ORIGINAL RESEARCH **Open access**

Benchmarking Without Reality: Dataset Construction Bias in Materials Evaluation

Wei Liu[1*], Zhang Min[1]

Abstract

In the rapidly evolving field of computational and data-driven materials engineering, machine learning models are increasingly deployed for property prediction, inverse design, and autonomous discovery. However, the integrity of these models hinges on the quality of training datasets, which often embed subtle biases arising from construction methodologies. This manuscript explores the conceptual underpinnings of dataset construction bias in materials AI evaluation, framing it as an epistemic challenge that distorts benchmarking outcomes and impedes genuine materials discovery. We introduce the Dataset Integrity Cascade (DIC) framework, a layered conceptual model that maps data curation processes to inference distortions, incorporating feedback mechanisms to reveal how biases propagate through representation learning, model training, and validation pipelines. By synthesizing recent advances in materials informatics, graph neural networks, and uncertainty quantification, the framework highlights systemic trade-offs between dataset scale and representational fidelity. Implications extend to high-throughput computation, closed-loop experimentation, and foundation models for science, suggesting pathways for more robust computational steering in materials design. This work underscores the need for integrative approaches that align dataset architectures with the inherent complexities of materials systems, fostering epistemically sound innovation without empirical validation.

Keywords Materials informatics, Representation learning, Dataset bias, Uncertainty propagation, Computational discovery, Machine learning evaluation

*Correspondence:
Wei Liu
wei.liu@gmail.com
[1] Department of Materials Informatics, School of Materials Science, Shanghai Jiao Tong University, Shanghai, China

Introduction

The advent of data-driven paradigms in materials engineering has transformed traditional discovery processes, shifting the epistemic foundation of the field from heuristic-guided experimentation toward predictive modeling infrastructures informed by vast computational and experimental datasets [1, 2]. Historically, materials innovation relied on iterative laboratory synthesis, phenomenological theory building, and intuition-guided compositional exploration. While this paradigm yielded foundational breakthroughs, it was constrained by slow throughput, limited sampling of chemical space, and an inability to systematically interrogate high-dimensional structure–property relationships. The rise of computational materials science—coupled with scalable data infrastructures—has reconfigured this landscape into one where machine learning systems operate as discovery accelerators, capable of screening millions of candidate materials across thermodynamic, electronic, and mechanical property domains.

Within this emergent ecosystem, machine learning enables the identification of novel materials with tailored functional properties, including engineered bandgaps for semiconductor applications, optimized adsorption energetics for catalysis, and transferable force fields for molecular simulations [3-5]. These predictive capabilities derive from architectures capable of encoding atomistic topology, electronic structure descriptors, and microstructural features into learnable representations. As a result, materials design has begun to transition from

retrospective analysis to prospective inference, where AI models not only predict properties but also guide inverse design strategies and closed-loop discovery pipelines.

Yet, as these methodologies proliferate, a critical oversight emerges: the foundational datasets underpinning AI training, validation, and benchmarking are frequently treated as neutral substrates rather than epistemically charged constructs. In practice, these datasets harbor construction biases that undermine the reliability, interpretability, and translational validity of model evaluations. Such biases arise from selective curation protocols, incomplete sampling of compositional and structural diversity, domain-specific filtering criteria, and mismatched representational schemes. Consequently, benchmarking exercises—often framed as objective performance assessments—may instead reflect alignment with dataset artifacts rather than fidelity to physical reality. This misalignment risks steering discovery efforts toward artifactual optima: materials that appear promising within computational manifolds yet fail under experimental or mechanistic scrutiny [6-8].

Evolution of Data-Driven Materials Paradigms

Historically, the transition toward data-centric materials science was catalyzed by advances in density functional theory (DFT), high-throughput computational screening, and automated workflow infrastructures capable of generating large property databases [9, 10]. Initiatives in computational materials genomics institutionalized the large-scale aggregation of calculated thermodynamic stability, electronic structure, and defect energetics data, enabling systematic exploration of chemical design spaces previously inaccessible through experimental means alone.

The integration of machine learning extended this paradigm beyond simulation throughput constraints. Architectures such as graph neural networks (GNNs), message-passing frameworks, and deep representation learning models enabled property prediction directly from structural graphs, bypassing the need for exhaustive first-principles calculations [11-14]. Models trained on stoichiometric compositions, density of states spectra, or local atomic environments demonstrated predictive utility across diverse tasks, including bandgap estimation, adsorption energy prediction, and elastic property forecasting [5, 15]. These developments reframed materials datasets not merely as repositories of computed values but as training substrates for emergent predictive intelligence.

However, the datasets fueling these models are not passive reflections of materials reality. They are constructed artifacts shaped by deliberate decisions in data sourcing, filtering thresholds, featurization pipelines, and augmentation strategies. Each construction step introduces systematic deviations that influence downstream inference. In porous materials genomics, for example, big-data aggregation reveals how curation heuristics amplify uncertainty in adsorption property predictions [16]. Similarly, in microstructure-informed learning, transfer learning studies highlight the difficulty of generalizing across heterogeneous imaging modalities and processing histories [17]. These observations expose a structural tension: as computational efficiency increases through dataset scaling, epistemic fidelity may erode due to latent construction biases. Thus, datasets increasingly function as mediating epistemic infrastructures—translating simulation outputs into AI-interpretable knowledge while simultaneously constraining interpretive validity [18, 19].

The Epistemic Role of Datasets in AI Evaluation

Datasets serve as the epistemic bridge linking materials phenomena to machine inference systems, encoding implicit assumptions about phase stability, bonding physics, synthesis feasibility, and measurement reliability within their structural organization [20, 21]. Their composition defines what constitutes “learnable reality” for AI systems. Consequently, bias in dataset construction manifests across multiple interdependent dimensions.

Sampling imbalances, for instance, often privilege thermodynamically stable crystalline phases, underrepresenting metastable, amorphous, or defect-rich systems that are technologically relevant yet computationally underexplored [6, 22]. Representational artifacts emerge when graph encodings omit long-range interactions, disorder states, or environmental couplings, constraining model expressivity despite architectural sophistication [11, 13]. Validation strategies introduce further distortions when dataset splits fail to preserve multimodal compositional or structural distributions, producing inflated performance metrics divorced from extrapolative deployment contexts [7, 23].

These compounded biases distort benchmarking interpretations. Models may appear highly performant on curated test sets yet fail under domain shift conditions, as observed in compound stability predictions and transferable force-field generalization tasks [3, 4, 6]. Uncertainty quantification frameworks attempt to mitigate these risks by estimating predictive variance, calibration error, or epistemic uncertainty [20, 21]. However, such techniques frequently operate downstream of dataset construction, quantifying model ignorance without interrogating upstream data distortions. This leads to a paradox of calibrated overconfidence, where uncertainty metrics are themselves conditioned on biased informational substrates [8].

In autonomous discovery systems, these epistemic distortions propagate recursively. Closed-loop pipelines—where AI models propose candidates, simulations validate them, and results retrain the model—can amplify initial dataset biases through iterative reinforcement [18, 24, 25]. Over successive cycles, discovery trajectories may converge toward narrow regions of materials space defined more by dataset topology than by physical opportunity. The conceptual gap, therefore, lies in the persistent treatment of datasets as passive inputs rather than active shapers of discovery logics. Construction decisions implicitly define the boundaries, densities, and accessibility gradients of explorable materials possibility spaces [26, 27].

Challenges in Benchmarking Integrity

Benchmarking infrastructures in materials AI are designed to standardize evaluation, enabling cross-model comparison through shared datasets and performance metrics. Platforms such as Matbench exemplify this effort, providing curated test suites spanning electronic, thermodynamic, and mechanical property prediction tasks [7]. While such benchmarks advance reproducibility and methodological transparency, they inherit construction biases embedded within their source databases.

For example, many benchmarking datasets overrepresent ordered crystalline compounds while neglecting defect-inclusive, disordered, or synthesis-condition-dependent materials representations [12, 28]. This structural skew privileges models optimized for idealized periodic systems, limiting translational validity in applied materials engineering contexts. In small-data regimes, these distortions intensify: limited samples amplify prior biases, encouraging overfitting to narrow compositional manifolds and inflating apparent predictive performance [23, 29].

Conversely, the emergence of large foundation models trained on expansive multimodal materials datasets introduces aggregation biases of a different order. Data harmonization across disparate repositories—each with distinct measurement standards, simulation fidelities, and curation logics—produces blended knowledge representations that obscure domain boundaries [30]. While such models promise broader generalizability, they risk embedding composite biases that are difficult to diagnose or disentangle.

Further complications arise in simulation–experiment coupling workflows. Discrepancies between computed and experimentally measured properties introduce domain shifts that propagate into benchmarking regimes [9, 19, 31]. Models evaluated on computational ground truths may exhibit degraded performance when confronted with real-world synthesis variability, measurement noise, or environmental dependencies. The resulting benchmarking landscape prioritizes apparent performance optimization over alignment with mechanistic and physical principles. In extreme cases, this may incentivize the pursuit of materials that exist only within biased data manifolds rather than experimentally realizable systems [6, 8, 10].

Addressing these challenges requires a systemic reframing of dataset ecosystems—not as static benchmarking substrates but as dynamic epistemic infrastructures whose construction logics shape downstream inference, evaluation, and discovery steering [2, 16, 32].

Framing the Conceptual Intervention

In response to these systemic distortions, this manuscript positions dataset construction bias as a core impediment to epistemically robust materials AI. Rather than treating bias as a peripheral data quality issue, we conceptualize it as a structural force that permeates representation learning, model validation, uncertainty quantification, and autonomous discovery architectures.

We therefore introduce a novel conceptual framework designed to interpret, map, and navigate dataset construction distortions across the materials AI lifecycle. By dissecting the cascades through which biases emerge,

propagate, and amplify—from data acquisition to benchmarking to closed-loop discovery—we aim to reorient computational workflows toward epistemic alignment. Such alignment is essential not only for improving predictive reliability but also for ensuring that AI-guided discovery remains tethered to physically realizable and scientifically meaningful innovation trajectories.

The theoretical foundations of computational materials engineering are increasingly inseparable from data-driven epistemologies, forming a hybridized discovery ecosystem in which algorithmic inference, simulation infrastructures, and curated datasets co-evolve as interdependent knowledge systems [1, 2]. Within this synthesis, machine learning does not operate merely as a predictive overlay but as an interpretive apparatus conditioned by the informational architectures that sustain it. Consequently, understanding how datasets are assembled—their provenance, filtering logics, representational encodings, and distributional geometries—becomes central to evaluating AI performance, particularly in benchmarking regimes where claims of generalizability and physical realism are adjudicated [6, 7].

This literature convergence reveals a structural shift: evaluation is no longer determined solely by model architecture or optimization strategy but by the epistemic topology of the datasets that scaffold training and validation. As such, theoretical inquiry into materials AI must interrogate dataset construction not as a preprocessing step but as a foundational act of scientific framing that shapes inference horizons, uncertainty landscapes, and discovery trajectories.

Dataset Curation in Materials Informatics

Dataset construction in materials informatics originates within high-throughput computational ecosystems, where density functional theory (DFT), molecular dynamics, and automated workflow engines generate large-scale property repositories spanning thermodynamic, electronic, and structural domains [9, 10, 31]. These infrastructures enable systematic sampling of chemical spaces, yet practical constraints—computational cost, convergence stability, and workflow standardization—inevitably bias sampling toward energetically favorable configurations, ordered crystalline phases, and chemically common element sets [6, 22].

Such skewed sampling introduces foundational imbalances. Low-energy ground states become overrepresented, while metastable phases, defect-rich systems, and kinetically accessible yet thermodynamically unfavorable materials remain underexplored. This imbalance narrows the representational bandwidth available to machine learning systems, constraining their capacity to infer beyond equilibrium-dominated manifolds.

Active learning frameworks attempt to address sampling inefficiencies by iteratively selecting high-value data points to refine interatomic potentials and property predictors [18, 24, 31]. However, literature demonstrates that these adaptive strategies may inadvertently entrench initial dataset biases. When acquisition functions are conditioned on biased seed datasets, exploration trajectories become self-reinforcing, privileging already-sampled regions while neglecting epistemically sparse zones.

In porous materials genomics, big-data curation initiatives assemble vast libraries of hypothetical and experimentally derived frameworks [16]. Yet filtering heuristics—such as synthetic feasibility constraints, pore size thresholds, or thermodynamic stability cutoffs—introduce selection artifacts that shape machine learning inference. What appears as predictive accuracy may instead reflect alignment with curated feasibility regimes rather than intrinsic adsorption physics.

Microstructure-sensitive domains further complicate curation logics. Datasets must capture morphological diversity across processing histories, imaging modalities, and scale hierarchies. Transfer learning offers partial mitigation by enabling knowledge transfer across domains, but studies reveal persistent biases inherited from source datasets, particularly when feature distributions diverge significantly [17].

Uncertainty quantification (UQ) scholarship reinforces these concerns by demonstrating how curation decisions propagate predictive variance and calibration errors through downstream models [8, 20, 21]. Emerging calls within the literature advocate for provenance-aware metrics capable of tracing uncertainty not only to model stochasticity but to dataset construction lineage itself.

Representation Learning and Bias Embedding

Representation learning constitutes the epistemic translation layer where curated datasets are transformed into machine-interpretable abstractions [11-13]. Within this layer, biases embedded during curation become encoded into feature spaces, latent manifolds, and graph topologies that govern model perception of materials reality.

Graph neural networks (GNNs) exemplify this representational synthesis, unifying molecular and crystalline systems through relational encodings of atomic connectivity and bonding environments [11, 14]. Their message-passing architectures enable cross-scale property prediction, yet their fidelity is bounded by the completeness of graph construction. Omission of defects, grain boundaries, disorder states, or dynamic fluctuations introduces structural blind spots that constrain representational expressivity [12, 29].

Stoichiometry-based deep learning models offer an alternative by predicting properties from composition alone, bypassing structural encodings [13]. While computationally efficient, such representations assume statistical regularities across compositional space that real-world datasets rarely satisfy. Distributional heterogeneity—stemming from synthesis feasibility, measurement availability, or computational prioritization—violates these assumptions, embedding bias directly into learned mappings.

In classification domains, deep learning models trained to identify crystal symmetries or structural prototypes demonstrate high accuracy within curated datasets [12]. Yet interpretive analyses reveal that these systems often internalize dataset regularities rather than crystallographic principles, producing skewed generalizations when confronted with out-of-distribution structures.

Feature engineering further modulates representational bias. Density of states embeddings, for instance, enhance adsorption property predictions by capturing electronic structure nuances [15]. However, their effectiveness depends on the fidelity and diversity of the underlying spectral datasets. Where spectral sampling is sparse or energetically constrained, representations amplify construction biases rather than resolve them.

Multimodal datasets introduce additional complexity. Integrating simulation outputs with experimental measurements promises richer embeddings but generates fusion biases when modalities differ in scale, noise, or sampling density [19, 30]. Models trained on such datasets may privilege high-volume modalities, marginalizing lower-density but epistemically critical signals.

Foundation models extend this paradigm through large-scale pretraining across aggregated materials repositories [27, 30]. While these architectures promise general-purpose representational capacity, their reliance on heterogeneous data sources risks homogenizing materials diversity. Without bias-aware curation protocols, scaling representation learning may amplify rather than mitigate epistemic distortions.

Model Training Dynamics and Inference Distortions

During model training, dataset construction biases crystallize into inference geometries that govern predictive behavior [3-5]. Optimization processes internalize distributional regularities, weighting overrepresented regions of feature space while underfitting sparse domains.

Force-field learning illustrates this dynamic. Machine-learned interatomic potentials trained on trajectory-limited datasets reproduce energy landscapes within sampled regimes yet exhibit unphysical extrapolations under novel thermodynamic or configurational conditions [3, 4, 14]. Here, bias is not merely statistical but mechanistic, distorting learned physical laws.

Bandgap prediction models employing multi-fidelity learning demonstrate how hierarchical datasets embed resolution biases [5]. High-fidelity quantum calculations anchor predictive accuracy, but their limited availability forces reliance on lower-fidelity approximations, introducing fidelity-induced distortions in training gradients.

Small-data learning exacerbates these issues. Feature selection, joint learning, and transfer strategies enhance predictive efficiency, yet limited sampling intensifies susceptibility to construction bias [23, 32]. Overfitting to biased priors becomes statistically probable, producing fragile generalizations.

Robustness studies confirm these vulnerabilities, revealing performance degradation when models trained on curated datasets encounter structurally or compositionally novel inputs [8, 22]. Data redundancy within large repositories offers partial mitigation by smoothing distributional

irregularities, but only when aggregation biases are explicitly recognized and corrected [22].

Inverse design systems amplify training distortions. Optimization algorithms navigating latent design spaces converge toward high-scoring candidates defined by biased datasets rather than physically realizable optima [26, 27]. Literature in solid-state chemistry underscores this limitation, emphasizing the need to interpret predictive outputs through integrative data–model interaction frameworks rather than algorithmic performance alone [2, 16].

Validation Pipelines and Benchmarking Fallacies

Validation infrastructures in materials AI aim to simulate real-world deployment through standardized benchmarks, yet their epistemic reliability is contingent upon dataset construction integrity [6, 7, 21]. Benchmark suites such as Matbench provide cross-property evaluation platforms, enabling comparative assessment across algorithms [7]. However, their test sets inherit structural biases embedded within source databases.

For instance, overrepresentation of thermodynamically stable crystalline compounds skews evaluation toward equilibrium-phase prediction accuracy while neglecting defect-mediated, metastable, or synthesis-conditioned phenomena [12, 28]. Consequently, models optimized for benchmark performance may underperform in applied discovery environments.

Compound stability prediction studies illustrate this fallacy. Formation energy models trained on curated datasets report high accuracy yet falter when evaluated against experimentally synthesized metastable compounds [6]. Apparent predictive success thus reflects dataset alignment rather than mechanistic validity.

Frameworks such as MODNet expose imbalance-induced distortions, demonstrating how benchmarking outcomes shift when dataset variance is explicitly incorporated into evaluation metrics [21]. Active learning validation loops in heterogeneous catalysis further reveal dynamic biases, where model-guided sampling reshapes validation distributions over time [25].

Small-data benchmarking compounds these risks. Efficiency-driven validation on limited subsets produces inflated generalization claims disconnected from broader materials diversity [23]. Epistemic risk emerges when benchmarks are treated as objective arbiters rather than constructed evaluation artifacts shaped by dataset lineage [8, 19].

Autonomous discovery systems intensify this concern. Closed-loop experimentation—where models propose, experiments validate, and results retrain algorithms—creates recursive validation ecosystems [18, 24, 25]. If initial datasets are biased, feedback amplification can distort entire discovery trajectories, reinforcing narrow exploration regimes under the guise of empirical validation.

Systemic Trade-offs in Discovery Ecosystems

Across the literature, systemic trade-offs emerge between dataset scale, fidelity, diversity, and computational feasibility [9, 19, 26]. While big-data paradigms promise improved generalizability, scale alone does not guarantee epistemic robustness. Compute-intensive curation pipelines introduce infrastructure biases tied to workflow standardization, convergence criteria, and simulation approximations [19].

Domain-specific case studies illustrate these tensions. In polymer membrane design and 2D materials discovery, machine learning accelerates screening yet remains constrained by datasets biased toward historically studied chemistries [26, 29]. Novelty detection becomes structurally limited by dataset lineage.

Coupled simulation frameworks—linking DFT with molecular dynamics or mesoscale modeling—expand property coverage but introduce construction dependencies where dataset assembly choices influence emergent aggregation behaviors [28]. Thus, even multiscale datasets remain epistemically conditioned.

Foundation models scale discovery further by aggregating multimodal repositories into unified training corpora [27, 30]. While enabling cross-domain transfer learning, such aggregation embeds global biases reflecting dominant research investments, measurement infrastructures, and computational accessibility patterns.

Uncertainty quantification and robustness analytics provide methodological tools to navigate these trade-offs, offering calibration metrics, out-of-distribution detection, and confidence-aware screening strategies [8, 20, 32]. Yet these approaches operate reactively, diagnosing distortions after dataset construction rather than governing curation at inception.

Synthesis and Conceptual Gap

This synthesis reveals a fragmented landscape in which dataset construction bias is acknowledged across curation, representation, training, and validation literatures but addressed in isolation. No unified theoretical architecture currently maps how construction decisions cascade across the full AI evaluation lifecycle—from data acquisition to benchmarking to autonomous discovery steering. **Table 1** summarizes a taxonomy of dataset construction bias across the materials AI lifecycle, linking upstream construction mechanisms to downstream benchmarking failure modes and the mitigation levers implied by prior literature.

Table 1. Taxonomy of dataset construction bias in materials AI: sources, mechanisms, and evaluation failure modes

Bias locus (pipeline stage)	Construction mechanism (what is done)	Typical mani materials
Sampling coverage (data generation)	Preferential sampling of low-energy / convergent structures; common chemistries	Overrepresenta crystals; metastable/defec regi
Filtering / feasibility curation	Removing “non-feasible,” “unstable,” or “hard-to-compute” cases	Selection artifac absence of synt relevant ca
Label fidelity / multi-fidelity mixing	Combining DFT levels, functionals, or computed + experimental labels without alignment	Inconsistent tar hidden label n shifted
Representation omission	Graph encodings omit defects, disorder, temperature/dynamics; coarse structural descriptors	“Clean crystal” r missing microstr
Modality imbalance (multimodal fusion)	Unequal volume/quality across modalities (sim dominates exp; imaging underweights chemistry)	Learned embedd by high-volur
Split / validation design	IID random splits; leakage via duplicates/prototypes; non-representative holdouts	Train–test overl families; narrow
Active-learning reinforcement	Acquisition conditioned on biased seed sets; exploitation-heavy selection	Iterative densifica sampled
Aggregation bias (foundation corpora)	Merging heterogeneous repositories; harmonization without bias control	Homogenized dominant doma othe
Small-data framing bias	Task definition + feature selection optimized on narrow datasets	Over-specialize fragile caus
Reporting / metric bias	Single-number metrics; no uncertainty/coverage context	Overclaims of u failure

The absence of such an integrative framework constitutes a critical conceptual void. Without systemic interpretive models, efforts to mitigate bias remain localized, reactive, and methodologically siloed. Bridging this gap requires reconceptualizing datasets as active epistemic infrastructures whose construction logics co-determine AI evaluation validity and discovery potential.

Proposed Conceptual Framework

To address the conceptual gaps identified, we introduce the Dataset Integrity Cascade (DIC) framework, an original layered model that interprets how biases in dataset construction propagate through materials AI pipelines, influencing evaluation integrity and discovery steering. The DIC conceptualizes dataset ecosystems as interconnected cascades, where upstream curation decisions flow into downstream distortions, mediated by feedback loops that either amplify or attenuate epistemic risks. This framework comprises three structural layers: the Curation Substrate, the Representation Conduit, and the Inference Apex, each interacting via data-model-discovery pipelines to reveal systemic dynamics.

The Curation Substrate layer focuses on foundational data assembly, where high-throughput computations and simulation couplings generate raw inputs [9, 10, 31]. Here, biases emerge from sampling logics that prioritize computational tractability over comprehensive coverage, such as favoring equilibrium states [6, 22]. The layer incorporates steering mechanisms to interpret how active learning feedbacks refine but potentially entrench initial selections [18, 24]. A key dynamic can be conceptualized as the Sampling Fidelity Trade-off, expressed as $S = \alpha \cdot D - \beta \cdot C$, where S captures the effective sampling scope, D Dsymbolizes dataset diversity (encompassing chemical and structural variances), C denotes computational cost constraints, and α,β \alpha, \beta α,β represent weighting factors for exploratory versus efficient curation. This formula interprets the interaction between diversity aspirations and resource limitations, highlighting how over-constrained C diminishes S, leading to biased substrates that misalign with materials complexity.

Transitioning to the Representation Conduit layer, biases are transformed through featurization and embedding processes [11-13]. Graph neural networks and stoichiometry-based encodings conduit raw data into model-ready forms, but incomplete representations embed distortions [13, 14, 17]. Feedback loops from uncertainty quantification inform iterative refinements, adjusting for multimodal mismatches [20, 21, 30]. The conduit's core interaction may be expressed as $R = f(G, U)$, where R denotes representational robustness, G encapsulates graph or feature granularity, and U signifies uncertainty propagation from upstream layers. This captures the interplay where coarse G amplifies U, resulting in conduits that channel biased signals into training, as seen in density of states or microstructure applications [15, 17].

At the Inference Apex, biases culminate in distorted benchmarking and discovery outcomes [6-8]. Models trained on cascaded data yield inferences that deviate from physical benchmarks, with closed-loop systems perpetuating cycles via autonomous feedbacks [18, 25, 26]. The apex integrates steering logics to interpret how validation pipelines recycle biases, potentially locking discovery into artifactual paths [19, 23, 27]. A unifying dynamic is the Bias Amplification Feedback, conceptualized as $Bn + 1 = Bn + \gamma \cdot (I - V)$, where B_n is bias at iteration n , I represents inference demands, V denotes validation rigor, and γ modulates loop sensitivity. This formula interprets recursive amplifications, where inadequate V escalates B, underscoring epistemic risks in iterative AI workflows.

Interlayer pipelines facilitate data flow: curation-to-representation transfers embed initial biases, representation-to-inference conduits model behaviors, and apex feedbacks recirculate to curation for adaptive steering. These pipelines incorporate computational logics that balance scale with fidelity, revealing trade-offs in small-data versus foundation model regimes [23, 30].

As conceptualized in **Figure 1**, the DIC framework diagrams these layers as cascading tiers with bidirectional arrows denoting feedbacks, circular nodes for steering points, and shaded gradients illustrating bias propagation intensity from substrate to apex. The figure's central axis aligns data pipelines with discovery trajectories, flanked by side panels depicting formulaic interactions to visualize systemic equilibria. This textual depiction emphasizes the framework's interpretive utility in navigating dataset-driven distortions without empirical assertions. **Figure 1** visualizes the Dataset Integrity Cascade (DIC) as a layered system in which upstream curation choices are transformed into representational distortions and culminate in benchmarking divergence, with feedback loops that can amplify or attenuate epistemic risk.

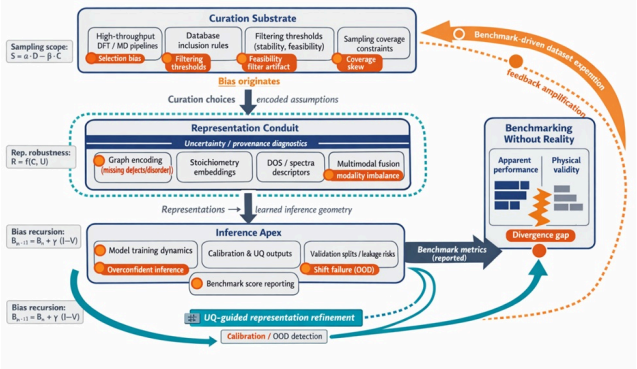

Figure 1. Dataset Integrity Cascade (DIC): a systems view of benchmarking without reality.

The DIC framework conceptualizes dataset construction bias as an upstream epistemic distortion that propagates through three coupled layers—Curation Substrate, Representation Conduit, and Inference Apex—shaping model training dynamics, uncertainty behavior, validation integrity, and ultimately benchmark outcomes. Orange markers indicate bias/risk entry points (e.g., selection filters, incomplete encodings, and benchmark illusions), while teal overlays denote diagnostic channels (uncertainty, provenance-aware checks, and out-of-distribution detection). Two feedback loops are highlighted: an inner loop enabling uncertainty-guided representational refinement, and an outer loop in which benchmark-driven dataset expansion can reinforce prior biases, increasing divergence between apparent performance and physical validity.

Table 2. operationalizes the DIC by mapping each layer to integrity criteria, diagnostic signals, and steering actions, clarifying how localized construction flaws cascade into downstream benchmarking distortions.

Bias locus (pipeline stage)	**Construction mechanism (what is done)**	**Typical mani materials**
Sampling coverage (data generation)	Preferential sampling of low-energy / convergent structures; common chemistries	Overrepresenta crystals; metastable/defec regi
Filtering / feasibility	Removing “non-feasible,” “unstable,”	Selection artifac absence of synt
curation	or “hard-to-compute” cases	relevant ca
Label fidelity / multi-fidelity mixing	Combining DFT levels, functionals, or computed + experimental labels without alignment	Inconsistent tar hidden label n shifted
Representation omission	Graph encodings omit defects, disorder, temperature/dynamics; coarse structural descriptors	“Clean crystal” r missing microstr
Modality imbalance (multimodal fusion)	Unequal volume/quality across modalities (sim dominates exp; imaging underweights chemistry)	Learned embedd by high-volur
Split / validation design	IID random splits; leakage via duplicates/prototypes; non-representative holdouts	Train–test overl families; narrow
Active-learning reinforcement	Acquisition conditioned on biased seed sets; exploitation-heavy selection	Iterative densifica sampled
Aggregation bias (foundation corpora)	Merging heterogeneous repositories; harmonization without bias control	Homogenized dominant doma othe
Small-data framing bias	Task definition + feature selection optimized on narrow datasets	Over-specialize fragile caus
Reporting / metric bias	Single-number metrics; no uncertainty/coverage context	Overclaims of u failure

Analytical Implications

The Dataset Integrity Cascade (DIC) framework offers interpretive lenses for dissecting representation–inference interactions and epistemic risk structures within materials AI pipelines. At the representational level, the conduit layer interprets how featurization choices—such as graph granularity or stoichiometry embeddings—interact with upstream curation to modulate signal fidelity, creating latent manifolds that constrain subsequent inference spaces [11-14]. This interaction reveals that representational robustness is not an intrinsic model property but a propagated outcome of layered decisions, where coarse encodings amplify uncertainties and narrow the effective hypothesis space for property mappings [20, 21]. Analytically, this implies that efforts to enhance generalizability, such as multimodal fusion or transfer mechanisms, must account for conduit-level distortions rather than treating them as downstream corrections [17, 30].

The framework further illuminates epistemic risk structures through the lens of feedback amplification. In closed-loop and autonomous discovery contexts, the apex layer's recursive dynamics show how validation pipelines can inadvertently reinforce curation biases, transforming initial sampling skews into systemic discovery locks [8, 18, 25]. This structure highlights infrastructure trade-offs: scaling datasets for foundation-model regimes increases representational capacity yet heightens exposure to aggregated construction artifacts, whereas small-data workflows trade breadth for heightened sensitivity to local biases [22, 23, 27]. The sampling fidelity trade-off equation introduced earlier captures this analytically, illustrating that maximizing diversity under computational constraints inevitably compresses the substrate, which then cascades into reduced inference reliability.

Computationally, the DIC exposes steering-logic dependencies, where active-learning feedbacks operate on biased priors, leading to interpretive cycles that privilege artifactual stability over unexplored chemical space [6, 24, 31]. Uncertainty quantification, positioned as a cross-layer diagnostic, interprets risk as cumulative rather than isolated, enabling systems-level insights into how representation–inference mismatches manifest as benchmarking divergences [8, 20, 21]. These implications underscore that dataset construction bias is not merely a preprocessing artifact but a foundational epistemic vector that reshapes discovery pipelines, demanding workflows that prioritize interpretive alignment over scale alone [2, 16, 19]. By framing these dynamics integratively, the framework guides conceptual redesign of data ecosystems toward greater resilience without invoking empirical metrics.

Results and Discussion

Integrating the DIC framework into the broader computational materials engineering landscape reveals systemic insights into how dataset construction choices permeate discovery logics. The literature on machine learning for solid-state materials consistently demonstrates that training dynamics inherit upstream biases, yet the cascade model interprets this as an interconnected rather than sequential process [1, 2, 16]. For instance, force-field and bandgap prediction efforts illustrate how representation conduits channel sampling artifacts into inference apexes, distorting validation outcomes even in high-throughput or closed-loop settings [3-5, 14, 25]. This interpretive view aligns with critiques of big-data optimism, emphasizing that infrastructure-level trade-offs—between compute feasibility and fidelity—persist across modalities and scales [8, 19, 26].

Epistemic risk structures become particularly salient in inverse design and foundation-model contexts, where the framework highlights feedback loops as double-edged mechanisms: they enable adaptive steering but risk locking trajectories into biased attractors if validation rigor is insufficient [26, 27, 30]. Multimodal and simulation–experiment couplings further exemplify this, as modality imbalances propagate through layers to create hybrid representations that deviate from physical priors [15, 19, 28]. The framework's analytical implications suggest that robustness examinations and uncertainty-aware refinements should be repositioned as cascade diagnostics rather than post-hoc patches, fostering workflows that explicitly track bias flow [8, 21, 22, 32].

At the ecosystem level, the DIC encourages a shift from isolated benchmark optimizations to holistic pipeline redesign, where steering logics prioritize epistemic alignment with materials complexity [6, 7, 10, 18, 24]. This does not diminish the value of existing advances in graph architectures, active learning, or small-data strategies but reframes them within a layered understanding of construction-induced distortions [11-13, 17, 23, 31]. The result is a conceptual scaffold for interpreting trade-offs that affect field-wide progress, from materials informatics to

autonomous discovery, without presuming resolution through scale or complexity alone.

Conclusion

The Dataset Integrity Cascade framework provides a conceptual architecture for interpreting dataset construction bias as a propagating epistemic challenge in materials AI evaluation. By delineating curation, representation, and inference layers alongside feedback and steering dynamics, it offers systems-level insights into how biases shape benchmarking integrity and discovery pipelines. The framework's emphasis on representation–inference interactions, epistemic risk structures, and infrastructure trade-offs advances an integrative perspective that complements existing computational methodologies while highlighting pathways for enhanced alignment in data-driven workflows. This interpretive contribution underscores the importance of viewing datasets as active mediators in the materials discovery ecosystem, guiding future conceptual developments toward epistemically robust computational practices.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 05 Aug 2022 Revised: 09 Sep 2022 Accepted: 16 Oct 2022
Published online: 18 March 2023

References

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1):54.

Schmidt J, Marques MAL, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller KR. Machine learning of accurate energy-conserving molecular force fields. Sci Adv. 2017;3(5):e1603015.

Huan TD, Batra R, Chapman J, Krishnan S, Chen L, Ramprasad R. A universal strategy for the creation of machine learning-based atomistic force fields. npj Comput Mater. 2017;3(1):37. https://doi.org/10.1038/s41524-017-0042-y.

Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156-63.

Bartel CJ, Chen CT, Tang Z, Zador L, Musgrave CB, Horton MK. A critical examination of compound stability predictions from machine-learned formation energies. npj Comput Mater. 2020;6(1):97.

Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. npj Comput Mater. 2020;6(1):138.

Li K, DeCost B, Choudhary K, Greenwood M, Hattrick-Simpers JR. A critical examination of robustness and generalizability of

machine learning prediction of materials properties. npj Comput Mater. 2023;9(1):55.

Perdew JP, Sun J, Ruzsinszky A, Peng H. The limited predictive power of density functional theory in materials discovery. npj Comput Mater. 2018;4(1):29.

Saal JE, Oliynyk AO, McCalla E. Machine learning in materials discovery: Confirmed stable ternary rare earth intermetallics from machine learning and DFT. Acta Mater. 2019;167:14-21.

Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. Sci Adv. 2017;3(12):e1701816.

Ziletti A, Kumar D, Scheffler M, Ghiringhelli LM. Insightful classification of crystal structures using deep learning. Nat Commun. 2018;9(1):2775.

Goodall REA, Lee AA. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nat Commun. 2020;11(1):6280.

Chen L, Tran H, Batra R, Kim C, Ramprasad R. Machine learning models for the prediction of energy, forces, and stresses for platinum. npj Comput Mater. 2021;7(1):19.

Fung V, Hu J, Ganesh P, Sumpter BG. Machine learned features from density of states for accurate adsorption energy prediction. Nat Commun. 2021;11(1):88.

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.

Goetz A, Klopotowski A, Sliwinski P, Ramprasad R. Addressing materials' microstructure diversity using transfer learning. npj Comput Mater. 2022;8(1):131.

Bernstein N, Csányi G, Deringer VL. De novo exploration and self-guided learning of potential-energy surfaces. npj Comput Mater. 2019;5(1):99.

Hattrick-Simpers J, Audus D, Green ML. Why big data and compute are not necessarily the path to big materials science. Commun Mater. 2022;3(1):55.

Janet JP, Duan C, Yang T, Nori A, Kulik HJ. A quantitative uncertainty metric controls error in data-driven materials discovery. npj Comput Mater. 2019;5(1):60.

De Wulf PMO, Rimella L, Ricci F, Schmidt J, Marques MAL, Hautier G, et al. Robust model benchmarking and bias-imbalance in data-driven materials science: Acase study on MODNet. npj Comput Mater. 2021;7(1):149.

Li K, Persaud D, Choudhary K, Gaultois MW, Hattrick-Simpers JR, DeCost B. Exploiting redundancy in large materials datasets for efficient machine learning with less data. Nat Commun. 2023;14(1):7282.

Hu C, Wang X, Duan D, Yang C, Qiu J, Tian Y. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):42.

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. 2019;5(1):21.

Kreitz B, Creighton JR, Kim H. Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt. Nat Commun. 2022;13(1):5183.

Goetz A, Garg A, Ramprasad R. Machine learning guided discovery of polymer membranes for solvent recovery. npj Comput Mater. 2023;9(1):81.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5. https://doi.org/10.1038/s41586-023-06735-9.

Lu S, Zhou Q, Guo X, Zhang Y, Wu J, Xiong J. Combining DFT, Monte Carlo and molecular dynamics simulations to systematically predict the aggregation behavior of polymer plasticizers in polyvinyl chloride. Adv Theory Simul. 2021;4(4):2000245.

Frey NC, Akinwande D, Jariwala D, Shenoy VB. Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing. ACS Nano. 2020;14(10):13402-17.

Omee SS, Hu S, Imam N, Louis SY, Edirisooriya M, Mir QH, et al. Material transformers: Deep learning language models for generative materials design. Mach Learn Sci Technol. 2023;4(1):015014.

Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Comput Mater Sci. 2017;140:171-80. https://doi.org/10.1016/j.commatsci.2017.08.031.

De Breuck PP, Hautier G, Rignanese GM. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. npj Comput Mater. 2021;7(1):83. https://doi.org/10.1038/s41524-021-00552-2.