How Far Can You Grow? Characterizing the Extrapolation Frontier of Graph Generative Models for Materials Science

Can Polat
can.polat@tamu.edu
Texas A&M University
College Station, USA

Erchin Serpedin
eserpedin@tamu.edu
Texas A&M University
College Station, USA
Texas A&M University at Qatar
Doha, Qatar

Mustafa Kurban*
kurbanm@ankara.edu.tr
Ankara University
Ankara, Turkiye
Texas A&M University at Qatar
Doha, Qatar

Hasan Kurban*
hkurban@hbku.edu.qa
Hamad Bin Khalifa University
Doha, Qatar

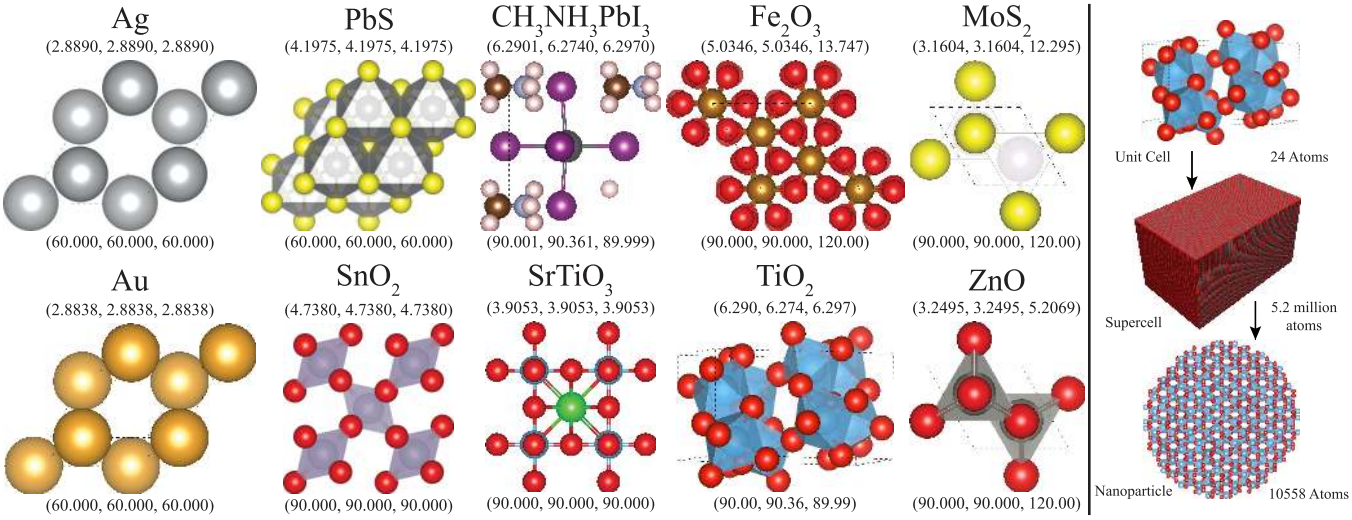

Figure 1: Detailed overview of RADII. (left) Unit cells of the materials (Ag, Au, $CH_3NH_3PbI_3$, Fe_2O_3, MoS_2, PbS, SnO_2, $SrTiO_3$, TiO_2, and ZnO) arranged left to right. The lattice constants a, b and c denote the cell edge lengths, while the angles α, β and γ specify the inter-edge angles between bc, ac and ab, respectively. (right) Workflow for generating radius-resolved nanoparticles.

Abstract

Every generative model for crystalline materials harbors a critical structure size beyond which its outputs quietly become unreliable; we call this the extrapolation frontier. Despite its direct consequences for nanomaterial design, this frontier has never been systematically measured. We introduce RADII, a radius-resolved benchmark of ~75,000 crystal-derived nanoparticle structures (33–11,298 atoms) that treats radius as a continuous scaling knob to trace generation quality from in-distribution to out-of-distribution regimes under leakage-free splits. Each model is conditioned on the target composition and atom count, isolating geometric extrapolation as the evaluation variable. RADII provides frontier-specific diagnostics: per-radius error profiles pinpoint each architecture's scaling ceiling, surface–interior decomposition tests whether failures originate at boundaries or in bulk, and cross-metric failure sequencing reveals which aspect of structural fidelity breaks first. Benchmarking five state-of-the-art architectures, we find that: (i) well-behaved models degrade

*Corresponding authors.

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818872

by ∼13% in global positional error beyond training radii, while divergent models exhibit poor absolute fidelity across scales; local bond fidelity varies sharply across architectures, from negligible degradation to more than 2× error growth; (ii) no two architectures share the same failure sequence, revealing the frontier as a multi-dimensional surface shaped by model family; and (iii) well-behaved models conform to the expected geometric scaling exponent $\alpha \approx 1/3$ whose in-distribution fit accurately predicts out-of-distribution error, making their frontiers quantitatively forecastable. Scaling MatterGen to its published parameter count stabilizes sampling but does not close the extrapolation frontier, while DiffCSP remains unstable even at published scale. These findings establish output scale as a first-class evaluation axis for geometric generative models. The dataset, generation pipeline, and implementations are available at https://github.com/KurbanIntelligenceLab/RADII.

CCS Concepts

• Applied computing → Chemistry; Physics.

Keywords

Crystal Generation, Benchmark, Quantum Chemistry, Equivariant Architectures, Graph Neural Networks

ACM Reference Format:
Can Polat, Erchin Serpedin, Mustafa Kurban, and Hasan Kurban. 2026. How Far Can You Grow? Characterizing the Extrapolation Frontier of Graph Generative Models for Materials Science. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3770855.3818872

1 Introduction

Generative models for crystalline materials are routinely evaluated at the scales on which they were trained, creating an illusion of reliability that shatters once output size departs from the training distribution. We refer to the critical size threshold at which this collapse occurs as the extrapolation frontier: a quantity that, despite its direct implications for nanomaterial design, has never been systematically measured. Nanostructured materials drive applications from photovoltaics to chemical sensing [59], with properties governed both by the periodic symmetry of primitive unit cells and by the finite morphologies of nanoparticles [7]. These regimes are conventionally treated separately: crystal simulations propagate ideal lattices, whereas nanoparticle workflows construct finite clusters through empirical or first-principles refinement [36]. Both approaches face steep scalability barriers across compositions, sizes, and orientations [61]. Density functional theory (DFT) [47] and its tight-binding approximation (DFTB) [17] deliver accurate energetics, but DFT's cubic scaling restricts exploration to modest system sizes [14]; DFTB reduces this cost yet remains limited for large-scale nanostructure generation [40, 53].

Machine learning offers a scalable alternative [27]. Graph neural networks [58] achieve strong predictive performance on molecular and crystalline benchmarks [11, 12]. Equivariant architectures [20, 42, 57] improve data efficiency through geometric symmetry constraints, and multimodal methods [50, 55] incorporate diverse input modalities. Generative models, including symmetry-preserving diffusion and flow approaches [28, 37], have advanced crystal structure generation considerably. However, existing evaluations focus predominantly on small molecules or bulk periodic crystals at roughly fixed output sizes. This evaluation paradigm conceals a fundamental failure mode: it remains unknown how far beyond their training distribution these models can reliably generate structures, and how generation quality degrades as output size increases.

RADII addresses this gap by systematically mapping the extrapolation frontier across architectures, materials, and metrics. Using radius as a continuous scaling knob, the benchmark links primitive unit cells extracted from published crystallographic data to nanoparticles across 25 size configurations spanning 6–30 Å, yielding approximately 75,000 structures containing 33–11,298 atoms. A leakage-free data split cleanly separates orientation interpolation (in-distribution, ID) from radius extrapolation (out-of-distribution, OOD), enabling precise identification of each model's scaling ceiling. Each generation task is conditioned on the target composition and atom count, so RADII isolates geometric extrapolation rather than conflating it with composition prediction or assignment ambiguity. For every model–material pair, RADII traces per-radius error profiles to locate the frontier, decomposes failures into surface versus interior contributions, and quantifies extrapolation gap severity across complementary metrics. Benchmarking state-of-the-art generative models reveals that all architectures degrade sharply, but the frontier location and failure signature vary systematically with material symmetry and architecture family. These findings establish output scale as a first-class evaluation axis for geometric generative models and demonstrate that current scaling limits are predictable rather than random. By providing a reproducible, geometry-grounded diagnostic testbed, RADII lays the foundation for developing architectures that generalize beyond their training horizon.

2 Related Work

2.1 Geometric Graph Generation for Materials

Generative modeling of atomic structures has advanced rapidly, yet nearly all progress has been measured at fixed output scales. Crystalline unit cells encode the symmetry operations and atomic motifs that generative models must reproduce [34], and when bulk crystals are truncated into finite clusters, their morphologies follow orientation-dependent surface energies governed by the Gibbs–Wulff theorem [2, 38, 54]. These geometric effects produce controlled, symmetry-consistent deviations from ideal bulk structure as system size varies, precisely the kind of structured distribution shift needed to characterize extrapolation frontiers. Quantum-confinement

phenomena [4] lie outside scope; RADII retains only the geometric variation needed to probe whether generation quality holds as output size departs from training conditions. By employing deterministic, symmetry-preserving construction rather than simulation-driven morphologies, the benchmark isolates scale as an independent variable for evaluating generative models [68].

Graph neural networks [58] and equivariant architectures [20, 42, 57] have achieved strong performance on molecular and crystalline benchmarks [11, 12], with multimodal methods [50, 55] further expanding input representations. Generative models, including diffusion-based approaches [24, 43], flow-matching methods [37], and symmetry-aware generators [28], have advanced crystal structure generation considerably. Recent symmetry-aware crystal structure prediction methods such as EquiCSP [39] and SGEquiDiff [10] further exploit space-group equivariance and Wyckoff-position priors to improve periodic structure generation. However, these periodic inductive biases assume translational symmetry that is explicitly broken in finite nanoparticles, and RADII's results empirically confirm that architectures designed under such assumptions can fail to scale to non-periodic clusters. More broadly, all these models are developed and validated on datasets where output size is approximately constant (e.g., QM9 at ∼9 atoms, MP-20 at ∼20 atoms per cell), meaning that their behavior under size extrapolation remains entirely uncharacterized. Unified geometric representation benchmarks such as Geom3D [41] have broadened evaluation across tasks and representations but likewise do not vary output scale. RADII is designed to fill exactly this diagnostic gap.

2.2 Scalability Limits of Physics-Based Methods

The extrapolation frontier matters in practice because physics-based alternatives cannot cover the size ranges that generative models are increasingly asked to target. Kohn–Sham DFT provides accurate energetics but scales as $\mathcal{O}(N^3)$, restricting routine simulations [5, 70]. Linear-scaling approaches such as ONETEP [1, 60], semi-empirical techniques like DFTB [30, 72], and classical or ML-based interatomic potentials [15, 44] extend this ceiling but cannot reliably generate structures approaching RADII's 11,298-atom upper bound. This computational bottleneck is precisely why ML-based generation is attractive for nanomaterial design, and why understanding where these models break under size extrapolation is urgent. Because RADII targets geometric scaling behavior rather than energetic accuracy, physics-based generation is both unnecessary and infeasible at benchmark scale; instead, deterministic symmetry-preserving construction provides the scalable ground truth needed to map extrapolation frontiers [35].

2.3 Evaluation Gaps in Existing Benchmarks

Benchmark datasets have driven geometric deep learning forward, yet none systematically probe size extrapolation. QM7 [6, 56], MD22 [13], PubChemQC [31], NablaDFT [29], and QH9 [69] target molecular energetics or Hamiltonians; Perov-5 [8, 9] and Carbon-24 [49] catalog crystalline frameworks; MatBench [16], OC20 [11], OC22 [62], LAMBench [48], and CrysMTM [51] benchmark property or force predictions on surface and bulk structures. All evaluate at approximately fixed output scale; none treats output size as a continuous evaluation axis, so extrapolation frontiers have never been measured. The broader OOD generalization literature for GNNs, including causality-based and augmentation-based approaches, underscores the importance of controlled distribution shifts; RADII provides exactly such a structured geometric shift and may serve as a complementary testbed for graph OOD methods. Power-law scaling relationships between error and system size are well-studied in neural scaling laws [26] and finite-size scaling in statistical physics [19], but have not been applied to characterize geometric generative models; the $\alpha \approx 1/3$ exponent identified in Section 4.6 connects to these traditions by relating generation error to the linear dimension of the structure. Closest to our setting, C2NP [52] formalizes a unit-cell-to-nanoparticle task with spherical truncation and rotation-stratified splits, evaluating multiple architectures and including a reverse (nanoparticle → unit cell) direction. RADII builds on this foundation with two distinct extensions: (i) frontier-specific diagnostics not present in C2NP (surface–interior decomposition, coordination correlation, cross-metric failure sequencing, and degradation ratios), and (ii) explicit scaling-law fits with OOD residual analysis that make frontiers quantitatively forecastable.

3 RADII Construction

3.1 Task Formulation

Given a primitive unit cell $\mathcal{U}_m = (\{\mathbf{b}_j, z_j\}_{j=1}^{M}, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ encoding basis positions, species, and lattice vectors, a target radius R, and the corresponding ground-truth species sequence $\mathbf{z} = (z_1, \ldots, z_{N(R)})$, the model generates a finite set of 3D atomic positions:

$$f : (\mathcal{U}_m,\ R,\ \mathbf{z}) \longmapsto \{\hat{\mathbf{x}}_i\}_{i=1}^{N(R)} \subset \mathbb{R}^3, \tag{1}$$

where $\hat{\mathbf{x}}_i \in \mathbb{R}^3$ is the predicted position of atom i, the atom count $N(R) = |\mathcal{S}(R)|$ is determined by R via the spherical truncation in Eq. (3), and atomic species are not predicted but supplied as inputs in $\mathbf{z}$. This conditioning establishes a natural one-to-one correspondence between predicted and reference atoms without requiring Hungarian matching or any other assignment algorithm. We emphasize that this formulation evaluates a geometry-only subtask of generative modeling: chemical identities are handled deterministically, and each model receives the exact stoichiometry and species ordering of the target nanoparticle. This design cleanly isolates geometric extrapolation as the variable of interest; extending the benchmark to an unconditioned track where models must additionally predict composition and ordering is a natural direction discussed in Section 5. Evaluating (1) across all 25 radii traces each model's quality from within the training distribution to beyond the extrapolation frontier.

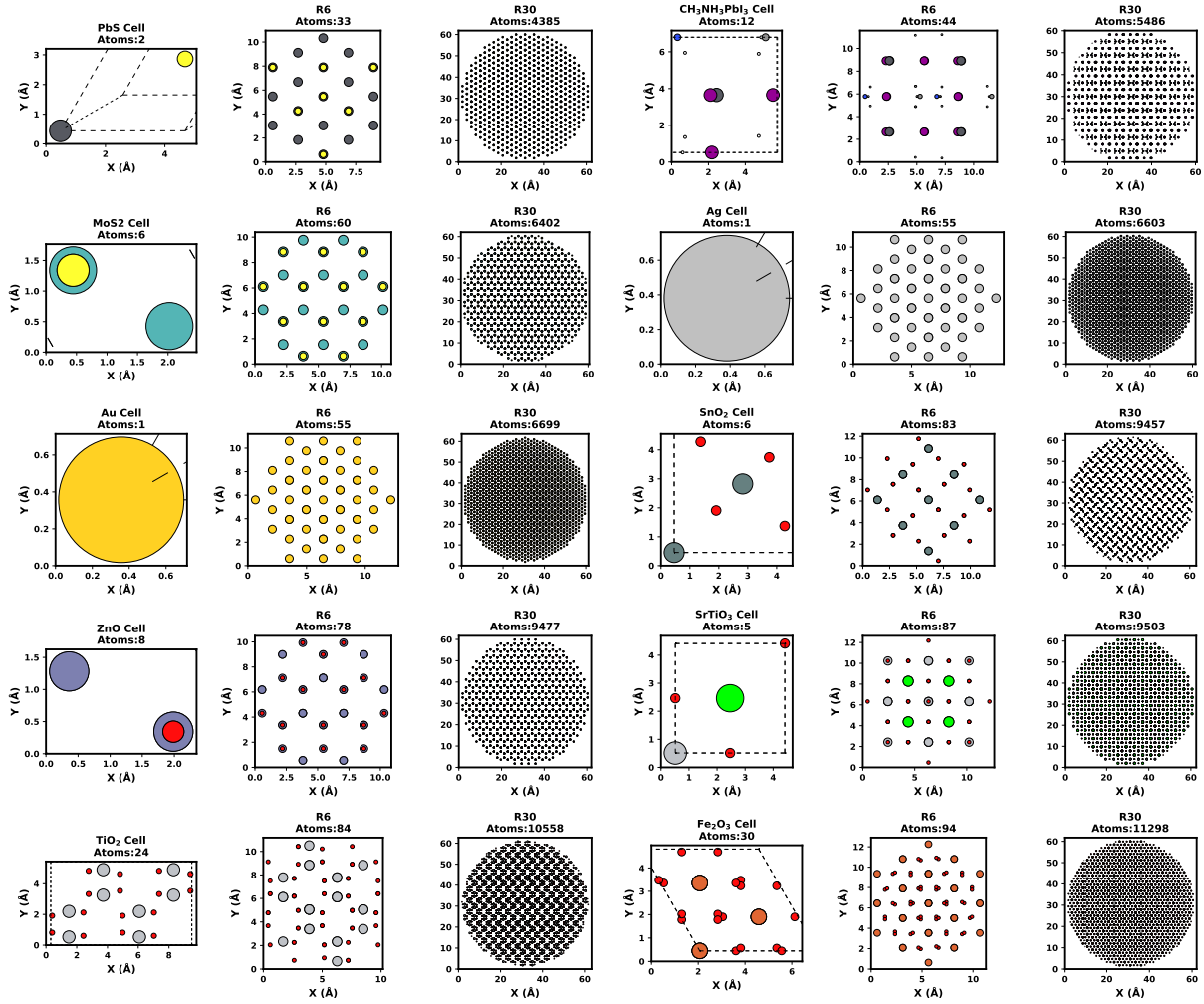

Figure 2: From primitive cell to radiuscontrolled nanoclusters. For each material in the dataset, the panels show, left to right, the primitive unit cell followed by its canonical $R = 6$ Å and $R = 30$ Ånanoparticles. Materials are arranged from top to bottom in ascending order of the atom count in their R=30 cluster, illustrating how coordination environments and bulklike cores emerge with increasing radius. All views share a common Ångström scale. Atom colours follow the conventional CPK palette.

Why no deterministic reconstruction baseline. Since the ground-truth nanoparticle is the deterministic spherical truncation of the unit cell (Eq. 3), a rule-based baseline that reconstructs the reference from the input lattice parameters would trivially achieve zero error: the lattice vectors and basis positions fully specify every atom's position. The task evaluates generation (producing atomic coordinates from learned representations), not reconstruction from explicit lattice parameters. A supervised coordinate-regression baseline (e.g., an MLP mapping unit-cell features to atom positions) faces the same triviality: with the unit cell, radius, and atom ordering all provided, the mapping is deterministic and learnable to near-zero error given sufficient capacity, providing no diagnostic value for measuring extrapolation. Instead, the ID performance of each generative model serves as its own architecture-specific reference against which OOD degradation is measured. The benchmark question is precisely whether a learned generative model can preserve and extrapolate these structural regularities through its learned representations as output size grows, an $O(N)$ coordinate-placement problem that an explicit lattice rule, but not a learned model, resolves for free.

3.2 Material Selection and Structure Generation

The benchmark spans FCC metals Ag [32], Au [33]), corundum Fe_2O_3 [18], a layered transition-metal dichalcogenide MoS_2 [21, 64], rock-salt PbS [65], rutile-type oxides SnO_2 [3] and TiO_2 [22], perovskites $SrTiO_3$ [46] and $CH_3NH_3PbI_3$ [63],

and wurtzite ZnO [66]. RADII is built entirely from experimentally reported crystal structures, with the ten primitive unit cells extracted from published CIFs and used as input conditioning for all models. Deterministic spherical truncation produces idealized crystal-derived nanoparticle references containing 33–11,298 atoms; after orientation sampling the benchmark comprises ∼75,000 structures, partitioned into 48,000 training, 13,500 ID-test, and 13,480 OOD-test structures. Figure 2 summarizes benchmark structure and key statistics, and Appendix A reports per-material and per-radius statistics.

Supercell construction. Let $N_{\text{rep}} = 60$. The periodic lattice of atomic sites is

$$\mathcal{L} = \Big\{ n_1\mathbf{v}_1 + n_2\mathbf{v}_2 + n_3\mathbf{v}_3 + \mathbf{b}_j \;\Big|\; n_1, n_2, n_3 \in \{0, \dots, N_{\text{rep}} - 1\},\ j \in \{1, \dots, M\} \Big\}, \tag{2}$$

where $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{R}^3$ are the primitive lattice vectors and $\{\mathbf{b}_j\}_{j=1}^{M}$ are the basis positions.

Scale-resolved nanoparticle. Selecting a central reference site $\mathbf{b}_0$ as the image of $\mathbf{b}_1$ in the central unit cell, the nanoparticle at radius R is

$$\mathcal{S}(R) = \{\mathbf{x} \in \mathcal{L} \mid \|\mathbf{x} - \mathbf{b}_0\|_2 \le R\}, \tag{3}$$

i.e., all lattice sites within a sphere of radius R. We sample $K = 25$ radii uniformly:

$$R_k = R_{\min} + (k-1)\Delta R, \qquad k = 1, \dots, 25, \tag{4}$$

with $R_{\min} = 6$ Å, $R_{\max} = 30$ Å, and $\Delta R = 1$ Å, yielding structures of 33–11,298 atoms.

3.3 Radius Split Protocol

The 25 radii ($R \in \mathcal{R} = \{6, 7, \dots, 30\}$ Å) are partitioned into three disjoint groups to cleanly separate interpolation from extrapolation. ID test radii $\mathcal{R}_{\text{ID}} = \{11, 13, 15, 17, 19, 21\}$ Å are held out from the training set but interleaved within the training radius range of 8–28 Å, and are evaluated under unseen orientations to measure interpolation quality.

The four OOD test radii $\mathcal{R}_{\text{OOD}} = \{6, 7, 29, 30\}$ Å lie strictly outside the training range, below (6, 7 Å) and above (29, 30 Å), probing extrapolation in both the small-particle and large-particle regimes. Because atom count scales as R^3, these radius offsets correspond to large shifts in structural complexity: the training radii span 81–9,148 atoms, whereas the OOD radii span 33–11,298 atoms, about 59% smaller than the smallest and 24% larger than the largest training structure. The leakage-free guarantee is two-fold: (i) no OOD or ID radius appears during training, and (ii) ID and OOD test orientations are excluded from training orientations via the angular exclusion constraint described below.

3.4 Quaternion-Based Orientation Sampling

For each nanoparticle $\mathcal{S}(R)$, we generate rigidly rotated copies using unit quaternions to provide diverse input orientations for training and evaluation.

Quaternion distance. Rotation is represented by $q = (q_x, q_y, q_z, q_w) \in \mathbb{S}^3$ with canonical sign $q_w \ge 0$. The geodesic distance is

$$d(q_1, q_2) = 2\arccos\big(|\langle q_1, q_2\rangle|\big) \in [0, \pi]. \tag{5}$$

Greedy angular separation. Given spacing $\Delta\theta > 0$, we construct a set $\mathcal{Q}$ by rejection sampling from the Haar-uniform distribution on SO(3), accepting candidates sequentially subject to

$$\max_{q' \in \mathcal{Q}} |\langle q', q\rangle| \le \cos(\Delta\theta/2), \tag{6}$$

guaranteeing $d(q_i, q_j) \ge \Delta\theta$ for all accepted pairs (Proposition 3.1).

Proposition 3.1 (Angular separation guarantee). For any $q_i, q_j \in \mathcal{Q}$ returned by the greedy procedure at spacing $\Delta\theta$,

$$|\langle q_i, q_j\rangle| \le \cos(\Delta\theta/2) \iff d(q_i, q_j) \ge \Delta\theta. \tag{7}$$

Proof. Follows directly from $d(q_i, q_j) = 2\arccos(|\langle q_i, q_j\rangle|)$ and the greedy acceptance rule. □

Split-specific construction and design rationale. RADII uses split-dependent angular spacings: $\Delta\theta_{\text{train}} = 16^\circ$, $\Delta\theta_{\text{ID}} = 14^\circ$, $\Delta\theta_{\text{OOD}} = 12^\circ$. The spacings decrease from train to test to provide progressively denser angular coverage for evaluation splits, ensuring thorough probing of orientation sensitivity at test time. The generation proceeds in two passes to enforce strict leakage prevention:

Pass 1 (Train + ID). A single global quaternion grid $\mathcal{Q}_{\text{train}}$ is generated once at spacing $\Delta\theta_{\text{train}}$; per-structure subsets are drawn via deterministic seeding to ensure reproducibility. ID quaternions are then sampled subject to an exclusion constraint against $\mathcal{Q}_{\text{train}}$:

$$\max_{q' \in \mathcal{Q}_{\text{train}}} |\langle q', q\rangle| \le \cos(\delta_{\text{ID}}/2), \tag{8}$$

with exclusion margin $\delta_{\text{ID}} = 8^\circ$.

Pass 2 (OOD). OOD quaternions are sampled subject to exclusion against the union $\mathcal{Q}_{\text{train}} \cup \mathcal{Q}_{\text{ID}}$ of all previously generated orientations:

$$\max_{q' \in \mathcal{Q}_{\text{train}} \cup \mathcal{Q}_{\text{ID}}} |\langle q', q\rangle| \le \cos(\delta_{\text{OOD}}/2), \tag{9}$$

with stricter exclusion margin $\delta_{\text{OOD}} = 10^\circ$ to provide a wider angular buffer against training orientations, reflecting the stronger isolation required for out-of-distribution evaluation.

Fixed left-multiplication offsets ($q_{\text{ID}} = \text{Euler}_{xyz}(20^\circ, 30^\circ, 45^\circ)$, $q_{\text{OOD}} = \text{Euler}_{xyz}(50^\circ, 70^\circ, 90^\circ)$) shift each split's quaternion candidates into a distinct region of SO(3) prior to exclusion checking. These offsets are chosen to be well-separated from each other and from the identity (which seeds the training grid), ensuring that each split explores a geometrically distinct portion of orientation space. The two-pass architecture guarantees that OOD orientations are excluded against both training and ID orientations, providing a strictly stronger leakage-free guarantee than independent sampling.

Highly symmetric structures that map multiple rotations to identical coordinates are deduplicated by hashing rounded coordinates at tolerance $\varepsilon = 10^{-6}$, retaining only unique orientations. After deduplication, the final dataset contains

74,980 structures in total, partitioned as 48,000 training structures, 13,500 ID test structures, and 13,480 OOD test structures.

3.5 Evaluation Metrics

RADII's metrics answer a central question: at what size does generation quality collapse, and what breaks first? They are organized into three tiers: generation quality measures tracked per radius, failure decomposition diagnostics that localize errors within each structure, and frontier characterization metrics that quantify scaling ceilings.

Correspondence guarantee. All metrics below rely on a one-to-one mapping between predicted and ground-truth atoms. As described in Section 3.1, this correspondence is guaranteed by construction: the conditioning provides the exact atom count $N(R)$ and species sequence, and atom ordering is inherited from the input. Kabsch alignment therefore operates on paired, equal-sized, species-matched point sets without requiring permutation search.

On assignment-free alternatives. Because the correspondence is guaranteed, assignment-free metrics such as Chamfer distance or earth mover's distance are not required for correctness. However, such metrics, along with radial distribution function (RDF) divergences, would provide complementary, correspondence-independent views of generation quality and could help cross-validate the failure sequences reported in Section 4.5. We discuss their inclusion in Section 5. The present metric suite is chosen to maximize diagnostic specificity: RMSD localizes global errors, BondMAE captures local chemical fidelity, and CoordCorr tracks topological preservation; aggregate distribution-level metrics would obscure these distinctions.

3.5.1 Generation Quality Measures. After centering both prediction P and ground truth G and computing the optimal Kabsch rotation $\mathbf{R}^\star$, we define: $\mathrm{RMSD}(P,G) = \sqrt{\frac{1}{N}\|\tilde{P} - \tilde{G}\|_F^2}$, where $\tilde{P} = \mathrm{center}(P)\mathbf{R}^\star$ and $\tilde{G} = \mathrm{center}(G)$. The one-to-one correspondence required by Kabsch alignment is guaranteed by the task conditioning: both P and G contain exactly $N(R)$ atoms with matching species, and ordering is preserved from the input.

Local bond-length MAE. Let $D_k(X) \in \mathbb{R}^{N\times k}$ be the k-nearest-neighbor distances via KD-tree. Flattening and sorting into vectors d_P, d_G:

$$\mathrm{BondMAE}_k(P,G) = \frac{1}{m}\sum_{i=1}^{m} |(d_P)_i - (d_G)_i|, \qquad m = \min(|d_P|, |d_G|). \tag{10}$$

We note that this formulation compares globally sorted distance vectors rather than per-atom neighbor lists, which may conflate distinct local environments; per-atom kNN matching or bond-angle distributions could provide finer localization and are considered for future inclusion (Section 5). As a robustness check, a per-atom variant that compares each atom's neighbor list individually yields identical model rankings (Spearman $\rho = 1.0$; Appendix C), confirming that global sorting does not mislead the architectural comparison. Tracking BondMAE alongside RMSD nonetheless distinguishes models that lose local chemical order from those that maintain short-range structure but produce incorrect global morphology.

3.5.2 Failure Decomposition Diagnostics.

Surface–interior error ratio. Let S and I index the outermost and innermost 25% of atoms by distance from the centroid, where shell membership is defined on the ground-truth structure and per-atom errors are computed from the Kabsch-aligned prediction using the guaranteed atom correspondence:

$$\mathrm{SurfIntRatio} = \frac{\mathrm{SurfRMSD}}{\mathrm{IntRMSD} + 10^{-8}}. \tag{11}$$

A ratio increasing with radius indicates boundary-driven collapse; a stable ratio signals uniform degradation. The 25% shell fraction is a default; Appendix C shows the ID–OOD conclusions hold across fractions from 15% to 30%.

Coordination preservation. Using KD-tree ball queries with cutoff r_c (default $r_c = 3.0$ Å; cross-architecture rankings are stable for $r_c \in [3.0, 4.5]$ Å, Appendix C), per-atom coordination numbers yield:

$$\mathrm{CoordCorr}(P,G) = \mathrm{corr}(c_P, c_G) \in [-1, 1]. \tag{12}$$

A sharp drop at a specific radius signals the model has exceeded the scale at which it maintains local structural rules.

3.5.3 Frontier Characterization. These metrics operate on per-radius error profiles $m(R)$ rather than individual structures.

ID–OOD degradation ratio.

$$\mathrm{Degrad}(m) = \frac{\frac{1}{|\mathcal{R}_{\mathrm{OOD}}|}\sum_{R\in\mathcal{R}_{\mathrm{OOD}}} m(R)}{\frac{1}{|\mathcal{R}_{\mathrm{ID}}|}\sum_{R\in\mathcal{R}_{\mathrm{ID}}} m(R) + 10^{-8}}. \tag{13}$$

Values near 1 indicate robust scaling; values $\gg 1$ indicate failure to generalize beyond training sizes.

Frontier radius. For quality threshold τ: $r^\star(m,\tau) = \max\{R \in \mathcal{R} : m(R) \le \tau\}$. Comparing $r^\star$ across models, materials, and metrics provides a compact summary of each architecture's scaling ceiling.

Reproducibility. All construction is fully deterministic: a global seed with per-structure FNV-1a hashing of (material, radius, split) ensures identical outputs across runs. The released repository includes CIF-to-nanoparticle scripts, split configurations, quaternion generation code, and all evaluation implementations under the MIT license.

4 Experiments

We evaluate five generative models (CDVAE [67], DiffCSP [23], FlowMM [45], MatterGen [71], and ADiT [25]) on the unit-cell → nanoparticle task (Eq. (1)). Each model receives the unit cell, target radius, and atom count as conditioning and generates exactly $N(R)$ atoms with the specified composition, ensuring one-to-one correspondence with the ground

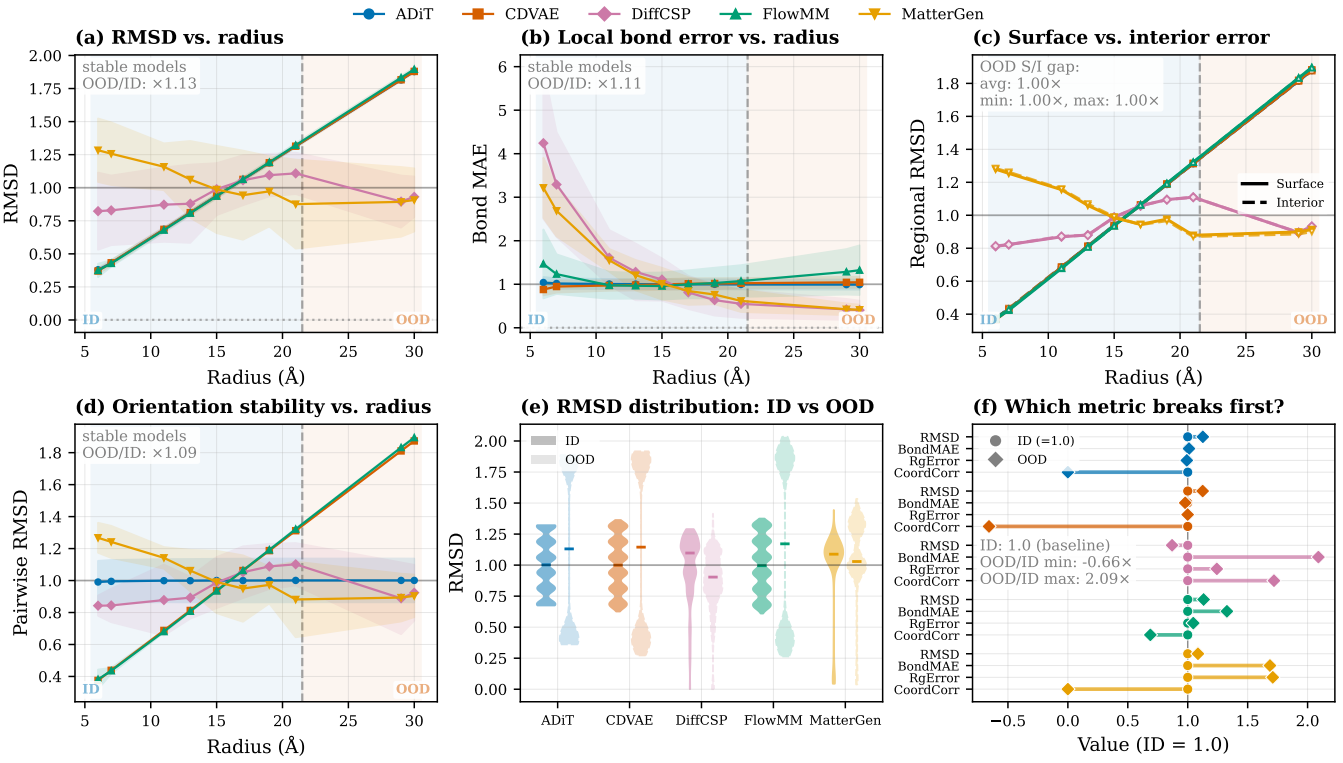

Figure 3: Extrapolation frontier across multiple dimensions of structural fidelity. (a) Global RMSD increases beyond the in-distribution boundary, revealing differing degrees of OOD degradation across models. (b) Local bond geometry errors diverge more strongly, showing that extrapolation behavior depends on the evaluation metric. (c) Surface atoms consistently exhibit higher errors than interior atoms, with similar trends from ID to OOD regimes, indicating uniform degradation. (d) Orientation consistency generally degrades alongside positional accuracy, though stability varies across architectures. (e) Distribution shifts in OOD samples show broader error tails compared to ID. (f) Multi-metric comparison highlights architecture-specific failure modes, with different models degrading along different structural dimensions. Shaded regions denote ID and OOD radii, and error bands indicate ±1 standard deviation across materials and seeds.

truth (Section 3.1). The species sequence and atom ordering are communicated to each model as an explicit input tensor that fixes the identity and position index of every atom; during generation, models produce coordinates for each index in this fixed sequence, so the output is order-aligned with the ground truth by construction rather than by post-hoc matching. All metrics are computed per radius, averaged over all orientations at that radius. We report both raw values (in Å) and normalized values where each model's ID mean is set to 1.0 so that OOD values directly express relative degradation; we stress that normalized ratios are interpretable only for models whose raw ID performance is structurally meaningful (see Section 4.1 for absolute-scale context). All experiments are repeated over three random seeds; we report means and ±1 standard deviation throughout. Seed-to-seed variability in the RMSD degradation ratio is small for the three well-behaved models (ADiT, CDVAE, FlowMM; std < 0.02); the divergent models show larger seed spread, itself a failure signal.

Training protocol. All five architectures are trained under a unified protocol to ensure a fair comparison (Table 1). All models share the same optimizer, batch size, gradient clipping, learning rate scheduler, and training until plateau. Radius conditioning is provided as an explicit scalar input alongside the unit cell. Architecture-specific parameters (hidden dimensions, cutoff radii, diffusion/flow schedules) follow published specifications scaled to a unified budget of ∼500–550K parameters to isolate architectural differences from capacity effects. We acknowledge that this budget may disadvantage architectures designed for larger scales; Section 5 discusses this trade-off, and Section 4.7 probes it directly by rerunning the divergent models at their published parameter counts. At the largest radii ($R = 30$ Å, up to 11,298 atoms), peak GPU memory ranged from ∼8 GB (CDVAE) to ∼18 GB (ADiT); all models used the same neighbor-cutoff-based graph construction (Table 1) without additional sparsification.

Table 1: Shared hyperparameters: Adam optimizer (lr = 10^{-4}), batch size 2, gradient clip norm 1.0, ReduceLROnPlateau (factor 0.5, patience 5), and 3 seeds.

	ADiT	CDVAE	DiffCSP	FlowMM	MatterGen
Hidden dim	24	92	120	120	130
Num layers	2	2	2	2	2
Cutoff (Å)	5.0	5.0	7.0	5.0	5.0
Latent dim	8	92	—	—	—
Diffusion steps	1000	—	1000	—	100
β range	$[10^{-4}, 0.02]$	—	$[10^{-4}, 0.02]$	—	—
σ range	—	$[0.01, 1.0]$	—	$[10^{-4}, —]$	—
Num Gaussians	—	50	50	50	50

4.1 Where Is the Frontier?

Figure 3(a) plots normalized RMSD versus radius. ADiT, CDVAE, and FlowMM show tightly clustered degradation ratios of 1.13, 1.12, and 1.13, corresponding to a consistent ∼13% RMSD increase from ID to OOD. DiffCSPs normalized OOD RMSD decreases to 0.87, but this apparent robustness is misleading: its raw ID RMSD exceeds 3,386 Å (∼290–470× larger than ADiT at 7.15 Å or FlowMM at 11.55 Å), indicating structurally incoherent outputs at all scales, with MatterGens 5,905 Å RMSD placing it in the same regime. Their normalized ratios therefore reflect relative change atop already failed baselines. At the shared ∼500K-parameter budget these two samplers are numerically divergent rather than merely inaccurate. We verified these magnitudes through multiple diagnostics: consistent Å units, valid Kabsch solutions ($\det(\mathbf{R}^\star) = +1$), broadly distributed per-atom errors rather than outliers, visual inspection showing globally disordered structures, and agreement from complementary metrics (DiffCSP: BondMAE 2.09×, RgError 1.24×; MatterGen: BondMAE 1.69×, RgError 1.71×). These models were originally designed for periodic crystals at smaller atom counts, suggesting a combination of task mismatch and parameter-budget constraints rather than a pipeline artifact. Figure 3(b) shows a contrasting frontier under local bond fidelity: BondMAE degradation diverges substantially (DiffCSP 2.09×, MatterGen 1.69×, FlowMM 1.33×) while ADiT (1.01) and CDVAE (0.98) remain near unity, demonstrating that the extrapolation frontier is metric-dependent. Models appearing robust under global RMSD may simultaneously lose bond-length distributions required for chemical validity.

4.2 Where Do Failures Originate?

A natural hypothesis is that extrapolation failures concentrate on under-coordinated surface atoms. Figure 3(c) partitions atoms into surface (outer 25%) and interior (inner 25%) shells defined on the ground-truth structure, with per-atom errors computed from the aligned prediction. Across all five models, the surface–interior gap ratio remains remarkably stable from ID to OOD, with changes bounded by ±0.0227 at the default 25% shell fraction (±0.0475 across fractions from 15% to 30%; Appendix C). This small shift rules out boundary-driven collapse: degradation propagates uniformly through the structure, indicating that the deficit lies in modeling how bulk structure extends with size rather than in surface geometry.

4.3 Does Orientation Stability Transfer Across Scale?

By evaluating each structure under multiple input rotations, RADII isolates orientation sensitivity as a dimension independent of positional accuracy. Figure 3(d) tracks orientation consistency across radii. CDVAE and FlowMM show orientation-consistency degradation ratios of 1.14 and 1.21; FlowMM's orientation degrades more steeply than its RMSD (1.13×), so orientation and positional accuracy are not interchangeable frontiers. ADiT keeps its absolute orientation error negligibly small at every scale (∼0.003–0.005 Å from ID to OOD), behaving as an effectively orientation-equivariant model even where positional accuracy degrades. Orientation consistency thus constitutes a secondary frontier, architecturally decoupled from the primary quality frontier.

4.4 How Are Errors Distributed?

Beyond mean degradation, distributional analysis exposes worst-case risk. Figure 3(e) shows full RMSD distributions under ID and OOD conditions. ADiT, CDVAE, and FlowMM exhibit matching tail ratios ($Q95_{OOD}/Q95_{ID}$) of 1.43, meaning worst-case OOD errors are 43% worse than worst ID cases. FlowMM shows the largest median shift (+0.176 in normalized RMSD) and highest normalized OOD maximum (2.03), indicating occasional catastrophic failures at extrapolation scales. DiffCSP's distribution narrows (tail ratio 0.91), consistent with collapse into a scale-insensitive failure mode where all outputs are uniformly poor.

4.5 Which Metric Breaks First?

Figure 3(f) compares ID → OOD degradation across four normalized metrics simultaneously. The failure sequences are qualitatively distinct: ADiT fails primarily on global RMSD (1.13×) while local chemistry is preserved; FlowMM breaks on BondMAE (1.33×) before RMSD; MatterGen suffers simultaneous RgError (1.71×) and BondMAE (1.69×) collapse; and DiffCSP is dominated by a BondMAE collapse (2.09×), with coordination correlation falling to ≈0. No two architectures exhibit the same sequence of metric failures, demonstrating that the frontier is a multi-dimensional surface shaped by model family.

To provide practitioner-ready comparisons, Table 2 instantiates the frontier radius $r^\star(m, \tau)$ for both RMSD and BondMAE: ADiT reaches the farthest under RMSD ($r^\star$ = 30 Å at τ = 15 Å), FlowMM is the only model with a finite BondMAE frontier, and DiffCSP and MatterGen have no finite frontier at any threshold (full threshold sweep in Appendix B). Material dependence reinforces this: RMSD degradation is nearly material-invariant (std = 0.005, range 1.118–1.139), whereas BondMAE spread widens dramatically (std = 0.208, range 0.962–1.701), with tetragonal oxides TiO_2 and SnO_2 hardest and cubic metals easiest, correlating with unit-cell complexity.

Table 2: Frontier radius $r^\star(m,\tau)$ in Å: the largest radius at which model m holds the metric within threshold τ; "–" means no radius qualifies. Full threshold sweep in Appendix B.

	RMSD $r^\star$ (Å)			BondMAE $r^\star$
Model	τ=5	τ=10	τ=15	τ=0.6
ADiT	11	21	30	–
CDVAE	–	7	13	–
DiffCSP	–	–	–	–
FlowMM	7	13	19	21
MatterGen	–	–	–	–

4.6 Scaling Laws for Nanostructure Generation

Figure 4 fits power-law relationships RMSD $\sim N^\alpha$ on ID radii. ADiT (α = 0.334, R^2 = 1.000), CDVAE (0.335, 1.000), and FlowMM (0.342, 1.000) exhibit nearly identical exponents near 1/3, indicating RMSD grows with the nanoparticles linear dimension ($N^{1/3} \propto R$). An exponent of 1/3 is the expected geometric baseline for a well-behaved model rather than a surprising finding; its diagnostic value lies in conformity versus deviation. Geometrically, this reflects spatially uniform positional error accumulating with structure size, corresponding to systematic scaling rather than abrupt failure. DiffCSP (α = 0.142, R^2 = 0.928) and MatterGen ($\alpha = -0.126$, R^2 = 0.897) deviate strongly, with MatterGens negative exponent arising from fitting noise under uniformly large errors. OOD residuals further distinguish predictable from unstable scaling: ADiT (0.001) and FlowMM (0.004) maintain near-zero residuals, meaning ID scaling accurately predicts OOD degradation, whereas DiffCSP (0.118) and MatterGen (0.050) diverge substantially. For models in the $\alpha \approx 1/3$ regime, performance at unseen sizes can therefore be estimated from ID fits alone. Appendix B reports the full scaling-law fits: the three well-behaved models follow a consistent RMSD power law while the divergent architectures exhibit unstable exponents. Seed-to-seed variability of α is below 0.002 for the conforming models, indicating the scaling law is a stable architectural property for them, whereas DiffCSP and MatterGen are seed-unstable (std > 0.09), itself a failure signal.

4.7 Does Model Capacity Close the Frontier?

The shared ~500K-parameter budget isolates architecture from capacity but may disadvantage models designed for larger scales. To separate the two effects, we retrained the two models that diverged at this budget (MatterGen and DiffCSP) at approximately their published parameter counts (MatterGen at 47M; DiffCSP at 12.4M), under the faster training settings detailed in Appendix D. The two outcomes differ. MatterGen at 47M recovers stable, bounded sampling (ID RMSD 17.63 Å, OOD RMSD 19.81 Å) where the 500K run had been numerically divergent; its shared-budget failure was therefore a capacity artifact. Yet the extrapolation frontier persists even at 47M: per-radius RMSD grows monotonically with R and is largest at the farthest OOD radii. DiffCSP at 12.4M, by contrast, still diverges: reverse-diffusion sampling saturates the numerical safety bound despite a normally converging training loss, so capacity is not its bottleneck. Together these runs decompose the failure modes: capacity governs whether sampling is stable at all, whereas the size-extrapolation frontier is architectural. Finally, as a benchmark rather than a model, RADII's analog of an architecture ablation is the sensitivity analysis of its own design choices (Appendix C), not an ablation of model internals.

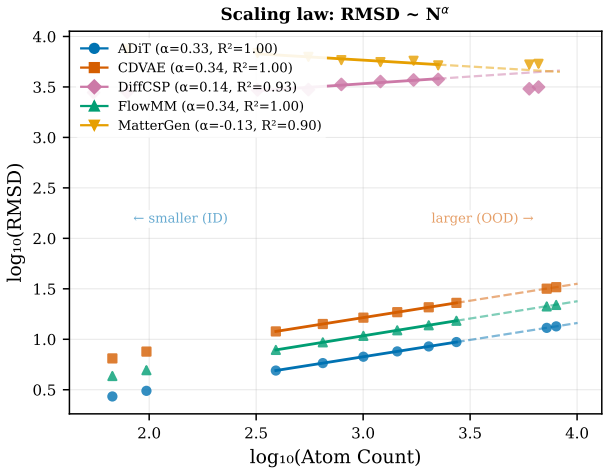

Figure 4: Power-law scaling relationships quantify the extrapolation frontier. A loglog plot of RMSD versus atom count shows approximate power-law behavior, RMSD $\sim N^\alpha$. Solid lines denote fits on in-distribution data, while dashed lines extrapolate into out-of-distribution regimes. Models with consistent scaling exhibit predictable extrapolation, whereas deviations indicate irregular or less predictable behavior beyond the training regime.

5 Limitations and Ethical Considerations

RADII evaluates generation on idealized nanoparticle geometries constructed by spherical truncation of bulk lattices without surface reconstruction, passivation, or finite-temperature relaxation, so performance reflects fidelity to ideal structures rather than experimentally realized surfaces, and both failure modes and success cases may change after structural relaxation. The current formulation conditions on atom count, species sequence, and ordering, isolating geometric extrapolation but narrowing generality; a natural extension is an unconditioned track evaluated with assignment-free metrics. The unified ~500–550K parameter budget enables controlled architectural comparison but may disadvantage models designed for larger scales or different inductive biases, and several evaluated methods were originally developed for periodic crystals rather than large finite clusters. Metric design also introduces limitations: BondMAE$_k$ compares globally sorted neighbor distances and may conflate

distinct local environments, while assignment-free measures such as Chamfer distance, earth mover's distance, or RDF divergences could provide complementary correspondence-independent validation. Finally, the ten selected materials span multiple bonding types and symmetries but do not exhaust inorganic chemistry, and the 6–30 Åradius range probes only the small/medium-nanoparticle regime, leaving open whether the observed scaling behavior persists at larger sizes.

Ethical considerations. RADII comprises only computational structural data derived from published crystallographic references; it involves no human subjects, personal data, or dual-use concerns, and all code and benchmark data are released under the MIT license. Because the benchmark targets are idealized crystal-derived references, results should not be over-interpreted as direct validation for real-world nanomaterial deployment without additional physical relaxation or experimental verification.

6 Conclusion

We introduced RADII, a radius-resolved benchmark that maps the extrapolation frontier of crystalline generative models across 25 size configurations, ten materials, and five architectures, revealing that well-behaved models degrade by ∼13% in global positional error beyond training radii while divergent models fail in absolute structural fidelity and local bond fidelity ranges from negligible degradation to more than 2Œ error growth, that no two architectures share the same failure sequence, that surface and interior errors grow in lockstep rather than from boundary effects, and that well-behaved models obey a power-law exponent $\alpha \approx 1/3$ whose in-distribution fit accurately predicts out-of-distribution error. These findings establish output scale as a first-class evaluation axis and reframe the extrapolation frontier as a diagnosable, forecastable quantity, though we recognize that the current benchmark evaluates a geometry-conditioned subtask on idealized references under a constrained parameter budget and that frontier locations may shift under more realistic conditions or larger model capacities. A first probe of the capacity question, rerunning the two divergent models at their published parameter counts, shows that greater capacity can stabilize sampling, as observed for MatterGen, but does not by itself close the size-extrapolation frontier; DiffCSP remains unstable even at published scale. Future work will incorporate DFT-relaxed and Wulff-shaped references, introduce an unconditioned evaluation track with assignment-free metrics, evaluate architectures at their intended scales alongside controlled-budget comparisons, explore size-conditioned training strategies to push the frontier outward, and investigate whether the scaling law and frontier-diagnostic framework transfer to proteins, biomolecular assemblies, and amorphous solids.

GenAI Disclosure

Generative AI tools were used during the preparation of this manuscript to assist with editing prose, refining mathematical notation, and debugging code. All scientific contributions, including benchmark design, experimental methodology, data generation, model evaluation, and interpretation of results, were conceived and executed entirely by the authors. All AI-generated text was critically reviewed, verified, and revised by the authors, who take full responsibility for the content of this work.

References

[1] Roi Baer and Martin Head-Gordon. 1997. Sparsity of the density matrix in Kohn-Sham density functional theory and an assessment of linear system-size scaling methods. Physical Review Letters 79, 20 (1997), 3962.

[2] Georgios D Barmparis, Zbigniew Lodziana, Nuria Lopez, and Ioannis N Remediakis. 2015. Nanoparticle shapes by using Wulff constructions and first-principles calculations. Beilstein Journal of Nanotechnology 6, 1 (2015), 361–368.

[3] W. H. Baur, R. A. Sass, et al. 1971. The rutile structure of SnO_2. Acta Crystallographica Section B 27 (1971), 2133.

[4] Debasis Bera, Lei Qian, Teng-Kuan Tseng, and Paul H Holloway. 2010. Quantum dots and their multimodal applications: a review. Materials 3, 4 (2010), 2260–2345.

[5] F Matthias Bickelhaupt and Evert Jan Baerends. 2000. Kohn-Sham density functional theory: predicting and understanding chemistry. Reviews in Computational Chemistry 15, 1 (2000), 1–86.

[6] L. C. Blum and J.-L. Reymond. 2009. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. J. Am. Chem. Soc. 131 (2009), 8732.

[7] Shuang Cao, Ning Sui, Peng Zhang, Tingting Zhou, Jinchun Tu, and Tong Zhang. 2022. TiO2 nanostructures with different crystal phases for sensitive acetone gas sensors. Journal of Colloid and Interface Science 607 (2022), 357–366.

[8] Ivano E Castelli, David D Landis, Kristian S Thygesen, Søren Dahl, Ib Chorkendorff, Thomas F Jaramillo, and Karsten W Jacobsen. 2012. New cubic perovskites for one-and two-photon water splitting using the computational materials repository. Energy & Environmental Science 5, 10 (2012), 9034–9043.

[9] Ivano E Castelli, Thomas Olsen, Soumendu Datta, David D Landis, Søren Dahl, Kristian S Thygesen, and Karsten W Jacobsen. 2012. Computational screening of perovskite metal oxides for optimal solar light capture. Energy & Environmental Science 5, 2 (2012), 5814–5819.

[10] Rees Chang, Angela Pak, Alex Guerra, Ni Zhan, Nick Richardson, Elif Ertekin, and Ryan P Adams. 2025. Space Group Equivariant Crystal Diffusion.

[11] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. 2021. Open catalyst 2020 (OC20) dataset and community challenges. Acs Catalysis 11, 10 (2021), 6059–6072.

[12] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. 2017. Machine learning of accurate energy-conserving molecular force fields. Science Advances 3, 5 (2017), e1603015.

[13] Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T Unke, Adil Kabylda, Huziel E Sauceda, Alexandre Tkatchenko, and Klaus-Robert Müller. 2023. Accurate global machine learning force fields for molecules with hundreds of atoms. Science Advances 9, 2 (2023), eadf0873.

[14] Aron J Cohen, Paula Mori-Sánchez, and Weitao Yang. 2008. Insights into current limitations of density functional theory. Science 321, 5890 (2008), 792–794.

[15] Murray S Daw and Michael I Baskes. 1984. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. Physical Review B 29, 12 (1984), 6443.

[16] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. 2020. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. npj Computational Materials 6, 1 (2020), 138.

[17] Marcus Elstner and Gotthard Seifert. 2014. Density functional tight binding. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 372, 2011 (2014), 20120483.

[18] L. W. Finger and R. M. Hazen. 1980. Crystal structure and isothermal compression of Fe_2O_3, Cr_2O_3, and V_2O_3 to 50 kbars. Journal of Applied Physics 51 (1980), 5362–5367. doi:10.1063/1.327451

[19] Michael E Fisher and Michael N Barber. 1972. Scaling theory for finite-size effects in the critical region. Physical Review Letters 28, 23 (1972), 1516.

[20] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. 2020. Se (3)-transformers: 3d roto-translation equivariant attention networks. Advances in Neural Information Processing Systems 33 (2020), 1970–1981.

[21] R. Grau-Crespo and R. Lopez-Cordero. 2002. MoS_2 structural properties. Phys. Chem. Chem. Phys. 4 (2002), 4078.

[22] M. Horn, C. R. Meagher, et al. 1972. Structure of Anatase TiO_2. Zeitschrift für Kristallographie 136 (1972), 273.

[23] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. 2023. Crystal structure prediction by joint equivariant diffusion. Advances in Neural Information Processing Systems 36 (2023), 17464–17497.

[24] Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. 2024. Space group constrained crystal generation.

[25] Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. 2025. All-atom Diffusion Transformers: Unified generative modelling of molecules and materials.

[26] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

[27] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. Nature Reviews Physics 3, 6 (2021), 422–440.

[28] Filip Ekström Kelvinius, Oskar B Andersson, Abhijith S Parackal, Dong Qian, Rickard Armiento, and Fredrik Lindsten. 2025. WyckoffDiff–A Generative Diffusion Model for Crystal Symmetry.

[29] Kuzma Khrabrov, Ilya Shenbin, Alexander Ryabov, Artem Tsypin, Alexander Telepov, Anton Alekseev, Alexander Grishin, Pavel Strashnov, Petr Zhilyaev, Sergey Nikolenko, et al. 2022. nabladft: Large-scale conformational energy and hamiltonian prediction benchmark and dataset. Physical Chemistry Chemical Physics 24, 42 (2022), 25853–25863.

[30] Byung-Hyun Kim, Jolla Kullgren, Matthew J Wolf, Kersti Hermansson, and Peter Broqvist. 2019. Multiscale modeling of agglomerated ceria nanoparticles: interface stability and oxygen vacancy formation. Frontiers in Chemistry 7 (2019), 203.

[31] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2025. PubChem 2025 update. Nucleic Acids Research 53, D1 (2025), D1516–D1525.

[32] H. W. King. 2002. CRC Handbook of Chemistry and Physics (83 ed.). CRC Press, Boca Raton, FL, USA. Standard phase data for silver (Ag).

[33] H. W. King. 2002. CRC Handbook of Chemistry and Physics (83 ed.). CRC Press, Boca Raton, FL, USA. Standard phase data for gold (Au).

[34] Charles Kittel and Paul McEuen. 2018. Introduction to solid state physics. John Wiley & Sons, Hoboken, NJ, USA.

[35] Mustafa Kurban, Can Polat, Erchin Serpedin, and Hasan Kurban. 2024. Enhancing the electronic properties of TiO2 nanoparticles through carbon doping: An integrated DFTB and computer vision approach. Computational Materials Science 244 (2024), 113248.

[36] Andrea C Levi and Miroslav Kotrla. 1997. Theory and simulation of crystal growth. Journal of Physics: Condensed Matter 9, 2 (1997), 299.

[37] Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. 2025. SymmCD: Symmetry-Preserving crystal generation with diffusion models.

[38] Rongjin Li, Xiaotao Zhang, Huanli Dong, Qikai Li, Zhigang Shuai, and Wenping Hu. 2016. Gibbs–Curie–Wulff theorem in organic materials: a case study on the relationship between surface energy and crystal growth. Advanced Materials 28, 8 (2016), 1697–1702.

[39] Peijia Lin, Pin Chen, Rui Jiao, Qing Mo, Jianhuan Cen, Wenbing Huang, Yang Liu, Dan Huang, and Yutong Lu. 2025. Equivariant diffusion for crystal structure prediction.

[40] Hongsheng Liu, Gotthard Seifert, and Cristiana Di Valentin. 2019. An efficient way to model complex magnetite: Assessment of SCC-DFTB against DFT. The Journal of chemical physics 150, 9 (2019), 094703.

[41] Shengchao Liu, Yanjing Li, Zhuoxinran Li, Zhiling Zheng, Chenru Duan, Zhi-Ming Ma, Omar Yaghi, Animashree Anandkumar, Christian Borgs, Jennifer Chayes, et al. 2023. Symmetry-informed geometric representation for molecules, proteins, and crystalline materials. Advances in neural information processing systems 36 (2023), 66084–66101.

[42] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. 2021. Spherical message passing for 3d graph networks.

[43] Youzhi Luo, Keqiang Yan, and Shuiwang Ji. 2021. Graphdf: A discrete flow model for molecular graph generation. In International Conference on Machine Learning. PMLR, PMLR, Virtual, 7192–7203.

[44] Avik Mahata, Tanmoy Mukhopadhyay, and Mohsen Asle Zaeem. 2022. Modified embedded-atom method interatomic potentials for Al-Cu, Al-Fe and Al-Ni binary alloys: From room temperature to melting point. Computational Materials Science 201 (2022), 110902.

[45] Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. 2024. Flowmm: Generating materials with riemannian flow matching. In Forty-first International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235). PMLR, Vienna, Austria, 35664–35686.

[46] Roger H Mitchell, Anton R Chakhmouradian, and Patrick M Woodward. 2000. Crystal chemistry of perovskite-type compounds in the tausonite-loparite series,(Sr1- 2 x Na x La x) TiO3. Physics and Chemistry of Minerals 27, 8 (2000), 583–589.

[47] Maylis Orio, Dimitrios A Pantazis, and Frank Neese. 2009. Density functional theory. Photosynthesis Research 102 (2009), 443–453.

[48] Anyang Peng, Chun Cai, Mingyu Guo, Duo Zhang, Chengqian Zhang, Antoine Loew, Linfeng Zhang, and Han Wang. 2025. LAMBench: A Benchmark for Large Atomic Models.

[49] Chris J Pickard. 2020. AIRSS data for carbon at 10GPa and the C+ N+ H+ O system at 1GPa. Materials Cloud Archive. doi:10.24435/materialscloud:2020.0026/v1

[50] Can Polat, Mustafa Kurban, and Hasan Kurban. 2024. Multimodal neural network-based predictive modeling of nanoparticle properties from pure compounds. Machine Learning: Science and Technology 5, 4 (2024), 045062.

[51] Can Polat, Erchin Serpedin, Mustafa Kurban, and Hasan Kurban. 2025. CrysMTM: a multiphase, temperature-resolved, multimodal dataset for crystalline materials. Machine Learning: Science and Technology 6, 3 (2025), 030603.

[52] Can Polat, Erchin Serpedin, Mustafa Kurban, and Hasan Kurban. 2026. C2NP: A Benchmark for Learning Scale-Dependent Geometric Invariances in 3D Materials Generation.

[53] Tingting Qi, Charles W Bauschlicher Jr, John W Lawson, Tapan G Desai, and Evan J Reed. 2013. Comparison of ReaxFF, DFTB, and DFT for phenolic pyrolysis. 1. Molecular dynamics simulations. The Journal of Physical Chemistry A 117, 44 (2013), 11115–11125.

[54] Emilie Ringe, Richard P Van Duyne, and Laurence D Marks. 2013. Kinetic and thermodynamic modified Wulff constructions for twinned nanoparticles. The Journal of Physical Chemistry C 117, 31 (2013), 15859–15870.

[55] Zachary A Rollins, Alan C Cheng, and Essam Metwally. 2024. MolPROP: Molecular Property prediction with multimodal language and graph fusion. Journal of Cheminformatics 16, 1 (2024), 56.

[56] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. Physical Review Letters 108 (2012), 058301.

[57] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 2021. E (n) equivariant graph neural networks. In International Conference on Machine Learning. PMLR, PMLR, Virtual, 9323–9332.

[58] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. 2018. Schnet–a deep learning architecture for molecules and materials. The Journal of Chemical Physics 148, 24 (2018), 241722.
[59] Arkadiy Simonov and Andrew L Goodwin. 2020. Designing disorder into crystalline materials. Nature Reviews Chemistry 4, 12 (2020), 657–673.
[60] Chris-Kriton Skylaris, Peter D Haynes, Arash A Mostofi, and Mike C Payne. 2005. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. The Journal of Chemical Physics 122, 8 (2005), 084119.
[61] Thomas Surek. 2005. Crystal growth and materials research in photovoltaics: progress and challenges. Journal of Crystal growth 275, 1-2 (2005), 292–304.
[62] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. 2023. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. ACS Catalysis 13, 5 (2023), 3066–3084.
[63] Aron Walsh, elds22, Federico Brivio, and Jarvist Moore Frost. 2019. WMD-group/hybrid-perovskites: Collection 1 (v1.0). https://doi.org/10.5281/zenodo.2641358. Hybrid perovskite $CH_3NH_3PbI_3$ structural data.
[64] R. W. G. Wyckoff. 1963. Crystal Structures Volume 1. Interscience Publishers, New York, NY, USA.
[65] R. W. G. Wyckoff. 1963. Crystal Structures Volume 1. Interscience Publishers, New York, NY, USA.
[66] R. W. G. Wyckoff. 1963. Crystal Structures Volume 1. Interscience Publishers, New York, NY, USA.
[67] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. 2021. Crystal diffusion variational autoencoder for periodic material generation.
[68] Ruo Xi Yang, Caitlin A McCandler, Oxana Andriuc, Martin Siron, Rachel Woods-Robinson, Matthew K Horton, and Kristin A Persson. 2022. Big data in a nano world: a review on computational, data-driven design of nanomaterials structures, properties, and synthesis. ACS Nano 16, 12 (2022), 19873–19891.
[69] Haiyang Yu, Meng Liu, Youzhi Luo, Alex Strasser, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. 2023. QH9: A Quantum Hamiltonian Prediction Benchmark for QM9 Molecules. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Track on Datasets and Benchmarks, Vol. 36. Curran Associates, Inc., Red Hook, NY, USA, 1–16.
[70] Haoyu S. Yu, Shaohong L. Li, and Donald G. Truhlar. 2016. Perspective: Kohn-Sham density functional theory descending a staircase. The Journal of Chemical Physics 145, 13 (10 2016), 130901. doi:10.1063/1.4963168
[71] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. 2023. Mattergen: a generative model for inorganic materials design.
[72] Guishan Zheng, Stephan Irle, and Keiji Morokuma. 2005. Performance of the DFTB method in comparison to DFT and semiempirical methods for geometries and energies of C20–C86 fullerene isomers. Chemical Physics Letters 412, 1-3 (2005), 210–216.

A Dataset Details

RADII is constructed from ten primitive unit cells taken from published, experimentally reported crystal structures. Table 3 summarizes the leakage-free split, and Table 4 reports the atom count of every (material, radius) pair on the ID and OOD test radii. Atom count scales as R^3, so the OOD radii correspond to structures well outside the training envelope: 33 atoms is 59% smaller than the smallest training structure (81 atoms) and 11,298 atoms is 24% larger than the largest (9,148 atoms).

Table 3: Leakage-free radius split. Atom counts are over all materials.

Split	Radii (Å)	Structures	Atom range
Train	15: 8–10, 12, 14, 16, 18, 20, 22–28	48,000	81–9,148
ID	6: 11, 13, 15, 17, 19, 21	13,500	203–3,858
OOD	4: 6, 7, 29, 30	13,480	33–11,298

Table 4: Atom count per (material, radius) on the ID and OOD test radii.

Radius (Å)	Ag	Au	$CH_3NH_3PbI_3$	Fe_2O_3	MoS_2	PbS	SnO_2	$SrTiO_3$	TiO_2	ZnO	Split
6	55	55	44	94	60	33	83	87	84	78	OOD
7	79	79	63	136	84	57	105	119	132	117	OOD
11	321	321	274	564	312	203	465	451	518	480	ID
13	555	555	402	914	522	365	777	781	862	747	ID
15	791	887	702	1400	792	515	1205	1227	1324	1185	ID
17	1205	1205	998	2032	1164	751	1717	1709	1908	1749	ID
19	1721	1721	1370	2862	1626	1045	2415	2387	2698	2415	ID
21	2243	2315	1886	3858	2178	1503	3249	3217	3614	3228	ID
29	5979	6051	4966	10180	5784	3887	8589	8537	9580	8565	OOD
30	6603	6699	5486	11298	6402	4385	9457	9503	10558	9477	OOD

B Full Quantitative Results

Table 5 reports every metric on the ID and OOD splits (mean ± standard deviation over three seeds), Table 6 the ID→OOD degradation ratios, Table 7 the full frontier-radius sweep, and Table 8 the power-law scaling fits. For DiffCSP and MatterGen the ∼500K-budget runs are numerically divergent (Appendix D), so their absolute magnitudes report the scale of divergent outputs rather than meaningful structural error.

Table 5: Per-model metrics on ID and OOD splits (mean ± std, three seeds). RMSD, BondMAE, RgError in Å; CoordCorr and SurfIntRatio dimensionless.

	RMSD		BondMAE		RgError		CoordCorr		SurfIntRatio	
Model	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ADiT	7.15±0.00	8.05±0.00	2.55±0.01	2.58±0.01	0.98±0.00	0.98±0.00	0.00	0.00	1.96±0.00	1.96±0.00
CDVAE	17.46±0.00	19.64±0.00	3.39±0.00	3.32±0.00	1.24±0.00	1.24±0.00	0.00	−0.00	1.09±0.00	1.09±0.00
DiffCSP	3386.5±134.2	2944.6±430.0	256.6±128.6	536.4±125.6	486.5±12.3	604.6±66.7	0.00	0.00	1.01±0.02	1.03±0.03
FlowMM	11.55±0.02	13.08±0.08	0.55±0.04	0.73±0.10	0.28±0.00	0.29±0.03	0.02±0.00	0.02±0.02	1.24±0.00	1.24±0.01
MatterGen	5904.6±1175.8	6408.6±873.4	1751.6±328.9	2953.3±313.7	884.0±162.9	1511.5±164.3	0.00	0.00	1.01±0.00	1.01±0.00

Table 6: ID→OOD degradation ratios (OOD mean / ID mean). CoordCorr is omitted: its ID and OOD values are ≈ 0 for several models, making the ratio uninformative (see absolute values in Table 5).

Model	RMSD	BondMAE	RgError	SurfIntRatio
ADiT	1.13	1.01	0.99	1.00
CDVAE	1.12	0.98	1.00	1.00
DiffCSP	0.87	2.09	1.24	1.02
FlowMM	1.13	1.33	1.05	1.00
MatterGen	1.09	1.69	1.71	1.00

Table 7: Frontier radius $r^\star(m,\tau)$ in Å across thresholds. "–" = no radius satisfies the threshold. ADiT extends farthest under RMSD; FlowMM is the only model with a finite BondMAE frontier; DiffCSP and MatterGen have none.

	RMSD threshold τ (Å)						BondMAE threshold τ (Å)					
Model	1.0	2.0	5.0	8.0	10.0	15.0	0.3	0.5	0.6	0.7	0.8	1.0
ADiT	–	–	11	17	21	30	–	–	–	–	–	–
CDVAE	–	–	–	7	7	13	–	–	–	–	–	–
DiffCSP	–	–	–	–	–	–	–	–	–	–	–	–
FlowMM	–	–	7	11	13	19	–	–	21	21	30	30
MatterGen	–	–	–	–	–	–	–	–	–	–	–	–

Table 8: Power-law fit $\log_{10}\mathrm{RMSD} = \alpha \log_{10} N + c$ on ID radii. OOD residual is the mean absolute prediction error when the ID fit is extrapolated to OOD radii.

Model	α	R^2	Intercept	OOD residual
ADiT	0.334	1.000	−0.176	0.001
CDVAE	0.335	1.000	0.209	0.004
DiffCSP	0.142	0.928	3.107	0.118
FlowMM	0.342	1.000	0.007	0.004
MatterGen	−0.126	0.897	4.145	0.050

C Metric Sensitivity Analysis

We sweep the two free parameters of the structural metrics. The CoordCorr cutoff r_c (Table 9) yields cross-architecture rankings that are identical to the default $r_c = 3.0$ Å for all $r_c \in [3.0, 4.5]$ Å (Spearman $\rho = 1.0$) and nearly identical at 5.0 Å ($\rho = 0.9$); only the very short cutoffs 2.0–2.5 Å, which capture partial first shells, reorder models. The SurfIntRatio shell fraction (Table 10) leaves the ID→OOD shift small across 15–30%: at the default 25% every model satisfies $|\Delta| \le 0.0227$, and the maximum over all fractions is 0.0475. Finally, a per-atom BondMAE that compares each atom's neighbor list individually (Table 11) reproduces the global-metric ranking exactly (Spearman $\rho = 1.0$). Per-material breakdowns of all three sweeps are consistent with these aggregates and are released with the benchmark repository.

Table 9: CoordCorr cutoff sweep: Spearman ρ of the cross-architecture ranking against the default $r_c = 3.0$ Å.

r_c (Å)	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Spearman ρ	−0.20	−0.10	1.00	1.00	1.00	1.00	0.90

Table 10: SurfIntRatio ID→OOD shift $\Delta = \mathrm{OOD} - \mathrm{ID}$ across shell fractions. The default fraction is 25%.

Fraction	ADiT	CDVAE	DiffCSP	FlowMM	MatterGen
0.15	+0.0463	+0.0001	+0.0475	−0.0019	+0.0023
0.20	+0.0379	−0.0012	+0.0276	−0.0017	+0.0032
0.25	+0.0012	−0.0025	+0.0227	−0.0023	+0.0037
0.30	−0.0066	−0.0025	+0.0146	−0.0033	+0.0031

Table 11: Per-atom vs. global BondMAE (Å). The per-atom variant compares each atom's neighbor list individually; model ranking is unchanged (Spearman $\rho = 1.0$).

	Per-atom		Global	
Model	ID	OOD	ID	OOD
ADiT	2.55	2.58	2.55	2.58
CDVAE	3.40	3.33	3.39	3.32
DiffCSP	256.60	536.35	256.60	536.35
FlowMM	0.74	0.91	0.55	0.73
MatterGen	1751.58	2953.33	1751.58	2953.33

D Published-Scale Capacity Experiment

To separate capacity limits from architectural ones, the two models that diverged at the shared ~500K budget were retrained near their published parameter counts: MatterGen at 47M parameters (hidden dimension 1100, 5 layers) and DiffCSP at 12.4M parameters (hidden dimension 465, 6 layers). To keep these larger runs tractable they used faster training settings than the main benchmark: half the training data (loaded_frac = 0.5) and 50 epochs, with the evaluation pipeline otherwise unchanged; the numbers here are therefore an indicative capacity probe rather than a capacity-matched re-benchmark.

Table 12 contrasts the two outcomes. MatterGen at 47M recovers stable, bounded sampling: per-radius RMSD grows monotonically with R (from 12.05 Å at R=11 to 23.19 Å at R=21) and peaks at the farthest OOD radii (33.19 Å at R=30), so the extrapolation frontier remains visible even at published scale. Per-material RMSD is uniform (ID 17.56–17.66 Å, OOD 19.74–19.84 Å), and predictions are bounded within ±113 Å (ID) and ±169 Å (OOD). DiffCSP at 12.4M, in contrast, still diverges: reverse-diffusion sampling saturates the ±10,000 Å numerical safety bound for all ten materials across both splits, even though training loss converges normally (final validation ≈ 1.10). Its per-radius RMSD shows no monotonic size trend, with error largest at the smallest radii, consistent with lattice-diffusion runaway rather than a size-generalization frontier. Capacity therefore governs whether sampling is stable at all (MatterGen) but does not, by itself, resolve the failure for every architecture (DiffCSP).

Table 12: Per-radius RMSD (Å) at published scale. MatterGen-47M is stable and grows monotonically with R; DiffCSP-12.4M remains bound-clipped and divergent, so its values report the magnitude of divergent outputs, not structural error.

Radius (Å)	Split	MatterGen-47M	DiffCSP-12.4M
6	OOD	6.44	2,314.9
7	OOD	7.54	2,236.1
11	ID	12.05	1,688.9
13	ID	14.28	1,708.0
15	ID	16.53	1,834.3
17	ID	18.74	2,060.9
19	ID	20.97	1,954.9
21	ID	23.19	1,781.1
29	OOD	32.07	1,669.2
30	OOD	33.19	1,655.1
Overall (ID / OOD)		17.63 / 19.81	1,838.0 / 1,968.8