ORIGINAL RESEARCH **Open access**

Latent Variable Leakage in Materials AI: A Conceptual Risk Framework

Fatima Zahra Amrani[1*], Youssef Benali[1]

Abstract

In the evolving landscape of materials artificial intelligence (AI), latent variables serve as compressed representations that underpin model architectures, facilitating the interpretation of complex material properties and behaviors. This manuscript explores the conceptual dimensions of latent-variable leakage, in which unintended informational flows within these representations may influence systemic outcomes in materials discovery and design. Through an integrative analysis of theoretical underpinnings, the discussion elucidates interaction dynamics between latent spaces and external variables, highlighting epistemic trade-offs in model transparency and generalization. The synthesis of recent literature reveals patterns in how leakage manifests across generative and predictive frameworks, emphasizing steering logics that balance representational fidelity with risk mitigation. A proposed conceptual framework interprets these dynamics as interconnected feedback structures, where leakage pathways intersect with domain-specific constraints in materials science. Ethical reasoning underscores the implications for equitable innovation, while systems-level insights advocate for reflexive approaches in AI deployment. This work contributes to scholarly discourse by framing leakage not as isolated anomalies but as inherent aspects of latent encoding, informing interpretive strategies for sustainable AI integration in materials research.

Keywords Materials AI, Conceptual framework, Latent variables, Information leakage, Epistemic risks, Model dynamics

*Correspondence:
Fatima Zahra Amrani
fatima.amrani@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, Mohammed V University, Rabat, Morocco

Introduction

The integration of artificial intelligence (AI) into materials science has reshaped the paradigms of discovery, design, and optimization, enabling the navigation of vast chemical and structural spaces that were previously intractable through traditional methodologies. At the core of many AI-driven approaches lies the utilization of latent variables—abstract, lower-dimensional encodings that capture essential features of high-dimensional data such as molecular configurations, crystal symmetries, or property spectra. These variables facilitate bridging observable phenomena with underlying patterns, enabling interpretive insights into material behaviors without direct empirical enumeration. However, the conceptual reliance on such encodings introduces subtle interaction dynamics, where the boundaries between intended representations and extraneous influences become permeable, potentially altering the trajectory of AI-assisted materials innovation.

In materials AI, latent variables often emerge from techniques such as variational autoencoders (VAEs) and generative models, which compress input data into probabilistic spaces for tasks ranging from property prediction to structure generation [1-3]. This compression inherently involves trade-offs, as the latent space must balance informativeness with dimensionality reduction, ensuring that key material attributes—such as band gaps, mechanical strengths, or catalytic efficiencies—are preserved while extraneous noise is filtered. Yet, the interpretive lens reveals that these spaces are not

neutral; they embody systemic biases from training datasets, algorithmic assumptions, and domain-specific constraints, leading to unintended informational interplays. For instance, in generative frameworks for crystalline materials, the latent encodings may inadvertently incorporate correlations from biased sampling, influencing the diversity and viability of proposed structures [4, 5].

The notion of leakage in this context extends beyond mere data contamination, encompassing epistemic dimensions where latent variables inadvertently reveal or propagate information that skews model interpretations. In broader machine learning discourse, leakage has been analyzed as a reproducibility concern, in which spurious correlations undermine the generalizability of scientific inferences [6]. In materials AI, this leakage manifests through feedback structures, where latent representations interact with external variables, such as environmental conditions or synthesis parameters, potentially amplifying uncertainties in downstream applications. Ethical reasoning further complicates this picture, as leakage may exacerbate disparities in access to reliable AI tools, particularly in resource-constrained research environments [7, 8].

Scholarly examinations of explainable AI in materials science underscore the need for transparency in latent constructs, revealing how opaque encodings can obscure causal relationships between material inputs and outputs [9, 10]. Systems-level insights suggest that leakage arises from the interplay of multiple factors: the heterogeneity of materials data, the stochastic nature of latent sampling, and the integration of multi-source information [11, 12]. For example, in fusion models that combine diverse datasets, latent variables may leak domain-specific artifacts, thereby altering the interpretability of unified predictions [11]. This dynamic necessitates a conceptual reframing: viewing leakage not as a flaw to be eradicated but as an intrinsic aspect that requires balanced steering logics.

The epistemic risks associated with latent variable leakage extend to the foundational assumptions of materials AI. As models increasingly rely on large language models (LLMs) and generative architectures for data extraction and insight generation, the latent spaces within these systems become conduits for unintended knowledge transfer [10, 13]. Interpretive analyses indicate that such leakage can distort the conceptual mapping of material properties, in which latent dimensions inadvertently encode non-physical variables such as dataset provenance or algorithmic priors [9, 14]. This distortion influences the broader ecosystem of materials research, where AI outputs inform decision-making in areas such as sustainable energy materials or advanced composites.

Moreover, the rapid evolution of AI methodologies in materials science—spanning from diffusion models for peptide generation to flow-based models for crystals—highlights the urgency of addressing leakage at a conceptual level [5, 15]. Interaction dynamics between latent variables and model objectives reveal trade-offs: enhancing expressiveness may increase leakage susceptibility, while constraining the space may limit innovative potential [3, 16]. Ethical considerations emphasize the responsibility of researchers to mitigate these risks, ensuring that AI-driven discoveries align with equitable and transparent principles [17, 18].

This manuscript synthesizes these threads into a cohesive interpretive framework, focusing on the conceptual risks of latent variable leakage in materials AI. By examining theoretical backgrounds and patterns in the literature, it elucidates how leakage shapes systemic outcomes and advocates for integrative strategies that prioritize reflexive awareness. The discussion remains steadfastly conceptual, drawing on analytical implications to foster deeper understanding without venturing into empirical validations. Through this lens, latent variable leakage emerges as a pivotal element in the maturation of materials AI, inviting scholarly reflection on its interpretive ramifications.

Theoretical Background & Literature Synthesis

Latent variables in al-driven materials representation: Latent variables form the interpretive backbone of many AI architectures in materials science, serving as intermediary constructs that distill complex data into manageable forms. These variables enable the conceptual bridging of atomic-scale details with macroscopic properties, facilitating insights into material functionalities without exhaustive enumeration. In generative models, for instance, latent spaces enable the

exploration of design possibilities, where encodings capture probabilistic distributions over structural motifs or chemical compositions [1-3]. The dynamics of these spaces involve continuous interactions between input features and learned representations, where compression trade-offs influence the fidelity of material interpretations.

Scholarly syntheses highlight how latent variables in VAEs and related frameworks encode hierarchical information, ranging from local bonding patterns to global symmetries in materials such as metamaterials and alloys [16, 19]. This encoding process introduces systemic interplays, as latent dimensions may inadvertently align with unobserved confounders, shaping the interpretive landscape of AI outputs. For example, in inverse-design models, latent variables guide navigation through property spaces, but their interactions with domain constraints can lead to skewed conceptual mappings [16]. Ethical reasoning in this context emphasizes representational equity, ensuring that latent constructs do not perpetuate biases arising from imbalanced datasets [7, 20].

Concepts of leakage in machine learning contexts: Leakage in machine learning encompasses the unintended transfer of information across model components, often manifesting as epistemic distortions that affect interpretive reliability. In scientific applications, this phenomenon has been analyzed through the lens of reproducibility and bias, where spurious signals in the training data propagate into inferences [6, 21]. The interaction dynamics of leakage involve feedback loops between data preprocessing, feature extraction, and validation protocols, potentially undermining the conceptual integrity of models [6].

Extending to materials AI, leakage interpretations reveal how latent variables can serve as conduits for such transfers, particularly in multi-modal or fused systems [11, 12]. Systems-level insights indicate that leakage arises from the permeability of model boundaries, where external variables infiltrate latent spaces, altering the balance between generalization and specificity [9, 14]. For instance, in explainable frameworks, leakage may obscure the traceability of decisions, complicating ethical evaluations of AI-assisted discoveries [9, 18]. Trade-offs emerge in steering these dynamics: enhancing model robustness might constrain latent expressiveness, while permissive designs risk amplified leakage [11, 22].

Bias and fairness implications in materials informatics: Bias in AI algorithms intersects with latent-variable dynamics, influencing the conceptual fairness of materials predictions and generation. Literature syntheses underscore how biases embedded in training data—stemming from historical sampling preferences or domain-specific emphases—can leak into latent representations, skewing interpretive outcomes [7, 17, 20]. In materials contexts, this leakage may manifest as uneven coverage of chemical spaces, where certain material classes are over- or under-represented in latent encodings [4, 23].

Ethical reasoning frames these implications as systemic challenges, advocating for integrative approaches that address leakage through reflexive model design [8, 18]. Interaction dynamics between biased inputs and latent variables highlight trade-offs in equity: prioritizing diverse datasets may introduce noise, while curated selections risk entrenching existing disparities [7, 17]. Scholarly discussions emphasize the epistemic value of mitigating such leakage, ensuring that AI in materials science contributes to inclusive innovation without perpetuating conceptual inequities [8, 20].

Generative models and latent space explorations in materials: Generative AI architectures, including diffusion and flow-based models, rely on latent variables to synthesize novel materials, interpreting data distributions through probabilistic lenses [4, 5, 15]. The synthesis of the literature reveals patterns in how these models handle latent dynamics, in which extrapolation predictions depend on the stability of the encoded spaces [3, 12]. Systems-level insights suggest that leakage in generative contexts arises from the interplay of stochastic sampling and boundary conditions, potentially leading to interpretive divergences in generated structures [2, 5].

In materials discovery, such models enable conceptual explorations of uncharted territories, but leakage introduces steering challenges, as latent variables may carry over artifacts from source domains [10, 13]. Trade-offs in model design —between creativity and control—underscore the need for balanced frameworks that interpret leakage as opportunities

for enhanced reflexivity [4, 16]. Ethical dimensions further integrate these discussions, highlighting how leakage affects the trustworthiness of AI-generated insights in collaborative research environments [10, 22].

Integration of multi-source data and epistemic risks: The fusion of heterogeneous data sources in materials AI amplifies conceptual risks associated with latent-variable leakage, as disparate inputs converge in shared representational spaces [11, 12]. Interpretive analyses indicate that this integration fosters complex feedback structures, in which leakage pathways emerge from mismatches in scale or modality [11, 14]. For materials applications, such as symmetry prediction or morphology forecasting, these dynamics influence the epistemic grounding of AI outputs [19, 23].

Literature patterns emphasize the trade-offs in multi-source approaches: while they enrich latent variables with comprehensive insights, they also heighten susceptibility to informational spills [12, 13]. Systems-level reasoning advocates for interpretive strategies that map these risks, ensuring that leakage does not erode the conceptual coherence of unified models [9, 21]. Ethical considerations reinforce this by framing leakage mitigation as a collective responsibility, promoting transparent dynamics in AI ecosystems [17, 18].

Proposed conceptual framework

The conceptual framework interprets latent-variable leakage in materials AI as an interwoven set of dynamics in which representational spaces interact with external influences through permeable boundaries. At its core, this framework elucidates steering logics that navigate the trade-offs between compression efficiency and informational integrity, framing leakage as a systemic feature rather than an aberration. Interaction dynamics are central: latent variables, as encoded abstractions, engage in feedback with input domains, model architectures, and output interpretations, potentially channeling unintended variables that reshape material insights.

Epistemic reasoning within the framework highlights how leakage affects the interpretive validity of AI in materials science, emphasizing reflexive awareness in design choices. For instance, in generative contexts, leakage may manifest as diffuse correlations, in which latent dimensions inadvertently integrate non-material factors such as dataset artifacts, thereby influencing the conceptual diversity of proposed designs. Systems-level insights integrate these elements, portraying the framework as a network of interconnected nodes—representing data sources, encoding processes, and deployment contexts—linked by leakage pathways that modulate overall system behavior.

Trade-offs emerge prominently: enhancing latent expressiveness for broader material exploration may amplify leakage risks, while restrictive encodings could limit innovative potential, creating a balanced tension in steering mechanisms. Ethical dimensions infuse the framework with considerations of equity, interpreting leakage as a vector for bias propagation that demands integrative safeguards. Analytical implications suggest that understanding these dynamics fosters more robust conceptual mappings, where materials AI evolves through heightened awareness of latent interplays. The systemic risk of latent variable leakage is conceptualized in **Figure 1** as a multi-layered ecosystem, where biases from data, algorithms, and domains permeate a model's core representations and distort its predictions and generations.

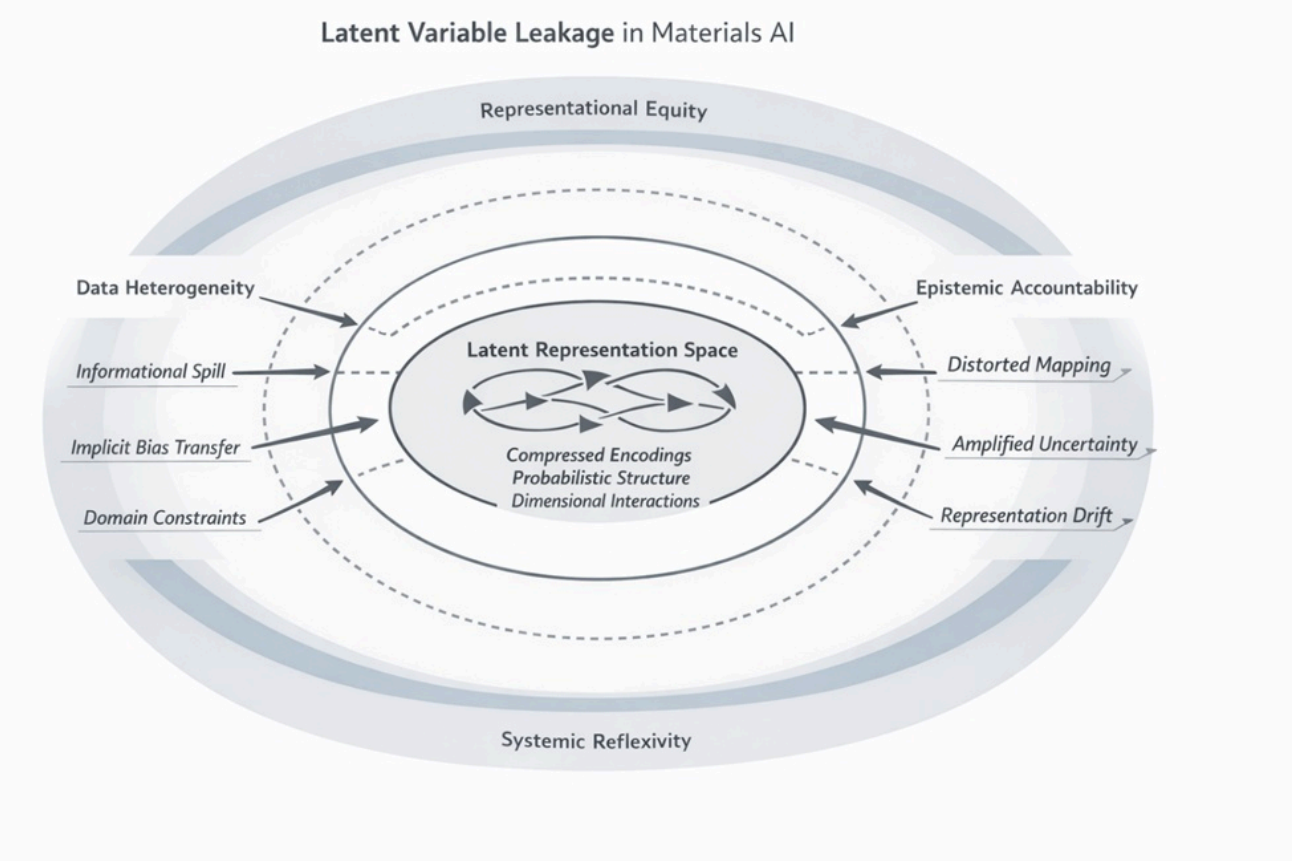

Figure 1. A conceptual schematic of latent variable leakage in materials AI. The diagram depicts leakage as a cyclical process originating in an elliptical latent space core, with permeable barriers mediating the ingress of external biases and the egress toward AI outputs, all enveloped by ethical and epistemic overlays.

The systemic risks associated with latent-variable leakage across representational, analytical, and ethical dimensions are summarized in **Table 1**.

Table 1. Conceptual risk framework for latent variable leakage in materials AI

Framework dimension	Source of leakage	Interaction dynamics	Epistemic/analytical risk	Steering logic (conceptual response)
Latent representation design	Compression trade-offs in VAEs, diffusion, and flow-based models	High-dimensional abstractions entangle intrinsic material features with dataset artifacts	Reordering of material salience; shift from physicochemical causality to representational correlation	Reflexive latent design prioritizing interpretability over maximal expressiveness
Training data and domain bias	Imbalanced material classes, historical sampling preferences	Biased distributions imprint asymmetries into latent geometries	Marginalization of underrepresented material domains; distorted exploratory attention	Representational equity through conceptual awareness of dataset–latent coupling

Generative exploration	Stochastic sampling and boundary extrapolation	Leakage enables cross-domain associations beyond intended constraints	Serendipity paired with epistemic uncertainty and speculative stabilization	Interpret leakage as an exploratory signal, not evidentiary confirmation
Multi-source data fusion	Heterogeneous modalities and mismatched scales	Latent variables act as convergence points for incompatible semantics	Loss of conceptual coherence; amplification of structural inconsistencies	Analytical mapping of fusion boundaries and modality-specific interpretive limits
Model objectives and optimization	Alignment between loss functions and latent structure	Optimization pressures reinforce certain leakage pathways	Feedback loops that stabilize artifactual correlations	Steering objectives toward epistemic resilience rather than performance alone
Explainability and transparency	Opaque latent encodings	Latent behaviors co-evolve with interpretive narratives	Obscured causal reasoning; reduced traceability	Latent-centric interpretive analysis instead of post-hoc explanation
Ethical and equity dimensions	Shared AI infrastructures and uneven access	Leakage propagates representational advantages across communities	Unequal knowledge production and interpretive authority	Governance-aware steering of latent representations as socio-epistemic assets
System-level feedback	Recursive interaction of data, models, and interpretation	Leakage operates as a persistent structural condition	Misattribution of confidence and stability in AI outputs	Conceptual governance treating leakage as diagnostic rather than anomalous

Analytical implications

The interpretive ramifications of latent-variable leakage in materials AI extend well beyond isolated modeling artifacts, reshaping the broader analytical landscape in which research practices assign meaning, relevance, and scientific priority. At an analytical level, interaction dynamics within latent spaces reveal that leakage subtly reorders the conceptual salience of material attributes. Unintended informational flows may amplify specific representational pathways while suppressing others, thereby redirecting analytical attention away from intrinsic physicochemical properties toward correlations that emerge primarily from representational entanglement rather than material causality [6, 14]. Such shifts challenge conventional assumptions of attribute independence and necessitate analytical frameworks that explicitly account for how latent encodings mediate interpretation.

From a systems perspective, these dynamics encourage a reevaluation of model-dependency structures in materials AI pipelines. Latent variable leakage foregrounds the reciprocal feedback between learned representations and domain interpretation, highlighting how analytical conclusions are co-produced by data distributions, model architectures, and interpretive conventions. Rather than treating latent spaces as neutral intermediaries, this perspective reframes them as active analytical agents that shape scientific narratives. Consequently, analytical strategies increasingly emphasize reflexive assessment of how representational overlap, compression, and abstraction influence downstream reasoning and conceptual framing.

Within generative modeling contexts, the analytical implications of leakage become particularly pronounced. Leakage may act as a catalyst for creative exploration by enabling unexpected associations across compositional, structural, or property domains, thereby supporting serendipitous hypothesis generation [3-5]. However, this creative potential is accompanied by heightened epistemic uncertainty, as emergent associations may lack clear grounding in established

physical mechanisms. Analytical reasoning must therefore navigate a delicate trade-off: leveraging leakage-induced novelty while resisting premature stabilization of speculative patterns into authoritative claims. This tension underscores the need for interpretive restraint and conceptual boundary-setting when integrating generative outputs into scientific analysis.

Ethical reasoning intersects with these analytical implications by drawing attention to the collective dimensions of interpretation. In collaborative and multi-institutional settings, shared AI tools embed latent representations that may differentially privilege certain data regimes, material classes, or epistemic assumptions [7, 17, 18]. Latent variable leakage can thus propagate representational asymmetries across research communities, influencing which materials questions are foregrounded and which are marginalized. Analytical equity, in this sense, depends on recognizing leakage not merely as a technical concern but as a socio-epistemic factor that shapes participation, interpretive authority, and knowledge circulation.

Steering logics emerge as a critical analytical response to these challenges. Rather than seeking to eliminate leakage outright, steering-oriented frameworks interpret leakage as a diagnostic signal—an indicator of where conceptual boundaries blur, representations overgeneralize, or domain assumptions remain underarticulated. Analytical reflexivity, informed by these signals, supports iterative refinement of conceptual models and interpretive mappings, aligning representational flexibility with epistemic accountability.

In multi-source and heterogeneous data environments, latent-variable leakage further heightens sensitivity to systemic interactions across datasets, modalities, and modeling layers [11, 18]. Leakage exposes vulnerabilities in data fusion processes, revealing how representational coupling across sources may introduce subtle biases or amplify structural inconsistencies. Analytical approaches attuned to these dynamics balance comprehensiveness with precision, treating leakage not as a disruptive anomaly but as evidence of underlying complexity within materials data ecosystems [9, 13]. Such an orientation encourages deeper interrogation of how material knowledge is synthesized across scales and representations.

Collectively, these analytical implications advocate for a paradigm that does not position latent-variable leakage solely as a risk to be mitigated, but rather as an interpretive phenomenon to be understood and strategically engaged. By integrating leakage awareness into analytical reasoning, materials AI research can transform representational ambiguity into an opportunity for enriched conceptual insight, fostering more reflexive, adaptive, and epistemically grounded scientific inquiry.

Results and Discussion

Integrating the conceptual elements of latent variable leakage reveals a dense web of interaction dynamics that permeate contemporary materials AI ecosystems. When synthesized with the theoretical foundations and the proposed framework, leakage emerges not as a singular modeling deficiency but as a structural condition of representation, in which latent spaces serve as nexus points for both informational convergence and divergence [1, 2, 10]. This perspective foregrounds feedback structures in which leakage pathways interact recursively with optimization objectives, data regimes, and interpretive conventions, collectively shaping the conceptual coherence of AI-driven materials insights without isolating linear or attributable causal chains.

Within this interpretive framing, representational trade-offs become central. Expansive latent spaces—designed to support broad exploration across composition, structure, and property domains—also increase susceptibility to informational spillover, entanglement, and abstraction drift [15, 16, 19]. These dynamics challenge assumptions that representational richness and epistemic reliability scale together. Instead, leakage exposes a tension between expressive capacity and epistemic containment, reinforcing the need for steering mechanisms that prioritize epistemic resilience over representational maximalism. Such steering logics do not seek representational austerity but rather cultivate controlled openness, in which leakage is rendered interpretable rather than suppressed.

Ethical considerations are inseparable from these representational dynamics. Latent variable leakage functions as a lens through which fairness and equity in materials AI can be examined, particularly in contexts where dominant material classes, data-rich chemistries, or well-studied structures implicitly shape latent geometries [8, 20]. Unintended informational flows may disproportionately marginalize underrepresented material domains, reinforcing asymmetries in exploratory attention and scientific visibility. From a systems-level standpoint, addressing these inequities requires reframing materials AI as adaptive socio-technical systems, where leakage-induced perturbations signal imbalances in representation, governance, and interpretive authority.

The literature synthesis further clarifies how leakage intersects with parallel advances in explainability and generativity. Analytical engagement with latent behaviors—rather than post hoc rationalization—supports more transparent interpretive practices by exposing how internal representations co-evolve with analytical claims [9, 21, 22]. In generative settings, leakage dynamics complicate conventional notions of extrapolation and generalization. Rather than treating extrapolative instability as a failure mode, this framework interprets leakage as a productive site for iterative conceptual refinement, where representational ambiguity invites epistemic caution, hypothesis re-articulation, and boundary re-negotiation [3, 12]. This reframing positions leakage awareness as an enabler of integrative reasoning rather than a constraint on generative ambition.

Epistemic reasoning deepens this discussion by interrogating the presumed neutrality of latent encodings. Leakage reveals that latent spaces are not passive compressions of material reality, but active participants in knowledge production, embedding value-laden assumptions about similarity, relevance, and significance [6, 23]. These effects intensify in multimodal and multisource environments, where leakage blurs the boundaries between experimental, simulated, and inferred data streams [11, 13]. While such blurring can enrich conceptual mappings by fostering cross-domain synthesis, it also complicates interpretive accountability, demanding heightened reflexivity in how material knowledge is assembled and communicated.

Collectively, the discussion suggests that latent-variable leakage should be interpreted not as an anomaly to be eliminated, but as an inherent feature of high-dimensional representations that requires conceptual governance. By engaging leakage through interpretive, ethical, and systems-level lenses, materials AI can mature beyond performance-centric paradigms toward more reflexive and sustainable modes of scientific inquiry.

Conclusion

This conceptual exploration of latent variable leakage in materials AI illuminates the interpretive depth of representational risks, reframing them as integral to the field’s ongoing epistemic evolution. Through analytical implications and systems-level synthesis, leakage is interpreted as a dynamic interplay among representation, interpretation, and governance—one that actively shapes how material knowledge is constructed, prioritized, and stabilized within AI-assisted research environments.

Trade-offs and steering logics emerge as pivotal organizing principles, guiding the balance between innovation and integrity in latent space design. Rather than privileging unbounded representational expressivity or rigid containment, the framework advocates reflexive steering, in which leakage signals inform conceptual refinement and interpretive restraint. Ethical reasoning reinforces this orientation, emphasizing that unmanaged leakage risks reproduce inequities in material visibility and scientific opportunity, while deliberate engagement can support more inclusive and equitable knowledge trajectories.

As AI increasingly integrates predictive and generative paradigms, these conceptual considerations become even more urgent. Leakage awareness fosters interpretive humility, encouraging researchers to interrogate not only what models predict or generate, but how latent structures condition the meanings ascribed to those outputs. In this way, potential

vulnerabilities are transformed into sources of interpretive richness, strengthening the epistemic foundations of AI-driven materials discovery.

In summary, this framework contributes to scholarly discourse by reinterpreting latent variable leakage through interactional and systemic lenses. By situating leakage within the broader ecology of materials AI, it offers a resilient conceptual foundation for future innovations—one that aligns representational power with epistemic responsibility, and technological advancement with sustainable scientific reasoning.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Jan 2022 Revised: 28 Mar 2022 Accepted: 05 May 2022
Published online: 18 July 2022

References

Lew AJ, Buehler MJ. Encoding and exploring latent design space of optimal material structures via a VAE-LSTM model. Forces Mech. 2021;5:100054.
https://doi.org/10.1016/j.finmec.2021.100054.

Ochiai T, Inukai T, Akiyama M, Furui K, Ohue M, Matsumori N, et al. Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. Commun Chem. 2023;6(1):249.
https://doi.org/10.1038/s42004-023-01054-6.

Hatakeyama-Sato K, Oyaizu K. Generative models for extrapolation prediction in materials informatics. ACS Omega. 2021;6(22):14566-74.
https://doi.org/10.1021/acsomega.1c01716.

De Breuck PP, Wang HC, Rignanese GM, Botti S, Marques MAL. Generative AI for crystal structures: A review. npj Comput Mater. 2025;11(1):370.
https://doi.org/10.1038/s41524-025-01881-2.

Luo X, Wang Z, Wang Q, Shao X, Lv J, Wang L, et al. CrystalFlow: A flow-based generative model for crystalline materials. Nat Commun. 2025;16(1):9267.
https://doi.org/10.1038/s41467-025-64364-4.

Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804.
https://doi.org/10.1016/j.patter.2023.100804.

Ferrara E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci. 2024;6(1):3.
https://doi.org/10.3390/sci6010003.

Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health. 2023;2(6):e0000278.
https://doi.org/10.1371/journal.pdig.0000278.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TYJ. Explainable machine learning in materials science. npj Comput Mater. 2022;8:204.
https://doi.org/10.1038/s41524-022-00884-7.

Jablonka KM, Ai Q, Al-Feghali A, Badhwar S, Bocarsly JD, Bran AM, et al. 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. Digit Discov. 2023;2(5):1233-50.
https://doi.org/10.1039/d3dd00113j.

Ravi SK, Comlek Y, Pathak A, Gupta V, Umretiya R, Hoffman A, et al. Interpretable multi-source data fusion through latent variable gaussian process. Eng Appl Artif Intell. 2025;145:110033.
https://doi.org/10.1016/j.engappai.2025.110033.

Chang R, Wang YX, Ertekin E. Towards overcoming data scarcity in materials science: Unifying models and datasets with a mixture of experts framework. npj Comput Mater. 2022;8(1):242.
https://doi.org/10.1038/s41524-022-00929-x.

Schilling-Wilhelmi M, Ríos-García M, Shabih S, Gil MV, Miret S, Koch CT, et al. From text to insight: Large language models for chemical data extraction. Chem Soc Rev. 2025;54(3):1125-50.
https://doi.org/10.1039/D4CS00913D.

Kayode GO, Montemore MM. Latent variable machine learning framework for catalysis: General models, transfer learning, and interpretability. JACS Au. 2024;4(1):80-91.
https://doi.org/10.1021/jacsau.3c00419.

Wang YJ. Artificial intelligence using a latent diffusion model enables the generation of diverse and potent antimicrobial peptides. Sci Adv. 2025;11(6):eadp7171.
https://doi.org/10.1126/sciadv.adp7171.

Tran TV, Nanthakumar SS, Zhuang X. Deep learning-based framework for the on-demand inverse design of metamaterials with arbitrary target band gap. npj Artif Intell. 2025;1(1):2.
https://doi.org/10.1038/s44387-025-00001-1.

Igbinovia MO, Danquah MM. Artificial intelligence algorithm bias in information retrieval systems and its implication for library and information science professionals: A scoping review. Tech Serv Q. 2025;42(4):253-79. https://doi.org/10.1080/07317131.2025.2512282.

Gervasi SS, Chen IY, Smith-McLallen A, Sontag D, Obermeyer Z, Vennera M, et al. The potential for bias in machine learning and opportunities for health insurers to address it. Health Aff. 2022;41(2):265-73. https://doi.org/10.1377/hlthaff.2021.01287.

Cao Z, Liu Q, Liu Q, Yu X, Kruzic JJ, Li X. A machine learning method to quantitatively predict alpha phase morphology in additively manufactured Ti-6Al-4V. npj Comput Mater. 2023;9(1):195. https://doi.org/10.1038/s41524-023-01152-y.

Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inform. 2024;154:104646. https://doi.org/10.1016/j.jbi.2024.104646.

Witman MD, Schindler P. MatFold: Systematic insights into materials discovery models' performance through standardized cross-validation protocols. Digit Discov. 2025;4(1):625-35. https://doi.org/10.1039/D4DD00250D.

Chávez-Angel E, Eriksen MB, Castro-Alvarez A, Garcia JH, Botifoll M, Avalos-Ovando O, et al. Applied artificial intelligence in materials science and material design. Adv Intell Syst. 2025;7(8):2400986. https://doi.org/10.1002/aisy.202400986.

Alghofaili YA, Alghadeer M, Alsaui AA, Alqahtani SM, Alharbi FH. Accelerating materials discovery through machine learning: Predicting crystallographic symmetry groups. J Phys Chem C. 2023;127(33):16645-53. https://doi.org/10.1021/acs.jpcc.3c03274.