CRYSTALGPD: BENCHMARKING THE CAPABILITY BOUNDARIES OF CRYSTAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Crystal large language models (LLMs) have emerged as a flexible paradigm for crystal structure generation, yet their capabilities remain difficult to characterize due to heterogeneous tasks, interfaces, representations, datasets, and evaluation protocols. To bridge this gap, we introduce CRYSTALGPD (Crystal Generation, Prediction, and Design), a unified benchmark for characterizing crystal LLM capability boundaries across de novo generation, crystal structure prediction, composition- and symmetry-constrained design, and symmetry-constrained design. CRYSTALGPD preserves native model interfaces while standardizing task instances, generation protocols, representation-specific parsing, and downstream evaluation. Our evaluation reveals recurring limitations in jointly achieving physical feasibility and novelty, maintaining validity and structural recovery as complexity increases, satisfying exact crystallographic symmetry, and generalizing from MP20 to MP52. In particular, composition is preserved more reliably than exact symmetry, and symmetry controllability degrades for lower-frequency space groups. Motivated by these findings, we further study CRYSTALREACT, a benchmark-driven agentic case study that provides on-demand crystallographic knowledge and verification during generation. CRYSTALREACT substantially improves physical validity and crystallographic constraint satisfaction, demonstrating that external knowledge can alleviate several limitations exposed by CRYSTALGPD. Together, CRYSTALGPD provides a unified framework for evaluating crystal LLMs and, more importantly, a systematic characterization of where their current capabilities succeed and break down, providing actionable guidance for developing more reliable crystal-generation systems. Anonymous code is available at `https://anonymous.4open.science/r/CrystalGPD`.

1 INTRODUCTION

The properties and stability of inorganic materials depend not only on their chemical compositions but also on their atomic arrangements and crystal symmetries (Jain et al., 2013). Identifying plausible crystal structures is therefore central to materials discovery. Traditional approaches based on density functional theory (DFT) (Kohn & Sham, 1965) provide accurate physical evaluation but are computationally expensive for exploring the vast crystal structure space. Generative methods (Xie et al., 2022; Jiao et al., 2024; Levy et al., 2025) offer a complementary paradigm by learning structural distributions from existing materials and efficiently proposing candidate crystals.

Recent crystal-generation methods span diffusion models, equivariant neural networks, and autoregressive language models (Jiao et al., 2023; Höllmer et al., 2025; Antunes et al., 2024). In particular, crystal LLMs (Gruver et al., 2024; Xu et al., 2026; Wu et al., 2026) have attracted increasing attention because they can express crystal structures through flexible textual representations and support increasingly diverse generation paradigms, ranging from direct generation to reasoning-based generation, as illustrated in Figure 1. This rapid development raises a fundamental question: *what can current crystal LLMs reliably accomplish, and where do their capabilities break down?*

Answering this question remains difficult. Existing crystal LLM studies differ substantially in datasets, output representations, and evaluation tasks, as summarized in Table 1. Consequently, results reported by individual studies reveal the performance of particular systems under particu-

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.

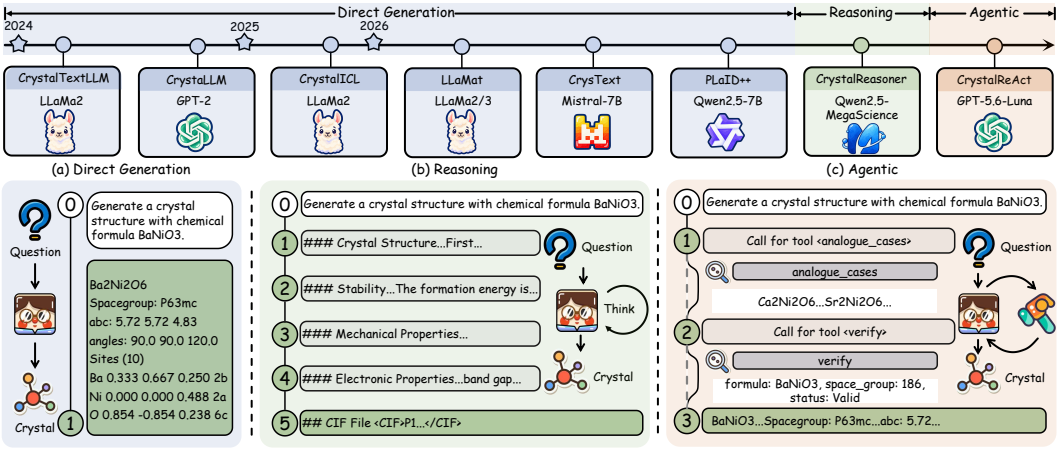


Figure 1: Evolution of crystal LLM paradigms and an illustration of the proposed agentic generation.

Table 1: Comparison of evaluation settings among crystal LLMs.

Paradigm	Model	Backbone	Representation	Dataset					Evaluation Task					
				MP20	MP30	MP52	P5	C24	DNG	CSP	CSD			
											Formula	Spacegroup	Form.+SG	E_{hull}
Direct generation	CrystalTextLLM	LLaMA-2-7B	Coordinates	✓	✓				✓		✓	✓		✓
	CrystaLLM	GPT-2	CIF	✓		✓	✓	✓	✓	✓	✓		✓	
	CrystalICL	LLaMA-2-7B	Coordinates	✓	✓		✓	✓	✓		✓	✓		
	LLaMat	LLaMA-2-7B / LLaMA-3-8B	Coordinates	✓					✓					
	CrysText	Mistral-7B-v0.3	CIF	✓						✓		✓		✓
	PLaID++	Qwen2.5-7B	Wyckoff	✓					✓			✓		
Reasoning	CrystalReasoner	Qwen2.5-3B-MegaScience	Reasoning + Coordinates	✓								✓	✓	
Agentic	CrystalReAct	GPT-5.6-Luna	Wyckoff	✓		✓			✓	✓	✓	✓	✓	

lar settings, but provide limited insight into their capability boundaries under a common evaluation view. Existing benchmarks (Wei et al., 2024; Duval et al., 2025) further focus primarily on individual generation scenarios, leaving several questions insufficiently characterized: whether crystal LLMs can jointly generate physically feasible and novel materials, how their structural accuracy scales with crystal complexity, how reliably they can satisfy compositional and crystallographic constraints, and how these capabilities generalize beyond training-aligned crystallographic conditions.

To bridge this gap, we introduce CRYSTALGPD (**Crystal Generation, Prediction, and Design**), a unified benchmark for systematically characterizing crystal LLM capability boundaries. CRYSTALGPD covers four complementary settings: de novo generation (DNG), crystal structure prediction (CSP), composition- and symmetry-constrained crystal structure design (CSD), and symmetry-constrained CSD. Rather than forcing heterogeneous models into a shared representation, CRYSTALGPD preserves their native prompts and output formats while standardizing task instances, generation protocols, and downstream evaluation for consistent system-level comparison.

Importantly, CRYSTALGPD goes beyond model ranking by revealing capability boundaries across crystallographic challenges. Across these settings, CRYSTALGPD reveals recurring limitations in achieving physically feasible and novel crystal discovery, maintaining physical validity as complexity increases, satisfying exact crystallographic constraints, and generalizing to less training-aligned conditions. Motivated by these limitations, we further study CRYSTALREACT as a benchmark-driven agentic case study, showing that on-demand access to crystallographic knowledge and verification can substantially improve physical validity and crystallographic constraint satisfaction.

Our main **contributions** are summarized as follows:

- We introduce CRYSTALGPD, a unified and representation-aware benchmark for crystal LLMs spanning de novo generation, crystal structure prediction, composition- and symmetry-constrained design, and symmetry-constrained design. CRYSTALGPD preserves native model interfaces while standardizing task instances, generation protocols, and downstream evaluation.
- We systematically characterize the capability boundaries of representative crystal LLMs across discovery quality, structural complexity, crystallographic controllability, and generalization. Our analyses reveal recurring limitations in jointly achieving physical feasibility and novelty, maintaining validity and structural recovery as complexity increases, and satisfying exact crystallographic symmetry under shifted and lower-frequency conditions.

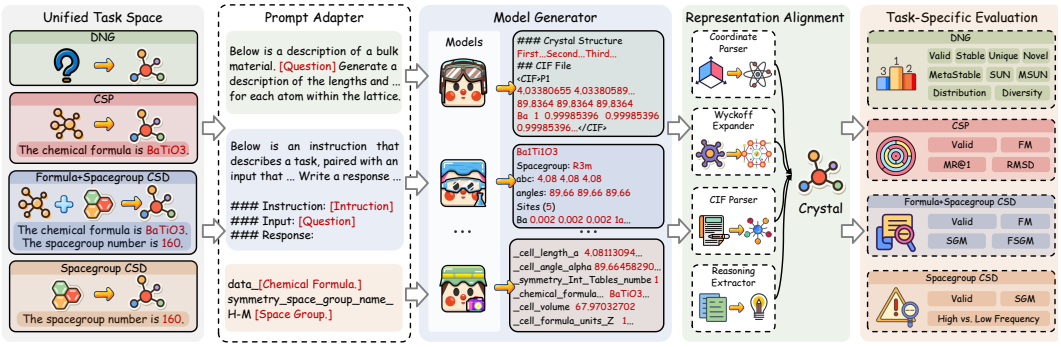


Figure 2: Overview of the CRYSTALGPD benchmark pipeline, which preserves native prompting and output formats of crystal LLMs while enabling consistent evaluation across diverse crystal tasks.

- Motivated by these benchmark findings, we study CRYSTALREACT as a benchmark-driven agentic case study and show that on-demand crystallographic knowledge access and verification can substantially improve physical validity and crystallographic constraint satisfaction.

2 RELATED WORK

LLM-Based Crystal Structure Generation Crystal LLMs have evolved from specialized models to general-purpose LLM adaptation with diverse training strategies. CrystaLLM (Antunes et al., 2024) trains GPT-2 from scratch on CIFs, while CrystalTextLLM (Gruver et al., 2024) shows that SFT enables LLM-based crystal generation. CrystalICL (Wang et al., 2025) extends this setting to in-context learning, and LLaMat (Ahlawat et al., 2026) is trained on large-scale crystal literature and materials data. PLaID++ (Xu et al., 2026) introduces DPO and substantially improves the SUN rate on the DNG task. CrysText (Mohanty et al., 2026) focuses on formula- and space-group-conditioned CSP, whereas CrystalReasoner (Wu et al., 2026) combines physical reasoning with GRPO to improve physical validity and stability. These systems differ in training data, conditioning, training objectives, and evaluation protocols, making their reported results difficult to compare directly.

Benchmarks for Crystal Generation Crystal generation is commonly evaluated on Perov-5 (Castelli et al., 2012), Carbon-24 (Pickard, 2020), MP-20 (Jain et al., 2013), and MPTS-52 using task-specific DNG and CSP metrics (Xie et al., 2022; Jiao et al., 2023). Matbench-genmetrics standardizes Materials Project splits and set-level generation metrics (Baird et al., 2024), while CSP-Bench evaluates conventional and learning-based CSP methods (Wei et al., 2024). Recent benchmarking efforts have also compared textual crystal representations under a common generation setting (Jia et al., 2025), while LeMat-GenBench provides a broader evaluation of DNG (Duval et al., 2025). However, existing benchmarks do not jointly compare crystal large language models across DNG, CSP, and CSD. CRYSTALGPD addresses this gap with a unified evaluation protocol.

3 THE CRYSTALGPD BENCHMARK

CRYSTALGPD characterizes crystal LLM capability boundaries across complementary generation settings, including DNG, CSP, composition- and symmetry-constrained CSD, and symmetry-constrained CSD. Rather than enforcing a single model interface, CRYSTALGPD preserves each system's native prompting and output representation while standardizing task instances, generation protocols, and downstream evaluation. This enables consistent system-level evaluation of discovery quality, physical validity, structural recovery, crystallographic controllability, and generalization.

3.1 BENCHMARK DESIGN PRINCIPLES

Native-interface preservation. Crystal LLMs differ substantially in their prompts and output formats, including CIF, Coordinate, Wyckoff-based, and reasoning-enhanced formats. These choices are integral to the systems being evaluated. CRYSTALGPD therefore retains each model's native prompt template and output format rather than forcing all systems into a shared interface.

Evaluation-side standardization. To enable consistent downstream evaluation, model outputs are converted into a crystal-structure object through representation-specific parsing. All models are then evaluated using the same task manifests and task-specific evaluators. This separates model-specific generation interfaces from the common crystallographic quantities used for evaluation.

Minimal model-specific post-processing. CRYSTALGPD evaluates model outputs without model-specific correction, filtering, or generation-side MLIP relaxation. Only representation-specific parsing is applied before evaluation. This design aims to measure the capabilities of the generated outputs themselves rather than those of additional repair or selection pipelines.

3.2 EVALUATION SCOPE AND MODEL CONFIGURATIONS

We organize the evaluated crystal LLMs according to their model-adaptation paradigms. CrystaLLM (Antunes et al., 2024) is trained from scratch on CIF sequences. LLaMat (Ahlawat et al., 2026) performs materials-domain continual pretraining (CPT) followed by supervised fine-tuning (SFT). CrystalTextLLM (Gruver et al., 2024), PLaID++ (Xu et al., 2026), CrysText (Mohanty et al., 2026), and CrystalReasoner (Wu et al., 2026) primarily adapt pretrained LLM backbones through post-training, using SFT alone or followed by preference optimization or reinforcement learning. Since CrysText and CrystalReasoner were not originally trained with unconditional generation prompts, we additionally construct CrysText++ and CrystalReasoner++ variants by incorporating unconditional prompts during training, enabling their evaluation on DNG.

The post-training-based models are independently retrained using their released implementations, whereas CrystaLLM and LLaMat are evaluated from publicly released checkpoints. DiffCSP (Jiao et al., 2023) and OMatG (Höllmer et al., 2025) are included as non-LLM reference methods for CSP, providing external anchors for physical validity and structure-recovery performance. CrystalICL is excluded from the main quantitative comparison because it follows a few-shot inference setting rather than the trained-generation setting considered here.

3.3 TASKS AND CAPABILITY DIMENSIONS

CRYSTALGPD evaluates complementary settings probing different crystal-generation capabilities:

De novo generation: discovery quality. DNG evaluates open-ended crystal discovery. We report **Valid** for structural validity, **Unique** for intra-set non-redundancy, and **Novel** for novelty relative to the reference set. **(Meta)Stable** quantify energetic feasibility, while **(M)SUN** measure the joint achievement of (meta)stability with uniqueness and novelty. ElemDiv, SGDiv, SizeDiv, and SiteDiv further characterize diversity across elements, space groups, crystal sizes, and Wyckoff sites.

Crystal structure prediction: structural recovery. CSP evaluates composition-conditioned structure prediction. **Valid** measures structural validity, **FM** measures valid formula matching, **MR@1** measures top-1 recovery of the reference structure, and **RMSD** measures structural deviation among matched prediction–reference pairs. Together, these metrics distinguish failures of basic physical generation from failures of exact structure recovery.

Composition- and symmetry-constrained design: joint controllability. This setting conditions generation jointly on a target chemical composition and space group. **Valid**, **FM**, **SGM**, and **FSGM** measure physical validity, valid formula matching, valid space-group matching, and valid generations satisfying both constraints, respectively. The decomposition allows us to distinguish composition preservation from exact symmetry realization.

Symmetry-constrained design: symmetry controllability and frequency sensitivity. This setting isolates space-group control by conditioning on a target space group while leaving the chemical composition unconstrained. We evaluate **SGM** across matched higher- and lower-frequency space groups within each crystal system, enabling analysis of how symmetry controllability varies with training-data exposure while partially controlling for differences among crystal systems.

3.4 UNIFIED EVALUATION PROTOCOL

Across all tasks, CRYSTALGPD standardizes target instances, generation settings, and evaluation while preserving model-native interfaces. Generated outputs undergo only the representation-specific parsing or Wyckoff expansion required to construct a common crystal-structure object. No

model-specific correction, filtering, or generation-side MLIP relaxation is applied before evaluation. Unless otherwise specified, all main evaluations use temperature 0.7, top-$p = 0.9$, and top-$k = 50$. Complete task definitions, metric formulations, dataset processing, model-training details, checkpoint provenance, and implementation settings are provided in the Appendix.

4 CRYSTALREACT: A BENCHMARK-DRIVEN AGENTIC CASE STUDY

The CRYSTALGPD evaluation exposes recurring difficulties in maintaining physical validity and satisfying exact crystallographic constraints. These observations motivate a complementary question: *can access to explicit crystallographic knowledge alleviate some of the capability limitations identified by the benchmark?* We study this question through CRYSTALREACT, an agentic crystal-generation framework that allows a reasoning model to request crystallographic information during inference. Our purpose is not to establish an optimal agentic architecture, but to use CRYSTALREACT as a diagnostic intervention for examining whether external crystallographic knowledge can improve model behavior on the capability dimensions exposed by CRYSTALGPD.

Inspired by ReAct-style reasoning and tool use (Yao et al., 2023; Li et al., 2025), CRYSTALREACT generates crystals through iterative interaction with a local crystallographic knowledge base. Given a generation task and the current inference state, the model may either request one piece of external information or produce a complete candidate crystal structure in the Wyckoff representation. Retrieved information is incorporated into the subsequent reasoning context, allowing the model to access structural, symmetry, and chemical knowledge when needed during generation.

Agentic crystallographic knowledge access. At each inference step, the model performs exactly one action: it either issues a tool request or generates a complete candidate structure. Tool responses are appended to the inference context and can inform subsequent structural reasoning. The underlying knowledge base contains reference crystal structures from the MP20 training split, crystallographic information for all 230 space groups, and elemental chemical properties such as oxidation states, electronegativity, and atomic radii.

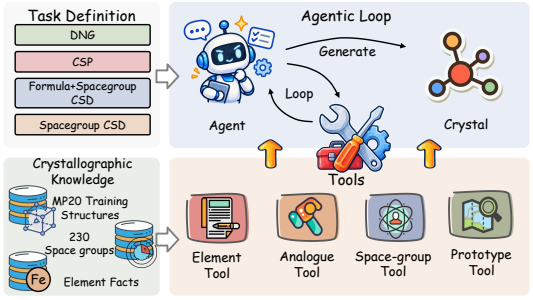


Figure 3: Overview of CRYSTALREACT. A reasoning model interacts with crystal-specific knowledge tools and produces a candidate crystal structure.

Crystal-specific knowledge tools. CRYSTALREACT provides complementary tools for accessing structural, symmetry, and chemical information. The *analogue* tool retrieves compositionally related structures from the MP20 training database, together with crystallographic descriptions (Ganose & Jain, 2019) and associated material properties. The *space-group* tool returns lattice constraints and valid Wyckoff positions for a requested symmetry, supporting symmetry-compatible site assignments. The *element* tool provides chemical properties including oxidation states, electronegativity, and atomic radii to support chemical compatibility reasoning. For DNG, the *prototype* tool additionally returns anonymous structural scaffolds that may guide open-ended generation. We additionally consider CRYSTALREACT W/VERIFY, which extends the available tool set with a verification tool. The verifier checks a generated candidate for structural validity and crystallographic consistency and returns violation feedback that the reasoning model can use to revise its generation.

Configuration. CRYSTALREACT uses GPT-5.6-Luna as its reasoning backbone. All inference runs use temperature 0.7 and top-p=0.9. Detailed tool schemas, knowledge sources, prompt templates, and implementation details are provided in Appendix E.1.

5 EXPERIMENTS

We use CRYSTALGPD to characterize four complementary capability dimensions of crystal LLMs: **RQ1** whether open-ended generation can jointly achieve physical feasibility and discovery quality; **RQ2** how physical validity and structural recovery scale with crystal complexity; **RQ3** how

Table 2: DNG results. Bold and underlined values indicate the best and second-best results among post-training-based crystal LLMs only. Complete results are reported in Appendix D.1.

Type	Model	Valid↑	Unique↑	Novel↑	Diversity				Stability		Metastability	
					ElemDiv↑	SGDiv↑	SizeDiv↑	SiteDiv↑	Stable↑	SUN↑	MetaStable↑	MSUN↑
From-scratch	CrystaLLM	91.96	86.95	10.18	0.68	0.45	0.26	16.92	11.79 (271)	0.26 (6)	84.34 (1939)	1.61 (37)
Continual-pretraining	LLaMat-2-CIF	89.28	87.41	22.27	0.71	0.24	0.22	7.30	0.85 (19)	0.00 (0)	52.28 (1167)	1.70 (38)
	LLaMat-3-CIF	94.76	82.14	14.77	0.70	0.23	0.22	6.57	1.56 (37)	0.34 (8)	65.85 (1560)	1.98 (47)
Post-training	CrystalTextLLM	71.00	90.82	34.08	**0.70**	0.18	0.23	5.26	0.34 (6)	0.00 (0)	24.62 (437)	0.51 (9)
	PLaID++-SFT	89.56	80.97	28.99	0.67	0.37	0.25	8.90	2.77 (62)	0.13 (3)	60.56 (1356)	3.08 (69)
	PLaID++-DPO	**95.84**	46.29	11.48	0.59	0.30	0.21	4.25	2.05 (49)	0.04 (1)	81.68 (1957)	1.71 (41)
	CrysText-SFT	47.76	93.19	90.97	0.64	0.19	0.25	11.07	0.17 (2)	0.00 (0)	5.62 (66)	1.02 (12)
	CrysText-GRPO	43.77	**94.51**	94.78	0.62	0.16	**0.26**	10.90	0.09 (1)	0.00 (0)	4.39 (48)	0.64 (7)
	CrysText++-SFT	83.87	82.10	12.17	0.69	**0.41**	0.25	12.97	**8.59** (180)	**0.19** (4)	**81.72** (1712)	0.76 (16)
	CrysText++-GRPO	83.68	82.17	13.48	0.69	**0.41**	0.25	13.09	8.22 (172)	0.10 (2)	79.68 (1667)	1.00 (21)
	CrystalReasoner-SFT	84.20	35.06	**96.86**	0.39	0.20	0.23	6.61	0.00 (0)	0.00 (0)	0.43 (9)	0.10 (2)
	CrystalReasoner-GRPO	93.28	14.92	94.17	0.37	0.09	0.22	5.45	0.64 (15)	0.00 (0)	1.67 (39)	0.21 (5)
	CrystalReasoner++-SFT	75.80	93.19	44.38	0.62	0.27	**0.26**	**14.48**	1.58 (30)	0.00 (0)	40.37 (765)	**3.48** (66)
	CrystalReasoner++-GRPO	88.04	86.69	33.85	0.63	0.25	0.25	12.40	2.50 (55)	0.00 (0)	49.43 (1088)	3.27 (72)
General LLM	GPT-5.6-Luna	66.96	73.48	57.59	0.61	0.15	0.24	10.88	0.24 (4)	0.00 (0)	15.95 (267)	0.72 (12)
Agentic	CrystalReAct	76.75	97.60	83.86	0.60	0.71	0.26	22.96	0.31 (6)	0.00 (0)	7.36 (141)	1.67 (32)
	CrystalReAct w/verify	79.04	98.53	84.46	0.65	0.72	0.26	22.36	0.35 (7)	0.05 (1)	6.98 (138)	2.02 (40)

reliably models satisfy composition and crystallographic symmetry constraints; and **RQ4** how crystallographic controllability changes under uneven training-data coverage. Across these analyses, we focus on recurring capability patterns rather than a single model ranking. We additionally report CRYSTALREACT as a benchmark-driven diagnostic intervention to examine whether access to external crystallographic knowledge can alleviate the identified limitations.

5.1 DE NOVO GENERATION: DISCOVERY QUALITY (RQ1)

Overall performance. Table 2 shows that no crystal LLM consistently optimizes all dimensions of open-ended generation. From-scratch and continual-pretraining models achieve relatively high validity and metastability but limited novelty. Post-training methods exhibit distinct optimization behaviors: PLaID++ provides balanced performance across stability, novelty, and uniqueness, CrysText++ achieves the highest metastability but relatively lower MSUN, while CrystalReasoner++ obtains the highest MSUN but relatively lower metastability. Compared with GPT-5.6-Luna, CRYSTALREACT explores a broader crystal space through crystallographic knowledge retrieval, achieving 0.71 spacegroup diversity and improving MSUN to 1.67%. The CRYSTALREACT W/VERIFY variant further improves MSUN to 2.02% by incorporating verification tool invocation.

Finding 1: Physical feasibility and novelty form a joint bottleneck for crystal discovery. Current crystal LLMs rarely generate structures that are simultaneously physically feasible and novel. Across evaluated models, the stable rate remains below 15%, and no model produces more than eight SUN structures. Relaxing the energy criterion to metastability increases feasible candidates, but many are removed by uniqueness and novelty requirements. Thus, the low SUN and MSUN rates reflect a compounded discovery bottleneck: generating energetically plausible structures is difficult, and discovering such structures beyond the reference distribution is harder still.

Finding 2: Improving individual generation objectives does not necessarily improve discovery yield. Existing post-training strategies optimize different aspects of crystal generation, but individual metric improvements do not consistently translate into higher SUN or MSUN. For example, PLaID++ improves validity and metastability while reducing uniqueness and novelty under DPO, whereas CrystalReasoner++ improves physical validity without improving joint discovery metrics. CrysText++, which is primarily optimized for structure prediction, maintains high uniqueness and diversity but generates few novel candidates. These patterns indicate that current post-training objectives remain only partially aligned with the multi-objective requirements of materials discovery.

5.2 CRYSTAL STRUCTURE PREDICTION: STRUCTURAL RECOVERY AND SCALING (RQ2)

Overall performance. Table 3 reports CSP performance on MP20 and MP52. Across crystal LLMs, performance decreases substantially from MP20 to MP52 in validity, formula matching, and top-1 structure recovery. Because model interfaces differ, these results should be interpreted as system-level comparisons: CrystaLLM, DiffCSP, and OMatG receive unit-cell compositions with

Table 3: CSP results on MP20 and MP52. Bold and underlined values indicate the best and second-best results among post-training-based crystal LLMs. Complete results are in Appendix D.2.

		MP20				MP52			
Type	Model	Valid	FM	MR@1	RMSD↓	Valid	FM	MR@1	RMSD↓
From-scratch	CrystaLLM	85.02	84.88	65.95	0.043	44.25	40.81	14.70	0.071
Continual-pretraining	LLaMat-2-CIF	81.46	75.13	45.81	0.069	47.63	35.69	11.59	0.131
	LLaMat-3-CIF	85.60	77.12	49.85	0.052	57.51	42.82	18.04	0.100
Post-training	CrystalTextLLM	57.03	55.91	25.53	0.090	25.69	23.13	2.99	0.217
	PLaID++-SFT	<u>78.27</u>	76.55	40.44	**0.048**	48.83	41.73	7.64	**0.098**
	PLaID++-DPO	76.41	74.62	39.19	**0.048**	45.81	39.52	7.13	**0.098**
	CrysText-SFT	36.44	36.10	5.24	0.251	24.64	22.85	2.00	0.290
	CrysText-GRPO	38.17	37.83	5.23	0.258	25.99	24.17	1.75	0.285
	CrysText++-SFT	74.31	74.08	<u>46.26</u>	0.058	38.14	37.35	**9.18**	0.117
	CrysText++-GRPO	76.46	76.33	**46.65**	0.057	39.66	38.81	<u>9.10</u>	0.113
	CrystalReasoner-SFT	72.16	71.76	32.26	0.058	41.35	39.43	5.35	0.120
	CrystalReasoner-GRPO	**79.73**	**79.36**	35.29	<u>0.055</u>	**50.46**	**47.82**	5.87	<u>0.104</u>
	CrystalReasoner++-SFT	71.91	71.19	32.25	0.059	41.97	40.00	5.49	0.117
	CrystalReasoner++-GRPO	78.24	<u>77.45</u>	34.71	0.056	<u>49.36</u>	<u>46.95</u>	5.87	<u>0.104</u>
Diffusion Flow Matching	DiffCSP	99.30	99.30	56.92	0.060	78.95	78.95	11.75	0.097
	OMatG	98.81	98.81	68.41	0.070	79.93	79.93	15.67	0.097
General LLM	GPT-5.6-Luna	77.71	76.01	14.09	0.060	72.72	68.75	2.78	0.102
Agentic	CrystalReAct	88.64	87.60	45.84	0.028	65.91	63.56	9.96	0.051
	CrystalReAct w/verify	92.93	92.64	46.88	0.027	82.36	81.42	10.18	0.054

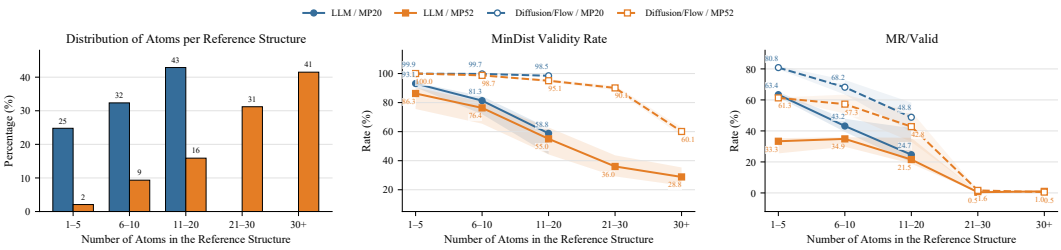


Figure 4: CSP performance by reference-structure size. Panels show the dataset distributions, MinDist validity, and MR/Valid. Solid lines/bands denote the median/interquartile range of post-training based crystal LLMs; dashed lines/bands denote the mean/range of DiffCSP and OMatG.

explicit atom counts, whereas other crystal LLMs operate from reduced formulas. Among post-training-based crystal LLMs, CrystalReasoner achieves the strongest validity and formula matching, while CrysText++ achieves the highest MR@1. The non-LLM DiffCSP and OMatG references retain stronger physical validity, providing an external anchor for CSP difficulty. Relative to GPT-5.6-Luna, CrystalReAct improves both physical validity and structure recovery, and the CrystalReAct w/verify variant further increases validity on both MP20 and MP52.

Finding 3: Physical validity becomes a scaling bottleneck as structural complexity increases. Figure 4 reveals a pronounced size dependence of physical validity. Across post-training-based crystal LLMs, median MinDist validity decreases from 93.1% to 58.8% on MP20 and from 86.3% to 28.8% on MP52. In contrast, the diffusion- and flow-based references degrade more gradually, retaining 90.1% validity for MP52 structures with 21–30 atoms and 60.1% for structures containing more than 30 atoms. The widening gap with atom count indicates that maintaining structurally validity becomes increasingly difficult for current crystal LLMs as complexity grows.

Finding 4: Structural recovery remains difficult even after successful physical generation. Differences in conditioning interfaces complicate direct comparison of absolute MR@1 across model families, particularly because some systems receive explicit unit-cell compositions while others receive only reduced formulas. To reduce the influence of invalid generations, Figure 4(c) reports MR/Valid, the match rate conditioned on producing a valid structure. Crystal LLMs still exhibit lower median MR/Valid than the diffusion- and flow-based references across shared atom-count ranges, indicating that physical validity alone does not guarantee accurate structural recovery. At the same time, selected reduced-formula systems, such as CrysText++, achieve competitive MR/Valid in several ranges (Appendix D.2), suggesting that the gap is not solely determined by the conditioning.

Table 4: Composition- and symmetry-constrained CSD results on MP20 and MP52. Bold and underlined values indicate the best and second-best results among post-training-based crystal LLMs only. Component-wise validity results are provided in Appendix D.3.

		MP20				MP52			
Type	Model	Valid	FM	SGM	FSGM	Valid	FM	SGM	FSGM
From-scratch	CrystaLLM	83.30	81.60	82.70	81.00	33.90	28.10	33.50	27.80
Continual-pretraining	LLaMat-2-CIF	83.30	76.80	33.70	33.50	47.70	35.40	6.20	5.80
	LLaMat-3-CIF	87.00	78.00	33.20	32.60	55.00	43.60	8.80	8.60
Post-training	CrystalTextLLM	59.80	58.80	25.50	25.40	25.10	22.20	2.40	2.30
	PLaID++-SFT	**84.00**	**82.80**	**81.90**	**80.90**	**41.50**	35.40	**34.80**	**30.30**
	PLaID++-DPO	81.50	80.10	79.80	78.70	35.70	29.30	31.70	27.10
	CrysText-SFT	77.70	77.10	70.20	69.90	32.20	30.60	18.30	17.70
	CrysText-GRPO	77.90	77.70	69.90	69.70	33.00	31.20	19.00	18.40
	CrysText++-SFT	76.90	76.80	61.40	61.40	32.50	31.70	14.00	14.00
	CrysText++-GRPO	78.10	78.10	63.80	63.80	39.40	**37.70**	15.50	15.50
	CrystalReasoner-SFT	71.60	71.30	61.00	60.80	34.20	31.60	20.80	19.40
	CrystalReasoner-GRPO	74.60	74.30	63.70	63.40	39.90	35.10	23.20	21.30
	CrystalReasoner++-SFT	71.80	70.70	61.20	60.50	34.60	31.60	22.30	20.90
	CrystalReasoner++-GRPO	76.10	75.10	64.40	63.60	39.30	35.90	23.90	22.30
General LLM	GPT-5.6-Luna	53.90	52.40	49.00	47.70	32.70	30.60	30.00	27.90
Agentic	CrystalReAct	84.20	84.30	82.00	80.30	47.40	40.80	46.00	39.70
	CrystalReAct w/verify	94.10	93.80	93.90	93.60	83.10	80.10	82.60	79.70

5.3 Composition- and Symmetry-Constrained Design: Joint Controllability (RQ3)

Overall performance. Table 4 evaluates the ability of crystal LLMs to jointly satisfy composition and space-group constraints. Performance decreases substantially from MP20 to MP52 across validity, formula matching, symmetry matching, and their joint satisfaction. Among post-training-based crystal LLMs, PLaID++ achieves the strongest joint constraint satisfaction on both datasets, while the remaining models exhibit different trade-offs between physical validity and symmetry control. GPT-5.6-Luna shows competitive symmetry matching on MP52 but lower overall validity. CrystalReAct improves both validity and constraint satisfaction relative to its backbone, while CrystalReAct w/verify further increases FSGM to 93.60% on MP20 and 79.70% on MP52.

Finding 5: Exact symmetry control is substantially harder than composition preservation. Table 5 decomposes joint constraint satisfaction into successive stages. From MP20 to MP52, median validity across post-training-based crystal LLMs falls from 76.90% to 34.60%. Conditional on producing a valid structure, however, formula preservation remains high: FM/Valid decreases only from 99.23% to 91.35%. By contrast, FSGM/FM decreases from 85.33% to 60.68%. Thus, after accounting for the large loss in physical validity, the dominant remaining challenge is realizing the requested symmetry rather than preserving chemical composition.

Table 5: Median stage-wise success rates across post-training-based crystal LLMs.

Dataset	Valid	FM/Valid	FSGM/FM	FSGM
MP20	76.90	99.23	85.33	63.60
MP52	34.60	91.35	60.68	19.40

Finding 6: Structural complexity explains only part of the MP20–MP52 generalization gap. Table 6 shows that crystal size contributes substantially to the observed degradation on MP52. Structures containing more than 20 atoms account for 70.8% of MP52 and achieve only 9.32% median FSGM, compared with 47.95% for structures with at most 20 atoms. However, controlling for atom count does not eliminate the gap: within the shared $\leq$ 20-atom range, MP52 still exhibits median decreases of 11.29 and 16.96 percentage points in Valid and FSGM relative to MP20. These results show that increasing structure size alone does not explain the generalization gap; current crystal LLMs also struggle under shifted combinations of composition and symmetry conditions.

Table 6: Median performance by atom count. The last row gives median paired MP20–MP52 differences for targets with at most 20 atoms.

Dataset subset	Proportion	Valid	FSGM
MP20	100.0	76.90	63.60
MP52 $\leq$ 20	29.2	63.01	47.95
MP52 $>$ 20	70.8	21.75	9.32
Difference	–	11.29	16.96

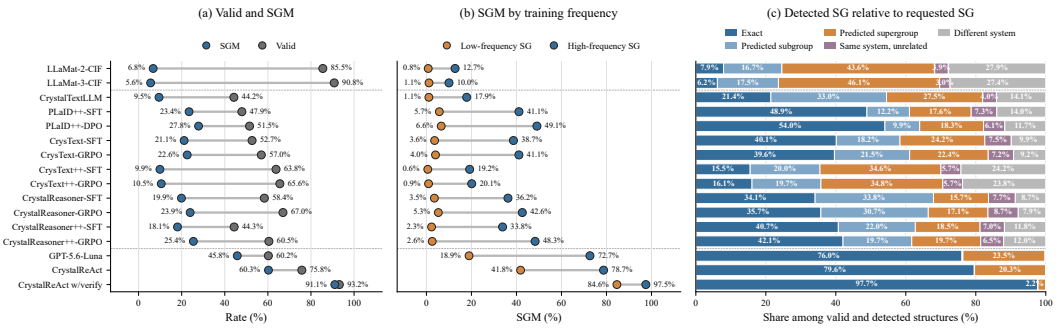


Figure 5: Space-group controllability in symmetry-constrained CSD. (a) Validity and space-group match rate (SGM). (b) Macro-averages SGM over the higher- and lower-frequency groups in each crystal-system pair. (c) Detected space groups relative to the requested space group among valid generations, categorized as exact, subgroup, supergroup, same-system unrelated, or different-system.

5.4 Symmetry-Constrained Design: Frequency Sensitivity (RQ4)

Overall performance. Figure 5(a) and (b) evaluate symmetry controllability independently. The results reveal large differences between structural validity and space-group realization: several models generate valid crystals at relatively high rates while matching the requested symmetry much less frequently. Post-training improves SGM primarily for the higher-frequency targets, with substantially smaller gains on the lower-frequency members of the matched crystal-system pairs. GPT-5.6-Luna demonstrates stronger symmetry controllability than the specialized crystal LLMs, while CrystalReAct further improves SGM, particularly on the lower-frequency targets. The CrystalReAct w/verify variant further improves both structural validity and symmetry satisfaction.

Finding 7: Symmetry controllability is strongly associated with training-set frequency. Figure 5(b) reveals a consistent gap between higher- and lower-frequency members of the seven crystal-system-matched pairs. Across the crystal LLMs, SGM ranges from 10.0% to 49.1% for higher-frequency targets but only from 0.6% to 6.6% for lower-frequency targets. Because the comparison is paired within each crystal system, it reduces variation from crystal-system differences and exposes a strong association between training-data frequency and exact symmetry controllability. Within the evaluated pairs, lower-frequency space groups remain substantially more difficult to realize.

Finding 8: Symmetry failures are often structured rather than arbitrary. Figure 5(c) further characterizes how generated space groups deviate from the requested symmetry. Although exact matches remain limited for many crystal LLMs, a substantial fraction of incorrect predictions fall into subgroup or supergroup relations with the target. Across the evaluated crystal LLMs, subgroup and supergroup predictions together account for a larger share than same-system-unrelated and different-system predictions. This pattern suggests that current models often retain partial information about the requested symmetry hierarchy, while failing to realize the exact constraint.

6 Conclusion

We present CrystalGPD, a unified benchmark for systematically characterizing crystal LLM capability boundaries across de novo generation, crystal structure prediction, composition- and symmetry-constrained crystal structure design, and symmetry-constrained crystal structure design. By preserving native prompts and output representations while standardizing task instances, generation protocols, representation-specific parsing, and downstream evaluation, CrystalGPD enables consistent comparison across heterogeneous crystal LLMs. Our evaluation reveals recurring limitations in achieving physically feasible and novel crystal discovery, maintaining validity and structural recovery as complexity increases, satisfying exact crystallographic constraints, and generalizing to less training-aligned conditions. Motivated by these findings, we further study CrystalReAct as a benchmark-driven agentic case study, showing that on-demand crystallographic knowledge and verification can improve physical validity and crystallographic constraint satisfaction. These results highlight the potential of integrating crystallographic knowledge into future crystal LLMs.

AI USE STATEMENT

In this work, we used generative AI tools for paper writing assistance, including language polishing and manuscript organization, as well as for code development and debugging. Additionally, we used generative AI tools to assist with the design and refinement of figures illustrating the proposed framework. We have reviewed all AI-assisted work and verified the correctness of the generated text, code, and figures through author inspection and revision. We take responsibility for the final content of this work, including all text, code, and figures.

REPRODUCIBILITY STATEMENT

We provide anonymous source code and implementation details to facilitate reproducibility. The code for the CRYSTALGPD benchmark, including dataset processing, evaluation protocols, and metric implementations, is available at `https://anonymous.4open.science/r/CrystalGPD`. The implementation of CRYSTALREACT, including the agent framework and inference pipeline, is available at `https://anonymous.4open.science/r/CrystalReAct`. Detailed descriptions of task definitions, datasets, evaluation settings, model configurations, and training procedures are provided in Sections 3 and 4, together with the corresponding appendices.

REFERENCES

Dhruv Ahlawat, Vaibhav Mishra, Somaditya Singh, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, Mausam, and NM Anoop Krishnan. A family of large language models for materials research with insights into model adaptability in continued pretraining. *Nature Machine Intelligence*, 8(3):435–448, 2026.

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024.

Kohei Shinohara Atsushi Togo and Isao Tanaka. Spglib: a software library for crystal symmetry search. *Sci. Technol. Adv. Mater., Meth.*, 4(1):2384822–2384836, 2024. doi: 10.1080/27660400.2024.2384822. URL `https://doi.org/10.1080/27660400.2024.2384822`.

Sterling G. Baird, Hasan M. Sayeed, Joseph H. Montoya, and Taylor D. Sparks. matbench-genmetrics: A python library for benchmarking crystal structure generative models using time-based splits of materials project structures. *J. Open Source Softw.*, 9(97):5618, 2024. doi: 10.21105/JOSS.05618. URL `https://doi.org/10.21105/joss.05618`.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html`.

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

Théo Cavignac, Jonathan Schmidt, Pierre-Paul De Breuck, Antoine Loew, Tiago F. T. Cerqueira, Hai-Chen Wang, Anton Bochkarev, Yury Lysogorskiy, Aldo H. Romero, Ralf Drautz, Silvana Botti, and Miguel A. L. Marques. AI-driven expansion and application of the Alexandria database. *Journal of Physics: Materials*, 9(2):025014, 2026. doi: 10.1088/2515-7639/ae6620. URL `https://doi.org/10.1088/2515-7639/ae6620`.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16,*

2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Alexandre Duval, Siddharth Betala, Samuel P. Gleason, Andy Xu, Georgia Channing, Daniel Levy, Ali Ramlaoui, Clémentine Fourrier, Chaitanya K. Joshi, Nikita Kazeev, Sékou-Oumar Kaba, Félix Therrien, Alex Hernández-García, Rocío Mercado, and N M Anoop Krishnan. Lemat-genbench: Bridging the gap between crystal generation and materials discovery. In *AI for Accelerated Materials Design - NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=ZfPGcTfDWn.

Alex M Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary W. Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=vN9fpfqoP1.

Philipp Höllmer, Thomas Egg, Maya M. Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian E. Roitberg, Mingjie Liu, George Karypis, Mark K. Transtrum, Richard G. Hennig, Ellad B. Tadmor, and Stefano Martiniani. Open materials generation with stochastic interpolants. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/hollmer25a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013. ISSN 2166-532X. doi: 10.1063/1.4812323. URL https://doi.org/10.1063/1.4812323.

Shuyi Jia, Aamod Varma, Pranav Manivannan, Dhruva Chayapathy, and Victor Fung. Benchmarking text representations for crystal structure generation with large language models. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/38b787fc530d0b31825827e2cc306656-Abstract-Conference.html.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=jkvZ7v4OmP.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=xnssGv9rpW.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.276. URL https://aclanthology.org/2025.emnlp-main.276/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Trupti Mohanty, Maitrey Mehta, Hasan M Sayeed, Bat-El Oded, Itay Pitussi, Arie Borenstein, Vivek Srikumar, and Taylor D Sparks. Crystext: a generative ai approach for text-conditioned crystal structure generation using llm. *Integrating Materials and Manufacturing Innovation*, 15(2):416–430, 2026.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *CoRR*, abs/2410.22570, 2024. doi: 10.48550/ARXIV.2410.22570. URL https://doi.org/10.48550/arXiv.2410.22570.

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. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2012.10.028. URL https://www.sciencedirect.com/science/article/pii/S0927025612006295.

Chris J Pickard. Airss data for carbon at 10gpa and the c+ n+ h+ o system at 1gpa. *(No Title)*, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65(11):1501–1509, 2013.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Martin Siron, Inel Djafar, Ali Ramlaoui, Etienne du Fayette, Amandine Rossello, Edvin Fako, Matthew McDermott, Felix Therrien, Luis Barroso-Luque, Flaviu Cipcigan, et al. Lematbulk: aggregating, and de-duplicating quantum chemistry materials databases. *arXiv preprint arXiv:2511.05178*, 2025.

Ruobing Wang, Qiaoyu Tan, Yili Wang, Ying Wang, and Xin Wang. CrystalICL: Enabling in-context learning for crystal generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods*

in Natural Language Processing, pp. 18429–18444, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.929. URL https://aclanthology.org/2025.emnlp-main.929/.

Lai Wei, Sadman Sadeed Omee, Rongzhi Dong, Nihang Fu, Yuqi Song, Edirisuriya Siriwardane, Meiling Xu, Chris Wolverton, and Jianjun Hu. Cspbench: a benchmark and critical evaluation of crystal structure prediction. *arXiv preprint arXiv:2407.00733*, 2024.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco S. Cohen, Abhishek Das, Sushree Jagriti Sahoo, Ammar Rizvi, Zachary W. Ulissi, and Larry Zitnick. UMA: A family of universal models for atoms. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/bbf23e81b0ad7637fe9a731d0b676ca6-Abstract-Conference.html.

Yuyang Wu, Stefano Falletta, Delia McGrath, and Sherry Yang. Crystalreasoner: Reasoning and RL for property-conditioned crystal structure generation. *CoRR*, abs/2605.14344, 2026. doi: 10.48550/ARXIV.2605.14344. URL https://doi.org/10.48550/arXiv.2605.14344.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=03RLpj-tc_.

Andy Xu, Rohan Desai, Larry Wang, Ethan T. Ritz, and Gabriel Hope. PLaID++: A preference aligned language model for targeted inorganic materials design. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wFThVGzmvq.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

APPENDIX CONTENTS

A Task and Metric Definitions **15**

A.1 De Novo Generation 15

A.2 Crystal Structure Prediction 15

A.3 Composition- and Symmetry-Constrained Crystal Structure Design 16

A.4 Symmetry-Constrained Crystal Structure Design 16

B Dataset Details **17**

B.1 LeMat-Bulk 17

B.2 Materials Project 17

C Implementation Details and Reproducibility **18**

C.1 Unified Evaluation Protocol 18

C.2 Model Preparation and Training Configuration 18

D Additional Results and Analyses **19**

D.1 Complete De Novo Generation Results 19

D.2 Complete Crystal Structure Prediction Results 20

D.3 Complete Composition- and Symmetry-Constrained Crystal Structure Design Results 21

D.4 Complete Symmetry-Constrained Crystal Structure Design Results 22

D.5 Additional Released-Checkpoint Results 25

E CrystalReAct **26**

E.1 CrystalReAct: Agentic Search for Crystal Generation 26

E.2 Crystallographic Information 26

E.3 Crystal-Specific Tools 27

E.4 Implementation Details 29

E.5 Additional Experiments on Smaller Models 29

E.6 Prompt Design 30

A Task and Metric Definitions

A.1 De Novo Generation

De novo generation (DNG) is an unconditional crystal structure generation task. Under the same unconditional input condition, each model uses rejection sampling to obtain 2,500 parseable structure outputs, which are evaluated following the LeMat-GenBench protocol (Duval et al., 2025). The task assesses the quality and materials-discovery potential of the generated crystal set in terms of validity, stability, uniqueness, novelty, and diversity.

Validity. Validity measures whether a generated crystal structure satisfies basic chemical and crystallographic constraints. Following the official LeMat-GenBench protocol (Duval et al., 2025), we evaluate validity through three complementary checks: Charge Neutrality (CN), Minimum-Distance Validity (MinDist), and Physical Plausibility (PhysPlaus). Charge Neutrality (CN) evaluates whether the generated composition admits an oxidation-state assignment with zero net charge. Minimum-Distance Validity (MinDist) checks whether all pairwise interatomic distances in the generated structure exceed 0.7 Å, preventing physically implausible atomic overlaps. Physical Plausibility (PhysPlaus) evaluates whether the generated structure satisfies basic crystallographic and physical constraints. Specifically, the structure must have a parseable crystallographic representation, a mass density between 0.01 and 25 g/cm^3, an atomic density between 10^{-5} and 0.5 atoms/Å^3, lattice lengths between 1 and 100 Å, lattice angles between 0° and 180°, and an identifiable space group assigned by `SpacegroupAnalyzer` (Atsushi Togo & Tanaka, 2024). Overall Valid is defined as the proportion of generated structures that pass all three checks. A structure is considered invalid if it fails any individual component.

Stability. Formation energy is the energy per atom relative to the elemental chemical potentials. The evaluator uses machine learning interatomic potentials (MLIPs), including MACE (Batatia et al., 2022), ORB (Neumann et al., 2024), and UMA (Wood et al., 2025), to evaluate the energetic properties of generated structures. Energy above the convex hull (E_{hull}) is the energy per atom above the lowest-energy combination of competing phases on the LeMat-Bulk convex hull. A crystal is stable when $E_{\text{hull}} \leq 0$ and metastable when $0 < E_{\text{hull}} \leq 0.1$ eV/atom. The evaluator additionally relaxes each generated structure with the MLIPs and reports the RMSD between the submitted and relaxed atomic positions. Lower formation energy, E_{hull}, and RMSD indicate higher potential stability.

Uniqueness, novelty, and discovery yield. Uniqueness measures structural non-redundancy within the valid generated set, whereas Novelty identifies valid generated structures that do not match any structure in LeMat-Bulk. SUN counts stable, unique, and novel candidates, whereas MSUN counts metastable, unique, and novel candidates. In our tables, Unique, Novel, Stable, SUN, Meta, and MSUN rates are normalized by the number of valid structures; parenthesized entries give the corresponding absolute counts.

Distribution, diversity, and supply concentration. All distribution metrics compare the valid generated set with LeMat-Bulk. JS denotes the Jensen–Shannon distance between discrete-property distributions, MMD is a kernel discrepancy between continuous-property distributions, and FID is the Fréchet distance between MLIP structural-embedding distributions; lower values indicate closer agreement. ElemDiv, SGDiv, SizeDiv, and SiteDiv describe coverage of elements, space groups, crystal sizes, and atomic sites, respectively, with higher values indicating greater diversity. Production (Prod) and reserve (Res) HHI are composition-weighted Herfindahl–Hirschman indices derived from elemental production and reserve concentration. Both are reported on a 0–10 scale, and lower values indicate lower supply concentration.

A.2 Crystal Structure Prediction

Crystal structure prediction (CSP) aims to generate a crystal structure from a given chemical composition. We evaluate CSP on the standard test splits of MP20 (Xie et al., 2022) and MPTS-52 (Jiao et al., 2023), hereafter abbreviated as MP52. The original test sets contain 9,046 and 8,096 target

Table 7: Space-group pairs used in symmetry-constrained CSD. Each crystal system contributes one higher-frequency and one lower-frequency space group, selected according to their occurrence counts in the MP20 training set. Each space group is evaluated with 500 generations.

	Higher-frequency		Lower-frequency	
Crystal system	Space group	Training count	Space group	Training count
Triclinic	2	676	1	470
Monoclinic	12	1,273	3	42
Orthorhombic	62	1,187	20	22
Tetragonal	139	1,233	137	24
Trigonal	166	1,076	150	21
Hexagonal	189	469	180	23
Cubic	225	3,960	217	20

structures, respectively. After removing MP52 structures whose Materials Project IDs occur in the MP20 training split, the evaluated MP52 set contains 6,298 targets.

Under our evaluation protocol, each model receives the chemical formula of a target structure and generates exactly one candidate. This single-sample setting defines the top-1 CSP evaluation. We use Match Rate at $k = 1$ (MR@1) and the root-mean-square deviation (RMSD) over matched prediction–target pairs as the primary metrics. Overall validity and Formula Match characterize the preceding stages of valid structure generation and composition preservation.

For N target structures $\{x_i\}_{i=1}^{N}$ and their corresponding predictions $\{\hat{x}_i\}_{i=1}^{N}$, MR@1 is defined as

$$\mathrm{MR@1} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\mathrm{Match}(\hat{x}_i, x_i)]\,. \tag{1}$$

Structural matching is performed with the `StructureMatcher` implementation provided by `pymatgen`(Ong et al., 2013), using $\mathrm{ltol} = 0.3$, $\mathrm{stol} = 0.5$, and an angular tolerance of 10°. MR@1 is computed over the complete test set, with each unsuccessful prediction assigned a zero match indicator. RMSD is averaged only over successfully matched prediction–target pairs. Valid denotes the proportion of structurally valid candidates, and Formula Match (FM) denotes the proportion of candidates that are both valid and consistent with the input formula.

A.3 Composition- and Symmetry-Constrained Crystal Structure Design

Composition- and symmetry-constrained crystal structure design (CSD) conditions generation jointly on a target chemical formula and space group. We evaluate all model configurations on the same fixed subsets, with 1,000 structures randomly sampled from MP20 and 1,000 structures randomly sampled from MP52 using seed 42. Performance is assessed by the ability of the generated structures to satisfy the prescribed composition and symmetry constraints.

Overall Validity (Valid) follows the validity definition above and measures the fraction of generation attempts that produce valid crystal structures. Formula Match (FM) is the proportion of prompts that produce valid structures whose reduced compositions match the requested formulas. Space-Group Match (SGM) is the proportion that produce valid structures with the requested space groups. The generated space-group number is determined using `SpacegroupAnalyzer` from PYMATGEN(Ong et al., 2013) with `symprec` $= 0.1$ Å. Formula-and-Space-Group Match (FSGM) is the proportion of prompts that produce valid structures satisfying both conditions.

A.4 Symmetry-Constrained Crystal Structure Design

The evaluation panel comprises 14 space groups spanning all seven crystal systems. We determine the frequency of each space group from the 27,136 structures in the MP20 training set and select one higher-frequency and one lower-frequency group from each crystal system, forming seven matched pairs (Table 7). Each model generates 500 candidate structures for every target space group, resulting in 7,000 generations in total. This paired design evaluates space-group control under different levels of training-data coverage while accounting for differences among crystal systems.

Validity and Space-Group Match. Overall Validity (Valid) follows the validity definition above and measures the fraction of generation attempts that produce valid crystal structures. For each valid structure, the detected space-group number is assigned using `SpacegroupAnalyzer` from `pymatgen` (Ong et al., 2013) with `symprec` $= 0.1$ Å. Space-Group Match (SGM) is the proportion of all generated candidates that are valid and whose detected space-group number equals the requested space-group number. We report Valid and SGM for each target space group and compute Overall SGM by pooling all 7,000 generations.

Frequency-Stratified Space-Group Match. Higher and Lower denote the macro-averaged SGM values across the seven higher-frequency and seven lower-frequency target groups, respectively. These frequency labels are assigned within each crystal-system pair according to the occurrence counts in the MP20 training split. For example, SG 2 and SG 1 form the triclinic pair, with SG 2 assigned to the Higher subset based on its 676 training structures and SG 1 assigned to the Lower subset based on its 470 training structures. Gap is defined as Higher minus Lower and is reported in percentage points.

Space-Group Relation Distribution. In addition to exact space-group matching, we characterize the relation between the requested space group G and the detected space group H among valid generations with an assigned space group. Each pair is categorized as *exact* when $H = G$, *predicted subgroup* when H is a subgroup of G, or *predicted supergroup* when H is a supergroup of G. The remaining pairs are categorized as *same-system unrelated* when H and G belong to the same crystal system and as *different-system* when they belong to different crystal systems. We report the proportion of each category among valid generations with an assigned space group.

B Dataset Details

B.1 LeMat-Bulk

LeMat-Bulk (Siron et al., 2025) is a large-scale unified bulk materials database constructed by aggregating and de-duplicating several quantum chemistry databases, including Materials Project (Jain et al., 2013), OQMD (Saal et al., 2013), and Alexandria (Cavignac et al., 2026). It contains over 5.3 million PBE-calculated inorganic crystal structures and additionally provides large collections of PBESol- and SCAN-level calculations. The dataset applies structure standardization and a dedicated graph-based fingerprinting approach to remove duplicated entries across databases. We use LeMat-Bulk (Siron et al., 2025) as the reference database to evaluate the novel crystal discovery performance of crystal LLMs on the de novo generation task.

B.2 Materials Project

We conduct the conditional-generation evaluations on MP20 and MP52, both derived from the Materials Project (Jain et al., 2013). MP20 follows the standard split introduced by CDVAE (Xie et al., 2022), comprising 27,136 training, 9,047 validation, and 9,046 test structures. We use the MP20 training split to train the evaluated models and treat it as the reference training distribution for subsequent analyses, including the identification of overlapping samples in MP52 and the estimation of space-group frequencies. MP52 follows the dataset split used by DiffCSP (Jiao et al., 2023) and contains 8,096 structures in its original test set. Compared with MP20, MP52 covers a broader range of structure sizes and is used to evaluate model generalization beyond the structural range represented during MP20 training.

Because the MP52 test set contains structures that also appear in the MP20 training split, we remove all overlapping structures according to their Materials Project IDs before evaluation. This filtering leaves 6,298 structures in the MP52 test set. The filtered set is used for the full MP52 CSP evaluation and also serves as the source set for sampling the MP52 targets used in composition- and symmetry-constrained CSD. MP20 provides the standard evaluation distribution adopted by prior crystal-generation studies, while the filtered MP52 set enables evaluation under a broader and less training-aligned structural distribution.

Table 8: Crystal LLMs evaluated in CRYSTALGPD.

Model	Variant	Backbone	Native output	Training Method
CrystaLLM	-	GPT-2	CIF	From Scratch
LLaMat2/3-CIF	-	LLaMA-2-7B / LLaMA-3-8B	Coordinates	CPT + SFT
CrystalTextLLM	-	LLaMA-2-7B	Coordinates	SFT
PLaID++	SFT / DPO	Qwen2.5-7B	Wyckoff	SFT + Iterative DPO
CrysText	SFT / GRPO	Mistral-7B-v0.3	CIF	SFT + GRPO
CrystalReasoner	SFT / GRPO	Qwen2.5-3B-MegaScience	Reasoning + Coordinates	SFT + GRPO

C IMPLEMENTATION DETAILS AND REPRODUCIBILITY

C.1 UNIFIED EVALUATION PROTOCOL

Unless otherwise specified, all main crystal-generation experiments use the same decoding parameters: a temperature of 0.7, top-p of 0.9, and top-k of 50. Coordinate- and Wyckoff-based representations use a maximum generation length of 500 tokens, increased to at least 800 tokens for MP52 to accommodate structures with larger numbers of atoms. The maximum generation lengths for CIF and reasoning+Coordinate outputs are 3,000 and 3,584 tokens, respectively. These representation-specific limits reduce output truncation while preserving the same number of candidates per target.

To improve comparability across model families, we standardize the task conditions and target manifests, the number of candidates generated for each target, the decoding parameters, the structure-level interpretation of parsed outputs, and the downstream evaluation pipeline. Each model retains its native prompt template and output representation because these components are part of the system being evaluated. Generated outputs undergo only the representation-specific parsing or Wyckoff expansion required to construct a common crystal-structure object. We apply no model-specific text correction, candidate filtering, reranking, or structural relaxation before evaluation. For example, the candidate-relaxation procedure in the original PLaID++ pipeline is excluded from generation-side post-processing. Structural relaxation required for energy-based metrics is instead performed uniformly by the shared evaluator and is therefore distinct from model-specific post-processing.

C.2 MODEL PREPARATION AND TRAINING CONFIGURATION

The main comparison includes two sources of model checkpoints. CrystalTextLLM, PLaID++, CrysText, and CrystalReasoner, together with the CrysText++ and CrystalReasoner++ variants, are independently retrained using their released implementations and subsequently evaluated with the unified CRYSTALGPD pipeline. In contrast, CrystaLLM, LLaMat-2-CIF, LLaMat-3-CIF, DiffCSP, and OMatG are evaluated from publicly released checkpoints. In all cases, we generate predictions from the selected checkpoints and recompute the reported metrics using the same benchmark datasets, structure-processing procedures, and downstream evaluators. Table 8 summarizes the evaluated crystal LLMs, representations, and training paradigms.

The use of released checkpoints reflects model-specific considerations. For CrystaLLM, we use the official `crystallm_mp_20_large` checkpoint to preserve its original from-scratch training regime and native conditioning interface, which differ from the post-training pipelines used by the other crystal LLMs. LLaMat is evaluated from its released LLaMat-2-CIF and LLaMat-3-CIF checkpoints because the data and resources required for controlled retraining are not fully available. DiffCSP and OMatG are included as specialized non-LLM CSP baselines rather than as models to be retrained in our study; we therefore evaluate their released task-specific checkpoints, using the `MP-20-CSP/Linear-ODE` configuration for OMatG. For the independently retrained models, training is conducted on four NVIDIA A40 GPUs with 40 GB of memory per GPU. We follow the settings described in the corresponding papers and released implementations; the only overridden fields are the SFT epoch count and effective batch size for CrystalTextLLM, the SFT learning rate and effective batch size for PLaID++, and the effective batch size for PLaID++ DPO. Evaluations of the publicly released checkpoints corresponding to these independently retrained model families are reported separately in Appendix D.5.

CrystalTextLLM SFT uses 10 epochs, an effective batch size of 32, and a learning rate of 5×10^{-4}. We use 4-bit FP4 LoRA(Hu et al., 2022) with rank 8, scaling factor 32, dropout 0.05, and

Table 9: Reproduced DNG quality results at temperature 0.7. Rates are percentages, parenthesized values are counts, energies are in eV/atom, and RMSD is in Å.

Model	Valid				Unique↑	Novel↑	Energy-based			Stability		Metastability	
	Overall↑	CN↑	MinDist↑	PhysPlaus↑			E_{form} (Std)↓	E_{hull} (Std)↓	RMSD (Std)↓	Stable↑	SUN↑	Meta↑	MSUN↑
CrystaLLM	91.96	98.88	92.84	99.52	86.95	10.18	-0.99 ± 0.92	0.12 ± 0.41	0.16 ± 0.35	11.79 (271)	0.26 (6)	84.34 (1939)	1.61 (37)
LLaMat-2-CIF	89.28	97.08	92.60	99.36	87.41	22.27	-0.56 ± 0.95	0.29 ± 0.58	0.35 ± 0.52	0.85 (19)	0.00 (0)	52.28 (1167)	1.70 (38)
LLaMat-3-CIF	94.76	97.60	97.08	99.88	82.14	14.77	-0.82 ± 0.99	0.18 ± 0.45	0.27 ± 0.39	1.56 (37)	0.34 (8)	65.85 (1560)	1.98 (47)
CrystalTextLLM	71.00	95.40	75.24	99.32	90.82	34.08	-0.21 ± 1.35	0.72 ± 1.02	0.62 ± 0.77	0.34 (6)	0.00 (0)	24.62 (437)	0.51 (9)
PLaID++-SFT	89.56	99.40	90.00	100.00	80.97	28.99	-0.68 ± 1.02	0.33 ± 0.72	0.35 ± 0.65	2.77 (62)	0.13 (3)	60.56 (1356)	3.08 (69)
PLaID++-DPO	95.84	99.92	95.88	100.00	46.29	11.48	-0.43 ± 0.53	0.13 ± 0.43	0.19 ± 0.36	2.05 (49)	0.04 (1)	81.68 (1957)	1.71 (41)
CrysText-SFT	47.76	97.11	49.80	98.29	93.19	90.97	0.96 ± 1.45	1.64 ± 1.29	1.58 ± 0.86	0.17 (2)	0.00 (0)	5.62 (66)	1.02 (12)
CrysText-GRPO	43.77	97.32	45.21	99.52	94.51	94.78	0.98 ± 1.43	1.66 ± 1.26	1.62 ± 0.83	0.09 (1)	0.00 (0)	4.39 (48)	0.64 (7)
CrysText++-SFT	83.87	98.84	85.07	99.76	82.10	12.17	-0.85 ± 0.92	0.18 ± 0.53	0.19 ± 0.44	8.59 (180)	0.19 (4)	81.72 (1712)	0.76 (16)
CrysText++-GRPO	83.68	98.76	84.92	99.68	82.17	13.48	-0.84 ± 0.99	0.20 ± 0.59	0.21 ± 0.44	8.22 (172)	0.10 (2)	79.68 (1667)	1.00 (21)
CrystalReasoner-SFT	84.20	94.08	88.80	100.00	35.06	96.86	1.24 ± 1.14	1.67 ± 1.12	1.47 ± 6.45	0.00 (0)	0.00 (0)	0.43 (9)	0.10 (2)
CrystalReasoner-GRPO	93.28	95.04	98.24	99.96	14.92	94.17	0.91 ± 0.98	1.31 ± 1.04	0.76 ± 0.93	0.64 (15)	0.00 (0)	1.67 (39)	0.21 (5)
CrystalReasoner++-SFT	75.80	97.08	78.00	100.00	93.19	44.38	-0.17 ± 1.09	0.56 ± 0.95	0.54 ± 0.78	1.58 (30)	0.00 (0)	40.37 (765)	3.48 (66)
CrystalReasoner++-GRPO	88.04	97.92	90.04	100.00	86.69	33.85	-0.29 ± 0.79	0.31 ± 0.56	0.35 ± 0.53	2.50 (55)	0.00 (0)	49.43 (1088)	3.27 (72)
CrystalReAct	76.75	99.08	77.47	99.92	97.60	83.86	-0.34 ± 1.73	1.45 ± 1.61	1.21 ± 1.06	0.31 (6)	0.00 (0)	7.36 (141)	1.67 (32)

the `q_proj` and `v_proj` target modules. Training uses a cosine learning-rate schedule with 100 warmup steps, zero weight decay, and a maximum sequence length of 2,048 tokens.

PLaID++ SFT uses 10 epochs, an effective batch size of 64, and a learning rate of 5×10^{-4}. It uses 4-bit FP4 LoRA(Dettmers et al., 2023) with rank 8, scaling factor 32, dropout 0.05, and the `q_proj` and `v_proj` target modules, together with a cosine learning-rate schedule, 100 warmup steps, zero weight decay, and a maximum sequence length of 2,048 tokens. The iterative DPO (Rafailov et al., 2023) stage uses seven rounds, one epoch per round, an effective batch size of 32, a learning rate of 10^{-6}, and $\beta = 0.1$.

CrysText retains its original 4-bit NF4 LoRA(Hu et al., 2022) configuration with rank 16, scaling factor 16, and zero dropout, applied to the attention and MLP projection modules. Its SFT stage uses 10 epochs, an effective batch size of 32, a learning rate of 2×10^{-4}, the 8-bit AdamW (Loshchilov & Hutter, 2019) optimizer, a cosine learning-rate schedule, five warmup steps, a weight decay of 0.01, and a maximum sequence length of 2,048 tokens. The subsequent GRPO (Shao et al., 2024) stage uses one epoch, an effective batch size of 12, a learning rate of 10^{-6}, six generations per prompt, $\beta = 0.001$, and $\epsilon = 0.2$.

CrystalReasoner retains its original BF16 full-parameter FSDP2 configuration with a maximum sequence length of 4,096 tokens. Its SFT stage uses two epochs, an effective batch size of 32, a learning rate of 10^{-4}, the Adam (Kingma & Ba, 2015) optimizer, a cosine learning-rate schedule, a warmup ratio of 0.1, and a weight decay of 0.01. Its GRPO stage uses one epoch, a prompt batch size of 64, eight generations per prompt, a PPO mini-batch size of 32, an actor learning rate of 10^{-5}, a rollout temperature of 1.0, top-$p = 1.0$, and a maximum completion length of 4,096 tokens.

To support unconditional generation, CrysText++ samples equally from unconditional, formula-conditioned, and formula-and-space-group-conditioned prompts. CrystalReasoner++ adds unconditional prompts to the original set of randomly selected conditional combinations and samples all prompt types with equal probability. Apart from these prompt-sampling schemes, the two variants retain the training settings of their corresponding base models.

D Additional Results and Analyses

D.1 Complete De Novo Generation Results

Tables 9,10 report the complete DNG results at temperature 0.7. Several models achieve high structural validity and favorable energetic properties under this sampling setting. PLaID++-DPO and CrystalReasoner-GRPO show particularly strong geometric validity, while the CrysText++ variants achieve more favorable energy-related metrics and generate larger numbers of stable or metastable candidates. These improvements do not always translate into higher uniqueness or novelty, indicating a quality–diversity trade-off among different generation objectives rather than a single ordering across all evaluation metrics.

Table 10: Distributional and diversity diagnostics for reproduced DNG at temperature 0.7. Downward and upward arrows indicate the preferred direction.

Model	Distribution			Diversity				HHI	
	JS↓	MMD↓	FID↓	ElemDiv↑	SGDiv↑	SizeDiv↑	SiteDiv↑	Prod↓	Res↓
CrystaLLM	0.27	0.011	2.23	0.68	0.45	0.26	16.92	3.53	2.62
LLaMat-2-CIF	0.34	0.004	1.58	0.71	0.24	0.22	7.30	4.08	2.97
LLaMat-3-CIF	0.35	0.004	1.96	0.70	0.23	0.22	6.57	3.96	2.99
CrystalTextLLM	0.36	0.006	2.24	0.70	0.18	0.23	5.26	4.00	2.76
PLaID++-SFT	0.36	0.006	1.93	0.67	0.37	0.25	8.90	3.32	2.45
PLaID++-DPO	0.47	0.038	3.80	0.59	0.30	0.21	4.25	3.77	2.42
CrysText-SFT	0.46	0.015	5.71	0.64	0.19	0.25	11.07	4.09	2.90
CrysText-GRPO	0.46	0.015	5.89	0.62	0.16	0.26	10.90	4.19	2.96
CrysText++-SFT	0.25	0.007	1.79	0.69	0.41	0.25	12.97	3.71	2.67
CrysText++-GRPO	0.26	0.008	1.72	0.69	0.41	0.25	13.09	3.69	2.68
CrystalReasoner-SFT	0.54	0.082	12.39	0.39	0.20	0.23	6.61	1.87	1.92
CrystalReasoner-GRPO	0.57	0.060	11.30	0.37	0.09	0.22	5.45	2.50	2.07
CrystalReasoner++-SFT	0.33	0.002	1.62	0.62	0.27	0.26	14.48	4.15	2.61
CrystalReasoner++-GRPO	0.38	0.002	1.78	0.63	0.25	0.25	12.40	4.40	2.77
CrystalReAct	0.36	0.074	22.09	0.58	0.67	0.26	20.67	2.10	1.78

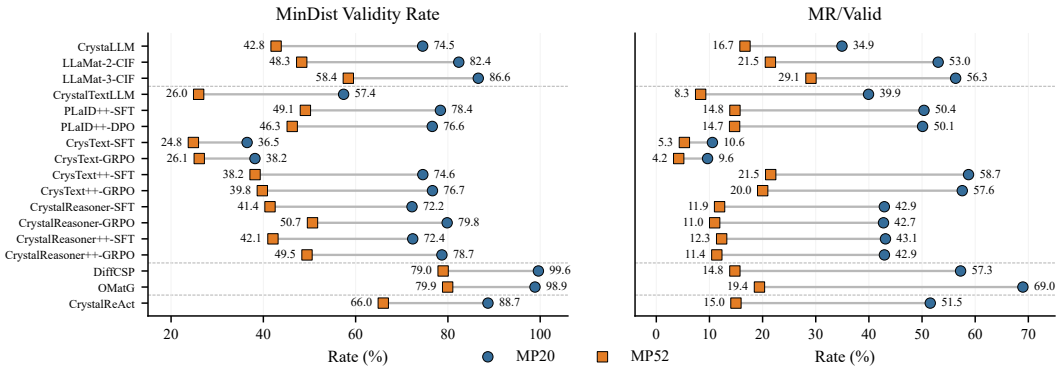


Figure 6: CSP task performance on MP20 and MP52. The left and right panels report MinDist validity and MR/Valid, respectively, where MR/Valid denotes the top-1 structure match rate among valid generations. Circle and square markers denote MP20 and MP52, and connecting segments pair the results for the same model across the two datasets. Dashed separators distinguish model groups trained under different paradigms.

The distributional diagnostics at temperature 0.7 further distinguish models whose validity scores are similar. The CrysText++ and CrystalReasoner++ variants generally provide better agreement with the reference distribution and broader space-group or site-level coverage, indicating that these model variants improve distributional coverage beyond DNG task performance.

D.2 Complete Crystal Structure Prediction Results

Table 11 supplements the primary CSP metrics with component-wise validity results on MP20 and MP52. The decomposition separates failures of geometric feasibility from failures to recover the requested composition or reference structure.

Before interpreting cross-model differences, we note that the composition representation provided to the models is not identical. CrystaLLM, DiffCSP, and OMatG receive the full chemical formula of the target unit cell, including the integer atom counts in that cell. The remaining models receive the reduced chemical formula, in which the atom counts are normalized to the simplest integer ratio. Thus, although all models are evaluated on the same target structures using the same downstream

Table 11: Complete CSP results on MP20 and MP52. Valid, CN, MinDist, PhysPlaus, FM, and MR@1 are percentages; RMSD is in Å.

	MP20							MP52						
Model	Valid	CN	MinDist	PhysPlaus	FM	MR@1	RMSD↓	Valid	CN	MinDist	PhysPlaus	FM	MR@1	RMSD↓
CrystaLLM	85.02	96.00	85.07	97.24	84.88	65.95	0.043	44.25	90.46	44.54	91.22	40.81	14.70	0.071
LLaMat-2-CIF	81.46	96.87	82.36	98.45	75.13	45.81	0.069	47.63	91.87	48.32	93.92	35.69	11.59	0.131
LLaMat-3-CIF	85.60	96.66	86.58	98.44	77.12	49.85	0.052	57.51	93.82	58.42	96.00	42.82	18.04	0.100
CrystalTextLLM	57.03	91.03	57.40	92.22	55.91	25.53	0.090	25.69	81.64	25.98	83.17	23.13	2.99	0.217
PLaID++-SFT	78.27	87.78	78.38	89.16	76.55	40.44	0.048	48.83	70.07	49.11	71.47	41.73	7.64	0.098
PLaID++-DPO	76.41	87.54	76.61	89.09	74.62	39.19	0.048	45.81	68.97	46.25	70.56	39.52	7.13	0.098
CrysText-SFT	36.44	87.96	36.48	87.59	36.10	5.24	0.251	24.64	86.14	24.83	85.79	22.85	2.00	0.290
CrysText-GRPO	38.17	96.68	38.18	97.10	37.83	5.23	0.258	25.99	94.47	26.10	95.14	24.17	1.75	0.285
CrysText++-SFT	74.31	93.48	74.56	94.37	74.08	46.26	0.058	38.14	85.04	38.20	85.42	37.35	9.18	0.117
CrysText++-GRPO	76.46	96.55	76.65	97.63	76.33	46.65	0.057	39.66	90.89	39.77	91.85	38.81	9.10	0.113
CrystalReasoner-SFT	72.16	95.20	72.22	96.62	71.76	32.26	0.058	41.35	88.73	41.44	90.19	39.43	5.35	0.120
CrystalReasoner-GRPO	79.73	96.19	79.84	97.70	79.36	35.29	0.055	50.46	91.35	50.65	92.98	47.82	5.87	0.104
CrystalReasoner++-SFT	71.91	94.84	72.37	96.27	71.19	32.25	0.059	41.97	89.66	42.08	91.39	40.00	5.49	0.117
CrystalReasoner++-GRPO	78.24	96.25	78.70	97.65	77.45	34.71	0.056	49.36	92.68	49.46	94.36	46.95	5.87	0.104
DiffCSP(Jiao et al., 2023)	99.30	98.47	99.61	99.61	99.30	56.92	0.060	78.95	98.24	78.96	99.65	78.95	11.75	0.097
OMatG(Höllmer et al., 2025)	98.81	98.14	98.87	99.51	98.81	68.41	0.070	79.93	94.73	79.95	94.44	79.93	15.67	0.097
CrystalReAct	88.64	95.70	88.68	97.14	87.60	45.84	0.028	65.91	91.09	65.99	92.78	63.56	9.96	0.051

metrics, the amount of composition information available at generation time differs across models. The results should therefore be understood as comparisons under a unified target set and evaluation pipeline while preserving model-specific input representations, rather than as comparisons under strictly identical input conditions.

Within this unified evaluation framework, overall validity closely follows the minimum-distance criterion. Producing a geometrically admissible crystal is therefore necessary but not sufficient for successful prediction: models with similar validity can still differ substantially in formula matching and structure prediction.

The cross-dataset comparison reveals an additional generalization challenge. MP52 is markedly more difficult than MP20, with degradation in validity, formula matching, and MR@1 and generally larger reconstruction errors. This pattern indicates that the difficulty is distributed across multiple stages rather than concentrated in a single evaluation criterion. DiffCSP and OMatG maintain a clear advantage in physical validity, particularly for larger or more challenging targets, but their advantage in MR@1 is not uniform across models and datasets. Overall, these results suggest that conditional text generation remains less reliable for exact CSP under larger structures and distribution shifts, even when some language-model-based methods achieve competitive structure-matching rates.

D.3 Complete Composition- and Symmetry-Constrained Crystal Structure Design Results

Table 12 reports the complete component-wise results for joint formula and space-group conditioning. This setting requires a generated structure to remain physically plausible while satisfying two coupled input constraints.

On MP20, formula match (FM) remains close to overall validity for most model configurations, whereas space-group match (SGM) is often substantially lower. This pattern indicates that satisfying the exact space-group constraint is generally more difficult than preserving the target composition after a valid structure has been generated. The small gap between SGM and FSGM further shows that most structures matching the requested space group also satisfy the target formula. Among the post-training-based crystal LLMs, PLaID++-SFT achieves the highest FSGM. However, the effects of RL post-training are not consistent across model families: improvements in validity or either individual constraint do not necessarily increase their joint satisfaction.

Performance decreases broadly on MP52, affecting validity, FM, SGM, and FSGM. For most post-training-based models, the gap between FM and SGM is also larger than on MP20, indicating that exact space-group control degrades more strongly than composition preservation under the shifted

Table 12: Complete composition- and symmetry-constrained CSD results on MP20 and MP52. All entries are percentages. Best values are bold and second-best values are underlined within each dataset.

	MP20							MP52						
Model	Valid	CN	MinDist	PhysPlaus	FM	SGM	FSGM	Valid	CN	MinDist	PhysPlaus	FM	SGM	FSGM
CrystaLLM	83.30	95.00	83.50	96.00	81.60	82.70	81.00	33.90	91.30	34.20	91.80	28.10	33.50	27.80
LLaMat-2-CIF	83.30	97.20	84.10	98.60	76.80	33.70	33.50	47.70	93.40	47.90	95.10	35.40	6.20	5.80
LLaMat-3-CIF	87.00	97.60	87.80	98.70	78.00	33.20	32.60	55.00	93.20	55.70	95.20	43.60	8.80	8.60
CrystalTextLLM	59.80	93.60	60.20	94.80	58.80	25.50	25.40	25.10	83.40	25.30	85.40	22.20	2.40	2.30
PLaID++-SFT	84.00	91.40	84.00	92.90	82.80	81.90	80.90	41.50	62.50	41.70	63.90	35.40	34.80	30.30
PLaID++-DPO	81.50	89.90	81.80	91.70	80.10	79.80	78.70	35.70	62.90	36.30	64.40	29.30	31.70	27.10
CrysText-SFT	77.70	95.60	78.00	96.50	77.10	70.20	69.90	32.20	85.20	32.20	86.10	30.60	18.30	17.70
CrysText-GRPO	77.90	96.70	78.00	98.20	77.70	69.90	69.70	33.00	92.70	33.00	94.00	31.20	19.00	18.40
CrysText++-SFT	76.90	93.30	77.00	94.40	76.80	61.40	61.40	32.50	84.40	32.80	85.10	31.70	14.00	14.00
CrysText++-GRPO	78.10	96.50	78.30	97.50	78.10	63.80	63.80	39.40	91.70	39.70	93.10	37.70	15.50	15.50
CrystalReasoner-SFT	71.60	95.50	71.60	96.90	71.30	61.00	60.80	34.20	89.70	34.40	91.50	31.60	20.80	19.40
CrystalReasoner-GRPO	74.60	96.50	74.60	98.00	74.30	63.70	63.40	39.90	92.40	40.10	94.10	35.10	23.20	21.30
CrystalReasoner++-SFT	71.80	95.60	72.00	96.80	70.70	61.20	60.50	34.60	91.70	34.70	93.30	31.60	22.30	20.90
CrystalReasoner++-GRPO	76.10	96.80	76.40	98.30	75.10	64.40	63.60	39.30	94.30	39.50	96.30	35.90	23.90	22.30
CrystalReAct	84.20	94.60	84.50	95.30	82.30	82.00	80.30	47.40	92.40	47.70	92.90	40.80	46.00	39.70

Table 13: Complete component-wise results for symmetry-constrained CSD. Overall SGM is computed over all target space groups; High and Low are macro-averages over the paired frequency groups, and Gap is their difference in percentage points. Best values are bold and second-best values are underlined; Gap is not ranked.

Model	Valid	CN	MinDist	PhysPlaus	Overall SGM	High	Low	Gap (pp)
LLaMat-2-CIF	85.50	96.43	89.09	98.21	6.77	12.74	0.80	11.94
LLaMat-3-CIF	90.79	96.96	93.46	99.56	5.59	10.03	1.14	8.89
CrystalTextLLM	44.23	86.29	45.93	88.80	9.49	17.89	1.09	16.80
PLaID++-SFT	47.93	66.51	48.44	67.11	23.43	41.11	5.74	35.37
PLaID++-DPO	51.50	69.20	51.73	69.46	27.83	49.09	6.57	42.51
CrysText-SFT	52.67	84.96	55.47	86.13	21.14	38.66	3.63	35.03
CrysText-GRPO	56.96	91.46	59.64	93.23	22.56	41.09	4.03	37.06
CrysText++-SFT	63.79	93.11	64.59	93.19	9.90	19.20	0.60	18.60
CrysText++-GRPO	65.57	96.33	66.44	96.60	10.53	20.11	0.94	19.17
CrystalReasoner-SFT	58.36	85.26	65.86	91.00	19.89	36.23	3.54	32.69
CrystalReasoner-GRPO	67.01	82.63	77.20	90.13	23.94	42.60	5.29	37.31
CrystalReasoner++-SFT	44.34	90.29	45.73	92.99	18.06	33.83	2.29	31.54
CrystalReasoner++-GRPO	60.47	94.36	61.56	95.90	25.43	48.29	2.57	45.71
CrystalReAct	75.76	94.69	76.10	95.06	60.27	78.71	41.83	36.89

target distribution. These mixed results demonstrate that neither overall validity nor either individual matching metric reliably predicts joint controllability, motivating the direct evaluation of FSGM.

D.4 Complete Symmetry-Constrained Crystal Structure Design Results

Table 13 reports the aggregate validity components and frequency-stratified SGM results. Tables 14 and 15 further resolve these averages into the individual space groups used for each crystal-system pair. It should be noted that CrystaLLM is not included because its native prompt does not support conditioning on a target space group without a specified chemical formula, as required by the symmetry-constrained CSD setting.

Across the evaluated configurations, SGM is consistently higher for the higher-frequency space groups than for their lower-frequency counterparts. This High–Low gap indicates a strong association between space-group frequency in the training data and exact symmetry controllability. However, the gap should not be interpreted solely as a consequence of lower structural validity. Validity

Table 14: Per-space-group Valid and SGM results for the higher-frequency members of the seven crystal-system pairs. All entries are percentages. Best values are bold and second-best values are underlined within each column.

	SG 2		SG 12		SG 62		SG 139		SG 166		SG 189		SG 225	
Model	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM
LLaMat-2-CIF	80.60	0.80	80.20	1.60	80.00	7.40	87.20	5.60	88.80	3.60	89.20	0.20	97.40	71.80
LLaMat-3-CIF	85.20	2.40	87.40	1.80	89.00	2.00	93.40	3.20	94.80	10.40	92.80	0.20	96.60	52.00
CrystalTextLLM	26.40	0.40	25.20	2.60	39.80	8.60	54.20	8.40	53.40	15.00	33.60	0.40	98.60	89.80
PLaID++-SFT	30.20	12.00	46.60	40.80	47.60	45.40	72.00	60.20	63.20	39.60	50.20	33.20	78.00	56.60
PLaID++-DPO	32.00	9.00	48.20	44.00	59.80	59.40	71.60	64.20	67.20	47.60	51.80	42.60	83.00	76.80
CrysText-SFT	32.40	25.20	20.40	16.40	57.60	18.20	56.20	46.40	69.80	60.00	67.00	47.20	75.80	57.20
CrysText-GRPO	33.80	25.40	22.20	19.40	64.00	17.40	69.20	53.80	75.80	65.80	72.60	50.60	70.20	55.20
CrysText++-SFT	50.00	12.20	63.00	20.20	47.00	8.60	63.80	27.40	66.00	28.40	74.20	2.00	78.40	35.60
CrysText++-GRPO	54.60	10.40	64.40	22.20	50.80	7.80	65.60	31.20	67.20	26.60	73.40	3.40	79.20	39.20
CrystalReasoner-SFT	53.80	42.60	53.00	40.00	88.40	70.60	53.00	19.60	51.20	23.60	54.60	49.80	63.80	7.40
CrystalReasoner-GRPO	80.80	67.80	66.60	50.60	84.60	72.40	45.00	17.80	45.00	11.80	74.00	72.60	62.60	5.20
CrystalReasoner++-SFT	37.60	27.00	43.60	36.20	42.40	33.60	54.60	31.40	65.40	44.40	51.00	23.80	65.00	40.40
CrystalReasoner++-GRPO	52.40	36.40	61.60	48.20	64.20	56.00	67.60	39.80	85.20	54.60	73.60	39.40	83.20	63.60
CrystalReAct	82.20	73.40	69.80	69.40	28.00	28.00	97.40	97.40	93.20	93.00	93.60	90.00	99.80	99.80

Table 15: Per-space-group Valid and SGM results for the lower-frequency members of the seven crystal-system pairs. All entries are percentages. Best values are bold and second-best values are underlined within each column.

	SG 1		SG 3		SG 20		SG 137		SG 150		SG 180		SG 217	
Model	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM	Valid	SGM
LLaMat-2-CIF	83.20	11.40	84.80	0.00	77.20	0.00	86.60	0.20	86.60	0.00	81.60	0.00	93.60	0.00
LLaMat-3-CIF	85.40	17.40	84.20	0.00	86.00	0.20	90.80	0.00	96.00	0.00	94.60	0.00	94.80	0.00
CrystalTextLLM	31.60	7.20	34.20	0.00	37.40	0.00	49.40	0.20	60.00	0.00	24.40	0.00	51.00	0.20
PLaID++-SFT	34.40	1.60	39.60	0.80	52.20	9.20	29.40	1.60	37.00	3.20	41.80	23.40	48.80	0.40
PLaID++-DPO	44.60	2.20	39.00	0.40	55.20	11.00	30.80	0.60	35.40	5.20	46.00	26.00	56.40	0.60
CrysText-SFT	34.20	22.00	37.20	1.40	40.20	0.00	69.00	0.00	87.60	0.20	38.80	0.80	51.20	1.00
CrysText-GRPO	37.00	26.20	41.40	0.60	45.60	0.20	71.40	0.00	88.60	0.40	52.80	0.40	52.80	0.40
CrysText++-SFT	53.40	3.40	44.60	0.00	70.80	0.00	59.00	0.20	69.00	0.20	77.20	0.00	76.60	0.40
CrysText++-GRPO	51.20	6.00	51.40	0.20	71.80	0.00	64.00	0.00	72.20	0.00	76.00	0.00	76.20	0.40
CrystalReasoner-SFT	38.40	15.00	54.20	2.60	74.00	3.40	79.40	0.00	75.40	0.80	70.40	3.00	7.40	0.00
CrystalReasoner-GRPO	66.80	23.40	75.80	2.40	82.80	8.20	85.60	0.00	89.20	0.00	73.60	3.00	5.80	0.00
CrystalReasoner++-SFT	37.80	11.80	40.00	0.80	36.00	0.80	58.20	0.40	39.20	0.60	32.60	1.00	17.40	0.60
CrystalReasoner++-GRPO	48.20	10.00	54.80	1.60	53.60	2.60	78.60	1.00	53.20	0.00	41.40	2.40	29.00	0.40
CrystalReAct	95.60	46.80	82.00	62.20	63.00	60.20	74.20	44.00	73.00	16.20	36.20	35.00	72.60	28.40

and exact space-group agreement are distinct outcomes: a model may generate physically plausible structures while still failing to realize the requested space group. This distinction is particularly clear for the CrysText++ variants, which achieve relatively high validity but substantially lower SGM than PLaID++-DPO. Thus, improving structural validity does not necessarily lead to stronger symmetry control.

The per-group results for the higher-frequency targets show that aggregate performance also conceals substantial variation across crystal systems. PLaID++ performs consistently across several space groups, while CrystalReasoner-GRPO performs particularly well for selected targets. In contrast, CrystalTextLLM achieves a very high SGM for SG 225 but performs much worse on most of the other higher-frequency groups. Therefore, strong performance on one frequent or structurally regular space group does not establish broad symmetry controllability, and model rankings can change substantially with the requested space group.

The lower-frequency targets are more difficult in terms of SGM for all evaluated models, including cases in which validity remains relatively high. For example, several models generate valid struc-

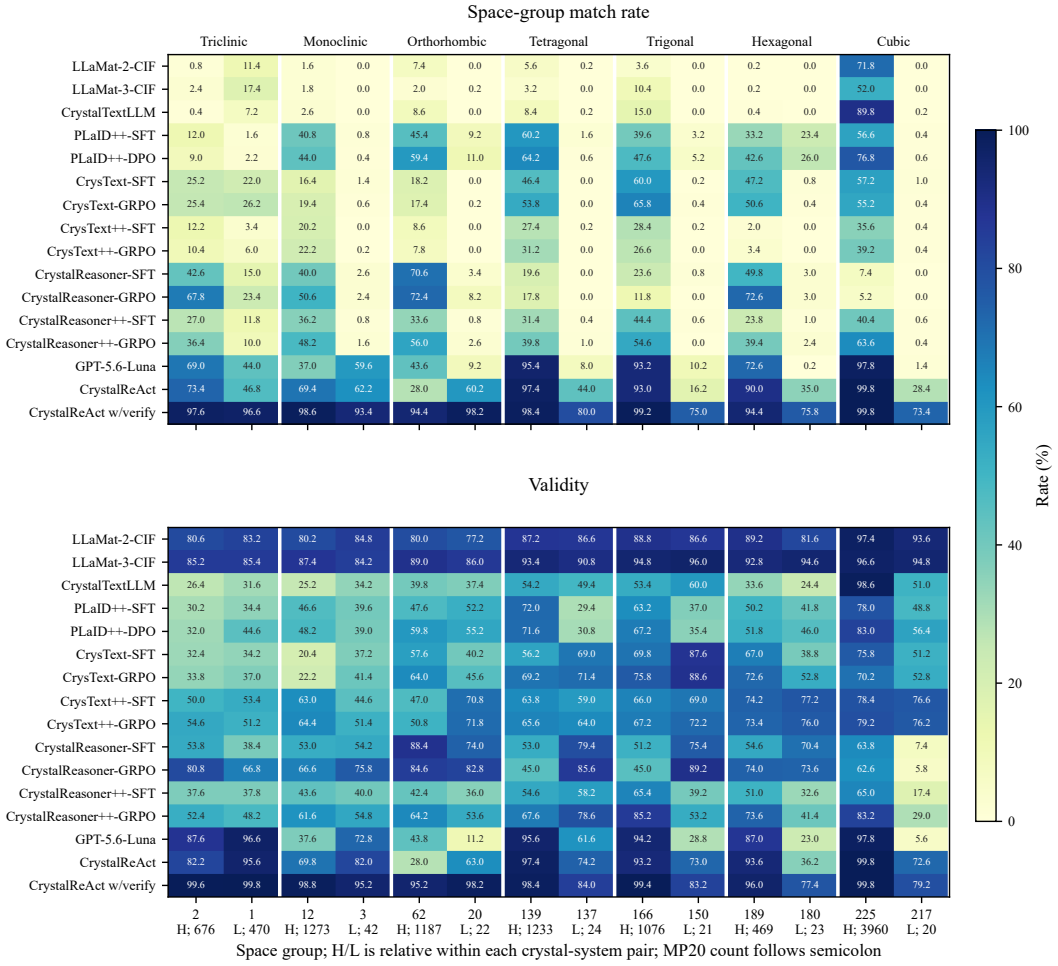


Figure 7: Per-space-group performance for symmetry-constrained CSD. The upper panel reports SGM and the lower panel reports Valid for the same 14 targets. Space groups are arranged in crystal-system pairs, with the higher- and lower-frequency members marked H and L.

tures for SGs 137, 150, or 217 but rarely recover the specified space group. These cases indicate that the dominant failure is conditional symmetry control rather than the inability to generate a physically admissible structure. Although some models achieve higher SGM on individual low-frequency groups, these improvements are not consistent across different crystal systems.

Figure 7 summarizes the two levels of variation across the 14 evaluated space groups. The paired high- and low-frequency comparison reveals a robust frequency-related trend, whereas the individual space-group results show that frequency alone does not determine performance. Some higher-frequency targets remain difficult for particular models, while certain lower-frequency targets can still be recovered under specific model configurations. The effect of frequency therefore varies across crystal systems and model–group pairs rather than acting as a uniform scaling factor.

These results also clarify how SGM should be interpreted. Because SGM requires both a valid generated structure and the correct target space group, a low SGM can arise either from invalid structure generation, incorrect symmetry control, or both. Conversely, a high aggregate SGM may reflect strong performance on a subset of frequently represented space groups without implying uniform generalization across the symmetry space. Overall SGM should therefore be examined jointly with validity and per-space-group results to distinguish failures of physical generation from failures of exact symmetry realization.

D.5 Additional Released-Checkpoint Results

Tables 16, 17, and 18 evaluate publicly released checkpoints of PLaID++, CrysText, and CrystalReasoner in terms of DNG quality, distributional and diversity characteristics, and conditional-generation performance, respectively. All checkpoints are evaluated using the same task instances, decoding settings, and evaluation pipeline as the main experiments. We report these results separately because the released checkpoints follow the original authors' training configurations rather than our unified reimplementation protocol.

DNG quality. Table 16 reports validity, uniqueness, novelty, energy, and discovery metrics at temperature 0.7. The released checkpoints exhibit different strengths. CrystalReasoner-GRPO achieves the highest validity (96.68%) and lowest mean relaxation RMSD (0.27 Å), but its low uniqueness (20.81%) and metastability rate (13.07%) limit MSUN to 0.25%. CrysText preserves much higher uniqueness and novelty, both exceeding 89%, but its validity remains below 47% and its metastability rates are below 7%. PLaID++-DPO achieves the lowest mean formation and hull energies and the highest metastability rate (58.76%), yielding the highest MSUN rate of 2.16%, compared with 0.25%–0.81% for the other checkpoints. Nevertheless, stable discovery remains limited across all models: PLaID++-DPO produces only one SUN structure, and the remaining checkpoints produce none. These results show that high validity or diversity alone does not ensure successful material discovery.

Table 16: Released-checkpoint DNG quality results at temperature 0.7. Rates are percentages, parenthesized values are counts, energies are in eV/atom, and RMSD is in Å. Best values are bold and second-best values are underlined.

Model	Valid				Unique↑	Novel↑	Energy-based						Stability		Metastability	
	Overall↑	CN↑	MinDist↑	PhysPlaus↑			E_{form} (Std)↓		E_{hull} (Std)↓		RMSD (Std)↓		Stable↑	SUN↑	Meta↑	MSUN↑
PLaID++-DPO	94.56	99.96	94.60	**100.00**	47.12	26.44	**-0.39** ±	0.74	**0.32** ±	0.64	0.45 ±	0.52	**0.72** (17)	**0.04** (1)	**58.76** (1389)	**2.16** (51)
CrysText-SFT	46.12	98.68	46.52	99.24	**95.32**	**90.11**	0.68 ±	3.63	1.52 ±	3.55	1.35 ±	0.80	0.26 (3)	0.00 (0)	6.50 (75)	0.43 (5)
CrysText-GRPO	44.68	98.96	45.20	99.76	95.08	89.35	0.57 ±	1.32	1.42 ±	1.13	1.35 ±	0.80	0.09 (1)	0.00 (0)	6.98 (78)	0.81 (9)
CrystalReasoner-SFT	93.08	**100.00**	93.08	99.92	30.47	43.23	0.62 ±	0.97	0.58 ±	0.97	0.39 ±	0.58	0.00 (0)	0.00 (0)	11.26 (262)	0.47 (11)
CrystalReasoner-GRPO	**96.68**	99.96	**96.72**	**100.00**	20.81	35.50	0.50 ±	0.65	0.45 ±	0.64	**0.27** ±	0.45	0.00 (0)	0.00 (0)	13.07 (316)	0.25 (6)

Distribution and diversity. Table 17 complements the quality metrics with distributional distances, diversity measures, and HHI diagnostics. No checkpoint performs best across all measures. CrysText achieves the lowest JS divergence and the highest elemental, size, and site diversity, consistent with its high uniqueness and novelty in Table 16. However, these advantages coexist with low physical validity and metastability. PLaID++-DPO achieves the lowest FID (3.55), despite having the highest MMD (0.056), the lowest space-group and size diversity, and the highest values of both HHI measures. CrystalReasoner-SFT achieves the lowest MMD (0.022) and highest space-group diversity (0.21), but its GRPO checkpoint worsens all three distributional distances and reduces space-group, size, and site diversity. Thus, the validity improvement from CrystalReasoner-SFT to GRPO is accompanied by reduced coverage of structural variation. The different rankings across these diagnostics further indicate that generation quality cannot be characterized by a single distributional or diversity metric.

Table 17: Distributional and diversity diagnostics for released checkpoints at temperature 0.7. Downward and upward arrows indicate the preferred direction. Best values are bold and second-best values are underlined.

Model	Distribution			Diversity				HHI	
	JS↓	MMD↓	FID↓	ElemDiv↑	SGDiv↑	SizeDiv↑	SiteDiv↑	Prod↓	Res↓
PLaID++-DPO	0.54	0.056	**3.55**	0.48	0.14	0.18	3.02	5.23	3.42
CrysText-SFT	**0.45**	0.034	52.92	**0.65**	0.20	**0.26**	10.43	3.08	2.58
CrysText-GRPO	0.46	0.034	4.56	**0.65**	0.18	0.25	**11.27**	2.94	2.47
CrystalReasoner-SFT	0.52	**0.022**	6.64	0.27	**0.21**	0.21	6.24	2.71	**1.70**
CrystalReasoner-GRPO	0.55	0.028	7.01	0.27	0.18	0.20	5.00	**2.58**	1.71

Conditional generation. Table 18 reports CSP matching, joint formula-and-symmetry satisfaction, and symmetry-only matching. CrystalReasoner-GRPO achieves the strongest results on both datasets for CSP and joint conditioning, as well as the highest overall SGM. Relative to its SFT checkpoint, GRPO increases CSP MR@1 from 31.34% to 34.04% on MP20 and from 9.42% to 10.35% on MP52, while improving FSGM from 59.60% to 62.50% and from 25.00% to 30.10%, respectively. PLaID++-DPO outperforms both CrysText checkpoints on CSP and joint conditioning, whereas its overall symmetry-only performance falls between CrysText-SFT and CrysText-GRPO. CrysText-GRPO improves joint conditioning and overall SGM over SFT, but its CSP changes are small and differ in direction across datasets. All checkpoints show substantial MP20-to-MP52 declines in MR@1 and FSGM. Symmetry-only results also reveal a pronounced frequency gap: high-frequency SGM ranges from 37.91% to 62.77%, whereas low-frequency SGM remains between 1.20% and 4.17%. Notably, CrystalReasoner-GRPO improves high-frequency SGM over SFT while slightly reducing low-frequency SGM, showing that aggregate improvements do not necessarily extend to underrepresented symmetry targets.

Table 18: Conditional-generation results obtained from publicly released checkpoints. CSP is summarized by MR@1, composition- and symmetry-constrained CSD by FSGM, and symmetry-constrained CSD by SGM.

	CSP MR@1		Formula+SG FSGM		SG-only SGM		
Model	MP20	MP52	MP20	MP52	Overall	High	Low
PLaID++-DPO	22.88	6.09	59.80	24.20	20.90	40.11	1.69
CrysText-SFT	6.16	3.71	50.50	19.40	20.60	37.91	3.29
CrysText-GRPO	6.37	3.59	51.80	22.00	21.63	39.09	4.17
CrystalReasoner-SFT	31.34	9.42	59.60	25.00	29.77	58.26	1.29
CrystalReasoner-GRPO	34.04	10.35	62.50	30.10	31.99	62.77	1.20

E CrystalReAct

E.1 CrystalReAct: Agentic Search for Crystal Generation

CrystalReAct implements an agentic search paradigm for crystal structure generation. During inference, the model can obtain crystallographic information relevant to the current structure-construction state and use the returned content in subsequent reasoning. The model therefore controls both the choice of evidence and the point at which it generate a complete Wyckoff-formatted structure.

The reasoning backend is GPT-5.6-Luna. Given the task condition and the current generation state, the model performs exactly one protocol action at each step: it either requests one tool call or outputs one complete Wyckoff-formatted candidate. After a tool call, the runtime executes the requested action, returns the corresponding structured result, and resumes generation with the result appended to the current inference context. The model is responsible for selecting the space group when it is not specified, decomposing the target composition into Wyckoff multiplicities, assigning elements to Wyckoff sites, and determining when to generate a candidate.

The generated structure is constructed from the task condition together with crystallographic information selected during inference. The model can use space-group information, element-level chemical information, and structural references when assigning sites and coordinates. DNG additionally supports sampled prototype conditions that provide an anonymous formula, a space group, and a Wyckoff prototype as a structural reference.

E.2 Crystallographic Information

CrystalReAct can access three complementary types of crystallographic information: reference structures, space-group information, and element-level chemical information. These sources provide structural, symmetry, and chemical context while leaving the final site assignment and coordinate construction to the reasoning backend.

Structural References. The structure case collection is built from the 27,136 structures in the MP20 training split. Each entry contains its Materials Project identifier, chemical and anonymous formulas, space group, lattice parameters, Wyckoff site assignments, complete Wyckoff-based structure representation, and available material properties, including formation energy, energy above the convex hull, band gap, stability, and metallicity. We additionally retain structure-level descriptions generated by Robocrystallographer (Ganose & Jain, 2019), which summarize local coordination environments and other crystallographic characteristics. From each structure, we derive anonymous formulas, space-group labels, and Wyckoff prototypes for retrieval.

Space group Information. The space-group information covers all 230 crystallographic space groups. Each record contains the space group number and Hermann–Mauguin symbol, crystal system, lattice constraints, and the complete set of Wyckoff positions. Every Wyckoff entry specifies its multiplicity, letter, site symmetry, and canonical coordinate expression. These records allow the model to inspect the admissible lattice geometry and site multiplicities of a candidate space group before constructing a structure.

Element-Level Chemical Information. The element-level chemical information extracted from PYMATGEN (Ong et al., 2013) contains common and ICSD-derived oxidation states, Pauling electronegativity, atomic radii, ionic radii, and Shannon radii together with their associated valence and coordination information. The model can use this information to reason about charge balance, chemically plausible substitutions, relative atomic sizes, and element-to-site assignments.

E.3 CRYSTAL-SPECIFIC TOOLS

CrystalReAct exposes crystallographic information through domain-specific tools. This section describes the knowledge provided by each tool and gives representative response examples. Each tool returns structured results corresponding to a specific crystallographic knowledge category.

E.3.1 STRUCTURAL ANALOGUES

The `analogue_cases` response provides a structural analogue with its Wyckoff-based structure description, crystallographic fields, Robocrystallographer description, and available material properties.


```
{
  "tool": "analogue_cases",
  "results": [
    {
      "structure": "
      BaTiO3
      Spacegroup: R3m
      abc: 5.70 5.70 13.98
      angles: 90.0 90.0 120.0
      Sites (15)
      Ba 0.000000 0.000000 0.000000 3a
      Ti 0.000000 0.000000 0.012000 3a
      O 0.500000 -0.500000 0.250000 9b",
      "robocrystallographer_description": "BaTiO3 is a rhombohedrally distorted perovskite
structure that crystallizes in the polar trigonal R3m space group. Ba2+ is coordinated by
twelve O2- atoms in a distorted BaO12 cuboctahedral environment, with the surrounding
oxygen framework displaced relative to the centrosymmetric cubic perovskite arrangement.
Ti4+ is coordinated by six O2- atoms to form distorted TiO6 octahedra, and the Ti
displacement along the polar three-fold axis produces inequivalent Ti-O bond lengths. The
TiO6 octahedra are connected through corner sharing along the three-fold direction, while
the rhombohedral distortion is associated with a cooperative polar displacement of the
cation and anion sublattices. O2- atoms coordinate both Ba2+ and Ti4+ and occupy the 9b
Wyckoff orbit required by the R3m symmetry.",
      "properties": {
        "formation_energy_per_atom_eV": -3.492237,
        "energy_above_hull_eV": 0.000041,
        "band_gap_eV": 2.508700
      }
    }
  ]
}
```

E.3.2 SPACE-GROUP CARD

The `spacegroup_card` tool returns the crystallographic information associated with one space group, including its number, Hermann–Mauguin symbol, crystal system, lattice constraints, complete Wyckoff table, multiplicities, site-symmetry information, and representative coordinates.


```
{
  "tool": "spacegroup_card",
  "results": [
    {
      "space_group_number": 160,
      "space_group_symbol": "R3m",
      "crystal_system": "trigonal",
      "lattice_constraints": "a=b; alpha=beta=90; gamma=120",
      "wyckoff_positions": "
      3a | 3m | (0.00, 0.00, z)
      9b | .m | (x, -x, z)
      18c | 1 | (x, y, z)"
    }
  ]
}
```


E.3.3 ELEMENT FACTS

The `element_facts` tool returns element symbols, likely oxidation states, Pauling electronegativities, atomic radii, ionic radii, and Shannon radii for the requested elements.


```
{
  "tool": "element_facts",
  "results": [
    {
      "element": "O",
      "likely_oxidation_states": [-2],
      "electronegativity_pauling": 3.44,
      "atomic_radius_A": 0.6,
      "ionic_radii_A": {
        "-2": {
          "II": 1.35,
          "III": 1.36,
          "IV": 1.38,
          "VI": 1.4,
          "VIII": 1.42
        }
      }
    }
  ]
}
```


E.3.4 PROTOTYPE CONDITION

The `sample_prototype_condition` tool is available for de novo generation. It returns an anonymous formula, a space group number, and a Wyckoff prototype. The sampled condition provides a structural starting point that the model can adopt, modify, discard, or use to guide a subsequent structural-reference request.


```
{
  "tool": "sample_prototype_condition",
  "results": [
    {
      "anonymous_formula": "ABC2D6",
      "space_group_number": 59,
      "space_group_symbol": "Pmmn",
      "wyckoff_prototype": "59_2a2b4e4e4f4f"
    }
  ]
}
```

Table 19: Effect of reasoning mode across different LLM backbones.

		CSP								CSD								SG				
		MP20				MP52				MP20				MP52								
Model	Variant	Valid	FM	MR@1	RMSD	Valid	FM	MR@1	RMSD	Valid	FM	SGM	Joint	Valid	FM	SGM	Joint	Valid	SGM	High	Low	Gap
Qwen3-8B	No Thinking	67.75	54.78	21.09	0.034	35.14	23.69	3.54	0.061	87.40	82.10	86.80	82.00	27.50	24.30	27.50	24.30	3.64	0.20	0.40	0.00	0.40
Qwen3-8B	Thinking	61.61	54.57	30.82	0.028	36.71	27.26	6.10	0.039	38.90	34.20	37.70	33.60	14.60	11.80	13.80	11.30	9.36	6.39	11.91	0.86	11.06
Qwen3-32B	No Thinking	47.11	42.53	27.61	0.026	24.71	17.74	5.86	0.046	67.30	61.20	66.10	60.80	25.50	18.20	23.60	17.70	21.37	10.07	18.03	2.11	15.91
Qwen3-32B	Thinking	45.17	42.64	22.16	0.031	20.88	18.09	3.78	0.048	32.40	30.40	28.10	26.40	9.90	7.50	8.00	6.40	40.03	21.06	33.69	8.43	25.26
Qwen3.8-27B	No Thinking	15.43	13.19	7.12	0.033	8.76	5.34	1.32	0.053	65.90	56.30	64.30	55.50	26.30	14.20	23.70	13.80	22.73	8.27	12.37	4.17	8.20
Qwen3.8-27B	Thinking	62.72	61.31	24.85	0.028	45.86	42.74	5.29	0.048	50.00	47.00	45.90	43.60	24.00	20.30	20.20	16.90	44.39	25.79	34.54	17.03	17.51

Table 20: Effect of the verification tool in CrystalReAct.

		CSP								CSD								SG				
		MP20				MP52				MP20				MP52								
Model	Variant	Valid	FM	MR@1	RMSD	Valid	FM	MR@1	RMSD	Valid	FM	SGM	Joint	Valid	FM	SGM	Joint	Valid	SGM	High	Low	Gap
Qwen3-8B	w/o verify	67.75	54.78	21.09	0.034	35.14	23.69	3.54	0.061	87.40	82.10	86.80	82.00	27.50	24.30	27.50	24.30	3.64	0.20	0.40	0.00	0.40
Qwen3-8B	w/ verify	63.46	48.98	12.22	0.041	30.85	19.70	1.94	0.071	85.70	82.30	85.10	82.30	26.90	23.50	26.50	23.50	9.87	3.59	8.23	0.11	8.12
Qwen3-32B	w/o verify	47.11	42.53	27.61	0.026	24.71	17.74	5.86	0.046	67.30	61.20	66.10	60.80	25.50	18.20	23.60	17.70	21.37	10.07	18.03	2.11	15.91
Qwen3-32B	w/ verify	54.40	48.86	30.78	0.026	28.64	20.34	6.56	0.052	66.00	59.20	64.20	58.50	25.30	16.60	20.90	16.30	41.76	12.07	21.54	2.60	18.94
Qwen3.8-27B	w/o verify	15.43	13.19	7.12	0.033	8.76	5.34	1.32	0.053	65.90	56.30	64.30	55.50	26.30	14.20	23.70	13.80	22.73	8.27	12.37	4.17	8.20
Qwen3.8-27B	w/ verify	13.42	12.22	6.61	0.04	8.27	6.03	1.67	0.065	53.20	48.50	52.60	48.40	25.00	15.60	23.30	15.30	4.27	2.86	2.37	3.20	-0.83
GPT-5.6-Luna	w/o verify	88.64	87.60	45.84	0.028	65.91	63.56	9.96	0.051	84.20	82.30	82.00	80.30	47.40	40.80	46.00	39.70	75.76	60.27	78.71	41.83	36.89
GPT-5.6-Luna	w/ verify	92.93	92.64	46.88	0.027	82.36	81.42	10.18	0.054	94.10	93.80	93.90	93.60	83.10	80.10	82.60	79.70	93.16	91.06	97.83	84.77	13.06

E.4 Implementation Details

CrystalReAct uses GPT-5.6-Luna as its reasoning model. The reported configuration uses temperature 0.7 and top-$p = 0.9$. Each instance permits at most fifteen inference steps and at most five model-initiated information requests. The model may output a complete candidate before either budget is exhausted.

E.5 Additional Experiments on Smaller Models

E.5.1 Experimental Setup

To further investigate the robustness of CrystalReAct beyond the primary GPT-5.6-Luna evaluation, we conduct additional experiments with smaller open-source reasoning models, including Qwen3-8B, Qwen3-32B-AWQ, and Qwen3.8-27B-AWQ-INT4. These experiments examine how reasoning and verification mechanisms affect crystal-generation performance under resource-constrained settings. We report results on CSP, composition- and symmetry-constrained CSD, and symmetry-constrained generation, using the same evaluation protocols as the main experiments.

E.5.2 Thinking vs. No Thinking

Table 19 compares reasoning-enabled and non-reasoning variants of smaller models. The effect of explicit reasoning varies across models and tasks. For Qwen3.8-27B-AWQ-INT4, enabling thinking substantially improves CSP performance, increasing both validity and formula matching on MP20 and MP52, while also improving symmetry controllability. However, reasoning does not consistently improve all settings: Qwen3-8B and Qwen3-32B show mixed changes across CSD and symmetry-constrained generation.

These results suggest that additional reasoning tokens can improve structural construction for some models, but their benefits depend on the underlying backbone and task requirements. In particular, reasoning alone does not guarantee better crystallographic constraint satisfaction, highlighting the complementary role of external crystallographic information.

E.5.3 No Verify Tool vs. With Verify Tool

Table 20 evaluates the effect of adding the verification tool. For GPT-5.6-Luna, verification consistently improves physical validity and crystallographic constraint satisfaction across CSP, CSD, and symmetry-constrained generation. For example, the verified variant substantially improves MP52

CSP validity and CSD joint satisfaction, indicating that explicit feedback can help revise structurally inconsistent candidates.

For smaller Qwen models, the effect of verification is more dependent on the backbone. While verification improves some metrics, such as symmetry controllability for Qwen3-32B, it may also introduce additional revision difficulty and reduce performance in other settings. These results indicate that verification is most effective when the reasoning model can appropriately interpret structural feedback and incorporate it into subsequent generation.

E.6 Prompt Design

The CrystalReAct prompt specifies the current generation task, the crystallographic information available to the model, the one-action-per-response interaction protocol, and the required crystal structure string format. The task-specific prompt is combined with the common generation instructions and the corresponding information-tool instructions.

E.6.1 General instruction.

The common generation instruction is:

```
You are a crystal-structure generation assistant with access to typed crystallographic
information tools. In every response, perform exactly one action: either request one
information item or output one complete crystal structure. Never place multiple
information requests in one response, and never combine an information request with a
crystal structure. To request information, write only <|begin_search_query|>{one JSON
object}<|end_search_query|>.

You may initiate at most <MAX_SEARCH_LIMIT> information requests, including malformed or
empty-result requests, and may stop earlier.

You choose the space group when it is not task-conditioned. You also choose the Wyckoff
multiplicity decomposition and element-to-orbit assignment. Use the returned
crystallographic information when making these decisions, and ensure that the chosen
multiplicities reproduce the requested composition.

When finished, always construct one complete Wyckoff-based crystal description. Do not
stop at a partial structure or a structural discussion. Use the following schema,
replacing every braced field:
<CIF>
{cell_formula}
Spacegroup: {Hermann-Mauguin symbol or number}
abc: {a} {b} {c}
angles: {alpha} {beta} {gamma}
Sites ({expanded_site_count})
{Element} {x} {y} {z} {WyckoffLabel}
</CIF>

The first line is the formula alone. Colons after Spacegroup, abc and angles are
mandatory. Prefer the Hermann-Mauguin space-group symbol returned by the space-group
information tool; a space-group number is also accepted. Each site row has exactly five
whitespace-separated fields: Element x y z WyckoffLabel. Give one row per asymmetric site.
N is the sum of the Wyckoff multiplicities, not the number of rows. Every Wyckoff label
must exist in the stated space group with the corresponding multiplicity. Do not occupy a
fixed Wyckoff orbit twice, and choose coordinates so that different site rows expand to
disjoint symmetry orbits. Always include the closing </CIF> tag.
```

E.6.2 Task-specific prompts.

For DNG, the task-specific instruction is:

```
This is unconditional de novo generation (DNG). Propose one plausible inorganic crystal
hypothesis. Choose elements through family-level chemical compatibility and charge balance
instead of starting from a remembered compound; revise the assignment if your first
formula is overly familiar. You also choose stoichiometry, symmetry, lattice, and sites.
Consider charge balance, oxidation states, coordination, radii, Wyckoff multiplicities,
and symmetry-expanded periodic contacts when constructing the candidate. No specific space
group or chemical family is predetermined. Choose them such that the resulting structure
is chemically plausible and has a valid Wyckoff-position assignment. A sampled anonymous
prototype is optional soft guidance: adopt, adapt, discard, or resample it. If it has more
```

```
than six orbit entries or unsupported free coordinates, prefer resampling or a structural
analogue over arbitrary dense packing.
```

For formula-conditioned CSP, the task-specific instruction is:

```
This is formula-conditioned CSP. Generate a crystal structure for the requested chemical
formula. Infer the candidate space group and element-to-orbit assignments yourself. You
may request space-group information, element-level chemical information, or structural
references before constructing the candidate.
```

For formula-and-space-group-conditioned CSD, the task-specific instruction is:

```
This is formula-and-space-group-conditioned CSD. Generate a crystal structure for the
requested chemical formula and space group. You may request space-group information,
element-level chemical information, or structural references. Structural references and
space-group information should remain consistent with the requested space group.
```

For space-group-conditioned generation, the task-specific instruction is:

```
This is space-group-conditioned generation. Generate a chemically plausible crystal
structure in the requested space group. Choose the formula and element-to-orbit assignment
yourself. You may request space-group information, element-level chemical information, or
structural references before constructing the candidate.
```

E.6.3 TOOL INSTRUCTIONS.

For conditional generation, the available information tools are:

```
A search action is a strict JSON object between the search tags. Choose exactly one of
these three typed shapes; angle-bracketed values below describe field types and are not
literal values:
- spacegroup_card: {"tool":"spacegroup_card","space_group":<integer>} returns space-group
information, including lattice constraints and the complete Wyckoff table.
- element_facts: {"tool":"element_facts","elements":[<element-symbol strings>]} returns
element-level chemical information, including likely oxidation states, electronegativity,
and radii.
- analogue_cases: {"tool":"analogue_cases", optional "formula":<chemical formula string>,
optional "anonymous_formula":<anonymous stoichiometry string>, optional
"wyckoff_prototype":<Wyckoff prototype string>, optional "space_group":<integer>, optional
"next_batch":<boolean>} returns up to five structural references. The request should
contain the composition or structural preference relevant to the current generation state.
Each action returns the corresponding crystallographic information in a compact structured
form.
```

For DNG, the available information tools are:

```
DNG exposes these four strict JSON search actions (angle brackets denote types):
- sample_prototype_condition: {"tool":"sample_prototype_condition"} returns one anonymous
formula, space group, and Wyckoff prototype.
- spacegroup_card: {"tool":"spacegroup_card","space_group":<integer>} returns space-group
information, including lattice constraints and the Wyckoff table.
- element_facts: {"tool":"element_facts","elements":[<element-symbol strings>]} returns
element-level chemical information, including likely oxidation states, electronegativity,
and radii.
- analogue_cases: {"tool":"analogue_cases", optional "formula":<chemical formula string>,
optional "anonymous_formula":<anonymous stoichiometry string>, optional
"wyckoff_prototype":<Wyckoff prototype string>, optional "space_group":<integer>, optional
"next_batch":<boolean>} returns up to five structural references.
Each action costs one attempt. Emit one action, then wait for the corresponding
information before issuing another action or producing the final crystal structure.
```

At runtime, this instruction is concatenated with the task-specific user prompt:

```
Crystal generation task:
{question}
```