Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Navigating Order-(Dis) Order Family Trees via Group-Subgroup Transitions

Shuya Yamazaki
Nanyang Technological University
Yuyao Huang
Nanyang Technological University
Martin Hoffmann Petersen
Nanyang Technological University
Wei Nong
Nanyang Technological University
Kedar Hippalgaonkar
kedar@ntu.edu.sg

Nanyang Technological University

Article

Keywords: Order-Disorder, Group-Subgroup Transitions, Space Group, Materials Discovery

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10128284/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

NAVIGATING ORDER-(DIS)ORDER FAMILY TREES VIA GROUP-SUBGROUP TRANSITIONS

Shuya Yamazaki
Nanyang Technological University

Yuyao Huang
Nanyang Technological University

Martin Hoffmann Petersen
Nanyang Technological University

Wei Nong
Nanyang Technological University

Kedar Hippalgaonkar*
Nanyang Technological University

ABSTRACT

As closed-loop materials discovery systems scale to produce millions of candidate compounds, the credibility of the novelty they reward becomes a critical concern. Novelty is commonly assessed against databases of ordered crystal structures, in which atomic sites are fully occupied. Yet, a predicted ordered structure may simply correspond to a particular ordering of a known disordered phase, whose sites are occupied by multiple species in the statistical average structure; we refer to such a structure as an ordered child of a disordered parent. Here, we introduce order-(dis)order family trees, a symmetry-based framework that organizes ordered and disordered structures through group-subgroup relations and enables novelty to be explicitly evaluated. We develop a high-throughput family matching procedure, to identify possible disordered parents and symmetry-related ordered relatives for a given ordered structure. As validation, we test our framework on synthesis-facing case studies (A-Lab), where it correctly recovers existing disordered parents for the targeted ordered structures. Extending this family-tree-based benchmark to experimental structure databases (ICSD), computational datasets (MP-20, Alex-MP-20, and GNoME), and crystal generative models further reveals that many ordered structures that appear novel as individual entries are, in fact, better understood as members of experimentally known order-(dis)order family trees. We also show that this is particularly evident in symmetry-agnostic all-atom generative models, which more frequently produce ordered structures derived from known disordered parents, whereas symmetry-constrained models are 2–4$\times$ less prone to this behavior. Our results establish order-(dis)order family trees as a key requirement for achieving genuine novelty in data-driven materials discovery.

Keywords Order-Disorder · Group-Subgroup Transitions · Space Group · Materials Discovery

1 Introduction

Materials discovery systems are pushing the field toward a closed loop between computational prediction and experimental synthesis. Agentic workflows now produce millions of candidate compounds at a rate that far outpaces experimental characterization, as they scale from proposing structures to directing robotic synthesis [1, 2]. Consequently, the central question is no longer simply whether we can predict stable structures beyond the training distribution, but whether the structures we predict are genuinely new to synthesis [3, 4]. In current practice, novelty is typically defined by absence from reference databases of ordered crystal structures [5, 6, 7]. Defining novelty in these systems has become a critical concern: a system that rediscovers known phases dressed as new compounds wastes synthesis cycles and corrupts its own reward.

Yet, answering whether a predicted structure is genuinely novel is more subtle than it first appears. Novelty is not an intrinsic property of a compound, but one defined relative to a reference: a structure is novel only with respect to the space against which it is compared, and the choice of that reference therefore shapes what counts as discovery. Although ordered structure databases are a natural reference given the data on which many discovery systems are

*Correspondence to kedar@ntu.edu.sg

trained, relying on them alone introduces a systematic blind spot. Experimentally synthesized materials frequently exhibit occupational disorder, in which the same crystallographic lattice sites are occupied by multiple species in the statistical average structure [8]; hereafter, disordered phase refers to this average crystallographic structure unless otherwise noted. In this setting, a predicted ordered structure may not correspond to a genuinely new compound, but instead to a specific ordering of a known disordered phase. Such a structure corresponds to one of many possible ordered configurations consistent with the same underlying disordered framework. We refer to such predictions as *ordered children* of *disordered parents* [3, 4]. When disordered phases are not included in the reference space, these ordered children are routinely labeled as novel, leading to a systematic overestimation of discovery rates relative to experimental reality.

Recent experimental efforts highlight this exact failure mode. In the A-Lab study [1], over 65% of the "successful" syntheses of GNoME compounds [2] were later found to lack the predicted cation ordering, and were instead more plausibly interpreted as already-known disordered parent phases [4]. An example is $MgTi_2NiO_6$: although the predicted structure was ordered, the synthesized product was more consistent with the known disordered ilmenite-type parent phase at the same composition. The same pattern has emerged in generative-model-driven discovery. In MatterGen [6], the experimentally synthesized product was not the predicted ordered phase but its disordered parent, subsequently identified as a known phase differing only slightly in site occupancy from previously reported structures [9].

These cases are unlikely to be isolated failures. Rather, they point to a structural gap in how computational predictions are interpreted. Ordered and disordered phases are often not unrelated alternatives, but members of the same crystallographic lineage [10]. This connection arises because atomic ordering directly governs crystal symmetry: a disordered phase, in which no specific arrangement of species over atomic sites is enforced, preserves the full symmetry of the underlying lattice, whereas the onset of ordering breaks that symmetry and produces a lower-symmetry structure. These symmetry reductions are not arbitrary but are constrained by group-subgroup relations in crystallography [10, 11, 12], governing the ordered structures that can descend from a given disordered parent. Multiple ordered structures can descend from the same parent through distinct symmetry-breaking pathways, forming a family of ordered siblings. When these relations are ignored, a structure can then be erroneously counted as a new compound even though it already occupies a position within a known order-disorder family.

Such relations are rarely available in a systematic, high-throughput form, and most reported examples have been uncovered only through detailed manual analysis of a small number of predicted structures subjected to experimental validation [3, 4, 9]. In principle, one could attempt exhaustive mapping by enumerating all symmetry-inequivalent ordered configurations of known disordered phases, for example using methods such as POCC [13], and then matching a query ordered structure against the resulting lookup space. In practice, however, extending such enumeration across the full set of known disordered compounds is computationally prohibitive and unsuitable for large-scale analysis. More fundamentally, this still does not provide a general framework for answering: (i) for a given ordered structure, what disordered parents it may derive from, (ii) what ordered siblings share the same parents, and, most importantly, (iii) whether its apparent novelty reflects a genuinely new family or merely a new member within an existing one.

To address this gap, we develop a symmetry-based framework that organizes ordered and disordered phases through **group-subgroup transitions** in space-group hierarchy: **order-(dis)order family trees**. In this view, a disordered parent is not only a possible precursor or competing interpretation of a predicted ordered phase, but also the root of a broader family tree whose branches include symmetry-lowered ordered descendants and related sibling structures. Group-subgroup transitions provide the natural language for navigating these trees, tracing how structural order emerges, splits, and reconnects across different levels of symmetry.

This family-tree perspective changes how computational discovery should be interpreted. Rather than treating each predicted structure as an isolated point, we view it as an element embedded in a manifold of related ordered and disordered phases. What is discovered computationally is therefore often not a standalone structure, but a position within an existing family tree. This distinction is particularly important during synthesis and characterization of predicted compounds, where experimentally realized phases may actually be disordered parents, or coexist with nearby siblings, or appear as partially ordered variants of the same family. In this framework, we develop a high-throughput procedure, built on the Symmetry and Wyckoff-sequence of Ordered and Disordered crystals representation (SWORD) [14], for identifying candidate disordered parents and symmetry-related family members of ordered query structures. Using these mappings, we then define disorder-aware family metrics that evaluate a structure set against both ordered and disordered known phases, rather than against ordered references alone. This yields a view of novelty that better reflects experimental reality by exposing cases where predicted ordered phases only appear novel because their known disordered counterpart is absent from the reference set, providing critical feedback to recent novelty assessments [5, 6].

We first validate our framework on the A-Lab cases, demonstrating that it recovers the disordered parent phases underlying a substantial fraction of the synthesis outcomes. We then introduce two new Family Tree-based novelty metrics, FT_{disorder} and FT_{order}, which are defined respectively as the fraction of ordered compounds connected through

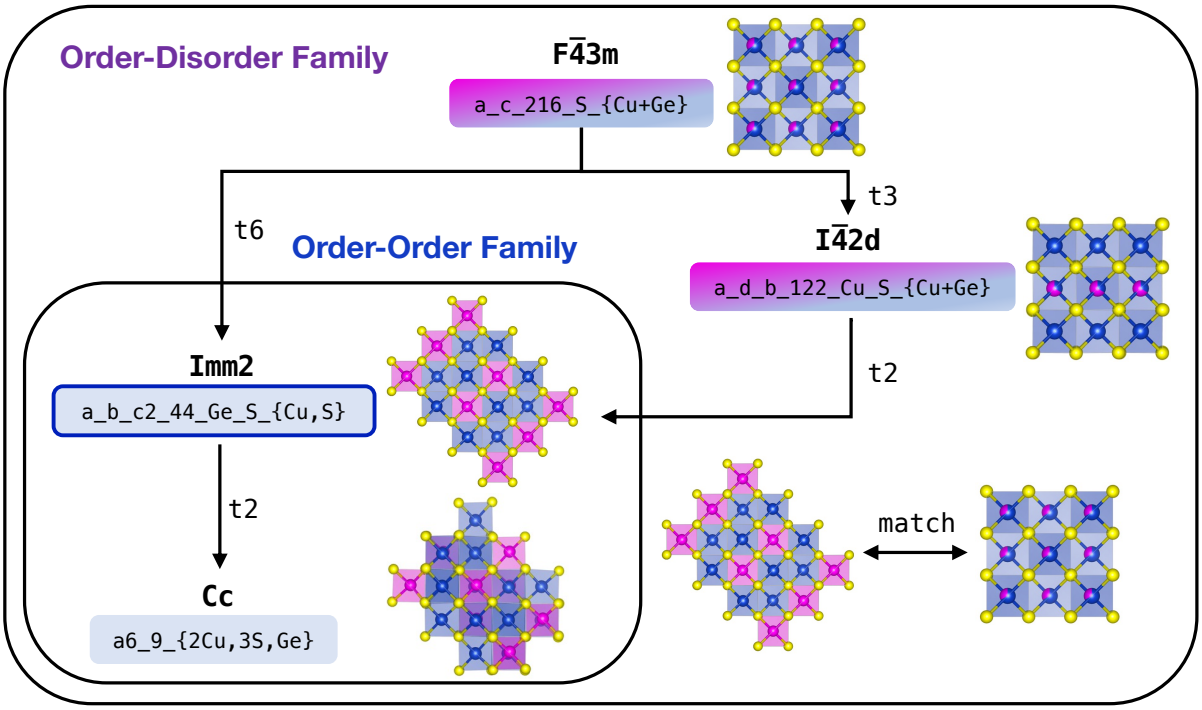


Figure 1: Order–(dis)order family tree recovered from the query structure Cu_2GeS_3 (`Imm2`, highlighted in bright blue). Each phase is labeled by its space group symbol and SWORD label, with parent-child links annotated by the subgroup index. The vertical arrangement follows point-group order, with higher-symmetry phases placed near the top. In this example, one of the disordered phases `a_c_216_S_{Cu+Ge}` occupies the highest-symmetry position and acts as the parent from which both ordered and disordered descendants are derived through group-subgroup transitions. Ordered children that share the same parent define an order-order subfamily within the broader order-(dis)order family tree. Two structures `match` if they share the same parent. Implementation details are provided in Appendix A.

order-disorder and order-order family trees to experimentally known disordered and ordered phases. We then apply these metrics across various experimental and computational structure databases (ICSD [15], MP-20 [5], Alex-MP-20 [6], GNoME [2]) and crystal generative models (DiffCSP [16], DiffCSP++ [17], FlowMM [18], MiAD [19], SymmCD [20], CrystalDiT [21], ADiT [22], MatterGen [6], Chemeleon2-RL [23], WyFormer [24]), finding that a consistent and substantial fraction of apparently novel structures are better understood as members of known order-(dis)order family trees. Therefore, we posit that novelty in materials discovery should be assessed not only by structural distinctness, but also by considering crystallographic lineage. Making order-(dis)order family trees explicit helps distinguish genuinely new families from symmetry-lowered rediscoveries of known phases, thereby providing a clearer bridge between computational prediction and experimental realization.

2 Methods

Systematic navigation of order–disorder family trees requires a crystal representation that is both physically meaningful and computationally tractable. Our framework builds on two components: (i) the symmetry hierarchy of crystals expressed through space groups, Wyckoff positions, and group-subgroup transitions [10, 25]; and (ii) a unified crystal representation describing both ordered and disordered structures within the same formal language. Together, these enable family-level reasoning where ordered children, disordered parents, and symmetry-related siblings are analyzed within a common framework.

2.1 Preliminaries

Space groups and Wyckoff positions The space group G constitutes an important aspect of a crystal structure, representing the set of symmetry operations that leave the lattice invariant. Each operation $g \in G$ acts on a point $\mathbf{r} \in \mathbb{R}^3$ via an affine transformation $\mathbf{r} \mapsto R\mathbf{r} + \mathbf{t}$, where R denotes a rotation or rotoinversion and $\mathbf{t}$ represents a translation. This group structure dictates the symmetry-equivalence of points within the crystal. For any specific point $\mathbf{r}$, the site-symmetry group is defined as the subgroup $G_{\mathbf{r}} = \{g \in G \mid g\mathbf{r} = \mathbf{r}\}$, containing all operations that leave $\mathbf{r}$ fixed. Points are formally categorized into Wyckoff positions, which are sets of symmetry-equivalent sites whose site-symmetry groups are conjugate within G [25]. Each Wyckoff position is uniquely characterized by its multiplicity, site symmetry, and fractional coordinates. Ultimately, Wyckoff positions provide a mathematically rigorous and compact description of crystal structures by grouping atomic sites according to their transformation properties under G.

Group-subgroup transitions Crystallographic relationships between structures are established using group–subgroup relations [10]. For two space groups G and H, the relation $H \leq G$ denotes that H is a subgroup of G, such that all symmetry operations in H are contained within G. A transition from G to H therefore corresponds to a reduction in symmetry. The degree of symmetry reduction is quantified by the subgroup index $[G : H] = |G|/|H|$, which counts how many cosets of H partition G. A particularly important relation to consider is the *translationengleiche* (t-type) subgroup relations, which preserve the translation group while lowering the point symmetry. These transitions are particularly relevant for tracing ordering processes in which previously equivalent sites become symmetry-distinct. Because t-type relations preserve the translational periodicity, this site-splitting accommodates ordering within the bounds of the original primitive cell, without introducing a supercell. Although *klassengleiche* (k-type) subgroup relations, which involve a loss of translational symmetry, can also be important, they are not the main focus here. Under a group–subgroup transition, Wyckoff positions may split, causing previously equivalent sites to become distinguishable. This splitting provides the crystallographic basis for linking parent, child, and sibling structures within the same family tree.

Disorder Experimental structures often deviate from an ideal fully ordered configuration and may exhibit disorder in either site occupancy or atomic position [26]. A broad and practically important class is *occupational disorder*, in which the occupation of a crystallographic site is statistical rather than uniquely assigned. This includes cases where multiple chemical species share the same site with fractional occupancies, as well as vacancy-containing cases where a site is only partially occupied. Another major form is *positional disorder*, in which atoms are distributed over multiple nearby positions instead of occupying a single well-defined site in the average structure. In this work, we focus primarily on occupational disorder and do not consider positional disorder, which constitutes less than 10% of ICSD entries [14].

2.2 SWORD Representation

Here, we use Symmetry and Wyckoff-sequence of Ordered and Disordered crystals (SWORD) [14], a unified and scalable representation where both ordered and disordered crystal structures are encoded. This enables consistent evaluation of family relations through space group symmetry-based analysis. SWORD expresses a crystal structure through symmetry-aware site assignments and occupancy information, making it particularly well suited for tracking how ordering, partial ordering, and disorder are distributed across Wyckoff sites. A SWORD label is written in the order of Wyckoff positions, space group number, and the atomic species occupying those positions. Using the disordered parent at the root of Figure 1 as an example, `a_c_216_S_{Cu+Ge}` denotes a structure in space group 216 in which Wyckoff position `WP a` is fully occupied by `Element S`, while Wyckoff position `WP c` is compositionally disordered between `Element Cu` and `Element Ge`. Likewise, one of its ordered children, `a_b_c2_44_Ge_S_{Cu, S}`, denotes a structure in space group 44 in which `WP a` is occupied by `Element Ge`, `WP b` by `Element S`, and the two distinct crystallographic orbits of `WP c` are occupied by `Element Cu` and `Element S`, respectively. This form of description is essential in our setting because the object of interest is not an isolated ordered phase, but the structural relation between an ordered derivative and its disordered or ordered relatives. By adopting a unified representation for both ordered and disordered phases, they can be compared at the level where order-disorder relationships are actually manifested.

2.3 Order-(Dis)Order Family Trees Framework

Family trees are constructed to recover, for a given structure (query), the set of crystallographically related phases connected by group-subgroup transitions and occupancy splitting, along which stoichiometry may vary. Starting from the query structure, candidate parent descriptions are systematically generated, and experimentally reported parent phases are subsequently identified using our `SWORDFamilyMatcher` module.

Figure 1 illustrates the concept of order–(dis)order family mapping using the family tree recovered from an ordered Cu_2GeS_3 query structure (`Imm2`, highlighted in bright blue) as an example. Each phase in this particular family tree corresponds to an experimentally reported ICSD entry and is labeled by its space group symbol and SWORD representation, while parent–child relations are defined by group–subgroup transitions and annotated with the corresponding subgroup index. The vertical arrangement follows point-group order, so that higher-symmetry phases appear higher in the tree. In this example, the disordered phase `a_c_216_S_{Cu+Ge}` occupies the highest-symmetry position and acts as the parent from which both ordered and disordered descendants are reached through group-subgroup transitions. From Figure 1, it is clear that apparently distinct structures cannot be treated as isolated phases: an ordered phase may be placed within a broader family containing higher-symmetry disordered parents and symmetry-related ordered siblings. Our framework turns this intuition into a symmetry-guided mapping problem, in which a query structure is assigned to a crystallographic lineage rather than evaluated only as an isolated structure. The full connected set defines an order–(dis)order family tree, while ordered descendants sharing the same parent form order–order subfamilies. The same family tree would be recovered regardless of which phase in the tree is used as the starting query. The query therefore serves only as an entry point to the family. Additional implementation details are provided in Appendix A.

This matching framework captures several cases within a single rule: a predicted ordered structure may match an experimentally synthesized disordered parent, and two or more ordered structures may be linked through a shared parent, such that structures that are not identical at the child level are still recognized as members of the same broader family. In this way, the mapping lifts the structure matching from the level of isolated crystals to the level of symmetry-related families. Rather than asking only whether a predicted structure has been seen before in exactly the same ordered or disordered form, we also ask whether it already resides within an existing order-(dis)order lineage.

Table 1: Comparison between ICSD Collection Codes of the previously reported disordered parents and those assigned by our framework for 22 A-Lab target phases. White rows (✓) denote agreement with previously reported parent assignments. Grey rows (□) indicate positional-disorder cases beyond the scope of the present framework. Green rows (∗) highlight newly identified disordered parents, demonstrating that our framework both recovers known parents and identifies previously unreported ones.

#	Target Phase	MP ID	ICSD Code[1]	ICSD Code[2]	Remark
1	Ba_2ZrSnO_6	mp-1228067	[43137]	[43137, 176330, 176331, 176329]	✓
2	$FeSb_3Pb_4O_{13}$	mp-1224890	[60805]	[60805]	✓
3	$Hf_2Sb_2Pb_4O_{13}$	mp-1224490	[62723]	[62723]	✓
4	$InSb_3Pb_4O_{13}$	mp-1223746	[41119]	[41119]	✓
5	KMn_3O_6	mp-1016190	[240249]	NaN	□
6	$KNaP_6(PbO_3)_8$	mp-1223429	[182501]	[182500, 182501, 182502]	✓
7	$KNaTi_2(PO_5)_2$	mp-1211611	[59284]	[67539, 71239]	□
8	$K_2TiCr(PO_4)_3$	mp-1224541	NaN	[280999]	∗
9	$KPr_9(Si_3O_{13})_2$	mp-1223421	NaN	[153272]	∗
10	$K_4MgFe_3(PO_4)_5$	mp-532755	[161484]	[161484]	✓
11	$K_4TiSn_3(PO_5)_4$	mp-1224290	[250088]	[250088, 72720, 91534, 250087]	✓
12	$NaCaMgFe(SiO_3)_4$	mp-1221075	[75294]	[263074, 263075, 263076, 263077, 263078, . . .]	□
13	$Mg_3MnNi_3O_8$	mp-1222170	[80306]	[80302, 80303, 80304, 80305, 80306, 80307]	✓
14	Mg_3NiO_4	mp-1099253	[290603]	[290603, 13774]	✓
15	$MgTi_2NiO_6$	mp-1221952	[171583]	[171583]	✓
16	$MgTi_4(PO_4)_6$	mp-1222070	[74287]	[74287, 290277]	✓
17	$MgV_4Cu_3O_{14}$	mp-1222158	[69731]	[69731, 69732]	✓
18	Mn_2VPO_7	mp-1210613	NaN	[250126]	∗
19	$Mn_4Zn_3(NiO_6)_2$	mp-1222033	[92223]	[92222, 92223, 92224]	✓
20	$Na_3Ca_{18}Fe(PO_4)_{14}$	mp-725491	[85103]	[85103]	✓
21	$Sn_2Sb_2Pb_4O_{13}$	mp-1219056	[62722]	[62722]	✓
22	$Zr_2Sb_2Pb_4O_{13}$	mp-1215826	[62721]	[62721]	✓

[1] Leeman et al. [4]
[2] This work

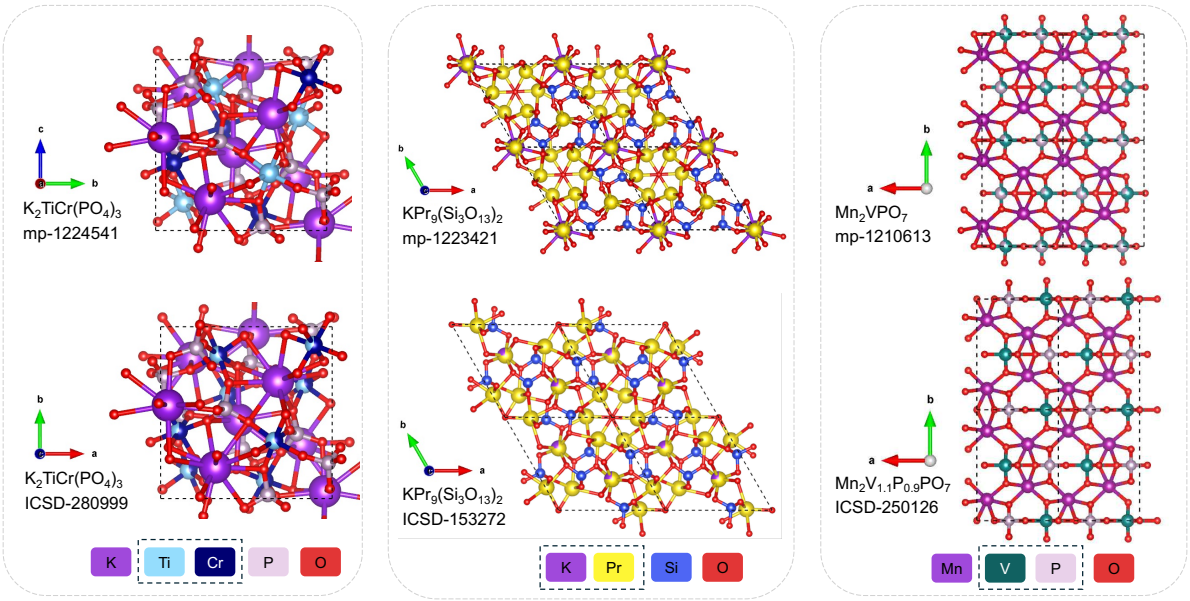


Figure 2: Recovery of experimentally reported disordered parents for three GNoME A-Lab target structures. Ordered GNoME structures are compared with the corresponding disordered parent phases reported in ICSD and recovered by our framework for $K_2TiCr(PO_4)_3$, $KPr_9(Si_3O_{13})_2$, and Mn_2VPO_7. The colored labels at the bottom denote the elemental species, and elements enclosed by dashed outlines indicate disordered species.

3 Results

3.1 Validation on A-Lab

We first validate that our framework can recover disordered parents for experimentally synthesized structures. Applying it to the 35 GNoME phases reported as "successfully" synthesized by A-Lab [1], we find 22 have experimentally reported disordered parent phases, either recovered by our framework or manually identified by Leeman et al [4]. Table 1 compares the ICSD Collection Codes of these previously reported disordered parents against those assigned by our framework for all 22 cases.

For 16 of the 22 (white-shaded ✓ rows), our framework systematically recovers the same disordered parents as Leeman et al. [4], directly validating the approach. Grey-shaded □ rows involve positional disorder, which lies outside the current scope since our method targets occupational disorder; nevertheless, even in several of these cases our framework recovers alternative substitutionally disordered parents not previously reported.

Most notably, for the three green-shaded $*$ phases ($K_2TiCr(PO_4)_3$, $KPr_9(Si_3O_{13})_2$, Mn_2VPO_7), prior analyses by Leeman et al. [4] suggested the possible existence of disordered parents, but did not identify any explicitly. In contrast, our framework recovers corresponding disordered parents for all three cases. Figure 2 compares the target GNoME A-Lab ordered structures with their corresponding experimentally reported disordered parent phases in ICSD for these cases. Together, these results demonstrate the framework's capacity to systematically navigate order–disorder family trees, including cases that are difficult to identify through manual inspection alone.

3.2 Order-(Dis)Order Family Tree Benchmark

Having validated the framework, we next use it as a benchmarking tool for family-level novelty assessment of ordered structures relative to experimental order–disorder family trees in ICSD. Rather than evaluating a structure set only through isolated child-level matches, we assess it through the family relations recovered by the proposed matching framework. To this end, we introduce two complementary Family Tree based metrics, FT_{disorder} and FT_{order}, that quantify how a given set of ordered structures is situated with respect to existing family trees. FT_{disorder} measures the fraction of ordered structures that fall within family trees rooted in existing disordered ICSD parents. A high value of FT_{disorder} indicates that many ordered structures can be traced to known disordered phases, suggesting that their

Table 2: Order-(dis)order family tree benchmark on known databases. For each dataset, FT_{disorder} represents the percentage of ordered structures that fall within order-(dis)order families containing an existing disordered parent, and FT_{order} represents the percentage that fall within families already spanned by existing ordered structures through group-subgroup relations. n^{FT}_{disorder} and n^{FT}_{order} are the corresponding numbers of structures assigned to these two categories, and n_{total} is the total number of ordered structures evaluated.

Dataset	FT_{disorder} (%)	FT_{order} (%)	n^{FT}_{disorder}	n^{FT}_{order}	n_{total}
$\text{GNoME}_{\text{A-Lab}}$	60.00	0.00	21	0	35
$\text{ICSD}_{\text{order}}$	6.13	10.37	3,564	6,026	58,116
$\text{MP-20}_{\text{train+val}}$	23.27	11.16	8421	4,039	36,183
$\text{Alex-MP-20}_{\text{train+val}}$	3.75	2.39	1,876	1,196	50,000
$\text{GNoME}_{\text{stable}}$	0.73	0.40	367	199	50,000

apparent ordering could be artificial and that experimentally realized compounds instead favor occupational disorder. As an example, in the set of 35 GNoME A-Lab structures, 21 are found by our framework to have existing disordered parents, corresponding to $FT_{\text{disorder}} = 60\%$. Meanwhile, FT_{order} measures the fraction of ordered query structures that fall within family trees already spanned by existing ordered ICSD structures, i.e., those sharing the same disordered parent as their root, whether or not such a parent is experimentally synthesized. Unlike FT_{disorder}, a high FT_{order} does not imply experimental suspicion; rather, it quantifies how much of the set of ordered structures lies within the known ordered family manifold formalized by group–subgroup relations, that is, structures that can in principle be derived from already synthesized ordered phases by rearranging atoms over specific crystallographic sites within a common host structural network, with accompanying symmetry lowering or raising. Conversely, a low FT_{order} suggests genuinely new family trees absent from experimentally known crystallographic records.

Benchmark on Databases. We evaluate FT_{disorder} and FT_{order} across both experimental and computational structure databases, including ICSD, MP-20, Alex-MP-20, and GNoME. Table 2 summarizes these statistics alongside the corresponding numbers of structures assigned to each category.

Ordered structures in ICSD, the largest experimental crystal structure database to date, exhibit $FT_{\text{disorder}} = 6.13\%$. This reveals a nontrivial regime in which a single underlying structural network has been realized in both ordered and disordered forms, connected through group–subgroup relations. One possible explanation is that a subset of these ordered structures reflects incomplete structural refinement, where compositional disorder was not fully resolved or reported [27]. More plausibly, however, these cases represent genuine order-disorder transitions within a group-subgroup crystallographic lineage [28]. In either case, the result highlights that even within the experimental record, a fraction of ordered compounds are not isolated from disorder-rooted families. A more pronounced effect is observed in MP-20 [5], a legacy Materials Project subset derived from ICSD structures with at most 20 atoms per unit cell. Here, FT_{disorder} increases sharply to 23.27%, indicating that a substantial fraction of structures coincide with experimentally realized disordered parent phases. We defer a more detailed analysis of this behavior to the Appendix B.

By contrast, computational structure databases exhibit substantially lower FT_{disorder} values. For 50k subsamples of Alex-MP-20 [6] and GNoME stable structures on the MP hull [2], we obtain $FT_{\text{disorder}} = 3.75\%$ and 0.73%, respectively. This reduction should not be interpreted as evidence that computational structures are intrinsically less prone to disorder. Rather, FT_{disorder} is best understood as a lower bound on disorder propensity: it only counts cases for which a potential disordered parent phase can be found in existing experimental references. Experimental databases are concentrated in well-explored chemical systems, where both ordered and disordered phases have often been identified and recorded. Computational datasets, in contrast, extend into sparsely explored regions of composition space. In these regions, an ordered structure could still belong to a broader disorder-rooted family, but the corresponding disordered parent might simply be absent from current databases. As a result, the lower FT_{disorder} observed for computational datasets more likely reflects incomplete coverage of disorder families in experimental references than a true absence of disorder. Consistent with this view, manual inspection of GNoME structures reveals that many candidates can be associated with plausible disordered parents in ICSD [3]; moreover, recent predictive studies also suggest that a majority of GNoME structures are in fact prone to occupational disorder [29].

A complementary perspective emerges from FT_{order}. In ICSD, more than 10% of ordered structures belong to order-order families, indicating that a noticeable fraction of experimentally reported phases within the same composition space are related through group-subgroup transitions on a common structural network. MP-20 shows a similar level of overlap, whereas Alex-MP-20 and GNoME exhibit lower fractions, potentially because their broader compositional sampling reduces the likelihood that multiple ordered members of the same family are simultaneously represented. At the same time, the lower values of FT_{order}, which explicitly account for chemical identity, may indicate that although

Table 3: Order-(dis)order family tree benchmark on 10,000 samples from crystal generative models trained on MP-20 and Alex-MP-20. For each model, FT_{disorder} represents the percentage of ordered structures that fall within order-(dis)order families containing an existing disordered parent, and FT_{order} represents the percentage that fall within families already spanned by existing ordered structures through group-subgroup relations. Lower values indicate weaker overlap with experimentally known family manifolds and therefore greater family-tree-based novelty.

Dataset	Model	$FT_{\text{disorder}} \downarrow (\%)$	$FT_{\text{order}} \downarrow (\%)$
MP-20	DiffCSP	4.88	10.13
	DiffCSP++	4.48	4.44
	FlowMM	4.23	7.79
	MiAD	14.24	25.43
	SymmCD	3.83	**3.26**
	CrystalDiT	5.05	13.69
	ADiT	12.61	28.12
	MatterGen	6.88	10.11
	Chemeleon2-RL	6.65	6.04
	WyFormer	**3.40**	4.00
Alex-MP-20	MatterGen	3.67	3.01
	WyFormer	**0.96**	**1.17**

a structure is new for a particular combination of elements, its underlying structure type has already been explored elsewhere [30].

Benchmark on Generative Models. We next benchmark two major classes of state-of-the-art crystal generative models trained on MP-20: (i) all-atom diffusion or flow-matching models and (ii) space group symmetry-based models, spanning 10 models in total. Category (i) includes DiffCSP, FlowMM, MiAD, CrystalDiT, ADiT, Chemeleon2-RL, and MatterGen, whereas category (ii) includes DiffCSP++, SymmCD, and WyFormer. Their central difference lies in the space over which generation is parameterized. All-atom models operate directly on the full set of atoms in the unit cell, without explicitly restricting the generative trajectory by crystallographic symmetry. Symmetry-based models instead operate in the asymmetric-unit, or Wyckoff subspace, representing crystals through discrete symmetry elements such as space group and site symmetry within a constrained parameterization. This both regularizes the search space and reduces the effective number of degrees of freedom. The two classes therefore offer a natural comparison between symmetry-agnostic generation and generation that is symmetry-aware by design.

Table 3 reports the order-(dis)order family tree benchmark for the 10 generative models trained on MP-20, together with MatterGen and WyFormer additionally trained on Alex-MP-20. The most immediate observation is the elevated FT_{disorder} of MiAD and ADiT, at 14.24% and 12.61%, respectively, followed by MatterGen. Notably, all three belong to the symmetry-agnostic all-atom models. These high values indicate that a substantial fraction of generated structures that appear novel at the ordered-child level may in fact fall within family trees whose disordered parent phases are already known experimentally. In other words, apparent novelty at the ordered level can mask a closer relationship to previously realized disorder-rooted families. This trend is particularly notable in light of the reported synthesis case of $\mathrm{TaCr_2O_6}$ generated by MatterGen [9]. Although initially proposed as a new ordered structure, the synthesized product was later found to correspond instead to a Cr-rich variant of an existing disordered parent phase. In this context, the high FT_{disorder} values of this model class suggest that the reported example may be symptomatic of a more general trend.

By contrast, WyFormer achieves the lowest FT_{disorder}, at 3.40%, followed by SymmCD, and both belong to the space group symmetry-based models. A similar pattern is observed when MatterGen and WyFormer are trained on Alex-MP-20: WyFormer again exhibits an FT_{disorder} nearly four times lower than that of MatterGen. Together, these results suggest that symmetry-regularized models are substantially less likely to generate ordered structures that fall within family trees of known disordered parents, and are therefore less prone to producing ordered candidates shadowed by experimentally reported disordered phases. A similar trend is observed for FT_{order}. All-atom models again show consistently higher values, indicating that their generations more frequently occupy regions of already known order-order family space. By contrast, the lower FT_{order} of symmetry-based models suggests that they are more likely to generate ordered structures that are novel not only as individual entries, but also at the family level.

This contrast is consistent with the design of the two model classes. WyFormer, one of the leading symmetry-constrained models, achieves the highest uniqueness and novelty in the Wyckoff subspace relative to symmetry-agnostic models

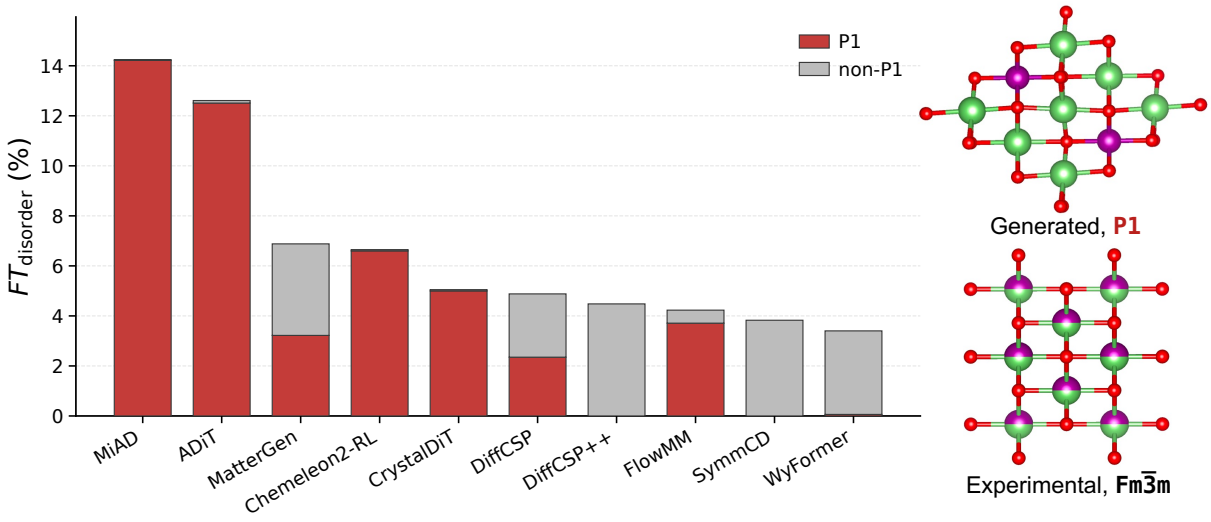


Figure 3: Space group decomposition of FT_{disorder} across crystal generative models. For each model, FT_{disorder} is decomposed by the space group of the ordered generated structure, with P1 highlighted in red. The dominant contribution of P1 in all-atom models points to the possibility that some of the overproduced P1 structures originate as ordered children of known disordered parent phases. Right, an example of generated P1 child and its matched ICSD disordered parent in Fm$\bar{3}$m.

such as ADiT and MatterGen [7]. Our results suggest that this advantage at the child level may translate into the family level: models designed to operate explicitly in a symmetry-constrained design space appear better positioned to discover genuinely new family-level configurations, rather than merely new ordered structures within already known families. This should be a consideration for future generative models in order to achieve the goal of synthesizing truly novel materials in experiments.

Stable P1. Space group P1 is the simplest crystallographic symmetry, containing no symmetry operations beyond identity and lattice translations. In nature, such trivial symmetry is rare: its occurrence in ICSD amounts to less than 1%, and even this fraction may be inflated by misreported cases [31, 32]. In contrast, all-atom generative models are known to overproduce P1 structures, which in some cases constitute nearly half of generated samples, far exceeding the experimental distribution. Whether these *novel* and *stable* P1 structures represent a byproduct of imperfect computational evaluation or a genuinely unexplored class of experimentally accessible structure types remains an open question in the community.

To examine this, we analyzed the space group distribution of ordered structures identified as children of existing disordered parents for each model. Figure 3 highlights the FT_{disorder} of each model, with P1 contributions marked in red. For all-atom models, including MiAD, ADiT, MatterGen, Chemeleon2-RL, CrystalDiT, DiffCSP, and FlowMM, P1 dominates FT_{disorder}, accounting for anywhere from nearly half to essentially all child structures, making it apparent that a part of P1 structures systematically appear as ordered configurations derived from existing disordered parent phases. An example is shown on the right side of Figure 3, where a P1 structure generated by DiffCSP is traced back to an ICSD disordered parent crystallizing in space group Fm$\bar{3}$m, using our framework.

Further DFT validation of P1 structures that have ICSD disordered parents, with MiAD and MatterGen taken as representative all-atom models, reveals that roughly 80–90% are metastable after relaxation, with energies above the hull below 100 meV/atom relative to the Materials Project database [33]. Details of the DFT calculations are provided in Appendix C. These findings, therefore, suggest that the overproduction of apparently stable P1 structures can be explained to some extent by their emergence as ordered children of high-symmetry disordered parent phases: configurations that, while compositionally disordered in experiment, are unlikely to be experimentally synthesized in their ordered P1 form.

Group-Subgroup Polymorphs. Order-order family trees, as formalized in Section 2.3, are governed by group-subgroup relations. In general, ordered structures belonging to the same FT_{order} family do not necessarily share the same stoichiometry, because symmetry lowering can split Wyckoff orbits in ways that alter the allowed occupancy

pattern. A particularly important subset arises when symmetry descent preserves the exact same stoichiometry. We refer to this subset as *group–subgroup polymorphs*.

These polymorphs represent a symmetry-grounded form of polymorphism: one ordered phase can be derived from another through a group-subgroup transition that lowers point-group symmetry while preserving composition. This perspective goes beyond treating polymorphs simply as distinct structures with the same stoichiometry, and instead isolates those connected by an explicit symmetry-descent pathway. A representative example is provided by the two ordered Cu_2GeS_3 phases in Figure 1, experimentally reported in space groups `Imm2` and `Cc`, where one can be obtained from the other through symmetry breaking along a group-subgroup transition.

This behavior is not rare in experimental data. In ICSD, among roughly 5×10^4 unique ordered compositions, about 10% exhibit polymorphism, and around 40% of those polymorphic systems contain at least one pair related by a group-subgroup transition. This indicates that a meaningful fraction of experimentally reported polymorphs are not merely alternative structural realizations of the same composition, but members of an explicit symmetry-connected lineage, namely an order-order family. Order–order family mapping therefore offers more than a descriptive framework for known polymorphs, but it could also provide a systematic basis for discovering new ones.

4 Discussion

The order-(dis)order family tree benchmark introduced here redefines novelty against experimentally grounded crystallographic lineage rather than isolated ordered endpoints. This family-level view provides a more meaningful notion of novelty in systems where multiple ordered phases descend from a common disordered parent. The benchmark is most informative when the relevant disordered parent is already present in the experimental record. For underexplored systems, the corresponding parent may simply remain unseen, and the absence of a family assignment marks regions of family space not yet experimentally charted rather than a genuine absence of disorder tendency.

More fundamentally, these results indicate that disorder is best treated not as a label to predict, but as a manifold to navigate. Extending access to unseen disordered parents is therefore a natural next step: it would improve how novelty is assessed, clarify what synthesis is actually likely to realize, and provide a broader structural basis for evaluating stability across experimentally relevant energetic landscapes. In this sense, the same family-manifold perspective should ultimately inform not only crystallographic novelty, but also how thermodynamic preference is estimated within an order-disorder family [34]. The order-(dis)order family tree framework provides a concrete structural basis for that shift.

5 Conclusion

Across A-Lab case studies, known structure databases (ICSD, MP-20, Alex-MP-20, GNoME), and state-of-the-art crystal generative models, we found that many ordered structures that appear novel under conventional child-level comparison are more naturally interpreted as members of already known family trees, often as ordered children of existing disordered parents. This effect is especially visible in synthesis-facing cases and in symmetry-agnostic all-atom generative models, where the labeled novelty can be traced to originate from rediscovery of ordered configurations of experimentally realized disordered parents. We further show that the excessive number of apparently novel and stable `P1` structures produced by all-atom models can, in part, be understood as ordered children of existing disordered parents. In parallel, the framework also exposes symmetry-connected order-order families and identifies a meaningful subset of experimentally reported polymorphs as group-subgroup polymorphs.

Our results further reveal that symmetry-constrained generative models provide a more promising route to genuine family-level discovery than symmetry-agnostic all-atom models. By operating directly in a symmetry-constrained design space, they are less likely to generate ordered structures that simply occupy already known order-(dis)order families, and are better positioned to explore new crystallographic lineage at the family level.

In this work, we introduced order-(dis)order family trees as a symmetry-grounded framework for organizing ordered and disordered crystal structures through group-subgroup relations. By shifting the unit of analysis from isolated structures to crystallographic families, the framework makes it possible to assess novelty at the level most relevant to experimental realization. We therefore argue that order-(dis)order family trees should be treated as a key requirement in evaluating novelty in materials discovery systems. Overall, these results position disorder-aware family matching as a practical foundation for future discovery pipelines and a necessary step toward identifying genuinely new, experimentally realizable materials.

Resource availability

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Kedar Hippalgaonkar (kedar@ntu.edu.sg).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The crystal structure data used in this study were obtained from the Inorganic Crystal Structure Database (ICSD), version 2025.2_v5.5.0. ICSD entries are available at https://icsd.products.fiz-karlsruhe.de under a commercial license. Data curation of the ICSD prior to analysis was performed using code available at https://github.com/YuyaoHuang330/SWORD. The implementation of the order-(dis)order family tree framework is available at https://github.com/shuyayamazaki/order-disorder-family-trees.

Acknowledgments

We thank Andrey Okhotin for providing data samples for DiffCSP, DiffCSP++, FlowMM, MiAD, SymmCD, and CrystalDiT. K.H. acknowledges funding from the MAT-GDT Program at A*STAR via the AME Programmatic Fund by the Agency for Science, Technology and Research under Grant No. M24N4b0034. M.H.P. acknowledges funding from the NTU AI4X Postdoctoral Fellowship.

Author contributions

S.Y. initiated the study, conceived the original idea, and developed the order-(dis)order family tree framework in close collaboration with Y.H. S.Y. designed and performed the validation on A-Lab cases, introduced the FT novelty metrics, and carried out the benchmarking and all data analyses. S.Y. wrote the initial manuscript. K.H., S.Y, M.P.H., Y.H., and W.N. contributed to reviewing and revising the manuscript.

Declaration of interests

The authors declare no competing interests.

References

[1] Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91, November 2023.

[2] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, November 2023.

[3] Anthony K. Cheetham and Ram Seshadri. Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery. *Chemistry of Materials*, 36(8):3490–3495, April 2024.

[4] Josh Leeman, Yuhan Liu, Joseph Stiles, Scott B. Lee, Prajna Bhatt, Leslie M. Schoop, and Robert G. Palgrave. Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy*, 3(1), March 2024.

[5] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation, 2021.

[6] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei

Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, January 2025.

[7] Siddharth Betala, Samuel P. Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K. Joshi, Sékou-Oumar Kaba, Félix Therrien, Alex Hernandez-Garcia, Rocío Mercado, N. M. Anoop Krishnan, and Alexandre Duval. Lemat-genbench: A unified evaluation framework for crystal generative models, 2025.

[8] Stephan Rühl. The inorganic crystal structure database (icsd): A tool for materials sciences, August 2019.

[9] Mikkel Juelsholt. Continued challenges in high-throughput materials predictions: Mattergen predicts compounds from the training dataset. January 2026.

[10] H. Bärnighausen. Group-subgroup relations between space groups: A useful tool in crystal chemistry. *MATCH Communications in Mathematical and in Computer Chemistry*, 9:139–175, 1980.

[11] S. Ivantchev, E. Kroumova, G. Madariaga, J. M. Pérez-Mato, and M. I. Aroyo. Subgroupgraph: a computer program for analysis of group–subgroup relations between space groups. *Journal of Applied Crystallography*, 33(4):1190–1191, August 2000.

[12] Yu Han, Chi Ding, Junjie Wang, Hao Gao, Jiuyang Shi, Shaobo Yu, Qiuhan Jia, Shuning Pan, and Jian Sun. Efficient crystal structure prediction based on the symmetry principle. *Nature Computational Science*, 5(3):255–267, February 2025.

[13] Kesong Yang, Corey Oses, and Stefano Curtarolo. Modeling off-stoichiometry materials with a high-throughput ab-initio approach. *Chemistry of Materials*, 28(18):6484–6492, September 2016.

[14] Yuyao Huang, Wei Nong, Shuya Yamazaki, Martin Hoffmann Petersen, Jianghai Wang, Ruiming Zhu, and Kedar Hippalgaonkar. Sword: Symmetry and wyckoff-sequence of ordered and disordered crystals, 2026.

[15] D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, and S. Rehme. Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. *Journal of Applied Crystallography*, 52(5):918–925, September 2019.

[16] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion, 2023.

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

[18] Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. 2024.

[19] Andrey Okhotin, Maksim Nakhodnov, Nikita Kazeev, Andrey E Ustyuzhanin, and Dmitry Vetrov. Miad: Mirage atom diffusion for de novo crystal generation, 2025.

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

[21] Xiaohan Yi, Guikun Xu, Xi Xiao, Zhong Zhang, Liu Liu, Yatao Bian, and Peilin Zhao. Crystaldit: A diffusion transformer for crystal generation, 2025.

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

[23] Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. 2025.

[24] Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[25] R. W. G. Wyckoff. *The Analytical Expression of the Results of the Theory of Space-Groups*, volume 318. Carnegie Institution of Washington, 1922.

[26] Dmytro Antypov, Chris M. Collins, Matthew S. Dyer, John B. Claridge, and Matthew J. Rosseinsky. Classification and statistical analysis of structural disorder in crystalline materials. *Journal of Applied Crystallography*, 58(3):659–677, May 2025.

[27] Anthony L. Spek. What makes a crystal structure report valid? *Inorganica Chimica Acta*, 470:232–237, January 2018.

[28] Denis Gratias and Marianne Quiquandon. Bicrystallography and beyond: Example of group–subgroup phase transformations. *Crystals*, 10(7):560, July 2020.

[29] Konstantin S. Jakob, Aron Walsh, Karsten Reuter, and Johannes T. Margraf. Learning crystallographic disorder: Bridging prediction and experiment in materials discovery. *Advanced Materials*, 38(5), October 2025.

[30] Hagen Eckert, Simon Divilov, Michael J. Mehl, David Hicks, Adam C. Zettel, Marco Esters, Xiomara Campilongo, and Stefano Curtarolo. The aflow library of crystallographic prototypes: Part 4. *Computational Materials Science*, 240:112988, May 2024.

[31] V. S. Urusov and T. N. Nadezhina. Frequency distribution and selection of space groups in inorganic crystal chemistry. *Journal of Structural Chemistry*, 50(S1):22–37, December 2009.

[32] Richard E. Marsh. P1 or p-1? or something else? *Acta Crystallographica Section B Structural Science*, 55(6):931–936, December 1999.

[33] Matthew K. Horton, Patrick Huck, Ruo Xi Yang, Jason M. Munro, Shyam Dwaraknath, Alex M. Ganose, Ryan S. Kingsbury, Mingjian Wen, Jimmy X. Shen, Tyler S. Mathis, Aaron D. Kaplan, Karlo Berket, Janosh Riebesell, Janine George, Andrew S. Rosen, Evan W. C. Spotte-Smith, Matthew J. McDermott, Orion A. Cohen, Alex Dunn, Matthew C. Kuner, Gian-Marco Rignanese, Guido Petretto, David Waroquiers, Sinead M. Griffin, Jeffrey B. Neaton, Daryl C. Chrzan, Mark Asta, Geoffroy Hautier, Shreyas Cholia, Gerbrand Ceder, Shyue Ping Ong, Anubhav Jain, and Kristin A. Persson. Accelerated data-driven materials science with the materials project. *Nature Materials*, 24(10):1522–1532, 2025.

[34] Simon Divilov, Hagen Eckert, David Hicks, Corey Oses, Cormac Toher, Rico Friedrich, Marco Esters, Michael J. Mehl, Adam C. Zettel, Yoav Lederer, Eva Zurek, Jon-Paul Maria, Donald W. Brenner, Xiomara Campilongo, Suzana Filipović, William G. Fahrenholtz, Caillin J. Ryan, Christopher M. DeSalle, Ryan J. Crealese, Douglas E. Wolfe, Arrigo Calzolari, and Stefano Curtarolo. Disordered enthalpy–entropy descriptor for high-entropy ceramics discovery. *Nature*, 625(7993):66–73, January 2024.

[35] Atsushi Togo, Kohei Shinohara, and Isao Tanaka. Spglib: a software library for crystal symmetry search. *Science and Technology of Advanced Materials: Methods*, 4(1), October 2024.

[36] 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, February 2013.

[37] Alex Ganose, Hrushikesh Sahasrabuddhe, Mark Asta, Kevin Beck, Tathagata Biswas, Alexander Bonkowski, Joana Bustamante, Xin Chen, Yuan Chiang, Daryl Chrzan, Jacob Clary, Orion Cohen, Christina Ertural, Max Gallant, Janine George, Sophie Gerits, Rhys Goodall, Rishabh Guha, Geoffroy Hautier, Matthew Horton, Aaron Kaplan, Ryan Kingsbury, Matthew Kuner, Bryant Li, Xavier Linn, Matthew McDermott, Rohith Srinivaas Mohanakrishnan, Aakash Naik, Jeffrey Neaton, Kristin Persson, Guido Petretto, Thomas Purcell, Francesco Ricci, Benjamin Rich, Janosh Riebesell, Gian-Marco Rignanese, Andrew Rosen, Matthias Scheffler, Jonathan Schmidt, Jimmy-Xuan Shen, Andrei Sobolev, Ravishankar Sundararaman, Cooper Tezak, Victor Trinquet, Joel Varley, Derek Vigil-Fowler, Duo Wang, David Waroquiers, Mingjian Wen, Han Yang, Hui Zheng, Jiongzhi Zheng, Zhuoying Zhu, and Anubhav Jain. Atomate2: Modular workflows for materials science. January 2025.

[38] G. Kresse and J. Furthmüller. Efficient iterative schemes forab initiototal-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, October 1996.

[39] G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59(3):1758–1775, January 1999.

[40] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, October 1996.

[41] Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, 2025.

A Order-(Dis)Order Family Matching Framework

Our order–(dis)order family matching framework is implemented in the `SWORDFamilyMatcher` module. For a query crystal structure S, the procedure first computes its SWORD child label,

$$\ell_{\mathrm{child}}(S) = \mathrm{SWORD}(S),$$

using symmetry analysis with `spglib` [35] through the SWORD labeling routine. In practice, the structure is first parsed into a `pymatgen` `Structure` [36], and partially occupied sites are renormalized when needed so that occupancies sum to unity before label construction. The overall procedure is summarized in Algorithm 1.

The central idea is to infer possible higher-symmetry parent descriptions by selectively removing chemical distinctions and recomputing the corresponding symmetry-aware label. If an ordered or partially ordered structure becomes compatible with a more symmetric description once certain species distinctions are masked, then that higher-symmetry description is taken as a plausible parent candidate.

Let $E(S)$ denote the set of distinct element types in S, with $n = |E(S)|$. For a chosen subset $M \subseteq E(S)$, we define a masking operation $\mathcal{M}_M(S)$ that replaces every element in M by a common dummy species X. Recomputing the SWORD label after masking gives a candidate parent label,

$$\ell^{(M)}_{\mathrm{parent}}(S) = \mathrm{SWORD}(\mathcal{M}_M(S)) \,.$$

Because the raw label obtained after masking is expressed in terms of the dummy species X, we then restore the chemically meaningful mixed-site description by replacing each occurrence of X with the corresponding masked element set. This yields the final parent label used for matching.

For an ordered query, all nontrivial masks with $2 \le |M| \le n$ are enumerated, giving

$$\sum_{k=2}^{n} \binom{n}{k} = 2^n - n - 1$$

possible masking patterns. For a query that is already disordered, masking is restricted to the mixed-element groups already present in the SWORD child label. This restriction is important because an element may appear simultaneously on ordered and disordered sites, so the relevant family relation is not obtained by arbitrary recombination of all species. Instead, the existing mixed-occupancy pattern encoded in the child label is treated as the chemically meaningful starting point, and only those disorder groups are further abstracted when generating parent candidates.

Accordingly, the parent-label set is written as

$$\mathcal{P}(S) = \left\{ \ell^{(M)}_{\mathrm{parent}}(S) \right\},$$

and the family label set of the query is

$$\mathcal{F}(S) = \{\ell_{\mathrm{child}}(S)\} \cup \mathcal{P}(S).$$

Given a reference structure R processed in the same way, we regard R as family-related to S whenever their family label sets intersect,

$$\mathcal{F}(S) \cap \mathcal{F}(R) \neq \varnothing.$$

For ordered queries, the implementation includes an additional vacancy-ordering augmentation step before parent enumeration. If the child label contains no mixed occupancy marker, the structure is analyzed for symmetry-compatible latent vacancy ordering using the `find_vacancy_ordered` routine. This routine identifies candidate void sites from a periodic Voronoi construction, evaluates their local environments against species-specific coordination templates, and fills only those sites whose geometry and nearest-neighbor spacing are consistent with the insertion of an existing species. When such a filled structure is obtained, parent labels are generated both from the original query and from the vacancy-filled variant. This step allows the matcher to recover family relations that may otherwise remain hidden when an experimentally relevant parent description is more naturally expressed after restoring an ordered vacancy configuration.

In the pairwise setting, two structures S_1 and S_2 are assigned to the same family if

$$\mathcal{F}(S_1) \cap \mathcal{F}(S_2) \neq \varnothing.$$

In database matching, the same criterion is used against every reference entry. If the matched reference child label contains mixed occupancy, the hit is recorded as a disordered-family match; otherwise, if the matched reference child label is ordered and distinct from the query child label, it is recorded as an order–order family match.

Algorithm 1 SWORDFAMILYMATCHER

Input: Query structure S or two structures S_1, S_2
Output: Family label set for S, or a Boolean family match for S_1, S_2

Function GETFAMILYLABELS(S)
parse S into a crystal structure object
renormalize partial occupancies if needed
$\ell_{\text{child}} \leftarrow \text{SWORD}(S)$
$\mathcal{S}_{\text{proc}} \leftarrow [S]$
if ℓ_{child} contains no mixed-occupancy marker **then**
  $S_{\text{fill}} \leftarrow$ FINDVACANCYORDERED(S)
  if S_{fill} is valid **then**
    append S_{fill} to $\mathcal{S}_{\text{proc}}$
  end if
end if
$\mathcal{P} \leftarrow \varnothing$
for each $\widetilde{S} \in \mathcal{S}_{\text{proc}}$ **do**
  $\widetilde{\ell}_{\text{child}} \leftarrow \text{SWORD}(\widetilde{S})$
  extract mixed-element groups from $\widetilde{\ell}_{\text{child}}$
  if mixed-element groups are present **then**
    define mask sets from those existing mixed groups only
  else
    define all element subsets $M \subseteq E(\widetilde{S})$ with $2 \leq |M| \leq |E(\widetilde{S})|$
  end if
  for each mask set M **do**
    construct $\mathcal{M}_M(\widetilde{S})$ by replacing elements in M with dummy species X
    $\ell_{\text{raw}} \leftarrow \text{SWORD}(\mathcal{M}_M(\widetilde{S}))$
    $\ell_{\text{parent}} \leftarrow$ restore masked element group into ℓ_{raw}
    add ℓ_{parent} to $\mathcal{P}$
  end for
end for
return $\mathcal{F}(S) = \{\ell_{\text{child}}\} \cup \mathcal{P}$

Function FIT(S_1, S_2)
$\mathcal{F}_1 \leftarrow$ GETFAMILYLABELS(S_1)
$\mathcal{F}_2 \leftarrow$ GETFAMILYLABELS(S_2)
if $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \varnothing$ **then**
  return `True`
else
  return `False`
end if

As a concrete example, for the ordered Cu_2GeS_3 query structure used in Figure 1, `SWORDFamilyMatcher` returns the following SWORD family dictionary:


```
{'child_label': 'a_b_c2_44_Ge_S_{Cu,S}',
 'parent_labels': ['a_c_216_S_{Cu+Ge}',
  'a_b_c2_44_Ge_{Cu+S}_{2(Cu+S)}',
  'a_b_c2_44_{Ge+S}_{Ge+S}_{Cu,Ge+S}',
  'a_227_{Cu+Ge+S}'],
 'matched_ordered_labels': ['a6_9_{2Cu,3S,Ge}'],
 'matched_ordered_ids': [85138, 146583],
 'matched_disordered_labels': ['a_d_b_122_Cu_S_{Cu+Ge}', 'a_c_216_S_{Cu+Ge}'],
 'matched_disordered_ids': [627781, 43531, 102963, 100955, 627773]}
```


This concrete example shows how the matcher encodes the query structure, its candidate parent descriptions, and the ordered and disordered ICSD relatives that together define the family recovered in Figure 1.

B MP-20

MP-20 has long been treated as one of the most representative training datasets for crystal generative models. It is defined as an ICSD-derived subset of Materials Project structures restricted to compounds with at most 20 atoms per unit cell, and contains roughly $45\,\mathrm{k}$ structures across the train, validation, and test splits. At the same time, when ordered ICSD entries are counted at the level of unique structures with no more than 20 atoms per unit cell, the total is only around $20\,\mathrm{k}$. This discrepancy is difficult to reconcile with a strict one-to-one view of MP-20 as a subset of experimentally reported ordered ICSD phases.

One plausible explanation is that the ICSD provenance inherited in *legacy* Materials Project data is not always cleanly recoverable at the level of unique current ICSD entries, and that MP-20 also includes Materials Project processed or relaxed versions of experimentally derived structures. Under this interpretation, MP-20 should be understood not as a strict snapshot of unique ordered ICSD phases, but as a legacy computational dataset with partial ICSD provenance and additional expansion introduced during downstream curation.

This distinction matters for interpreting its family-tree statistics. In particular, the fact that FT_{disorder} reaches 23.27% indicates that more than one in five ordered structures in MP-20 have existing disordered parents in ICSD. This suggests that a non-negligible portion of MP-20 lies in crystallographic families for which disorder is already an experimentally established realization. As a consequence, models trained on MP-20, particularly all-atom generative models, may in part be learning to reproduce ordered children of existing disordered parents rather than structures that are genuinely novel at the family level. At a high level, this observation further motivates disorder-aware dataset construction and generative modeling in crystal structure generation.

C DFT details

We use DFT settings from Materials Project `https://docs.materialsproject.org/methodology/materials-methodology/calculation-details/gga+u-calculations/parameters-and-convergence` for structure relaxation and energy computation. In particular, we do GGA and GGA+U calculations with `atomate2.vasp.flows.mp.MPGGADoubleRelaxStaticMaker` [37], which in turn relies on `pymatgen.io.vasp.sets.MPRelaxSet` and `pymatgen.io.vasp.sets.MPStaticSet` [36]. Computations themselves were done with VASP [38] version 5.4.4. with the plane-wave basis set [38]. The electron-ion interaction is described by the projector augmented wave (PAW) pseudo-potentials [39]. The exchange-correlation of valence electrons is treated with the Perdew-Burke-Ernzerhof (PBE) functional within the generalized gradient approximation (GGA) [40]. The raw total energies computed by DFT were corrected with `MaterialsProject2020Compatibility` before putting into the `PhaseDiagram` to obtain the DFT E_{hull}. We used the MP convex hull `2023-02-07-ppd-mp.pkl.gz` distributed by `matbench-discovery` [41] as the reference hull.