ORIGINAL RESEARCH **Open access**

Scientific Claims Under Distribution Shift: A Conceptual Theory for Robust Materials AI Inference

Lucas Meyer[1*], Anna Schmid[2], Stefan Braun[1]

Abstract

The integration of artificial intelligence into materials science has accelerated property prediction, inverse design, and discovery pipelines. Yet, the reliability of resulting scientific claims remains vulnerable to distribution shifts—systematic differences between training and inference data distributions arising from variations in synthesis protocols, characterization instruments, environmental conditions, or sampling biases. This purely conceptual manuscript develops a novel theoretical framework for robust materials AI inference in the presence of such shifts. We posit that distribution shifts do not merely degrade predictive accuracy but fundamentally alter the epistemic status of scientific claims by introducing unaccounted covariances between material descriptors and latent generative processes. The framework reconceptualizes inference as a multi-layered epistemic process: (i) shift ontology delineation, (ii) value-laden alignment of data representations with domain invariants, and (iii) claim robustness via counterfactual stabilization. By synthesizing insights from materials informatics, machine learning theory on distribution shifts, and philosophical analyses of epistemic values in science, we argue that robust inference requires explicit modeling of shift-induced epistemic uncertainty rather than mitigation as a post hoc engineering concern. This theory provides a conceptual scaffold for evaluating the validity of AI-derived materials claims across heterogeneous datasets, advancing a shift from performance-centric to epistemically grounded AI deployment in materials science.

Keywords Materials informatics, Data bias, Distribution shift, Scientific claims, Epistemic values, Robust inference

*Correspondence:
Lucas Meyer
lucas.meyer@gmail.com
[1] Department of Materials Modeling and Artificial Intelligence, Faculty of Engineering, ETH Zurich, Zurich, Switzerland
[2] Department of Data-Driven Materials Science, Faculty of Engineering, University of Bern, Bern, Switzerland

Introduction

Materials science stands at the threshold of a data-driven paradigm, where artificial intelligence (AI) systems increasingly mediate the formulation of scientific claims about structure-property relationships, phase stability, and functional performance [1, 2]. The Materials Genome Initiative and subsequent high-throughput computational and experimental efforts have generated vast repositories of materials data, enabling machine learning models to predict properties with apparent high fidelity [3, 4]. Yet this apparent success masks a critical vulnerability: distribution shifts, in which the statistical properties of data encountered during inference diverge from those used in model training. To formalize this divergence, distribution shift may be expressed as a discrepancy between the joint training and deployment data distributions:

$$Ptrain\,(X,Y) \neq Pdeploy\,(X,Y) \quad (1)$$

where X denotes material descriptors (e.g., composition, structure, processing variables) and Y represents observed properties or performance outcomes. This inequality captures the breakdown of the independent and identically distributed (i.i.d.) assumption that underwrites conventional machine learning generalization.

Such shifts arise ubiquitously in materials contexts—from variations in precursor purity across laboratories, to instrument-specific artifacts in spectroscopy, to sampling biases favoring stable crystalline phases over metastable or amorphous structures [5, 6].

Distribution shifts challenge the foundational assumption of i.i.d. (independent and identically distributed) data that underpins most AI applications in materials informatics. When a model trained on density functional theory (DFT) relaxations from one database is applied to experimental synthesis outcomes from another, or when predictions for high-entropy alloys are extrapolated to compositions outside the training range, the resulting inferences may appear statistically confident yet scientifically spurious [7, 8]. This problem is not merely technical; it strikes at the heart of scientific claim-making. In traditional materials research, claims are vetted through reproducibility, mechanistic understanding, and alignment with physical laws. AI-derived claims, however, often rest on correlational patterns that may not generalize under shift, leading to overconfident assertions about material behavior in untested regimes [9, 10].

The prevalence of distribution shifts in materials AI is amplified by the field's inherent heterogeneity. Materials data span multiple scales (atomic to macroscopic), modalities (images, spectra, compositions), and generative processes (computational vs. experimental). Covariate shifts occur when input descriptors (e.g., elemental features) differ in distribution between source and target; label shifts when property measurements vary due to different testing conditions; and concept shifts when the underlying structure-property mapping evolves (e.g., due to temperature-dependent phase transitions) [11, 12]. Recent syntheses of materials informatics highlight that while models achieve low error on in-distribution benchmarks, out-of-distribution performance collapses, undermining claims of "discovery" or "prediction" [13, 14].

This vulnerability has profound implications for scientific epistemology. Scientific claims in materials science carry normative weight: they inform synthesis decisions, guide resource allocation, and influence policy on sustainable materials. When AI inferences falter under shift, they risk propagating epistemic errors—false positives in property prediction or overlooked failure modes—that erode trust in data-driven materials research [15, 16]. Moreover, shifts often embed human and societal values: datasets overrepresent commercially viable materials from well-funded labs, introducing selection biases that reflect economic rather than scientific priorities [17, 18].

Prior responses to distribution shifts in AI have focused on empirical mitigations—such as domain adaptation, invariant learning, and uncertainty quantification [19, 20]. While valuable, these approaches remain downstream of a deeper conceptual deficit: a lack of theoretical articulation of how shifts interact with the epistemic structure of materials claims. This manuscript addresses that gap through a purely conceptual lens. We develop a novel theoretical framework that reconceptualizes robust materials AI inference as an epistemic process rather than a predictive one. The framework emphasizes:

- Ontology of shifts as perturbations to the generative model of materials data.
- Alignment of inference with epistemic values (e.g., generalizability, mechanistic plausibility).
- Stabilization of claims via counterfactual reasoning under plausible shifts.

By foregrounding conceptual foundations, we aim to elevate materials AI from a tool of correlation to a partner in robust scientific reasoning. This theory is timely, as materials informatics matures toward autonomous discovery systems where human oversight diminishes [21, 22]. Without such a framework, AI risks amplifying rather than alleviating epistemic uncertainties in materials claims.

The remainder proceeds as follows. We first synthesize the theoretical background, integrating literature on materials informatics, distribution shift theory, data bias, and values in science. We then propose the conceptual framework, detailing its components and a textual representation of its schematic structure. This work lays the groundwork for future formalizations and empirical validations, but remains strictly conceptual in scope.

Theoretical Background and Literature Synthesis

Materials informatics: From data-driven prediction to claim generation

Materials informatics has progressed from early statistical learning approaches—largely confined to correlational analysis within curated property databases—toward complex artificial intelligence ecosystems capable of inverse design, closed-loop optimization, and semi-autonomous discovery [1, 3]. Foundational contributions established machine learning as a pragmatic strategy for navigating the combinatorial complexity of composition–structure–property relationships, introducing hand-crafted descriptors based on elemental statistics, crystallographic features, and thermodynamic proxies to render materials spaces computationally tractable [23, 24]. These early models primarily supported predictive tasks, offering probabilistic estimates of material properties conditional on observed features.

Over the past decade, however, the field has undergone a qualitative shift. Recent surveys emphasize the rise of deep learning architectures—including graph neural networks, attention-based transformers, and generative models—that encode relational inductive biases and enable scalable exploration of high-dimensional chemical spaces [2, 4]. These methods increasingly serve not only as predictors but as engines of *claim generation*, underpinning assertions about material superiority, design optimality, or mechanistic relevance. In autonomous or semi-autonomous workflows, such claims are propagated downstream to guide experimental prioritization, synthesis decisions, and resource allocation.

Crucially, this expansion from prediction to epistemic claim-making is predicated on assumptions of data consistency and stationarity. Large-scale computational repositories derived from density functional theory (DFT), such as Materials Project and OQMD, provide internally coherent labels generated under standardized approximations and numerical settings. In contrast, experimental datasets exhibit substantial heterogeneity arising from divergent measurement protocols, sample preparation practices, environmental conditions, and operator judgment [25, 26]. When AI-generated claims—such as assertions of enhanced strength, stability, or functional performance—are extrapolated across heterogeneous regimes without explicit accounting for distributional shifts, discrepancies between predicted and observed behavior become systematic rather than incidental [27, 28]. This tension marks a fundamental challenge for materials informatics as it matures into a knowledge-producing scientific infrastructure.

Distribution shift: Taxonomy and mechanisms in materials contexts

Distribution shift theory, rooted in statistical learning and pattern recognition, formalizes deviations between training and deployment environments through distinctions such as covariate shift (changes in $P(X)$ with invariant $P(Y \mid X)$ label shift (changes in $P(Y)$ with invariant $P(X \mid Y)$, and concept shift (changes in the conditional relationship P(Y|X) [11, 29]. While originally articulated in abstract learning settings, these categories acquire concrete and domain-specific instantiations in materials AI. Formally, these shift categories may be distinguished through distributional transformations:

$$Covariate\ shift: Ptrain(X) \neq Pdeploy(X), P(Y \mid X)\ invariant \tag{2}$$

$$Label\ shift: Ptrain(Y) \neq Pdeploy(Y), P(X \mid Y)\ invariant \tag{3}$$

$$\begin{array}{l} Concept\ shift \\ :\ Ptrain\,(Y \\ |\ X) \\ \neq Pdeploy\,(Y \\ |\ X) \end{array} \quad (4)$$

Within materials science, these probabilistic divergences correspond to physically grounded perturbations, including compositional sampling bias, protocol-dependent measurement variation, and evolving structure–property mappings under environmental or processing change.

Covariate shifts commonly arise from compositional, structural, or chemical biases embedded in databases—for example, the systematic overrepresentation of well-studied inorganic crystals in the Materials Project relative to experimentally reported structures in the ICSD [30]. Label shifts manifest when nominally identical properties are measured under divergent protocols, such as mechanical testing conducted at different strain rates, temperatures, or atmospheres, thereby altering the empirical distribution of outcomes [31]. Concept shifts are particularly salient in materials science, where stability relationships, phase boundaries, and property–structure mappings are intrinsically contingent on external conditions such as pressure, temperature, or processing history [32].

Empirical investigations consistently demonstrate that such shifts induce severe performance degradation. Models trained exclusively on simulated data often fail to generalize to experimental spectra due to instrument-specific noise and calibration artifacts. At the same time, property predictors collapse when queried on compositions lying outside the convex hull of the training distribution [5, 33]. From a theoretical standpoint, these phenomena violate the independent and identically distributed (i.i.d.) assumption underpinning standard generalization guarantees, rendering conventional performance metrics epistemically fragile indicators of reliability [34, 35]. In materials AI, distribution shift thus operates not merely as a technical nuisance but as a structural threat to inferential validity. To consolidate how canonical shift categories manifest in materials AI and why they undermine scientific claims, **Table 1** summarizes the dominant types of distribution shift, their materials-specific instantiations, and associated epistemic consequences.

Table 1. Distribution shift types in materials AI and their epistemic implications

Shift type	Formal definition	Materials-specific manifestation	Representative examples	Epistemic consequence for claims
Covariate shift	Change in input distribution P(X) with invariant P(Y)	X	Compositional or structural bias across databases	Overrepresentation of simple oxides in the Materials Project relative to ICSD
Label shift	Change in output distribution P(Y) with invariant P(X)	Y	Property measurements under differing protocols	Tensile strength measured at different strain rates or temperatures
Concept shift	Change in conditional mapping P(Y)	X	Physics-driven evolution of structure–property relations	Phase stability changes under pressure or thermal cycling
Instrumental shift	Change induced by measurement modality	Instrument- or preprocessing-dependent artifacts	SEM intensity variation affecting microstructure classification	Spurious feature–property correlations
Environmental shift	Contextual variation external to material	Operational or deployment conditions	Humidity effects in photovoltaic efficiency prediction	Overconfident extrapolation beyond tested regimes

Data bias: Sources and epistemic consequences

Beyond distributional mismatch, materials informatics is shaped by entrenched forms of data bias arising from historical trajectories, methodological conventions, and broader socioeconomic forces. Existing datasets disproportionately emphasize oxide ceramics, low-complexity crystalline phases, and materials aligned with established industrial applications, while systematically underrepresenting disordered systems, metastable phases, and high-entropy compositions [17]. Selection bias further compounds this imbalance: published datasets tend to record only successful syntheses and stable outcomes, introducing survivorship bias that obscures failure modes and exploration boundaries [18].

Measurement infrastructures introduce additional layers of bias. Instrument-specific artifacts—such as variations in scanning electron microscopy (SEM) intensity, detector sensitivity, or preprocessing pipelines—can imprint systematic distortions on microstructural representations used for learning tasks [32]. When absorbed into training data, these distortions are reified by AI models as spurious regularities rather than contingent measurement effects.

The epistemic consequences of such biases are profound. Models trained on skewed datasets internalize distorted priors, generating predictions that appear statistically confident yet remain tethered to artifacts of data production rather than underlying material reality [32]. Under distribution shift, these distortions are amplified, yielding claims that lack robustness across contexts. As a result, materials AI risks producing epistemically misleading outputs—claims that reflect the structure of datasets rather than the structure of the material world—thereby undermining the scientific legitimacy of AI-mediated inference [31].

Epistemic values in science and AI

Philosophical analyses of scientific practice distinguish epistemic values—such as truth, consistency, explanatory adequacy, and predictive reliability—from non-epistemic values tied to social utility, economic priorities, or institutional incentives [21]. In AI-augmented science, this distinction becomes operationally salient: values are embedded not only in interpretive judgments but in upstream decisions regarding data curation, model architecture, loss functions, and evaluation metrics [25, 26].

In materials informatics, value commitments manifest through preferences for computational tractability, data availability, and commercial relevance, often privileging materials classes amenable to high-throughput simulation or aligned with near-term technological goals [27]. Distribution shifts intensify these tensions. Models optimized for in-distribution accuracy may implicitly reinforce non-epistemic values encoded in training data, such as industrial relevance or historical research focus, at the expense of epistemic virtues like generality or causal insight [28, 29].

Recent scholarship calls for value-sensitive design principles in materials AI, advocating transparency, reflexivity, and alignment with scientific norms of justification and reproducibility [30, 31]. Nevertheless, these efforts remain largely programmatic. A unified theoretical account that explicitly connects distribution shift, data bias, and value embedding to the epistemic status of AI-generated claims in materials science has yet to be articulated.

Synthesis: The need for a conceptual theory

Taken together, the existing literature addresses fragments of a shared problem. Materials informatics research documents methodological advances and application successes; distribution shift theory elucidates formal failure modes; bias analyses expose structural distortions in data ecosystems; and philosophy of science interrogates the role of values in knowledge production. What remains missing is a conceptual synthesis that theorizes how distribution shifts operating under biased, value-laden data regimes systematically undermine the credibility of AI-mediated material claims—and, critically, how robustness might be redefined under these conditions.

This gap motivates the present framework. By integrating insights from machine learning theory, materials informatics practice, and epistemic analysis, the proposed conceptual structure reframes robustness not as a purely statistical property but as an epistemic achievement contingent on bias awareness, value alignment, and shift-sensitive inference. In doing so, it provides a foundation for re-evaluating how materials AI systems generate, justify, and sustain scientific claims.

Proposed conceptual framework: Epistemic Robust Inference under Shift (ERIS)

The proposed framework— epistemic robust inference under shift (ERIS)—reconceptualizes materials AI inference as a hierarchical epistemic process rather than a flat mapping from inputs to predictions. Departing from conventional accuracy-centric paradigms, ERIS frames inference as the structured transformation of heterogeneous data into qualified scientific claims that remain defensible under plausible distributional shifts. Robustness, in this view, is not a statistical artifact but an epistemic achievement.

ERIS is organized into four interdependent layers:

(1) Shift ontology,

(2) Value alignment,

(3) Counterfactual stabilization, and

(4) Claim formulation.

These layers operate sequentially while remaining tightly coupled through feedback loops, ensuring that downstream claims are continuously revised in light of distributional uncertainty, value commitments, and domain constraints.

Layer 1: Shift ontology

Inference within ERIS begins with the explicit articulation of a distributional shift. Rather than treating shifts as ex post performance anomalies, this layer constructs a shift ontology that maps observed data distributions to underlying material generative processes, including synthesis pathways, characterization modalities, and environmental conditions.

Shifts are classified along materials-specific axes—compositional, structural, environmental, and instrumental—thereby situating statistical deviations within physically interpretable contexts. This ontology functions as a diagnostic scaffold, revealing covariances between learned descriptors and latent causal factors. By embedding shift awareness upstream, ERIS prevents the conflation of distributional artifacts with intrinsic material behavior.

Layer 2: Value alignment

The second layer addresses the normative structure of inference by explicitly aligning epistemic values—such as generalizability, mechanistic fidelity, and explanatory coherence—with data representations and modeling choices. Non-epistemic influences, including commercial prioritization or dataset availability biases, are surfaced and critically examined rather than implicitly absorbed.

Alignment is operationalized through value-weighted reparameterization of representations, privileging invariant and physically grounded features (e.g., symmetry-preserving descriptors, conservation-consistent embeddings) over shift-sensitive correlates. Domain theory thus acts as a constraint on representation learning, ensuring that inference remains anchored to scientific norms rather than opportunistic correlations.

$$f^* = Arg_{min} f \ [\ L_{pred(f;\ X,\ Y)} + \ \ \lambda \ L \ epistemic(f) \ ] \quad (5)$$

where L_{pred} denotes predictive loss, $L_{epistemic}$ encodes penalties for violations of epistemic values (e.g., mechanistic inconsistency, lack of invariance), and λ represents the weighting of epistemic commitments within the inference process. This formulation is interpretive rather than prescriptive, illustrating how value alignment constrains model optimization beyond accuracy alone.

Layer 3: Counterfactual stabilization

Robustness in ERIS is achieved through counterfactual stabilization, wherein claims are evaluated across families of plausible alternative distributions consistent with known material physics. Unlike conventional data augmentation, this layer emphasizes conceptual invariance: a claim must remain stable under perturbations that reflect realistic variations in synthesis, environment, or measurement.

Uncertainty is framed epistemically—as the degree of claim fragility across counterfactual scenarios—rather than solely probabilistically. This shift reframes uncertainty from a numerical confidence interval into a structured assessment of inferential resilience.

Layer 4: Claim formulation

The final layer translates stabilized inferences into modulated scientific claims. Outputs are not point predictions but qualified assertions explicitly conditioned on identified shifts, value alignments, and stabilization results. Claims take the form of epistemically annotated statements—for example, “Property X is robust under compositional shift within range Y, conditional on synthesis pathway Z.”

By embedding qualifiers directly into outputs, ERIS resists overconfidence and aligns AI-mediated inference with norms of scientific accountability and transparency. As shown in **Figure 1**, the epistemic flow in materials AI proceeds not through a linear data pipeline, but through a layered process of shift ontology, value alignment, and counterfactual stabilization, culminating in the formulation of scientifically qualified claims.

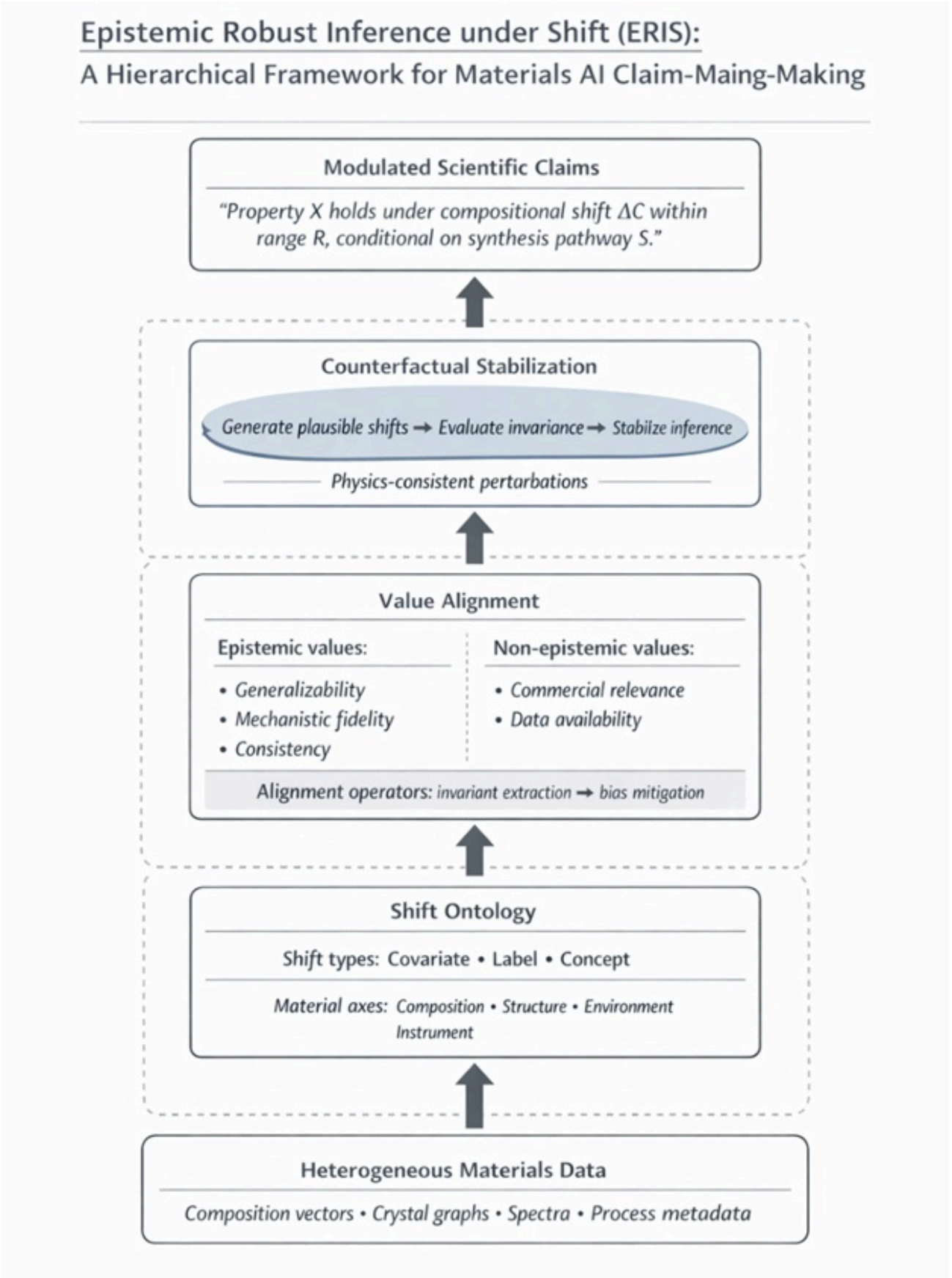
Epistemic Robust Inference under Shift (ERIS):
A Hierarchical Framework for Materials AI Claim-Maing-Making
Modulated Scientific Claims
"Property X holds under compositional shift ΔC within range R, conditional on synthesis pathway S."
Counterfactual Stabilization
Generate plausible shifts → Evaluate invariance → Stabilze inference
Physics-consistent pertarbations
Value Alignment
Epistemic values:
• Generalizability
• Mechanistic fidelity
• Consistency
Non-epistemic values:
• Commercial relevance
• Data availability
Alignment operators: invariant extraction → bias mitigation
Shift Ontology
Shift types: Covariate • Label • Concept
Material axes: Composition • Structure • Environment
Instrument
Heterogeneous Materials Data
Composition vectors • Crystal graphs • Spectra • Process metadata

Figure 1. Epistemic robust inference under shift (ERIS): A hierarchical framework for materials AI claim-making

ERIS thus provides a conceptual lens for robust inference, ensuring materials AI claims transcend correlational artifacts to embody scientific rigor. While **Figure 1** visualizes ERIS as an epistemic workflow, **Table 2** explicates how each layer addresses specific inferential failure modes induced by distribution shift, clarifying the framework's role in stabilizing scientific claims.

Table 2. Mapping ERIS layers to distribution-shift–induced failure modes and epistemic corrective functions.

ERIS layer	Primary inferential risk addressed	Typical failure without this layer	Epistemic corrective introduced	Effect on scientific claims
Shift ontology	Unarticulated distributional mismatch	Conflation of dataset artifacts with physical causation	Explicit mapping between data shifts and material generative processes	Prevents misattribution of correlations as intrinsic properties
Value alignment	Bias amplification under shift	Optimization for convenience or commercial relevance	Prioritization of epistemic values and domain invariants	Enhances generalizability and mechanistic fidelity
Counterfactual stabilization	Fragility under plausible perturbations	High in-distribution accuracy with out-of-support collapse	Evaluation across physics-consistent counterfactual shifts	Elevates resilience and epistemic warrant
Claim formulation	Overconfident point predictions	Unqualified assertions detached from the inference context	Modulated, shift-conditioned scientific claims	Restores epistemic humility and accountability

Conceptual contribution

Collectively, ERIS reframes materials AI from a predictive instrument into an epistemic system designed for responsible claim-making. Robust inference emerges not from isolated model performance but from the integration of shift awareness, value sensitivity, and counterfactual reasoning. In doing so, ERIS provides a principled foundation for scientific inference under distributional uncertainty.

Propositions

Drawing on the ERIS framework, the following propositions articulate its epistemic implications for materials AI. These propositions are deductive extensions of the framework's layers and serve as conceptual anchors rather than empirically testable hypotheses.

Proposition 1: Shift ontology as an epistemic prerequisite

Explicit delineation of distributional shift is a necessary epistemic condition for valid materials inference. Without mapping shifts to material generative processes, AI outputs risk reifying dataset artifacts as physical causation [1, 13]. In alloy design, for example, unmodeled compositional covariate shifts can yield spurious claims of enhanced ductility that reflect database bias rather than intrinsic behavior [24].

Proposition 2: Value alignment mitigates bias propagation

Epistemic values must actively guide representation alignment to counteract bias-induced shifts. When non-epistemic values dominate data curation, inferential fragility is amplified, particularly for underrepresented material classes such as

amorphous or metastable systems [17, 23]. Value-weighted invariants enhance claim generalizability by prioritizing mechanistic fidelity over correlational convenience.

Proposition 3: Counterfactual stabilization ensures claim resilience

Claims stabilized across counterfactual distributions consistent with material physics possess higher epistemic warrant than those justified by in-distribution metrics alone [11]. In photovoltaic materials prediction, counterfactuals that incorporate environmental variability (e.g., humidity or temperature shifts) yield more resilient estimates of efficiency and stability [15].

Proposition 4: Modulated claims express epistemic humility

Scientific claims generated by materials AI should embed qualifiers derived from shift ontology and value alignment, transforming predictions into conditioned assertions. Unqualified outputs foster overconfidence, undermining the normative standing of AI-assisted claims within scientific discourse [4].

Proposition 5: Robustness is an emergent epistemic property, not a model attribute

Robustness in materials AI should be understood as an emergent property of the inference process rather than an intrinsic characteristic of a trained model. Within ERIS, robustness arises from the interaction of shift ontology, value alignment, and counterfactual stabilization, rather than from architectural complexity or benchmark performance alone. Models that exhibit high in-distribution accuracy may nevertheless yield epistemically fragile claims when embedded in biased data regimes or deployed under unarticulated shifts [34]. This proposition asserts that robustness is achieved only when inferential claims remain stable across epistemically plausible distributions, reframing robustness as a systems-level epistemic accomplishment rather than a statistical metric.

Proposition 6: Epistemic feedback is necessary for sustained validity under iterative discovery

In iterative materials discovery workflows—such as active learning and autonomous experimentation—epistemic validity cannot be preserved without structured feedback between claims and upstream inferential layers. ERIS posits that claim formulation must recursively inform shift ontology and value alignment, enabling the system to recalibrate its assumptions as new data regimes emerge. Absent such epistemic feedback, early mischaracterizations of shift or value commitments become path-dependent, leading to progressive erosion of claim validity despite apparent model improvement [18]. This proposition maintains that sustained robustness under acceleration requires reflexive epistemic updating, not merely continual data accumulation.

Results and Discussion

The ERIS framework advances a conceptual reorientation of materials AI inference, positioning distribution shifts not as mere technical hurdles but as epistemic challenges intertwined with values and biases. By layering shift ontology, value alignment, counterfactual stabilization, and claim formulation, ERIS transcends performance-oriented paradigms and aligns AI with the rigorous claim-making traditions of materials science [2, 4, 23].

Conceptually, this framework addresses gaps in existing literature. While prior syntheses emphasize empirical adaptations to shifts [19, 20], ERIS foregrounds their ontological and axiological dimensions. For example, shift ontology explicates how generative processes—such as DFT approximations versus experimental noise—introduce covariances that undermine i.i.d. assumptions [7, 8, 30]. This resonates with philosophical critiques of epistemic values, where unexamined biases in data reflect societal priorities, potentially skewing AI toward economically viable materials [18].

Value alignment, in turn, conceptualizes representation as a normative act. Unlike invariant learning techniques that seek statistical independence [35], ERIS incorporates epistemic criteria such as explanatory power, ensuring that alignments

honor domain invariants (e.g., symmetry principles in crystal graphs) [25, 26]. This mitigates propagation of biases, such as survivorship effects in synthesis data, fostering more equitable scientific claims.

Counterfactual stabilization introduces a novel epistemic tool: reasoning under hypothetical shifts to probe the fragility of claims. This extends concept-shift analyses [12] by proposing that robustness derives from invariance to physically plausible perturbations, rather than solely from data augmentation [5, 6]. In practice, this could reconceptualize benchmarks in materials informatics, shifting from mean accuracy to epistemic uncertainty metrics [9, 10].

Finally, modulated claims embody epistemic humility, qualifying inferences with shift contexts to prevent spurious generalizations [14, 16]. This has broader implications for autonomous materials discovery, where diminished human oversight amplifies shift vulnerabilities [21, 22].

Limitations of ERIS include its conceptual abstraction, which remains to be formalized (e.g., via probabilistic graphical models). Nonetheless, it provides a scaffold for integrating shifts into epistemic workflows, potentially informing value-sensitive AI design in materials science [3, 13]. Future extensions could explore intersections with multi-scale modeling, where shifts cascade across atomic to macroscopic levels [24, 25]. Overall, ERIS advocates for an epistemically grounded materials AI that enhances the reliability of scientific claims amid data heterogeneity.

Conclusion

In summary, distribution shifts pose profound epistemic threats to scientific claims in materials AI, necessitating a conceptual framework like ERIS to foster robustness. By delineating shift ontologies, aligning values, stabilizing via counterfactuals, and formulating modulated claims, ERIS reconceptualizes inference as an epistemic endeavor. This theory bridges materials informatics with philosophical insights on values and bias, promoting claims that withstand distributional perturbations. Ultimately, it paves the way for trustworthy AI in materials science, ensuring inferences contribute meaningfully to knowledge advancement.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 May 2025 Revised: 01 Jul 2025 Accepted: 26 Aug 2025
Published online: 18 January 2026

References

Li C, Zheng K. Methods, progresses, and opportunities of materials informatics. InfoMat. 2023;5(2):e12345.

Sivan D, Satheesh Kumar K, Abdullah A, Raj V. Advances in materials informatics: A review. J Mater Sci. 2024;59(3):1123-45.

Wang Z, Chen A, Tao K, Han Y, Li J. MatGPT: A vane of materials informatics from past, present, to future. Adv Mater. 2024;36(6):2306733.

Chaikittisilp W, Yamauchi Y, Ariga K. Material evolution with nanotechnology, nanoarchitectonics, and materials informatics. Adv Mater. 2022;34(15):2105678.

Ling C. A review of the recent progress in battery informatics. npj Comput Mater. 2022;8(1):33.

Jha D, Gupta V, Ward L, Yang Z, Wolverton C. Enabling deeper learning on big data for materials informatics applications. Sci Rep. 2021;11(1):4244.

Sha W, Li Y, Tang S, Tian J, Zhao Y, Guo Y, et al. Machine learning in polymer informatics. InfoMat. 2021;3(4):353-61.

Oaki Y, Igarashi Y. Materials informatics for 2D materials combined with sparse modeling and chemical perspective. Bull Chem Soc Jpn. 2021;94(3):845-56.

Choudhary K, DeCost B, Chen C, Jain A. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8(1):59.

Wang Z, Sun Z, Yin H, Liu X, Wang J, Zhao H. Data-driven materials innovation and applications. Adv Mater. 2022;34(22):2104113.

Hu J, Liu D, Fu N, Dong R. Realistic material property prediction using domain adaptation based machine learning. Digit Discov. 2024;3(1):45-56.

Zou D, Liu S, Miao S, Fung V. GeSS: Benchmarking geometric deep learning under scientific applications with distribution shifts. Adv Neural Inf Process Syst. 2024;37.

Jablonka KM, Ongari D, Moosavi SM, Smit B. Big-data science in porous materials: Materials genomics and machine learning. Chem Rev. 2020;120(16):8066-129.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50:71-103.

Lu T, Li M, Lu W, Zhang TY. Recent progress in the data-driven discovery of novel photovoltaic materials. J Mater Inform. 2022;2(1):11.

Liu Y, Yang Z, Zou X, Ma S, Liu D. Data quantity governance for machine learning in materials science. Natl Sci Rev. 2023;10(5):nwac269.

Sarkisov L, Bueno-Perez R, Sutharson M. Materials informatics with PoreBlazer v4.0 and the CSD MOF database. Chem Mater. 2020;32(23):9849-67.

Saal JE, Oliynyk AO, Meredig B. Machine learning in materials discovery: Confirmed predictions and their underlying approaches. Annu Rev Mater Res. 2020;50:49-69.

Li Q, Miklaucic N, Hu J. Out-of-distribution materials property prediction using adversarial learning based fine-tuning. arXiv preprint arXiv:2408.09297. 2024.

Unni R, Zhou M, Wiecha PR, Zheng Y. Advancing materials science through next-generation machine learning. Curr Opin Solid State Mater Sci. 2024;30:101157.

Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. Nat Rev Mater. 2021;6(8):655-78.

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):42.

Rowan-Robinson RM, Leong Z, Carpio S, Oh C. Material informatics for functional magnetic material discovery. AIP Adv. 2024;14(1):015301.

Pilania G. Machine learning in materials science: From explainable predictions to autonomous design. Comput Mater Sci. 2021;193:110360.

Flovik V. Quantifying distribution shifts and uncertainties for enhanced model robustness in machine learning applications. arXiv preprint arXiv:2405.01978. 2024.

Xi W, Lee YJ, Yu S, Chen Z, Shiomi J, Kim SK. Ultrahigh-efficient material informatics inverse design of thermal metamaterials for visible-infrared-compatible camouflage. Nat Commun. 2023;14(1):7357.

Noh J, Gu GH, Kim S, Jung Y. Machine-enabled inverse design of inorganic solid materials. Chem Sci. 2020;11(19):4871-81.

Badini S, Regondi S, Pugliese R. Unleashing the power of artificial intelligence in materials design. Materials. 2023;16(17):5927.

Alvarado R. AI as an epistemic technology. Sci Eng Ethics. 2023;29(5):35.

Russo F, Schliesser E, Wagemans J. Connecting ethics and epistemology of AI. AI Soc. 2024;39(2):361-83.

Coeckelbergh M. Democracy, epistemic agency, and AI: Political epistemology in times of artificial intelligence. AI Ethics. 2023;3(1):99-107.

Tolsgaard MG, Boscardin CK, Park YS. The role of data science and machine learning in health professions education: Practical applications, theoretical contributions, and epistemic beliefs. Adv Health Sci Educ. 2020;25(5):1057-86.

Hancox-Li L, Kumar IE. Epistemic values in feature importance methods: Lessons from feminist epistemology. Proc ACM Conf Fair Account Transpar. 2021:817-26.

Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics. 2021;47(5):329-35.

Carabantes M. Black-box artificial intelligence: An epistemological and critical analysis. AI Soc. 2020;35(2):309-17.