

ORIGINAL RESEARCH

Open access

Search without Coverage: Exploration Blind Spots in AI-Guided Materials Discovery

Claire Martin^{1*}, Julien Robert², Sophie Bernard¹

Abstract

The advent of computational and data-driven materials engineering has transformed the landscape of materials discovery, leveraging machine learning algorithms and high-throughput simulations to accelerate the identification of novel compounds and properties. Within this paradigm, AI-guided systems integrate representation learning, graph neural networks, and uncertainty quantification to navigate vast chemical spaces, yet persistent exploration blind spots arise from incomplete coverage in data infrastructures and model architectures. These blind spots manifest as epistemic gaps where AI-driven searches fail to probe underrepresented regions of materials possibility spaces, potentially overlooking breakthrough innovations. This manuscript introduces the Coverage Dynamics Framework (CDF), a conceptual lens that dissects the interplay between data modalities, representational embeddings, and discovery steering logics to illuminate these blind spots. By framing exploration as a dynamic interplay of coverage vectors and feedback mechanisms, the CDF highlights systemic trade-offs in AI-guided pipelines, such as the tension between exploitation of known datasets and exploration of sparse domains. Implications extend to enhancing autonomous discovery systems, fostering multimodal data integration, and refining uncertainty-aware workflows in materials informatics. Ultimately, this framework advocates for infrastructure-level interventions to mitigate blind spots, promoting more comprehensive and resilient AI-assisted materials engineering ecosystems.

Keywords Materials informatics, Uncertainty quantification, AI-guided materials discovery, Representation learning, Data-driven materials engineering, Exploration blind spots

*Correspondence:

Claire Martin

claire.martin@outlook.com

¹ Department of Computational Materials Research, Faculty of Engineering, University of Lyon, Lyon, France

² Department of Materials Data Analytics, Faculty of Engineering, University of Strasbourg, Strasbourg, France

Introduction

The field of computational and data-driven materials engineering has emerged as a cornerstone of modern materials science, driven by the exponential growth in computational power and the proliferation of large-scale datasets. This paradigm shift began with the integration of high-throughput density functional theory calculations, enabling the systematic screening of thousands of candidate materials for targeted properties such as electronic bandgaps, mechanical strength, or catalytic efficiency [1, 2]. As computational infrastructures evolved, materials informatics formalized the application of data mining and statistical methods to extract patterns from these datasets, facilitating predictive modeling without exhaustive experimental validation [3, 4]. The role of AI and machine learning has been pivotal, transforming passive data analysis into active discovery engines that optimize design cycles and reduce time-to-market for advanced materials in sectors like energy storage, semiconductors, and biomaterials [5, 6].

Central to this ecosystem are data-driven approaches that harness multimodal datasets, encompassing structural, compositional, and functional attributes derived from simulations, experiments, and literature mining [7, 8]. High-throughput computation, exemplified by databases like the Open Quantum Materials Database, has democratized access to precomputed properties, allowing researchers to query vast repositories for inverse design tasks where desired functionalities dictate material compositions [9]. Machine learning models, particularly deep learning architectures, have amplified this capability by learning hierarchical representations from atomic-scale features, enabling predictions across diverse material classes without reliance on explicit physical models [10, 11]. Graph neural networks have proven especially adept at capturing topological relationships in crystal structures, outperforming traditional descriptors in tasks like property prediction and phase stability assessment [12, 13].

Yet, despite these advances, the current discovery models exhibit inherent limitations rooted in epistemic and computational constraints. Exploration in AI-guided systems often prioritizes densely populated regions of chemical space, where abundant data supports robust model training, leading to a form of confirmation bias that reinforces known paradigms while neglecting sparse or anomalous domains [14, 15]. This manifests in blind spots—areas of potential innovation obscured by incomplete data coverage, biased representations, or inadequate uncertainty handling. For instance, representation learning architectures may embed materials in low-dimensional manifolds that inadvertently compress variability in underrepresented classes, such as metastable phases or hybrid organic-inorganic compounds [16, 17]. High-throughput infrastructures, while scalable, frequently overlook the coupling between simulation fidelity and experimental realities, resulting in discovery pipelines that propagate errors from idealized models to real-world applications [18, 19].

Uncertainty quantification emerges as a critical safeguard, yet its integration remains uneven, often limited to aleatoric noise rather than epistemic gaps arising from dataset imbalances or model extrapolation [5, 20]. Autonomous discovery systems, including closed-loop experimentation platforms, promise self-correcting workflows but struggle with steering logics that fail to adapt to exploration frontiers, where data scarcity amplifies risk [21, 22]. These constraints are compounded by the siloed nature of materials data ecosystems, where interoperability between simulation-derived and experiment-sourced modalities hinders comprehensive coverage [23, 24]. Epistemic challenges further arise from the black-box nature of deep learning models, which obscure interpretability and limit human oversight in guiding explorations toward uncharted territories [25, 26].

In response, this manuscript positions a novel conceptual framework to address these blind spots, emphasizing the dynamics of search coverage in AI-guided materials discovery. By conceptualizing exploration as an interplay of representational, data, and inference layers, the framework elucidates systemic mechanisms for enhancing coverage without empirical interventions. This approach integrates insights from materials informatics and computational design, offering a pathway to more equitable navigation of materials spaces.

Theoretical Background & Literature Synthesis

Materials data infrastructures

The foundation of data-driven materials engineering lies in robust infrastructures that aggregate, curate, and disseminate multimodal datasets essential for AI applications. These infrastructures have evolved from isolated repositories to interconnected ecosystems, enabling high-throughput access to properties computed via density functional theory or extracted from experimental archives [1, 9]. Key advancements include toolkits like Matminer, which facilitate data mining and feature engineering for inorganic materials, bridging raw data to machine-learnable formats [16]. Multimodal integration has been emphasized, incorporating textual knowledge from scientific literature through unsupervised embeddings that uncover latent relationships in materials synthesis parameters [7, 8]. However, challenges persist in ensuring data quality and diversity; datasets often exhibit biases toward stable, high-symmetry structures, limiting generalizability to novel compositions [14, 24]. This underscores the need for infrastructures that prioritize coverage across chemical spaces, including sparse regions prone to discovery blind spots.

Representation learning architectures

Representation learning forms the computational backbone of materials AI, transforming atomic and molecular descriptors into embeddings suitable for downstream tasks. Early efforts focused on stoichiometry-based models that predict properties without explicit structural knowledge, leveraging deep neural networks to infer chemistry from elemental compositions [10, 11]. Graph neural networks have advanced this by encoding crystal graphs as relational structures, capturing local environments and long-range interactions for accurate property forecasting [12, 13]. Innovations in transfer learning have enabled cross-property predictions, where models trained on abundant data adapt to scarce domains through shared representations [15, 23]. Despite these strides, architectures often embed materials in manifolds that favor dense clusters, potentially masking underrepresented subspaces and creating exploration voids [17, 27]. Foundation models for science further amplify this by scaling to vast datasets, yet they risk entrenching biases if training corpora lack diversity [25].

AI-guided discovery

Systems AI-guided systems orchestrate discovery by coupling predictive models with optimization algorithms, steering searches toward optimal materials candidates. Active learning strategies have been instrumental, using uncertainty estimates to select informative samples for iterative refinement [5, 20]. Closed-loop experimentation integrates robotics and real-time feedback, accelerating synthesis and characterization cycles [19, 21, 22]. Autonomous platforms exemplify this, employing Bayesian frameworks for adaptive sampling in thin-film materials discovery [28]. However, these systems often operate within constrained search spaces, where initial biases propagate through feedback loops, leading to suboptimal coverage of peripheral regions [6, 29]. The challenge lies in balancing exploitation of high-confidence predictions with exploration of uncertain frontiers, a dynamic that current paradigms address unevenly [15].

Computational design

Paradigms Inverse design paradigms invert traditional workflows, starting from desired properties to generate candidate structures via generative models or optimization routines [2, 6, 30]. High-throughput computation supports this by screening virtual libraries, while machine learning accelerates evaluations through surrogate models [4, 18]. Deep learning frameworks facilitate design space exploration, incorporating data augmentation to simulate variability [15]. Yet, these paradigms encounter limits in handling multi-objective trade-offs, where conflicting properties (e.g., stability versus reactivity) create design blind spots [26]. Coupling simulations with experiments remains a bottleneck, as discrepancies in fidelity introduce epistemic uncertainties that distort design outcomes [19]. Emerging approaches emphasize hierarchical design, layering atomic-scale simulations with macroscopic predictions to enhance comprehensiveness [3].

Uncertainty & interpretability

Uncertainty quantification is vital for trustworthy AI in materials, distinguishing between model confidence and inherent variability [5, 14]. Bayesian deep learning and ensemble methods provide probabilistic outputs, guiding decisions in sparse data regimes [20]. Interpretability complements this by elucidating model decisions, such as through attention mechanisms in graph networks that highlight influential atomic features [12, 13]. However, interpretability tools often fall short in complex architectures, where black-box behaviors obscure the roots of prediction failures [25, 26]. In discovery contexts, unaddressed uncertainties amplify blind spots, particularly in extrapolative scenarios beyond training distributions [23, 31]. Literature synthesis reveals a consensus on the need for integrated uncertainty-aware infrastructures, yet gaps remain in scaling these to multimodal, high-dimensional settings [18, 32].

Proposed conceptual framework

To address the exploration blind spots identified in AI-guided materials discovery, we introduce the Coverage Dynamics Framework (CDF), an original systems-level construct that models search processes as interdependent layers of data ingestion, representational mapping, and inference steering. The CDF conceptualizes discovery not as a linear pipeline

but as a dynamic network where coverage—defined as the extent of probed possibility space—emerges from interactions across these layers. At its core, the framework posits three structural layers: the Data Coverage Layer (DCL), which handles multimodal input modalities and their interoperability; the Representation Coverage Layer (RCL), which embeds materials features into navigable manifolds; and the Discovery Steering Layer (DSL), which orchestrates feedback loops for adaptive exploration. These layers interconnect via coverage vectors, abstract entities representing the density and diversity of information flow.

In the DCL, data infrastructures are viewed as filters that modulate input coverage, where multimodal fusion (e.g., simulation-experiment coupling) expands accessible domains but risks dilution if modalities conflict. The RCL transforms these inputs into embeddings, emphasizing the trade-off between dimensionality reduction and information preservation; over-compression can create representational voids, analogous to unexplored subspaces. The DSL integrates outputs from prior layers to guide searches, employing steering logics that balance local optimization with global probing. Feedback loops within the CDF recirculate insights, allowing the system to self-adjust coverage deficits, such as by prioritizing uncertainty-driven queries in sparse regions. This structure illuminates blind spots as emergent from layer misalignments, for instance, when RCL embeddings fail to capture DCL diversity, leading DSL to reinforce biased paths.

The CDF can be visualized as a networked diagram with nodes for each layer and edges denoting coverage vectors, as conceptualized in **Figure 1**.

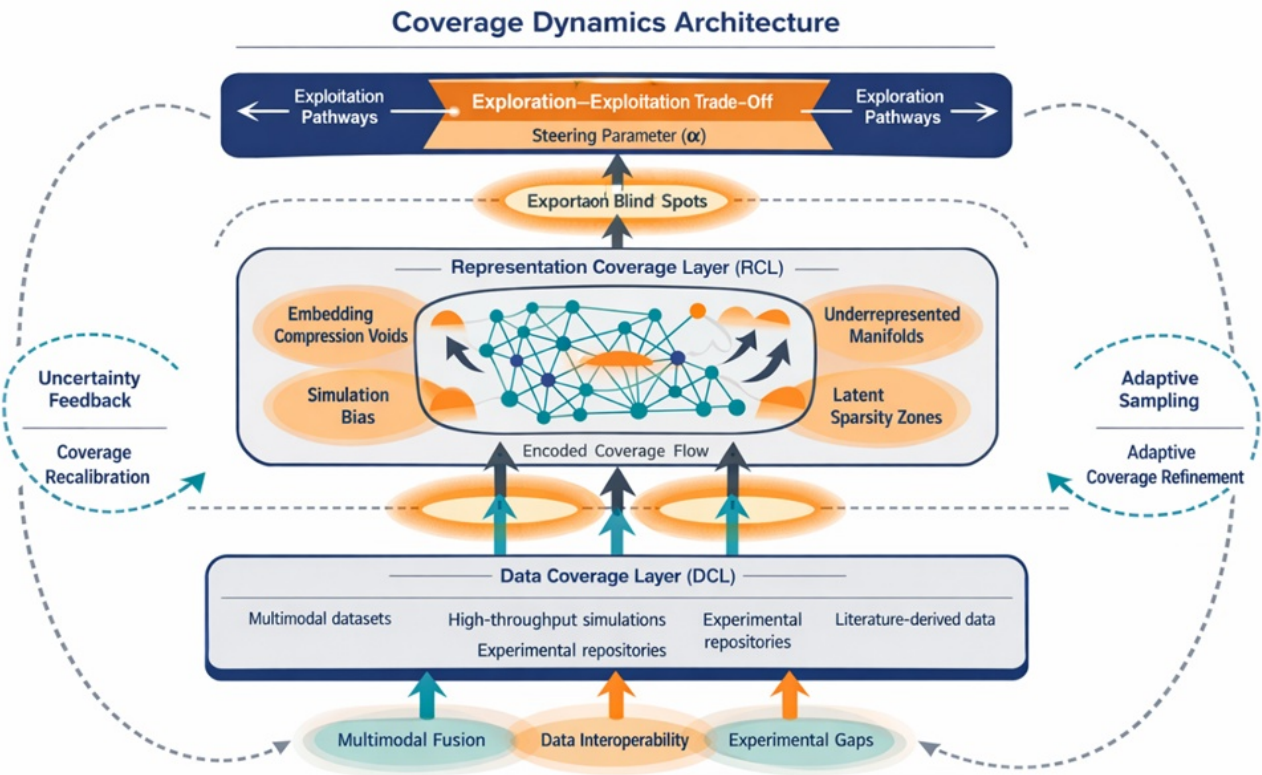


Figure 1. Coverage Dynamics Framework (CDF): Layered Architecture of Exploration Coverage in AI-Guided Materials Discovery

Conceptual architecture of the Coverage Dynamics Framework (CDF) illustrating exploration coverage across three interdependent layers: the Data Coverage Layer (DCL), Representation Coverage Layer (RCL), and Discovery Steering Layer (DSL). Coverage vectors propagate upward from multimodal data infrastructures into embedding manifolds and

decision systems, while adaptive feedback loops recalibrate exploration priorities. Shaded regions denote systemic blind spots emerging at layer interfaces, including data sparsity, representational compression voids, and steering trade-off tensions. The framework visualizes discovery as a dynamic coverage network rather than a linear search pipeline. The structural roles and coverage determinants of each CDF layer are summarized in **Table 1**.

Table 1. Structural Layers of the Coverage Dynamics Framework and Their Coverage Functions

CDF Layer	Core Function	Coverage Determinants	Blind Spot Risks	Mitigation Logics
Data Coverage Layer (DCL)	Aggregates multimodal materials datasets	Dataset diversity, modality interoperability, simulation–experiment coupling	Sparse chemical domains, experimental underrepresentation, simulation bias	Coverage-aware database curation, multimodal fusion, targeted data acquisition
Representation Coverage Layer (RCL)	Embeds materials into latent manifolds	Embedding dimensionality, feature compression, topology learning	Manifold sparsity, compression-induced information loss, cluster bias	Adaptive embedding scaling, sparsity-aware training, uncertainty-weighted mapping
Discovery Steering Layer (DSL)	Guides exploration and optimization	Active learning policies, steering parameters, search heuristics	Over-exploitation, constrained search spaces, novelty suppression	Exploration-weighted steering, uncertainty sampling, adaptive optimization
Feedback Coupling System	Recirculates insights across layers	Coverage diagnostics, uncertainty propagation	Reinforced bias loops, delayed recalibration	Iterative validation loops, coverage recalibration engines

To formalize key dynamics, the interaction between data diversity and representational fidelity in mitigating blind spots may be expressed as: $C = \int_D R(d) \cdot U(d) dd$ where C denotes overall coverage, D the data domain, R(d) the representational mapping function for datum d, and U(d) an uncertainty weighting that amplifies contributions from underrepresented d. This integral captures the accumulation of coverage across the data space, emphasizing how uncertainty-aware representations enhance exploration breadth.

Further, the trade-off in steering logics between exploitation and exploration can be conceptualized as:

$$S = \frac{\alpha E}{1 + (1 - \alpha) X} \tag{1}$$

with S as steering output, E exploitation utility (e.g., confidence in known regions), X exploration potential (e.g., novelty in sparse domains), and α \alpha a dynamic parameter modulated by feedback from prior layers. This expression highlights the tunable balance required to avoid blind spots.

Finally, feedback loop efficacy in the CDF may be captured as:
$$F = \sum_{i=1}^n \beta_i \Delta C_i$$
 where F is feedback strength, β_i layer-specific coefficients, and ΔC_i coverage increments from iteration i , underscoring iterative improvements in systemic coverage. These formulas provide interpretive tools for understanding CDF dynamics, without implying empirical validation.

Analytical implications

The Coverage Dynamics Framework (CDF) offers a lens for dissecting the systemic implications of exploration blind spots in AI-guided materials discovery, revealing trade-offs and interactions that inform computational workflows. By framing coverage as an emergent property of layered dynamics, the CDF highlights how misalignments in data ingestion and representational mapping can propagate through discovery pipelines, leading to inefficient resource allocation in high-throughput systems [1, 4]. For instance, in materials informatics, where multimodal datasets drive model training, the DCL's emphasis on interoperability implies that fragmented infrastructures exacerbate blind spots by restricting the flow of diverse modalities, such as coupling high-fidelity simulations with sparse experimental validations [7, 19]. This suggests a need for steering logics in the DSL that prioritize coverage expansion over precision in densely sampled regions, potentially reallocating computational efforts toward epistemic frontiers [5, 20].

Representationally, the RCL's role in embedding materials features underscores implications for graph neural networks and deep learning architectures, where manifold structures may inadvertently prioritize topological similarities at the expense of outlier detection [12, 13]. Analytical insights from the CDF indicate that enhancing coverage requires dynamic adjustments to embedding dimensions, balancing compression for efficiency against preservation of variability in underrepresented classes like amorphous or defective structures [10, 17]. In inverse design paradigms, this translates to refined workflows where generative models incorporate coverage vectors to sample beyond local optima, mitigating the risk of design stagnation [6, 30]. Uncertainty quantification integrates here as a modulator, where epistemic uncertainties signal blind spots, prompting feedback loops to recalibrate searches [14, 18].

At the systems level, the CDF illuminates infrastructure trade-offs in autonomous discovery, such as the tension between closed-loop efficiency and comprehensive exploration [21, 22]. Feedback mechanisms within the framework imply that iterative refinements can amplify coverage if tuned to detect layer-specific deficits, fostering resilient ecosystems that adapt to evolving data landscapes [23, 24]. This has broader implications for foundation models in science, where scaling data volumes must be coupled with coverage-aware training to avoid entrenching biases [8, 25].

The coverage accumulation dynamic in the CDF can be further conceptualized as a vectorized interaction:

$$\vec{C} = R \cdot \vec{D} + \lambda \vec{U}$$

where $\vec{C}$ is the coverage vector, R the representation matrix transforming data vector $\vec{D}$, and $\lambda \vec{U}$ a scaled uncertainty vector that perturbs embeddings to probe blind spots. This expression captures the additive role of uncertainty in expanding coverage, providing an interpretive tool for workflow optimization.

Similarly, the feedback-induced adaptation may be expressed as:

$$\Delta S = \gamma \left(\vec{C}_{t+1} - \vec{C}_t \right)$$

with ΔS denoting steering shift, γ an adaptation rate, and $\vec{C}_t$ coverage at time t , illustrating how incremental coverage gains drive discovery redirection. These formulations underscore the CDF's utility in guiding computational steering without prescriptive metrics.

Results and Discussion

The Coverage Dynamics Framework (CDF) advances contemporary understanding of exploration blind spots by integrating disparate yet interdependent elements of computational materials engineering into a unified systems architecture. Existing literature has largely treated data scarcity, representation bias, and uncertainty modeling as separable technical constraints; however, the CDF reframes these not as isolated deficiencies but as emergent properties of coupled infrastructures operating across discovery pipelines [3, 15]. By situating blind spots within relational coverage dynamics rather than static dataset limitations, the framework surfaces structural discontinuities that remain obscured in high-throughput paradigms optimized primarily for volumetric expansion rather than epistemic balance [16, 26].

This repositioning has significant theoretical implications. High-throughput ecosystems often presume quasi-uniform sampling of chemical and structural spaces, implicitly equating scale with comprehensiveness. The CDF challenges this equivalence by demonstrating that data abundance can coexist with representational sparsity when coverage vectors cluster around historically studied material families—such as transition-metal oxides or conventional battery chemistries—while leaving peripheral domains underrepresented [9, 31]. Consequently, predictive confidence may be artificially inflated in densely sampled zones while remaining systematically fragile at the exploratory frontier. This reframing calls for a shift from throughput-centric evaluation metrics toward coverage-sensitive discovery diagnostics.

From an infrastructural standpoint, the framework introduces actionable conceptual logics for next-generation materials database design. Rather than curating repositories solely through volumetric aggregation, coverage-aware architectures would embed vectorized mapping of compositional, structural, and thermodynamic search spaces. Such infrastructures could algorithmically identify lacunae—regions where simulation density, experimental validation, and representational fidelity are simultaneously low—thereby guiding targeted data acquisition strategies [2, 11]. In practical discovery environments, this may enhance robustness in property prediction for chemically exotic alloys, metastable phases, or low-symmetry perovskite derivatives that lie beyond dominant sampling regimes.

The CDF further contributes interpretive clarity to the coupling of simulation and experimental ecosystems. Closed-loop discovery platforms frequently assume fidelity continuity between computational predictions and laboratory validation; however, blind spots often emerge precisely where this continuity fractures. For instance, simulation datasets may privilege equilibrium structures under idealized thermodynamic assumptions, while experimental synthesis operates under kinetic constraints and defect-mediated realities. The layered architecture of the CDF—particularly the interaction between the Data Coverage Layer (DCL) and Representation Coverage Layer (RCL)—highlights how such fidelity mismatches propagate into steering decisions, amplifying exploratory risk [18, 19].

By proposing interleaved validation loops, the framework extends beyond conventional active learning cycles. Traditional active learning prioritizes uncertainty reduction within existing representation spaces; the CDF instead incorporates steering logics that evaluate whether the representation space itself is coverage-deficient. This meta-exploratory orientation enables dynamic recalibration of discovery priorities, where exploration is not merely directed toward uncertain predictions but toward structurally under-mapped domains. Such steering could reduce computational overhead in resource-constrained infrastructures by avoiding redundant sampling in already saturated regions [5, 20].

Nevertheless, the conceptual nature of the framework introduces inherent limitations. As an interpretive systems model, the CDF does not prescribe algorithmic implementations or quantitative coverage thresholds. Its operationalization depends on domain expertise to define meaningful coverage axes—whether compositional gradients, crystallographic topologies, or processing conditions—and to adjudicate trade-offs between exploratory breadth and predictive depth [14, 24]. Furthermore, while the framework diagnoses structural blind spots, it does not independently resolve the epistemic biases embedded within legacy datasets, such as functional approximations in density functional theory or publication biases favoring high-performance materials.

Future theoretical expansions may explore multi-scale coverage integrations, linking atomistic representations with mesostructural morphologies and macroscopic performance envelopes. Such vertical coupling could illuminate how blind

spots propagate across scale transitions—for example, when nanoscale defect physics remains underrepresented in bulk property optimization models [17, 32]. However, increasing scale dimensionality also complicates coverage mapping, potentially generating combinatorial sparsity that challenges visualization and steering infrastructures.

Beyond technical ecosystems, the CDF carries broader epistemological and ethical implications. By foregrounding coverage asymmetries, the framework intersects with emerging discourses in AI ethics and equitable innovation. Discovery systems that over-optimize abundant datasets risk perpetuating material monocultures—prioritizing well-funded application domains such as energy storage while neglecting materials relevant to low-resource or sustainability-critical contexts [4, 8]. Embedding coverage governance within discovery pipelines may therefore promote more socially distributed innovation trajectories.

In autonomous research platforms, these insights translate into safeguards against over-optimization. Closed-loop laboratories trained on narrow discovery priors may converge prematurely on locally optimal materials families, constraining innovation diversity. Coverage-aware steering logics could counteract such convergence by injecting exploratory perturbations into underrepresented design regions, fostering resilient and pluralistic discovery ecosystems [21, 22].

Collectively, the CDF reframes exploration management from a reactive process—addressing blind spots post hoc—to a proactive infrastructural design principle. By embedding coverage diagnostics across data ingestion, representation learning, and discovery steering, the framework enriches the conceptual discourse on sustainable AI-guided materials engineering and positions exploration equity as a core systems objective rather than a peripheral corrective [23, 25].

Conclusion

In summary, exploration blind spots constitute a foundational constraint within AI-guided materials discovery, limiting the epistemic reach and innovative capacity of computational design ecosystems. These blind spots do not arise solely from insufficient data volume but from uneven coverage distributions spanning datasets, representation architectures, and steering mechanisms. As discovery infrastructures scale, the risk of such asymmetries intensifying—rather than dissipating—becomes increasingly pronounced.

The Coverage Dynamics Framework offers a novel conceptual instrument to interpret and address these systemic discontinuities. By structuring exploration across layered coverage domains—data, representation, and steering—the framework elucidates how blind spots emerge through feedback misalignments and fidelity discontinuities. This layered systems perspective moves beyond reductionist diagnostics, enabling a holistic understanding of discovery robustness [1, 3].

Importantly, the CDF formalizes exploration trade-offs through symbolic expressions that articulate the balance between coverage breadth, representational depth, and steering precision. These abstractions provide conceptual scaffolding for optimizing discovery workflows without prescribing rigid algorithmic pathways. As such, the framework remains adaptable across diverse materials informatics contexts, including inverse design, generative modeling, and autonomous experimentation platforms [6, 12].

The infrastructural implications are substantial. Coverage-aware database curation, representation architectures sensitive to sparsity gradients, and steering systems calibrated to exploratory risk could collectively enhance the resilience of AI-driven discovery. Such infrastructures would not only accelerate materials innovation but also broaden its epistemic inclusivity—ensuring that emergent materials spaces are explored with intentional comprehensiveness rather than incidental accessibility.

Looking forward, embedding coverage governance into discovery pipelines may become as critical as accuracy optimization or computational efficiency. As AI systems assume greater autonomy in scientific exploration, frameworks

like the CDF will be essential for safeguarding against systemic blind spots that constrain innovation horizons.

Ultimately, the Coverage Dynamics Framework advocates for a paradigm shift: from discovery acceleration alone toward discovery completeness. By foregrounding exploration coverage as a first-class infrastructural concern, it lays conceptual groundwork for more robust, equitable, and forward-looking materials engineering ecosystems—capable not only of discovering faster, but of discovering wiser.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 03 Oct 2022

Revised: 23 Feb 2023

Accepted: 20 Apr 2023

Published online: 18 September 2023

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-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. *npj Comput Mater.* 2017;3:54.
<https://doi.org/10.1038/s41524-017-0056-5>.
- Balachandran PV, Young J, Lookman T, Rondinelli JM. Learning from data to design functional materials without inversion symmetry. *Nat Commun.* 2017;8:14282.
<https://doi.org/10.1038/ncomms14282>.

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>.

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>.

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput Mater*. 2019;5:21.
<https://doi.org/10.1038/s41524-019-0153-8>.

Tagade PM, Adiga SP, Pandian S, Park MS, Hariharan KS, Kolake SM. Attribute driven inverse materials design using deep learning Bayesian framework. *npj Comput Mater*. 2019;5:127.
<https://doi.org/10.1038/s41524-019-0263-3>.

Kim E, Tomala A, Huang K, Strubell E, Matthews S, Saunders A, et al. Machine-learned and codified synthesis parameters of oxide materials. *Sci Data*. 2017;4:170127.
<https://doi.org/10.1038/sdata.2017.127>.

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>.

Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM*. 2013;65(11):1501–9.
<https://doi.org/10.1007/s11837-013-0755-4>.

Goodall REA, Lee AA. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat Commun*. 2020;11:6280.
<https://doi.org/10.1038/s41467-020-19964-7>.

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;8:17593.
<https://doi.org/10.1038/s41598-018-35934-y>.

Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. *Commun Mater*. 2022;3:93.
<https://doi.org/10.1038/s43246-022-00315-6>.

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett*. 2018;120:145301.

Sutton C, Boley M, Ghiringhelli LM, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. *Nat Commun*. 2020;11:4428.
<https://doi.org/10.1038/s41467-020-17112-9>.

Kim Y, Kim Y, Yang C, Park K, Gu GX, Ryu S. Deep learning framework for material design space exploration using active transfer learning and data augmentation. *npj Comput Mater*. 2021;7:140.
<https://doi.org/10.1038/s41524-021-00609-2>.

Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput Mater*. 2016;2:16028.

Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci.* 2022;2:718-28.
<https://doi.org/10.1038/s43588-022-00349-3>.

Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A critical review of machine learning for energy materials. *Adv Energy Mater.* 2020;10(8):1903242.
<https://doi.org/10.1002/aenm.201903242>.

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

Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integr Mater Manuf Innov.* 2017;6:207-17.

Häse F, Roch LM, Aspuru-Guzik A. Chimera: Enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem Sci.* 2018;9:7642-55.
<https://doi.org/10.1039/C8SC02239A>.

Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature.* 2023;624(7990):86–91.

Gupta V, Choudhary K, Tavazza F, Campbell C, Liao WK, Choudhary A, et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat Commun.* 2021;12:6595.
<https://doi.org/10.1038/s41467-021-26921-5>.

Choudhary K, Garrity KF, Reid ACE, DeCost B, Biacchi AJ, Hight Walker AR, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput Mater.* 2020;6:1-13.

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>.

Aykol M, Herring P, Anapolsky A. Perspective—combining physics and machine learning to tackle grand challenges in energy storage. *J Electrochem Soc.* 2020;167:060507.

Batra R, Tran HD, Kim C, Chapman J, Chandrasekaran A, Ramprasad R. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput Mater.* 2020;6:141.

Balachandran PV. Machine learning guided design of functional materials with targeted properties. *Comput Mater Sci.* 2019;164:82-90.

Kadulkar S, Sherman ZM, Ganesan V, Truskett TM. Machine learning—assisted design of material properties. *Annu Rev Chem Biomol Eng.* 2022;13(2022):235-54.

Mahjoub R, Laws KJ, Ferry M. Amorphous phase stability and the interplay between electronic structure and topology. *Acta Mater.* 2017;131:131-40.

Wang AY-T, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J. Machine learning for materials scientists: An introductory guide toward best practices. *Chem Mater.* 2020;32(12):4954-65.
<https://doi.org/10.1021/acs.chemmater.0c01907>.

Varley JB, Miglio A, Ha VA, van Setten MJ, Rignanese GM, Hautier G. High-throughput design of non-oxide p-type transparent conducting materials: Data mining, search strategy, and identification of boron phosphide. *Chem Mater.* 2017;29(6):2568-73.