

ORIGINAL RESEARCH

Open access

Throughput without Accountability: Responsibility Dilution in Autonomous Materials Innovation

Maria Gonzalez^{1*}, Javier Ruiz¹, Lucia Torres²

Abstract

The integration of machine learning, robotics, and high-performance computing has transformed computational and data-driven materials engineering, shifting discovery from sequential human-led campaigns to autonomous, closed-loop pipelines capable of evaluating thousands of candidates per cycle. This paradigm delivers unprecedented throughput, yet it simultaneously disperses decision authority across data pipelines, inference engines, and robotic agents, creating a structural dilution of responsibility that existing frameworks have not systematically addressed. Current literature excels at accelerating prediction, synthesis, and characterization but treats governance as an external overlay rather than an intrinsic computational dynamic. The result is a growing epistemic risk: high-velocity discovery without traceable stewardship. This conceptual manuscript reframes throughput as a governance problem. We synthesize the data-driven materials ecosystem and autonomous laboratory architectures to expose how responsibility fragments across layered pipelines. To resolve this, we introduce the Dilution Cascade Framework, an original systems model that maps accountability propagation through data–model–discovery layers, formalizes feedback steering logics, and identifies computational interventions to restore traceability without sacrificing velocity. The framework offers infrastructure-level insights for embedding governance in next-generation autonomous platforms. Its implications extend to the design of materials innovation ecosystems that remain both high-throughput and epistemically accountable, ensuring that accelerated discovery serves as a foundation for responsible scientific infrastructure rather than a vector for diffused agency.

Keywords Self-driving laboratories, Autonomous materials discovery, Epistemic accountability, Responsibility dilution, Data-driven governance, Computational pipelines

*Correspondence:

Maria Gonzalez
maria.gonzalez@gmail.com

¹ Department of Materials Informatics, Faculty of Engineering, University of Granada, Granada, Spain

² Department of Computational Materials Simulation, Faculty of Engineering, University of Seville, Seville, Spain

Introduction

The acceleration imperative in computational materials engineering

Materials discovery has long been constrained by the scale of experimental validation relative to the vastness of chemical and structural space. Over the past decade, computational and data-driven approaches have dismantled this bottleneck. Machine learning models now predict properties across the periodic table with increasing fidelity [1–3], while graph neural networks and large-scale generative models map structure–property relationships at unprecedented resolution [3, 4].

These advances have culminated in autonomous laboratories—integrated platforms that couple robotic synthesis, in-situ characterization, and closed-loop decision engines [5–7]. Systems such as the A-Lab [7] and related self-driving

architectures demonstrate the capacity to propose, execute, and iterate synthesis routes with minimal human intervention, achieving discovery rates orders of magnitude higher than traditional workflows. Throughput, once measured in compounds per year, is now discussed in terms of candidates per day or even per hour.

This acceleration is not merely quantitative. It represents a qualitative reconfiguration of the discovery pipeline: from human-centric hypothesis testing to distributed, algorithmically steered exploration. The field has moved from informatics as a supportive tool [1, 8] to informatics as the operational core of innovation [9].

Throughput as infrastructure, not just efficiency

The literature frames this shift predominantly through the lens of workflow optimization. Reviews emphasize scaling laws in deep learning for materials [6], the fusion of simulation and experiment in closed loops [5], and the standardization of data infrastructures [10, 11]. Benchmarks focus on prediction accuracy, experimental success rates, and cycle times [12].

Yet throughput is more than a performance metric. In autonomous systems, it becomes an infrastructural property that reshapes the epistemic and organizational architecture of discovery. When robotic agents select precursors, models interpret spectra, and active learning loops decide the next experiment, the locus of scientific agency fragments. Each component contributes to the outcome, but no single node retains comprehensive oversight.

This fragmentation is not a bug in the system; it is an emergent feature of high-throughput design. As pipelines grow more parallel and layered, the cognitive and ethical load of any individual decision becomes diluted across the ensemble [13]. The field has optimized for velocity while leaving the governance substrate underdeveloped.

The governance vacuum in autonomous discovery

Responsibility in materials innovation has traditionally rested with the principal investigator or research team, who design experiments, interpret results, and bear accountability for claims. In autonomous pipelines, this model erodes. Data curation may occur upstream in curated repositories [10], model training on federated datasets [14], inference in cloud-based agents, and execution in distributed robotic facilities [15, 16].

When a novel material emerges, tracing causal responsibility becomes computationally and organizationally non-trivial. Was the breakthrough attributable to the generative model, the active learning policy, the robotic execution fidelity, or the initial dataset curation? Current literature acknowledges ethical dimensions in passing [13] but does not provide computational mechanisms to embed accountability as a native system property.

This creates a structural mismatch: the infrastructure supports exponential throughput, yet lacks corresponding mechanisms for epistemic stewardship. The result is responsibility dilution—a progressive attenuation of traceable agency as discovery velocity increases.

Positioning the contribution

This manuscript addresses the gap by reframing throughput as a governance challenge rather than a workflow metric. We synthesize the foundational literature on data-driven materials ecosystems and autonomous platforms to reveal the systemic nature of dilution. Building on this analysis, we propose the Dilution Cascade Framework, a novel conceptual architecture that models responsibility propagation, identifies dilution nodes, and formalizes steering logics to restore accountability within the computational fabric of discovery.

The framework is interpretive and integrative, offering systems-level insights for the design of future autonomous materials platforms. It positions governance not as an external constraint but as an intrinsic computational dynamic essential for sustainable high-throughput innovation.

Theoretical Background & Literature Synthesis

Foundations of machine learning in materials informatics

The integration of machine learning into materials science began with descriptor-based models for property prediction [1, 8] and evolved rapidly toward graph-based and transformer architectures capable of capturing complex many-body interactions [3, 4]. Foundational works established that models trained on computational databases could generalize across chemical space [2, 11], while text-mining approaches extracted latent knowledge from the literature itself [17].

These developments created the data substrate for autonomous systems. Large-scale repositories and standardized benchmarks [12] enabled the training of models that now serve as the inference engines of self-driving laboratories. The literature consistently highlights the shift from small-data to big-data paradigms [18, 19], yet the governance of these data ecosystems—curation standards, provenance tracking, and bias mitigation—remains secondary to performance metrics [10, 13].

The emergence of autonomous laboratory architectures

Autonomous platforms represent the convergence of robotics, AI planning, and materials-specific heuristics. Early demonstrations in chemistry were followed by integrated systems for inorganic synthesis [7] and thin-film optimization [16]. These platforms close the design–make–test–analyze loop through active learning agents that select experiments, robotic execution that minimizes human intervention, and real-time feedback that refines models [5, 20, 21].

Reviews document the scaling potential of such systems and their capacity to navigate multi-dimensional parameter spaces far beyond human capability [22]. However, the same literature reveals a consistent pattern: the focus remains on throughput and discovery rate, with human oversight positioned as a supervisory layer rather than an embedded computational component [13, 23]. The infrastructure for autonomy is mature; the infrastructure for accountability is not.

Data governance and epistemic infrastructure in computational materials

Parallel to hardware and algorithmic advances, the field has developed sophisticated data infrastructures. Initiatives emphasize FAIR principles, ontology-driven knowledge graphs, and interoperable platforms [3, 10]. Network analyses of synthesizable materials [21, 24] and universal interatomic potentials [3] illustrate how shared data ecosystems amplify discovery velocity.

Yet these infrastructures prioritize accessibility and reusability over traceability of agency. When multiple models, datasets, and agents interact in a federated pipeline [14, 15], the provenance of a discovery outcome becomes distributed. Literature on small-data challenges [13, 18] and uncertainty quantification [12] acknowledges epistemic risks, but frames them as technical hurdles rather than governance failures. The result is a mature computational ecosystem optimized for scale but structurally indifferent to responsibility allocation.

The unaddressed dimension: Responsibility as a computational dynamic

Across the synthesized literature, responsibility appears peripherally—often in discussions of human-in-the-loop safeguards [23] or ethical guidelines for deployment [13]. No framework treats accountability as a quantifiable, propagatable property of the discovery pipeline itself.

Dilution emerges precisely at the interfaces where agency transfers: from curated datasets to trained models, from models to planning agents, from agents to robotic execution, and from execution back to model refinement. Each transfer layer introduces opacity and diffusion. The current paradigm excels at accelerating discovery but lacks native mechanisms to track, mitigate, or restore the epistemic chain of custody.

This synthesis reveals a coherent conceptual gap: the materials community has engineered high-throughput autonomous systems without equivalent investment in governance primitives. The Dilution Cascade Framework addresses this gap by providing a systems-level model that renders responsibility visible, propagatable, and steerable within the computational architecture of discovery.

Proposed Conceptual Framework

The dilution cascade framework: Core philosophy

We introduce the Dilution Cascade Framework (DCF) as an original interpretive architecture for autonomous materials innovation ecosystems. The DCF reconceptualizes discovery pipelines as layered responsibility propagation systems in which throughput acts as both accelerator and diluent. Rather than treating governance as an external ethical overlay, the framework embeds accountability as a native computational dynamic that evolves alongside data, models, and experiments.

The DCF comprises four structural layers connected by bidirectional pipelines and explicit feedback loops. It formalizes dilution as a measurable attenuation of traceable agency and provides steering logics to counteract it. The framework is conceptual and infrastructure-oriented, designed to guide the engineering of next-generation platforms that maintain epistemic integrity at scale.

Structural layers and pipeline dynamics

The DCF defines four interdependent layers that together form the backbone of any autonomous materials discovery system. The data ecosystem layer serves as the foundational substrate, where heterogeneous sources ranging from high-fidelity computational databases to literature-derived knowledge graphs are aggregated and curated. This layer establishes the initial responsibility baseline by enforcing rich provenance metadata and uncertainty annotations that travel with every data element. Building directly upon this foundation, the model inference layer transforms static data into dynamic, actionable hypotheses through predictive and generative architectures that increasingly operate with minimal human calibration. Here, agency begins its first significant migration from human-defined rules to algorithmic inference. The autonomous experimentation layer then operationalizes these hypotheses through networks of robotic platforms and embedded sensors, where physical actions become decoupled from direct human oversight and decision latency drops to seconds rather than days. Finally, the discovery output layer consolidates experimental outcomes into validated materials knowledge, closing the loop by feeding refined insights back into the upstream layers. Throughout this vertical cascade, data and candidate materials flow predominantly downward, while accountability is designed to propagate bidirectionally, though in practice the forward acceleration of throughput consistently outpaces the backward restoration of traceability.

Feedback loops and computational steering logics

To counteract the natural tendency toward dilution, the DCF incorporates three interlocking classes of feedback mechanisms that operate continuously across the layers. Epistemic traceability loops carry uncertainty estimates and provenance signatures from downstream execution back to upstream data and model components, ensuring that no decision is ever made in isolation from its full causal history. Governance steering loops function as dynamic regulators, continuously modulating model confidence thresholds, experiment selection policies, and robotic operational envelopes in response to real-time assessments of accountability erosion. Human-in-the-architecture loops, in turn, create persistent, interpretable intervention ports at every layer boundary, allowing scientific stewards to inject contextual judgment or reclaim oversight without interrupting the overall velocity of the pipeline. These loops are not ancillary safety features but core computational primitives, governed by steering logics that treat accountability itself as an optimizable state variable within the system.

Formalization of key dynamics

The progressive erosion of responsibility across successive layers can be conceptualized as a multiplicative attenuation process expressed as

$$\begin{aligned} R_{\{l+1\}} \\ &= R_l \\ &\cdot (1 - \delta_l) \end{aligned} \quad (1)$$

where R_l represents the residual responsibility traceability at layer l , and δ_l denotes the layer-specific dilution coefficient (bounded between 0 and 1) that quantifies the fraction of agency ceded to automated components without retained provenance metadata. When integrated over the full cascade, the final accountability state becomes

$$\begin{aligned} R_{final} \\ &= R_{initial} \prod_l \\ &= \frac{1}{N} (1 - \delta_l) R_{\{final\}} \end{aligned} \quad (2)$$

providing a symbolic scaffold for diagnosing and preempting fragmentation in any given pipeline configuration.

The restorative action of governance feedback can be captured through the iterative adjustment

$$\begin{aligned} G_{\{t+1\}} \\ &= G_t \\ &+ \eta \left(\frac{T_t \cdot U_t}{A_t} \right) \end{aligned} \quad (3)$$

in which G is the governance steering signal at time step t , T_t is the instantaneous throughput, U_t is the prevailing epistemic uncertainty, A_t is the current accountability level, and η functions as a tunable gain parameter that balances responsiveness against system stability. This relation formalizes the countervailing dynamic whereby elevated throughput amplifies uncertainty unless actively counterbalanced by explicit accountability interventions.

The overarching equilibrium between discovery velocity and epistemic integrity is then described by the composite utility function

$$\begin{aligned} E \\ &= \alpha T \\ &+ \beta (1 - A) \end{aligned} \quad (4)$$

where E denotes the system's epistemic utility, T and A are normalized throughput and accountability, respectively, and the weighting coefficients α and β encode the infrastructure designer's relative prioritization of scale versus stewardship. These expressions remain purely interpretive devices that illuminate the internal logic of the cascade and guide the principled engineering of governance-native platforms.

Visualization and interpretation

As conceptualized in **Figure 1**, responsibility attenuates across the vertical discovery cascade as throughput accelerates, necessitating embedded governance steering loops.

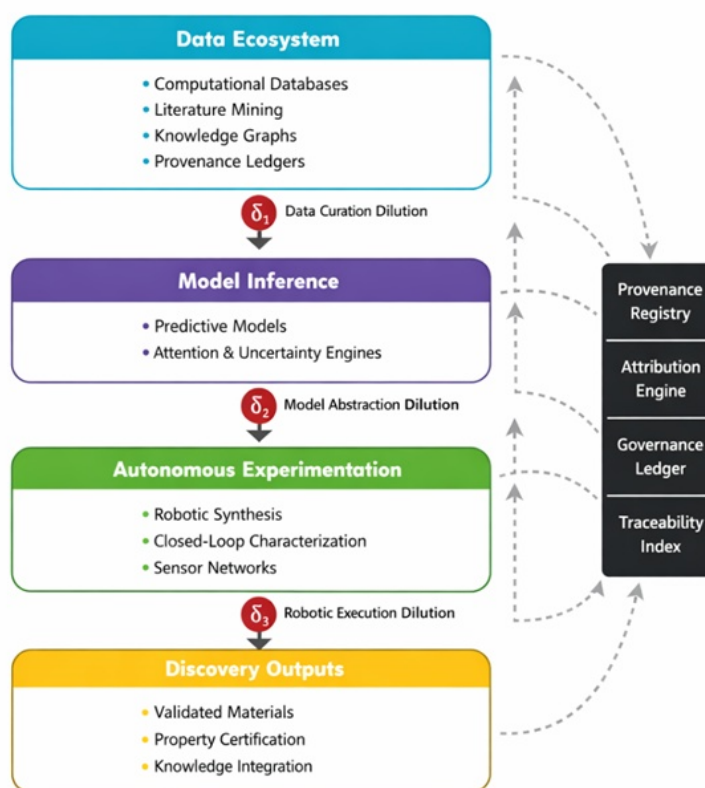


Figure 1. Dilution Cascade Framework (DCF)

A vertical systems architecture illustrating responsibility propagation across four computational layers—data ecosystems, model inference, autonomous experimentation, and discovery outputs. Downward pipelines represent high-velocity discovery throughput, while upward dashed feedback channels encode governance restoration loops. Attenuation nodes (δ_i) at each interface mark responsibility dilution points. A parallel accountability core accumulates provenance signals and redistributes steering interventions through human and algorithmic oversight ports.

Analytical implications

The Dilution Cascade Framework supplies a coherent analytical apparatus for dissecting the internal mechanics of autonomous materials innovation systems and for prescribing targeted architectural interventions that preserve epistemic integrity at industrial scales of operation. When applied to the data ecosystem layer, the framework immediately surfaces the latent dilution that originates in the aggregation of heterogeneous sources. Large-scale repositories, even when meticulously curated according to emerging standards, frequently propagate incomplete lineage information across federated datasets. The DCF treats this incompleteness not as a data-quality footnote but as the primordial coefficient that seeds every subsequent layer. Consequently, platform architects are compelled to embed persistent provenance schemas—blockchain-inspired ledgers or graph-based attribution graphs—at the point of ingestion, ensuring that every training example or literature-derived embedding carries an explicit responsibility vector that survives model compression and transfer. This requirement reframes data infrastructure from a passive repository into an active accountability substrate, one whose design directly modulates the baseline of the entire cascade.

Moving upward into the model inference layer, the framework exposes a second critical dilution node that arises from the very success of contemporary deep-learning architectures. Graph networks and foundation models achieve remarkable generalization precisely by abstracting away the granular provenance of their training data [3, 4]. The DCF interprets this

abstraction as an engineered that trades explanatory fidelity for predictive power. Analytical application of the responsibility attenuation equation reveals that even modest increases in model scale can produce disproportionately large drops in traceability unless counteracted by native interpretability mechanisms—such as attention provenance maps or counterfactual attribution layers—that propagate responsibility signals alongside predictions. In practice, this implies that next-generation materials foundation models must be co-designed with governance primitives rather than retrofitted, a shift that alters the entire training objective from pure loss minimization to a multi-objective optimization that explicitly penalizes untraceable inference paths.

At the autonomous experimentation layer, the framework’s analytical lens illuminates the physical manifestation of dilution in robotic execution. When a planning agent dispatches a synthesis protocol to a distributed fleet of manipulators, the causal chain splinters across hardware controllers, sensor fusion modules, and safety interlocks. Each additional degree of parallelism multiplies the opportunities for δ_3 . The DCF therefore advocates for the insertion of lightweight, real-time accountability telemetry—compact digital twins of decision provenance—that travel with every robotic command and are reconciled against experimental outcomes at sub-second latencies. Such telemetry transforms the experimentation layer from a black-box actuator into a transparent extension of the upstream governance core, allowing the steering loops to modulate robotic envelopes dynamically whenever accountability metrics breach predefined thresholds.

The discovery output layer, often treated as a mere aggregation step in conventional pipelines, emerges under the DCF as the critical reconciliation junction where cumulative dilution either crystallizes into irreversible epistemic debt or is actively reversed through feedback. Here the framework’s utility function E becomes operational: system designers can simulate alternative pipeline configurations and observe how small adjustments in η or in the relative weighting of α and β produce qualitatively different discovery regimes—one optimized for sheer volume, another for traceable, auditable knowledge. This analytical capability extends beyond single laboratories to networked ecosystems in which multiple autonomous platforms interoperate. In such federated settings, the DCF supplies the formal scaffolding for cross-platform responsibility contracts, enabling institutions to share discovery outputs without relinquishing epistemic custody.

Taken together, these implications redefine the engineering brief for autonomous materials platforms. Rather than bolting governance onto mature workflows as an afterthought, the DCF demands that accountability be treated as a first-class computational resource, comparable in architectural status to compute, data, or model capacity. The result is a new class of infrastructure whose key performance indicators include not only candidates synthesized per hour but also responsibility retention ratios across the cascade. By rendering these ratios explicit and optimizable, the framework converts what has been an invisible structural vulnerability into a design variable that can be engineered with the same rigor historically reserved for materials properties themselves. Key responsibility attenuation mechanisms across the discovery stack are systematized in [Table 1](#).

Table 1. Responsibility Dilution Nodes across Autonomous Materials Discovery Pipelines

Pipeline Layer	Primary Agents	Dilution Mechanism	δ Coefficient Driver	Governance Intervention Lever
Data Ecosystems	Repositories, literature mining systems, federated datasets	Loss of provenance granularity	Dataset aggregation scale	Persistent attribution ledgers
Model Inference	GNNs, foundation models, generative engines	Abstraction of training lineage	Model complexity + compression	Explainable inference mapping
Experimentation	Robotic synthesis platforms, sensor arrays	Execution opacity	Hardware parallelism	Accountability telemetry twins
Orchestration	Planning agents, workflow schedulers	Decision delegation layering	Automation depth	Policy-encoded decision gates

Discovery Outputs	Validation systems, certification protocols	Attribution collapse	Multi-agent convergence	Responsibility reconciliation registries
-------------------	---	----------------------	-------------------------	--

Results and Discussion

The Dilution Cascade Framework situates itself at the intersection of two rapidly maturing domains: the engineering of high-throughput autonomous laboratories and the broader societal imperative for responsible artificial intelligence in scientific discovery. Its central contribution lies in demonstrating that the governance deficit observed in current materials innovation ecosystems is not an accidental byproduct of rapid technological advance but a predictable consequence of architectural choices that privilege forward velocity over bidirectional traceability. By providing a unified language and symbolic apparatus for describing this phenomenon, the framework bridges the largely technical discourse of computational materials science with the emerging requirements of research integrity in an age of algorithmic agency. As illustrated in **Figure 2**, responsibility does not merely cascade—it diffuses laterally across distributed computational actors, creating networked opacity zones.

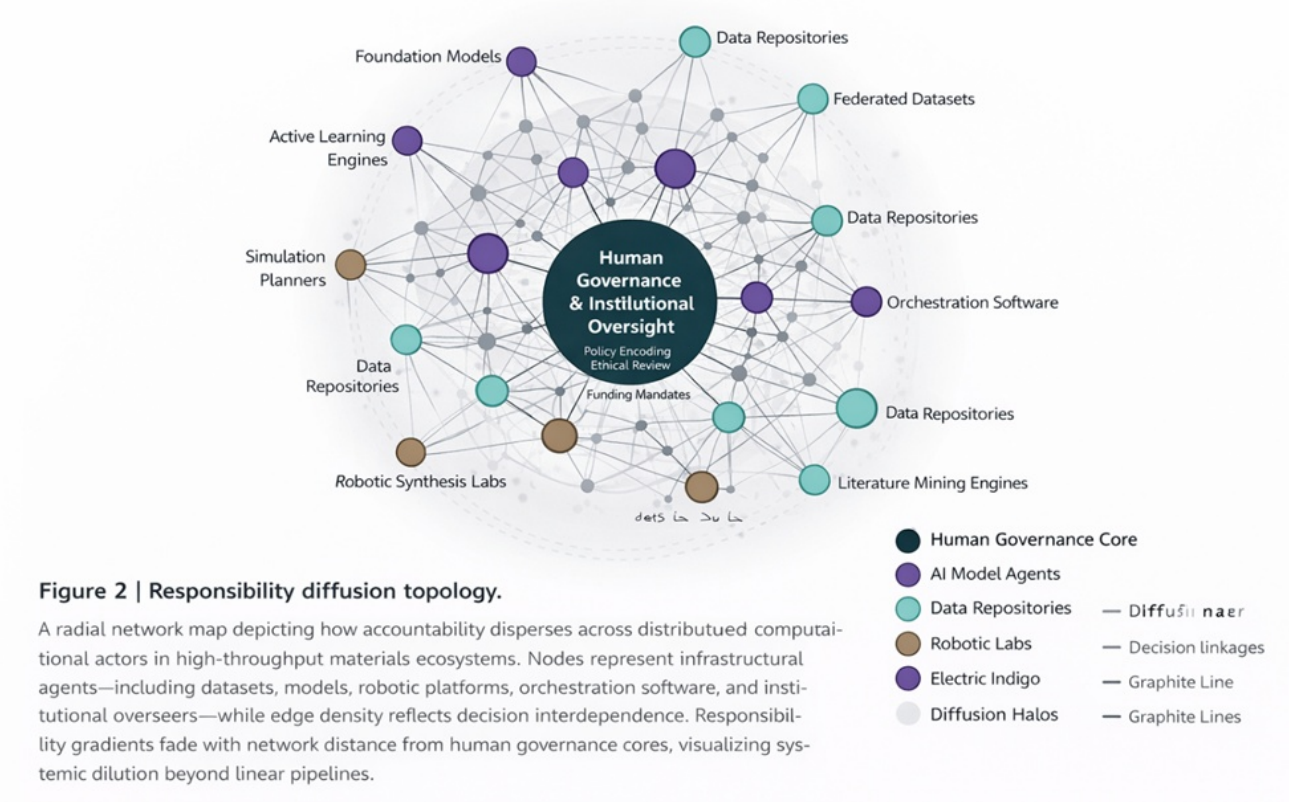


Figure 2. Responsibility Diffusion Topology in High-Throughput Autonomous Discovery Networks

One particularly salient implication concerns the evolution of standards and best practices within the field. Initiatives that have successfully standardized data formats and ontology schemas [10] could be extended to encompass accountability metadata as a native element of the FAIR principles—perhaps under a new “FAIR-A” designation where the additional “A” denotes accountability. Such an extension would enable repositories, model hubs, and robotic orchestration layers to interoperate under a shared responsibility protocol, much as materials databases today interoperate under common

schemas. The DCF supplies the conceptual blueprint for implementing these protocols without imposing prohibitive computational overhead, thereby offering a practical pathway from aspiration to infrastructure.

At the same time, the framework invites a sober assessment of inherent tensions. Embedding persistent provenance and steering mechanisms inevitably introduces additional latency and storage demands. In high-velocity regimes where decisions must be rendered in milliseconds, the marginal cost of accountability telemetry must be weighed against the marginal benefit of further throughput gains. The DCF's utility function E provides a principled mechanism for navigating this trade-off, yet real-world deployment will require community-wide consensus on acceptable thresholds for coefficients—thresholds that may differ across application domains ranging from energy-storage materials to quantum computing components. These normative choices lie beyond the scope of any purely conceptual model and will necessarily emerge through iterative dialogue among researchers, funders, and infrastructure providers.

The framework also surfaces questions about the evolving role of human scientists within increasingly autonomous ecosystems. Far from rendering researchers obsolete, the DCF repositions them as strategic stewards who operate at the governance core, intervening at high-leverage nodes rather than micromanaging every experimental cycle. This shift demands new competencies—fluency in both domain science and computational governance—that current training paradigms are only beginning to address. Over the coming decade, graduate programs and professional development initiatives may therefore need to incorporate modules on pipeline stewardship and epistemic risk management, treating these as core literacies alongside density-functional theory or phase-field modeling.

Finally, the DCF carries implications that extend beyond the materials community. As autonomous discovery platforms proliferate across chemistry, biology, and manufacturing, the responsibility dilution phenomenon is likely to manifest in analogous forms. The framework's layered, feedback-rich architecture offers a transferable template for other domains, suggesting that the governance challenges of high-throughput science are fundamentally infrastructural rather than discipline-specific. In this sense, the materials field—long accustomed to serving as an early adopter of computational paradigms—once again finds itself in a position to shape broader norms of accountable innovation in the algorithmic era.

Conclusion

The Dilution Cascade Framework reframes the spectacular throughput gains of autonomous materials innovation as a profound governance challenge rather than a simple success story of engineering efficiency. By mapping the propagation and attenuation of responsibility across the canonical layers of data, model, experimentation, and discovery, the framework renders visible the structural dynamics that have remained largely implicit in the literature. Its original formalizations—capturing multiplicative dilution, iterative governance steering, and epistemic utility—provide symbolic tools that translate abstract concerns about accountability into concrete design parameters for next-generation platforms.

The ultimate value of the DCF lies in its capacity to guide the construction of materials innovation ecosystems that are simultaneously faster and more trustworthy. When responsibility is engineered as a native property of the computational fabric rather than an external ethical constraint, high-throughput discovery ceases to be a vector for diffused agency and becomes instead a foundation for robust, auditable scientific infrastructure. The framework therefore stands as both diagnostic instrument and prescriptive blueprint, offering the materials community a pathway to accelerate responsibly at the very moment when acceleration has never been more urgent.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Jan 2025

Revised: 19 Feb 2025

Accepted: 05 May 2025

Published online: 18 September 2025

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ramprasad R, Batra R, Pilania G, Mannodi-Kanakithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. *npj Comput Mater.* 2017;3(1):54.
<https://doi.org/10.1038/s41524-017-0056-5>.
- Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater.* 2019;5:83.
<https://doi.org/10.1038/s41524-019-0221-0>.
- Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci.* 2022;2(11):718-28.
- Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater.* 2019;31(9):3564-72.
<https://doi.org/10.1021/acs.chemmater.9b01294>.
- 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>.
- 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>.

Fei Y, Gallant M, Persson K, Szymanski NJ, Rendy B, Kumar RE, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*. 2023;624(7986):86–91.
<https://doi.org/10.1038/s41586-023-06734-w>.

Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. *InfoMat*. 2019;1(3):338–58.
<https://doi.org/10.1002/inf2.12028>.

Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat Rev Mater*. 2021;6(8):655–78.
<https://doi.org/10.1038/s41578-020-00255-y>.

Butler KT, Choudhary K, Csanyi G, Ganose AM, Kalinin SV, Morgan D. Setting standards for data driven materials science. *npj Comput Mater*. 2024;10:231.
<https://doi.org/10.1038/s41524-024-01411-6>.

Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: T Matbench test set and Automatminer reference algorithm. *npj Comput Mater*. 2020;6:138.
<https://doi.org/10.1038/s41524-020-00406-3>.

Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*. 2016;533(7601):73–6.

Nandy A, Duan C, Kulik HJ. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Curr Opin Chem Eng*. 2022;36:100779.
<https://doi.org/10.1016/j.coche.2021.100779>.

Moosavi SM, Nandy A, Jablonka KM, Ongari D, Janet JP, Boyd PG, et al. Understanding the diversity of the metal-organic framework ecosystem. *Nat Commun*. 2020;11(1):4068.
<https://doi.org/10.1038/s41467-020-17755-8>.

Jha D, Ward L, Paul A, Liao WK, Choudhary A, Wolverton C, et al. ElemNet: Deep learning the chemistry of materials from only elemental composition. *Sci Rep*. 2018;8(1):17593.
<https://doi.org/10.1038/s41598-018-35934-y>.

Aykol M, Hegde VI, Suram S, Hung L, Sumpter BG, Persson KA. Network analysis of synthesizable materials discovery. *Chem Mater*. 2019;31(4):1373–82.

Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*. 2019;571(7763):95–8.
<https://doi.org/10.1038/s41586-019-1335-8>.

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>.

Zhou Q, Chen X, Wang J. Machine learning assisted material discovery: A small data approach. *Acc Mater Res*. 2022;3(6):685–94.
<https://doi.org/10.1021/accountsmr.1c00236>.

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett*. 2018;120(14):145301.
<https://doi.org/10.1103/PhysRevLett.120.145301>.

Dinic F, Voznyy O. Unconstrained machine learning screening for new li-ion cathode materials enhanced by class balancing. *Adv Theory Simul*. 2023;6(6):2300081.

Persaud D, Ward L, Hatrick-Simpers J. Reproducibility in materials informatics: lessons from 'A general-purpose machine learning framework for predicting properties of inorganic materials'. *Digit Discov.* 2024;3(2):281-6.

Choudhary K, DeCost B, Tavazza F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys Rev Mater.* 2018;2(8):083801.
<https://doi.org/10.1103/PhysRevMaterials.2.083801>.

Pyzer-Knapp EO, Pitera JW, Staar PWJ, Takeda S, Laino T, Sanders DP, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater.* 2022;8:84.
<https://doi.org/10.1038/s41524-022-00765-z>.