ORIGINAL RESEARCH **Open access**

Bias in Materials Datasets without Datasets: A Conceptual Account of How Bias Enters Before Any Modeling

Lucas Pereira[1*], Bruno Martins[1], Renata Azevedo[2], Pedro Costa[1]

Abstract

Artificial intelligence (AI) in materials science is often treated as a pipeline in which bias primarily emerges during model training, evaluation, or deployment. This framing is structurally incomplete. Many distortions later labeled as "dataset bias" are already introduced before any dataset is formally assembled, labeled, cleaned, or modeled. This conceptual manuscript advances a theory-first account of pre-dataset bias: systematic misrepresentation that originates upstream of data tables through decisions about what counts as a material instance, a property definition, a valid operating regime, and an actionable target. We argue that early bias is not merely a statistical artifact but an epistemic and procedural commitment that shapes what becomes observable, measurable, and publishable. We introduce a novel framework—the bias before data (BBD) framework—which decomposes pre-dataset bias into five coupled mechanisms: problem framing bias, regime availability bias, measurement–proxy bias, curation–visibility bias, and legitimacy bias. BBD provides a structured vocabulary for identifying where bias enters, why it persists despite technical improvements, and how it constrains the legitimacy of scientific claims even when predictive performance appears strong.

Keywords Materials informatics, Distribution shift, Dataset bias, Measurement bias, Proxy targets, Scientific validity

*Correspondence:
Lucas Pereira
lucas.pereira@gmail.com
[1] Department of Materials Modeling and AI Systems, Faculty of Engineering, University of Minho, Braga, Portugal
[2] Department of Intelligent Materials Analytics, Faculty of Engineering, University of Porto, Porto, Portugal

Introduction

Bias in artificial intelligence is frequently discussed as a problem of skewed samples, label noise, missingness, and uneven coverage. In materials science, the bias discussion often follows a similar pattern: the training set under-represents certain chemistries; benchmarks privilege particular structure families; property values depend on inconsistent protocols; datasets are patched together from incompatible sources; or models fail under distribution shift. These are real concerns, and they matter for scientific reliability and decision-making in materials design [1–4]. However, the prevailing narrative often assumes that bias begins once a dataset exists, and that corrective effort should focus on balancing, cleaning, and reweighting that dataset. This manuscript argues that such a narrative is structurally incomplete: in materials AI, much of what later appears as "dataset bias" is introduced earlier—before any spreadsheet, repository, or benchmark is created.

Materials datasets do not arise spontaneously from the physical world. They are the output of a chain of choices: what phenomena are considered "in scope," which materials families are worth testing, which property definitions are accepted as canonical, which regimes are practically measurable, which instruments and protocols are treated as authoritative, and which outcomes are publishable or valuable to collect. These upstream decisions are not neutral background conditions.

They actively shape what becomes observable, recordable, and actionable for downstream AI [2, 5–7]. In this sense, "dataset bias" in materials informatics should be understood not only as a sampling distortion inside a dataset, but also as a pre-dataset distortion in the construction of the epistemic space from which any dataset is drawn.

This point is not merely philosophical. Contemporary materials AI increasingly operates as a design partner, screening candidate compositions, proposing microstructures, optimizing processing conditions, and accelerating discovery through surrogate modeling and active learning [3, 4, 8, 9]. Such workflows depend on the assumption that data encode stable, comparable meaning across contexts—yet material properties are often measurement-defined, protocol-dependent, and sensitive to hidden variables such as impurities, processing history, morphology, and test conditions [10–12]. A model can therefore perform well while still inheriting structural blind spots: it may generalize over what is frequent and measurable rather than what is physically comprehensive or decision-relevant. This mismatch becomes acute when AI outputs are treated as evidence for claims about mechanisms, optimality, feasibility, or deployability. In those cases, bias is no longer merely an error source; it becomes a threat to the legitimacy of the claim.

We propose that the most consequential biases in materials datasets stem from upstream constraints and incentives that determine what is collected in the first place. High-throughput experiments and automated computational workflows can increase volume, but volume does not guarantee representativeness. If collection is driven by convenience, institutional capability, historically popular systems, or funding priorities, then coverage expands unevenly, leaving the dataset dense in some regions and nearly empty in others [8, 13, 14]. Even when community databases enable scale through shared repositories and standardized workflows, the data remain bounded by the regimes selected, the definitions accepted, and the measurement proxies treated as equivalents of the underlying scientific construct [5, 6, 15]. As a result, bias is not merely something the model "learns"; it is something the scientific process "permits."

Recent developments in representation learning and foundation-model-like approaches have further intensified this issue. When models learn rich embeddings from large, heterogeneous corpora of materials information—compositions, structures, text-mined knowledge, or computed properties—the learned representation can appear broadly transferable while still encoding historical and disciplinary biases about which families, properties, and regimes are "typical" [16–18]. Such systems can deliver impressive predictive performance, but their outputs may remain anchored to the visibility structure of the upstream data: what was measured, what was archived, what was standardized, and what was recognized as a legitimate target [6, 17, 19]. If a field systematically under-measures certain failure modes, toxicity outcomes, long-term degradation pathways, or edge-case regimes, then the model's competence will mirror that absence—not as random noise, but as an organized silence.

This manuscript, therefore, develops a conceptual account of bias before data. We define pre-dataset bias as a systematic distortion introduced by upstream decisions that shape observability, collectability, comparability, and legitimacy, before any modeling step. Pre-dataset bias differs from traditional dataset bias in three ways. First, it is not reducible to statistical imbalance; it is rooted in how the scientific world is discretized into "instances" and "targets." Second, it is resistant to common fixes (reweighting, augmentation, balancing) because it may reflect what was never measured or never made visible. Third, it often remains undetected because it hides behind performance metrics evaluated on similarly biased benchmarks [1, 2, 9]. These properties make pre-dataset bias a foundational concern for applied AI in materials science, especially when systems are used to inform design decisions with safety, sustainability, and cost consequences.

To address this, we introduce a novel theoretical framework—the bias before data (BBD) framework—that maps how bias enters before any dataset exists. BBD is not a checklist of "best practices," nor a repackaging of existing governance guidelines. Instead, it is a causal-structural decomposition of upstream bias mechanisms that produce downstream dataset artifacts. The framework distinguishes five coupled sources: (1) problem framing bias, where research questions encode implicit priorities and omit competing objectives; (2) regime availability bias, where feasible measurement and computation constrain what is considered learnable; (3) measurement–proxy bias, where operational targets substitute for scientific constructs; (4) curation–visibility bias, where publishing, archiving, and database practices determine what

becomes countable; and (5) legitimacy bias, where community norms decide what claims are acceptable, thereby shaping what data are worth collecting [5, 6, 10, 15].

The contribution of this paper is conceptual: it provides a new vocabulary and structure for diagnosing upstream bias in materials AI. By making pre-dataset bias explicit, BBD enables more defensible interpretation of model results, more transparent articulation of validity scope, and more responsible use of AI outputs in design-centric settings where errors carry downstream consequences [2, 4, 11, 12]. Importantly, BBD does not claim that bias can be eliminated; rather, it argues that bias must be typed, located, and bound to the legitimacy of claims, so that predictive success is not mistaken for scientific completeness or decision readiness.

Theoretical Background & Literature Synthesis

Materials informatics as a socio-technical data construction pipeline

Materials informatics is frequently described as an integration of physics, chemistry, data science, and domain expertise, enabling accelerated mapping from composition and structure to properties and performance [3, 4]. Yet at a deeper level, materials informatics is also a data construction enterprise: it depends on translating physical systems into standardized descriptors, translating experimental outcomes into scalar targets, and translating diverse records into interoperable forms that can be learned by algorithms [5, 6]. This translation is unavoidably selective. Not all phenomena can be measured at scale, not all regimes can be safely explored, and not all properties can be represented without loss of context. Therefore, what counts as a “materials dataset” is already the outcome of selective constraints and normalization choices.

Community databases and pipeline toolchains have dramatically improved accessibility and standardization in recent years. Domain repositories, programmatic workflows, and descriptor libraries have enabled reproducibility and large-scale data-driven modeling [5, 20]. At the same time, the growth of computational databases—especially those derived from high-throughput density functional theory (DFT)—has created a dominant modality of “available” materials data, which shapes research direction by making some properties easy to obtain and others comparatively scarce [13, 14]. Even when high-throughput data are accurate within their stated approximations, their presence can bias the field toward computationally convenient objectives and away from difficult-to-measure operational constraints.

Benchmark culture and the illusion of neutral coverage

A central driver of materials AI progress is benchmarking: standardized splits, leaderboards, and common predictive tasks facilitate rapid comparison across architectures and representations [6, 9]. However, benchmark culture can also conceal structural bias by defining success against targets that are historically available, easy to compute, or popular within the community. In such cases, models become optimized for benchmark-likeness rather than for decision legitimacy under real constraints. This concern becomes sharper when benchmark properties are treated as proxies for downstream goals (e.g., stability proxies for device lifetime, idealized strength proxies for deployable performance) without explicitly stating what is lost in the proxy mapping [10–12].

Moreover, benchmark generalization may be circular: training and evaluation often share the same underlying construction biases. When both sets reflect the same visibility structure—same measurement protocols, same favored chemistries, same publication incentives—good test performance may only confirm internal consistency rather than robustness to regime change. This is closely related to broader concerns about distribution shift and out-of-distribution generalization, which have become increasingly recognized as central limitations in applied AI systems [1, 2, 9]. In materials science, distribution shifts can arise from changes in synthesis routes, microstructural states, measurement conditions, and compositional constraints, often without explicit labels in datasets [11, 12]. These shifts are not “edge cases”; they are normal conditions of scientific practice.

Representation choices as epistemic commitments

A distinctive feature of materials AI is its reliance on representations that convert physical systems into machine-readable forms, such as compositional vectors, structure graphs, local environment descriptors, or learned embeddings [16, 20, 21]. Representation is often framed as a technical step—improving expressivity, reducing dimensionality, or achieving invariance. Yet representation is also an epistemic commitment: it encodes assumptions about what aspects of a material are causal, relevant, stable, or comparable across contexts.

Descriptor libraries and representation toolkits have made representation engineering more systematic and accessible [20, 22]. However, the ease of generating features can create a subtle bias: what is representable becomes what is treated as real in the modeling worldview. For example, composition-only representations may implicitly treat processing and microstructure as noise, whereas structure-only representations may omit synthesis feasibility and defect landscapes that dominate real performance [11, 12]. Learned representations can mask these omissions by producing smooth latent spaces that appear general, even if the latent structure reflects data availability rather than physical completeness [16, 17]. In such settings, bias is not only an imbalance of samples; it is a mismatch between the representation's semantics and the downstream claim's meaning.

Property targets, proxies, and measurement-defined reality

Material properties are not always direct observables. Many targets used in materials AI are operational definitions produced by protocols: hardness depends on indenter type and load; corrosion resistance depends on environment and test duration; battery performance depends on cycling procedures; and mechanical properties depend on specimen preparation and strain rate [10–12]. This makes the target variable itself a potential bias channel: two measurements with the "same name" may not mean the same scientific construct.

As a result, models often learn from mixed definitions that have been collapsed into a single scalar label. This introduces bias via semantic compression: variability due to protocol differences becomes hidden within the label distribution and can be incorrectly attributed to intrinsic material behavior. Recent work emphasizing uncertainty quantification and careful dataset construction in materials ML highlights that predictive confidence is not only a modeling output; it is a function of how targets are defined and how noise is introduced upstream [2, 23]. If the target is a proxy for a deeper construct (e.g., DFT formation energy as a proxy for stability), then bias can be introduced by conflating proxy validity with scientific validity [13, 14, 24].

Database visibility, negative results, and missingness by design

Bias in materials data is also created by what is not recorded. Negative results, failed synthesis attempts, and unstable processing regimes are often absent from published datasets, despite being highly informative for design decisions. This missingness is not random; it is structured by publication incentives, reporting norms, and database schema constraints [5, 6]. Consequently, AI systems may learn that certain chemistries or processing windows "do not exist," when in reality they exist as unreported failures or unarchived trials.

The emergence of large open databases and community-driven data infrastructures has improved accessibility. Still, it also creates a filter: what is accepted into a database becomes what can be learned. In practice, database inclusion criteria often privilege clean, standardized entries, which may systematically exclude messy but realistic conditions, thereby producing an optimism bias in model outputs [6, 15]. This can be especially problematic when AI is used for recommending actionable candidates: the model may overestimate feasibility because infeasibility evidence is invisible upstream.

Generalization, uncertainty, and the limits of "More Data"

A common response to dataset bias is to collect more data. Yet “more data” does not necessarily repair structural distortions if the new data are collected under the same upstream constraints and incentives. In such cases, scaling can intensify bias by making the dominant regimes even denser and the underrepresented regimes still sparse. This is one reason uncertainty-aware modeling has become central in materials ML: it provides tools to express where predictions may not be warranted [2, 23]. However, uncertainty alone cannot diagnose why regimes are missing, why proxies were used, or why certain constraints were never represented. Uncertainty can flag fragility, but it cannot reconstruct absent causal structure.

Recent discussions of robust evaluation, shift-aware learning, and scientifically grounded ML have emphasized that generalization must be interpreted relative to the data-generating process and the boundaries of validity [1, 2, 9]. In materials informatics, this implies that bias analysis must begin upstream of modeling, because the most important limitation may be that the space of possible observations was structured in a biased way before the dataset was assembled.

Proposed conceptual framework

The Bias Before Data (BBD) framework: Core idea

The BBD framework explains how bias can enter materials datasets before any modeling step—before splitting into train/test, before feature engineering, and even before a dataset is formally defined. The central thesis is:

In materials AI, many “dataset biases” are downstream manifestations of upstream decisions that shape what becomes measurable, recordable, comparable, and legitimate to claim.

Accordingly, BBD treats pre-dataset bias as a structured cascade: upstream commitments determine the observational field; the observational field determines what is collectible; collectability determines what is curated; curation determines what becomes a dataset; and the dataset determines what models can learn and what claims appear justified.

BBD decomposes pre-dataset bias into five coupled mechanisms, each representing a distinct upstream decision class:

Module 1 — Problem Framing Bias (PFB): Bias from “What counts as the question?”

Definition

Distortion is introduced when research questions encode implicit priorities and omit competing values.

Mechanism

The chosen objective (e.g., maximize strength, minimize bandgap error, improve catalytic activity) defines what is collected, while excluded objectives (toxicity, durability, recyclability, manufacturability) become invisible.

Key effect

The dataset becomes “complete” only with respect to the framed goal, enabling models that are accurate but normatively narrow [4, 6, 12].

Module 2 — Regime Availability Bias (RAB): Bias from feasible measurement and computation

Definition

Distortion is created when the accessible regime of measurement/computation constrains what the dataset can represent.

Mechanism

Data arise from what labs can synthesize, what instruments can measure reliably, and what computations can execute at scale.

Key effect

The dataset reflects capability geography (what is feasible) rather than *scientific geography* (what matters) [13–15].

Module 3 — Measurement–Proxy Bias (MPB): Bias from substituting targets

Definition

Distortion is created when measured or computed proxies are treated as equivalent to a scientific construct.

Mechanism

Operational metrics replace underlying constructs, conveying a false equivalence: “property value” becomes a stand-in for real performance across contexts.

Key effect

Models learn proxy-optimized behavior, while scientific meaning silently shifts underneath [10–12, 24-28].

Module 4 — Curation–Visibility Bias (CVB): Bias from what becomes countable

Definition

Distortion is introduced through selective archiving, publishing, database inclusion rules, and schema design.

Mechanism

“Clean” data survive, negative/failed trials vanish, and heterogeneous records are excluded because they are hard to standardize.

Key effect

The dataset systematically over-represents success conditions and under-represents feasibility boundaries [5, 6].

Module 5 — Legitimacy Bias (LB): Bias from what claims are allowed

Definition

Distortion arises when community legitimacy norms determine which data are worth collecting.

Mechanism

If the field rewards novelty, certain families become oversampled; if it rewards performance metrics, durability conditions remain scarce.

Key effect

“What exists” in the dataset becomes a reflection of “what is rewarded,” and models inherit this as reality [6, 9, 15, 29-31].

Figure 1 shows how upstream decisions create downstream dataset bias.

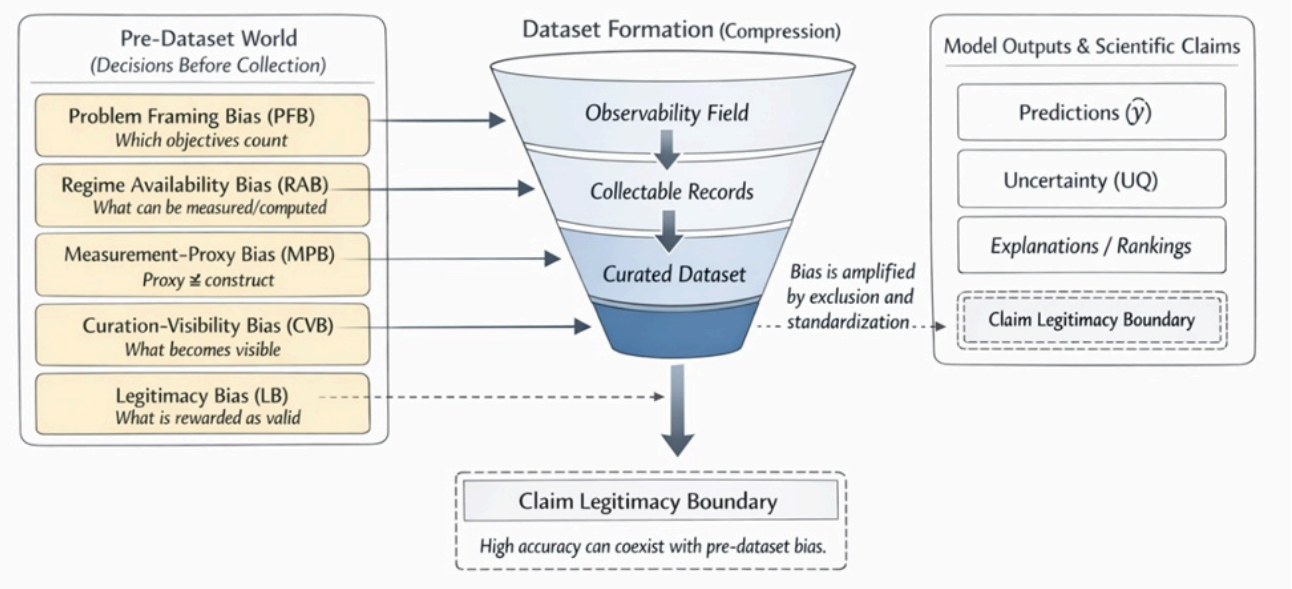

Figure 1. Bias before data (BBD): how upstream decisions create downstream dataset bias

Mechanisms of pre-dataset bias

The bias before data (BBD) Framework claims that what later appears as “dataset bias” often originates before dataset assembly. This section specifies how upstream distortions arise as repeatable mechanisms rather than accidental imperfections. The goal is to provide a scientific account of pre-dataset bias as a structural feature of modern materials knowledge production under high-throughput, database-centered, and benchmark-driven regimes [3–6].

Selective observability: Bias from what can be seen

Every materials dataset is limited by observability, defined here as the intersection of (i) experimental feasibility, (ii) computational tractability, (iii) measurable operational definitions, and (iv) institutional capacity. Observability is not merely a technical constraint; it is a boundary that shapes what becomes “real” to the learning system, because it can only learn from recorded instances [5, 6]. In computational pipelines, observability is often governed by model approximations and standardized workflows that determine which properties can be computed at scale and which cannot [7–9]. In experimental pipelines, observability depends on instrumentation, sample preparation, throughput limitations, and protocol variability that influence whether outcomes are comparable across studies [10–12].

A key consequence is that regions of materials space that are difficult to synthesize, expensive to characterize, unstable under ambient conditions, or hazardous to test become underrepresented before any sampling plan is defined. This produces regime availability bias (RAB) as a mechanism of epistemic narrowing: the dataset becomes a record of what is tractable rather than what is scientifically or technologically decisive [4, 5, 11]. Even when large repositories are used, their contents remain shaped by what was observable within existing research systems and incentives [6, 13].

Semantic freezing: Bias from converting concepts into labels

A second mechanism is semantic freezing, where a complex scientific construct (e.g., “stability,” “strength,” “activity,” “lifetime”) is converted into a dataset label that appears stable, scalar, and universally comparable. Yet many material properties are not entity-like objects; they are measurement-defined outcomes that depend on conditions and protocols [10–12]. Once the field adopts a label as a “target,” the definition becomes operationally fixed, even if the underlying meaning varies across contexts.

Semantic freezing is intensified when machine learning requires consistent targets across samples. Heterogeneous measurements are often merged into a unified label distribution to enable predictive tasks, which can quietly compress multiple meanings into one number. The resulting dataset supports prediction, but may not support interpretation, portability, or mechanistic inference because the target no longer corresponds to a single coherent construct [2, 11, 12]. This mechanism is the core of Measurement–Proxy Bias (MPB): models optimize for the proxy label while the scientific meaning drifts underneath.

In computational datasets, semantic freezing also appears when computed quantities become treated as definitive ground truth for physical behaviors, despite known limitations of approximations and missing real-world factors [7, 8, 14]. This does not invalidate computational data; rather, it locates its validity within defined boundaries. BBD emphasizes that bias enters when those boundaries are forgotten, even as the numbers remain the same.

Feasibility silence: Bias from missing negatives and missing constraints

A third mechanism is feasibility silence: the systematic absence of failed synthesis, unstable processing routes, negative experimental outcomes, and unpublishable intermediate trials. Such missingness is rarely random. It is produced by incentives that reward success, novelty, and clean comparability rather than failure documentation and constraint reporting [5, 6, 15]. Feasibility silence makes the dataset appear more optimistic than reality, particularly when models are used for recommendation and design.

This mechanism is central to curation–visibility bias (CVB). Visibility is not merely “what exists”; it is what is made countable. Database schema choices, inclusion criteria, and cleaning filters often exclude noisy, ambiguous, or condition-rich entries that would complicate learning but encode crucial feasibility boundaries [6, 13]. As a result, AI models trained on visible data can recommend candidates that are theoretically interesting but practically infeasible, not because the model is defective, but because evidence of infeasibility is absent upstream [11, 12].

Benchmark Lock-In: Bias from standardized tasks becoming reality

A fourth mechanism is benchmark lock-in, in which popular predictive tasks are treated as the field’s scientific core. Benchmarks are valuable for accelerating algorithmic progress and enabling fair comparison [6, 16]. However, when benchmark tasks dominate publication and attention, they can reshape what data gets collected, which properties are prioritized, and what forms of generalization are pursued.

Benchmark lock-in can create a feedback loop: the community collects data that supports benchmark tasks; benchmark tasks select models that perform well under the same construction constraints; and success encourages further data collection in the same directions. This produces Legitimacy Bias (LB): the dataset becomes aligned with what the community recognizes as legitimate achievement rather than with the full space of decision-critical constraints [6, 16]. Under this loop, “high accuracy” may represent mastery of the benchmark regime rather than scientific readiness for deployment decisions under shift [1, 2].

Bias accumulation: Why pre-dataset bias compounds across stages

A defining feature of pre-dataset bias is compounding. A small framing choice at the outset (e.g., optimizing a single performance metric) can propagate into regime selection, measurement design, and database inclusion, ultimately creating a dataset structurally incapable of supporting certain claims. This compounding resembles what has been discussed in broader ML reliability literature: errors under dataset shift are not only model failures; they are failures of alignment between training conditions and real conditions [1, 2].

BBD therefore reframes bias as an accumulated distortion of the observation-to-claim chain. The consequence is not merely reduced accuracy under shift; it is the production of scientifically persuasive outputs that exceed their legitimacy boundary [32-35].

Implication

Bias produces overclaiming pathways

In materials AI manuscripts, the most common harm of bias is not that predictions are "wrong," but that predictions are right for reasons that do not warrant the conclusion being stated. This distinction is crucial. A model can successfully predict an experimental label, while the label itself encodes proxy substitution, semantic compression, and missing contextual variables [10–12]. In such settings, accuracy is compatible with invalid scientific inference.

BBD identifies a systematic pathway:

1. Upstream distortion shapes what becomes visible and measurable,

2. The dataset encodes that partial world as "ground truth,"

3. Model learns consistent patterns,

4. Manuscript interprets patterns as scientific insight,

5. The scientific claim exceeds the dataset's warrant for construction.

This pathway is intensified by the success of modern representations and architectures that provide smooth latent structures and high benchmark performance [16–18]. Such systems can appear transferable even when their competence is anchored to the visibility regime.

The distinction between predictive utility and scientific legitimacy

Materials informatics often operates under two competing evaluation logics:

· Predictive utility: Does the model reproduce labels?

· Scientific legitimacy: Does the model justify a claim about the material world?

Predictive utility is necessary for screening and ranking, and it is a valid scientific achievement [3, 4]. However, scientific legitimacy requires that what is being predicted corresponds to the construct invoked in the conclusion. When proxy targets stand in for deeper constructs, legitimacy requires explicit boundary statements: what exactly is guaranteed, under which conditions, and what is not supported [2, 11, 12].

This problem becomes sharper when models are used beyond regression: e.g., candidate suggestion, inverse design, optimization, and discovery claims. In these settings, the model output is treated as action-relevant. Yet action legitimacy depends on whether feasibility constraints, stability regimes, and operational conditions are represented upstream [11, 12]. If they were excluded through CVB and RAB, then action guidance is structurally biased.

Bias and generalization under distribution shift

Distribution shift is a central limitation in applied ML. During a shift, uncertainty estimates may degrade, calibration may fail, and predictive confidence may become misleading [1, 2]. In materials science, shift can arise from changes in synthesis routes, microstructural regimes, measurement protocols, and operational environments—often without explicit metadata [10–12]. Pre-dataset bias makes these shifts more likely because the dataset's construction omits exactly the variables that drive them.

BBD emphasizes that "robustness" cannot be assessed only as a model property. It must be understood as a property of the observation system that generated the data. When the observation system is biased toward convenient regimes,

generalization claims are fragile by design.

Why bias is not solved by scaling alone

Recent work in representation learning and large models suggests that scale can improve performance and transferability [16–18]. Yet BBD argues that scale can also amplify pre-dataset bias. If new data are generated under the same observability constraints, measurement proxies, and visibility rules, then the dataset becomes larger but not more epistemically complete.

In other words, scale increases density within visible regimes, but does not automatically add missing regimes. This produces a coverage illusion: the dataset appears comprehensive because it is large, while its blind spots remain structurally stable. Thus, addressing pre-dataset bias requires explicit treatment of missing regimes and proxy meanings, not only larger corpora.

Limitations and Future Conceptual Directions

This manuscript is deliberately conceptual and therefore has three limitations that should be acknowledged as part of scholarly completeness. First, BBD does not claim to provide a numeric metric for bias severity. Pre-dataset bias is not always quantifiable because its core mechanism is the absence of regimes and meanings that never became data. The framework is therefore a theory of where bias enters and how it constrains claims, not a scoring system.

Second, BBD does not argue that pre-dataset bias is fully avoidable. Many distortions arise from unavoidable constraints: safety restrictions, instrumentation limits, cost barriers, and genuine scientific difficulties. The purpose of BBD is not to accuse data producers, but to make upstream selectivity legible and therefore governable in how claims are written.

Third, BBD does not replace existing concerns about dataset imbalance, noise, and uncertainty. Instead, it reframes them as downstream symptoms that require upstream interpretation. Uncertainty quantification and robust evaluation remain essential, but they operate within the visibility structure that pre-dataset choices create [1, 2, 23]. Where that structure omits decisive regimes, uncertainty can flag fragility, but cannot recover missing knowledge.

Future conceptual work can extend BBD in three directions. (i) Bias interaction theory: formalizing how the five modules compound nonlinearly in real pipeline ecosystems. (ii) Claim-typing extensions: mapping which categories of scientific claims are most vulnerable to each bias module (predictive, mechanistic, prescriptive). (iii) Institutional ecology: analyzing how funding, publication norms, and benchmark culture create predictable visibility structures that produce recurring bias patterns in different materials subfields [6, 15, 16].

Conclusion

This conceptual manuscript argued that bias in materials datasets can arise before they are sampled, cleaned, and modeled. The dominant assumption that bias is primarily a modeling-stage artifact is incomplete for materials AI, where targets are measurement-defined, and regimes are feasibility-bound. Database visibility is shaped by community incentives. We introduced the Bias Before Data (BBD) Framework to explain how upstream decisions structure the observation field and thereby pre-structure dataset content. BBD decomposes pre-dataset bias into five coupled mechanisms: Problem Framing Bias, Regime Availability Bias, Measurement–Proxy Bias, Curation–Visibility Bias, and Legitimacy Bias.

The central claim is not that materials datasets are “bad,” but that datasets are constructed views of the material world, shaped by what can be observed, what is worth recording, and what can be standardized. When models learn from these constructed views, predictive success can coexist with systematic blind spots. The scientific risk is therefore not only

error, but overclaiming—the tendency to treat benchmark accuracy as evidence for conclusions that exceed what the upstream construction makes legitimate.

BBD contributes a new vocabulary for diagnosing upstream distortion and for binding scientific claims to the limits of observability, proxy meaning, and visibility. This perspective strengthens materials informatics as a scientific discipline by shifting evaluation from performance alone toward claim validity under construction constraints. By making pre-dataset bias explicit, the framework supports more defensible interpretations of AI outputs, more careful communication of generalization boundaries, and a clearer distinction between predictive utility and scientific legitimacy.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 28 Nov 2023 Revised: 05 Jan 2024 Accepted: 02 Mar 2024
Published online: 18 July 2024

References

Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst. 2020;33:13991–4002.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn. 2021;110:457–506.

Himanen L, Jäger MOJ, Morooka EV, Federici Canova F, Ranawat YS, Gao DZ, et al. DScribe: library of descriptors for machine learning in materials science. Comput Phys Commun. 2020;247:106949.

Laakso J, Himanen L, Pouillon Y, Jäger MOJ, Foster AS. Updates to the DScribe library: new descriptors and derivatives. J Chem Phys. 2023;158(23):234802.

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

Wei J, Chu X, Sun X, Xu K, Deng H, Chen J, et al. Machine learning in materials science. InfoMat. 2020;2:338–58.

Choudhary K, DeCost B, Tavazza F. Machine learning with force-field-inspired descriptors for materials: fast screening and interpretation. Phys Rev Mater. 2021;5:063802.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2021;7:25.

Musil F, De S, Yang J, Campbell JE, Ceriotti M. Physics-inspired structural representations for molecules and materials. Chem Rev. 2021;121:9759–815.

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

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;32:3564–75.

Goodall R, Lee AA. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nat Commun. 2020;11:6280.

Kauwe SK, Graser J, Murdock RJ, Sparks TD. Can machine learning find extraordinary materials? Comput Mater Sci. 2020;174:109498.

Bartel CJ, Millican SL, Deml AM, Rumptz JR, Tumas W, Weimer AW, et al. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. Nat Commun. 2020;11:4485.

Draxl C, Scheffler M. NOMAD: the FAIR concept for big data-driven materials science. MRS Bull. 2020;45:713–9.

Moosavi SM, Nandy A, Jablonka KM, Ongari D, Lan G, Smit B. Understanding the diversity of the metal–organic framework ecosystem. Nat Commun. 2020;11:4068.

Rosen AS, Iyer SM, Ray D, Yao Z, Aspuru-Guzik A, Gagliardi L, et al. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. Matter. 2021;4:1578–95.

Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60:3770–80.

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

Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater. 2020;6:39.

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. 2020;8:030902.

Choudhary K, DeCost B, Tavazza F. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8:59.

Stanev V, Oses C, Kusne AG, Rodriguez E, Paglione J, Curtarolo S, et al. Machine learning modeling of superconducting critical temperature. npj Comput Mater. 2021;7:107.

Jha D, Ward L, Paul A, Liao W-K, Choudhary A, Agrawal A, et al. ElemNet: deep learning the chemistry of materials from only elemental composition. Sci Rep. 2020;10:9836.

Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater. 2020;6:25.

Chen L, Kim C, Batra R, Lightstone JP, Wu C, Li Z, et al. Frequency-dependent dielectric constant prediction of polymers using machine learning. npj Comput Mater. 2020;6:61.

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

Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. Nature. 2020;533:73–6.

Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 2020;571:95–8.

Kononova O, Huo H, He T, Rong Z, Botari T, Sun W, et al. Text-mined dataset of inorganic materials synthesis recipes. Sci Data. 2020;7:203.

Rickman JM, Lookman T, Kalidindi SR. Materials informatics: from the atomic-level to the continuum. Acta Mater. 2021;201:420–44.

Stanev V, Kusne AG, Rodriguez E, Oses C, Paglione J, Curtarolo S, et al. Data-driven discovery and synthesis of superconductors. Nat Rev Mater. 2021;6:657–75.

Park CW, Wolverton C. Developing an improved crystal structure descriptor for machine learning models of formation energies. Phys Rev Mater. 2020;4:063801.

Fung V, Zhang J, Juarez E, Sumpter BG. Benchmarking graph neural networks for materials chemistry. npj Comput Mater. 2021;7:84.

Chen W, Zhang Y, Huang Z, Wang L, Ong SP. Learning the properties of materials from scratch with large-scale graph neural networks. Nat Mach Intell. 2023;5:234–45.