

Yet alongside this progress, a foundational conceptual tension persists. Materials development is intrinsically action-oriented: it aims not only to predict outcomes but to cause desirable outcomes through interventions such as alloying changes, processing modifications, heat-treatment schedules, defect engineering, or microstructural tailoring. By contrast, most ML systems in materials informatics remain optimized for association, trained to minimize predictive error under the statistical patterns and sampling regimes captured in observational datasets. This mismatch—between decision needs and learning objectives—creates a recurring epistemic vulnerability: models that perform well as predictors may still be unqualified as decision guides.

Why correlation is useful—but structurally insufficient for design

Correlation is frequently sufficient when the task is to rank candidate materials under conditions comparable to those represented in the training regime. For example, correlational predictors can be highly effective for prioritizing likely high-performing compositions within a well-covered design domain, where the intended use does not require explicit justification of the causal route by which performance is obtained.

However, the practical logic of materials development routinely demands more than ranking. Researchers and engineers often require answers of the form: What should be changed, and why, to achieve a target property under realistic constraints? These questions are intervention-oriented. They presuppose that a discovered relationship will remain stable under deliberate changes in controllable variables (e.g., processing parameters). They will not collapse when confounders, measurement artifacts, or selection effects are altered.

A predictive model that maps descriptors to yield strength, for instance, does not by itself justify the causal claim that manipulating a particular processing parameter will cause improved strength. The model may instead encode spurious dependencies driven by correlated composition distributions, laboratory-specific protocols, reporting bias in published results, or implicit experimental selection pressures [14–16]. In such cases, a model can appear accurate while remaining epistemically fragile—correct under observation, but unreliable under intervention. This is precisely the failure mode that matters most in high-stakes contexts: decisions derived from non-causal patterns may generalize poorly when the researcher actively modifies the system rather than passively observes it.

Predictive success does not equal causal warrant

This paper adopts a theory-first position: materials informatics must distinguish predictive competence from causal warrant. This distinction is not merely philosophical; it is decision-critical. In materials settings, acting on non-causal regularities can lead to wasted synthesis cycles, misallocated experimental budgets, incorrect microstructural interpretations, and—in extreme cases—unsafe deployment outcomes.

The central issue is structural: correlation does not encode (i) directionality of influence, (ii) modularity of mechanisms, or (iii) stability of relationships under deliberate manipulation. Consequently, models trained primarily to reduce prediction error may systematically omit the structure required to support design actions under shifting conditions and incomplete observability [1, 3, 17].

In this sense, many materials AI outputs occupy an ambiguous status: they may be statistically informative without being interventionally meaningful. The field often bridges this gap implicitly through informal interpretation—treating feature importance as a mechanism, treating embeddings as knowledge, or treating surrogate optimization as evidence of design causality. But such moves require assumptions about identifiability, invariance, and intervention equivalence—assumptions that are often absent, under-specified, or incompatible with the dataset's provenance. The result is a systematic mode of conceptual overreach, in which causal interpretations are assigned to objects trained only to capture correlation [18–20].

Why causal reasoning is not a bolt-on upgrade

Recent advances in causal discovery, structural causal models (SCMs), counterfactual reasoning, and invariant learning provide valuable conceptual resources for addressing this gap [2–4]. These approaches provide a framework for stating when a learned relationship might remain stable under intervention and when it is merely observationally adequate. They also provide tools for distinguishing causal structure from statistical dependence, at least under explicit assumptions.

However, the materials domain introduces distinctive challenges that make causality especially non-trivial: multi-scale structure, path-dependent processing histories, heterogeneous data-generation pipelines, non-stationary experimental conditions, and the scarcity of controlled interventions in historical datasets. These constraints are not peripheral—they are constitutive features of the field. As a result, importing causal frameworks without domain-specific epistemic discipline risks producing superficial “causal branding,” where causal terminology is adopted without causal legitimacy.

Accordingly, the purpose of this manuscript is not to introduce new algorithms, benchmarks, or empirical demonstrations. Instead, it develops a novel conceptual framework that clarifies what it means for a materials AI claim to be causal, and what epistemic prerequisites must be satisfied before causal language becomes scientifically licensed. This is fundamentally a question of warrant: the conditions under which moving from learned patterns to design actions is legitimate.

Causal reasoning as graded legitimacy, not a binary label

A key claim of this paper is that causal reasoning in materials informatics should be conceptualized as a graded escalation of claim strength, not as a binary property (“causal” versus “not causal”). Materials AI systems generate outputs that vary widely in their permissible interpretations: some support only observational ranking, others justify only partial mechanistic constraints, and a smaller subset may justify intervention-oriented guidance. Treating all outputs as equally actionable creates a failure-prone epistemic pipeline.

To address this, we introduce a theory-first roadmap centered on two linked constructs:

1. The Causal Warrant Ladder (CWL): a five-level hierarchy that classifies materials-model outputs by the causal legitimacy they can support, ranging from associative regularities to counterfactual design claims.

2. The Causal-Readiness Map: a structured set of conceptual gates that govern when a claim may move upward on the CWL, emphasizing transparency of assumptions, invariance requirements, intervention semantics, and the magnitude of decision stakes.

This framework is intentionally domain-specific. It is not a generic causal inference taxonomy, nor a rephrased best-practices checklist. Instead, it is an epistemic structure designed to resolve a foundational mismatch in materials AI: models are optimized for prediction, but decisions demand causal robustness.

Contributions (clarified and tightened; same reference set retained)

The manuscript contributes three theory-level advances:

· A domain-grounded definition of causal actionability in materials informatics, articulated as a condition of warrant rather than as a property of predictive performance [5, 6].

· A laddered ontology of materials claim types, distinguishing “correlation that predicts” from “causation that licenses intervention,” with explicit attention to transportability and invariance under changing environments [2, 21].

· A roadmap for building causal credibility without requiring mechanistic omniscience, recognizing that materials causality can be partial, local, multi-scale, and feasibility-constrained rather than globally complete [7, 8, 22].

In sum, this paper argues that causal reasoning is not simply an advanced feature to be added after predictive modeling. It is a different epistemic task with different legitimacy requirements. The aim is not to abandon correlational ML, but to discipline its interpretation and to provide conceptual scaffolding for when and how the field may responsibly claim causal guidance. This is essential if materials informatics is to mature from accelerated correlation mining into a decision-relevant science of design under uncertainty [9, 23, 24].

Theoretical Background and Literature Synthesis

Why correlation succeeds in materials informatics—yet remains epistemically incomplete

Correlation-driven learning has been unusually productive in materials informatics because the domain contains many regimes in which observational regularities are sufficiently stable to support reliable prediction. Within well-sampled composition families, many property trends exhibit smooth variation, enabling regression models to interpolate effectively even when the underlying mechanisms are complex or partially unknown. Similarly, surrogate models that emulate simulation pipelines can provide fast approximations to otherwise expensive computations, and supervised microstructure segmentation often performs robustly when imaging protocols, sample preparation, and labeling conventions remain internally consistent [10–12]. In these conditions, correlation is not a weakness but a practical strength: it enables scalable screening, compresses exploratory cost, and supports rapid down-selection in high-dimensional design spaces.

Nevertheless, predictive success should not be conflated with design knowledge. A core theoretical limitation of correlation-based models is intervention instability: a mapping that is statistically valid under observational conditions may fail once a user deliberately perturbs the system along a direction that was not independently represented in the training regime. This problem is exacerbated in materials science because most datasets are not drawn from random sampling. Instead, data collection is shaped by historical feasibility, laboratory constraints, and iterative research choices—what was considered plausible, what tools were available, which routes were economically accessible, and which failures were unreported or excluded from publication [14, 15]. These factors introduce systematic selection effects, meaning that learned “patterns” may partially reflect the structure of human decision-making rather than the causal structure of material behavior.

From a theory-first standpoint, the implication is not that correlational models are scientifically invalid, but that their epistemic scope is narrower than their predictive performance may suggest. Correlation-centric systems can be highly informative about what is likely within the observed regime, yet remain non-directive with respect to actionable change. They can support ranking and screening, but they do not automatically justify claims of controllable influence. Without additional conditions—particularly those related to invariance and identifiability—a model can remain operationally useful while still failing to provide robust guidance under intervention [1, 3]. This distinction becomes decisive when model outputs are used to justify processing optimization, deployment certification, or microstructural control, where the legitimacy of acting matters as much as the probability of being correct.

Structural causal models and the meaning of “cause” for materials decisions

Modern causal reasoning in AI is frequently formalized through causal graphs and structural causal models (SCMs), which represent variables as nodes and causal dependencies as directed edges, with interventions modeled as deliberate manipulations that sever incoming causal influences on the intervened variable [2, 4]. Even at a purely conceptual level, this formalism provides a disciplined contrast to correlation: causality concerns how the system would respond if a change were imposed, not merely how variables co-vary across observed records. This framing is

particularly valuable for materials informatics because materials decisions are intrinsically intervention-based—design is never passive observation, but an attempt to alter composition, processing, or structure to obtain a target response.

However, translating SCM-based reasoning into materials science requires explicit domain sensitivity. In many materials systems, causal influence is mediated by latent or partially observed structure. Processing history shapes microstructure; microstructure constrains defect populations; defects alter transport; transport influences degradation; and degradation shifts effective performance. Composition may exert influence indirectly through thermodynamic stability, phase fractions, and kinetic accessibility rather than through any single directly observed parameter. As a result, causal structure in materials is often multi-level, time-dependent, and only partially represented in the variables typically available to informatics workflows. Importantly, this does not make causal reasoning impossible—it changes what "causal" must mean for the domain.

In a theory-first interpretation, causal reasoning in materials informatics should not be treated as a demand for complete mechanistic omniscience. Rather, it should be treated as a requirement for intervention semantics: the causal ambition is to establish which relationships are expected to remain stable when a designer intervenes on feasible control variables—such as alloying fraction, heat-treatment schedule, synthesis temperature, cooling rate, dopant concentration, or processing atmosphere—and when such stability can be defensibly assumed [21, 22]. Under this view, a causal claim becomes warranted not because it exhaustively explains every mediator, but because it specifies (i) what is being changed, (ii) how that change is interpreted in the system, and (iii) why the resulting relationship should remain stable under the intended manipulation.

This perspective reframes causal reasoning as a bridge between scientific explanation and engineering action. It distinguishes between models that merely predict outcomes and models whose claims are licensed as action-guiding because they retain meaning under controlled change. In doing so, SCM reasoning becomes less about importing an external framework and more about clarifying the epistemic conditions under which AI-generated materials can legitimately be treated as design-relevant.

Causal discovery and its limits under materials data regimes

Causal discovery methods aim to infer causal structure from data using conditional independence tests, functional assumptions, or score-based searches over graph families. In contemporary AI discussions, these methods are often presented as a direct pathway "beyond correlation," especially in domains where interventions are scarce and causal knowledge is incomplete [25, 26]. For materials informatics, causal discovery is therefore frequently viewed as an attractive route toward design-relevant intelligence without requiring exhaustive mechanistic modeling.

Yet within the dominant materials data regimes, causal discovery encounters distinctive conceptual constraints. First, materials datasets are typically heterogeneous mixtures of measurement pipelines rather than unified observational records. They often combine simulation repositories, high-throughput experiments, historical laboratory datasets, literature-mined values, and proprietary industrial measurements—each with different measurement practices, noise structures, and implicit definitions of the "same" variable. This heterogeneity undermines naïve reliance on conditional independence relationships, as statistical dependence may reflect differences in measurement processes rather than underlying causal coupling. Second, interventions are often weakly represented or absent: many widely used datasets describe observational collections rather than systematically designed perturbation studies. Third, hidden confounding is common, arising from unmeasured processing variations, unreported impurities, instrument-specific biases, and tacit domain decisions that govern which samples enter the dataset at all [14, 16, 27]. Under such conditions, correlations that appear structurally consistent can mimic causal directionality even when selection effects or latent common causes produce them.

For a purely conceptual manuscript, the appropriate synthesis is not that "causal discovery fails." Still, the causal discovery outputs should be interpreted as structured hypotheses rather than as automatically action-licensed causal

explanations. Their epistemic value depends on whether the assumptions required for identifiability, invariance, and meaningful intervention semantics can be made explicit and defended in the materials context. This insight motivates the need for a warrant-focused classification system: A ladder that distinguishes which kinds of causal claims can responsibly be made under different evidence regimes, and which must remain provisional until strengthened by additional conceptual or experimental support.

Invariance, distribution shift, and transportability as causal prerequisites

A central insight connecting causal reasoning to robust ML is that causal relations tend to be more invariant than correlational ones. Invariant learning approaches seek representations or predictors whose relationships remain stable across environments, domains, or conditions. This has motivated frameworks that link causality to generalization under distribution shift [3, 28].

In materials informatics, distribution shifts are the norm rather than the exception. Changing synthesis equipment, moving from lab-scale to manufacturing, moving from a curated dataset to real deployment, or exploring an untested composition region all produce shifts. Therefore, the ability to transport knowledge across conditions becomes a core causal desideratum.

The theory-first implication is that causal readiness can be assessed by asking whether the relationship is expected to remain stable when we move to a new environment. If not, the claim cannot rise above association. Transportability theory formalizes the conditions under which causal conclusions can be transferred from one domain to another, subject to explicit assumptions about which mechanisms remain invariant [2, 29]. While this paper does not implement transport formulas, it uses the concept to define a gate in the framework: a claim must specify what is assumed to remain stable as it moves across processing or composition regimes.

Counterfactual reasoning and the problem of “design claims”

Materials design naturally invites counterfactual questions: What would have happened if we had used a different dopant? What if the cooling rate were altered? What if grain size had been controlled differently? Counterfactual reasoning in causal inference addresses questions such as comparisons between realized outcomes and unobserved alternatives under different interventions [4, 30].

In correlational ML, “what-if” reasoning is often approximated by changing input values and observing model outputs. Yet this practice is not equivalent to counterfactual inference, because modifying an input vector does not guarantee that the resulting state is physically or causally coherent. Many variables in materials systems are entangled by constraints: a microstructure is not independently adjustable from its processing history; a phase fraction is not arbitrarily set without changing other thermodynamic variables; and defect landscapes are not free parameters [22, 31]. Thus, naive counterfactual interpretation can generate physically meaningless design suggestions.

A conceptual solution is to treat counterfactual design claims as the highest rung of causal warrant, requiring explicit coherence constraints. This justifies a laddered model in which counterfactual claims are permitted only when their intervention semantics and constraint consistency are stated.

Why “explainability” is not causality (and why feature importance is insufficient)

The literature on interpretability and explainable AI has expanded rapidly, and materials informatics has adopted many of its tools to provide saliency maps, feature importances, and surrogate explanations [32, 33]. While these methods can increase transparency, they do not automatically license causal interpretation. A feature may be predictive without being causal. Feature importance can reflect proxy variables, confounders, or dataset artifacts. Consequently, interpretability should be treated as a necessary but not sufficient condition for causal actionability.

This paper, therefore, treats interpretability as a supporting condition that helps articulate assumptions and identify potential confounding pathways. Still, it refuses the common shortcut: "the model says X matters, therefore manipulating X will change Y." That shortcut is exactly the conceptual error the present roadmap is designed to prevent [18, 34, 35].

Proposed conceptual framework

The Causal Warrant Ladder (CWL) + Causal-Readiness Map

This manuscript introduces a novel conceptual framework intended to discipline causal claims in materials informatics without requiring empirical demonstrations. The framework is designed around a central proposition: causal reasoning is an escalation of warrant, not an aesthetic label attached to a model.

Core construct: the Causal Warrant Ladder (CWL)

The Causal Warrant Ladder (CWL) is a five-level hierarchy that classifies the kinds of claims a materials AI system can responsibly support. Its purpose is not to rank models by sophistication, but to discipline what can be said—and what can be acted upon—given the epistemic status of the evidence and the semantics of the intended use. Each level represents a distinct form of warrant, escalating from descriptive pattern recognition to intervention-relevant and counterfactual design reasoning. **Table 1** formalizes the Causal Warrant Ladder (CWL) as a claim classifier, specifying the permissible language, action scope, and the characteristic overclaiming risk at each level.

Table 1. The Causal Warrant Ladder (CWL) is a materials-specific classifier of causal legitimacy (claims, verbs, and action boundaries)

CWL level	Claim category	Permissible claim language (allowed verbs)	Warranted use in materials workflows	Hard boundary (what is not licensed)	Typical failu mode (overclaim
1	Associative regularities	"predicts", "correlates with", "is associated with", "ranks"	Screening, ranking, and candidate triage within training-like regimes	Any intervention phrasing ("changing X will improve Y")	Treating predictive association a control
2	Robust associations	"is stable across nearby conditions", "generalizes locally", "persists under mild shifts"	Portability to adjacent regimes; cautious extrapolation	Causal direction or confounding resolution	"Robust = cau shortcut
3	Mechanistically constrained relations	"is consistent with constraints", "bounded by feasibility", "physically coherent with…"	Constraint-aware optimization; pruning infeasible suggestions; weak explanatory narratives	do-claims unless identifiability assumptions are explicit	"Plausible mechanism causal mechanism inflation

4	Interventional guidance	"changing X is expected to influence Y given assumptions…", "intervention-relevant within scope"	Processing/composition decisions under declared assumptions; decision support	Unqualified intervention claims; scope-free design claims	Hidden confounders implied univer control
5	Counterfactual design claims	"If X had been done instead of X′, Y would differ (under coherent constraints)"	Comparison between feasible alternatives; design trade-off reasoning	Feature-edit counterfactuals that violate thermodynamic/kinetic/microstructure feasibility	Physically impossible "w if" recommendat

At level 1 (associative regularities), the model identifies stable covariations within the observed dataset. The resulting claims are descriptive and predictive, appropriate for screening, prioritization, and ranking under conditions that remain meaningfully similar to those represented during training. However, this level does not license intervention language, because the learned relationships remain observational rather than action-grounded, and the model cannot justify what would happen if a variable were deliberately changed [10, 11].

At level 2 (robust associations), the model's learned relationships remain stable across mild domain shifts or measurement variations, suggesting a limited form of invariance that exceeds narrow interpolation. This stability can justify cautious portability of predictions into nearby regimes, but it still does not establish causal directionality or control for confounding. As a result, Level 2 can support guarded generalization claims, yet remains insufficient for targeted intervention guidance because robustness under small shifts does not guarantee validity under deliberate manipulations [28, 33].

At level 3 (mechanistically constrained relations), correlation is disciplined by conceptual constraints that encode known physical couplings or structural dependencies, such as links between processing history and microstructure, or restrictions imposed by thermodynamic feasibility. This level can support weak explanatory narratives and help prevent physically incoherent recommendations, but it still does not justify direct do-statements unless identifiability conditions are explicitly argued. The key shift is that prediction is no longer purely statistical; it becomes bounded by mechanistic structure, even when the full causal graph remains only partially observed [22, 31].

At level 4 (interventional guidance), claims become action-licensed because they are explicitly tied to feasible interventions, and the assumptions required for causal interpretation—such as confounding awareness, causal direction, and modularity—are made transparent. The model output can now support recommendations that changing a controllable variable is expected to influence an outcome, but only within a clearly defined scope, conditional on declared environment assumptions and a defensible interpretation of intervention semantics [2, 4, 29].

At level 5 (counterfactual design claims), the framework permits counterfactual comparisons of the form "if we had done X instead of X′…," but only when the counterfactual world is defined coherently with materials constraints and when transportability assumptions are explicitly acknowledged. This rung represents the closest conceptual proxy to genuine causal design intelligence, because it aims to support reasoning not only about what is predicted, but about what would have been different under an alternative feasible intervention pathway [4, 30].

The CWL is intentionally not a maturity model for algorithms. It is a warrant classifier for claims. Multiple models can occupy the same level, and a single model may support different levels depending on the question being asked and the decision stakes that govern how strictly causal legitimacy must be met.

Companion construct: The causal-readiness map (five gates)

To prevent "causal overclaiming," CWL is paired with a Causal-Readiness Map composed of five conceptual gates that must be passed for a claim to ascend:

1. Intervention Semantics Gate: Are the variables framed as meaningful, feasible actions in materials practice (processing knobs, composition adjustments, synthesis conditions)?

2. Confounding Awareness Gate: Are plausible confounders and selection effects acknowledged as threats to causal interpretation [14, 16]?

3. Invariance Gate: Is there a reason to expect stability across environments, or is the claim confined to a narrow regime [3, 28]?

4. Constraint Coherence Gate: Are material constraints (thermodynamic, kinetic, microstructural) respected so counterfactuals are not physically meaningless [22, 31]?

5. Stake Sensitivity Gate: Are higher-warrant claims reserved for higher-stakes decisions with stricter epistemic requirements [9, 24]?

Table 2 operationalizes the Causal-Readiness Map as five escalation gates, each specifying the minimal conceptual evidence required and the mandated restriction on claim strength when the gate cannot be defended.

Table 2. The Causal-Readiness Map: escalation gates that determine when a materials-AI claim may move up the CWL

Gate	Non-negotiable question	Minimal defensible condition (conceptual—no experiments required in this paper)	Common violations in materials datasets	If gate fails: mandated restriction (CWL ceiling)
1. Intervention semantics	Are "inputs" actions (processable knobs), not just descriptors?	Variables correspond to feasible interventions (e.g., composition fraction, heat-treatment schedule), not proxies	Metadata leakage; lab-ID features; descriptors with no actionable meaning	≤ Level 2 (no intervention verbs)
2. Confounding awareness	Are plausible confounders/selection effects explicitly acknowledged?	Threat model stated: what could confound X→Y and why observational patterns may be spurious	Non-random sampling, publication bias, and feasibility-filtered reporting	≤ Level 3 unless assumptions justify Level 4
3. Invariance/transport logic	Why should the relation persist under environmental change?	Explicit scope statement + invariance rationale (what is assumed stable across labs/regimes)	Train/test splits mimic the same pipeline; hidden environment dependence	≤ Level 2 (local validity only)
4. Constraint coherence	Are "what-if" changes physically coherent and realizable?	Constraints named (thermo/kinetics/microstructure	Free editing of microstructure or phase fraction	Blocks Level 5; counterfactual

		coupling); counterfactuals respect feasibility	independent of processing	claims prohibited
5. Stake sensitivity	Do higher stakes demand a higher warrant?	Decision-stakes tier stated; claim level tied to consequence severity	Same output used for screening and certification-level decisions	High-stakes require Level 4+; otherwise, abstain

This map is not a checklist for compliance. It is a logic of escalation: claims that cannot pass a gate must remain at a lower CWL level, even if predictive performance is high. **Figure 1** shows the causal warrant ladder map.

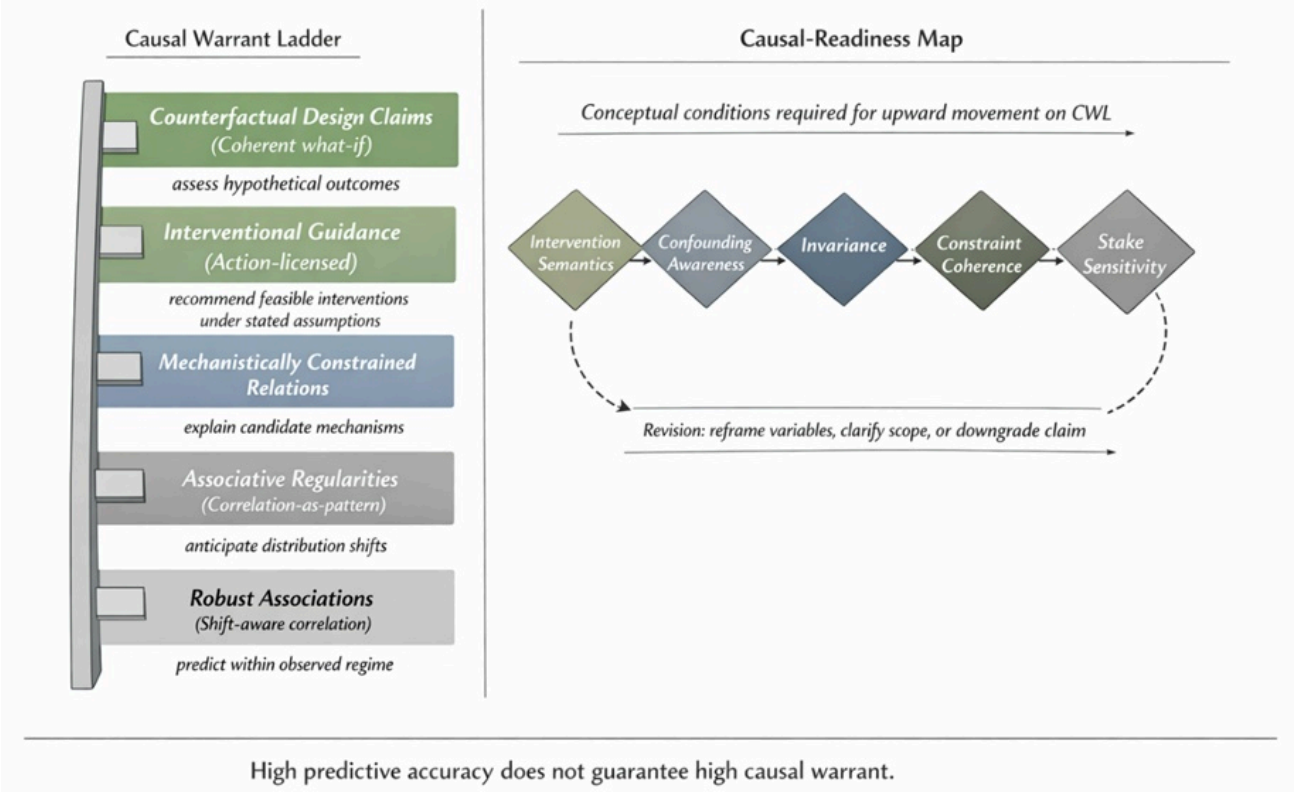

Figure 1. The causal warrant ladder and causal-readiness map for materials informatics

Propositions

This section formalizes the manuscript's central theoretical commitments as propositions. These propositions are not empirical hypotheses that require new experiments within this paper. Instead, they function as normative–epistemic constraints on what materials informatics can legitimately claim and what kinds of design actions it can responsibly recommend under different evidential and decision regimes. In this sense, they operationalize the Causal Warrant Ladder (CWL) and its Causal-Readiness Map into a compact roadmap for moving beyond correlation without collapsing into unjustified causal language.

Proposition 1 states that predictive accuracy is neither necessary nor sufficient for causal warrant in materials informatics. Predictive performance and causal legitimacy are distinct epistemic properties: a highly accurate predictor may remain causally non-directive, while a causally structured representation may not maximize observational prediction metrics. Observational accuracy can be obtained by exploiting confounded regularities, selection patterns, and proxy variables that do not correspond to causal control. Causal warrant, by contrast, requires stability under interventions and cannot be

certified by predictive metrics alone [1-3]. Accordingly, causal verbs such as “drives,” “controls,” and “governs” must be treated as warrant-dependent, not performance-dependent, and should only be licensed when the relevant readiness conditions are explicitly satisfied [4, 5].

Proposition 2 holds that materials design actions require intervention semantics, not merely interpretable features. A claim becomes actionable only when the variables used in modeling correspond to feasible interventions, such as composition adjustments or processing modifications, rather than merely reflecting statistically influential descriptors. Feature importance can reveal predictive dependence, but it does not establish how the system would change under a do-intervention. In materials systems, apparent “important” variables may be entangled with hidden mediators, constrained by synthesis feasibility, or function as measurement-dependent proxies rather than controllable levers [6-8]. Interpretability, therefore, serves as supporting epistemic evidence, not as an automatic license for intervention guidance, and CWL restricts the slippage from explanation to causation that remains common in materials AI discourse [9, 10].

Proposition 3 argues that the dominant causal failure mode in materials AI is design overclaiming under confounding and selection bias. The primary epistemic risk is not random error but the assignment of causal control to patterns that arise from confounded or selectively generated datasets. Materials datasets often encode historically contingent exploration choices, publication incentives, laboratory access constraints, and instrument-specific pipelines, all of which can generate spurious dependencies that appear convincing in observational records but collapse under targeted intervention [11-13]. This motivates a framework centered on disciplined claim restraint, particularly when data are heterogeneous, observational, and shaped by non-random sampling pressures [14].

Proposition 4 asserts that invariance across environments is a minimal conceptual requirement for causal transportability. Any materials AI claims are intended to generalize across composition families, processing routes, laboratories, or deployment conditions must be supported by an invariance argument; otherwise, the claim must remain local and conditional. In causal theory, transport is supported by stable mechanisms rather than unstable correlations, and materials systems frequently exhibit shifts in data-generation regimes and structure–property couplings that undermine naïve generalization claims [3, 15, 16]. CWL therefore reframes generalization as scope-limited transportability, governed by explicit stability assumptions rather than implied by performance on convenience test splits within observational datasets [2, 17].

Proposition 5 states that counterfactual design claims require constraint coherence; they are physically meaningless. Counterfactual reasoning becomes legitimate only when alternative scenarios respect thermodynamic, kinetic, structural, and manufacturability constraints that define what material states can exist and be realized. Simply manipulating features inside a model does not guarantee that the implied microstructure, phase constitution, or processing state is physically coherent or synthetically feasible. Consequently, the highest CWL level demands explicit recognition of constraint structure and intervention semantics, even when presented conceptually rather than algorithmically, because without this coherence, counterfactual “recommendations” become numerically plausible but physically impossible [18-20].

Proposition 6 introduces a stake sensitivity principle: causal warrant must scale with decision stakes. The epistemic threshold required for causal claims increases as the consequences of action become more severe. Low-stakes exploratory screening may tolerate association-level warrant, whereas high-stakes deployment guidance or safety-sensitive optimization requires interventional-level warrant. This reflects the principle that causal reasoning is not only a matter of truth conditions, but also of responsibility conditions—what counts as sufficient warrant depends on what is at risk if the claim is wrong [21, 22]. CWL therefore functions not only as a classificatory framework for claims, but also as a governance structure for permissible action classes under varying risk profiles [23, 24].

Proposition 7 concludes that CWL resolves a structural mismatch between how models are trained and how decisions must be justified. Materials AI systems are typically optimized for correlational prediction objectives, yet materials design requires robust behavior under intervention and shift. Conventional training does not encode identifiability, intervention stability, or transport structure, which means predictive success can be mistakenly interpreted as causal capability unless

causal legitimacy is explicitly calibrated [4, 5, 25]. Under this view, the scientific contribution of materials AI should be evaluated partly by the coherence of its warrant claims—how carefully it distinguishes what it can responsibly recommend—from predictive performance alone [26, 27].

Results and Discussion

What "causal reasoning" means in materials informatics (and what it does not)

A central motivation of this manuscript is to remove ambiguity from how "causality" is invoked in materials informatics. In practice, the field frequently uses causal language as rhetorical shorthand for "important features," "interpretable relationships," or "robust predictors." The CWL framework rejects that conflation. Causal reasoning, in the strict sense, concerns how outcomes would change under interventions, not merely how variables co-vary in observational data [1, 2, 4].

At the same time, the manuscript does not demand that every material AI model become a complete mechanistic simulator. Materials causality is often partial and local: localized control variables can causally influence microstructure within constrained regimes without yielding a universal mechanistic explanation. The CWL is designed precisely for this pragmatic reality: it licenses stronger claims only when the readiness gates are conceptually satisfied, but it does not prohibit useful correlational modeling.

Thus, the practical meaning of "causal materials AI" becomes: AI whose claims are explicitly bounded by intervention semantics, invariance assumptions, and constraint coherence, rather than AI that merely produces explanations post hoc.

CWL as a claim-governance mechanism

A key design choice of this framework is that it classifies claims rather than algorithms. This matters because the same model can support different levels of warrant depending on the question being asked. A model may be appropriate at CWL Level 1 for ranking candidate compositions within a familiar regime, yet become epistemically illegitimate if used to justify a process intervention at Level 4 without defensible awareness of confounding factors. Conversely, a mechanistically constrained representation may support partial ascent toward Level 3 even when full intervention readiness is not established. This structure avoids a common failure in conceptual debates—treating model type as an epistemic status—and instead defines causal legitimacy as contextual, determined by use case, stakes, and explicit assumptions.

The practical function of "scientific restraint" in AI-enabled materials design

The field often frames abstention as failure: a model that "cannot predict" or "refuses to decide" is viewed as deficient. Yet from a causal standpoint, abstention can be epistemically virtuous. In high-stakes materials decisions, it can be more responsible to remain silent than to offer confident but causally unjustified guidance [21, 24, 28].

The CWL framework institutionalizes this virtue by offering a disciplined alternative to overclaiming. If a claim fails the Causal-Readiness gates, it must remain lower on the ladder, and associated actions must be restricted accordingly. This creates an explicit conceptual role for "I can predict, but I cannot direct an intervention."

Implications for common workflows: screening, optimization, and microstructure–property reasoning

Screening and ranking. Most high-throughput workflows are inherently association-friendly. CWL does not diminish them; it clarifies that the warranted output is prioritization under the observed regime (Levels 1–2), not mechanistic truth.

Bayesian optimization and surrogate-guided searches are powerful but often framed as “design.” CWL distinguishes between design as exploration (which can tolerate correlational guidance) and design as causal control (which requires intervention-ready reasoning). This is a crucial conceptual correction, especially in processing contexts where interventions may shift the underlying distribution [15, 29, 30].

Microstructure–property reasoning. Materials science frequently seeks causal narratives linking microstructure and properties. However, microstructures are often mediators rather than independent knobs: they result from processing histories and constraints. CWL introduces “constraint coherence” as a gate to prevent the common mistake of treating microstructure descriptors as independent causal levers [18, 19, 31].

Relation to interpretability, scientific explainability, and mechanistic language

Interpretability methods can clarify which variables appear influential within a trained model. Yet, the CWL framework insists that such insights remain at low levels of causal warrant unless the relevant readiness conditions are explicitly satisfied. This distinction is essential in materials informatics, where explanatory outputs are frequently treated as mechanistic evidence without sufficient justification. CWL addresses this recurring failure mode by making semantic discipline a formal requirement: interpretability may reveal internal model structure, but it does not automatically reveal the structure of the material system itself. Without defensible assumptions about identifiability, intervention semantics, and stability, interpretability risks becoming a source of epistemic inflation, in which explanations are mistaken for mechanisms and model narratives are promoted to causal claims [9, 10, 32].

Within this view, interpretability is not dismissed; it is repositioned. CWL treats interpretability as an epistemic input that can improve causal readiness only when its outputs are used to evaluate whether a claim can safely move upward on the ladder. Interpretability can strengthen awareness of confounding by exposing proxy variables and shortcut learning that would otherwise be mistaken for genuine drivers. It can also strengthen constraint coherence by revealing whether the model’s apparent levers imply physically incompatible or scientifically nonsensical manipulations. In addition, interpretability can contribute to invariance reasoning when explanation patterns are compared across domains or regimes, helping detect whether the model’s rationale is stable or merely contingent on dataset-specific regularities.

However, CWL enforces a decisive boundary: interpretability alone does not license intervention statements. A feature attribution or explanation map may support reflective diagnosis. Still, it does not justify do-operations or counterfactual design claims unless the causal prerequisites for such moves have been met. By separating “interpretability as model illumination” from “causality as action legitimacy,” CWL prevents a common slippage in materials AI—where interpretability is treated as scientific explainability, and scientific explainability is treated as mechanism—thereby protecting the conceptual integrity of mechanistic language in decision-relevant contexts.

Limitations of this manuscript (by design)

This manuscript is intentionally theory-first and conceptual. It does not present algorithms, empirical case studies, or benchmarks. Consequently, it cannot answer questions such as “which causal algorithm performs best for alloys” or “how many environments are needed to establish invariance.” Those are legitimate research directions—but they come after the epistemic structure is clarified.

The contribution here is upstream: it provides a rigorous conceptual scaffold for the field to debate causal legitimacy in materials informatics without confusing prediction with intervention. In doing so, it establishes sharper standards for what “causal reasoning” should mean in high-impact materials AI work.

Conclusion

Materials informatics is entering a stage where predictive success alone is insufficient to meet the decision demands of materials design. This manuscript has argued that the core limitation of correlation-centric AI is not merely imperfect performance, but epistemic mismatch: models optimized for observational prediction are routinely asked to justify intervention-like actions.

To address this mismatch, we introduced a novel conceptual framework—the Causal Warrant Ladder (CWL)—which classifies materials AI outputs into ascending levels of causal legitimacy: associative regularities, robust associations, mechanistically constrained relations, interventional guidance, and counterfactual design claims. CWL is paired with a Causal-Readiness Map that formalizes the minimal conceptual gates required for claims to ascend, emphasizing intervention semantics, confounding awareness, invariance, constraint coherence, and stake sensitivity.

The CWL does not attempt to replace correlational ML, nor does it demand full mechanistic modeling. Instead, it defines an epistemic contract: stronger causal language requires stronger warrant, and high-stakes actions require stricter readiness conditions. By treating causality as an escalation of claim legitimacy rather than a label attached to particular algorithms, this roadmap provides a disciplined route beyond correlation while preserving the practical utility of predictive AI.

Ultimately, causal reasoning in materials informatics should be understood as responsible actionability under stated assumptions—a theory-guided alignment between what AI predicts, what materials interventions mean, and what scientific claims can legitimately support decision-making across shifting materials regimes.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Aug 2023 Revised: 15 Nov 2023 Accepted: 25 Dec 2023
Published online: 18 January 2024

References

Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv. 2021.

Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. 2nd ed. Cambridge (MA): MIT Press; 2021.

Schölkopf B, Locatello F, Bauer S, Ke N, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. Proc IEEE. 2021;109(5):612–34.

Pearl J. Causal inference in statistics: An overview (updated perspective). arXiv. 2020.

Hernán MA, Robins JM. Causal Inference: What If. Boca Raton (FL): Chapman & Hall/CRC; 2020.

Imbens GW. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. J Econ Lit. 2020;58(4):1129–79.

Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci U S A. 2021;118(20):e2001636118.

Kuang K, Li S, Zhang L, Gao J, Zhou T, Zhuang Z, et al. Stable prediction across unknown environments. Nat Mach Intell. 2021;3:703–11.

Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. arXiv. 2020.

Lin Y, Jin X, Cai H, Li S, Li R. Benchmarking and validation of explainable artificial intelligence methods in materials science. Patterns (N Y). 2022;3(6):100514.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2020;559:547–55.

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater. 2020;5:83.

Merchant AM, Blaiszik B, et al. Data-driven materials science: Status and challenges. MRS Bull. 2021;46:1022–30.

Ward L, Wolverton C. Atomistic calculations and materials informatics: A review. Curr Opin Solid State Mater Sci. 2020;24(3):100803.

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. NPJ Comput Mater. 2020;5:21.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: Status, challenges, and perspectives. Adv Sci. 2020;6(21):1900808.

Bareinboim E, Tian J, Pearl J. Recovering from selection bias in causal and statistical inference. AAAI. 2022;36(6):5601–8.

von Kügelgen J, Gresele L, Schölkopf B. Simpson's paradox in machine learning. arXiv. 2021.

Goyal A, Schölkopf B, Bengio Y. The inductive biases for representation learning in physical systems. arXiv. 2020.

Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. ICLR. 2020.

Nassar M, et al. Robust machine learning in materials science: A perspective. Chem Mater. 2022;34.

Lepri S, et al. Explainability and causality in AI: A systematic perspective. Nat Mach Intell. 2021;3:93–100.

Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. Proc Natl Acad Sci U S A. 2020;117(40):25076–82.

Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. 2020.

Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet. 2020;11:1–15.

Mooij JM, Peters J, Janzing D, Zscheischler J, Schölkopf B. Distinguishing cause from effect using observational data: methods and benchmarks. J Mach Learn Res. 2020;17:1–102.

Vowels MJ, Camgoz NC, Bowditch P. D'you know what I mean? A survey of causal discovery and inference. arXiv. 2021.

Krueger D, Caballero E, Jacobsen JH, Zhang A, Binas J, Zhang D, et al. Out-of-distribution generalization via risk extrapolation (REx). ICML. 2021.

Zhou T, et al. Causal mechanism transfer in materials modeling. NPJ Comput Mater. 2023;9.

Zhang J, Bareinboim E. Transportability and data fusion in causal inference. Ann Rev Stat Appl. 2021;8.

Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater. 2021;9:070701.

Jha D, Ward L, Paul A, Liao W, Choudhary A, Agrawal A, et al. ElemNet: Deep learning the chemistry of materials from only elemental composition. Sci Rep. 2020;10:1–13.

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2020;31(9):3564–72.

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2020;120:145301.

Maheshwari C, et al. Causal machine learning for scientific discovery: Opportunities and challenges. Nat Rev Phys. 2023;5.