ORIGINAL RESEARCH **Open access**

Data Lineage and Scientific Traceability in Computational Materials Pipelines

Lucas Andrade[1*], Mariana Lopes[1]

Abstract

In the evolving landscape of computational and data-driven materials engineering, the integration of high-throughput simulations, machine learning models, and autonomous discovery systems has accelerated materials innovation. However, the complexity of these pipelines often obscures the origins and transformations of data, leading to challenges in reproducibility, error propagation, and epistemic accountability. This conceptual manuscript addresses the critical need for robust data lineage and scientific traceability mechanisms within computational materials workflows. We introduce a novel framework, the Integrated Traceability Architecture (ITA), which conceptualizes traceability as a multilayered system embedding provenance tracking across data generation, model training, and discovery iterations. By synthesizing recent advancements in materials informatics, representation learning, and uncertainty quantification, the framework elucidates how lineage-aware pipelines can enhance decision-making in inverse design and closed-loop experimentation. Implications extend to fostering reliable multimodal datasets, optimizing simulation-experiment couplings, and mitigating risks in foundation models for materials science. This work provides a systems-level perspective on traceability, promoting infrastructure designs that balance computational efficiency with scientific integrity, ultimately steering towards more transparent and accelerated materials discovery paradigms.

Keywords Materials informatics, Uncertainty quantification, Representation learning, Data lineage, Scientific traceability, Computational materials pipelines

*Correspondence:
Lucas Andrade
lucas.andrade@outlook.com
[1] Department of Materials Data Science, Faculty of Engineering, Federal University of Minas Gerais, Belo Horizonte, Brazil

Introduction

The advent of data-driven approaches in materials engineering has fundamentally transformed traditional discovery paradigms, shifting the field from empirically intensive trial-and-error practices toward computationally orchestrated exploration strategies [1]. This transition has been catalyzed by the rapid proliferation of machine learning methodologies tailored to materials-specific challenges, including property prediction, structural optimization, synthesis planning, and performance forecasting across multiscale systems [2-4]. Within computational materials science, discovery pipelines now span a broad methodological spectrum—from high-throughput density functional theory calculations to advanced representation learning architectures capable of encoding atomic and crystallographic complexity. Graph neural networks, in particular, have enabled structurally faithful embeddings that support predictive inference across expansive chemical and compositional spaces [5, 6].

As these pipelines expand in sophistication—integrating transfer learning regimes, multimodal datasets, and autonomous decision loops—the interpretability of scientific workflows becomes increasingly obscured. Analytical outputs often emerge from deeply layered transformations whose intermediate steps remain poorly documented or epistemically inaccessible. In this context, data lineage—defined as the comprehensive record of data origins, transformations, dependencies, and downstream

utilizations—emerges as a foundational requirement for ensuring the reliability, reproducibility, and interpretability of computational outputs [7, 8]. Far from a metadata convenience, lineage functions as a structural backbone that anchors predictive inference to verifiable evidentiary histories.

Evolution of computational materials ecosystems

Over the past decade, materials informatics has matured into a central pillar of contemporary materials engineering, leveraging large-scale datasets and artificial intelligence to decode structure–property relationships with unprecedented resolution [1, 9]. High-throughput computational infrastructures have democratized access to simulated materials repositories, enabling rapid screening of candidate compounds across application domains such as energy storage, catalysis, and structural alloys [10-12]. Parallel advances in deep learning architectures have reshaped representational paradigms. Crystal graph convolutional neural networks and related frameworks now translate atomic configurations into relational embeddings capable of capturing bonding environments, symmetry operations, and lattice topologies with high predictive fidelity [5, 13].

These computational advances are increasingly complemented by autonomous discovery systems that integrate robotic experimentation with AI-guided computational steering, effectively closing the loop between simulation and empirical validation [14-17]. Such systems embody a shift toward self-optimizing discovery infrastructures in which hypothesis generation, testing, and recalibration occur within recursive feedback cycles. However, the growing interconnectivity of these components amplifies systemic risks. Untracked data modifications, undocumented preprocessing steps, or opaque model recalibrations can propagate uncertainties throughout the pipeline. These risks are particularly acute in inverse design contexts, where target properties guide compositional exploration and even minor provenance gaps can distort optimization trajectories [18, 19]. Scientific traceability—extending beyond data provenance to include epistemic justifications for methodological choices—thus becomes indispensable for sustaining trust in computational inference [3, 20].

Challenges in data-driven pipelines

Contemporary computational materials pipelines confront multifaceted challenges in managing data integrity and traceability across distributed analytical environments. Multimodal datasets exemplify this complexity. Integrating experimental spectra, simulated electronic structures, thermodynamic datasets, and literature-derived descriptors requires harmonized frameworks capable of preserving lineage across heterogeneous formats [21-23]. Without such infrastructures, relational coherence between modalities becomes fragile, undermining integrative inference.

Uncertainty quantification introduces additional traceability pressures. Predictive reliability assessments—particularly under sparse data regimes—often remain decoupled from the underlying data histories that shape error propagation [10, 20, 24, 25]. This disconnect complicates efforts to attribute predictive variance to specific sources such as sampling bias, representational insufficiency, or model overfitting. Closed-loop discovery systems further intensify lineage demands. In these architectures, experimental feedback iteratively refines model parameters and screening priorities, creating dynamic dependencies that evolve over successive cycles [14, 15, 26, 27]. Absent robust traceability mechanisms, such feedback loops risk reinforcing biased discovery pathways rather than correcting them.

Empirical accounts within the literature underscore these vulnerabilities. Provenance discontinuities in high-entropy alloy exploration and polymer phase prediction workflows have necessitated retrospective validation efforts, introducing inefficiencies and delaying translational deployment [11, 26]. The emergence of foundation models for scientific applications compounds these concerns. Pretrained on vast, heterogeneous corpora, such architectures aggregate knowledge without always preserving transparent training lineages or dataset genealogies, complicating interpretability and auditability in materials contexts [9, 21]. These developments collectively signal the need for a structural reevaluation of pipeline design—one that embeds traceability as an intrinsic architectural principle rather than a retrospective documentation layer.

The need for traceability-centric frameworks

Existing initiatives within materials research infrastructures have emphasized interoperability, database federation, and

knowledge graph integration to support property inference and discovery analytics [22, 28, 29]. While these efforts enhance data accessibility and relational querying, they frequently underemphasize systematic lineage capture across analytical stages. Event-sourced computational architectures represent a partial advance, enabling provenance tracking within accelerated discovery environments by logging simulation states and analytical transitions [8, 30]. However, such systems often fall short in capturing epistemic rationales—why specific modeling decisions were enacted, how uncertainty thresholds were defined, or under what assumptions screening filters were applied—particularly within distributed computational ecosystems.

This limitation constrains the full potential of simulation–experiment couplings, where traceable data flows could otherwise optimize resource allocation, minimize rediscovery redundancies, and enhance collaborative reproducibility [14, 17, 21]. Against this backdrop, the present manuscript advances the position that data lineage and scientific traceability must be reconceptualized not as auxiliary technical features but as foundational enablers of sustainable computational materials ecosystems. By framing traceability through a systems-level lens, we seek to bridge informatics silos, enabling discovery pipelines that remain resilient to data ambiguities while adaptive to emergent AI paradigms [1, 3, 4].

This introduction establishes the conceptual groundwork for a deeper synthesis of the relevant literature, culminating in the proposal of the Integrated Traceability Architecture (ITA). The ITA framework positions traceability as a pivotal mediating infrastructure within computational materials pipelines—integrating provenance capture with discovery logics to enhance epistemic robustness, reproducibility, and translational reliability across the full spectrum of data-driven materials engineering.

Theoretical Background & Literature Synthesis

The theoretical foundations of data lineage and scientific traceability in computational materials engineering emerge from the convergence of informatics, machine learning, and systems engineering frameworks that collectively shape how knowledge is generated, transformed, and validated across discovery ecosystems. Data lineage, in its most formalized articulation, refers to structured metadata that documents the lifecycle of data entities—from acquisition and preprocessing through successive transformations to downstream analytical utilization. Within computational materials environments, this concept expands dramatically in scale and complexity. High-throughput simulations, automated characterization systems, and multimodal repositories generate vast volumes of interdependent data whose interpretive value depends not solely on numerical outputs but on the contextual scaffolding surrounding their production. Provenance therefore functions not merely as archival record but as epistemic infrastructure, preserving interpretability, reliability, and reusability across iterative computational workflows.

Scientific traceability extends this foundation beyond documenting data pathways to encompass the reasoning architectures embedded within discovery systems. It captures the methodological rationale underlying model selection, feature engineering strategies, and uncertainty handling protocols. In doing so, traceability operationalizes reproducibility not only at the level of datasets but at the level of analytical decision logic. As computational materials engineering increasingly incorporates autonomous and semi-autonomous pipelines, the ability to reconstruct inferential pathways becomes essential for validating discoveries, auditing algorithmic behavior, and ensuring alignment with physical principles.

Foundations in materials informatics and representation learning

Materials informatics provides the operational substrate within which traceability demands intensify. Machine learning models trained on compositional, structural, and processing datasets depend on representation learning to transform raw materials descriptors into model-operational embeddings. Graph neural networks exemplify this paradigm by encoding crystal lattices, bonding environments, and symmetry operations into relational vector spaces that support property prediction across thermal, catalytic, and electronic domains. These representational systems are particularly consequential in inverse design settings, where latent embeddings guide the identification of candidate materials optimized for target functionalities.

Yet each representational transformation—dimensionality reduction, feature augmentation, latent compression—introduces epistemic distance between original measurements and predictive outputs. Without embedded

lineage infrastructures, such transformations risk obscuring bias sources, masking sparsity artifacts, or amplifying spurious correlations. Consequently, explainable AI frameworks are increasingly conceptualized not only as transparency mechanisms but as lineage-preserving architectures capable of logging decision pathways alongside predictive results.

High-throughput computation and autonomous systems

Parallel traceability pressures emerge from high-throughput computational frameworks that automate density functional theory calculations, phase stability analyses, and large-scale property screenings. These infrastructures accelerate discovery by orchestrating thousands of simulations concurrently, often embedded within machine learning–guided optimization loops. Autonomous discovery systems extend this paradigm by coupling hypothesis generation, simulation execution, and experimental validation into recursive feedback cycles.

Within such environments, provenance must capture not only static datasets but dynamic state transitions—how candidate materials were proposed, filtered, recalibrated, or experimentally rejected. Event-sourced data architectures exemplify this approach by recording each analytical state change as a traceable transaction, enabling retrospective interrogation of discovery trajectories. However, distributed computational ecosystems complicate provenance preservation. Multi-tenant simulation clusters, federated databases, and cross-institutional workflows introduce heterogeneity in metadata standards, fragmenting traceability practices. Self-driving laboratories further amplify this challenge, as adaptive experimentation requires persistent lineage tracking across iterative cycles to maintain epistemic continuity.

Uncertainty quantification and multimodal integration

Uncertainty quantification introduces an additional lineage layer within materials AI. Predictive systems frequently operate under sparse sampling regimes, anharmonic effects, or extrapolative inference beyond training distributions. Probabilistic modeling frameworks—particularly Bayesian approaches—address these uncertainties by attaching confidence intervals and posterior distributions to predictions. When integrated with lineage infrastructures, these probabilistic signals become traceable artifacts linking predictive uncertainty back to data quality, representational adequacy, and model assumptions.

Small-data learning paradigms reinforce this necessity, as each datapoint exerts amplified influence on model behavior. Multimodal integration compounds traceability demands further. Contemporary discovery pipelines routinely fuse computational simulations, experimental measurements, and literature-derived textual datasets. Aligning these heterogeneous modalities requires relational infrastructures capable of preserving provenance across format boundaries. Knowledge graphs serve this integrative function by embedding materials entities, properties, synthesis conditions, and performance outcomes within queryable relational networks, enabling lineage navigation across multimodal evidence chains.

Simulation–experiment coupling and foundation models

The coupling of simulations with experimental platforms represents one of the most consequential frontiers for traceability. Real-time integration enables adaptive steering of discovery workflows, where experimental feedback recalibrates simulation priorities and machine learning hypotheses. Foundation models pretrained on expansive scientific corpora introduce new opportunities within this coupling by enabling transferable representations and cross-domain inference. However, their inferential opacity necessitates rigorous traceability scaffolds to mitigate hallucination risks and ensure physical plausibility in materials contexts.

Reinforcement learning–driven inverse design illustrates this requirement clearly. Reward functions, constraint embeddings, and policy updates must remain auditable to confirm that optimization trajectories remain anchored to thermodynamic and kinetic realities. Collaborative discovery ecosystems spanning institutions and industrial stakeholders further underscore the necessity of shared traceability standards capable of preventing informational silos and enabling reproducible cross-platform innovation.

Infrastructure and epistemic considerations

At the infrastructural scale, tensions emerge between computational scalability and provenance fidelity. Large-scale screening systems prioritize throughput and storage efficiency, often relegating detailed lineage logging due to computational overhead. Yet recommender systems for compound discovery, attribute-driven design platforms, and AI-mediated materials selection engines depend fundamentally on curated datasets whose reliability is inseparable from their traceability.

As AI life-cycle integration expands—from molecular simulations to industrial deployment—end-to-end lineage becomes indispensable for validating performance claims, ensuring regulatory compliance, and sustaining reproducibility across translational stages. Physical computing integrations and autonomous production environments further heighten this requirement, demanding traceability frameworks capable of spanning digital–physical interfaces.

Synthesizing these intersecting literatures reveals a structural gap. While discrete tools exist for provenance capture in simulations, interpretability in machine learning, metadata management in databases, and audit trails in autonomous laboratories, their integration remains fragmented. A unified systems-level framework capable of embedding lineage across data acquisition, representation learning, model inference, uncertainty attribution, and experimental feedback remains underdeveloped. Addressing this gap requires reconceptualizing traceability not as an auxiliary metadata layer but as foundational epistemic infrastructure embedded across the full stack of computational materials discovery.

Proposed conceptual framework

To address the identified gaps, we propose the Integrated Traceability Architecture (ITA), a novel conceptual framework that embeds data lineage and scientific traceability as core components of computational materials pipelines. The ITA conceptualizes pipelines as multilayered systems, comprising data generation, model inference, and discovery steering layers, interconnected through feedback loops that propagate traceability metadata. This architecture ensures that every computational artifact—be it a simulated property or a learned representation—carries an immutable lineage record, facilitating epistemic audits and adaptive refinements.

At the foundational data generation layer, raw inputs from high-throughput simulations or multimodal sources are tagged with provenance descriptors, including computational parameters and uncertainty estimates. This layer interfaces with representation learning modules, where transformations (e.g., graph embeddings) are logged to preserve structural fidelities. The model inference layer then leverages these traceable representations for predictions, incorporating uncertainty quantification to flag potential lineage breaks. Finally, the discovery steering layer orchestrates inverse design and closed-loop iterations, using traceability to evaluate feedback efficacy and steer towards unexplored material spaces. Feedback loops span these layers, allowing upstream revisions based on downstream insights, such as recalibrating simulations from experimental discrepancies. The functional distribution of traceability mechanisms across pipeline layers and their associated epistemic implications are summarized in **Table 1**.

Table 1. Traceability Functions Across Computational Materials Pipeline Layers

Pipeline Layer	Core Computational Functions	Traceability Artifacts Captured	Ep W L
Data Generation	High-throughput simulations; experimental acquisition; literature mining	Simulation parameters; instrument metadata; acquisition timestamps	pro repr
Representation Learning	Feature engineering; graph embeddings; multimodal fusion	Transformation logs; embedding genealogy; feature mappings	La am s n
Model Inference	Property prediction; inverse design; foundation model integration	Training datasets; hyperparameters; uncertainty outputs	C pre ha co
Discovery Steering	Autonomous planning; RL	Feedback histories; reward	M ex

	optimization; recommendation systems	functions; experimental outcomes	feed
Closed-Loop Integration	Simulation–experiment recalibration; adaptive retraining	Iteration histories; recalibration triggers	E re ine

Central to ITA is the computational steering logic, which dynamically routes data flows based on lineage quality metrics. For instance, paths with incomplete traceability may trigger alternative branches, ensuring robustness in autonomous systems. This can be conceptualized as a traceability-weighted decision function, where the propagation of a data entity D through a pipeline P is modulated by its lineage completeness $S(D) = P(D) \cdot \alpha(L)$ Here, α represents a scaling factor that interprets lineage integrity, capturing the interaction between data fidelity and pipeline reliability without implying empirical measurement.

Further, ITA introduces epistemic risk structures via a conceptual aggregation of dependencies. Dependencies across layers are modeled as a network where nodes denote artifacts and edges encode transformations, with traceability manifesting as edge annotations. The risk of epistemic drift—arising from untraced propagations—may be expressed as: $R = \sum_{e \in E} (1 - t_e) \cdot w_e$ where t_e is the traceability score of edge e (conceptually between 0 and 1), and w_e its weight reflecting dependency criticality. This formula highlights trade-offs in infrastructure design, illustrating how partial traceability amplifies risks in interconnected pipelines.

An additional dynamic within ITA involves representation-inference interactions, formalized through a loop efficiency construct. The iterative refinement of models via traceable feedback can be captured as: $E = \beta \cdot \int (M \circ T)\, dI$ with M denoting model updates, T traceability inputs, and I inference steps; β interprets convergence tendencies. These formulas underscore ITA's emphasis on interpretive systems dynamics, promoting workflows where traceability informs rather than constrains discovery.

The overall structure is visualized as a schematic with layered modules connected by bidirectional arrows representing feedback, and traceability threads weaving through each component, as conceptualized in **Figure 1**. This framework thus provides a blueprint for lineage-aware ecosystems, balancing computational demands with scientific accountability.

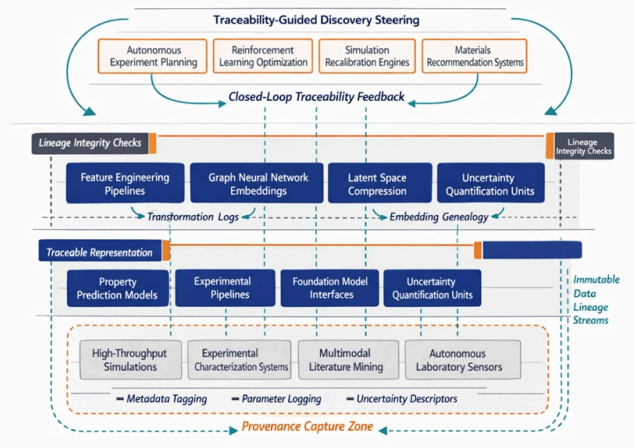

Figure 1. Integrated Traceability Architecture (ITA) for Computational Materials Pipelines.

Schematic representation of the multilayered ITA framework embedding data lineage and scientific traceability across discovery infrastructures. The architecture spans data generation, representation learning, model inference, and discovery steering layers, interconnected through immutable lineage streams and bidirectional feedback loops. Provenance descriptors, transformation logs, and uncertainty annotations propagate across the stack, enabling epistemic audits, risk monitoring, and adaptive recalibration within autonomous materials discovery ecosystems.

Analytical implications

The Integrated Traceability Architecture (ITA) offers a lens through which to interpret the dynamics of computational materials pipelines, revealing implications for workflow optimization and epistemic resilience. By embedding lineage as a structural element, ITA shifts the focus from isolated computations to interconnected systems where traceability informs every stage of materials discovery. This interpretive approach elucidates how data transformations influence model behaviors, particularly in representation learning where atomic embeddings evolve through layered processing [5, 13]. For instance, in graph neural networks applied to crystal structures, traceable lineages allow for

the dissection of feature propagations, highlighting how initial data assumptions cascade into prediction outcomes without necessitating empirical validations [3, 5]. The modulation of closed-loop discovery feedbacks by lineage depth and traceability continuity is conceptually illustrated in **Figure 2**.

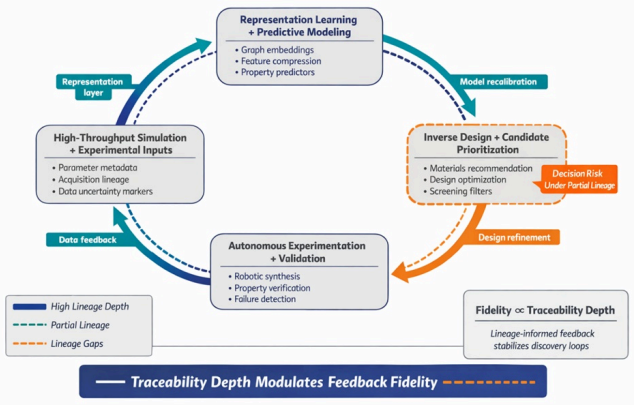

Figure 2. Traceability-Modulated Feedback Dynamics in Closed-Loop Materials Discovery.

Conceptual schematic depicting how lineage depth governs feedback fidelity across simulation, modeling, design, and experimental validation loops. Traceable data flows stabilize model recalibration and candidate prioritization, whereas lineage discontinuities introduce epistemic risk, bias amplification, and discovery inefficiencies within autonomous materials exploration systems.

Workflow dynamics and feedback integration

In high-throughput computational environments, ITA's layered structure implies enhanced feedback mechanisms that adaptively refine pipelines [6, 30]. Data generation layers, when augmented with provenance tags, enable selective routing of inputs to model inference, mitigating the dilution of signal in noisy multimodal datasets [21, 22, 31]. This dynamic can be interpreted through a conceptual flow equilibrium, where the balance between input volume V and traceability depth T modulates output fidelity $F = \frac{V}{1+e-kT}$ Here, k represents a conceptual sensitivity parameter, capturing the sigmoid-like interaction where deeper traceability asymptotically improves fidelity, illustrating trade-offs in resource-intensive simulations [10, 12]. Such implications extend to autonomous discovery systems, where closed-loop iterations benefit from lineage-aware steering, ensuring that experimental feedbacks realign computational paths without epistemic disconnects [14-16].

Epistemic risk management

ITA's emphasis on risk structures provides insights into managing uncertainties in materials AI [4, 20]. By conceptualizing dependencies as annotated networks, the framework interprets how partial traceability amplifies propagation errors in inverse design [18, 19]. For example, in attribute-driven designs for alloys, untraced anharmonic effects could skew property predictions, but ITA implies risk-minimizing logics that prioritize high-lineage paths [10, 11, 26]. This is further formalized in a dependency aggregation model: $D = \prod i = 1n (li) pi$ where l_i denotes lineage completeness for component i, and p_i its positional weight in the pipeline, expressing multiplicative interactions that underscore vulnerabilities in chained inferences [3, 25]. These interpretations promote infrastructures resilient to data scarcities, as seen in small-data regimes where traceable priors enhance generalization [4].

Discovery steering and infrastructure trade-offs

At the discovery level, ITA implies steering logics that leverage traceability for efficient exploration of chemical spaces [9, 24]. In recommender systems or foundation model applications, lineage records facilitate the interpretation of discovery trajectories, revealing biases in phase separation predictions or thermal conductivity estimates [13, 25]. Trade-offs emerge in distributed platforms, where brokering traceability across tenants balances scalability with detail retention [28, 29]. Conceptualizing this as a cost-benefit equilibrium: $B = C - \gamma \cdot (S - T)$ with C as computational cost, S scalability, T traceability, and γ a trade-off coefficient, captures how over-emphasizing speed may erode benefits in long-term discovery [27, 30]. Overall, these analytical implications frame ITA as a tool for interpreting pipeline efficiencies, fostering designs that integrate traceability to elevate materials engineering paradigms [1, 21].

Results and Discussion

The conceptual framing of data lineage and scientific traceability via ITA invites broader reflections on the maturation of computational materials engineering. While the framework addresses core gaps in provenance management, it also surfaces tensions inherent to data-driven ecosystems [1, 7, 8]. One key discussion point revolves around the integration of traceability in evolving AI architectures, such as foundation models that aggregate multimodal knowledge [9, 21]. ITA's layered approach suggests that without embedded lineage, these models risk opaque inferences, yet implementing comprehensive tracking could introduce computational overheads, prompting a reevaluation of efficiency in high-throughput settings [6, 30, 31].

Interoperability and collaborative ecosystems

In collaborative materials research platforms, ITA underscores the value of standardized traceability for knowledge sharing [22, 28, 29]. Event-sourced systems exemplify this, but extending to epistemic annotations could bridge simulation-experiment divides, as in self-driving laboratories where traceable feedbacks enhance collective discovery [8, 15-17]. However, challenges arise in distributed computations, where varying tenant protocols might fragment lineages, implying a need for meta-frameworks that harmonize traceability without stifling innovation [27, 28]. This discussion highlights how ITA could inform policy in materials acceleration platforms, promoting interoperability that accelerates inverse design while safeguarding scientific integrity [14, 18, 19].

Limitations and conceptual boundaries

Conceptually, ITA is bounded by its focus on interpretive systems, eschewing empirical metrics that might quantify traceability impacts [3, 4]. This limits direct applicability to performance benchmarking, yet it strengthens its role in guiding infrastructure designs amid data ambiguities [20, 23]. In uncertainty-laden domains like anharmonic dynamics or corrosion modeling, the framework's risk structures offer interpretive tools, but real-world variabilities—such as experimental noise—may necessitate hybrid extensions [10, 26]. Furthermore, in machine learning for alloys or polymers, representation-inference interactions imply adaptive logics, but without addressing ethical dimensions of data sourcing, traceability alone may not fully mitigate biases [2, 11, 13].

Future directions in materials informatics

Looking ahead, ITA paves the way for traceability-infused advancements in autonomous systems and recommender engines [14, 17, 24]. By interpreting feedback loops through lineage lenses, the framework could inspire designs that optimize closed-loop experimentation, reducing rediscovery in vast material spaces [9, 12, 25]. Discussions in community perspectives emphasize this, advocating for platforms where traceability fosters trust in AI-led discoveries [14, 21]. Ultimately, while ITA remains conceptual, its implications encourage a shift towards lineage-centric paradigms, balancing computational prowess with epistemic accountability in materials engineering [1, 20].

Conclusion

In summary, this manuscript has explored the pivotal role of data lineage and scientific traceability in computational materials pipelines, introducing the Integrated Traceability Architecture (ITA) as a novel conceptual framework. By synthesizing advancements across materials informatics, representation learning, and autonomous systems, we have illuminated how traceability can be woven into pipeline layers to enhance workflow dynamics and mitigate epistemic risks. The analytical implications reveal interpretive insights into feedback integrations, risk managements, and infrastructure trade-offs, formalized through symbolic expressions that capture system interactions.

Discussion points further contextualize ITA within broader ecosystems, addressing interoperability, limitations, and future trajectories in data-driven discovery. This work underscores that traceability is not merely a technical adjunct but a foundational element steering towards transparent, efficient materials engineering. As computational paradigms evolve, adopting lineage-aware architectures like ITA promises to bolster reliability in inverse design, uncertainty handling, and multimodal integrations, ultimately accelerating innovations in fields from energy materials to advanced alloys. By prioritizing systems-level interpretations, this conceptual contribution invites ongoing refinements to foster resilient discovery pipelines.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 15 Apr 2023 Revised: 03 Sep 2023 Accepted: 04 Nov 2023
Published online: 18 March 2024

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.

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

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TYJ. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):204.

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

Lee J, Asahi R. Transfer learning for materials informatics using crystal graph convolutional neural network. Comput Mater Sci. 2021;190:110314.

Luo S, Li T, Wang X, Faizan M, Zhang L. High-throughput computational materials screening and discovery of optoelectronic semiconductors. Wiley Interdiscip Rev Comput Mol Sci. 2021;11(1):e1489.

Soedarmadji E, Stein HS, Suram SK, Guevarra D, Zhou J, Shinde A, et al. Tracking materials science data lineage to manage millions of materials experiments and analyses. npj Comput Mater. 2019;5(1):79.

Bousquet RR, Oses C, Stein HS, Curtarolo S, Gregoire JM. ESAMP: Event-sourced architecture for materials provenance management and application to accelerated materials discovery. Digit Discov. 2023;2(5):1238-52.

Wang X, Gong G, Li N. Artificial-intelligence-led revolution of construction materials: From molecules to Industry 4.0. Matter. 2023;6(6):1832-59.

Xu P, Chen J, Li S, Lu W. Uncertainty and anharmonicity in thermally activated dynamics. Comput Mater Sci. 2021;193:110390.

Roy A, Balasubramanian G. Predictive descriptors in machine learning and data-enabled explorations of high-entropy alloys. Comput Mater Sci. 2021;190:110381.

Yan Y, Zhang L, Li S, Liang H, Qiao Z. Adsorption behavior of metal-organic frameworks: From single simulation, high-throughput computational screening to machine learning. Comput Mater Sci. 2021;190:110383

Hiraide K, Hirayama K, Endo K, Muramatsu M. Application of deep learning to inverse design of phase separation structure in polymer alloy. Comput Mater Sci. 2021;190:110278.

Szymanski NJ, Bartel CJ, Zeng Y, Tu Q, Ceder G. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26.

MacLeod BP, Parlane FGL, Brown AK, Hein JE, Berlinguette CP. What is a minimal working example for a self-driving

laboratory? Matter. 2022;5(4):1141-56.

Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LPE, Hein JE, et al. ChemOS 2.0: An orchestration architecture for chemical self-driving laboratories. Matter. 2024;7(5):1883-97.

Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milikisiyants D, et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nat Commun. 2023;624(1):86-91.

Tagade PM, Adiga SP, Pandian S, Kolake SM, Song S, Joshi S. Attribute driven inverse materials design using deep learning Bayesian framework. npj Comput Mater. 2019;5(1):103.

Cui Z, Gao T, Talamadupula K, Ji Q. Knowledge-augmented deep learning and its applications: A survey. IEEE Trans Neural Netw Learn Syst. 2023;36(2):2133-53.

Mussmann S, Liang P. On the relationship between data efficiency and error for uncertainty sampling. In International Conference on Machine Learning. PMLR; 2018. p. 3674-82.

Chen C, Zuo Y, Ye W, Li J, Deng Z, Ong SP. AI applications through the whole life cycle of material discovery. Matter. 2020;3(4):991-1014.

Court CJ, Cole JM. Propnet: A knowledge graph for materials science. Matter. 2020;2(3):551-61.

Oliveira L, Zaera-Polo J, Calatayud M, Kisielowski C, Specht P, Algarra AG, et al. Accelerating the adoption of research data management strategies. Matter. 2022;6(3):696-709.

Muroga A. Recommender system for discovery of inorganic compounds. npj Comput Mater. 2022;8(1):249.

Zhang S, He J, Zhao W, Liu P. Predicting lattice thermal conductivity via machine learning: A perspective. npj Comput Mater. 2023;9(1):19.

Zhou X, Li L, Zhang Z, Li W, Chen Y, Zhao M, et al. Nobility vs. mobility: Insights into molten salt corrosion mechanisms of high-entropy alloys via high-throughput experiments and machine learning. Matter. 2024;7(5):1898-918.

Statt A, Kovalenko A, Tron A, Curtarolo S, Amsler M, Goedecker S, et al. Physical computing for materials acceleration platforms. Matter. 2023;6(3):710-33.

Ghiringhelli LM, Levchenko SV, Heinen J, Impagnatiello A, Nardi F, Vitale V, et al. Brokering between tenants for an international materials acceleration platform. Matter. 2023;6(10):3345-67.

Aykol M, Montoya JH, Hummelshøj J. The materials research platform: Defining the requirements from user stories. Matter. 2019;1(6):1433-8.

Frey CE, Muñoz JA, Amsler M. Mkite: A distributed computing platform for high-throughput materials simulations. Comput Mater Sci. 2023;230:112480.

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