ORIGINAL RESEARCH **Open access**

Representation Is Not Reality: Epistemic Limits of Learned Materials Embeddings in Computational Design Systems

Li Zhang[1*], Wei Chen[1]

Abstract

The rapid evolution of computational and data-driven materials engineering has transformed materials discovery from traditional trial-and-error approaches to sophisticated AI-integrated pipelines. Within this paradigm, learned embeddings serve as foundational representations that encode complex material properties, structures, and behaviors into latent spaces amenable to machine learning algorithms. However, these embeddings, while powerful for predictive modeling and high-throughput screening, introduce epistemic limits that challenge the fidelity of computational design systems. This manuscript explores the disconnect between representational abstractions and physical reality, emphasizing how embedding-induced biases, dimensionality reductions, and generalization assumptions constrain the reliability of AI-guided materials innovation. We introduce a novel conceptual framework, the Epistemic Representation Cascade (ERC), which dissects the multi-layered interactions between data infrastructures, learning architectures, and discovery workflows to reveal inherent epistemic risks. By integrating insights from materials informatics and representation learning, the ERC highlights feedback mechanisms that amplify or mitigate these limits, offering systems-level guidance for enhancing interpretability and robustness in autonomous design ecosystems. Implications extend to closed-loop experimentation and inverse design, advocating for infrastructure-aware strategies that prioritize epistemic alignment over mere predictive accuracy. This work underscores the need for balanced computational steering in materials AI, fostering more trustworthy pathways for next-generation materials engineering.

Keywords Materials informatics, Epistemic constraints, Representation learning, AI in materials science, Computational discovery, Data-driven design

*Correspondence:
Li Zhang
li.zhang@outlook.com
[1] Department of Materials Informatics, School of Materials Engineering, Tsinghua University, Beijing, China

Introduction

The advent of computational and data-driven materials engineering marks a pivotal shift in how scientists and engineers approach the design and discovery of new materials. Historically, materials development relied on empirical experimentation and intuition-driven synthesis, often constrained by time, cost, and the vastness of chemical space. In recent years, however, the integration of high-performance computing, machine learning, and large-scale databases has enabled unprecedented acceleration in this field. High-throughput computational methods, such as density functional theory (DFT) simulations coupled with automated workflows, now allow for the virtual screening of thousands of candidate materials in fractions of the time required for physical testing [1, 2]. This computational paradigm not only expands the explorable design space but also facilitates the identification of materials with tailored properties for applications ranging from energy storage to catalysis.

Central to this transformation is the role of AI and data ecosystems in materials science. Machine learning models,

trained on multimodal datasets encompassing structural, electronic, and thermodynamic data, predict material behaviors with remarkable efficiency [3, 4]. For instance, graph neural networks and other deep learning architectures have become instrumental in processing crystal structures and molecular graphs, enabling predictions of properties like band gaps, stability, and reactivity [5, 6]. These tools are embedded within broader informatics infrastructures that aggregate data from simulations, experiments, and literature, forming the backbone of materials informatics [7]. Such ecosystems support inverse design strategies, where desired properties guide the generation of novel compositions rather than vice versa [8, 9]. The synergy between data-driven models and computational pipelines has led to breakthroughs, such as the rapid screening of solid-state electrolytes or perovskites for photovoltaic applications [10, 11].

Despite these advances, high-throughput infrastructures reveal inherent limitations in current discovery models. While computational workflows excel at scaling predictions, they often overlook the epistemic gaps between modeled representations and actual material realities. Learned embeddings—vectorial encodings of materials data—simplify complex phenomena into lower-dimensional spaces for algorithmic efficiency, but this abstraction can distort underlying physics [12, 13]. For example, in representation learning, assumptions about translational invariance or feature invariance may not hold across diverse material classes, leading to biased inferences [14]. Moreover, the reliance on curated datasets introduces selection biases, where underrepresented materials or edge cases skew model generalizations [15, 16]. These constraints are exacerbated in autonomous discovery systems, where closed-loop integrations between prediction, synthesis, and validation amplify errors if epistemic limits are not addressed [17].

Epistemic and computational constraints further complicate this landscape. Uncertainty quantification, essential for reliable design, remains challenging in deep learning frameworks applied to materials, as models may confidently predict erroneous outcomes due to overfitting or domain shifts [18, 19]. In simulation-experiment coupling, discrepancies arise from the idealizations in computational models, such as neglecting defects or environmental factors, which embeddings fail to fully capture [20, 21]. The push toward foundation models for science, inspired by large language models, promises broader applicability but risks compounding these issues through opaque black-box behaviors [22]. As materials engineering increasingly depends on AI-guided decisions, understanding these limits becomes critical to avoid over-reliance on representations that diverge from physical truths.

This manuscript positions a new interpretive lens on these challenges, focusing on the epistemic limits of learned materials embeddings in computational design systems. By synthesizing recent developments in materials informatics and AI architectures, we reveal how representational choices influence discovery outcomes. The introduction of the Epistemic Representation Cascade (ERC) framework provides a systems-level analysis of these dynamics, emphasizing the interplay between data infrastructures, learning mechanisms, and workflow integrations. Through this framework, we explore computational steering logics that can mitigate epistemic risks, fostering more robust and interpretable materials design pipelines. Ultimately, this work advocates for a paradigm where representations are viewed not as proxies for reality but as tools requiring continuous epistemic scrutiny.

Theoretical Background & Literature Synthesis

Materials data infrastructures

The foundation of computational and data-driven materials engineering lies in robust data infrastructures that enable the aggregation, curation, and utilization of vast datasets. These infrastructures encompass multimodal materials datasets derived from high-throughput computations, experimental validations, and literature mining, forming repositories that fuel machine learning applications [23, 24]. For instance, databases integrating DFT-calculated properties with structural descriptors allow for systematic exploration of material spaces, supporting tasks like property prediction and anomaly detection [25]. The evolution of these systems has emphasized interoperability, with knowledge graphs and ontologies facilitating semantic connections across disparate data sources [26]. Such infrastructures not only standardize representations but also enable feedback loops in autonomous discovery, where data from closed-loop experimentation refines subsequent models [27, 28].

However, these data ecosystems introduce epistemic challenges related to completeness and fidelity. Curated datasets often suffer from biases toward stable or

synthesizable materials, limiting the representation of metastable or novel compounds [29]. Multimodal integration, while enriching embeddings, can propagate inconsistencies if simulation data diverges from experimental realities [30]. Literature highlights the need for dynamic infrastructures that adapt to emerging data, yet current paradigms struggle with scalability and provenance tracking [31].

Representation learning architectures

Representation learning has emerged as a cornerstone of AI in materials science, transforming raw structural and compositional data into embeddings suitable for downstream tasks. Graph neural networks, in particular, have proven effective for encoding crystal lattices and molecular topologies, capturing local and global features through message-passing mechanisms [5, 32]. These architectures extend to universal models that handle diverse material types, incorporating alchemical distributions or structural invariants to enhance generalizability [3, 14]. Deep learning variants, such as generative adversarial networks, further enable sampling of chemical spaces for inverse design, generating embeddings that guide synthesis proposals [7, 8].

Despite their utility, these architectures impose epistemic limits through dimensionality reduction and feature selection. Embeddings may overlook subtle interactions, like long-range electronic effects, leading to incomplete representations of material behaviors [12, 19]. Uncertainty quantification techniques, integrated into these models, attempt to address this by estimating confidence in embeddings, but they often rely on assumptions that do not align with physical variability [18, 21]. Synthesis of recent works underscores the trade-offs between architectural complexity and interpretability, where overparameterized models risk epistemic opacity [6, 22].

AI-Guided discovery systems

AI-guided discovery systems integrate representation learning with computational workflows to automate materials innovation. These systems leverage machine learning for high-throughput screening, prioritizing candidates based on predicted properties and feasibility [4, 10]. Autonomous platforms incorporate closed-loop experimentation, where AI directs robotic synthesis and characterization, iteratively refining embeddings through real-time data [27, 28]. Inverse design paradigms invert this process, using embeddings to map target functionalities back to material compositions [9, 11, 20].

Epistemic constraints in these systems arise from the reliance on learned representations that may not faithfully mirror reality. For example, in perovskite discovery, models trained on embeddings can overlook kinetic barriers, resulting in theoretically stable but practically unrealizable materials [13, 17]. Literature emphasizes the need for hybrid approaches that couple simulations with experiments to ground embeddings, yet discrepancies persist due to representational abstractions [15, 25].

Computational design paradigms

Computational design paradigms in materials engineering encompass inverse strategies and optimization frameworks that exploit embeddings for efficient exploration. Techniques like genetic algorithms combined with machine learning accelerate polymer or alloy design by evolving embeddings toward optimal property landscapes [9, 11]. High-throughput paradigms scale this to entire classes of materials, using parallel computations to populate design spaces [2, 31].

However, these paradigms highlight epistemic limits in how embeddings influence design decisions. Generalization across material domains can fail if embeddings encode domain-specific biases, leading to suboptimal or misleading designs [16, 29]. The integration of foundation models aims to broaden applicability, but without epistemic safeguards, they may amplify uncertainties in computational steering [22, 30].

Uncertainty & interpretability

Uncertainty quantification and interpretability are critical for addressing epistemic limits in learned embeddings. Bayesian approaches and ensemble methods provide probabilistic insights into model predictions, revealing variances in embedding spaces [18, 21]. Interpretability tools, such as attention mechanisms in neural networks, elucidate how representations contribute to outcomes, aiding in the detection of epistemic gaps [19, 32].

Yet, challenges remain in translating these into actionable insights for materials design. Literature notes that while uncertainty metrics improve reliability, they often do not capture systemic biases inherent in data infrastructures [24, 26]. Enhancing interpretability requires balancing computational efficiency with transparency, ensuring that

embeddings serve as interpretable bridges rather than opaque barriers [14, 23].

Proposed conceptual framework

To address the epistemic limits of learned materials embeddings in computational design systems, we propose the Epistemic Representation Cascade (ERC) framework. This original systems-level construct dissects the multi-layered dynamics of data processing, model inference, and discovery integration, revealing how representational choices cascade through workflows to influence outcomes. The ERC conceptualizes materials engineering pipelines as interconnected cascades, where each layer transforms inputs while introducing or propagating epistemic risks. At its core, the framework identifies three structural layers: the Data Ingestion Layer, the Embedding Transformation Layer, and the Discovery Steering Layer. These layers interact via bidirectional feedback loops that modulate information flow, ensuring adaptive responses to epistemic discrepancies. The structural roles, transformation logics, and epistemic exposures of each ERC layer are systematized in **Table 1.**

Table 1. Structural Architecture of the Epistemic Representation Cascade (ERC)

ERC Layer	Core Function	Primary Computational Mechanisms	E Ris
Data Ingestion Layer	Aggregation and preprocessing of multisource materials knowledge	Data fusion pipelines; ontology mapping; preprocessing standardization	San pr as me nois inco
Embedding Transformation Layer	Conversion of raw materials data into latent computational representations	GNNs; VAEs; transformer encoders; manifold learning; dimensional compression	Ab los o in mis di
Discovery Steering Layer	Integration of embeddings into design, screening, and experimentation workflows	High-throughput screening; inverse design; autonomous experimentation; optimization engines	am ove in synt
Cross-Layer Feedback System	Recursive recalibration of cascade distortions	Uncertainty triggers; retraining loops; data augmentation; representation re-weighting	F l mis ur

The Data Ingestion Layer serves as the entry point, aggregating multimodal datasets from simulations, experiments, and informatics repositories. Here, raw materials data—such as atomic coordinates, property descriptors, and thermodynamic profiles—are preprocessed into standardized formats. Epistemic risks emerge from selection biases and incompleteness, where underrepresented data skews subsequent representations. Feedback from downstream layers refines ingestion by prioritizing gap-filling acquisitions, fostering dynamic infrastructure evolution.

Transitioning to the Embedding Transformation Layer, this component employs representation learning architectures to encode ingested data into latent spaces. Graph-based or deep learning models compress complexities into vectors, enabling efficient predictions. However, dimensionality reductions and invariance assumptions introduce abstractions that diverge from physical realities, manifesting as epistemic distortions. The transformation process in this layer can be conceptualized as $\begin{matrix} E \\ \oplus \epsilon \end{matrix} = \phi\left(D\right)$, where *E* denotes the resulting embedding, ϕ represents the learning function applied to input data *D*, and ϵ symbolizes the inherent epistemic distortion arising from abstraction mechanisms. This expression captures the interaction between data fidelity and representational simplification, highlighting how distortions accumulate without corrective measures. The ERC incorporates uncertainty quantification mechanisms within this layer to flag potential misalignments, with loops feeding back to data ingestion for targeted enrichment.

The Discovery Steering Layer integrates transformed embeddings into computational workflows, guiding high-throughput screening, inverse design, and closed-loop experimentation. Steering logics—algorithmic rules for decision-making—leverage embeddings to prioritize candidates, but epistemic limits can lead to amplified errors

in autonomous systems. Feedback loops connect this layer to upstream components, allowing real-time adjustments that mitigate risks through iterative refinement.

Central to the ERC are the feedback loops that enable cascade resilience. Positive loops amplify reliable representations by reinforcing validated pathways, while negative loops dampen distortions by triggering reinterpretations or data augmentations. Computational steering logics within these loops prioritize epistemic alignment, such as weighting embeddings based on uncertainty metrics or interpretability scores. This ensures that discovery pipelines remain grounded, balancing efficiency with fidelity.

The ERC's layered structure and loops are visualized as a cascading flowchart, with data flowing downward through transformations while feedback arrows arc upward for corrections, as conceptualized in **Figure 1**.

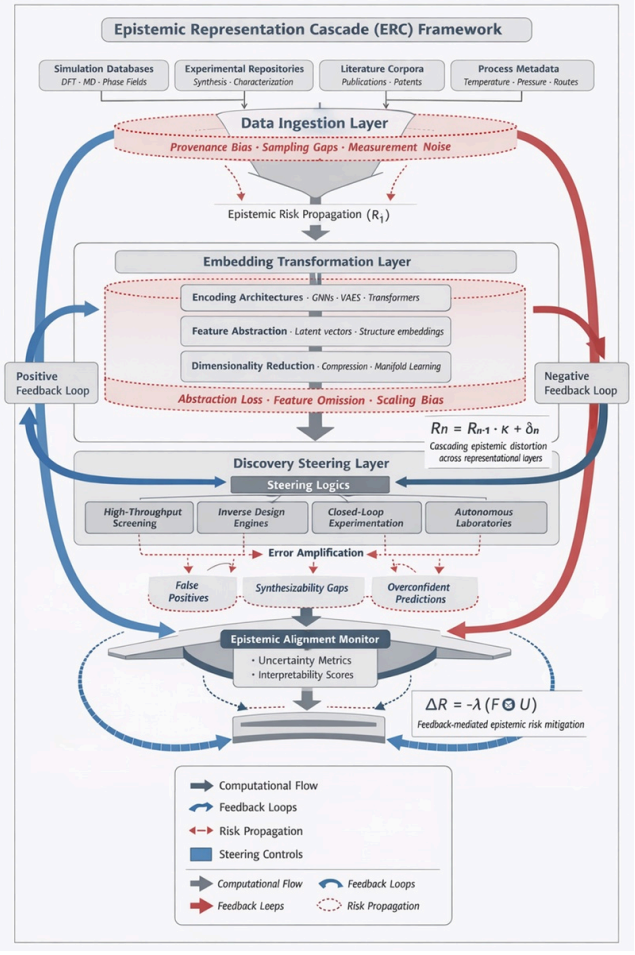

Figure 1. Epistemic Representation Cascade (ERC): Layered Transformations, Steering Logics, and Feedback-Driven Risk Modulation in Materials AI

The Epistemic Representation Cascade (ERC) framework conceptualizes materials AI discovery as a layered transformation pipeline linking data ingestion, embedding generation, and computational steering workflows. Epistemic risks propagate downward through representational compression while feedback loops enable cascade resilience by reinforcing validated pathways and dampening distortions through uncertainty-conditioned recalibration.

This framework provides interpretive insights into representation-inference interactions, highlighting infrastructure trade-offs that inform more robust materials AI ecosystems.

Analytical implications

The Epistemic Representation Cascade (ERC) framework generates layered analytical implications for computational and data-driven materials engineering by revealing how epistemic limits are structurally embedded within discovery infrastructures rather than emerging solely at model outputs. In this view, epistemic distortions are not discrete anomalies but cumulative conditions that propagate through sequential representational transformations. By disaggregating AI workflows into ingestion, embedding, and steering strata, the ERC foregrounds representation–inference coupling as a central determinant of scientific reliability, enabling a systems-level interpretation of how infrastructural design choices shape downstream discovery trajectories.

Within the Data Ingestion Layer, epistemic risk originates in the provenance structures of materials knowledge itself. Simulation-derived datasets, for instance, embed thermodynamic assumptions, boundary constraints, and methodological approximations that subtly condition the statistical landscapes from which embeddings are generated. When such datasets dominate ingestion pipelines, their latent biases cascade into representational priors, influencing latent geometries and inferential confidence structures. This dynamic implies that epistemic distortions may be infrastructurally inherited rather than algorithmically introduced, particularly where underrepresented chemistries, metastable states, or process-dependent materials remain sparsely sampled. Analytical scrutiny therefore supports the development of provenance-sensitive steering logics capable of dynamically weighting data sources according to epistemic

credibility, diversity coverage, and experimental grounding. In this framing, ingestion becomes an epistemic governance process rather than a passive aggregation mechanism.

The Embedding Transformation Layer introduces a second locus of analytical consequence through abstraction-induced distortion. Learning architectures, particularly graph neural networks, encode relational structures via message-passing operations that privilege local bonding environments. While such architectures enhance predictive resolution for atomistic properties, they may attenuate global or emergent phenomena, including mesoscale ordering effects or synthesis-dependent phase behaviors. Dimensional compression further intensifies this abstraction tension. Reduced latent spaces enable scalable screening and computational tractability, yet they risk collapsing high-variance physicochemical signals into homogenized manifolds, thereby obscuring rare but scientifically consequential features. From an epistemic standpoint, representation learning thus constitutes both an inferential enabler and a distortion amplifier.

The accumulation of epistemic distortion across representational strata may be formalized as a cascading risk dynamic:

$$Rn = Rn - 1 \cdot \kappa + \delta n \quad (1)$$

where Rn is the risk at layer n, κ a propagation factor, and δn the layer-specific distortion increment. This formula captures the interaction between inherited risks and new introductions, illustrating how unmitigated distortions compound in sequential transformations. The ERC's feedback loops offer a mechanism for mitigating these, by enabling iterative recalibration where high-uncertainty embeddings trigger upstream data augmentations, thus refining inference pathways in high-throughput systems [4, 27]. Layer-specific distortion pathways and their corrective mitigation logics are synthesized in **Table 2.**

Table 2. Epistemic Distortion Cascades and Feedback Mitigation Mechanisms

Cascade Stage	Distortion Mechanism	Manifestation in AI Workflows
Data Provenance Encoding	Dataset bias; underrepresentation; experimental sparsity	Skewed training distributions; incomplete chemical coverage
Representational Compression	Dimensional reduction; latent abstraction	Loss of rare physicochemical signals; homogenized embeddings
Architectural Encoding Bias	Model inductive priors; invariance assumptions	Overemphasis on local bonding environments
Steering Optimization Bias	Objective function narrowing; screening thresholds	Overprioritization of high-confidence predictions
Closed-Loop Amplification	Autonomous retraining on biased outputs	Recursive distortion reinforcement
Cross-Modal Misalignment	Simulation–experiment divergence	Inconsistent embedding calibration

At the Discovery Steering Layer, implications extend to workflow integrations, where epistemic risks manifest in design outcomes. Inverse design pipelines, reliant on embeddings for property-to-composition mapping, may generate candidates that align with representational ideals but diverge from synthesizable realities [8, 20]. The ERC interprets this as a steering logic challenge, advocating for hybrid rules that incorporate interpretability metrics alongside predictive scores [19, 32]. Such logics could prioritize embeddings with lower epistemic distortion, fostering more robust closed-loop experimentation by aligning computational predictions with experimental feedback [17, 28]. The feedback loop mitigation can be conceptualized as:

$$\Delta R = -\lambda (F \otimes U) \quad (2)$$

where ΔR represents risk reduction , λ the loop efficacy, F the feedback signal, and U the uncertainty vector. This notation formalizes the trade-off in corrective dynamics, showing how feedback interacts with uncertainty to dampen epistemic divergences.

At the Discovery Steering Layer, epistemic distortions transition from representational conditions into actionable design consequences. Inverse design systems, which map target properties onto candidate compositions through embedding navigation, may generate outputs that align with representational optima while diverging from synthesizable or scalable material realities. This divergence reflects a steering logic misalignment between predictive abstraction and experimental feasibility. Analytical interpretation therefore supports hybrid steering architectures in which interpretability metrics, synthesizability constraints, and uncertainty thresholds operate alongside predictive performance scores. Such integrated logics enable prioritization of candidates whose embeddings exhibit lower epistemic distortion, fostering closer alignment between computational exploration and laboratory validation.

Broader systems-level analysis reveals intensified cascade vulnerabilities in multimodal infrastructures where simulation, experimental, and literature datasets converge. Epistemic inconsistencies across modalities—arising from measurement noise, reporting bias, or simulation approximations—may interact multiplicatively rather than additively if left unmediated. Feedback-conditioned harmonization mechanisms thus become essential for maintaining cross-modal representational coherence. This systems perspective further surfaces infrastructural trade-offs in autonomous discovery environments, where scalability, automation, and compression efficiencies must be balanced against epistemic fidelity. Accelerated screening pipelines, while operationally efficient, risk propagating distortions at velocities that outpace corrective oversight.

Representation learning architectures also shape discovery horizons through latent space topology. Encoding schemes emphasizing invariance or symmetry may constrain exploration to historically validated materials classes, limiting extrapolative discovery potential. ERC interprets this as representation-bounded exploration, wherein the geometry of embeddings subtly steers the epistemic directionality of materials innovation itself [22].

Collectively, these analytical implications reposition embeddings from passive computational intermediaries to active epistemic agents embedded within discovery infrastructures. Embeddings function simultaneously as carriers, amplifiers, and modulators of epistemic risk, shaping both inferential confidence and exploratory direction. This reconceptualization supports computational design paradigms in which uncertainty quantification, interpretability diagnostics, and feedback conditioning are integrated directly into AI discovery architectures. Through iterative feedback evolution, representational systems progressively align with experimental realities, enhancing physical congruence and inferential robustness across materials AI ecosystems.

Results and Discussion

The ERC framework integrates epistemic limits into the broader discourse of computational materials engineering, revealing how learned embeddings mediate between data infrastructures and discovery outcomes. This interpretive approach aligns with ongoing efforts in materials informatics to enhance system robustness, particularly in light of representation-induced constraints [7, 16]. By emphasizing cascade dynamics, the ERC extends discussions on AI-guided systems, where closed-loop integrations often amplify epistemic gaps if representational fidelity is overlooked [27, 28]. For example, in high-throughput screening, embeddings derived from biased datasets can steer workflows toward suboptimal paths, echoing literature concerns about generalization limits [2, 10].

A key discussion point revolves around representation-inference interactions, where the ERC's layers illuminate trade-offs in learning architectures. Deep learning models, while adept at capturing complex patterns, introduce abstractions that challenge interpretability in design contexts [5, 19]. This resonates with critiques of black-box behaviors in foundation models, suggesting that epistemic scrutiny could inform more transparent architectures without sacrificing performance [22, 30]. Furthermore, the framework's feedback loops contribute to dialogues on uncertainty quantification, proposing adaptive mechanisms that align computational predictions with physical validations [18, 21].

In terms of discovery steering logics, the ERC fosters debate on balancing efficiency and reliability in inverse

design. Traditional paradigms prioritize rapid candidate generation, yet epistemic distortions may lead to infeasible materials, as observed in perovskite or polymer applications [11, 13]. The ERC interprets this as an opportunity for infrastructure-aware strategies, where steering incorporates epistemic risk assessments to refine workflows [20, 32]. This perspective also intersects with autonomous discovery, advocating for systems that evolve through data-model interactions, mitigating limits in simulation-experiment coupling [25, 31].

Broader implications for materials AI ecosystems highlight the need for integrative approaches that transcend isolated components. The ERC's systems-level insights encourage cross-disciplinary synthesis, drawing from catalysis informatics and nanocluster modeling to address universal epistemic challenges [14, 15]. By framing embeddings as epistemic conduits rather than definitive realities, the framework prompts reevaluation of data infrastructures, ensuring they support resilient cascades [23, 26]. Ultimately, this discussion underscores the ERC's role in advancing computational design, promoting workflows that prioritize epistemic harmony for sustainable innovation in materials engineering.

Conclusion

In computational and data-driven materials engineering, learned embeddings have revolutionized discovery pipelines, yet their epistemic limits underscore a fundamental disconnect: representation is not reality. The Epistemic Representation Cascade (ERC) framework provides a novel interpretive structure to navigate these limits, dissecting data-model-discovery interactions through layered cascades and feedback mechanisms. By revealing epistemic risk structures and infrastructure trade-offs, the ERC offers systems-level insights that enhance computational steering logics, fostering more interpretable and robust AI ecosystems.

This work integrates recent advancements in materials informatics, representation learning, and autonomous systems, highlighting the need for epistemic-aware strategies in high-throughput and inverse design paradigms. As materials science progresses toward foundation models and closed-loop automation, the ERC advocates for balanced workflows that mitigate distortions while leveraging representational strengths. Future directions may explore extending the cascade to emerging multimodal integrations, ensuring computational design remains grounded in physical fidelity.

Ultimately, embracing these epistemic considerations will empower materials engineers to transcend current constraints, driving innovative discoveries with greater trustworthiness and efficiency.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Jun 2021 Revised: 28 Sep 2021 Accepted: 19 Oct 2021
Published online: 18 March 2022

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, Shi J, Borlido P, Chen L, Botti S, Marques MAL. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. npj Comput Mater. 2017;3(1):52.

Faber FA, Christensen AS, Huang B, von Lilienfeld OA. Alchemical and structural distribution based representation for universal quantum machine learning. npj Comput Mater. 2018;4(1):32.

Ahmad Z, Xie T, Maheshwari C, Grossman JC, Viswanathan V. Machine learning enabled computational screening of inorganic solid electrolytes for all-solid-state batteries. npj Comput Mater. 2018;4(1):15.

Chen C, Ye W, Zuo Y, Zheng C. Graph networks as a universal machine learning framework for molecules and crystals. npj Comput Mater. 2019;5(1):60.

Pilania G, Iverson CN, Lookman T, Welck H. Machine learning for molecular and materials science. npj Comput Mater. 2019;5(1):82.

Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse materials design. npj Comput Mater. 2020;6(1):84.

Park CW, Wolverton C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. npj Comput Mater. 2020;6(1):63.

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

Fung V, Hu G, Ganesh P, Sumpter BG. Machine learned features from density functional theory: Models for materials synthesis predictions. npj Comput Mater. 2021;7(1):11.

Kim C, Batra R, Chen L, Tran H, Ramprasad R. Polymer design using genetic algorithm and machine learning. npj Comput Mater. 2021;7(1):7.

Zhu Y, Xu T, Zhang H, Hu X, Zhao L, Lu C, et al. Inverse design of two-dimensional materials with invertible neural networks. npj Comput Mater. 2021;7(1):184.

Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat Commun. 2018;9(1):3405.

Zeni C, Rossi K, Glielmo A, Fekete Á, Gaston N, Baletto F, et al. Building machine learning force fields for nanoclusters. Nat Commun. 2018;9(1):3747.

Toyao T, Maeno Z, Takakusagi S, Kamachi T, Takigawa I, Shimizu KI. Machine learning for catalysis informatics: Recent applications and prospects. Nat Commun. 2020;11(1):357.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. Nat Commun. 2018;9(1):5100.

Saal JE, Oliynyk AO, Meredig B. Machine learning in materials discovery: Confirmed predictions and their underlying approaches. Nat Commun. 2020;11(1):5536.

Chen D, Bai Y, Zhao W, Wu F, Ma J, Shen C. Deep reasoning networks for unsupervised pattern de-mixing with constrained convolution loss. Nat Mach Intell. 2021;3(9):758-67.

Oviedo F, Ren Z, Sun S, Settens C, Liu Z, Hartono NTP, et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. Nat Mach Intell. 2019;1(11):532-40.

Saidi P, Campbell Z, Ghadbeigi L. Inverse design of solid-state materials via a continuous representation. Matter. 2019;1(5):1175-94.

St John PC, Phillips C, Kemper TW, Wilson AN, Crowley MF, Nimlos MR, et al. Message-passing neural networks for high-throughput polymer screening. Matter. 2019;1(4):980-97.

Rosen AS, Iyer S, Ray D, Yao Z, Aspuru-Guzik A, Gagliardi L. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. Matter. 2021;4(5):1578-97.

Friedman B, Li Y, Sanchez-Lengeling B, Aspuru-Guzik A. Machine learning for inverse design of solid-state materials. Matter. 2019;1(5):1138-40.

Szymanski NJ, Zeng Y, Huo M, Bartel CJ, Kim H, Ceder G. Toward autonomous design and synthesis of novel inorganic materials. Matter. 2021;4(9):2704-35.

Mrdjenovich D, Horton M, Montoya JH, Walsworth VL, Blanchard PER, Gagliardi L. Propnet: A knowledge graph for materials science. Matter. 2020;2(2):464-80.

Pendleton IM, Cattabriga G, Li Z, Najeeb MA, Friedler SA, Norquist AJ, et al. Experiment specification, capture and laboratory automation technology (ESCALATE): A software pipeline for automated chemical experimentation and data management. Matter. 2019;1(6):1575-92.

Moosavi SM, Nearing GS, Chhatre K, Novikov P, Smit B. Autonomous materials discovery driven by data automation. Matter. 2021;4(7):2220-37.

Abolhasani M, Kumara K. Autonomous closed-loop experimental design for materials discovery. Matter. 2020;2(4):850-2.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Chem Rev. 2018;118(10):5107-44.

Deringer VL, Bartók AP, Proserpio DM, Day GM, Csányi G, Pickard CJ. Machine learning for crystal identification and discovery. Chem Rev. 2021;121(16):10073-141.

Meredig B, Antono E, McCulloch S, Rajan K, Lopez SA. High-throughput machine learning recommendation system for materials discovery. Comput Mater Sci. 2018;145:285-94.

Stein HS, Gregoire JM. Progress and prospects for accelerating materials science with automated and autonomous workflows. Adv Intell Syst. 2019;1(6):1900048.