ORIGINAL RESEARCH **Open access**

Discovery Pipelines as Epistemic Filters: What Computational Workflows Exclude

Hiroshi Tanaka[1], Yuki Sato[1*], Kenji Mori[2]

Abstract

In the evolving landscape of computational and data-driven materials engineering, discovery pipelines integrate machine learning, high-throughput computations, and autonomous systems to accelerate the identification of novel materials. These workflows, encompassing materials informatics, representation learning, and inverse design, operate as structured sequences that process vast datasets to infer properties and guide experimentation. However, inherent in their design are epistemic filters—mechanisms that selectively emphasize certain knowledge pathways while excluding others, potentially limiting the breadth of scientific insight. This manuscript addresses this conceptual gap by examining how computational architectures, such as graph neural networks and foundation models, impose exclusions through representation biases, uncertainty handling, and feedback dynamics. We introduce the Epistemic Filtration Framework (EFF), a novel systems-level model that maps data ingestion, model inference, and discovery steering to reveal excluded epistemic domains. By interpreting pipeline interactions, the framework highlights trade-offs in multimodal integration and simulation-experiment coupling, offering insights into enhancing workflow inclusivity. Implications extend to materials research ecosystems, fostering more comprehensive discovery logics without empirical validation. This conceptual analysis underscores the need for reflective infrastructure design in AI-augmented materials science, balancing efficiency with epistemic completeness.

Keywords Materials informatics, Uncertainty quantification, Representation learning, Discovery pipelines, Computational workflows, Epistemic filters

*Correspondence:
Yuki Sato
yuki.sato@outlook.com
[1] Department of Computational Materials Engineering, Faculty of Engineering, University of Tokyo, Tokyo, Japan
[2] Department of Data-Driven Materials Discovery, Faculty of Engineering, Kyoto University, Kyoto, Japan

Introduction

The emergence of computational paradigms in materials engineering

The field of materials engineering has undergone a profound transformation with the advent of computational and data-driven approaches, shifting from traditional trial-and-error methods to systematic, algorithmically guided discovery [1, 2]. Historically, materials innovation unfolded through iterative cycles of synthesis, testing, and refinement, processes that were both time-intensive and constrained by laboratory throughput. While such empirically grounded strategies produced foundational advances across structural alloys, semiconductors, and functional ceramics, their exploratory reach remained inherently bounded by physical experimentation. The emergence of computational paradigms has fundamentally reoriented this landscape, enabling virtual interrogation of materials behaviors prior to synthesis and thereby repositioning computation as a primary engine of discovery rather than a supplementary analytical tool.

This paradigm leverages vast computational resources to simulate material behaviors, predict properties, and optimize designs, thereby reducing the time and cost associated with physical experimentation. Multiscale simulations, electronic structure modeling, and thermodynamic calculations now operate in concert with

machine learning systems, forming hybridized discovery environments capable of traversing expansive compositional and structural design spaces. Central to this shift are discovery pipelines—integrated, multi-stage computational workflows that process input data through predictive, representational, and optimization layers to yield actionable scientific insights. These pipelines represent not simply technical infrastructures but orchestrated systems of inference in which data transformation, modeling, and decision logic unfold sequentially and iteratively.

Such pipelines incorporate high-throughput screening architectures in which thousands of candidate materials are evaluated virtually through automated simulation stacks [3, 4]. Machine learning models embedded within these workflows learn from historical datasets to forecast functional properties, stability windows, and performance envelopes, augmenting simulation outputs with predictive generalization [5, 6]. In perovskite materials design, for instance, integrated workflows combine density functional theory calculations with supervised learning predictors to identify stable compositions, defect tolerances, and optoelectronic properties across vast chemical permutations [7, 8]. Comparable computational ecosystems are now operational across catalysis, battery materials, biomaterials, and quantum functional systems, collectively redefining the scale and speed of materials exploration.

Beyond efficiency gains, these infrastructures enable access to previously inaccessible regions of materials possibility space. Computational enumeration and predictive modeling allow researchers to probe metastable phases, high-entropy alloys, and extreme-condition compounds that would be experimentally prohibitive to explore exhaustively. In this sense, computational discovery pipelines do more than accelerate research—they expand the ontological boundaries of what can be systematically investigated within materials science.

Epistemic filters within computational workflows

Yet, as these pipelines increase in complexity and autonomy, they introduce layered abstractions that actively shape the epistemic landscape of materials research. Discovery workflows do not merely transmit knowledge from data to insight; they structure how knowledge is constituted, filtered, and prioritized within computational environments.

Epistemic filters refer to the implicit mechanisms embedded within workflows that privilege certain interpretations, inference pathways, or discovery trajectories over others, often rooted in algorithmic design choices, representational encodings, and data integration logics [9, 10]. These filters operate through multiple computational strata. At the data level, curation decisions determine which materials systems are represented. At the representation level, feature engineering and embedding architectures encode selective structural and chemical attributes. At the modeling level, objective functions and training regimes privilege particular predictive targets.

Within computational materials science, such filters manifest through descriptor selection in machine learning pipelines, parameterization strategies in atomistic simulations, and discretization choices in numerical modeling environments, each influencing what forms of knowledge are computationally amplified or attenuated [11]. Feature spaces constructed around crystallographic symmetry or stoichiometric ratios, for example, may insufficiently capture microstructural disorder, defect clustering, or processing-induced heterogeneity.

The proliferation of deep learning architectures has intensified these epistemic dynamics. Graph neural networks, attention-based materials models, and physics-informed representation systems embed structural priors directly into model architectures, shaping how relational atomic information is processed and interpreted [12, 13]. While these systems enhance predictive performance, they simultaneously encode inductive biases that delimit interpretive flexibility. Consequently, representational power becomes inseparable from epistemic selectivity within computational discovery ecosystems.

Challenges in data-driven discovery ecosystems

Despite their transformative potential, data-driven discovery ecosystems face structural challenges arising from the integration of heterogeneous data infrastructures. One of the most significant complexities lies in the harmonization of multimodal datasets, where experimental measurements, simulation outputs, structural descriptors, and processing metadata must be rendered interoperable within unified analytical frameworks [14, 15].

Representation learning architectures are frequently deployed to encode these disparate modalities into shared

latent vector spaces, enabling cross-modal inference and similarity mapping. However, such embeddings may compress or obscure mechanistic subtleties if cross-modal dependencies are inadequately captured [16, 17]. Latent representations optimized for predictive accuracy may fail to preserve synthesis pathways, kinetic barriers, or metastability signatures that are critical to experimental realization.

Uncertainty quantification introduces an additional epistemic dimension within these workflows. While probabilistic modeling techniques exist to characterize predictive confidence, many operational pipelines emphasize deterministic outputs, privileging point estimates over distributional inference [18, 19]. This orientation risks filtering out exploratory trajectories characterized by high epistemic uncertainty, even though such regions may harbor unconventional or high-impact materials candidates.

Autonomous discovery and reinforced filtration dynamics

The integration of autonomous experimentation platforms further compounds these epistemic dynamics. Closed-loop discovery systems couple machine learning predictors with robotic synthesis and high-throughput characterization infrastructures, enabling iterative hypothesis generation and validation cycles [20, 21]. These systems dramatically accelerate optimization processes, allowing materials candidates to be refined through rapid feedback iterations.

However, autonomy introduces reinforcement effects within epistemic filtration processes. Models trained on historically bounded datasets may preferentially explore high-confidence regions of design space, reinforcing existing knowledge contours while neglecting anomalous or low-density regimes [22, 23]. In this way, closed-loop optimization may inadvertently narrow exploratory diversity even as it accelerates convergence.

High-throughput computational infrastructures impose further filtration through scalability logics. Materials classes amenable to rapid simulation workflows—particularly crystalline solids with periodic structures—are disproportionately represented, whereas disordered systems, amorphous phases, and complex interfacial materials remain computationally marginalized due to resource intensity [24, 25].

Inverse design frameworks introduce additional epistemic compression. When target properties are specified and generative or optimization algorithms back-propagate candidate compositions, solution spaces are often collapsed into scalar objective landscapes, potentially obscuring degeneracy structures, competing trade-offs, and multi-objective equilibria [26]. As a result, discovery pathways become streamlined but epistemically narrowed.

Simulation–experiment coupling and foundation model mediation

Coupled simulation–experimental ecosystems add further layers of epistemic mediation. Experimental validation does not operate as a neutral arbiter but as a selective reinforcement mechanism shaped by instrumentation limits, measurement noise, and throughput constraints. The integration of experimental feedback into computational pipelines thus introduces filtration based on detectability and feasibility.

The emergence of foundation models for scientific discovery amplifies these infrastructural abstractions. Trained on expansive corpora of materials data, literature, and simulation records, such models promise cross-domain transferability and generative hypothesis formation [27, 28]. Yet their scale embeds historical research biases, dominant chemistries, and prevailing theoretical assumptions within latent model structures.

Consequently, rare materials classes, unconventional processing pathways, or sparsely studied phenomena may remain underrepresented within model priors. The epistemic inclusivity promised by scale may therefore coexist with structural exclusion of edge-case knowledge domains.

Positioning the conceptual inquiry

These intersecting dynamics reveal a foundational conceptual gap within computational materials engineering. Discovery pipelines, while optimized for predictive acceleration and design efficiency, function simultaneously as epistemic filtration systems that shape the boundaries of computationally accessible knowledge [29]. Their architectures encode inclusion criteria, prioritization logics, and interpretive compressions that influence which materials phenomena are rendered visible or invisible within algorithmic search regimes.

This manuscript therefore positions discovery pipelines not merely as instruments of technological advancement but as epistemic infrastructures that actively configure knowledge production in materials engineering. By synthesizing contemporary developments across computational workflows, representation learning, and autonomous experimentation, we identify recurrent filtration patterns embedded within pipeline architectures.

To systematically interrogate these dynamics, we introduce the Epistemic Filtration Framework (EFF), an original conceptual model that dissects discovery workflows into layered epistemic strata encompassing data ingestion, representation encoding, model inference, and optimization steering. Through this systems-level lens, the framework elucidates how computational design choices filter epistemic content, shaping discovery trajectories and interpretive horizons.

In articulating these filtration mechanisms, the manuscript advances a broader theoretical repositioning of AI-guided materials discovery—from a paradigm defined solely by acceleration to one equally characterized by epistemic structuring. Such a reframing opens pathways toward more reflexive, inclusive, and epistemically aware computational materials engineering infrastructures.

Theoretical Background & Literature Synthesis

Foundations of Materials Informatics and Machine Learning Integration Materials informatics has emerged as a cornerstone of data-driven materials engineering, providing the infrastructural backbone for organizing and analyzing vast repositories of material data [1, 2]. This discipline integrates machine learning to extract patterns from structured datasets, enabling predictions of properties such as bandgaps, thermal conductivities, and mechanical strengths [3, 4]. Key to this integration is representation learning, where atomic structures are encoded into machine-readable formats, often using descriptors like Coulomb matrices or graph-based representations [5, 6]. These representations facilitate the application of deep learning architectures, including convolutional and graph neural networks, which model hierarchical relationships in material systems [7, 8]. Literature emphasizes how such integrations streamline discovery by bridging atomic-scale simulations with macroscopic predictions, yet they inherently select for data modalities that align with the chosen encoding schemes [9, 10].

High-throughput computational methods further extend this foundation by enabling parallel evaluations of material candidates [11]. These approaches, often coupled with density functional theory, generate extensive datasets for training models, accelerating the identification of functional materials like perovskites or metallic glasses [12, 13]. However, the synthesis reveals that high-throughput pipelines prioritize computational tractability, often excluding materials with complex electronic structures that demand higher fidelity simulations [14, 15].

Autonomous Systems and Closed-Loop Dynamics Autonomous discovery systems represent an advanced evolution, incorporating robotics and real-time experimentation into computational loops [16, 17]. These systems use active learning to select experiments that maximize information gain, iteratively refining models based on feedback [18, 19]. In materials contexts, such as battery electrolyte design or organic luminophore synthesis, closed loops couple predictive models with automated synthesis platforms [20, 21]. The literature synthesizes these as self-correcting infrastructures that enhance discovery efficiency, but they introduce epistemic steering where model confidence directs exploration, potentially bypassing low-probability yet innovative pathways [22, 23].

Inverse design paradigms complement autonomy by starting from desired properties and optimizing toward viable compositions [24]. Generative models, including variational autoencoders and generative adversarial networks, sample chemical spaces to propose candidates [25, 26]. Synthesis of these works highlights their role in navigating high-dimensional spaces, yet exclusions arise from the optimization objectives, which may favor local minima over global diversity [27, 28].

Uncertainty Quantification and Multimodal Integration Uncertainty quantification is pivotal in ensuring the robustness of computational predictions, employing techniques like Bayesian inference or ensemble methods to capture aleatoric and epistemic uncertainties [5, 18]. In materials AI, this involves propagating uncertainties through workflows to inform decision-making [6, 19]. The synthesis indicates that while these methods improve reliability, they can filter out exploratory inferences if uncertainty thresholds are set conservatively [29].

Multimodal datasets, combining textual descriptions, images, and numerical simulations, pose integration challenges [14]. Foundation models adapt natural language processing techniques to scientific data, enabling cross-domain inferences [27]. However, literature points to exclusions in coupling simulations with experiments, where discrepancies in data fidelity lead to filtered alignments [15, 20]. Overall, this synthesis underscores that while computational workflows advance materials discovery, their layered designs impose epistemic filters through selective representation, steering, and integration, necessitating a framework to interpret these dynamics.

Proposed conceptual framework

The Epistemic Filtration Framework (EFF) To interpret the exclusions inherent in computational discovery pipelines, we introduce the Epistemic Filtration Framework (EFF), an original systems-level model that conceptualizes workflows as multi-layered filters operating on epistemic content. The EFF dissects pipelines into three structural layers: data ingestion, model inference, and discovery steering. At the data ingestion layer, raw inputs from multimodal sources—such as structural databases, simulation outputs, and experimental measurements—are filtered through representation schemes that prioritize certain features, excluding others based on encoding biases. The model inference layer applies architectures like graph neural networks to process these representations, generating predictions while marginalizing uncertainties or emergent interactions not captured in the training manifold. Finally, the discovery steering layer incorporates feedback loops to guide iterative refinement, often reinforcing dominant pathways at the expense of alternative epistemic domain, revealing that exclusions are produced at multiple strata rather than at a single model stage **(Table 1)**.

Table 1. Epistemic filtration mechanisms across discovery pipeline layers in the Epistemic Filtration Framework (EFF)

EFF layer	Typical pipeline components	Filtration mechanism (how selection occurs)
Data ingestion	Databases, simulation corpora, experimental repositories, literature-derived data	Coverage and availability constraints; dataset priors embedded by historical sampling
Representation encoding (R)	Descriptor design, graph construction choices, modality fusion embeddings	Dimensional compression; modality weighting; feature selection and featurization priors
Model inference (I)	GNNs, deep predictors, surrogates, foundation models	Inductive bias and manifold constraint; objective-aligned learning
Uncertainty handling	Ensembles, Bayesian approximations, calibration layers	Thresholding and decision gating; reducing distributions to point outputs
Discovery steering	Active learning, inverse design optimization, acquisition functions	Convergence pressure; scalarization of trade-offs; exploitative policies
Simulation–experiment coupling	Closed-loop pipelines,	Detectability/feasibility gating; coupling

	feasibility filters, lab constraints	misalignment; fidelity mismatch

Central to the EFF are feedback loops that propagate exclusions across layers. For instance, uncertainty signals from inference can loop back to ingestion, adjusting data selection to favor low-uncertainty regimes. This dynamic creates a self-reinforcing filtration, where pipelines evolve toward efficiency but narrow epistemic breadth. Computational steering logics, embedded in autonomous systems, further modulate this by directing resources toward high-confidence discoveries, excluding speculative explorations.

The interaction between representation and inference can be conceptualized as a filtration function, where the epistemic output E_{out} is a subset of the input epistemic space E_{in}, modulated by a representation matrix R and inference operator I: $E_{out} = I(R \cdot E_{in})$. This captures the reductive nature of workflows, where dimensional compression in R excludes latent variables, and I amplifies selected patterns.

In high-throughput contexts, trade-offs emerge in pipeline scalability, expressed as a balance between computational depth D and breadth B: Efficiency ≈ D / B, where increasing D (detailed simulations) reduces B (explored candidates), filtering out diverse but computationally intensive options.

Multimodal integration introduces another dynamic, where coupling strength C between simulation and experiment modalities may be expressed as C = ∫ (alignment factors) ds, integrating over shared epistemic spaces s; weak C leads to exclusions in uncoupled domains.

These elements are interconnected as conceptualized in **Figure 1**, which illustrates the EFF as a directed acyclic graph with nodes for layers and edges for feedback loops, overlaid with filtration funnels representing exclusions at each stage. The figure depicts data flowing from ingestion through inference to steering, with dashed lines indicating excluded epistemic branches, emphasizing the framework's interpretive role in revealing hidden workflow dynamics.

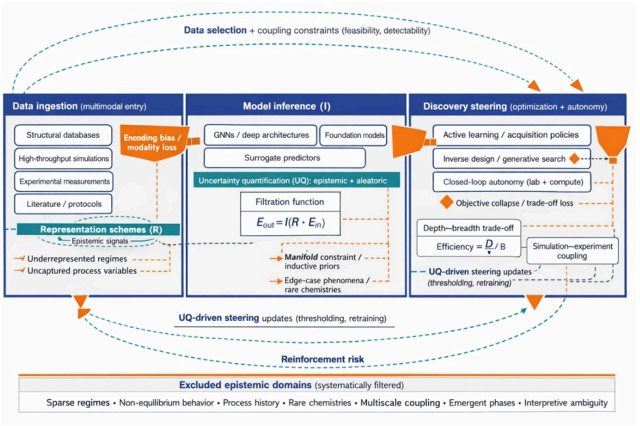

Figure 1. Epistemic Filtration Framework (EFF) conceptualizes discovery pipelines as layered epistemic filters

Data ingestion converts multimodal sources into computational representations (R), model inference applies learned operators (I) with uncertainty quantification overlays, and discovery steering implements optimization and autonomous closed-loop decisions that prioritize specific search trajectories. At each layer, filtration funnels depict excluded branches of epistemic content (e.g., underrepresented regimes, emergent interactions, edge-case phenomena) that are structurally marginalized by encoding choices, inductive priors, and steering convergence. Teal dashed feedback loops illustrate how uncertainty-driven updates and simulation–experiment coupling can reinforce exclusions over iterations, yielding self-reinforcing narrowing of the computationally accessible epistemic space, consistent with the filtration relation.

Analytical implications

Implications for Representation Learning and Model Architectures The Epistemic Filtration Framework (EFF) illuminates how representation learning in computational workflows acts as an initial gatekeeper, shaping the epistemic content available for downstream inference [1, 2]. In materials informatics, representations such as graph-based encodings capture atomic connectivity but may exclude dynamic properties like thermal fluctuations if static descriptors dominate [3, 4]. This filtration implies a trade-off in workflow design: enhancing representational fidelity increases computational overhead, potentially restricting pipeline scalability [5, 6]. For graph neural networks, the framework suggests that message-passing mechanisms amplify local interactions while filtering global emergent behaviors, leading to epistemic narrowing in large-scale

structures [7, 8]. Analytically, this can be expressed as a propagation decay function, where epistemic retention R decays with layer depth L: R = e^{-α L}, with α representing filtration strength due to architectural choices; higher α implies greater exclusion of distant interactions.

In deep learning architectures applied to materials, the EFF highlights implications for handling multimodal data, where fusion layers integrate disparate inputs but often prioritize dominant modalities [9, 10]. This results in workflows that steer toward data-rich domains, excluding underrepresented regimes such as rare-earth compounds [11, 12]. The analytical insight here is that integration strategies must account for epistemic imbalances, fostering designs that adaptively weight modalities to mitigate filtration biases [13, 14].

Dynamics of Uncertainty Quantification in Pipelines Uncertainty quantification intersects with EFF layers by modulating inference reliability, yet it introduces its own filtrations [15, 16]. In high-throughput computations, probabilistic models like Bayesian neural networks provide variance estimates, but conservative thresholding can exclude high-uncertainty candidates that may harbor innovative materials [17, 18]. The framework implies that uncertainty feedback loops, when tightly coupled, reinforce epistemic convergence, limiting diversity in discovery outputs [19, 20]. This dynamic may be captured as an uncertainty amplification loop: $U_{n+1} = U_n + \beta (F - T)$, where U is uncertainty at iteration n, F is filtration factor, T is threshold, and β is loop gain; positive β escalates exclusions over iterations.

For autonomous systems, the implications extend to closed-loop experimentation, where uncertainty steers resource allocation [21, 22]. The EFF suggests that over-reliance on low-uncertainty paths filters out exploratory branches, implying a need for hybrid logics that periodically inject epistemic perturbations to broaden the filtration aperture [23, 24].

Steering Logics and Infrastructure Trade-offs Discovery steering logics, as per the EFF, govern how pipelines navigate chemical spaces, often optimizing for efficiency at the cost of comprehensiveness [25, 26]. In inverse design, gradient-based optimizations filter toward local optima, excluding multi-modal landscapes [27, 28]. Analytically, this trade-off can be framed as an optimization boundary: S = argmax (Utility) subject to E_excluded ≤ γ, where S is steering decision, Utility balances performance and cost, and γ bounds acceptable exclusions; this formalizes the implicit epistemic cost in workflow infrastructure.

The framework's implications for simulation-experiment coupling reveal that misalignments in data fidelity lead to filtered inferences, particularly in foundation models where pre-training biases exclude domain-specific nuances [29]. This underscores analytical needs for adaptive infrastructures that dynamically recalibrate filtrations, enhancing overall epistemic inclusivity in materials engineering ecosystems.

Results and Discussion

The Epistemic Filtration Framework (EFF) provides a lens for interpreting the systemic exclusions in computational discovery pipelines, integrating insights from materials informatics to autonomous systems [1, 2]. By framing workflows as layered filters, the EFF reveals how design choices— from data ingestion to steering—shape epistemic outcomes, often prioritizing targeted efficiency over broad exploration [3, 4]. This interpretation aligns with observed patterns in machine learning applications, where representation biases in graph neural networks exclude certain structural motifs, as seen in perovskite and glass discoveries [5-8]. This exclusion is not incidental but produced by interacting filtration drivers distributed across workflow strata **(Table 2)**.

Table 2. Epistemic exclusion map: excluded domains, pipeline drivers, and where they originate in EFF

Excluded epistemic domain	Typical driver inside computational workflows	Primary EFF origin (layer)	ex pe (fe l
Underrepresented compositional regimes	Training-set density and benchmark availability	Data ingestion	S regi h unc unc g p sea

			to re
Non-equilibrium and kinetic phenomena	Preference for equilibrium descriptors and static structure encodings	Representation encoding (R)	If ki abs emb inf c “rec st op w repre
Process history and synthesis pathway dependence	Lack of standardized processing metadata; modality imbalance	Data ingestion + R	M me pr co moc str pr sh
Emergent multiscale coupling	Local inductive biases (e.g., message passing locality)	Model inference (I)	M le pri s sir an lo
Degenerate solutions and Pareto diversity	Scalarized objectives; convergence pressure in inverse design	Discovery steering	Opti co mu fe re repe
Out-of-distribution behavior	Foundation-model priors; surrogate extrapolation	I + UQ handling	Ca m v unc be reje exp
Simulation–experiment mismatch	Fidelity gaps; detectability constraints; feasibility gating	Coupling interface	Loop pr w mea co
			st shap
Interpretive ambiguity in multimodal fusion	Fusion methods privileging dominant modality	R + I	Mis mo are we unc ma em smo

A key discussion point is the interplay between feedback loops and epistemic risks. In closed-loop systems, iterative refinements can entrench initial biases, leading to a homogenized discovery space [9-12]. This raises questions about workflow resilience: how can infrastructures incorporate mechanisms to detect and counteract filtrations without compromising automation? The EFF suggests interpretive strategies, such as modular layers that allow for epistemic audits, enabling researchers to trace exclusions back to specific components [13, 14].

Furthermore, the framework engages with uncertainty dynamics, where quantification methods serve dual roles as enablers and filters [15, 16]. In multimodal contexts, incomplete integrations exclude cross-domain insights, as in simulation-experiment mismatches [17, 18]. Discussion extends to high-throughput paradigms, where scalability trade-offs filter resource-intensive explorations, potentially overlooking breakthrough materials [19, 20]. Inverse design logics amplify this, as generative approaches sample constrained subspaces [21, 22].

Broader ecosystem implications involve balancing computational steering with epistemic diversity. Foundation models, while versatile, embed filtrations from training corpora, suggesting a need for reflective design practices [23, 24]. The EFF encourages integrative thinking, where workflow architects consider exclusion mappings alongside performance goals [25, 26]. Challenges remain in applying such frameworks conceptually, as they require shifting from empirical validation to systems-level analysis [27, 28].

Ultimately, the discussion posits that recognizing pipelines as epistemic filters fosters more equitable materials research, mitigating risks of overlooked innovations [29] This conceptual shift could inform future infrastructures, promoting dynamics that adaptively widen filtration apertures.

Conclusion

In computational and data-driven materials engineering, discovery pipelines function as epistemic filters, selectively channeling knowledge while excluding alternative pathways. The Epistemic Filtration Framework (EFF) offers an original interpretive model to dissect these dynamics, highlighting interactions across data, model, and steering layers. Through analytical implications, it reveals trade-offs in representation, uncertainty, and infrastructure that shape discovery logics.

This conceptual analysis underscores the importance of reflective workflow design, encouraging systems that balance efficiency with epistemic inclusivity. By addressing filtrations, materials research can evolve toward more comprehensive ecosystems, enhancing innovation without empirical mandates.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Jul 2022 Revised: 02 Oct 2022 Accepted: 30 Nov 2022
Published online: 18 March 2023

References

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1):54.

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83. https://doi.org/10.1038/s41524-019-0221-0.

Mazhnik E, Oganov AR. Machine learning directed search for ultraincompressible, superhard materials. npj Comput Mater. 2019;5(1):1-7.

Chen L, Tran H, Batra R, Kim C, Ramprasad R. Machine learning models for the prediction of effective thermal properties of porous media. npj Comput Mater. 2019;5(1):1-8.

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

Janet JP, Duan C, Kulik HJ. Uncertainty quantification in molecular property prediction using faith-based artificial neural networks. npj Comput Mater. 2022;8(1):1-13.

Chen G, Tao X, Hajfathalian M, Cao G, Chen Z, Wang J, et al. A critical examination of robustness and generalizability of machine learning prediction of materials properties. npj Comput Mater. 2023;9(1):55.

Konstantakopoulos M, Fadel ER, Prost V, Mannan S, Malgras V, Hayashi K, et al. Interpretable discovery of semiconductors with machine learning. npj Comput Mater. 2023;9(1):1-10.

Jennings PC, Lysgaard S, Hummelshøj JS, Vegge T, Bligaard T. Genetic algorithms for computational materials discovery accelerated by machine learning. npj Comput Mater. 2019;5(1):46.

Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156-63.

Meredig B, Antono E, McCulloch S, Rajan K, Ling Y. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Comput Mater Sci. 2020;171:109203.

Li W, Jacobs R, Morgan D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. Comput Mater Sci. 2018;150:454-63. https://doi.org/10.1016/j.commatsci.2018.04.033.

Elton DC, Boukouvalas Z, Fuge MD, Chung PW. Deep learning for molecular design—a review of the state of the art. Comput Mater Sci. 2019;162:433-46.

Kondo R, Yamakawa S, Masuoka Y, Tajima S, Asahi R. Microstructure recognition using convolutional neural networks for prediction of ionic conductivity in ceramics. Acta Mater. 2017;141:29-38.

Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse materials design. Nat Commun. 2020;11(1):1-9.

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

Zhong W, Yan K, Liu H, Zhao X, Tang Y, Liu X, et al. Machine learning guided discovery of barrierless transitions in two-dimensional heterogeneous catalysis. Nat Commun. 2022;13(1):1-9.

Lu X, Qian P, Chen N, He Z, Wang X, Wu J, et al. Machine learning accelerated discovery of functional glass via human-machine collaboration. Nat Commun. 2022;13(1):1-11.

Sumita M, Yang X, Ishihara S, Tamura R, Tsuda K. Machine learning in materials discovery: Confirmed efficient synthesis of new organic luminophores. Nat Commun. 2018;9(1):1-7.

Zakutayev A, Wunder N, Schwarting M, Perkins JD, White R, Munch K, et al. An open experimental database for exploring inorganic materials. Sci Adv. 2018;4(4):eaaq1566.

Ren F, Ward L, Williams T, Laws KJ, Wolverton C, Hattrick-Simpers J, et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. Sci Adv. 2018;4(4):eaat6049.

Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. Sci Adv. 2017;3(12):e1701816.

Dave A, Mitchell J, Kandasamy K, Wang H, Burke B, Paria B, et al. Autonomous discovery of battery electrolytes with robotic experimentation and machine learning. Cell Rep Phys Sci. 2020;1(12):100264.

Keith JA, Vassilev-Galindo V, Cheng B, Chmiela S, Gastegger M, Müller K-R, et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. Chem Rev. 2021;121(15):9816-72.

Borg CKH, Muckley ES, Nyby C, Saal JE, Ward L, Mehta A, et al. Quantifying the performance of machine learning models in materials discovery. Digit Discov. 2023;2(2):327-38.

Yang L, Yang L, Fan G, Ma W, Zheng S, Mavromatis V, et al. Machine learning-enabled high-throughput material screening for MOF-based membrane gas separation. Digit Discov. 2023;2(2):415-27.

Xia W, Song J, Chen H, Han Z, Zhao G, Zhang J, et al. Accelerating materials discovery using integrated deep machine learning approaches. J Mater ChemA. 2023;11(48):25973-82. https://doi.org/10.1039/D3TA03771A.

Abolhasani M, Kumara KSRN. Autonomous materials synthesis by machine learning and robotics. Nat Mach Intell. 2023;5(1):4-6.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55. https://doi.org/10.1038/s41586-018-0337-2.