ORIGINAL RESEARCH **Open access**

Failure Visibility and Epistemic Accountability in Self-Driving Materials Engineering

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

Abstract

Self-driving laboratories have emerged as a cornerstone of computational and data-driven materials engineering, fusing automated high-throughput experimentation with machine-learning-driven decision engines to compress discovery timelines from years to weeks. This paradigm shift reconfigures the materials pipeline into a closed-loop system in which data generation, model inference, and experimental steering operate with minimal human intervention. Yet the very autonomy that accelerates discovery simultaneously obscures the epistemic foundations of the knowledge it produces. Failures—whether arising from underrepresented chemical spaces, model extrapolation beyond training distributions, or unacknowledged aleatoric–epistemic uncertainty boundaries—often remain latent until downstream validation, eroding trust in autonomous outputs. Current uncertainty quantification and explainability techniques, while technically sophisticated, are typically deployed in isolation and rarely propagate failure signals across the full discovery stack. We articulate a conceptual architecture, the Epistemic Visibility and Accountability Framework (EVAF), that treats failure not as an anomaly to be minimized but as a structured signal to be surfaced and attributed at every layer of the self-driving pipeline. By integrating multi-scale representation tracking, inference-trace logging, and risk-propagation mapping, EVAF establishes a computational substrate for epistemic accountability: the systematic assignment of responsibility for knowledge claims to specific data, model, or orchestration components. The framework reframes self-driving systems from opaque optimizers into transparent epistemic engines, enabling materials engineers to maintain intellectual oversight without sacrificing autonomy. Its implications extend to infrastructure design, regulatory readiness for autonomous discovery platforms, and the long-term reliability of data-intensive materials science.

Keywords Data-driven discovery, Self-driving laboratories, Uncertainty propagation, Computational materials engineering, Epistemic accountability, Failure visibility

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

Introduction

The rise of autonomous discovery pipelines

The integration of robotics, machine learning, and high-throughput characterization has produced a new class of research infrastructure: self-driving laboratories (SDLs) that autonomously propose, execute, and interpret experiments in materials chemistry and solid-state physics [1, 2]. These systems operate at the intersection of closed-loop optimization and generative modeling, continuously refining surrogate representations of vast design spaces [3, 4]. In contrast to traditional Edisonian or even high-throughput screening approaches, SDLs embed inference directly into the experimental loop, allowing the discovery trajectory itself to become an optimizable object [5].

This computational re-engineering of the materials pipeline has delivered measurable acceleration in domains ranging from thin-film photovoltaics to solid-state electrolytes [2, 4]. Yet the literature reveals a persistent asymmetry: while hardware and algorithmic throughput have scaled dramatically, the epistemic scaffolding required to interpret and trust the resulting knowledge claims has not kept pace.

Epistemic friction in data-driven workflows

Every autonomous decision in an SDL rests on a chain of representations—composition spaces, property manifolds, uncertainty surfaces—and a sequence of inferences that link these representations to experimental actions. When these chains contain undetected weaknesses (sparse training regions, unmodeled physics, or compounding numerical approximations), the system can generate confident but erroneous trajectories. The resulting failures are frequently invisible to both the orchestration layer and the human supervisor until significant resources have been expended [6, 7].

Uncertainty quantification (UQ) methods have become standard in materials machine learning, distinguishing aleatoric noise from epistemic ignorance [8, 9]. However, most implementations remain localized: a Bayesian neural network may flag high epistemic uncertainty for a single prediction, yet that signal rarely propagates backward to question the validity of the upstream dataset curation or forward to adjust the experimental budget allocation [10, 11]. The consequence is a discovery process that is statistically aware but epistemically opaque.

Failure visibility as an infrastructure problem

Visibility of failure is not merely a diagnostic convenience; it is a structural requirement for any system that claims to generate reliable scientific knowledge autonomously. In conventional materials research, failure is rendered visible through peer scrutiny, replication attempts, and the slow accumulation of contradictory evidence. In SDLs, these social and temporal buffers are deliberately removed. The system must therefore internalize mechanisms that make its own epistemic limits computationally legible in real time.

Current orchestration platforms such as ChemOS and A-Lab provide elegant abstractions for experiment scheduling and data management [2, 12], yet their logging architectures are optimized for throughput rather than epistemic traceability. When a campaign terminates with unexpected phase purity or degraded performance, tracing the causal chain across weeks of autonomous decisions remains a manual, error-prone forensic exercise.

The accountability gap

Epistemic accountability—the capacity to assign responsibility for a knowledge claim to the specific computational and experimental choices that produced it—has received even less systematic attention. Accountability here is not a legal or ethical abstraction but a practical requirement for iterative improvement of the discovery engine itself. Without it, SDLs risk becoming black-box generators of irreproducible knowledge, undermining the very reproducibility crisis that data-driven methods were intended to solve [13, 14].

The field therefore faces a dual imperative: to preserve the acceleration promised by autonomous systems while embedding computational structures that render their epistemic weaknesses visible and attributable.

Positioning the epistemic visibility and accountability framework

The Epistemic Visibility and Accountability Framework (EVAF) addresses this imperative by redesigning the self-driving pipeline as a layered epistemic substrate. Rather than treating uncertainty as a post-hoc diagnostic, EVAF embeds failure-visibility primitives at the level of data representation, model inference, orchestration logic, and human–system interface. The framework introduces computational mechanisms for propagating epistemic risk across these layers, enabling the system to surface, localize, and attribute failures as intrinsic features of the discovery process.

The following sections synthesize the theoretical foundations that motivate EVAF and then articulate its conceptual architecture in detail.

Theoretical Background & Literature Synthesis

Self-driving laboratory architectures and their epistemic assumptions

Contemporary SDLs are built on three interdependent layers: (i) robotic execution and characterization, (ii) machine-learning surrogates for property prediction and experiment selection, and (iii) orchestration software that closes the loop [1, 5, 15]. Early implementations relied on predefined design spaces and relatively simple Bayesian optimization routines [2]. More recent systems incorporate generative models and large-scale pre-trained representations, allowing exploration of chemically open-ended spaces [4, 16].

Across these architectures, a common epistemic assumption persists: that the surrogate model's confidence estimates, when properly calibrated, provide a sufficient proxy for the reliability of downstream decisions. This assumption holds reasonably well within densely sampled regions of design space but breaks down at the boundaries where truly novel materials are most likely to be discovered [17, 18]. Literature on autonomous experimentation increasingly acknowledges this boundary problem [3, 19], yet solutions remain fragmented—typically adding more data or more sophisticated UQ without reconsidering the information architecture itself.

Representation–inference interactions in materials machine learning

The predictive power of materials ML derives from the interplay between chosen representations (composition-based, graph-based, or physics-informed) and the inference procedures applied to them [13, 20]. Recent scaling laws in materials foundation models demonstrate that larger, more diverse training corpora improve in-distribution performance dramatically [16]. However, the same studies reveal pronounced degradation in out-of-distribution regimes, precisely the regimes SDLs are designed to explore.

This creates a structural tension: the representations that enable rapid discovery are also the representations most prone to silent failure when the system ventures beyond its epistemic comfort zone. Uncertainty quantification methods attempt to mitigate this by estimating predictive variance or using ensemble disagreement [6, 7, 11]. Yet these techniques rarely interrogate the deeper question of whether the representation itself remains valid for the queried chemistry. Epistemic visibility, in this context, requires mechanisms that can flag not only high uncertainty but also representation drift or inference–representation mismatch.

Epistemic risk structures across the discovery stack

Epistemic risk—the possibility that a knowledge claim is systematically wrong due to incomplete or biased information—manifests differently at each stage of an SDL campaign [8, 9]. At the data-acquisition layer, risk arises from undersampling of the chemical space or from sensor calibration drift. At the surrogate-model layer, risk accumulates through compounding approximation errors and unmodeled physical mechanisms. At the orchestration layer, risk is introduced by the experiment-selection policy itself, which may systematically avoid regions where the model is most uncertain (a phenomenon sometimes termed "uncertainty aversion" in autonomous systems).

Literature on multi-fidelity and active learning has developed sophisticated methods for balancing exploration and exploitation [11], yet these methods optimize for information gain rather than epistemic legibility. The result is an optimization landscape in which the system can achieve high campaign efficiency while simultaneously accumulating undetected epistemic debt. EVAF's contribution lies in treating this debt as a first-class observable that must be tracked and reported alongside property predictions.

Explainability, interpretability, and the limits of post-hoc analysis

Considerable effort has been devoted to making materials ML models more interpretable through feature attribution, counterfactual generation, and mechanistic probes [10, 14]. These approaches are valuable for human understanding but operate primarily as retrospective tools. In an SDL operating at 10–100 experiments per day, post-hoc interpretability cannot keep pace with the rate of decision-making.

Moreover, most explainability techniques focus on individual predictions rather than on the cumulative epistemic state of the entire campaign. A model may correctly explain why it predicted a particular band gap for a given composition, yet fail to communicate that the entire

local region of composition space was extrapolated from a chemically dissimilar training set. Epistemic accountability therefore demands a shift from per-prediction explanations to system-level epistemic accounting.

Infrastructure trade-offs and the reproducibility imperative

The push toward autonomous discovery has coincided with renewed emphasis on reproducibility and data stewardship in materials science [13, 18]. FAIR data principles and open-source orchestration platforms have improved metadata capture [12], yet metadata alone does not guarantee epistemic traceability. When an SDL campaign fails to reproduce a literature result, the question is rarely "which experiment failed?" but rather "at which point in the multi-week autonomous trajectory did the system's representation of the problem diverge from physical reality?"

Current infrastructure largely lacks the primitives required to answer such questions efficiently. This gap is not a failure of individual research groups but a consequence of the rapid evolution of SDL hardware and software outpacing the development of corresponding epistemic infrastructure.

Synthesis: The need for a new computational substrate

Collectively, the literature reveals a mature set of components—autonomous orchestration, calibrated uncertainty quantification, scalable representations, and emerging explainability methods—yet these components remain loosely coupled. What is missing is an integrative architecture that treats failure visibility and epistemic accountability as first-order design constraints rather than secondary diagnostics. The Epistemic Visibility and Accountability Framework (EVAF) supplies this missing substrate by re-engineering the information flow within self-driving systems to make epistemic risk both computationally observable and operationally actionable. Layer-specific epistemic risks and their associated visibility mechanisms are systematized in **Table 1**.

Table 1. Epistemic Failure Modes and Visibility Primitives across the EVAF Discovery Stack

Discovery Layer	Primary Epistemic Risk	Failure Visibility Primitive	Propag Pathv
Experimental Data Substrate	Measurement noise, calibration drift, sparse sampling	Anomaly detection nodes, sampling density maps	Upwa represe fidelity
Representation Fidelity	Manifold distortion, descriptor incompleteness, topology gaps	Fidelity tracking engine, embedding divergence metrics	Upwa inferer downw dat acquis
Inference & Surrogate Modeling	Extrapolation error, model bias, uncertainty miscalibration	Ensemble disagreement monitors, variance fields	Bidirec wit orchest laye
Orchestration Governance	Acquisition bias, uncertainty aversion, budget misallocation	Decision attribution registry, steering logic audits	Downw experi execu
Human–System Interface	Interpretability gaps, oversight latency	Epistemic dashboards, risk heatmaps	Feedba orchest polic

Table 1. Layer-resolved epistemic failure modes and visibility primitives within the Epistemic Visibility and Accountability Framework (EVAF). The table maps discovery-stack risks to computational monitoring mechanisms, propagation pathways, and accountability outputs, illustrating how failure signals are rendered observable and attributable across autonomous materials engineering infrastructures.

Proposed framework

The Epistemic Visibility and Accountability Framework (EVAF)

EVAF reconceptualizes the self-driving materials engineering pipeline as a stack of interconnected epistemic layers, each equipped with native visibility primitives and accountability channels. The framework does not replace existing SDL components; it augments them with a parallel

epistemic computation graph that runs alongside the primary discovery loop.

At its core, EVAF maintains a dynamic epistemic state vector that evolves with every experimental iteration. This vector aggregates signals from four primary channels: representation fidelity, inference confidence, risk propagation, and decision attribution. Rather than producing a single scalar uncertainty score, the framework generates a structured failure signature—a sparse, high-dimensional encoding that localizes epistemic weaknesses to specific pipeline stages and data subsets.

The architecture operates through continuous, lightweight monitoring of representation–inference interactions. For example, when a new composition is proposed, the system evaluates not only the predicted property distribution but also the degree to which the active representation diverges from the training manifold in a chemically meaningful embedding space. Significant divergence triggers an epistemic flag that propagates both upward to the orchestration layer (potentially triggering a change in acquisition strategy) and downward to the data layer (flagging the need for targeted supplemental measurements).

Crucially, every decision made by the system—whether experiment selection, model retraining, or campaign termination—is logged with an explicit attribution trace linking the decision to the epistemic state at the moment it was taken. This trace forms the basis for post-campaign epistemic audits and, more importantly, enables the system to learn from its own epistemic failures across campaigns. Over time, the orchestration layer can develop meta-policies that preferentially allocate experimental budget to regions where previous campaigns accumulated high epistemic debt.

The framework's power lies in its systems-level integration. By making failure visible at the infrastructure level, EVAF transforms the self-driving laboratory from a statistical optimizer into a self-reflective epistemic instrument. Materials engineers retain the ability to intervene at any point with full contextual awareness of why the system is uncertain or has failed, without needing to reconstruct thousands of intermediate states manually. The layered architecture of the Epistemic Visibility and Accountability Framework and its cross-stack failure propagation mechanisms are illustrated in **Figure 1**.

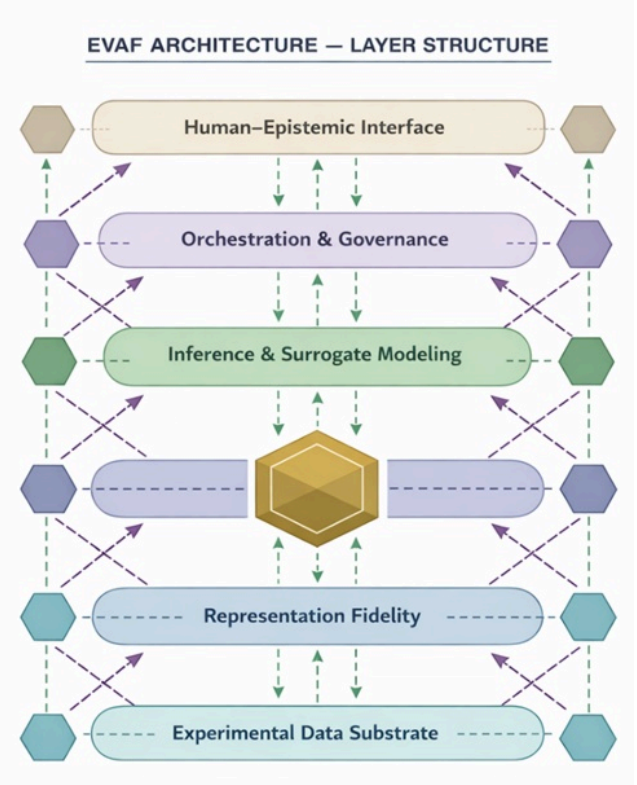

Figure 1. Positioning the Epistemic Visibility and Accountability Framework

Epistemic Visibility and Accountability Framework (EVAF) architecture for self-driving materials engineering. The framework augments autonomous discovery pipelines with a parallel epistemic computation stack spanning data acquisition, representation learning, surrogate inference, orchestration governance, and human oversight. Embedded visibility primitives track representation fidelity, inference confidence, and experimental anomalies, while a central epistemic state vector aggregates structured failure signatures. Bidirectional risk-propagation pathways transmit epistemic debt across layers, and attribution conduits establish decision provenance. Together, these mechanisms render autonomous discovery epistemically transparent, enabling real-time failure localization and post-campaign accountability audits.

Analytic implications

The Epistemic Visibility and Accountability Framework (EVAF) reconfigures the analytic landscape of self-driving materials engineering by embedding failure visibility and epistemic accountability as core computational primitives rather than ancillary diagnostics. This integration generates layered insights into representation–inference interactions,

risk propagation dynamics, discovery steering logics, and infrastructure trade-offs, each of which reshapes how knowledge claims are generated, validated, and sustained within autonomous pipelines [6-8].

Representation fidelity as a computable observable

At the foundation of EVAF lies the treatment of representation fidelity as a continuously monitored state variable. In conventional self-driving laboratories, material representations—whether composition vectors, graph embeddings, or physics-informed descriptors—are selected during initialization and assumed to remain valid across the explored chemical space [13, 20]. EVAF augments these representations with a parallel fidelity tracker that computes local manifold consistency metrics at every inference step. When a proposed composition falls into a region where the active embedding exhibits topological distortion relative to the accumulated experimental manifold, the framework registers a representation-fidelity deficit. This deficit is not a scalar flag but a structured vector that encodes the specific dimensions of divergence (for example, under-represented coordination environments or unmodeled electronic correlations).

The analytic implication is profound: inference is no longer performed on a static representational substrate but on a dynamically validated one. High-fidelity regions permit aggressive exploitation, while low-fidelity zones trigger conservative exploration policies that prioritize data acquisition over property optimization. Over the course of a campaign, this mechanism prevents the accumulation of epistemic debt in peripheral regions of design space, a common pathology in purely uncertainty-aware optimizers [9, 11]. The result is a discovery process whose representational health is as legible as its predicted property distributions, enabling engineers to interpret campaign trajectories not merely as sequences of experiments but as evolving maps of epistemic coverage.

Inference-trace logging and granular failure attribution

EVAF maintains an immutable inference-trace ledger that records, for each decision epoch, the exact chain of data subsets, model parameters, and orchestration rules that informed the action. When a downstream validation reveals a discrepancy—such as a synthesized material deviating from predicted stability—the ledger supports automated back-propagation to isolate the originating epistemic weakness. Attribution might localize to a particular training cluster that lacked sufficient diversity, to a surrogate update that amplified extrapolation error, or to an acquisition function that systematically discounted high-uncertainty candidates [10, 14].

This traceability transforms failure analysis from retrospective forensics into prospective epistemic accounting. Patterns emerge across campaigns: certain classes of materials consistently induce representation drift, or specific orchestration heuristics repeatedly generate overconfident trajectories. The framework can therefore surface meta-insights—such as the identification of “epistemic attractors” where the system converges on misleading local optima—without requiring human reconstruction of thousands of intermediate states. The analytic consequence is a shift in the unit of analysis from individual experiments to the epistemic genealogy of entire discovery campaigns, revealing how seemingly minor representational choices compound into systemic knowledge distortions [5, 12].

Risk propagation mapping and systemic epistemic dynamics

Epistemic risk in self-driving pipelines is not confined to single predictions; it propagates and amplifies across layers. EVAF formalizes this propagation through a directed risk graph that connects data-layer uncertainties, model-layer approximations, orchestration-layer policies, and execution-layer measurement noise—an outlook consistent with multiscale uncertainty propagation perspectives in ICME and computational mechanics communities [21, 22]. Each edge in the graph carries a calibrated propagation factor derived from historical campaign traces, while node-level attribution can be decomposed into feature-importance and uncertainty signatures to localize risk sources within materials models [23]. When a new epistemic flag is raised—say, from representation drift—the framework updates the global risk state and redistributes experimental budget to mitigate downstream amplification [7, 8].

At the systems level, this creates a form of computational self-awareness. The orchestration layer can detect when the aggregate epistemic risk exceeds a campaign-specific threshold and initiate corrective actions, such as inserting calibration experiments or invoking a higher-fidelity surrogate, consistent with multi-fidelity and benchmark-driven uncertainty management practices [11, 21]. Over

multiple campaigns, the risk graph itself evolves, learning which propagation pathways are most prone to instability in particular material families, including settings where data scarcity is the dominant failure driver [24]. The analytic insight here is that discovery efficiency is not solely a function of optimization speed but of epistemic homeostasis—the capacity of the pipeline to maintain bounded risk while exploring—especially in chemically diverse families such as high-entropy materials where uncertainty gradients can be steep [25]. EVAF makes this homeostasis observable and tunable, elevating it from an implicit property of well-designed systems to an explicit design objective.

Reconfiguration of discovery steering logics

Traditional steering logics in self-driving laboratories balance exploration and exploitation through information-theoretic or acquisition-function criteria [2, 11]. EVAF augments these logics with an epistemic-utility term that rewards actions expected to reduce global epistemic debt, aligning steering with uncertainty-aware design strategies used in practical Bayesian optimization workflows for functional materials [26]. The resulting steering policy favors campaigns that systematically close representational gaps even when immediate property gains appear modest. This does not slow discovery; rather, it channels exploration toward regions where subsequent exploitation will be more reliable, including regimes where small-data constraints demand conservative, evidence-building trajectories [24].

The analytic implication is a qualitative change in campaign character. Instead of producing isolated high-performing candidates embedded in sparsely validated neighborhoods, EVAF-guided pipelines generate clusters of materials whose surrounding chemical space is epistemically well-mapped. This clustering effect supports downstream scale-up and transfer learning, as the knowledge claims carry explicit provenance and confidence bounds, and can be anchored to community reference infrastructures that have historically accelerated reproducible materials innovation [27]. Steering logics thus become instruments of epistemic curation, shaping not only what is discovered but the epistemic robustness with which it is discovered.

Infrastructure trade-offs and human–system symbiosis

Implementing EVAF imposes measurable computational overhead—maintenance of the fidelity tracker, trace ledger, and risk graph—but this overhead is offset by substantial reductions in wasted experimental cycles and improved campaign reproducibility, particularly when uncertainty must be propagated across coupled models and scales [21, 22]. The framework's visibility primitives also reshape the human–system interface. Rather than receiving only final candidate lists, engineers interact with a live epistemic dashboard that displays current risk hotspots, attribution summaries (e.g., feature-importance-linked uncertainty drivers), and projected campaign health trajectories [23]. Interventions become targeted and informed, preserving autonomy while restoring intellectual oversight [3, 15].

At the infrastructure level, EVAF encourages the design of modular epistemic services that can be plugged into existing orchestration platforms. The trade-off is between added complexity and enhanced trustworthiness, a balance that favors adoption in high-stakes domains such as energy materials or biomedical devices where epistemic failures carry material consequences. This modularity is increasingly important as pipelines incorporate modern interatomic potential stacks—where equivariant architectures and training stability can materially shape uncertainty behavior [28]—and as "test-arena" style datasets for potentials motivate more explicit governance over generalization claims [29]. The broader implication is that self-driving systems evolve from throughput optimizers into epistemic instruments whose outputs are accompanied by verifiable accountability structures.

Results and Discussion

The Epistemic Visibility and Accountability Framework addresses a structural lacuna in contemporary self-driving materials engineering: the absence of computational mechanisms that render the epistemic foundations of autonomous discovery legible and attributable at infrastructure scale. By synthesizing insights from uncertainty quantification [6-8, 21, 22], explainable machine learning [10, 14, 23], and autonomous orchestration [1, 2, 5, 12], EVAF offers a systems-level response to the growing recognition that acceleration without accountability risks eroding the scientific character of data-driven discovery.

Implementation pathways are deliberately modular. The framework's primitives—representation trackers, inference ledgers, and risk graphs—can be realized as lightweight middleware layers atop existing platforms such as ChemOS or A-Lab extensions [12]. Early integration could

target high-throughput thin-film or polymer workflows [2, 4], where experimental throughput is already high but epistemic traceability remains manual. Standardization of epistemic metadata formats would further accelerate community adoption, aligning with ongoing efforts toward FAIR data principles in materials science [13, 18] and with the field's broader movement toward shared, reference-grade innovation infrastructures [27].

For human oversight, EVAF restores a form of distributed cognition in which engineers and autonomous agents share a common epistemic language. The dashboard interface transforms supervision from exception handling to strategic epistemic stewardship, allowing domain experts to redirect campaigns toward underrepresented subspaces or to certify high-confidence regions for downstream applications. This symbiosis mitigates the risk of over-delegation while preserving the speed advantages of autonomy, including in small-data regimes where human priors and targeted evidence accumulation remain decisive [24].

Broader implications extend to regulatory and community dimensions. As self-driving laboratories move toward industrial deployment, epistemic accountability becomes a prerequisite for regulatory acceptance, particularly in sectors with safety or reproducibility requirements. The framework provides a computational substrate for audit-ready discovery, where every knowledge claim is traceable to its originating data, model, and decision context, including uncertainty propagation across scales typical of ICME-aligned workflows [22]. Within the research community, EVAF encourages a shift in publication norms: alongside predicted materials, authors would report epistemic coverage metrics and failure signatures, enriching the collective knowledge base and reducing irreproducible claims [13, 14]. Practical exemplars—such as uncertainty-aware Bayesian optimization for targeted cathode design [26] and governance-aware exploration in high-entropy energy materials spaces [25]—illustrate how such reporting could become actionable rather than performative.

Challenges remain. The computational cost of continuous trace logging and risk mapping must be managed through efficient data structures and selective activation. Calibration of fidelity and propagation metrics will require community benchmarks—especially for modern potential-model stacks where architecture choices (e.g., equivariance) alter generalization and error structure [28]—and the framework's effectiveness in chemically open-ended spaces—where representation drift is most severe—will demand iterative refinement. Nevertheless, these challenges are infrastructural rather than conceptual; they are addressable through the same iterative, data-driven ethos that SDLs themselves embody.

In the longer term, EVAF points toward a new generation of self-driving systems that are not merely autonomous but epistemically reflective. Such systems will learn not only which materials to synthesize but how to maintain the integrity of the knowledge they produce. This reflective capacity represents a qualitative advance in computational materials engineering, aligning autonomous discovery more closely with the self-correcting nature of scientific inquiry.

Conclusion

The Epistemic Visibility and Accountability Framework (EVAF) establishes failure visibility and epistemic accountability as foundational design principles for self-driving materials engineering. By embedding structured mechanisms for representation tracking, inference tracing, risk propagation, and decision attribution, EVAF transforms autonomous pipelines from opaque optimizers into transparent epistemic engines. The resulting systems generate knowledge claims whose provenance and limitations are computationally legible, enabling sustained intellectual oversight without compromising discovery speed.

This conceptual architecture addresses the central tension of the data-driven paradigm: the compression of discovery timelines must be accompanied by commensurate advances in epistemic infrastructure. EVAF supplies that infrastructure, integrating seamlessly with existing SDL components while introducing new primitives that make epistemic risk an observable and manageable feature of the discovery process. Its implications span workflow dynamics, human–system collaboration, infrastructure design, and the long-term reliability of autonomous materials science.

As self-driving laboratories become central to the materials innovation ecosystem, frameworks that prioritize epistemic accountability will determine whether acceleration translates into trustworthy knowledge or merely accelerated uncertainty. EVAF offers a computational path toward the former, positioning self-driving systems as genuine partners

in the scientific enterprise—capable of rapid exploration while remaining accountable for the knowledge they produce. The framework thus contributes a systems-level foundation for the next era of computational and data-driven materials engineering, one in which autonomy and epistemic integrity are not competing objectives but mutually reinforcing dimensions of discovery infrastructure.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Sep 2024 Revised: 07 Nov 2024 Accepted: 29 Nov 2024
Published online: 18 March 2025

References

Tom G, Schmid SP, Baird SG, Cao Y, Darvish K, Hao H, et al. Self-Driving laboratories for chemistry and materials science. Chem Rev. 2024;124(16):9633-732. https://doi.org/10.1021/acs.chemrev.4c00055.

MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, Day S, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Advs. 2020;6(20):eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.

Hung L, Yager JA, Monteverde D, Baiocchi D, Kwon H, Sun S, et al. Autonomous laboratories for accelerated materials discovery: a community survey and practical insights. Digit Discov. 2024;3(7):1273-9. https://doi.org/10.1039/D4DD00059E.

Fei Y, Gallant M, Persson K, Szymanski NJ, Rendy B, He T, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. Nature. 2023;624(7990):86-91. https://doi.org/10.1038/s41586-023-06734-w.

Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, et al. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26. https://doi.org/10.1016/j.matt.2021.06.036.

Tavazza F, DeCost BL, Choudhary K. Uncertainty Prediction for Machine Learning Models of Material Properties. ACS Omega. 2021;6(48):32431-40. https://doi.org/10.1021/acsomega.1c03752.

Psaros AF, Meng X, Zou Z, Guo L, Karniadakis GE. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. J Comput Phys. 2023;477:111902. https://doi.org/10.1016/j.jcp.2022.111902.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn. 2021;110(3):457-506. https://doi.org/10.1007/s10994-021-05946-3.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion. 2021;76:243-97. https://doi.org/10.1016/j.inffus.2021.05.008.

Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and

chemistry. Acc Mater Res. 2022;3(6):597-607. https://doi.org/10.1021/accountsmr.1c00244.

Tran A, Tranchida J, Wildey T, Thompson AP. Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys. J Chem Phys. 2020;153(7):074705. https://doi.org/10.1063/5.0015672.

Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LPE, Hein JE, et al. ChemOS: An orchestration software to democratize autonomous discovery. PLoS ONE. 2020;15(4):e0229862.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. https://doi.org/10.1038/s41557-021-00716-z.

Canty RB, Bennett JA, Brown KA, Buonassisi T, Kalinin SV, Kitchin JR, et al. Science acceleration and accessibility with self-driving labs. Nat Commun. 2025;16(1):3856. https://doi.org/10.1038/s41467-025-59231-1.

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

Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. Nanoscale Adv. 2020;2(8):3115-30. https://doi.org/10.1039/D0NA00388C.

Wang Z, Sun Z, Yin H, Liu X, Wang J, Zhao H, et al. Data-Driven Materials Innovation and Applications. Adv Mater. 2022;34(16):2104113. https://doi.org/10.1002/adma.202104113.

Tobias AV, Wahab A. Autonomous 'self-driving' laboratories: a review of technology and policy implications. R Soc Open Sci. 2025;12(7):250646. https://doi.org/10.1098/rsos.250646.

Tao Q, Xu P, Li M, Lu W. Machine learning for perovskite materials design and discovery. npj Comput Mater. 2021;7(1):23.

Shields MDUncertainty quantification in computational solid and structural materials modeling. USACM Report. 2023.

Chen W. Multiscale and multidimensional uncertainty quantification in integrated computational materials engineering. Acta Mater (series extensions 2018–2022).

Liu Z, Singh A, Li Y. Feature importance and uncertainty quantification of machine learning model in materials science. ASME Int Mech Eng Congr Expo. 2023.

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):42. https://doi.org/10.1038/s41524-023-01000-z.

Mashhadimoslem H, Karimi P, Elkamel A, Yu A. Toward high entropy material discovery for energy applications using computational and machine learning methods. npj Comput Mater. 2025;12:50.

Park S, Shim Y, Hur J, Ji S, Jeon D, Yuk JM, et al. Element mapping-based Bayesian optimization framework enabling direct materials design:A case study on NASICON-type cathode materials. npj Comput Mater. 2025.

Wang Y, Liu Y, Song S, Yang Z, Qi X, Wang K, et al. Accelerating the discovery of insensitive high-energy-density materials by a materials genome approach. Nat Commun. 2018;9(1):2444.

Yang Z, Wang X, Li Y, Lv Q, Chen CY, Shen L. Efficient equivariant model for machine learning interatomic potentials. npj Comput Mater. 2025;11:49.

Cărare V, Thiemann FL, Morrow JD, Wales DJ, Pyzer-Knapp EO, Dicks L. Global properties of the energy landscape: A testing and training arena for machine learned potentials. npj Comput Mater. 2025.