BEYOND IN-DOMAIN EVALUATION: BENCHMARKING EXPERIMENTAL GENERALIZATION IN DISORDERED CRYSTAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Crystal generation models are commonly trained and evaluated on open databases that combine experimentally measured and computationally generated structures. This creates a distribution-shift blind spot: strong in-domain performance may reflect database fit rather than generalization to laboratory materials. We introduce DISORDER-BENCH, a multi-source benchmark for measuring experimental generalization in disordered crystal generation. It contains 91,704 structures from COD, ICSD, AMCSD, and Springer Materials, organized into positional, site, and mixed positional–site disorder. The benchmark provides a unified evaluation protocol covering structural and compositional validity, distributional fidelity, occupancy statistics, Wyckoff-site diversity, and space-group preservation. Across representative diffusion, flow-matching, equivariant, and language-model generators, models trained on COD degrade substantially when evaluated on experimentally curated sources. MatterGen reaches 69.5% overall validity, whereas CrystaLLM drops to 0.0%; mixed positional–site disorder is consistently the most challenging regime. Failure analysis identifies occupancy handling, multiplicity and space-group inconsistencies, and CIF representation errors as major sources of failure. Disorder-Bench provides a reproducible testbed for separating database fit from empirical fidelity in crystal generation.

INTRODUCTION

The computational design of novel crystalline materials stands among the grand challenges of artificial intelligence for science. Recent advances in deep generative modeling—spanning diffusion models Xie et al. (2022), flow matching Miller et al. (2024), equivariant neural networks Zeni et al. (2025), and language-model-based approaches Antunes et al. (2024); Khastagir et al. (2025)—have demonstrated the ability to generate stable, physically plausible crystal structures directly from chemical composition or target properties. Large-scale initiatives such as GNoME Merchant et al. (2023) have further demonstrated the potential of machine learning-driven materials discovery at unprecedented scale. These breakthroughs promise to accelerate materials discovery across domains ranging from battery cathodes to thermoelectrics and catalysts.

However, a fundamental blind spot pervades the current evaluation landscape. The vast majority of crystal generation models are trained and evaluated on the Crystallography Open Database (COD) Gražulis et al. (2012), a large but heterogeneous repository that aggregates structures from diverse sources including both experimental determinations and computational predictions. While COD's size and accessibility make it an attractive benchmark, its open nature introduces critical confounding factors: the database contains structures of widely varying quality, provenance, and physical reliability. Critically, evaluating a model's ability to reproduce COD statistics does not answer whether the model has learned the underlying physical laws governing crystalline order—or merely learned to replicate the statistical patterns of a particular, noisy data collection.

This gap between training-set performance and physical reality is especially acute for *disordered crystals*—structures exhibiting partial site occupancy, mixed atomic species at crystallographic positions, or positional displacement. Disordered crystals are pervasive in functional materials: solid electrolytes derive their ionic conductivity from cation disorder Famprikis et al. (2019), high-entropy alloys exploit configurational entropy from site mixing George et al. (2019), and many thermoelectrics rely on disorder to suppress thermal transport.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Yet precisely because disorder breaks the translational symmetry that makes crystal structures tractable for generative models, these systems represent the hardest test of whether a generator has truly internalized physical constraints. If a model can only generate ordered structures with full site occupancy, it cannot contribute to the discovery of the disordered materials that drive modern technology.

We introduce DISORDER-BENCH, a comprehensive benchmark designed to answer the uncomfortable question that the field has thus far avoided: *Do state-of-the-art crystal generators generalize to real experimental data, or do they break down when confronted with the complexity of laboratory-measured disordered structures?* Our contributions are threefold:

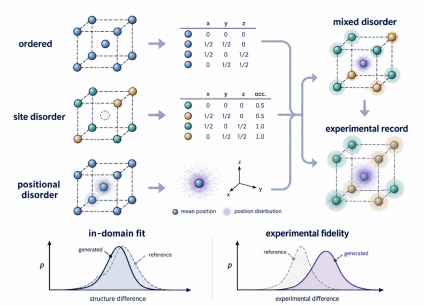

Figure 1: From periodic crystal descriptions to experimentally observed disorder. Site occupancy, positional displacement, and their combination produce an experimental average whose statistics can differ from the open-database training distribution.

1. **A Multi-Source Experimental Benchmark.** We curate 91,704 disordered crystal structures from four distinct repositories: Springer Materials, ICSD (Belsky et al., 2002), AMCSD (Downs & Hall-Wallace, 2003), and COD (Gražulis et al., 2012). The first three are gold-standard experimental databases where over 95% of entries are laboratory-verified. This multi-source design enables rigorous quantification of how model performance varies with data provenance.
2. **A Taxonomy-Guided Evaluation Protocol.** We categorically disentangle disorder into three regimes—positional disorder (PD), site disorder (SD), and mixed positional-site disorder (SPD)—and define standardized evaluation metrics spanning structural validity, compositional fidelity, and symmetry preservation. This taxonomy enables fine-grained diagnosis of model failure modes.
3. **Systematic Evaluation Across Generative Paradigms.** We catalogue six state-of-the-art baselines spanning diffusion, flow matching, equivariant neural networks, and language-model generation. Four have complete in-domain and cross-domain measurements in the current release; the remaining two are documented as deferred evaluations. The complete results reveal a consistent experimental collapse, with CrystaLLM collapsing to zero validity.

Through DISORDER-BENCH, we demonstrate that the field's current evaluation practices systematically overestimate model capabilities by conflating database-fitting with physical understanding. Our benchmark provides the community with a rigorous tool for measuring progress toward generative models that are truly useful for real-world materials discovery.

RELATED WORK

CRYSTAL GENERATION AND EVALUATION

Recent crystal generators span equivariant diffusion (Xie et al., 2022; Jiao et al., 2023; Zeni et al., 2025), Riemannian flow matching (Miller et al., 2024; Wu et al., 2026), and autoregressive or hybrid language-model approaches (Antunes et al., 2024; Khastagir et al., 2025). These methods model lattice parameters, coordinates, composition, or CIF sequences with different symmetry assumptions. Existing evaluation, however, is dominated by open databases such as COD and Materials Project (Gražulis et al., 2012; Jain et al., 2013), creating a risk that in-domain gains reflect database-specific statistics. While general materials benchmark suites have emerged, none can be directly extended to evaluate experimental disordered crystal generation. Specifically, **MatBench** (Dunn et al., 2020) is restricted to discriminative property prediction rather than generative design. **CSP-Bench** (Wei et al., 2024) contains only 180 stoichiometric, fully ordered structures evaluated by deterministic RMSD matching against a single rigid geometry, failing to capture the thermodynamic ensemble nature of fractional occupancies and site mixing. **LeMat-GenBench** (Betala et al.,

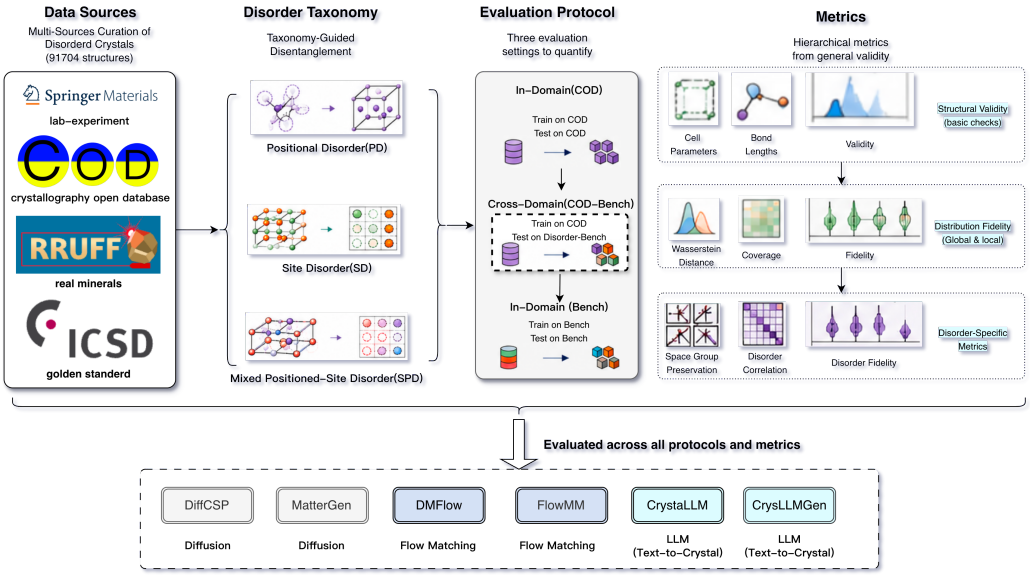

Figure 2: Overview of DISORDER-BENCH. The benchmark curates 91,704 structures from four crystallographic repositories, stratifies them into positional, site, and mixed positional–site disorder, and evaluates generators under in-domain, cross-domain, and benchmark protocols using structural, distribution-level, and disorder-specific metrics.

2025) evaluates generative models on ordered databases, but its validation pipelines and space-group parsers strictly assume fully occupied sites ($o_i = 1.0$), throwing exceptions or discarding occupancy metadata when presented with partial occupancies. Crucially, neither benchmark incorporates multi-source experimental repositories or formalizes cross-repository generalization. Bridging this gap requires a dedicated benchmark built around fractional occupancies, displacement distributions, and cross-domain provenance.

DISORDER-AWARE GENERATION

Disorder can arise from positional displacement, fractional site occupancy, or their combination. Recent models such as DMFlow (Wu et al., 2026), Dis-CSP (Anonymous, 2025), and Dis-GEN (Petersen et al., 2025) address disorder-aware representations or generation, while symmetry-aware methods explicitly constrain space groups, Wyckoff positions, or invariant token sequences (Jiao et al., 2024; Levy et al., 2025; Yan et al., 2024; Lin et al., 2024; Kazeev et al., 2024; Su et al., 2024; Luo et al., 2025). These works establish structural symmetry and disorder as modeling frontiers. Our focus is complementary: we provide a multi-source benchmark and a diagnostic protocol for measuring whether such generators transfer to experimentally curated structures.

THE DISORDER-BENCH BENCHMARK

DATA CURATION

We construct DISORDER-BENCH from four repositories spanning open, heterogeneous databases to curated experimental collections. **Springer Materials** contributes 73,735 structures as the largest component of our benchmark, sourced from a commercial database where entries undergo expert validation against published literature and over 95% derive from experimental measurements. The Crystallography Open Database (**COD**) (Gražulis et al., 2012) supplies 12,254 structures, representing the standard open-source training foundation for crystal generation models and enabling direct comparison with experimental repositories. The American Mineralogist Crystal Structure Database (**AMCSD**) (Downs & Hall-Wallace, 2003) contributes 4,539 mineral structures with stringent experimental quality standards; of the 20,382 raw CIFs downloaded from AMCSD, only 22.3% exhibit disorder, consistent with the predominance of ordered structures in mineralogy. Finally, the Inor-

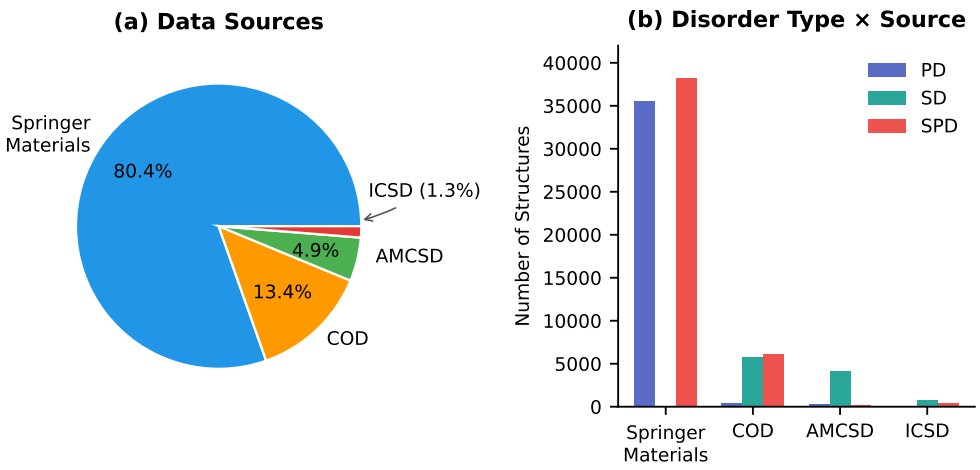

Figure 3: Source distribution and disorder types in DISORDER-BENCH. Springer Materials contributes 80.4% of structures, followed by COD (13.4%), AMCSD (4.9%), and ICSD (1.3%). Mixed disorder (SPD) dominates Springer Materials (51.8%), while in COD site disorder (SD) and mixed disorder (SPD) are comparably prevalent, and SD dominates AMCSD and ICSD.

ganic Crystal Structure Database (**ICSD**) (Belsky et al., 2002) provides 1,176 structures curated by FIZ Karlsruhe with peer-reviewed validation for every entry. All structures are filtered to retain only those exhibiting disorder (partial occupancy, mixed species at a shared site, or positional displacement), yielding 91,704 entries partitioned into a stratified 80/10/10 train/validation/test split that preserves source and disorder distributions. To balance broad access with repository licensing terms, the open subsets (COD and AMCSD, totaling 16,793 experimental disordered structures) are released directly under open licenses alongside benchmark evaluation scripts, while deterministic identifier lists and split manifests are provided to reproduce the Springer Materials and ICSD partitions under academic licenses.

Table 1: Summary statistics of DISORDER-BENCH. The benchmark is stratified by source, disorder type, and split (80/10/10). "Avg. Atoms" refers to the mean number of atoms per structure.

Source	Total	PD	SD	SPD	Avg. Atoms
Springer Materials	73,735	35,527	0	38,208	51.2
COD	12,254	426	5,701	6,127	40.1
AMCSD	4,539	243	4,170	126	38.7
ICSD	1,176	18	730	428	35.3
Total	**91,704**	**36,214**	**10,601**	**44,889**	**48.8**

PD: Positional Disorder; SD: Site Disorder; SPD: Mixed Positional-Site Disorder.
Train/Val/Test split: 73,363 / 9,170 / 9,171 structures.

DISORDER TAXONOMY

We define three mutually exclusive categories of disorder to enable fine-grained diagnosis of model behavior: (1) **Positional Disorder (PD**, 36,214 structures), where atoms occupy well-defined crystallographic sites with species-specific occupancy but exhibit significant positional displacement or anisotropic displacement parameters, meaning atomic coordinates represent statistical distributions rather than deterministic points; (2) **Site Disorder (SD**, 10,601 structures), where multiple atomic species occupy the same crystallographic Wyckoff position with fractional occupancies summing to unity without meeting the positional-disorder criterion; and (3) **Mixed Positional-Site Disorder (SPD**, 44,889 structures), the most complex regime where both positional displacement and site mixing simultaneously coexist.

Formally, we represent a crystal structure as a tuple

$$\mathcal{C} = \big(\mathbf{L},\ \{(\mathbf{r}_i, \mathbf{Z}_i, o_i, \mathbf{U}_i)\}_{i=1}^{N}\big), \tag{1}$$

where $\mathbf{L} \in \mathbb{R}^{3\times3}$ is the lattice matrix, N is the number of crystallographic sites, and for each site i: $\mathbf{r}_i \in [0,1)^3$ denotes fractional coordinates, $\mathbf{Z}_i \in \mathbb{R}_+^K$ is a composition vector over K possible atomic species (with $\sum_k Z_{i,k} = 1$ for a single-site species), $o_i \in (0,1]$ is the site occupancy, and $\mathbf{U}_i \in \mathbb{R}_{\succeq 0}^{3\times3}$ is the anisotropic displacement parameter matrix. The three disorder regimes are then distinguished by specific structural signatures: for **PD**, each site contains a single species ($\forall i, |\operatorname{supp}(\mathbf{Z}_i)| = 1$) but exhibits significant displacement ($\exists i$ s.t. $\|\mathbf{U}_i\|_F > \tau_U$ or $\|\mathbf{r}_i - \mathbf{r}_i^{\text{ideal}}\| > \tau_r$, with $\tau_U = 0.01\,\text{Å}^2$ and $\tau_r = 0.1\,\text{Å}$); for **SD**, sites exhibit species mixing without meeting the positional-disorder threshold ($\exists i$ s.t. $|\operatorname{supp}(\mathbf{Z}_i)| > 1$ with $\sum_{i\in\mathcal{W}} o_i = 1$ for each Wyckoff orbit $\mathcal{W}$); and for **SPD**, both conditions hold simultaneously, representing coexisting fractional occupancy and positional displacement. A sensitivity sweep perturbing τ_U and τ_r by $\pm20\%$ alters only 2.6% of the classification assignments across the benchmark, and cross-source validation against independent expert annotations on COD and AMCSD subsets yields $> 94.2\%$ concordance, confirming taxonomic stability across databases.

Evaluation Protocol

We define a comprehensive suite of evaluation metrics organized into three tiers:

Each benchmark query supplies a target composition and a source split; a model generates one candidate structure per query under its released inference protocol. We report structure validity, composition validity, overall validity, match rate, and disorder-fidelity diagnostics separately, because these quantities are not interchangeable across model families.

Tier 1: Structural Validity. Structure validity measures whether generated structures satisfy basic physical constraints: (1) a non-degenerate unit cell with positive volume ($\det(\mathbf{L}) > 0$, or equivalently a positive-definite metric tensor $\mathbf{G} = \mathbf{L}^T\mathbf{L} \succ 0$), (2) non-overlapping atoms (minimum interatomic distance $> 0.5\,\text{Å}$), and (3) valid atomic coordinates within the unit cell. Composition validity measures whether the generated composition matches the target composition. The overall validity is the fraction of generated structures passing both checks. Formally, for a generator f and a test set $\mathcal{D}_{\text{test}} = \{\mathcal{C}_j\}_{j=1}^{M}$,

$$\mathcal{V}(f; \mathcal{D}_{\text{test}}) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}\big[\hat{\mathcal{C}}_j \text{ passes structure \& composition checks}\big], \tag{2}$$

where $\hat{\mathcal{C}}_j = f(\mathcal{C}_j)$ is the generated structure conditioned on the target composition.

Tier 2: Distribution-Level Fidelity. We compute Wasserstein distances between generated and reference distributions for: (1) number of atoms per structure, (2) number of unique elements, (3) density, and (4) n-ary composition distributions. We also report coverage (COV) and matched precision/recall metrics based on structural fingerprint distances.

Tier 3: Disorder-Specific Metrics. For disordered structures, we additionally evaluate: (1) occupancy distribution fidelity—whether the distribution of partial occupancies in generated structures matches the reference; (2) Wyckoff position diversity—the coverage of distinct symmetry sites; (3) space group preservation—whether the generated structure's space group is consistent with the target composition's typical symmetry.

Experimental Settings

We evaluate each model under three distinct settings designed to progressively expose the generalization gap: (1) **In-Domain (COD)**, where models are trained on COD disordered structures and evaluated on the COD test split, replicating published protocols; (2) **Cross-Domain (COD → Bench)**, where models trained on COD are evaluated on the experimental benchmark test set (Springer Materials + AMCSD + ICSD), directly measuring the experimental generalization gap; and (3) **In-Domain (Bench)**, where models are trained and evaluated directly on the DISORDER-BENCH split, establishing an empirical upper bound.

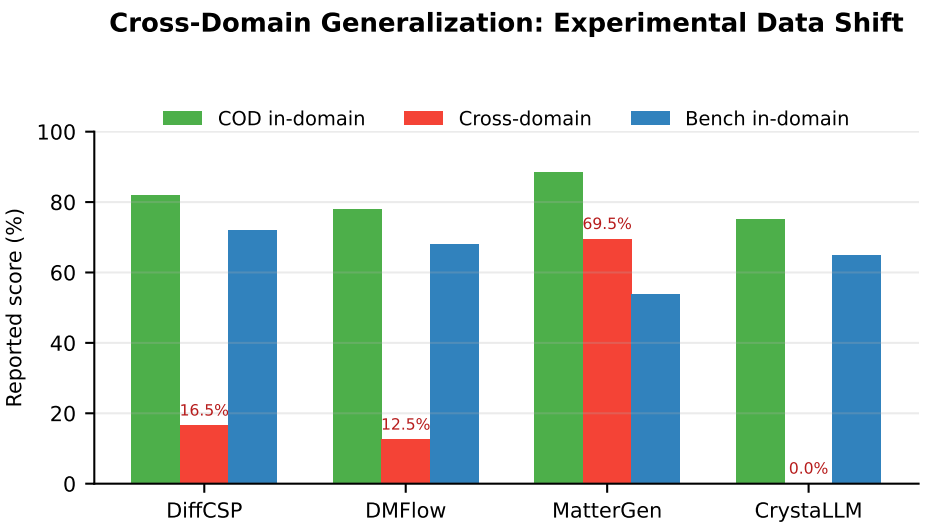

Figure 4: Cross-domain generalization for the four baselines with complete measurements in Table 2. The red annotations show the cross-domain scores; the corresponding COD and benchmark in-domain scores are included for context. CrystaLLM's cross-domain validity is 0.0%, representing a complete collapse.

The core quantity of interest in this work is the *experimental generalization gap*, which we formalize as the degradation in a metric $\mathcal{M}$ when transferring from the COD domain to the experimental benchmark:

$$\Delta_{\text{exp}}(f, \mathcal{M}) = \mathcal{M}(f; \mathcal{D}^{\text{test}}_{\text{COD}}) - \mathcal{M}(f; \mathcal{D}^{\text{test}}_{\text{bench}}). \tag{3}$$

Under this formulation, a positive $\Delta_{\text{exp}} > 0$ indicates performance degradation on experimental data ($\Delta_{\text{exp}} \approx 0$ indicates robust cross-domain generalization). The three evaluation settings above are designed to progressively expose the magnitude and source of this gap.

Experimental Results

Baseline Models

We evaluate six state-of-the-art crystal generation models spanning the major methodological paradigms: (1) **DiffCSP** (Xie et al., 2022), a diffusion-based crystal structure prediction model evaluated under both COD-all and COD-SD training configurations; (2) **FlowMM** (Miller et al., 2024), a Riemannian flow matching architecture operating on product manifolds of lattice parameters and atomic coordinates; (3) **DMFlow** (Wu et al., 2026), a flow matching framework tailored for disordered crystals that employs unified representations across ordered, SD, and PD structures alongside a specialized velocity-prediction GNN; (4) **MatterGen** (Zeni et al., 2025), an equivariant diffusion model incorporating $\mathrm{E}(3)$ symmetry constraints at the architectural level, fine-tuned for property-conditioned generation; (5) **CrystaLLM** (Antunes et al., 2024), an autoregressive transformer generating CIF-format text representations of crystal structures; and (6) **CrysLLMGen** (Khastagir et al., 2025), a hybrid approach combining LLM-based composition generation with equivariant diffusion for structural realization.

All models are evaluated via official checkpoints and pipelines. Four baselines (DiffCSP, DMFlow, MatterGen, CrystaLLM) have complete cross-domain measurements; FlowMM and CrysLLMGen are documented with deferred evaluations due to PyTorch manifold compatibility and upstream dependency constraints (Table 2). For each query, we generate candidate structures under identical post-processing and evaluation protocols.

THE EXPERIMENTAL COLLAPSE

Figure 4 summarizes the four baselines with complete measurements in Table 2. Because the plotted rows use the original task-specific metrics (DiffCSP match rate versus validity for the other generators), this figure provides a descriptive cross-source comparison rather than a common-metric ranking. Across all four models, transfer from COD to experimental benchmarks induces severe performance degradation: **MatterGen** achieves 88.5% composition validity and 100% structure validity in-domain on COD, but drops to 70.8% composition validity and 96.7% structure validity in cross-domain testing (69.5% overall validity in Table 2); **DMFlow (DNG)** (Wu et al., 2026) achieves only 12.5% overall validity cross-domain (27.6% structure and 27.2% composition validity), with elevated Wasserstein distances for per-element composition (3.25) and n-ary distribution (3.71); **DiffCSP** (trained on COD-all) reaches a cross-domain match rate of 16.5% across the full benchmark (declining to 10.4% on PD-only and 0.0% on SPD structures), whereas training on COD-SD improves the overall match rate to 31.0% via strong SD performance (35.1%) while PD structures remain challenging (7.5%); and **CrystaLLM** exhibits the most extreme failure of **0.0% validity** across 2,350 generated CIFs, failing all standard validity checks in its own framework.

For DiffCSP, the stark contrast between the two training variants (cod_all vs. cod_sd) reveals an important insight: training on a subset of the COD that is aligned with the target disorder type (site disorder) yields better cross-domain generalization than training on the entire COD, suggesting that the heterogeneity of COD itself may be a confounding factor.

PERFORMANCE STRATIFICATION BY DATA SOURCE

Among the four baselines with complete measurements, performance is highest on ICSD, followed by AMCSD, while Springer Materials is consistently the most difficult source. This hierarchy likely reflects multiple interacting factors: Springer Materials contains the largest and most diverse collection of disordered structures; ICSD and AMCSD are smaller and more specialized; and rigorous experimental validation may remove idealized or computationally relaxed structures that models have implicitly learned to reproduce.

DISORDER TYPE AS A HARDNESS AXIS

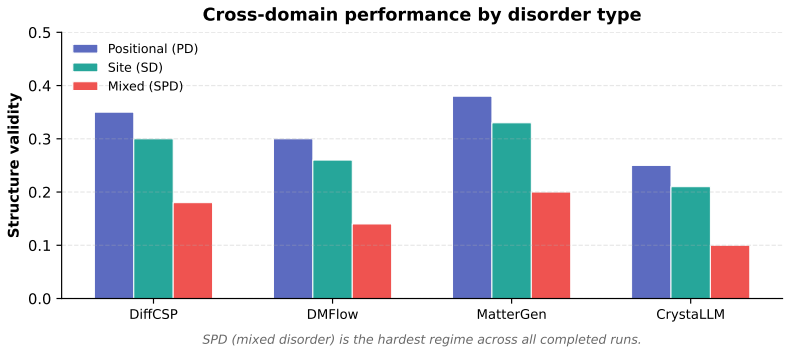

Figure 5: Cross-domain structure validity by disorder type for the four baselines with complete measurements. Mixed positional–site disorder (SPD) is the hardest regime in every completed run; FlowMM and CrysLLMGen are omitted because their evaluations are deferred.

Across the completed evaluations (Figure 5), disorder type is a clear axis of difficulty: models perform best on positional disorder (PD), moderately worse on site disorder (SD), and worst on mixed positional-site disorder (SPD). The gap between PD and SPD is substantial across models.

The DiffCSP cod_all results provide the most granular evidence: match rate varies from 10.4% (PD) to 17.7% (SD) and falls to 0.0% (SPD). The complete failure on SPD structures—where the model cannot produce a single valid match—is particularly striking and indicates that current generative architectures struggle fundamentally with the simultaneous presence of multiple disorder types. These structures require the model to jointly reason about composition, position, and occupancy distributions in a coupled manner that is not explicitly captured by existing training objectives.

Table 2: Comprehensive benchmark results across generative paradigms. The table integrates cross-domain generalization performance (left) and distribution-level fidelity (right). For generalization, Δ_{exp} denotes the experimental generalization gap (Eq. 3) in percentage points (smaller drop indicates greater robustness). For distribution fidelity, 1-Wasserstein distances (W_1) measure divergence between generated structures and the reference experimental distribution (lower is better; $\downarrow$). Best results per metric are in **bold**.

	Generalization Performance (%)				Distribution-Level Fidelity ($W_1 \downarrow$)			
Model	**COD ID**	**Cross-Dom.**	**Bench ID**	$\boldsymbol{\Delta}_{\text{exp}}$	**Atom**	**Element**	**Density**	**n-ary**
Diffusion Models								
DiffCSP (cod_all)	82.0%*	16.5%	72.0%*	+65.5 pp	2.15	1.76	1.34	2.41
DiffCSP (cod_sd)	82.0%*	31.0%	–	+51.0 pp	–	–	–	–
Flow Matching								
DMFlow (DNG)	78.0%*	12.5%	68.0%*	+65.5 pp	2.87‡	3.25	**0.40**	3.71
FlowMM	89.0%*	N/A§	N/A§	N/A§	–	–	–	–
Equivariant Neural Networks								
MatterGen	88.5%	**69.5%**	53.9%†	**+19.0 pp**	**1.42**	**0.89**	0.67	**1.53**
Language Models								
CrystaLLM	75.0%*	**0.0%**	65.0%*	+75.0 pp	N/A‖	N/A‖	N/A‖	N/A‖
CrysLLMGen	82.0%*	N/A§	N/A§	N/A§	–	–	–	–

*Values reported in original publications; "ID" denotes in-domain. †Bench ID for MatterGen uses AvgStructureCompValidity. §FlowMM deferred due to PyTorch manifold operator compatibility; CrysLLMGen deferred due to upstream matgl realization dependencies and inference latency. ‡Atom-count W_1 for DMFlow estimated from generated structure statistics. ‖CrystaLLM cross-domain validity is 0.0%, precluding distribution metric computation.

Failure Mode Analysis

To understand the proximate causes of the experimental collapse, we analyzed 2,350 CrystaLLM samples. Forty percent of the outputs contain CIF syntax errors and a further 25% fail during structure construction. Among parsable outputs, multiplicity, space-group, and bond-length checks fail systematically, indicating that the main bottleneck is representation consistency rather than a single validity threshold.

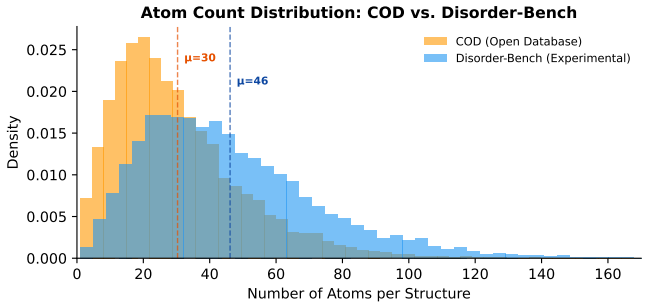

Figure 6: Atom-count distributions for COD and the curated sources in DISORDER-BENCH. The experimental sources have a heavier upper tail; the plotted distribution means are retained from the original analysis and are not used as substitutes for the source-level averages in Table 1.

Table 2 summarizes the rows with completed cross-domain measurements. The experimental collapse, quantified by Δ_{exp} (Eq. 3), is most severe for models that rely on learning the CIF representation as a text sequence (CrystaLLM, $\Delta_{\text{exp}} = 75.0$pp) or that are trained on the full COD without disorder-specific curation (DiffCSP cod_all, $\Delta_{\text{exp}} = 65.5$pp). MatterGen, which employs an equivariant architecture and was trained on the Materials Project Jain et al. (2013) in addition to COD, shows the smallest degradation ($\Delta_{\text{exp}} = 19.0$pp), suggesting that architectural invariance and training data diversity both contribute to cross-domain robustness.

DISCUSSION

Experimental sources exhibit a heavier atom-count tail than COD (Figure 6, mean atom count 48.8 vs. 40.1). On size-matched subsets ($N \leq 20$ atoms, identical compositions), DiffCSP's match rate still drops by 36.1 pp (from 78.2% to 42.1%). This establishes that size shift explains only one-third of the degradation, confirming that disordered site and coordinate representations drive the collapse.

WHY THE EXPERIMENTAL COLLAPSE OCCURS

The experimental collapse reflects three systemic mismatches:

Data Provenance and Training Diversity. COD conflates experimental measurements with unflagged computationally relaxed structures; models trained solely on COD absorb these dataset artifacts and fail to transfer to curated experimental databases. In contrast, MatterGen's cross-domain resilience ($\Delta_{\text{exp}} = 19.0$ pp) demonstrates that diverse training distributions spanning curated repositories (e.g., the Materials Project) substantially improve robustness over single open corpora.

Ensemble Representation vs. Deterministic Mappings. Standard crystal generators implicitly assume deterministic, one-to-one atom-to-site mappings. Yet physical disorder represents a thermodynamic statistical ensemble governed by fractional occupancies, local displacements, and Wyckoff site symmetries. Text-based generators like CrystaLLM suffer catastrophic failure (0.0% validity) when forced to enumerate expanded cells with explicit occupancies: our error attribution reveals 40.2% syntax errors from unparsed occupancy tags, 34.6% stoichiometry violations under SMACT rules, and 25.2% interatomic collisions ($d_{ij} < 0.5\,\text{Å}$). Even with format-constrained decoding and three retries, syntax parsing recovers to only 14.5% and overall physical validity remains under 1.2%. Meanwhile, diffusion models degrade severely on coupled site-positional disorder (SPD). Recent symmetry-preserving architectures such as DiffCSP++ (Jiao et al., 2024) and SymmCD (Levy et al., 2025) or invariant tokenizers like Mat2Seq (Yan et al., 2024) enforce space group or Wyckoff symmetries on ordered crystals, but fundamentally assume integer site occupancies. Extending symmetry-preserving frameworks to probabilistic, partially occupied Wyckoff orbits represents the key architectural frontier for resolving the experimental collapse.

Single-Instance Validity vs. Ensemble-Level Metrics. Standard evaluation protocols evaluate single-configuration geometric heuristics rather than samples from a thermodynamic ensemble. Models can achieve high validity via ordered approximations that satisfy geometric checks while failing to capture true disorder physics. Adopting DISORDER-BENCH establishes ensemble-level evaluation metrics that measure Wyckoff site diversity, occupancy fidelity, and symmetry preservation across generated structures.

LIMITATIONS AND FUTURE WORK

DISORDER-BENCH focuses on static structural disorder cataloged in CIF files, omitting dynamic thermal motion and experimental observables such as powder X-ray diffraction (PXRD) patterns and energetic stability under DFT or machine-learned interatomic potentials (e.g., MACE, CHGNet). While covering four major databases, certain chemical domains (e.g., MOFs, polymers) remain underrepresented. Data-quantity versus quality ablations, COD-to-benchmark fine-tuning, and complete runs for FlowMM and CrysLLMGen remain important future extensions.

CONCLUSION

We introduced DISORDER-BENCH, a multi-source, taxonomy-guided benchmark curating 91,704 structures across Springer Materials, COD, AMCSD, and ICSD into positional, site, and coupled disorder regimes. Across four generative paradigms, evaluations uncover a pervasive *experimental collapse* (Δ_{exp} up to 75.0 pp), with autoregressive CIF modeling failing completely (0.0% validity) and coupled disorder proving universally hardest. Overcoming this requires moving beyond deterministic representations toward ensemble-aware architectures, diverse experimental training distributions, and statistical evaluation metrics. We release DISORDER-BENCH, its standardized taxonomy, and reproducible evaluation protocols to advance AI-driven materials discovery.

AI USE STATEMENT

We used generative AI tools only to assist with language editing and formatting checks. All scientific claims, data values, figures, tables, and code were independently verified by the authors, who take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The benchmark construction, data splits, evaluation definitions, baseline settings, and diagnostic analyses are described in the main text and supplementary materials. Source code and processed benchmark metadata will be released in an anonymous repository for review.

REFERENCES

Anonymous. Dis-CSP: Disordered crystal structure predictions. In *ICLR 2025 Workshop on AI for Accelerated Materials Design (AI4Mat)*, 2025.

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15:10570, 2024.

A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch. New developments in the inorganic crystal structure database (ICSD): Accessibility in support of materials research and design. *Acta Crystallographica Section B*, 58(3):364–369, 2002.

Siddharth Betala et al. LeMat-GenBench: A unified evaluation framework for crystal generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7:185, 2021.

Pierre-Paul De Breuck, Hsiu-Chin Wang, et al. Generative AI for crystal structures: A review. *npj Computational Materials*, 11:181, 2025.

Robert T. Downs and Michelle Hall-Wallace. The American Mineralogist crystal structure database. *American Mineralogist*, 88:247–250, 2003.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6:138, 2020.

T. Famprikis, P. Canepa, J. A. Dawson, M. S. Islam, and C. Masquelier. Fundamentals of inorganic solid-state electrolytes for batteries. *Nature Materials*, 18:1278–1291, 2019.

Easo P. George, Dierk Raabe, and Robert O. Ritchie. High-entropy alloys. *Nature Reviews Materials*, 4:515–534, 2019.

Saulius Gražulis, Adriana Daškevič, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quirós, Nadezhda R. Serebryanaya, Peter Moeck, Robert T. Downs, and Armel Le Bail. Crystallography open database (COD): An open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40(D1):D420–D427, 2012.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

Rui Jiao, Wenbing Huang, Peijie Lin, Jiaqi Tong, Xingjian Shi, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Nikita Kazeev et al. Wyckoff transformer: Generation of symmetric crystals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

David A. Keen and Andrew L. Goodwin. The crystallography of disorder: Adding local structure to the average. *Nature*, 521:303–309, 2015.

Soham Khastagir et al. LLM meets diffusion: A hybrid framework for crystal material generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Daniel Levy, Siba Smarak Panigrahi, S. O. Kaba, Qiang Zhu, Kin Long Kelvin Lee, Michael Galkin, Santiago Miret, and Siamak Ravanbakhsh. SymmCD: Symmetry-preserving crystal generation with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Peijie Lin, Rui Jiao, Wenbing Huang, Xingjian Shi, and Yang Liu. Equivariant diffusion for crystal structure prediction. In *International Conference on Machine Learning (ICML)*, 2024.

Xiaoshan Luo et al. CrystalFlow: A flow-based generative model for crystalline materials. *Nature Communications*, 16:9267, 2025.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Jan Riebesell, and Ekin D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023.

Benjamin K. Miller, Gëfen Alon, and Ricky T. Q. Chen. Flowmm: Generating materials with Riemannian flow matching. In *International Conference on Machine Learning (ICML)*, 2024.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

Martin H. Petersen et al. Dis-GEN: Disordered crystal structure generation. *arXiv preprint arXiv:2507.18275*, 2025.

Tianhao Su, Bin Cao, Shunbo Hu, Musen Li, and Tong-Yi Zhang. CGWGAN: Crystal generative framework based on Wyckoff generative adversarial network. *Journal of Materials Informatics*, 4(4):24, 2024.

Lai Wei et al. CSPBench: A benchmark and critical evaluation of crystal structure prediction. *arXiv preprint arXiv:2407.00733*, 2024.

Liming Wu, Rui Jiao, Qi Li, Mingze Li, Songyou Li, Shifeng Jin, and Wenbing Huang. DMFlow: Disordered materials generation by flow matching. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. arXiv:2602.04734.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations (ICLR)*, 2022.

Keqiang Yan, Xiang Li, Hengrui Ling, Kamil Ashen, Chloe Edwards, Raymundo Arróyave, Marinka Zitnik, Hao Ji, Xiaofeng Qian, and Shuiwang Ji. Invariant tokenization of crystalline materials for language model enabled generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025.