ORIGINAL RESEARCH **Open access**

Data Density as Discovery Bias: Uneven Sampling in Computational Materials Exploration

Fatima Zahra Amrani[1*], Youssef Benali[1]

Abstract

Computational materials engineering has evolved into a data-intensive ecosystem in which high-throughput screening, machine learning surrogates, and autonomous discovery pipelines increasingly dictate the pace and direction of materials innovation. Within this paradigm, the spatial and compositional density of available computational data emerges as a previously under-examined source of systematic bias. Uneven sampling—whether arising from historical focus on well-studied chemical spaces, computational cost gradients, or architectural preferences of representation-learning models—creates “data deserts” that skew downstream inference, limit generalization of generative architectures, and constrain the reach of closed-loop optimization. The present conceptual analysis identifies data density as an epistemic rather than merely statistical bias and introduces the Density-Compensated Epistemic Exploration Framework (DCEF) as an interpretive infrastructure that reframes discovery pipelines through explicit density mapping, bias quantification, and adaptive steering logics. By integrating representation-learning dynamics with uncertainty propagation and feedback-driven re-sampling, the DCEF offers a systems-level lens for diagnosing and mitigating discovery bias without invoking empirical validation or performance metrics. Its implications extend to the design of next-generation materials data infrastructures, the governance of foundation models for science, and the epistemic transparency of simulation–experiment coupling.

Keywords Materials informatics, Representation learning, Computational discovery, Data density bias, Uneven sampling, Epistemic infrastructure

*Correspondence:
Fatima Zahra Amrani
fatima.amrani@gmail.com
[1] Department of Computational Materials Science, Faculty of Sciences and Engineering, Mohammed V University, Rabat, Morocco

Introduction

The past decade has witnessed the maturation of computational materials engineering from isolated density-functional-theory calculations into a densely interconnected ecosystem of high-throughput infrastructures, machine-learning surrogates, and closed-loop discovery platforms [1-3]. Materials informatics has transitioned from descriptor-based regression to graph-neural-network architectures capable of learning latent representations directly from atomic graphs or stoichiometric fingerprints [4-6]. Generative models now propose candidate compositions or structures with minimal human supervision [7, 8], while active-learning loops iteratively refine surrogate models by querying regions of high predictive uncertainty [9, 10]. These developments have compressed research timelines, expanded accessible chemical spaces, and enabled inverse design objectives that were previously intractable.

Yet this acceleration has exposed a structural limitation that remains largely unarticulated: the non-uniform density of the underlying computational data corpus. High-throughput campaigns have historically prioritized systems with established databases, modest unit-cell sizes, or favorable convergence properties, producing dense clusters of data points around conventional alloys, perovskites, and metal–organic frameworks while leaving vast regions of

composition–structure space sparsely sampled [11-13]. Representation-learning models trained on such corpora inherit these density gradients; graph convolutions and equivariant transformers amplify local correlations while under-representing extrapolation regimes [6, 14]. Consequently, generative proposals and uncertainty estimates become systematically biased toward data-rich subspaces, creating a self-reinforcing cycle wherein “easy” materials dominate published literature and autonomous platforms.

Current discovery models largely treat data sparsity as a technical nuisance to be addressed through imputation or transfer learning rather than as an epistemic constraint intrinsic to the sampling process itself [3, 9, 15]. Active-learning strategies, for instance, typically optimize acquisition functions under the assumption that the feature space is homogeneously explorable, ignoring that acquisition cost scales non-linearly with local data density [9, 10]. Multi-fidelity hierarchies similarly propagate uncertainty without explicit accounting for the underlying density landscape [3]. The result is a latent discovery bias that distorts the perceived Pareto front of materials properties and limits the diversity of experimentally actionable candidates.

The conceptual gap lies in the absence of a unified interpretive framework that positions data density as a first-order driver of discovery bias rather than a second-order statistical artifact. Such a framework must operate at the infrastructure level, bridging representation learning, uncertainty quantification, and pipeline steering without reducing the problem to benchmark metrics or empirical retraining protocols. The present work addresses this gap by introducing the Density-Compensated Epistemic Exploration Framework (DCEF). The DCEF conceptualizes discovery pipelines as density-modulated dynamical systems in which bias arises from the interaction between sampling density, representation geometry, and inference uncertainty. It provides analytical lenses for tracing bias propagation across layers and identifies steering logics that restore epistemic balance through compensatory re-sampling and representation re-weighting. By reframing uneven sampling as an epistemic rather than merely computational phenomenon, the DCEF offers a systems-level vocabulary for redesigning materials data infrastructures and governing the deployment of generative models in autonomous discovery.

Theoretical Background & Literature Synthesis

Materials data infrastructures

Contemporary materials data infrastructures function as the epistemic substrate upon which all computational discovery workflows are constructed. These infrastructures aggregate outputs from high-throughput density functional theory (DFT) campaigns, curated experimental repositories, combinatorial synthesis datasets, and increasingly, multimodal characterization archives integrating spectroscopy, microscopy, and process metadata [8, 11, 16]. The architectural expansion of such infrastructures has enabled unprecedented coverage of crystallographic prototypes and compositional permutations, yet this apparent scale masks profound structural unevenness. Density heterogeneity emerges not merely as a statistical artifact but as a systemic property of how materials knowledge is generated.

Metal–organic frameworks (MOFs), for instance, benefit from algorithmic enumeration strategies that systematically recombine linker and node chemistries, producing vast yet structurally coherent design spaces [8, 11]. In contrast, high-entropy alloys (HEAs) and compositionally complex ceramics remain sparsely sampled because their configurational degrees of freedom expand combinatorially, rendering exhaustive enumeration computationally prohibitive [12, 17, 18]. This asymmetry is further amplified by thermodynamic metastability, where many theoretically plausible configurations fail to converge in simulation workflows, leading to selective archival of only stable or near-stable structures.

Density gradients are also shaped by socio-technical forces embedded in the scientific publication ecosystem. Computational cost barriers privilege simulations involving smaller unit cells, lighter elements, or higher symmetry lattices, thereby skewing representation toward energetically tractable systems [3, 19]. Publication bias reinforces this skew, as successful convergence, novel functionality, or record-breaking properties are more likely to be reported and subsequently ingested into databases. Failed calculations, unstable phases, or negative results rarely achieve infrastructural visibility. The resulting corpus is therefore not a neutral sampling of chemical space but a historically contingent cartography whose peaks and voids encode decades of methodological preference, funding priorities, and computational feasibility. All downstream

machine learning inference is consequently conditioned on this inherited topography.

Representation learning architectures

Representation learning architectures translate infrastructural density into computable structure. Graph neural networks (GNNs), message-passing networks, and symmetry-equivariant transformers have emerged as dominant paradigms because they encode atomic connectivity, geometric periodicity, and rotational invariances directly into model structure [4, 6, 20]. By embedding physical symmetries into the learning process, these architectures achieve superior scaling behavior and predictive performance across large crystallographic datasets.

Yet their epistemic grounding remains inseparable from the density contours of the training corpus. Message-passing operations aggregate local neighborhood statistics, meaning that frequently observed coordination motifs—such as octahedral metal–oxygen environments or tetrahedral covalent networks—become statistically reinforced within latent space geometry [6, 14]. Rather than learning universal chemical rules, models learn weighted statistical regularities conditioned on sampling frequency.

In low-density regimes, this statistical grounding weakens. Sparse coordination chemistries or rare compositional motifs produce embeddings characterized by inflated variance and unstable neighborhood relations. Generative architectures operating atop these embeddings confront a bifurcation: either collapse toward dense latent clusters where reconstruction loss is minimized, or extrapolate into chemically implausible regions unsupported by training evidence [4, 7]. Diffusion and autoregressive generators partially alleviate this by smoothing latent transitions, yet they remain bounded by the probability mass of the original corpus.

Self-supervised pre-training strategies—such as masked stoichiometry prediction or contrastive learning on crystal graphs—extend representational coverage by leveraging unlabeled structures [14]. However, pre-training redistributes attention rather than neutralizing density bias. The foundational signal remains anchored in infrastructural sampling, ensuring that representational geometry continues to mirror historical data concentration.

AI-Guided discovery systems

AI-guided discovery systems operationalize learned representations into iterative search mechanisms. Active learning frameworks, evolutionary algorithms, Bayesian optimization loops, and generative adversarial networks collectively form the algorithmic core of autonomous materials exploration pipelines [7, 9, 10, 21]. These systems are designed to optimize acquisition efficiency by prioritizing candidates expected to yield maximal information gain or property improvement.

Acquisition functions in active learning are typically formulated around predictive variance, entropy, or expected improvement metrics [9]. Implicit within these formulations is an assumption of uniform navigability across feature space. When density gradients are introduced, this assumption collapses. High-uncertainty candidates tend to cluster along the peripheries of dense regions, where models possess partial contextual grounding, rather than deep within sparse interiors where uncertainty becomes epistemically unbounded [9, 10]. Consequently, exploration trajectories trace the edges of known domains instead of penetrating uncharted compositional regimes.

Genetic algorithms and reinforcement learning controllers exhibit analogous tendencies. Fitness landscapes inferred from dense data guide mutation and recombination toward historically sampled chemistries. Generative adversarial networks (GANs), widely deployed for inverse design, inherit mode-covering biases that replicate the density distribution of training datasets [7]. Even when diversity regularizers are introduced, generated candidates disproportionately occupy latent basins corresponding to well-represented structural families. Autonomous discovery loops thus become self-reinforcing epistemic circuits, amplifying infrastructural imbalances rather than correcting them.

Computational design paradigms

Inverse design, phase stability prediction, and property optimization workflows exemplify how density bias propagates into applied computational design [7, 12, 13]. Within dense subspaces—such as perovskites, layered oxides, or binary alloys—models demonstrate high predictive fidelity, enabling rapid screening and optimization. However, this performance does not generalize isotropically across materials space.

In sparse regimes, predictive ranking deteriorates systematically. Candidates exhibiting unconventional

bonding environments or compositional asymmetries are frequently mis-ranked due to representational extrapolation errors [12, 18]. This produces an illusion of design efficiency: models appear highly accurate within historically explored regions while silently excluding transformative outliers.

Multi-fidelity computational pipelines, which combine low-accuracy surrogate simulations with selective high-accuracy recalculations, further entrench density effects [3]. Surrogate models trained on dense datasets propagate their uncertainty structures into fidelity-bridging layers. High-accuracy recalibration therefore concentrates on already dense zones where surrogate confidence is highest, leaving sparse regimes epistemically under-refined. The cumulative outcome is a progressive narrowing of the explored design manifold toward computationally tractable, historically favored regions.

Uncertainty & interpretability

Uncertainty quantification frameworks have advanced significantly, evolving from ensemble variance heuristics to Bayesian neural networks, Monte Carlo dropout, and deep ensemble inference [3, 15]. These approaches estimate epistemic uncertainty arising from parameter indeterminacy and model structure. However, they rarely disentangle this from aleatoric uncertainty induced by infrastructural sparsity.

In density-skewed datasets, predictive variance conflates model ignorance with data absence. Regions of sparse sampling produce high epistemic uncertainty not solely because the model lacks knowledge, but because the infrastructural record itself is incomplete [15]. Without explicit density-aware decomposition, uncertainty estimates risk misguiding acquisition strategies by overstating confidence in dense regimes and understating ignorance in sparse ones.

Explainable AI techniques—including attention mapping, gradient attribution, and feature saliency analyses—offer partial transparency into model reasoning [15, 22]. Yet interpretability outputs reveal their own infrastructural conditioning. Feature importance scores disproportionately emphasize chemical motifs, coordination environments, and compositional ratios that dominate training corpora. Rather than neutral arbiters of mechanistic insight, interpretability tools become mirrors reflecting density concentration.

This paradox underscores a critical epistemic limit: interpretability can diagnose density bias but cannot, in its current form, mitigate it. The explanatory layer exposes infrastructural imbalance while remaining structurally dependent upon it.

Synthesis perspective

Taken collectively, these literatures reveal density not as a peripheral dataset characteristic but as a systems-level force shaping representation geometry, exploration dynamics, design outcomes, and epistemic confidence. Materials data infrastructures encode uneven historical sampling; representation architectures translate this unevenness into latent structure; discovery systems operationalize it into search trajectories; computational design workflows amplify its predictive asymmetries; and uncertainty frameworks struggle to disentangle its epistemic consequences. Density thus functions as an invisible steering field guiding the direction, speed, and scope of AI-driven materials discovery.

The systemic pathways through which uneven sampling propagates discovery bias across infrastructures, models, and autonomous pipelines are synthesized in **Table 1**.

Table 1. Density-Modulated Bias Pathways Across Computational Materials Discovery Pipelines

Pipeline Layer	Density Condition	Mechanism of Bias Emergence	Episten Consequ
Materials Data Infrastructures	Clustered sampling around well-studied chemistries	Historical simulation focus, convergence feasibility, publication bias	Skewe knowlec cartograp materia space
Representation Learning Architectures	Latent reinforcement of dense coordination motifs	Message passing aggregation and symmetry-conditioned embedding statistics	Distort latent geometry extrapola instabil

AI-Guided Discovery Systems	Exploration constrained to density peripheries	Variance-driven acquisition anchored near known clusters	Bounda explorat without d spars penetrat
Computational Design Workflows	Predictive fidelity concentrated in dense regimes	Surrogate training bias and ranking extrapolation errors	Illusory de efficien
Uncertainty & Interpretability Layers	Variance conflation between ignorance and sparsity	Lack of density-aware uncertainty decomposition	Mislead episten confider
Autonomous Discovery Feedback Loops	Iterative reinforcement of sampled domains	Generative and optimization mode collapse	Narrow innovati trajector

Proposed conceptual framework

The Density-Compensated Epistemic Exploration Framework (DCEF) reframes computational materials discovery as a density-modulated dynamical system. It consists of four interdependent layers connected by explicit feedback loops: (1) Density Assessment Layer, (2) Bias Identification Layer, (3) Compensation Layer, and (4) Steering Layer. The framework operates on the principle that discovery bias is not an external artifact but an emergent property of the interaction between sampling density $\rho(x)$, representation geometry, and inference uncertainty.

The Density Assessment Layer maps the computational corpus onto a continuous density field $\rho(x)$ defined over the chosen feature space (e.g., composition–structure–property manifold). This mapping is performed conceptually through kernel density estimation or graph-based neighborhood aggregation, yielding a scalar field that quantifies local data availability.

The Bias Identification Layer quantifies the deviation between observed density and an idealized uniform or target density field. This deviation can be conceptualized as

$$B(x) = 1 - \rho obs(x)\, \rho ideal(x) \quad (1)$$

where B(x) serves as an epistemic bias intensity metric. High values of B(x) indicate regions where inference is systematically under-constrained.

The Compensation Layer introduces re-weighting and adaptive re-sampling operators that counteract bias propagation. Representation learning within this layer incorporates density-aware attention mechanisms or loss terms that penalize over-reliance on dense clusters. Uncertainty propagation is modulated by a density-dependent scaling factor so that surrogate predictions in sparse regions carry appropriately inflated epistemic uncertainty.

The Steering Layer translates bias and compensation signals into pipeline-level actions: selective query generation, surrogate retraining triggers, and generative-model conditioning. Feedback loops ensure that each discovery iteration updates the density field, closing the loop between observation and steering.

These interactions are captured by the conceptual relation $\frac{dD}{dt} = f(B(x), C(x), U(x))$ where D denotes discovery potential, C(x) is the computational cost of compensation, and U(x) is propagated uncertainty. The function f embodies the steering logic that seeks to maximize D subject to epistemic balance rather than raw throughput.

A third relation formalizes the equilibrium condition toward which the framework converges:

$\min \rho \left(\int B(x)\, 2dx + \lambda \int C(x)\, dx \right)$ here λ is a conceptual trade-off parameter reflecting infrastructure constraints.

The layered interactions between density fields, bias diagnostics, compensation operators, and steering feedbacks are visualized in **Figure 1**.

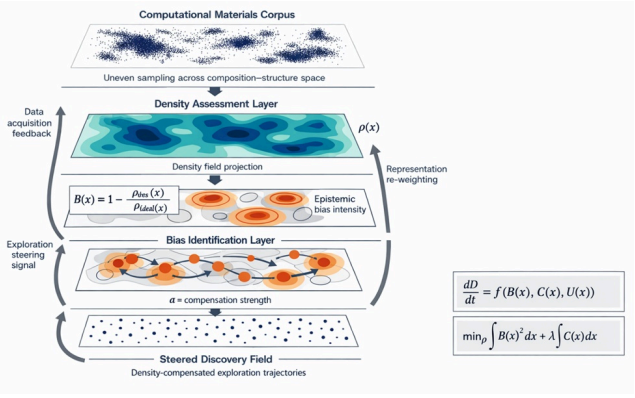

Figure 1. Density-Compensated Epistemic Exploration Framework (DCEF): A Layered Systems Model of Data Density–Driven Discovery Bias

The DCEF thus provides an interpretive infrastructure for diagnosing and steering around data-density-induced discovery bias. It does not prescribe specific algorithms but supplies the conceptual primitives—density fields, bias metrics, compensation operators, and steering equilibria—required to redesign materials data infrastructures and autonomous platforms at the epistemic level.

Analytical implications

The Density-Compensated Epistemic Exploration Framework (DCEF) carries several analytical consequences for how computational materials discovery pipelines are conceptualized and operated.

First, it shifts the locus of bias analysis from model architecture or training protocol to the upstream density landscape of the data corpus itself. Rather than attributing poor generalization or mode collapse solely to architectural limitations or optimization dynamics [4, 6, 7], the framework locates a primary source of distortion in the non-uniform ρ(x) field. This implies that even optimally tuned graph neural networks or equivariant transformers will exhibit systematic extrapolation errors in sparse regimes unless density modulation is introduced at the representation or inference stage.

Second, the bias metric B(x) defined in the framework provides a scalar diagnostic that can be tracked across pipeline iterations. Because B(x) is density-relative rather than absolute, it remains agnostic to the choice of feature representation—whether compositional fingerprints, graph embeddings, or multimodal descriptors—allowing consistent monitoring irrespective of evolving representation-learning practices [5, 6, 20]. This diagnostic role reveals how quickly (or slowly) autonomous systems escape historical density attractors and whether steering interventions produce measurable epistemic re-balancing.

Third, the compensation layer's density-aware mechanisms imply a fundamental trade-off between local fidelity and global coverage. Re-weighting schemes or adaptive sampling that penalize dense-cluster dominance necessarily reduce effective sample efficiency within well-sampled regions [9, 10]. The conceptual equilibrium condition

$$\min_{\rho}\left(\int B(x)^2 dx + \lambda \int C(x)\, dx \right) \quad (2)$$

thus surfaces an infrastructure-level Pareto frontier: aggressive bias compensation increases computational overhead (via targeted high-fidelity queries in sparse zones) while yielding broader epistemic coverage. Materials data platforms that prioritize short-term throughput over long-term diversity implicitly operate at high-B(x) equilibria, whereas those pursuing structural novelty must accept elevated C(x) costs.

Fourth, the steering layer's feedback dynamics suggest that discovery potential D is not monotonically increasing with data volume. Incremental addition of points to already dense regions can saturate representational capacity without meaningfully advancing epistemic reach, whereas targeted acquisition in high-B(x) zones produces disproportionate gains in D. This non-linearity reframes the value of new computational data: its marginal epistemic utility is inversely related to local density, providing a principled rationale for prioritizing sparse-regime exploration even when immediate property-prediction accuracy would favor continued exploitation of dense subspaces.

Finally, the framework exposes an under-appreciated interaction between uncertainty quantification and density modulation. Standard epistemic uncertainty estimates grow in sparse regions [3, 15], yet without explicit density compensation they frequently fail to drive acquisition into true deserts because boundary effects around dense clusters dominate variance-based acquisition functions [9]. By coupling uncertainty with B(x), the DCEF enables more faithful propagation of epistemic risk, ensuring that uncertainty genuinely reflects knowledge gaps rather than mere proximity to historical sampling patterns.

Results and Discussion

The introduction of data density as a primary epistemic bias driver invites a systemic reconsideration of several entrenched design logics in computational materials engineering. Rather than functioning as a secondary dataset characteristic, density emerges within the Density-Compensated Epistemic Exploration Framework (DCEF) as a structuring field that silently governs exploration trajectories, inference confidence, and discovery valuation. Recognizing this structuring role destabilizes throughput-

centric paradigms that have historically defined the field's computational expansion.

High-throughput screening infrastructures, originally engineered to maximize candidate enumeration within computationally tractable chemical subspaces [1, 2], operate through convergence-optimized allocation strategies. Workflow schedulers preferentially route computational resources toward compositions, symmetries, and unit-cell complexities that minimize electronic relaxation cost and maximize calculation success rates. While operationally efficient, this logic recursively amplifies density gradients: the fastest-converging chemistries become the most represented, the most modeled, and ultimately the most discoverable. Within the DCEF perspective, throughput ceases to be an epistemically neutral metric. Instead, it becomes a density amplifier.

The framework therefore suggests that throughput-oriented design criteria require supplementation with density-aware coverage indices capable of tracking exploration saturation across compositional and structural manifolds. Campaigns guided by such indices would deliberately allocate computational budget toward under-sampled phase spaces—even when convergence likelihood is low or property yield uncertain. This introduces a fundamental optimization tension between immediate discovery volume and long-horizon epistemic diversification. Screening infrastructures would thus transition from purely accelerative engines into density-balancing discovery systems.

Autonomous and closed-loop discovery platforms extend this tension into algorithmic governance. Contemporary systems optimize short-horizon objectives—predictive improvement, uncertainty reduction, or threshold attainment—without explicit awareness of global density topology [9, 10, 21]. Acquisition functions, even when uncertainty-weighted, remain locally conditioned by model familiarity. From the DCEF standpoint, such acquisition logics inadvertently reward exploration near dense knowledge basins while penalizing epistemic excursions into sparse regimes.

Incorporating density-compensated steering would necessitate reformulating acquisition functions to include coverage penalties and sparsity incentives. A density term could function as a counter-gradient, discouraging clustering in low-bias regions and rewarding penetration into high-bias zones. The resulting exploration trajectories would appear slower in early operational cycles, as models allocate queries toward epistemically distant candidates with lower predicted success probability. However, over extended discovery horizons, such steering could yield structurally diverse candidate portfolios and mitigate convergence toward historically saturated chemistries. This reframing raises unresolved evaluation questions: should autonomous systems be assessed by short-term optimization yield or long-term discovery topology diversification?

Representation-learning research communities have historically prioritized symmetry preservation, computational scalability, and expressive latent geometry [4, 6, 20]. The DCEF indicates that density robustness constitutes an equally foundational architectural axis. Current embedding strategies encode sampling frequency implicitly, allowing dense coordination environments to dominate latent manifold curvature. Future architectures may therefore incorporate density modulation layers capable of dynamically re-weighting neighborhood aggregation. Adaptive attention mechanisms could down-scale contributions from over-represented motifs, while contrastive or entropy-maximizing objectives could explicitly expand sparse-region embedding resolution. Importantly, such modulation would need to operate without eroding the physically grounded inductive biases—periodicity, equivariance, and locality—that underpin predictive reliability.

The emergence of foundation models for materials science intensifies the scale at which density imprinting operates. Large-scale pre-training across heterogeneous simulation and experimental corpora promises universal representation transferability [26]. Yet these models aggregate density gradients across infrastructures rather than neutralizing them. When scaled to billions of structural tokens, density bias does not dissipate; it consolidates. The DCEF thus implies that post-training alignment or downstream fine-tuning cannot adequately correct foundational density imprinting. Instead, density-compensated pre-training curricula—incorporating stratified sampling, sparsity up-weighting, or continual density diagnostics—may be required to prevent large models from institutionalizing infrastructural blind spots at generative scale.

Beyond algorithmic design, the framework foregrounds a broader epistemic governance challenge. As simulation–experiment coupling tightens within self-driving laboratory ecosystems [11, 21], density-conditioned inference begins

to shape experimental decision hierarchies. AI systems nominate candidates for synthesis based on predicted stability, functionality, and confidence. Because predictive confidence correlates with infrastructural density, experimentally actionable candidates disproportionately originate from well-sampled subspaces. Sparse-regime candidates—despite potentially transformative properties—receive lower prioritization due to elevated uncertainty. Density bias thus propagates from computational representation into laboratory materialization.

From a governance standpoint, this translation converts infrastructural sampling imbalance into experimental selection bias. Without density-compensated steering, autonomous laboratories risk reinforcing historically privileged chemistries while neglecting structurally novel domains. Infrastructure-level integration of density diagnostics—embedded within candidate ranking dashboards, experimental queue design, and funding allocation heuristics—therefore becomes essential to maintaining exploratory pluralism. Responsible deployment of AI-guided discovery systems will increasingly depend on whether they expand or contract the epistemic horizon of materials innovation.

Conclusion

Computational materials exploration has reached a developmental inflection point at which increases in dataset volume, algorithmic complexity, and model scale no longer guarantee proportional expansion in discovery breadth. The Density-Compensated Epistemic Exploration Framework reframes uneven sampling not as a peripheral data-quality artifact but as a foundational epistemic mechanism that governs what becomes discoverable, how confidence is distributed, and which regions of chemical space remain structurally occluded.

By introducing density fields, bias gradients, compensation operators, and steering equilibria, the DCEF establishes an interpretive infrastructure capable of tracing density imprinting across the full discovery stack—from data aggregation and representation learning to acquisition logic and experimental translation. This systems perspective reveals that predictive accuracy and exploratory coverage do not scale synchronously. Gains in local reliability may coincide with contractions in global search diversity, generating discovery ecosystems that are simultaneously precise and epistemically narrow.

The analytical implications of this reframing extend across infrastructural, architectural, and governance domains. High-throughput campaigns require density-aware allocation strategies; autonomous loops demand coverage-sensitive acquisition logics; representation architectures must incorporate sparsity-robust embedding dynamics; and foundation models necessitate density-compensated scaling curricula. Collectively, these redesign imperatives position density not as noise to be minimized but as structure to be governed.

Although the DCEF remains deliberately conceptual—eschewing empirical benchmarking and algorithmic prescription—it establishes data density as a first-order variable in the epistemology of computational materials discovery. Future progress in AI-guided materials engineering will depend not solely on accelerating computation or enlarging datasets, but on cultivating infrastructural awareness of how knowledge is distributed within those datasets. In this sense, the next phase of the field's evolution is unlikely to be defined by how much data we generate, but by how deliberately we navigate its density contours.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 04 Jan 2022 Revised: 31 Mar 2022 Accepted: 20 Apr 2022
Published online: 18 September 2022

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.

Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156-63. https://doi.org/10.1016/j.commatsci.2016.12.004.

Jensen S, Jacobsen TCS, Reuter K, Thygesen KS. Data-driven discovery of 2D materials by deep generative models. npj Computat Mater. 2022;8(1):232. https://doi.org/10.1038/s41524-022-00923-3.

Chen L, Tran H, Batra R, Kim C, Ramprasad R. Machine learning models for the prediction of energy, forces, and stresses for molecules and materials. npj Comput Mater. 2021;7(1):19.

Zheng P, Liu H, Wang J, Yu B, Gu X, Lu S. Enhancing geometric representations for molecules with equivariant transformer networks. npj Comput Mater. 2021;7(1):200.

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. npj Comput Mater. 2020;6(1):84

Zheng Z, Xu Z, Hu Y, Yaghi OM. Machine-Learning-guided morphology engineering of nanoscale metal-organic frameworks. Matter. 2020;3(4):1104-14.

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(1):21. https://doi.org/10.1038/s41524-019-0153-8.

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. https://doi.org/10.1038/s41524-019-0181-4.

Rosen AS, Iyer SM, Ray D, Yao Z, Aspuru-Guzik A, Gagliardi L, et al. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. Matter. 2021;4(5):1578-97. https://doi.org/10.1016/j.matt.2021.02.015.

Huang W, Martin P, Zhuang HL. Machine-learning phase prediction of high-entropy alloys. Acta Mater. 2019;169:225-36. https://doi.org/10.1016/j.actamat.2019.03.012.

Zou C, Li J, He Q, Liang D, Luo Y, Tong H, et al. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. Acta Mater. 2021;202:211-21. https://doi.org/10.1016/j.actamat.2020.10.056.

Park CW, Wolverton C. Self-supervised machine learning for alloy composition prediction using x-ray absorption spectroscopy. npj Comput Mater. 2022;8(1):1.

Li X, Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Yao Y. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):204. https://doi.org/10.1038/s41524-022-00884-7.

Saidi P, Zadkhast P, Sasani F, Shad E, Srivastava A. Machine learning-enabled discrete element method: A parallel computing perspective. Comput Mater Sci. 2021;197:110626.

Kim G, Diao H, Lee C, Samaei AT, Phan T, de Jong M, et al. First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation. Acta Mater. 2019;181:124-38. https://doi.org/10.1016/j.actamat.2019.09.026.

Vazquez G, Singh P, Sauceda D, Batzner S, Kozinsky B. Efficient machine-learning model for fast assessment of elastic properties of high-entropy alloys. Acta Mater. 2022;232:117927. https://doi.org/10.1016/j.actamat.2022.117927.

Möller JJ, Körner W, Krugel G, Urban DF, Elsässer C. Compositional optimization of hard-magnetic phases with machine-learning models. Acta Mater. 2018;153:53-61. https://doi.org/10.1016/j.actamat.2018.03.051.

Dunn A, Wang Q, Ganpule S, Wang D, Jain A. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm. npj Comput Mater. 2020;6(1):138. https://doi.org/10.1038/s41524-020-00406-3.

Saeki A, Ueda M, Matsunaga Y, Furusawa M, Hui JK, Kim MW, et al. Machine learning identification of experimental conditions for the synthesis of single-phase white phosphors. Matter. 2021;4(12):4040-57. https://doi.org/10.1016/j.matt.2021.10.004.

Fung V, Hu G, Ganesh P, Sumpter BG. Machine learned features from density of states for accurate adsorption energy prediction. Nat Commun. 2021;12(1):88.