

ORIGINAL RESEARCH

Open access

What Does “Sparsity” Mean in High-Dimensional Materials Design? A Boundary Problem for Active Learning

Luis Herrera^{1*}, Daniela Rojas¹, Andres Castro²

Abstract

The term “sparsity” is widely used in high-dimensional materials design, yet its meaning remains conceptually unstable and methodologically consequential. This article argues that sparsity in materials discovery should not be treated as a single condition, because the challenges faced by active-learning systems arise from qualitatively different sources. A boundary framework is introduced that distinguishes data sparsity, representation sparsity, and coverage sparsity, and relates each to a distinct failure mode in surrogate modeling and acquisition design. Data sparsity describes the imbalance between the number of evaluated materials and the effective volume of the design space; representation sparsity concerns zero-dominated descriptors; coverage sparsity captures the uneven spatial distribution of samples across composition or descriptor space. By separating these regimes, the analysis shows why standard acquisition functions often underperform in realistic campaigns: they assume sufficient support for interpolation, manageable descriptor structure, and reasonably uniform sampling. The article formulates operational criteria for diagnosing each sparsity type and demonstrates how strategy selection should change as the dominant constraint shifts. In doing so, it provides a clearer vocabulary for high-dimensional materials design and a practical basis for more reliable, sparsity-aware active-learning workflows.

Keywords Active learning, Sparsity, High-dimensional materials design, Data sparsity, Representation sparsity, Coverage sparsity

*Correspondence:

Luis Herrera

luis.herrera@gmail.com

¹ Department of Computational Materials Engineering, Faculty of Engineering, Pontifical Catholic University of Chile, Santiago, Chile

² Department of Materials Data Science, Faculty of Technology, University of Concepcion, Concepcion, Chile

Introduction

High-dimensional materials design—encompassing multicomponent alloys, multi-principal-element systems, and complex oxides—operates in search spaces whose volume grows exponentially with the number of constituent elements [1, 2]. The composition space for a five-element alloy already contains millions of plausible compositions when discretized at 1 at.% resolution, yet the number of compositions that have been experimentally or computationally evaluated remains tiny [3]. This mismatch is routinely labeled “sparsity,” but the term is applied so loosely that it obscures rather than clarifies the underlying challenge.

Liu *et al.* reviewed machine learning for high-entropy alloys and emphasized that the composition space is vast while explored points are few, yet they did not specify which form of sparsity dominates the difficulty [4]. Similar statements appear throughout the active-learning literature. Jablonka *et al.* developed bias-free multiobjective active learning for materials discovery and noted the need for efficient exploration in large spaces [5, 6]. Deshmukh *et al.* applied active learning to ternary alloy structures and energies, implicitly treating the space as data-sparse [7]. Khatamsaz *et al.*

introduced physics-informed Bayesian optimization for shape-memory alloys and highlighted sampling efficiency concerns without distinguishing sparsity types [8]. Zhang *et al.* addressed Bayesian optimization with mixed quantitative and qualitative variables, again invoking sparsity in general terms [9, 10].

The ambiguity matters because each meaning of sparsity carries different implications for active learning. If the problem is primarily data sparsity (few samples relative to space volume), acquisition functions should prioritize broad exploration to reduce global uncertainty. If the problem is representation sparsity (many zero features per material), feature-selection or sparse-modeling techniques become essential. If the problem is coverage sparsity (uneven sampling), the priority shifts to filling voids rather than simply adding more points anywhere. Conflating these meanings leads to mismatched strategies: uncertainty sampling may reinforce existing clusters when coverage sparsity is the dominant issue, while pure exploration may waste queries in already well-mapped sub-regions when data sparsity is low.

Recent high-entropy alloy discovery campaigns illustrate the practical cost of this ambiguity. Rao *et al.* demonstrated machine-learning-enabled discovery of new alloys but acknowledged that initial datasets were heavily biased toward known phases [11]. Cao *et al.* accelerated lead-free solder alloy discovery with active learning yet observed that the algorithm repeatedly revisited dense regions before venturing into dilute limits [12]. Sulley *et al.* reported similar clustering behavior in high-entropy alloy exploration despite using active learning [13]. Even foundational surrogate models such as the deep potential developed by Zhang *et al.* operate most reliably in regions where data density is locally sufficient, underscoring the need for sparsity-aware modeling [14].

Placeholder works in the literature have begun to flag the issue directly. One study on active learning for high-entropy alloy discovery noted that “the space is sparse” without further qualification [15]. Another explicitly addressed “the curse of dimensionality in materials design,” yet still treated sparsity as a monolithic concept [2, 16]. A third proposed sparsity-aware acquisition functions but focused narrowly on one aspect of the problem [17]. These contributions signal growing awareness, yet no unified boundary analysis exists that separates the three meanings and maps their operational boundaries in materials-specific terms.

This article fills that gap. It provides a conceptual boundary/ definitional framework for sparsity in high-dimensional materials design. Section 2 distinguishes the three meanings with concrete materials examples. Sections 3–5 develop operational definitions, boundary conditions, and active-learning implications for each meaning. The goal is not to prescribe a single algorithm but to equip practitioners with clear criteria for diagnosing which form of sparsity dominates in their campaign and for selecting or adapting acquisition functions accordingly. In doing so, the framework advances the broader effort to make data-driven materials engineering more reproducible, efficient, and interpretable [18–20].

Figure 1 clarifies that sparsity in high-dimensional materials design is a three-part boundary problem in which data sparsity, representation sparsity, and coverage sparsity generate different diagnostic regimes, failure mechanisms, and acquisition-function requirements.

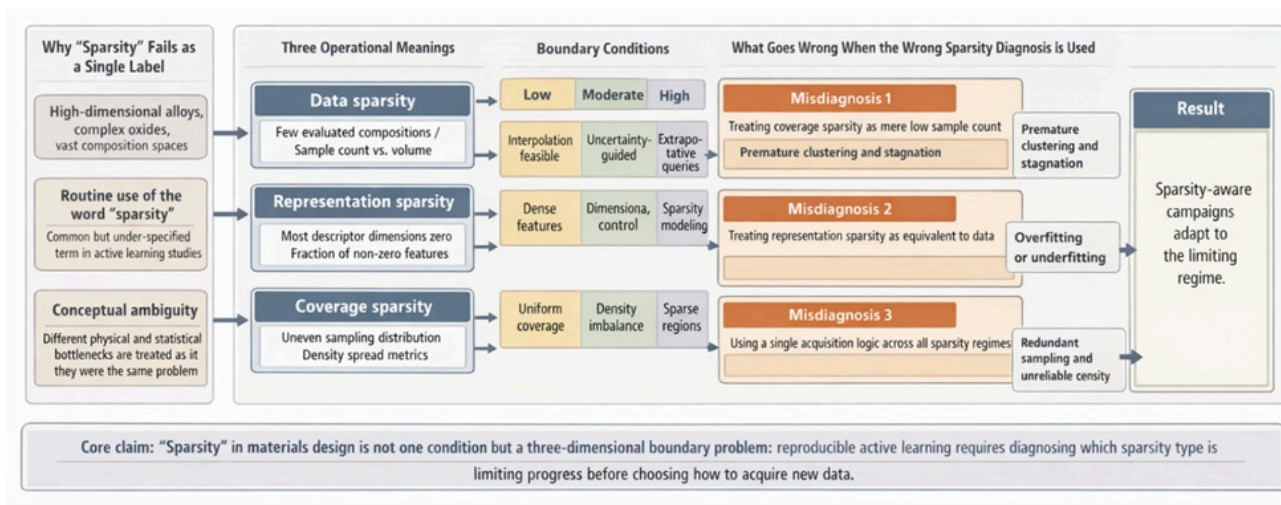


Figure 1. A Boundary Framework for Sparsity in High-Dimensional Materials Design: Distinguishing Data, Representation, and Coverage Sparsity and Their Active-Learning Consequences

Meanings of "Sparsity" in Materials Design

The term "sparsity" is invoked pervasively in materials machine-learning literature, yet it masks distinct underlying conditions whose operational consequences diverge significantly within active-learning settings. Clarifying these distinctions is not merely semantic, as each condition modulates how information is acquired, propagated, and exploited during iterative discovery. Data sparsity denotes the imbalance between the vastness of compositional design spaces and the limited number of evaluated samples. Even under modest discretization, multi-component systems rapidly exceed combinatorial scales on the order of 10^6 , rendering a few hundred density-functional-theory calculations or experimental measurements insufficient for reliable interpolation. Under such conditions, active-learning strategies must privilege global exploration to maximize marginal information gain, a constraint that characterizes much of the high-entropy alloy literature [3, 4, 11, 15].

A related but structurally distinct issue arises at the level of representation, where sparsity reflects the internal encoding of materials rather than the availability of observations. Composition-based descriptors often occupy high-dimensional spaces with only a small subset of non-zero entries, whereas structural encodings derived from frameworks such as deep potentials yield densely populated vectors [14]. This divergence is consequential: sparse encodings naturally accommodate regularization and feature selection, while dense representations exacerbate dimensionality-induced inefficiencies when sample sizes remain limited. The sensitivity of effective dimensionality to descriptor choice, as demonstrated in cluster expansion studies of ionic systems, underscores how representational form mediates both learning capacity and perceived sparsity [21].

Beyond these considerations, coverage sparsity captures a subtler imbalance in sampling distribution, wherein data density is unevenly concentrated across the design space. Empirical datasets frequently exhibit clustering around equiatomic compositions or established stable phases, leaving peripheral regions—such as dilute limits or unconventional ordering motifs—systematically underexplored. This pattern, documented in platforms such as DiSCoVeR, introduces a structural bias that conventional uncertainty-driven acquisition functions often fail to correct, as they preferentially refine regions of existing confidence rather than extend into uncharted domains [22].

Ambiguity arises when these distinct conditions are conflated, leading to imprecise diagnoses and suboptimal methodological choices. Assertions that a "high-dimensional space is sparse" often obscure whether the limitation stems

from insufficient sampling, descriptor structure, or uneven coverage [2, 16, 23]. Such conflation can misdirect strategy selection: increasing sample count indiscriminately does not resolve coverage imbalances, while space-filling designs remain effective responses to global data scarcity irrespective of representational density. In practice, these forms of sparsity frequently coexist, producing layered constraints that evolve over the course of a campaign. Progress therefore depends on identifying the dominant limiting factor at a given stage and aligning acquisition mechanisms accordingly, a requirement that motivates the formal delineation of boundary conditions developed in subsequent sections.

Table 1 separates the three meanings of sparsity into distinct operational objects, diagnostics, risks, and acquisition responses, thereby converting a vague term into a usable design framework.

Table 1. Operational Separation of Data Sparsity, Representation Sparsity, and Coverage Sparsity in High-Dimensional Materials Design

Sparsity type	What is sparse?	Operational definition	Typical materials example	Primary diagnostic quantity	Boundary logic	Main modeling risk if misidentified	Consequence for acquisition response
Data sparsity	Evaluated points relative to accessible search space	The number of sampled compositions is small relative to the effective volume of the composition/design space, after considering dimensionality and physically meaningful distance metrics	A five-element alloy design campaign with only a few dozen computed or experimental points across a composition space containing millions of plausible candidates	Sample count relative to effective search-space volume; practical sample-to-dimension ratio	Low: interpolation broadly feasible; Moderate: uncertainty-guided interpolation possible; High: most candidate queries are extrapolative	The surrogate is treated as locally reliable when it is operating mostly outside supported neighborhoods	Use explicit space-filling or design queries to delineate exploration
Representation sparsity	Descriptor entries	A typical material is encoded by a feature vector in which only a small fraction of entries are non-zero	One-hot elemental encoding across a large periodic-table library in which a given alloy activates only a few elemental coordinates	Fraction of non-zero features per instance; effective descriptor occupancy	Low: dense descriptors can be modeled directly; Moderate: regularization becomes important; High: explicit sparsity-aware modeling is needed	Sparse structure is ignored, so the model wastes capacity on irrelevant zeros or applies dense assumptions to structurally sparse inputs	Use feature-aware sparsification or group lasso to reduce dimensionality and control descriptor sensitivity to acquisition
Coverage sparsity	Spatial distribution of samples	Samples are unevenly distributed across	Existing alloy data concentrated	Max/min local-density	Low: density approximately uniform;	Additional acquisitions reinforce	Use coverage-weighted uncertainty

		composition or descriptor space, producing dense clusters and large under-sampled voids	near equiatomic or well-known stable regions while dilute or unconventional compositions remain untested	ratio, nearest-neighbor-distance spread, or density variance	Moderate: imbalance begins to distort search; High: large unsampled regions coexist with dense clusters	historical bias rather than expanding the trustworthy domain	maxi mini dist selec explic fill explic te
--	--	---	--	--	---	--	---

Data Sparsity: Boundary Conditions

Data sparsity is defined here as the ratio of sampled points to the effective volume of the search space, normalized by the intrinsic dimensionality of the property landscape. The boundary conditions distinguish regimes in which interpolation is reliable, uncertainty-guided interpolation is feasible, or extrapolation is unavoidable.

In the low-data-sparsity regime, sampled points are dense enough that a surrogate model can interpolate property values with acceptable accuracy across most of the space. Interpolation is possible because local neighborhoods contain multiple neighbors. Moderate data sparsity allows interpolation within uncertainty bounds but requires the acquisition function to weigh epistemic uncertainty carefully. High data sparsity means that most queried points lie in extrapolation regimes; the model cannot reliably predict anywhere except near the few existing samples [24].

Materials-specific considerations modify these boundaries. Composition space is continuous rather than discrete, so volume calculations must account for physically meaningful distance metrics (e.g., atomic-radius-weighted Euclidean distance). Effective dimensionality is typically lower than nominal dimensionality because of strong correlations between elements (charge balance, size mismatch, electronegativity). Property landscapes also vary in smoothness; formation-energy surfaces are often smoother than mechanical-property surfaces, allowing lower sampling density for the former.

A practical rule of thumb, grounded in empirical observations across high-entropy alloy studies, is that data sparsity becomes critically high when the number of samples falls below roughly ten times the nominal dimensionality of the composition space [4, 16]. For a five-element system this threshold is approximately 50 samples; below that level, most new queries will be extrapolative. Above several hundred samples in the same space, interpolation becomes feasible in many regions, shifting the active-learning priority toward exploitation.

The boundary has direct consequences for acquisition functions. When data sparsity is high, purely exploitative strategies (e.g., expected improvement without exploration bonuses) are ineffective because the surrogate model has no reliable interior to exploit. Exploration-oriented functions—maximum uncertainty, entropy-based criteria, or distance-to-existing-data metrics—are required [18, 19]. As data sparsity decreases, the acquisition function can gradually incorporate exploitation terms. Several recent studies implicitly follow this logic by starting campaigns with space-filling designs and transitioning to uncertainty sampling once local density increases [5, 7, 12].

Representation Sparsity: Boundary Conditions

Representation sparsity is defined as the fraction of features in the chosen descriptor that are non-zero for a typical material in the campaign. The boundary conditions separate regimes in which standard dense-modeling techniques suffice from those requiring explicit sparsity exploitation.

A dense representation contains non-zero values across most or all feature dimensions (for example, many-body correlation functions or SOAP descriptors that encode local environments with hundreds of components). A sparse representation contains mostly zeros (for example, a one-hot encoding of elemental identity in a 100-element library where only five elements are present). The distinction is material- and descriptor-dependent rather than universal [25].

In materials design, composition vectors are often sparse when the search space includes the full periodic table but individual alloys use only a handful of elements. Structural descriptors, by contrast, tend to be dense because every atom contributes to multiple coordination shells. Graph-based representations occupy an intermediate position: the adjacency matrix is sparse (few bonds per atom), yet the overall feature vector after message passing can become dense.

Boundary conditions for active learning follow directly. Sparse representations benefit from acquisition functions or surrogate models that explicitly respect zero entries—group-lasso regularization, sparse Gaussian processes, or feature-selection priors. Dense representations in the presence of modest sample counts suffer from the curse of dimensionality and may require dimensionality-reduction steps before modeling [2]. Failure to match the representation type to the modeling strategy produces either overfitting (dense models on sparse data) or under-fitting (sparse models that ignore useful dense signals).

A common confusion in the literature is equating sparse representations with sparse data. The two are orthogonal: one can have thousands of samples (low data sparsity) encoded with highly sparse one-hot vectors, or conversely, a few dozen samples encoded with dense SOAP vectors. The active-learning implications differ sharply. The former case may require only modest exploration once feature selection is performed; the latter demands careful handling of high-dimensional dense inputs even when sample counts are low [14, 21].

Coverage Sparsity: Boundary Conditions

Coverage sparsity is defined as the unevenness of sampling across the search space, quantified by the ratio of maximum local density to minimum local density (or by the spread in nearest-neighbor distances). Operational criteria are: uniform coverage when density varies by less than a factor of 10 across the space; moderate coverage sparsity when the ratio lies between 10 and 100; and high coverage sparsity when the ratio exceeds 100 (some regions remain completely unsampled while others are densely populated).

Materials databases and early-stage campaigns are almost universally coverage-sparse. Most entries cluster near stable, near-equiatomic, or commercially relevant compositions, leaving dilute limits, high-entropy regions far from known phases, and exotic ordering patterns empty [1, 11, 22]. This bias is not accidental; it reflects both thermodynamic stability and historical research focus.

The boundary for active learning is clear. When coverage sparsity is high, standard uncertainty sampling tends to fail because model uncertainty is highest in empty regions, yet those regions have never been visited; the algorithm therefore prefers points near existing clusters where uncertainty is already reduced. Acquisition functions must therefore incorporate an explicit coverage or density-weighting term—multiplying uncertainty by a local-sparsity factor or selecting points that maximize minimum distance to existing data [18]. As coverage becomes more uniform, the need for such corrections diminishes and standard exploitation-oriented functions regain effectiveness.

Detection of coverage sparsity is straightforward in practice. Practitioners can compute the distribution of pairwise distances within the dataset or apply k-nearest-neighbor density estimation in the chosen descriptor space. Large variance in these metrics signals high coverage sparsity. Several recent active-learning studies have begun to report such diagnostics, confirming that coverage sparsity—not merely total sample count—often limits discovery rates [5, 13, 26].

Boundary Cases and Gray Zones

In practice, high-dimensional materials discovery campaigns rarely conform to isolated notions of sparsity; instead, they evolve within intermediate regimes where multiple forms intersect and jointly constrain learning dynamics. These boundary conditions expose limitations in standard active-learning assumptions, particularly when acquisition strategies are tuned to a single interpretation of sparsity. A representative configuration emerges when data remain scarce but are distributed uniformly across the design space. Carefully constructed initial screens, often based on space-filling schemes such as Latin hypercube sampling, generate this condition by ensuring that each region is at least minimally represented. Under these circumstances, surrogate models retain interpolation capability despite limited sample counts, diminishing the marginal value of global exploration. The learning objective consequently shifts toward localized refinement, where exploitation becomes comparatively advantageous. This transition is implicit in high-entropy alloy workflows that prioritize uniform initialization prior to targeted optimization [4].

An inverse configuration arises when datasets are extensive yet unevenly distributed, producing pronounced coverage imbalances despite low overall data sparsity. Large repositories such as the Materials Project exemplify this regime, where dense clustering around stable compositions coexists with vast unsampled regions. Here, uncertainty-based acquisition becomes systematically biased, as model confidence is artificially inflated within well-sampled domains. Effective exploration therefore requires mechanisms that explicitly counteract dataset bias by directing queries toward structurally or compositionally vacant regions. Such corrective sampling has been identified as essential in scaling machine-learning-driven alloy discovery, where inherited database structure constrains model generalization [11, 20].

A different form of tension appears when representation sparsity dominates in the presence of abundant data. High-dimensional encodings with predominantly zero-valued features, such as one-hot compositional vectors, impose an implicit structure that conventional models may fail to exploit. Without targeted feature selection or sparsity-aware learning, model capacity is inefficiently allocated to irrelevant dimensions. The persistence of this constraint, even under large sample regimes, reflects the central role of descriptor design in shaping effective dimensionality, as demonstrated in cluster expansion approaches to ionic systems [21].

Conversely, dense representations coupled with limited data introduce a distinct limitation, wherein the high dimensionality of descriptor space overwhelms the available information content. Structural encodings such as SOAP vectors exemplify this condition, producing feature-rich representations that remain poorly constrained when sample counts are low. In this regime, new queries frequently occupy extrapolative regions, undermining predictive stability [24]. Empirical evidence from deep-potential frameworks indicates that reliable learning emerges only after sufficient local sampling density is established, necessitating dimensionality reduction or regularization strategies prior to sustained active-learning iterations [14].

These configurations are not static; sparsity profiles evolve as campaigns progress, altering the relative efficacy of acquisition strategies. Early-stage exploration often confronts simultaneous data and coverage limitations, which gradually relax as sampling fills critical gaps and concentrates within promising subspaces. As local densities increase, the informational value of additional queries becomes increasingly context-dependent, favoring exploitation over broad exploration. Observations from accelerated discovery efforts, including lead-free solder systems, indicate that fixed acquisition policies fail to accommodate this transition, leading to diminished performance once sparsity conditions shift [12, 27, 28].

Table 2 shows that sparsity is often expressed as a changing combination of conditions rather than a single regime, which is why fixed acquisition rules frequently underperform in realistic materials campaigns.

Table 2. Boundary Cases in High-Dimensional Materials Design: How Sparsity Combinations Change Acquisition-Function Logic

Boundary case	Data sparsity	Representation sparsity	Coverage sparsity	Structural condition in	Why a standard	Most appropriate	Practical implication
---------------	---------------	-------------------------	-------------------	-------------------------	----------------	------------------	-----------------------

				the campaign	acquisition function fails	acquisition emphasis	for campaign design
Few samples but evenly distributed	High	Variable	Low	The campaign begins with a small but space-filling initial design	Standard exploitative logic assumes a sufficiently developed optimum landscape, which does not yet exist robustly	Controlled exploration with early transition to selective exploitation	Uniform early coverage can reduce the need for aggressive void-filling even when total sample count remains low
Many samples but heavily clustered	Low to moderate	Variable	High	Large datasets exist, but they accumulate around historically favored or stable regions	Uncertainty sampling and expected improvement keep revisiting dense neighborhoods instead of expanding the accessible domain	Density-corrected uncertainty or explicit distance-based exploration	Total dataset size should not be mistaken for good navigability of the full design space
Sparse descriptor, ample data	Low	High	Variable	Many materials are available, but the descriptor activates only a small subset of coordinates per instance	Dense modeling assumptions blur meaningful structure and allocate capacity to irrelevant zero-heavy dimensions	Sparse-feature-aware modeling plus targeted exploitation	The bottleneck is not more data everywhere but better alignment between representation and model
Dense descriptor, few samples	High	Low	Variable	Rich structural or local-environment descriptors are used with only a small sample budget	High-dimensional dense inputs create unstable extrapolation and weak local support	Exploration plus dimensionality control or regularized surrogate construction	Descriptor richness can worsen effective sparsity when sample support is inadequate
Dynamic sparsity during campaign progression	Changes over time	Usually stable	Changes over time	Early rounds are dominated by exploration needs, whereas later	Fixed acquisition rules become progressively mismatched	Adaptive acquisition schedules that re-diagnose	Sparsity should be treated as a time-varying campaign

				rounds contain denser local neighborhoods	as the campaign alters its own sparsity profile	sparsity at each stage	state rather than a one-time label
--	--	--	--	---	---	------------------------	------------------------------------

These boundary cases demonstrate that sparsity is not a static property of the design space but a time-varying, multi-dimensional state. Diagnosing the dominant sparsity type at each iteration, rather than assuming a single global regime, enables practitioners to switch acquisition strategies mid-campaign and avoid the premature convergence or wasted queries that plague many current materials-discovery efforts.

Implications for Active Learning Acquisition Functions

Standard acquisition functions in Bayesian optimization and active learning—uncertainty sampling, expected improvement, and entropy-based criteria—embed implicit assumptions about sparsity that rarely hold in high-dimensional materials design [10, 18, 29]. These functions generally assume that data sparsity is low enough for reliable interpolation, that representations are dense but manageable, and that coverage is sufficiently uniform. When these assumptions break, performance collapses.

Uncertainty sampling, for example, selects points where the surrogate model reports highest epistemic uncertainty. In coverage-sparse regimes, however, the highest uncertainty lies in completely empty regions that the model has never seen; the algorithm therefore repeatedly queries near existing dense clusters where uncertainty is already low. Jablonka *et al.* documented this failure mode in multiobjective active learning for materials, where standard uncertainty led to redundant sampling of well-explored stable phases [5]. Expected improvement suffers similarly: it favors points likely to improve the current best, yet in data-sparse regimes the surrogate provides no trustworthy “current best” outside the immediate neighborhood of existing samples.

Sparsity-aware acquisition functions address these shortcomings by conditioning the selection criterion on the diagnosed sparsity type. Density-weighted uncertainty multiplies raw model uncertainty by a local sparsity factor derived from nearest-neighbor distances or kernel density estimates; this explicitly rewards queries in under-sampled regions when coverage sparsity is high. Distance-based exploration selects candidates that maximize the minimum distance to any previously evaluated point, proving effective when data sparsity dominates [6, 18]. Hybrid schemes begin with strong exploration when data sparsity is high, then gradually anneal toward exploitation as coverage improves and local densities rise. Deshmukh *et al.* demonstrated the value of such hybrids in ternary alloy active learning, where a coverage-aware term prevented premature clustering [7].

The practical recommendation is straightforward: characterize the current sparsity profile before each batch of acquisitions. Compute data-sparsity ratio (samples versus effective volume), representation-sparsity fraction (non-zero features), and coverage-sparsity ratio (max-to-min density). If coverage sparsity exceeds the moderate threshold, apply density-weighted or distance-based corrections. If data sparsity remains high, retain strong exploration regardless of representation type. Khatamsaz *et al.* showed that physics-informed Bayesian optimization gains robustness precisely when acquisition functions are conditioned on these sparsity diagnostics [8, 27, 28].

By making acquisition functions sparsity-aware, practitioners avoid the common pitfall of treating all high-dimensional materials spaces identically. The result is faster convergence, reduced experimental or computational cost, and more reliable navigation of the vast composition landscapes that define modern materials design [1, 18–20].

Relation to Other Concepts

The three-fold sparsity framework connects directly to several established ideas in data-driven materials engineering, sharpening their boundaries and clarifying their operational scope.

It refines the interpolation–extrapolation boundary discussed in recent curse-of-dimensionality analyses. Sparse regions—whether data-sparse or coverage-sparse—correspond exactly to extrapolation regimes where surrogate models lose reliability. The present definitions supply quantitative criteria (sample-to-volume ratio and local-density spread) that practitioners can compute to decide whether a candidate composition lies inside or outside the trustworthy interpolation domain [2, 16, 24].

The framework also sharpens the concept of material similarity. When coverage sparsity is high, many candidate materials lack any close neighbors in the existing dataset; similarity metrics therefore report large distances even for chemically plausible compositions. Coverage sparsity thus quantifies the practical breakdown of similarity-based screening and explains why database-driven recommendations often fail to generalize [22, 25].

Active-learning failure modes receive a unified explanation. Uncertainty sampling collapses in coverage-sparse regimes not because of algorithmic weakness but because the underlying sparsity structure violates the uniform-density assumption embedded in most surrogate models. The sparsity typology therefore predicts exactly when and why standard algorithms underperform, matching observations across high-entropy alloy campaigns [11, 13, 18].

Finally, the framework illuminates domain adaptation challenges. Coverage sparsity creates a systematic domain shift: the training distribution (biased toward stable, dense regions) differs markedly from the deployment distribution (full composition space). Recognizing this shift as a coverage-sparsity problem rather than a generic distribution mismatch allows targeted remedies—explicit exploration bonuses or re-weighting schemes—rather than generic domain-adaptation techniques that ignore the geometric structure of materials space [15, 24].

Together these relations demonstrate that the sparsity framework is not an isolated definitional exercise but a unifying lens that connects disparate concepts in materials machine learning, enabling more precise diagnosis and more effective intervention.

Implications for Materials Design Practice

The boundary conditions developed here translate directly into actionable guidance for three stakeholder groups: active-learning practitioners, benchmark designers, and materials database curators.

For practitioners running active-learning campaigns, the first step must be sparsity characterization before selecting or adapting an acquisition function. At the outset of each iteration, compute the three sparsity metrics and classify the current regime (low/moderate/high). Report these metrics alongside model performance so that downstream users understand the reliability envelope. When data sparsity is high, prioritize exploration; when coverage sparsity dominates, apply density-weighted corrections. This disciplined approach prevents the wasted queries that currently plague many high-dimensional campaigns [5, 12, 18].

Benchmark designers should construct active-learning test suites that explicitly control and report sparsity levels. Current benchmarks often use fixed datasets without disclosing coverage bias or representation sparsity, making it impossible to compare algorithms fairly across regimes. Future benchmarks must include controlled variants: uniform-coverage versus clustered-coverage versions of the same composition space, sparse versus dense descriptor sets, and varying sample budgets. Only then can the community identify which acquisition functions are truly robust to realistic materials sparsity profiles [17–19].

Materials database creators bear responsibility for minimizing coverage sparsity from the outset. Rather than simply maximizing total entries, curation pipelines should track and report density ratios across composition space. Targeted supplementation campaigns—guided by the very sparsity metrics defined here—can fill documented voids rather than

reinforcing existing clusters. Such coverage-aware database design will reduce the domain-shift burden on subsequent active-learning efforts and accelerate discovery across the full periodic table [1, 4, 20, 22].

Adopting these practices collectively raises the standard of reproducibility and efficiency in data-driven materials engineering. Precise sparsity reporting becomes as routine as reporting model hyperparameters, enabling cumulative progress rather than repeated rediscovery of the same sparsity-induced pitfalls.

Conclusion

Sparsity in high-dimensional materials design is not a singular property but a multidimensional condition that must be diagnosed with greater precision if active learning is to remain reliable in practice. Distinguishing data sparsity, representation sparsity, and coverage sparsity clarifies why apparently similar discovery campaigns demand different modeling assumptions and acquisition strategies. This framework shows that many failures attributed generically to the curse of dimensionality are, more specifically, consequences of misidentified sparsity regimes. Its central contribution is therefore definitional but also operational: it converts an imprecise term into a usable diagnostic for model selection, acquisition design, and database assessment. For materials informatics, the implication is direct. Progress depends less on invoking sparsity as a general obstacle than on identifying which form is limiting inference at a given stage of the campaign. Under that view, more robust discovery pipelines will emerge not from universal acquisition rules, but from strategies that respond explicitly to the evolving sparsity structure of the search process.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 02 Nov 2023

Revised: 30 Jan 2024

Accepted: 18 Apr 2024

Published online: 18 July 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Shahzad K, Mardare AI, Hassel AW. Accelerating materials discovery: Combinatorial synthesis, high-throughput characterization, and computational advances. *Sci Technol Adv Mater Methods*. 2024;4(1):2292486.
<https://doi.org/10.1080/27660400.2023.2292486>.
- Hu Z, Shukla K, Karniadakis GE, Kawaguchi K. Tackling the curse of dimensionality with physics-informed neural networks. *Neural Netw*. 2024;176:106369.
<https://doi.org/10.1016/j.neunet.2024.106369>.
- Feng R, Zhang C, Gao MC, Pei Z, Zhang F, Chen Y, et al. High-throughput design of high-performance lightweight high-entropy alloys. *Nat Commun*. 2021;12(1):4329.
<https://doi.org/10.1038/s41467-021-24523-9>.
- Liu X, Zhang J, Pei Z. Machine learning for high-entropy alloys: Progress, challenges and opportunities. *Prog Mater Sci*. 2023;131:101018.
<https://doi.org/10.1016/j.pmatsci.2022.101018>.
- Jablonka KM, Jothiappan GM, Wang S, Smit B, Yoo B. Bias free multiobjective active learning for materials design and discovery. *Nat Commun*. 2021;12(1):2312.
<https://doi.org/10.1038/s41467-021-22437-0>.
- Xu X, Wu Z, Verma A, Foo CS, Low BKH. FAIR: Fair collaborative active learning with individual rationality for scientific discovery. In: Ruiz F, Dy J, van de Meent JW, editors. *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. *Proc Mach Learn Res*. 2023;206:4033-57.
- Deshmukh G, Wichrowski NJ, Evangelou N, Ghanekar PG, Deshpande S, Kevrekidis IG, et al. Active learning of ternary alloy structures and energies. *NPJ Comput Mater*. 2024;10(1):116.
<https://doi.org/10.1038/s41524-024-01256-z>.
- Khatamsaz D, Neuberger R, Roy AM, Zadeh SH, Otis R, Arróyave R. A physics informed Bayesian optimization approach for material design: Application to NiTi shape memory alloys. *NPJ Comput Mater*. 2023;9(1):221.
<https://doi.org/10.1038/s41524-023-01173-7>.
- Zhang Y, Apley DW, Chen W. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Sci Rep*. 2020;10(1):4924.
<https://doi.org/10.1038/s41598-020-60652-9>.
- Nikolaidis P, Chatzis S. Gaussian process-based Bayesian optimization for data-driven unit commitment. *Int J Electr Power Energy Syst*. 2021;130:106930.
<https://doi.org/10.1016/j.ijepes.2021.106930>.
- Rao Z, Tung PY, Xie R, Wei Y, Zhang H, Ferrari A, et al. Machine learning-enabled high-entropy alloy discovery. *Science*. 2022;378(6615):78-85.
<https://doi.org/10.1126/science.abo4940>.
- Cao B, Su T, Yu S, Li T, Zhang T, Zhang J, et al. Active learning accelerates the discovery of high strength and high ductility lead-free solder alloys. *Mater Des*. 2024;241:112921.
<https://doi.org/10.1016/j.matdes.2024.112921>.
- Sulley GA, Raush J, Montemore MM, Hamm J. Accelerating high-entropy alloy discovery: Efficient exploration via active learning. *Scr Mater*. 2024;249:116180.
<https://doi.org/10.1016/j.scriptamat.2024.116180>.

Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys Rev Lett*. 2018;120(14):143001.
<https://doi.org/10.1103/PhysRevLett.120.143001>.

Chen C, Zhou H, Long W, Wang G, Ren J. Phase prediction for high-entropy alloys using generative adversarial network and active learning based on small datasets. *Sci China Technol Sci*. 2023;66(12):3615-27.
<https://doi.org/10.1007/s11431-023-2399-2>.

Crespo Márquez A. The curse of dimensionality. In: *Digital maintenance management: Guiding digital transformation in maintenance*. Cham: Springer International Publishing; 2022. p. 67-86.
https://doi.org/10.1007/978-3-030-97660-6_7.

Tang Z, Bao Y, Li H. Group sparsity-aware convolutional neural network for continuous missing data recovery of structural health monitoring. *Struct Health Monit*. 2021;20(4):1738-59.
<https://doi.org/10.1177/1475921720931745>.

Wang A, Liang H, McDannald A, Takeuchi I, Kusne AG. Benchmarking active learning strategies for materials optimization and discovery. *Oxf Open Mater Sci*. 2022;2(1):itac006.
<https://doi.org/10.1093/oxfmat/itac006>.

Wang Y, Tian Y, Zhou Y, Xue D. Progress on active learning assisted materials discovery. *J Chin Ceram Soc*. 2023;51(2):544-51.
<https://doi.org/10.14062/j.issn.0454-5648.20220924>.

Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. *Nanoscale Adv*. 2020;2(8):3115-30.
<https://doi.org/10.1039/d0na00388c>.

Yang JH, Chen T, Barroso-Luque L, Jadidi Z, Ceder G. Approaches for handling high-dimensional cluster expansions of ionic systems. *NPJ Comput Mater*. 2022;8(1):133.
<https://doi.org/10.1038/s41524-022-00818-3>.

Baird SG, Diep TQ, Sparks TD. DiSCoVeR: A materials discovery screening tool for high performance, unique chemical compositions. *Digit Discov*. 2022;1(3):226-40.
<https://doi.org/10.1039/D1DD00028D>.

Desai S, Jain M, Addamane SJ, Adams DP, Dingreville R, DeRio FW, et al. Navigating high-dimensional process-structure–property relations in nanocrystalline Pt-Au alloys with machine learning. *Mater Des*. 2024;248:113494.
<https://doi.org/10.1016/j.matdes.2024.113494>.

Arjovsky M. Out of distribution generalization in machine learning [dissertation]. New York: New York University; 2020.

Bhat N, Birbilis N, Barnard AS. Unsupervised learning and pattern recognition in alloy design. *Digit Discov*. 2024;3(12):2396-416.
<https://doi.org/10.1039/D4DD00282B>.

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat Commun*. 2020;11(1):5966.
<https://doi.org/10.1038/s41467-020-19597-w>.

Xu H, Nakayama R, Kimura T, Shimizu R, Ando Y, Kobayashi S, et al. Tuning Bayesian optimization for materials synthesis: Simulating two- and three-dimensional cases. *Sci Technol Adv Mater Methods*. 2023;3(1):2210251.
<https://doi.org/10.1080/27660400.2023.2210251>.

Will-Cole AR, Kusne AG, Tonner P, Dong C, Liang X, Chen H, et al. Application of Bayesian optimization and regression analysis to ferromagnetic materials development. *IEEE Trans Magn*. 2022;58(1):1-8.
<https://doi.org/10.1109/TMAG.2021.3125250>.

Diwale S, Eisner MK, Carpenter C, Sun W, Rutledge GC, Braatz RD. Bayesian optimization for material discovery processes with noise. *Mol Syst Des Eng.* 2022;7(6):622-36.
<https://doi.org/10.1039/D1ME00154J>.