ORIGINAL RESEARCH **Open access**

Defining “Chemical Space Coverage” for Generative Materials Models: A Boundary Problem for Diversity Metrics

Victor Hugo[1*], Daniel Cruz[1], Javier Salazar[2]

Abstract

Generative materials models promise to accelerate discovery by systematically exploring vast regions of chemical space, yet the core concept of “chemical space coverage” remains poorly defined and inconsistently applied across the literature. Researchers routinely claim that their models achieve high “diversity” or “coverage,” but these statements rest on incompatible assumptions about what chemical space actually encompasses. This boundary/definitional article clarifies the term by distinguishing three primary dimensions—compositional space, structural space, and property space—and demonstrates how current diversity metrics conflate or ignore these dimensions, rendering cross-study comparisons unreliable. Drawing on recent advances in generative modelling for crystals and molecules, the analysis shows that a model may report excellent elemental coverage while entirely neglecting novel crystal prototypes or property combinations, or conversely achieve broad structural diversity within a narrow compositional slice. To resolve these ambiguities, the article proposes an operational definition of chemical space coverage built around four explicit, computable metrics: compositional coverage (C_{comp}), structural coverage (C_{struct}), property coverage (C_{prop}), and joint coverage (C_{joint}). Each metric is accompanied by practical boundary conditions that define thresholds for “broad,” “comprehensive,” or “exploratory” coverage. The framework further articulates five essential boundary conditions for sufficiency—task dependence, reference dependence, sparsity adjustment, validity trade-off, and diminishing returns—thereby transforming coverage from a vague aspirational term into a precise evaluative criterion. Adoption of this multi-dimensional framework will enable consistent benchmarking, prevent over-optimistic claims, and guide the responsible development of generative models that truly expand the frontiers of materials design rather than merely resampling known regions. The proposed definitions and boundaries therefore constitute a necessary foundation for the next generation of inverse design methodologies in computational materials engineering.

Keywords Inverse design, Generative materials models, Chemical space coverage, Diversity metrics, Compositional space, Structural space

*Correspondence:
Victor Hugo
victor.hugo@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, National University of Colombia, Bogota, Colombia
[2] Department of Materials Data Analytics, Faculty of Technology, University of Antioquia, Medellin, Colombia

Introduction

Generative models for materials claim to explore “chemical space” and generate “diverse” structures [1-3]. But what does “chemical space coverage” actually mean? Does it mean generating many different elements? Many different crystal prototypes? Many different properties? The term is used vaguely, and diversity metrics vary widely across papers. A model that claims 90% coverage by one metric may have 10% coverage by another. This paper provides a

boundary/definitional analysis of chemical space coverage for generative materials models, distinguishing compositional, structural, and property coverage, and proposing operational definitions.

Recent years have witnessed an explosion of deep generative approaches applied to inorganic crystals, organic molecules, and hybrid materials [4-6]. These models are routinely described as “navigating chemical space,” “sampling unseen regions,” or “maximising diversity” [7-9]. Yet the very concept they claim to optimise remains operationally undefined. Without a shared boundary definition, it is impossible to determine whether one model genuinely covers more chemical space than another or whether reported gains are merely artefacts of different evaluation protocols. The present work therefore treats “chemical space coverage” as a boundary problem rather than a performance metric.

The boundary perspective is essential because chemical space is not a single entity but a composite of multiple interdependent subspaces. A generative model may saturate one subspace while barely touching another, yet still receive high marks under prevailing evaluation practices [10, 11]. For instance, a model trained only on binary oxides may generate thousands of distinct compositions within that slice yet fail to venture into ternary or quaternary regimes [12]. Conversely, a model that enumerates every known space group within a narrow elemental window may appear structurally diverse while remaining compositionally trivial [13]. These discrepancies are not minor technicalities; they undermine the central promise of inverse design—that generative models can discover materials outside the training distribution.

Current diversity metrics exacerbate the problem. Some papers rely on simple counts of unique elements or space groups [14, 15]. Others employ geometric descriptors such as pairwise distances in SOAP or graph-based embeddings [16, 17]. Still others focus exclusively on the spread of predicted properties [18, 19]. Each choice implicitly privileges one dimension of chemical space and neglects the others. As a result, claims of “high coverage” are incomparable across studies, and the field lacks a common language for assessing progress toward the goal of comprehensive materials exploration.

Figure 1 organises the manuscript’s core argument as a hierarchical framework, showing how ambiguous usage of “chemical space coverage” is resolved into explicit dimensions, operational metrics, boundary conditions, and reporting consequences.

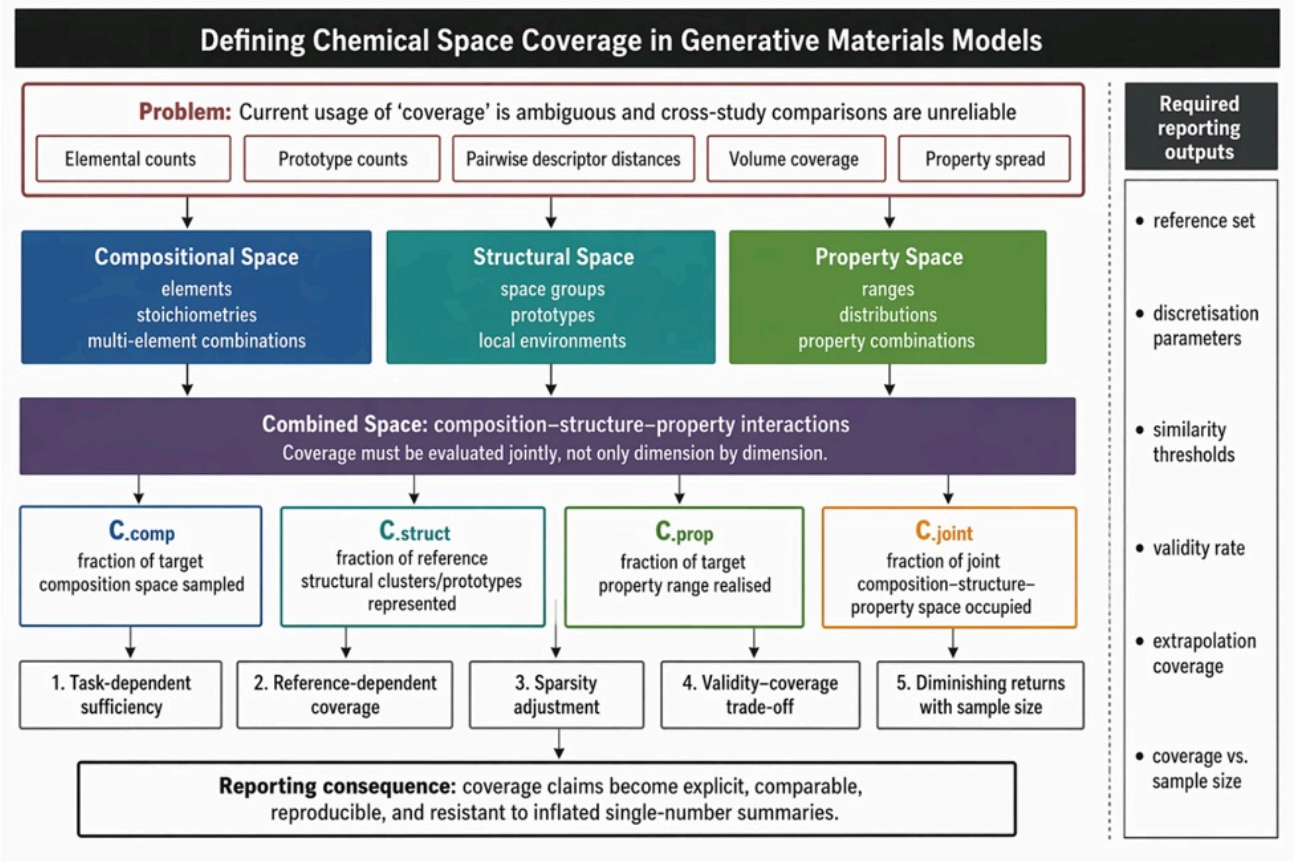

Figure 1. Hierarchical framework for defining and evaluating chemical space coverage in generative materials models

The analysis remains strictly conceptual: no new datasets, no simulations, and no empirical benchmarks are introduced. Instead, the work synthesises existing literature to expose logical inconsistencies and to construct a coherent definitional framework. By establishing clear boundaries around what counts as coverage, what counts as sufficient coverage, and what counts as meaningful diversity, the article equips the community with the conceptual tools needed to evaluate generative materials models rigorously and reproducibly. The ultimate aim is not to prescribe a single numerical score but to make explicit the assumptions that have hitherto remained implicit, thereby enabling more honest, comparable, and ultimately more productive claims about how far generative models have truly ventured into chemical space.

Table 1 clarifies why current coverage claims are often incomparable by separating the dominant usages in the literature from the dimensions they capture, neglect, or systematically distort.

Table 1. Competing usages of chemical space coverage in generative materials research and the specific dimensional blind spots they introduce

Dominant usage in the literature	What is actually being measured	Dimension primarily captured	Dimensions neglected or underrepresented	Why the usage is analytically insufficient	Typical interpretive error produced
Elemental coverage	Fraction of elements represented in	Compositional space at a very coarse level	Stoichiometry, oxidation state, coordination environment, structure, property	Presence of many elements does not imply meaningful combinatorial or	Mistaking broad elemental inclusion for broad

	generated outputs			functional exploration	chemical-space exploration
Prototype or space-group coverage	Number of distinct prototypes or symmetry classes generated	Structural space	Composition-specific variation, local chemical environment, emergent properties	Raw counts treat all prototypes as equally informative and ignore near-equivalence or clustering	Equating structural labels with genuine structural diversity
Pairwise descriptor distance	Average or median distance in SOAP, RDF, graph, or embedding space	Local geometric dissimilarity	Global dispersion, clustering structure, reference dependence, property relevance	High pairwise distance can coexist with poor global reach or heavy redundancy	Mistaking local separation for global coverage
Volume coverage in discretised space	Fraction of bins or cells occupied in a chosen space	Usually compositional or low-dimensional mixed space	Sensitivity to discretisation choice, sparsity structure, validity, task relevance	Coverage percentage is highly unstable across binning schemes and dimensional scales	Treating discretisation-dependent occupancy as an objective measure
Property diversity or range	Spread of predicted or computed property values	Property space	Composition novelty, structural novelty, reliability of surrogate predictions	Broad ranges may be driven by a few outliers and do not imply systematic exploration	Mistaking property spread for coverage of chemical space as a whole
Training-set distance / novelty proxy	Distance from generated samples to known training examples	Extrapolation relative to reference data	Internal diversity, joint coverage, validity, task-specific usefulness	Being far from training data is not equivalent to covering a meaningful region of feasible materials space	Equating novelty with coverage
Validity rate alone	Fraction of generated structures passing filters	Feasibility or plausibility, not coverage itself	Breadth of explored space, novelty, dimensional reach	Validity is a necessary companion metric but cannot stand in for coverage	Mistaking physically plausible generation for broad exploration

Current Usage and Confusions

Elemental coverage often serves as a convenient proxy for chemical space exploration [2, 14], defined simply as the fraction of periodic table elements appearing in generated structures, as when authors note that “Our model generates materials with 30 different elements” [15]. Its appeal lies in immediate interpretability, yet this metric entirely neglects stoichiometry, oxidation states, and coordination environments; a model can thus achieve complete elemental coverage through repeated unary phases or minor dopant variations of a single binary compound, yielding negligible advance in genuine compositional diversity [12].

Shifting emphasis to structural variety, prototype coverage instead tallies distinct crystal prototypes or space groups [4, 13], with claims such as “Generated structures span 50 space groups” [16]. Although this acknowledges architectural diversity, it assumes equal separation among all prototypes, despite closer similarity between two monoclinic structures than between monoclinic and cubic ones, while overlooking how identical prototypes can produce markedly different properties across compositions [17].

A further evolution employs pairwise distance metrics within continuous descriptor spaces—using SOAP, RDF, or graph kernels—to quantify local dissimilarity, often summarized as “Average pairwise SOAP distance = 0.8” [5, 10, 11]. Such approaches still leave global scale undefined and remain insensitive to clustering, so that tightly grouped yet separated clusters register the same average distance as uniformly dispersed points, obscuring true coverage extent [18].

Volume-based coverage attempts a more comprehensive discretization of composition space, reporting fractions such as “Model covers 20% of ternary composition space” [1, 3, 19]; however, exponential growth in high-dimensional volumes renders uniform sampling intractable beyond ternaries, while arbitrary bin sizes produce inconsistent percentages for identical structure sets [20]. Property diversity, in turn, examines ranges of computed attributes like band gap or formation energy [8, 21, 22], yet remains downstream of structure and composition, vulnerable to outlier dominance and surrogate model inaccuracies [23].

Across these usages, “coverage” therefore denotes fundamentally different phenomena, allowing models to excel on one axis while faltering on others and rendering cross-paper comparisons anecdotal rather than cumulative [1, 2, 24]. Without explicit decomposition of chemical space subspaces and standardized operational metrics, progress in generative materials design risks remaining fragmented.

Dimensions of Chemical Space

Compositional space encompasses all feasible combinations of elements and their stoichiometric ratios, forming a high-dimensional, continuous yet inherently sparse domain [2, 12]. Coverage in this dimension therefore extends beyond mere elemental presence to interrogate which specific ratios and multi-element pairings are realised, through metrics such as element frequency distributions, co-occurrence matrices, and occupied fractions of the composition simplex. The factorial growth of this space with increasing elements renders exhaustive enumeration impossible, even for ternaries, compelling any meaningful metric to accommodate sparsity without penalising models for unexplored ratios [25].

Once composition is fixed, structural space addresses the geometric arrangements of atoms, integrating discrete features like space groups and prototypes with continuous parameters such as lattice constants and fractional coordinates [4, 13, 16]. Here, coverage demands not simple enumeration of distinct prototypes or coordination environments via bond-length histograms and polyhedral motifs, but similarity-weighted assessments that avoid inflating counts with near-duplicates, recognising that minor positional differences may be functionally inconsequential while identical space groups can mask profound local chemical variations [17, 26].

These structural and compositional realisations give rise to property space, the manifold of emergent physical and chemical attributes whose distributions—ranges, shapes, and novel combinations—ultimately matter for discovery [8, 21, 22]. Yet properties remain downstream, contingent on computational or experimental resolution, so that varying levels of theory can reshape observed coverage while identical property values may arise from structurally distinct origins, underscoring the limitations of property-focused evaluation in isolation [23].

True coverage therefore cannot be assessed marginally across these subspaces, as generative models may span diverse compositions and prototypes yet remain trapped within narrow property regimes, or achieve broad property ranges while recycling a single compositional family such as perovskites [7, 27]. Only joint statistics—mutual information between dimensions and realised fractions of composition–structure–property triplets—reveal whether models have

genuinely expanded the materials universe or merely reorganised known regions, completing a framework that demands simultaneous, interdependent evaluation rather than isolated claims.

Proposed Coverage Metrics

Compositional coverage quantifies the fraction of a target composition space sampled through discretisation into fixed-concentration bins, defined as the ratio of occupied to total bins within fixed-element subspaces, where values exceeding 0.5 indicate broad coverage [2, 12, 19]. This formulation remains computationally tractable yet anchors compositional claims with clarity and reproducibility, even as its restriction to selected subspaces acknowledges the prohibitive scaling across arbitrary element counts.

Structural coverage advances this logic by measuring the fraction of prototype space sampled under similarity-weighted clustering, employing measures such as SOAP or graph-edit distance to derive the ratio of represented clusters against a reference library, with thresholds above 0.3 denoting diversity [4, 13, 16]. The clustering mechanism prevents inflation from near-duplicates, delivering a more faithful assessment of structural variety than unweighted counts, although it inherits dependence on the chosen reference and similarity criteria.

Property coverage subsequently evaluates the realised fraction of target property ranges via the normalised span between predicted extrema, offering a direct index of functional exploration in materials design [8, 21, 22]. Sensitive to the definition of reference bounds—particularly for emergent material classes—this metric privileges utility over isolated structural or compositional extent.

Joint coverage completes the framework by discretising the combined composition–structure–property space into hypercubes and computing the occupied fraction, thereby exposing whether models achieve meaningful intersections rather than merely additive marginal coverage, despite dimensionality constraints that favour low-dimensional projections [1, 7, 27]. Requiring all four values alongside explicit parameters replaces ambiguous claims of “high coverage” with precise, comparable statements that distinguish genuine expansion of the known materials universe from superficial reorganisation.

Boundary Conditions for “Sufficient” Coverage

Task-dependent sufficiency underscores that coverage thresholds derive meaning only from the generative model’s intended application: compositional coverage dominates when seeking novel stoichiometries, structural coverage when uncovering new mechanisms, and property coverage when optimising functional performance [3, 9, 24]. Sufficiency must therefore be framed relative to downstream goals rather than asserted in absolute terms.

A related imperative is reference-dependent coverage, which insists that every numerical claim explicitly define the baseline—whether the training distribution, the set of all known materials, or a theoretical enumeration—since impressive interpolation within training data can masquerade as genuine exploration when measured against looser references [1, 20, 25].

Beyond these anchors, sparsity-adjusted coverage recognises the intrinsic sparsity of high-dimensional chemical space, requiring that reported fractions be contextualised against the coverage expected from an equal number of uniform random samples drawn from the same reference [10, 28]. This comparison distinguishes meaningful progress from statistical inevitability.

Coverage must further be evaluated alongside validity, as models that achieve high breadth by populating composition space with predominantly unstable or non-physical structures deliver little practical advance; joint reporting of coverage and validity therefore exposes the true cost of breadth [5, 11, 29]. Finally, diminishing returns dictate that coverage be

presented as a function of sample size, revealing the power-law saturation typical of such explorations and allowing assessment of whether further scaling remains productive [6, 26].

Together these boundary conditions convert coverage from an unbounded aspiration into a bounded, context-sensitive instrument. Applied alongside the operational metrics, they establish a rigorous yet adaptable standard for assessing how genuinely generative models expand the materials universe.

Table 2 consolidates the manuscript's theoretical contribution by linking each proposed metric to its governing boundary conditions, characteristic failure mode, and minimum reporting requirement.

Table 2. Integrative framework linking coverage metrics, sufficiency conditions, gray-zone failures, and mandatory reporting outputs

Metric	Formal evaluative question	Primary sufficiency condition(s)	Most likely gray-zone failure if used alone	What must be reported alongside it	Why it matters theoretically
Compositional coverage (C_{comp})	How much of the defined composition space has been sampled?	Task dependence; reference dependence; sparsity adjustment	Apparent breadth driven by coarse elemental variety while stoichiometric richness remains narrow	Reference composition space, discretisation scheme, bin size, fixed element-count subspace, random baseline	Establishes whether the model truly explores compositional possibilities rather than merely listing many elements
Structural coverage (C_{struct})	How much of the reference prototype or structural-cluster space is represented?	Reference dependence; validity trade-off	Inflated diversity due to near-duplicate structures or raw prototype counting	Prototype library, clustering rule, similarity threshold, validity-filter definition	Distinguishes true structural reach from repetition in symmetry-labelled space
Property coverage (C_{prop})	How much of the target property range or distribution is realised?	Task dependence; reference dependence	Wide range caused by outliers or surrogate-model artefacts rather than systematic exploration	Property bounds, prediction method, uncertainty statement, whether single-property or multi-property reporting is used	Connects coverage to downstream functional relevance rather than structural appearance alone
Joint coverage (C_{joint})	How much of the combined composition–structure–property space is occupied?	Task dependence; sparsity adjustment; reference dependence	Misleadingly high marginal coverage masking narrow occupation of meaningful combinations	Joint-space definition, dimensional projection choice, hypercube resolution, feasibility constraints	Provides the strongest test of whether a model expands the materials universe rather than reshuffling marginals
Validity rate	What fraction of generated samples are	Validity trade-off	High reported coverage despite	Crystallographic filters, energetic filters, stability criteria	Prevents breadth from being mistaken for

	physically or chemically plausible?		non-physical outputs		scientifically useful exploration
Extrapolation coverage	What fraction of generated samples lie meaningfully beyond the training distribution?	Reference dependence; task dependence	Novelty claims based only on interpolation within familiar regions	Descriptor choice, nearest-neighbour rule, extrapolation threshold	Separates genuine out-of-distribution exploration from resampling of known regions
Coverage versus sample size	Does coverage continue to grow meaningfully as generation scales?	Diminishing returns	Static one-number claims that hide saturation or inefficiency	Curve or benchmark points, sample counts, generation protocol held constant	Reveals whether additional generation yields real exploration or only redundant accumulation

Boundary Cases and Gray Zones

A particularly insidious failure mode arises when generative models achieve high coverage of easy, low-energy regions such as common high-symmetry families while neglecting rare or high-energy prototypes, producing apparently strong structural scores through hundreds of cubic perovskites yet zero triclinic structures critical for exotic properties [4, 13, 16, 26]. In such cases the aggregate metric appears robust, yet genuine diversity remains shallow because the model merely saturates the simplest subspaces of structural space.

A related pitfall occurs when models remain confined to the training distribution, generating structures that register high pairwise distances and descriptor-space variety yet constitute rearrangements of known composition–structure pairs rather than true extrapolation [1, 7, 20, 27]. Here, separate reporting of out-of-distribution fractions—defined by distance to the nearest training example exceeding a chosen threshold—becomes essential to isolate genuine exploration from disguised interpolation.

This limitation intensifies with redundancy, where thousands of outputs collapse into a few tight clusters; pairwise metrics may still suggest respectable diversity while global reach stays severely restricted [5, 10, 11, 17, 18]. Cluster-aware evaluation is therefore required to penalise local over-sampling explicitly.

Finally, coverage of unstable structures occupies a subtle boundary: models may span many compositions and prototypes yet yield only positive formation energies or imaginary phonon modes, inflating scores without populating the thermodynamically accessible domain [2, 8, 12, 21, 23, 29]. Stratifying coverage into stable and overall subsets clarifies utility relative to the modelling objective, whether mapping stability landscapes or identifying synthesizable targets.

These boundary cases demonstrate that raw coverage numbers mislead without contextual stratification by difficulty, extrapolation depth, redundancy, and stability; only such layered analysis prevents over-interpretation and ensures claims reflect substantive advancement rather than model biases or evaluation artefacts.

Relation to Other Concepts

Relation to diversity in generative models builds directly on earlier reviews of diversity metrics [5, 10]. Those reviews catalogued goal-directed generators and similarity measures, yet stopped short of supplying operational definitions tied to

the multi-dimensional nature of chemical space. The present framework extends that work by adding explicit, dimension-specific coverage metrics that can be computed uniformly and compared across studies, thereby converting diversity from a qualitative descriptor into a set of bounded, reportable quantities.

Relation to chemical space in drug discovery highlights a telling contrast. In molecular design, chemical space is comparatively well delineated by graph-based enumerations and fingerprint similarity [18, 25]. Crystal chemical space, however, is more complex because periodicity, space-group constraints, and variable stoichiometry introduce additional discrete and continuous degrees of freedom [4, 13, 16]. The framework therefore adapts lessons from molecular chemical space while explicitly addressing the extra layers of complexity unique to periodic solids.

Relation to sparsity acknowledges that high-dimensional chemical spaces are intrinsically sparse [1, 20, 28]. Random sampling baselines yield vanishingly small coverage fractions, so modest absolute coverage can still represent meaningful progress. The proposed metrics incorporate sparsity adjustment by requiring comparison against a random-sampling reference, ensuring that reported values are interpreted in light of the underlying dimensionality rather than against an unrealistic uniform benchmark.

Relation to extrapolation clarifies that coverage of regions far from the training set is fundamentally harder than interpolation [7, 27]. Earlier discussions of out-of-distribution performance in generative models noted this distinction but lacked quantitative coverage tools. The framework resolves the gap by demanding separate reporting of extrapolation coverage, thereby making explicit whether a model has merely reorganised known data or has truly ventured into uncharted chemical territory.

Taken together, these relations embed the new coverage definitions within the broader literature while sharpening their applicability to materials-specific challenges. The result is a conceptual bridge between existing diversity concepts, chemical-space thinking in adjacent fields, sparsity considerations, and extrapolation requirements.

Implications for Generative Model Evaluation

For model developers the framework imposes a clear reporting obligation: every publication must supply coverage scores across all four metrics—compositional, structural, property, and joint—together with the chosen reference sets and discretisation parameters [1-3]. Developers can no longer rely on a single headline diversity number; instead, they must demonstrate balanced exploration across dimensions. This requirement discourages cherry-picking of favourable metrics and encourages architectural choices that genuinely expand chemical space rather than overfit narrow subspaces [4, 13, 16].

For benchmark designers the implications are equally direct. Existing generative-model benchmarks must incorporate the four coverage metrics as core evaluation criteria rather than optional add-ons [5, 10, 20]. Benchmark suites should also furnish standardised reference chemical spaces—curated prototype libraries, composition grids, and property ranges—so that coverage numbers become directly comparable across competing models. Without such standardisation, benchmark results will continue to reflect evaluation artefacts more than scientific progress [11, 17, 26].

For practitioners who deploy generative models in real discovery campaigns the framework supplies decision-making clarity. Coverage metrics allow rapid assessment of whether a model explores the precise regions required for a given application—novel compositions for high-entropy alloys, novel prototypes for metastable phases, or novel property combinations for optoelectronics [8, 21, 22]. Practitioners can therefore reject models that achieve high marginal coverage yet fail on joint coverage, avoiding wasted synthesis effort on redundant or unstable candidates. The multi-dimensional view also guides iterative refinement: if compositional coverage is adequate but structural coverage lags, attention can shift to prototype-aware training objectives [7, 27, 29].

Collectively, these implications shift generative-model evaluation from an ad-hoc, metric-of-the-month practice to a structured, transparent standard. Model developers, benchmark designers, and practitioners gain a common language that replaces vague “high diversity” claims with precise, multi-faceted coverage profiles. The ultimate outcome is more reliable literature, faster scientific consensus, and accelerated translation of generative ideas into laboratory-validated materials.

Proposed Reporting Standard

The required coverage report for generative materials papers adopts a compact but explicit format designed to enforce transparency and comparability. Rather than relying on a single aggregate indicator, authors must report coverage across four distinct dimensions: compositional, structural, property, and joint space. Each dimension is expressed through a clearly defined metric and evaluated relative to a specified reference space.

For the **compositional dimension**, the metric C_{comp} is reported as the percentage of a defined compositional reference space that is occupied by generated samples. The reference space may correspond to the training distribution, a curated database, or a discretised compositional grid, but must be explicitly stated. Authors must additionally indicate whether the reported value satisfies a predefined threshold for sufficiency.

For the **structural dimension**, the metric C_{struct} is defined as the fraction of reference structural prototypes or similarity-based clusters that are represented in the generated set. The construction of the reference prototype library and the similarity threshold used for clustering must be disclosed. As with compositional coverage, a binary threshold assessment must be provided alongside the reported value.

For the **property dimension**, the metric C_{prop} quantifies the proportion of a specified target property range that is realised by generated samples. Target ranges must be justified, either by reference to known materials, application-driven constraints, or physically plausible limits. Where multiple properties are considered, authors must clarify whether coverage is reported per property or jointly across properties.

For the **joint dimension**, the metric C_{joint} captures coverage over the combined composition–structure–property space. This is operationalised by discretising the joint space into multi-dimensional regions and reporting the fraction that is occupied by generated samples. Because of the high dimensionality involved, authors may restrict analysis to well-defined subspaces, but such restrictions must be clearly specified.

In addition to these four primary metrics, three supplementary quantities must be reported. First, the validity rate of generated structures must be provided as the percentage of samples that satisfy basic crystallographic and energetic constraints [5, 11, 29]. Second, extrapolation coverage must be quantified as the fraction of generated samples whose descriptor-space distance from the training set exceeds a defined threshold, thereby distinguishing genuine out-of-distribution exploration from interpolation [7, 27]. Third, authors must report coverage as a function of sample size, either as a curve or through representative points (for example, after 10k and 100k generated samples), in order to reveal diminishing returns in exploration efficiency [6, 26].

The reporting standard further requires full disclosure of all methodological parameters used in computing coverage metrics. This includes discretisation schemes, bin sizes, similarity thresholds, clustering algorithms, and the specific reference datasets or libraries employed. Such transparency eliminates hidden degrees of freedom that currently permit divergent coverage claims for identical sets of generated structures [1, 20, 25].

To ensure consistency across the field, journals and conferences are encouraged to require this structured coverage report as a mandatory component of the Methods section or supplementary materials. The format is intentionally lightweight and does not impose a substantial additional burden on authors. However, it substantially improves the interpretability, reproducibility, and comparability of reported results.

By replacing ambiguous claims of “high diversity” with explicit, multi-dimensional coverage metrics, the proposed standard enables reviewers to evaluate generative models against a consistent checklist and allows readers to understand precisely which regions of chemical space have been explored. More broadly, it supports cumulative scientific progress by ensuring that improvements in reported coverage reflect genuine expansion of chemical space rather than artefacts of metric selection. In this way, the reporting standard operationalises the boundary definitions developed in this work, translating conceptual clarity into routine scientific practice.

Conclusion

“Chemical space coverage” is vaguely defined. Three dimensions—compositional, structural, and property—must be distinguished before any coverage claim can be evaluated meaningfully. Current diversity metrics conflate these dimensions and render cross-study comparisons unreliable. The present boundary/definitional analysis resolves the ambiguity by proposing four operational metrics: compositional coverage (C_{comp}), structural coverage (C_{struct}), property coverage (C_{prop}), and joint coverage (C_{joint}). Each metric is accompanied by explicit boundary conditions that define thresholds for broad, comprehensive, or exploratory coverage.

Five essential boundary conditions further qualify sufficiency: task dependence, reference dependence, sparsity adjustment, coverage–validity trade-off, and diminishing returns. Boundary cases—easy versus hard regions, interpolation versus extrapolation, redundancy, and unstable structures—illustrate how raw numbers can mislead unless contextualised. The framework relates cleanly to existing diversity concepts, chemical-space thinking in drug discovery, sparsity considerations, and extrapolation requirements, while extending them specifically to crystalline materials.

Implications for model developers, benchmark designers, and practitioners are immediate and practical: report multi-dimensional coverage, specify references, and present coverage as a function of sample size. The proposed reporting standard translates these requirements into a compact, mandatory table that enforces transparency without added burden.

Community adoption of these multi-dimensional coverage metrics will replace anecdotal claims with cumulative, comparable knowledge. Generative materials models can then be judged not by how loudly they proclaim “diversity” but by how rigorously they document which parts of chemical space they have actually opened. The definitions and boundaries offered here therefore constitute a necessary foundation for the next generation of trustworthy inverse design in computational materials engineering.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Apr 2025 Revised: 11 Aug 2025 Accepted: 20 Nov 2025
Published online: 18 January 2026

References

Zhang J, Mercado R, Engkvist O, Chen H. Comparative study of deep generative models on chemical space coverage. J Chem Inf Model. 2021;61(6):2572-81.
https://doi.org/10.1021/acs.jcim.0c01328.

Türk H, Landini E, Kunkel C, Margraf JT, Reuter K. Assessing deep generative models in chemical composition space. Chem Mater. 2022;34(21):9455-67.
https://doi.org/10.1021/acs.chemmater.2c01860.

Yue T, Tao L, Varshney V, Li Y. Benchmarking study of deep generative models for inverse polymer design. Digit Discov. 2025;4(4):910-26.
https://doi.org/10.1039/D4DD00395K.

Qiu Z, Jin L, Du Z, Chen H, Mao G, Cen Y, et al. Massive discovery of crystal structures across dimensionalities by leveraging vector quantization. npj Comput Mater. 2025;11(1):184.
https://doi.org/10.1038/s41524-025-01613-6.

Renz P, Luukkonen S, Klambauer G. Diverse hits in de novo molecule design: Diversity-based comparison of goal-directed generators. J Chem Inf Model. 2024;64(15):5756-61.
https://doi.org/10.1021/acs.jcim.4c00519.

Azzouzi M, Worakul T, Corminboeuf C. NaviDiv: A web app for monitoring chemical diversity in generative molecular design. Digit Discov. 2026;5:1579-89.
https://doi.org/10.1039/D5DD00487J.

Luo X, Wang Z, Wang Q, Shao X, Lv J, Wang L, et al. CrystalFlow: A flow-based generative model for crystalline materials. Nat Commun. 2025;16(1):9267.
https://doi.org/10.1038/s41467-025-64364-4.

Zagribelnyy B, Aladinskiy V, Bondarev N, Ilin I, Malkov M, Vasileva A, et al. Molecular LEGION: Incalculably large coverage of chemical space around the NLRP3 target. Sci Data. 2026;13(1):576.
https://doi.org/10.1038/s41597-026-06850-y.

Gao W, Luo S, Coley CW. Generative AI for navigating synthesizable chemical space. Proc Natl Acad Sci U S A. 2025;122(41):e2415665122.
https://doi.org/10.1073/pnas.2415665122.

Hu X, Liu G, Yao Q, Zhao Y, Zhang H. Hamiltonian diversity: Effectively measuring molecular diversity by shortest Hamiltonian circuits. J Cheminform. 2024;16(1):94.
https://doi.org/10.1186/s13321-024-00883-4.

Mrugalla F, Franz C, Alber Y, Mogk G, Villalba M, Mrziglod T, et al. Generating diversity and securing completeness in algorithmic retrosynthesis. J Cheminform. 2025;17(1):72.
https://doi.org/10.1186/s13321-025-00981-x.

Shi Y, von Itzstein M. How size matters: Diversity for fragment library design. Molecules. 2019;24(15):2838.
https://doi.org/10.3390/molecules24152838.

Park H, Li Z, Walsh A. Has generative artificial intelligence solved inverse materials design? Matter. 2024;7(7):2355-67.
https://doi.org/10.1016/j.matt.2024.05.017.

Noh J, Kim J, Stein HS, Sanchez-Lengeling B, Gregoire JM, Aspuru-Guzik A, et al. Inverse design of solid-state materials via a continuous representation. Matter. 2019;1(5):1370-84.
https://doi.org/10.1016/j.matt.2019.08.017.

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 design of inorganic materials. npj Comput Mater. 2020;6(1):84.
https://doi.org/10.1038/s41524-020-00352-0.

Ren Z, Tian SIP, Noh J, Oviedo F, Xing G, Li J, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. Matter. 2022;5(1):314-35.
https://doi.org/10.1016/j.matt.2021.11.032.

Medina-Franco JL, Chávez-Hernández AL, López-López E, Saldívar-González FI. Chemical multiverse: An expanded view of chemical space. Mol Inform. 2022;41(11):e2200116.
https://doi.org/10.1002/minf.202200116.

Coley CW. Defining and exploring chemical spaces. Trends Chem. 2021;3(2):133-45.
https://doi.org/10.1016/j.trechm.2020.11.004.

Medina-Franco JL, Sánchez-Cruz N, López-López E, Díaz-Eufracio BI. Progress on open chemoinformatic tools for expanding and exploring the chemical space. J Comput Aided Mol Des. 2022;36(5):341-54.
https://doi.org/10.1007/s10822-021-00399-1.

Szymanski NJ, Bartel CJ. Establishing baselines for generative discovery of inorganic crystals. Mater Horiz. 2025;12(19):8000-11.
https://doi.org/10.1039/D5MH00010F.

Zhao Y. Deep learning based generative materials design [dissertation]. University of South Carolina; 2022.

Parida C, Roy D, Lastra JMG, Bhowmik A. Mining chemical space with generative models for battery materials. Batter Supercaps. 2026;9(1):e202500309.
https://doi.org/10.1002/batt.202500309.

Wei G, Huang Y, Duan C, Song Y, Du Y. Navigating chemical space with latent flows. Adv Neural Inf Process Syst. 2024;37:58663-97.

Park H, Onwuli A, Walsh A. Exploration of crystal chemical space using text-guided generative artificial intelligence. Nat Commun. 2025;16(1):4379.
https://doi.org/10.1038/s41467-025-59636-y.

Lavecchia A. Navigating the frontier of drug-like chemical space with cutting-edge generative AI models. Drug Discov Today. 2024;29(9):104133.
https://doi.org/10.1016/j.drudis.2024.104133.

De Breuck PP, Wang HC, Rignanese GM, Botti S, Marques MAL. Generative AI for crystal structures: A review. npj Comput Mater. 2025;11(1):370.

https://doi.org/10.1038/s41524-025-01881-2.

Nguyen TM, Truong TN. Inverse design using goal-conditioned reinforcement learning for organic semiconductor materials from benzene and thiophene-based polycyclic aromatic compounds. npj Comput Mater. 2025;11(1):372. https://doi.org/10.1038/s41524-025-01852-7.

Kneiding H, Morán-González L, Kuriakose N, Nova A, Balcells D. Inverse design of inorganic compounds with generative AI. arXiv preprint arXiv:2604.11827. 2026. https://doi.org/10.48550/arXiv.2604.11827.

Wang J, Wang Y, Chen Y. Inverse design of materials by machine learning. Materials (Basel). 2022;15(5):1811. https://doi.org/10.3390/ma15051811.