PLAID++: A PREFERENCE ALIGNED LANGUAGE MODEL FOR TARGETED INORGANIC MATERIALS DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning from Verifiable Rewards (RLVR) has emerged as a promising approach to improve correctness in LLMs, however, in many scientific problems, the objective is not necessarily to produce *the* correct answer, but instead to produce a diverse array of candidates which satisfy a set of constraints. We study this challenge in the context of materials generation. To this end, we introduce PLaID++, an LLM post-trained for stable and property-guided crystal generation. We find that performance hinges on our crystallographic representation and reward formulation. First, we introduce a compact, symmetry-informed Wyckoff text representation which improves computational efficiency and encourages generalization from physical priors. Second, we demonstrate that temperature scaling acts as an entropy regularizer which counteracts mode collapse and encourages exploration. By encoding symmetry constraints directly into text and guiding model outputs towards desirable chemical space, PLaID++ generates structures that are thermodynamically stable, unique, and novel at a $\sim$50% greater rate than prior methods and conditionally generates structures with desired space group properties. Our work demonstrates the potential of adapting post-training techniques from natural language processing to materials design, paving the way for targeted and efficient discovery of novel materials.

1 INTRODUCTION

The discovery of new solid-state materials is the foundation of many transformative technologies including solar cells (Green et al., 2014), batteries (Zhao et al., 2020), and carbon capture (Sriram et al., 2024a). However, the search for new materials is constrained by the immense scale of chemical space—previous explorations have only uncovered a fraction of the total number of potential stable inorganic compounds (Davies et al., 2016). Generative models offer a promising avenue for accelerating technological breakthroughs by efficiently discovering novel and stable structures in unexplored regions of chemical space.

Previous work has applied variational autoencoders (Xie et al., 2021) and denoising models (Zeni et al., 2025; Jiao et al., 2023; Miller et al., 2024) to generate stable and novel structures. However, these works either do not explicitly encode crystallographic symmetry or use computationally inefficient and complex representations, limiting the efficacy of their models.

Meanwhile, language models (Gruver et al., 2024; Sriram et al., 2024b) have emerged as a promising alternative. Their pretrained knowledge makes them data-efficient, often requiring far fewer domain-specific examples than training models from scratch (Gruver et al., 2024). Moreover, their natural-language interface allows one model to be applied to many tasks, including unconditional and conditional generation, infilling, and crystal structure prediction (Gruver et al., 2024).

Symmetry is a defining aspect of crystal structures. The set of rotations, reflections, inversions, and translations exhibited by a crystal lattice form its *space group*. A crystal's space group and symmetries are not merely a mathematical construct but critical to many optical, electrical, and magnetic properties like piezoelectricity (Malgrange et al., 2014; Yang et al., 2005). One way to define a crystal's structure from its symmetries is via Wyckoff positions, whereby one can specify a few key atomic coordinates and have the remaining atomic positions filled by applying symmetry

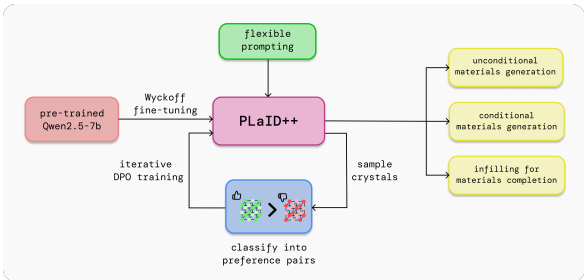

Figure 1: Overview of the PLaID++ pipeline, highlighting Wyckoff fine tuning and iterative DPO.

operations (Hahn et al., 1983). We train an LLM to learn and exploit structural parameters from the symmetries in each crystal. By allowing the model to discover patterns implicitly through training on Wyckoff-based text representations, we see significant improvements in generation performance.

To guide our search space towards chemically useful structures, we introduce Reinforcement Learning from Interatomic Potentials (RLIP), a reinforcement learning framework for physically grounded materials generation. To do so, we adapt Direct Preference Optimization (DPO) (Rafailov et al., 2023), a computationally efficient reinforcement learning method. Our pipeline is as follows: first, we perform supervised finetuning on a base LLM with Wyckoff-based text encodings of crystals. Then we further fine-tune the LLM via multiple rounds of DPO on generated structures categorized by their stability, novelty, and space group. To prevent entropy collapse and increase the diversity of generated structures, we increase the sampling temperature across successive iterations of DPO.

By incorporating inherent crystal symmetries and the feedback of a Machine Learning Interatomic Potential (MLIP), we significantly increase the rate of stable, unique, and novel (S.U.N.) materials. Our experiments demonstrate that our method **PLaID++ generates materials at a state of the art stability rate and generates S.U.N materials at a $\sim$50% higher rate** than prior models while retaining the flexibility of natural language prompting. By using representations that encourage learning from physical priors and diversity-aware training, we also highlight the effectiveness of reinforcement learning for materials discovery, **achieving $\sim$115% and $\sim$50% improvements in unconditional and space group conditioned generation, respectively, compared to fine-tuning alone.** Our contributions are as follows:

1. We introduce Reinforcement-Learning from Interatomic Potentials (RLIP), a diversity-aware framework for fine-tuning LLMs for material generation.
2. Motivated by mode collapse after preference optimization on a coordinate-based representation, we develop a novel, symmetry-informed text representation for crystal structures which is compact, performant, and physically-motivated.
3. We demonstrate that unified training across conditional and unconditional generation tasks are mutually beneficial for data-sparse regimes. We achieve state-of-the-art results for generating novel and stable materials in both settings.

2 Related Work

Variational autoencoders (VAEs) were among the early approaches to crystal structure generation, using learned latent spaces to produce physically plausible atomic arrangements (Xie et al., 2021; Court et al., 2020). Specifically, CDVAE introduced a continuous representation of crystal structures, however, VAEs face challenges in generating structures with strict crystallographic constraints, as their representations often lack direct symmetry considerations.

Diffusion models like DiffCSP and FlowMM iteratively denoise atomic positions to generate crystal structures, achieving higher stability and validity than earlier VAE approaches (Jiao et al., 2023; Miller et al., 2024). While these models excel at capturing complex chemical landscapes, they face significant inference cost due to multi-step denoising, which limits their scalability for rapidly screening a broad sample of material candidates.

Language models are scalable and capable when leveraging extensive pre-training on vast text corpora. In particular, CrystalLLM (Gruver et al., 2024) pioneered the use of fine-tuned LLMs for crystal generation, showcasing the effectiveness of text-based representations for materials prediction.

Symmetry-aware language models have introduced Wyckoff-position-based representations, leveraging inherent crystallographic symmetries to more effectively constrain the generative process within physical reality. WyFormer and CrystFormer (Kazeev et al., 2024; Cao et al., 2024) utilized transformer-based architectures to predict atomic positions conditioned explicitly on predefined space groups and chemical formulas. Although these symmetry-aware methods showed improvements in structural validity, their reliance on explicit formula and space group constraints considerably limits their utility in exploratory *de novo* generation tasks.

3 BACKGROUND

3.1 CRYSTALLOGRAPHY

Crystal structures are periodic arrangements of atoms defined by repeating unit cells. Each unit cell is fully described by lattice parameters—consisting of side lengths (l_1, l_2, l_3) and angles $(\theta_1, \theta_2, \theta_3)$—and atomic positions within the cell, characterized by their fractional coordinates (x_i, y_i, z_i) and elemental identities e_i.

The properties of crystals are heavily influenced by their underlying symmetry, captured mathematically by their space groups. Each space group comprises a set of symmetry operations, such as rotations, reflections, and inversions, that map the crystal structure onto itself. Atomic positions consistent with these symmetry operations are defined through Wyckoff positions (Hahn et al., 1983), significantly reducing the degrees of freedom required to specify a structure. Consequently, employing Wyckoff positions in generative modeling can greatly enhance the validity and physical plausibility of generated structures. We present additional mathematical background on Wyckoff positions in Appendix 7.6.

In computational materials science, thermodynamic stability is a critical metric which can be described using the energy above hull (E^{hull}), which quantifies how energetically favorable a material is compared to known competing phases (Sun et al., 2016). Materials with $E^{\text{hull}} \leq 0$ eV/atom at a given temperature and pressure are considered thermodynamically stable, while those slightly above zero (typically $E^{\text{hull}} < 0.1$ eV/atom) are metastable and potentially synthesizable (Jain et al., 2013).

3.2 REINFORCEMENT LEARNING FOR LLM FINE-TUNING.

RL methods align LLMs to human preferences by maximizing a learned reward while regularizing toward a reference model to prevent mode collapse (Bai et al., 2022). Actor–critic variants such as PPO reduce gradient variance but add overhead via explicit reward and value models (Schulman et al., 2017). Direct Preference Optimization (DPO) bypasses these components by fitting an implicit reward from preference pairs and optimizing a KL-regularized objective directly(Rafailov et al., 2023; Widatalla et al., 2024). DPO uses observed preference pairs of responses (y_w, y_l), generated from π_{ref}, where y_w denotes the preferred response. DPO optimizes an implicit reward function $r^*(x, y)$ by maximizing the likelihood of preferences under a Bradley-Terry preference model: $p(y_w \succ y_l \mid x) = \exp(r^*(y_w|x))/(\exp(r^*(y_w|x)) + \exp(r^*(y_l|x)))$. The authors show that a KL-regularized objective can be optimized without directly instantiating $r^*(x, y)$, leading to the following objective:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right] \tag{1}$$

To effectively generate stable structures, we employ a reinforcement learning framework based on DPO. By curating a dataset of stability and novelty-ranked crystal pairs evaluated via MLIP relaxations, our model iteratively learns to favor generation of structures with lower predicted energies in unseen crystal space, thus aligning generative behavior towards chemically stable and experimentally realizable new materials.

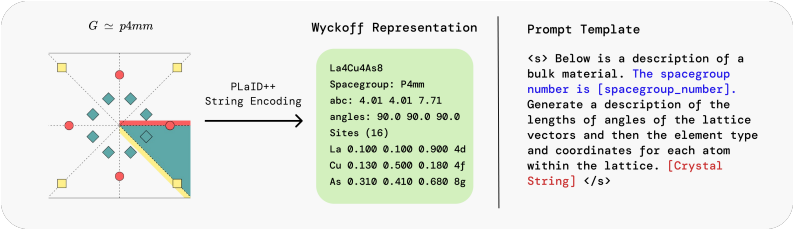

Figure 2: **Left:** An example crystal highlighting the $p4mm$ space group symmetry, where colors represent atoms of different elements. We represent the entire asymmetric unit of the crystal in our Wyckoff-based text representation by leveraging the crystal's symmetry. **Right:** The template from which we generate prompts used for training. For conditional generation, we include the blue conditioning information, and for unconditional generation, we remove it from our prompt. In all prompts, the crystal string is replaced with the encoding on the left.

4 METHOD

4.1 BASE MODEL

Our approach begins with a pre-trained large language model (Qwen-2.5 7B) (Yang et al., 2024), which has demonstrated strong capabilities in text generation and reasoning. Following the methodology outlined by (Gruver et al., 2024), we fine-tune the base version of Qwen-2.5 7B to model the crystal generation process via parameter-efficient fine-tuning (Hu et al., 2021). Unlike prior works that rely on explicit space group constraints or predefined templates, our method enables more flexible and generalizable unconditional and conditional structure generation through the use of a natural language prompt.

The training objective is to maximize the likelihood that the model generates the correct token sequence representing stable crystal structures. We provide the prompt template used to fine-tune our model on both objectives in Figure 2. We fine-tune two versions of the base model: one where the crystal structures are represented as text-based 3D coordinates and another based on a new text-based Wyckoff representation.

4.2 DIRECT PREFERENCE OPTIMIZATION (DPO)

To explicitly align our model toward generating crystals with desired properties, we employ an iterative form of Direct Preference Optimization (DPO) (Rafailov et al., 2023), a reinforcement learning technique that directly optimizes model outputs based on ranked preferences as highlighted by Equation 1. We jointly optimize our model on three explicit objectives: thermodynamic stability, novelty with respect to a reference crystal set, and space group conditioning.

Stability To train the model to prefer stable crystals, we use EquiformerV2 86M (eqV2) (Barroso-Luque et al., 2024) as a reward signal. eqV2 is a Machine Learning Interatomic Potential (MLIP) that predicts relaxed formation energies for materials, serving as an efficient proxy for stability assessment. To guide the model towards generating more stable materials, we categorize a crystal's energy above the hull into one of three buckets: stable (≤ 0 eV/atom E^{hull}), metastable (≤ 0.08 eV/atom E^{hull}), or unstable (> 0.08 eV/atom E^{hull}). We create stability preference pairs (x, y_w, y_l), where x is the model prompt and y_w and y_l are accepted and rejected crystals. Concretely, we sample preference pairs to form (stable, metastable) and (stable/metastable, unstable) sets, forming a tiered dataset that encodes relative preferences over increasing degrees of thermodynamic stability. This allows us to incorporate fine-grained (stable, metastable) reward signals into our training set without overfitting to exact eqV2-predicted E^{hull} values and without relying heavily on stable generations from the base model, which are too infrequent to provide sufficient training signal as seen in Appendix 7.2.

Uniqueness and Novelty To incorporate novelty into our reward, we construct preference pairs that distinguish between (stable & novel, stable & not novel). Incorporating uniqueness into a pairwise reward is non-trivial and presents many challenges. Namely, since uniqueness is defined at the group level (i.e. whether or not crystals within a set of rollouts are considered the "same" crystal as defined by Pymatgen's StructureMatcher (Ong et al., 2013)), pairwise reward computation of

uniqueness is ambiguous. However, we can control sample diversity during rollouts via the sampling temperature. As temperature increases, uniqueness and novelty tend to increase, while stability and validity of generated crystals decrease. Based on this observation, we find that increasing the sampling temperature across successive iterations of DPO can prevent overoptimization for stability, preventing the model from collapsing on narrow regions of chemical space.

Space Group To condition our model on generating crystals with a desired space group, we add an additional constraint during preference pair construction. For each space group, we use a conditional generation prompt indicating the target space group number and compare crystals that match this group. Preference pairs are then formed by contrasting (1) stable or metastable crystals with the correct space group against other stable or metastable crystals with the incorrect space group, and (2) stable or metastable structures against unstable structures. This strategy requires stability as a base condition for generation, but also encourages the model to align for structural symmetry.

Fine-Tuning We emphasize that both the S.U.N. and space group preference pairs are jointly used in the same DPO objective across the unconditional prompt and space group specific conditional prompts. This joint optimization encourages the model to learn to generate both generally stable crystals and space-group-specific structures in a unified pipeline.

To generate this dataset, we sample 10,000 crystal structures from our fine-tuned Qwen model for unconditional generation and sample 1,000 crystal structures across each of seven space groups. More details about our space group conditional generation is provided in Section 5.1; From here, we categorize our samples into our preference datasets based off the above metrics, resulting in a universal dataset that covers both unconditional and conditional generation.

We apply DPO to fine-tune Qwen on this curated preference dataset. We adopt an iterative DPO approach in which $\pi_{\text{ref}} = \pi_{\theta-1}$, with generations from $\pi_{\theta-1}$ used to construct the set of preference pairs for π_θ. We have additional information about our DPO hyperparameters, our dataset creation process, as well as other potential DPO variants and ablations in Appendix 7.2.

4.3 Wyckoff Representation

When RLIP is naively trained on 3D coordinate representations, PLaID++ models tend to overfit preference pairs and undergo *mode collapse*, producing stable but repetitive crystals (See Section 5.3). We address this failure by introducing a novel Wyckoff representation.

Crystal structures inherently follow symmetry constraints that can be represented by space groups and Wyckoff positions. To improve the model's ability to generate valid and stable structures, our Wyckoff representation encodes these symmetries directly into text. Unlike prior approaches such as DiffCSP++ (Jiao et al., 2024), which rely on searching template structures for Wyckoff sites, our method predicts them directly within the generative process.

Our representation is computationally efficient: on the MP-20 dataset, Wyckoff encodings average 185.5 tokens per crystal compared to 214.7 for coordinate-based strings, a 14% reduction that translates to better sample efficiency and faster training. Moreover, the format scales efficiently as our crystals grow more complex, since additional atoms at the same Wyckoff site require only a multiplicity update rather than new coordinate lines. By tying atoms to Wyckoff sites, a single change propagates through symmetry operations, forcing the model to generalize physical priors. Together, these properties make Wyckoff encoding a compact, symmetry-aware foundation that both improves efficiency and mitigates diversity collapse during RL fine-tuning.

5 Experiments

5.1 Setup

We trained our model on the well-established MP-20 dataset (Xie et al., 2021), a collection of 45,231 inorganic metastable crystalline materials from the Materials Project (Jain et al., 2013) with up to 20 atoms. We follow the methodology of Gruver et al. (2024) by independently fine-tuning a pre-trained Qwen-2.5 7B model using 4-bit quantization and Low-Rank Adapters (LoRA) (Hu et al., 2021), implemented with PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020). Symmetry information for our Wyckoff representation is calculated by Pyxtal Fredericks et al. (2021). Following

Table 1: Results for unconditional materials generation on the MP-20 dataset. Our flagship PLaID++ variant uses a Wyckoff text representation, DPO on stability and novelty-based preference pairs, and dynamic temperature adjustment. Column-best figures are in **bold**.

Method	Params	Validity (%) (↑)		Stability (%) (↑)	S.U.N. (%) (↑)
		Structural	Composition		
CDVAE	–	**100.00**	86.70	1.6	–
DiffCSP	–	**100.00**	83.25	5.0	3.3
FlowMM	–	96.85	83.19	4.6	2.8
FlowLLM	–	99.94	90.84	13.9	4.7
MatterGen-MP	–	–	–	13.0	–
CrystalLLM-70B	–	99.60	95.40	5.28	–
Jointly-trained ADiT	–	99.74	92.14	15.4	5.3
PLaID++ (Non-DPO Types)	3D Coord	98.93	88.94	7.20	2.81
	Wyckoff	99.80	91.22	7.17	3.58
	Iterative SFT	99.85	94.54	9.85	4.13
PLaID++ (DPO Types)	3D Coord	96.62	89.94	13.94	1.69
	Wyckoff	98.81	92.57	15.59	6.25
	Stability + Novelty	99.78	95.64	16.72	6.22
PLaID++	–	99.75	**97.34**	**22.27**	**7.74**

supervised fine-tuning, we apply DPO on the generated preference dataset to further guide generation towards stable structures. Full hyperparameters details are provided in Appendix 7.2.

For unconditional generation, we sample 10,000 structures from each fine-tuned model, parsing a CIF from the generated string. We resample if a CIF cannot be parsed from the string, which guarantees all samples can be interpreted as crystals but does not guarantee validity.

Similarly, for space-group conditioned generation, we sample 1,000 structures for each of seven space groups. These space groups were chosen in accordance with (Zeni et al., 2025) as they represent each of the seven crystal systems with varying levels of symmetry. Specifically, we sample from space groups $P1$, $C2/c$, $Amm2$, $I\bar{4}m2$, $P3$, $P6_3/mmc$, and $F\bar{4}3m$.

To validate the stability and accuracy of our results, we performed energy relaxations on samples generated by our flagship model fine-tuned with DPO and the Wyckoff encoding. Although accurate, Density Functional Theory (DFT) is computationally expensive, scaling typically as $O(N^3)$ with respect to the number of atoms N. Therefore, we leverage machine learning interatomic potentials (MLIPs) such as eqV2 (Barroso-Luque et al., 2024) and eSEN (Fu et al., 2025) as efficient proxies for DFT. These MLIPs predict the relaxed atomic positions and energies with significantly reduced computational overhead. To verify our eSEN results, we compute stability using DFT on a random 1,000 crystal subset of our generated crystals. Detailed information regarding our DFT setup are in Appendix 7.2.

5.2 Metrics

To initially evaluate the quality of our generated crystal structures, we focus on structural and compositional validity, as defined by Xie et al. (2021). These metrics provide an effective proxy for assessing the quality of generated crystals before conducting more computationally expensive stability evaluations. We explain more details about these metrics in Appendix 7.1

Our primary metric for evaluating generated crystal structures is the **S.U.N Rate** from Miller et al. (2024), which measures the percentage of crystals that are **stable, unique**, and **novel**. Stability is assessed by comparing a crystal's energy to a convex hull of previously computed energies from Riebesell et al. (2023). Crystals on or below the hull (≤ 0 eV/atom E^{hull}) are deemed *stable*. We evaluate **uniqueness**—differentiation from other generated crystals—and **novelty**, which measures diversity from the training data to calculate S.U.N.

For our space group conditional generation task, we adopt the **S.S.U.N Rate** metric from Zeni et al. (2025), which measures the percentage of strucutures that are the correct symmetry group as deemed by Pyxtal (Fredericks et al., 2021) and are metastable (≤ 0.1 eV/atom E^{hull}), unique and novel.

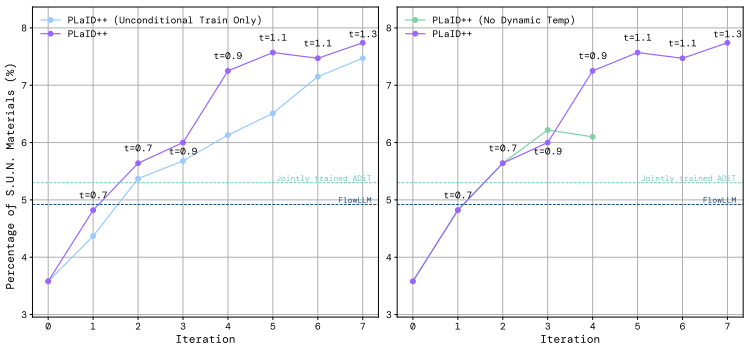

Figure 3: Evolution of S.U.N. percentage of PLaID++ variants' over DPO iterations. Reference lines represent S.U.N. rates from ADiT and FlowLLM's flagship models. **Left:** Ablation over joint training. **Right:** Ablation over dynamic temperature.

Due to compute constraints, we assess stability for unconditional and conditional generation primarily via the eSEN MLIP (Fu et al., 2025). Note that we specifically **use different MLIPs**—eqV2 and eSEN—for preference dataset creation and crystal generation evaluation to avoid reward hacking or overfitting to a specific MLIP.

5.3 Results

Unconditional Generation We compare our model to seven prior methods, as reported from Joshi et al. (2025): CDVAE (Xie et al., 2021), DiffCSP (Jiao et al., 2023), FlowMM (Miller et al., 2024), FlowLLM (Sriram et al., 2024b), MatterGen-MP (Zeni et al., 2025), CrystalLLM (Gruver et al., 2024), and ADiT (Joshi et al., 2025). Our main results are presented in Table 1. On stability and S.U.N, the most important metrics, PLaID++ outperforms all prior methods. For our best overall PLaID++ model, 22.27% of the relaxed structures are deemed stable, out of which 72% are novel, and 55% are unique, leading to a S.U.N. rate of 7.74%. **PLaID++ achieves the highest stability rate and a $\sim$50% higher S.U.N rate than the best previous method.**

On the proxy metrics, PLaID++ achieves performance on par with other methods. Our Wyckoff encoding specifically increases compositional validity for both the base and DPO fine-tuned models, highlighting symmetry-based encodings as a natural and intuitive mechanism to increase the validity and stability of generated structures. Though many of these metrics have saturated, we report PLaID++'s performance for comparison and completeness.

We also observe that PLaID++ is significantly more computationally efficient than comparable methods. On a singular NVIDIA 80GB H100 GPU, PLaID++ samples 10,000 crystals in approximately 23 minutes, yielding 27.17 S.U.N. crystals per minute. **This generation speed is 5x faster than FlowLLM**, which takes 89.6 minutes to generate 10,000 materials on a NVIDIA 80GB A100 GPU, yielding 5.25 S.U.N. crystals per minute (Sriram et al., 2024b). Such high throughput is essential in real-world discovery, as faster sampling dramatically shortens the design–build–test cycle, enabling rapid screening of candidates and speedups of experimental pipelines (Szymanski et al., 2023; Curtarolo et al., 2013).

Conditional Generation To test the ability of PLaID++ to generalize towards targeted structure synthesis, we measure the S.S.U.N. percentage for four model variants: (i) the 3D coordinate fine-tuned Qwen model (ii) the Wyckoff fine-tuned Qwen model, (iii) DPO with data only from the conditional generation task (iv) DPO trained jointly on conditional and unconditional generation data (flagship PLaID++). As shown in Figure 4, applying DPO to space-group pairs increases S.S.U.N. by an average of 22% compared to the base Wyckoff model. Incorporating additional preference data from unconditional generation further boosts performance, resulting in a total increase of 47% over the base model—more than doubling the improvement from space-group conditioning data alone.

Groups with low training data occurrences ($<$ 400 samples each) like ($Amm2$, $Im2$, $P6_3/mmc$) show little improvement or sometimes slight degradation after DPO. This supports the elicitation hypothesis of reinforcement learning (Lambert, 2023), whereby RL amplifies existing valuable

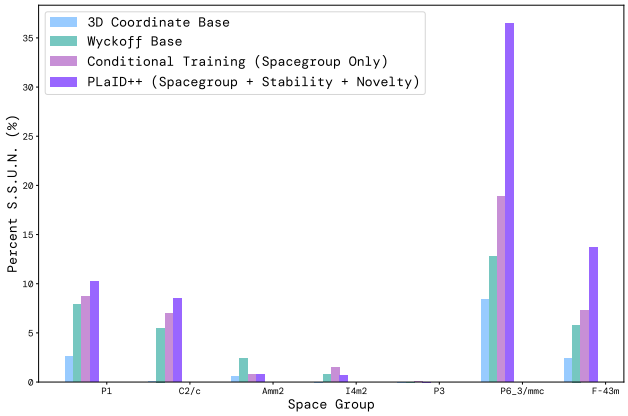

Figure 4: Histogram of S.S.U.N. results across models for space group conditional generation tasks.

behaviors in the base model rather than teaching models new ones. With too few examples, neither the space-group nor the stability objective can generate a meaningful training signal for the model to learn from.

Representation Choice is Critical to Prevent Mode Collapse

In Table 2, we demonstrate the importance of our crystallographic representation. As we perform more iterations of DPO with the 3D coordinate representation, the diversity of the distribution quickly collapses and the model struggles to generalize from the DPO examples seen during training. Each iteration of the post-trained model has a progressively lower S.U.N. rate even as stability increases. This shows that **with the incorrect representation, RL leads the model to memorize rather than generalize from reward feedback.** We hypothesize that our Wyckoff representation works because of its higher information density, where local atom type or coordinate changes propagate to globally meaningful structure differences, expanding the diversity of valid structures the model can learn and generate. This is further reinforced by the ablation on the SFT base models in Table 1, where despite similar stability rates, the uniqueness and novelty of the Wyckoff representation is significantly higher, leading to a $\sim$25% increase in S.U.N.

Metric	It 1	It 2	It 3
S.U.N. (%)	2.93	2.23	1.69
Stability (%)	10.35	13.47	15.94

Table 2: PLaID 3D coordinate DPO's S.U.N. and stability percentage across iterations of RLIP

Reinforcement Learning from Interatomic Potentials (RLIP) Across multiple iterations of Direct Preference Optimization (DPO), we observe in Figure 3 that DPO successively improves the number of generated S.U.N. materials, demonstrating that reinforcement learning is an effective and robust method to improve generation performance. On unconditional and space group conditioned generation, **we improve S.U.N. by $\sim$115% and $\sim$50% respectively compared to fine tuning alone**.

From our ablation on temperature sampling in Figure 3, we highlight that removing dynamic temperature updates across DPO iterations leads to mode collapse, where the S.U.N. rate degrades even as stability continues to increase after 3-4 iterations. This suggests that increasing temperature acts as an entropy regularizer which encourages exploration and counteracts model collapse.

Joint Training Improves Performance Table 1 and Figure 4 show that our DPO model trained together on the unconditional and space group conditioned generation tasks outperforms the DPO models trained individually on each respective task. This validates the efficacy of a universal foundational model that can be leveraged across multiple downstream objectives. The performance gains are particularly pronounced in conditional generation, where only a few hundred samples are available in the training set. This suggests that the additional training signal derived from unconditional generation feedback can be successfully transferred to conditional generation.

PLaID++ Crystal Novelty Analysis To better understand how reinforcement learning reshapes generative behavior, we visualize the compositional embedding space of S.U.N. crystals from the

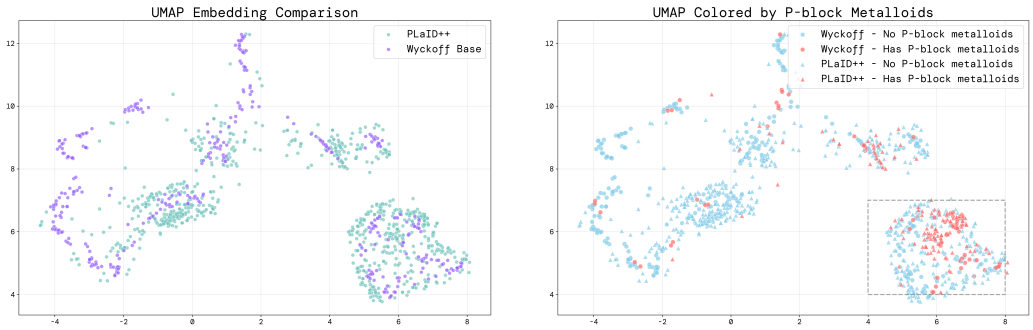

Figure 5: **Left:** UMAP visualizations of compositional embeddings of S.U.N. crystals generated by our PLaID++ and 3D Coord Base models. **Right:** Identical data, with crystals containing P-block metalloids being highlighted in red and those without P-block metalloids being in blue.

Wyckoff Base and PLaID++ models using UMAP (McInnes et al., 2018) (Figure 5). The left plot highlights a clear distributional shift: while both models occupy similar structural regions, PLaID++ expands into areas underrepresented in the base model, providing concrete evidence that RLIP enables local exploration around novel materials in our supervised baseline. In the right plot, we analyze the largest crystal cluster and observe that PLaID++ not only produces substantially more samples but also preferentially generates crystals containing P-block metalloids. Together, these visualizations demonstrate that PLaID++ is not merely memorizing preferences but uncovering out-of-distribution materials with meaningful structural diversity.

Density Functional Theory Performance Validation Relaxing 1,000 structures using DFT, we find a stability rate of approximately 19.1% and a S.U.N. rate of 13%. These results are consistent with our findings using the eSEN machine learning interatomic potential (Fu et al., 2025), providing strong empirical support for the quality of our generated materials. Notably, the DFT-based S.U.N. rate is higher than that obtained from the full 10,000-sample evaluation using eSEN. This discrepancy is expected due to the smaller DFT sample size, which yields a higher relative uniqueness rate.

6 Discussion

In this paper, we introduce a novel Wyckoff-based text encoding for crystal structures and introduce a new reinforcement learning framework, Reinforcement Learning from Interatomic Potentials (RLIP), to guide generation towards chemically stable and useful structures. Our method achieves state-of-the-art performance across stability and S.U.N. metrics. Current random structure search methods achieve less than a 1% success rate (Pickard and Needs, 2011) in identifying stable materials—PLaID++ represents a significant acceleration from traditional approaches to future methods with greater stability and real-world utility.

Limitations & Future Work While PLaID++ achieves state-of-the-art stability and S.U.N. rates, it has several limitations that point toward promising avenues for future work. We have fine-tuned and evaluated exclusively on the MP-20 dataset ($\sim$45K structures), and scaling to larger, more diverse datasets such as Alexandria ($\sim$2.6M structures) (Schmidt et al., 2024) or LeMaterial ($\sim$6.7M structures) (Siron et al., 2024) may further broaden the model's coverage of chemical space. Our experiments also focus on unconditional and space-group-conditioned generation, but extending generation to other objectives such as band-gap optimization would demonstrate wider practical utility. We also only explored Direct Preference Optimization (DPO) as our RL method; investigating alternative algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) or Group Relative Policy Optimization (GRPO) (Shao et al., 2024) might uncover different stability–diversity trade-offs.

Impact Statement Our approach has the potential to accelerate the discovery of novel materials for renewable energy, electronics, and carbon-capture applications—paving the way for technologies that benefit society. However, deploying generative models carries risks, including the synthesis of dangerous compounds and inequitable access to these tools.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *arXiv preprint arXiv:2403.15734*, 2024.

Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535, 2020.

Stefano Curtarolo, Gus LW Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. The high-throughput highway to computational materials design. *Nature materials*, 12(3):191–201, 2013.

Daniel W Davies, Keith T Butler, Adam J Jackson, Andrew Morris, Jarvist M Frost, Jonathan M Skelton, and Aron Walsh. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627, 2016.

Robert A Evarestov and Vyacheslav P Smirnov. *Site symmetry in crystals: theory and applications*. Springer Berlin Heidelberg, 1997.

Scott Fredericks, Kevin Parrish, Dean Sayre, and Qiang Zhu. Pyxtal: A python library for crystal structure generation and symmetry analysis. *Computer Physics Communications*, 261:107810, 2021. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2020.107810. URL http://www.sciencedirect.com/science/article/pii/S0010465520304057.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S Levine, Meng Gao, Misko Dzamba, and C Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. *arXiv preprint arXiv:2502.12147*, 2025.

Martin A Green, Anita Ho-Baillie, and Henry J Snaith. The emergence of perovskite solar cells. *Nature photonics*, 8(7):506–514, 2014.

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

Theo Hahn, Uri Shmueli, and JC Wilson Arthur. *International tables for crystallography*, volume 1. Reidel Dordrecht, 1983.

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, 2021.

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), 7 2013. doi: 10.1063/1.4812323.

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

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *arXiv preprint arXiv:2402.03992*, 2024.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W. Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials, 2025. URL https://arxiv.org/abs/2503.03965.

Nikita Kazeev, Ruiming Zhu, Ignat Romanov, Andrey E Ustyuzhanin, Shuya Yamazaki, Wei Nong, and Kedar Hippalgaonkar. Wyckofftransformer: Generation of symmetric crystals. In *AI for Accelerated Materials Design-NeurIPS 2024*, 2024.

Nathan Lambert. Elicitation, the simplest way to understand post-training. https://www.interconnects.ai/p/elicitation-theory-of-post-training, November 2023.

Cécile Malgrange, Christian Ricolleau, and Michel Schlenker. *Symmetry and physical properties of crystals*. Springer, 2014.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. *arXiv preprint arXiv:2406.04713*, 2024.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL http://arxiv.org/abs/1912.01703. cite arxiv:1912.01703Comment: 12 pages, 3 figures, NeurIPS 2019.

Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.

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. *Advances in Neural Information Processing Systems*, 36, 2023.

Janosh Riebesell, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*, 2023.

Jonathan Schmidt, Tiago F.T. Cerqueira, Aldo H. Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel A.L. Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024. ISSN 2542-5293. doi: https://doi.org/10.1016/j.mtphys.2024.101560. URL https://www.sciencedirect.com/science/article/pii/S2542529324002360.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Martin Siron, Inel Djafar, Lucile Ritchie, Etienne Du-Fayet, Amandine Rossello, Ali Ramlaoui, Leandro von Werra, Thomas Wolf, and Alexandre Duval. Lemat-bulk dataset, 2024. URL https://huggingface.co/datasets/LeMaterial/LeMat-Bulk.

Anuroop Sriram, Sihoon Choi, Xiaohan Yu, Logan M Brabson, Abhishek Das, Zachary Ulissi, Matt Uyttendaele, Andrew J Medford, and David S Sholl. The open dac 2023 dataset and challenges for sorbent discovery in direct air capture. *ACS Central Science*, 10(5):923–941, 2024a. doi: 10.1021/acscentsci.3c01629. URL https://doi.org/10.1021/acscentsci.3c01629.

Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=0bFXbEMz8e.

Wenhao Sun, Stephen Dacek, Shyue Ong, Geoffroy Hautier, Anubhav Jain, William Richards, Anthony Gamst, Kristin Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science Advances*, 2:e1600225–e1600225, 11 2016. doi: 10.1126/sciadv.1600225.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization. *bioRxiv*, pages 2024–05, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jiashi Yang et al. *An introduction to the theory of piezoelectricity*, volume 9. Springer, 2005.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, pages 1–3, 2025.

Qing Zhao, Sanjuna Stalin, Chen-Zi Zhao, and Lynden A Archer. Designing solid-state electrolytes for safe, energy-dense batteries. *Nature Reviews Materials*, 5(3):229–252, 2020.

7 APPENDIX

7.1 PROXY METRIC DETAILS

Validity provides a computationally efficient check on whether a generated crystal is physically plausible. We assess this through two criteria: *structural validity*, which ensures that no two atoms are closer than 0.5 Å, and *positional validity*, which verifies charge neutrality.

Stability Rate assesses how many generated materials are thermodynamically stable. Stability is determined by the energy above the hull (E^{hull}) metric, which measures the energy difference between a material and the convex hull of competing phases with the same composition. A material is considered stable if $E^{\text{hull}} \leq 0$ eV/atom, while those with $E^{\text{hull}} < 0.$ eV/atom are classified as metastable. We compute $\text{E}^{\overline{\text{hull}}}$ by relaxing generated structures using eSEN using the Materials Project dataset (February 2023) as a reference. Total energies were corrected using the MP2020 compatability scheme, which maintains consistency across various functionals (DFT/DFT+U).

S.U.N. Rate extends the stability metric to assess both novelty and uniqueness. A structure is novel if it is not structurally similar to any training set material, determined using Pymatgen's StructureMatcher (Ong et al., 2013). Uniqueness ensures that duplicate generations are not counted separately by grouping structurally similar outputs into equivalence classes. The S.U.N. rate is defined as the fraction of generated structures that are Stable, Unique, and Novel:

$$\text{Stability Rate} = \frac{N_{\text{stable}}}{N_{\text{gen}}} \times 100\% \tag{2}$$

$$\text{SUN Rate} = \frac{N_{\text{SUN}}}{N_{\text{gen}}} \times 100\% \tag{3}$$

Together, these metrics provide a more rigorous assessment of generative model performance by quantifying both the stability and originality of discovered materials.

7.2 ADDITIONAL EXPERIMENT DETAILS

Wyckoff Crystal Representation. We represent each crystal structure in PLaID++ as a structured text sequence that encodes key physical and symmetry-related attributes. The representation begins with the chemical formula, expressed as a concatenation of each element and its corresponding multiplicity. This is followed by the space group number, calculated using Pyxtal (Fredericks et al., 2021) with a symmetry tolerance of 0.01. Next, we include the lattice parameters: side lengths and angles, each rounded to two decimal places. Finally, for each atom in the structure, we list the element type, fractional coordinates (rounded to three decimal places), Wyckoff site label, and site multiplicity also from Pyxtal (Fredericks et al., 2021). This format provides a comprehensive and consistent description of a crystal's composition, geometry, and symmetry. We provide additional examples of generated structures in Appendix 7.7.

Supervised Finetuning. We perform supervised finetuning (SFT) on Qwen to adapt the model for crystal structure generation. For our fine-tuning, the model performs infilling one-third of the time. The remaining two-thirds of the time, we generate structures *de novo*, where we either generate crystals unconditionally, or randomly append between one and five additional properties in our prompt. The properties we include are the chemical formula, energy above the hull, formation energy per atom, band gap and space group although space group generation was the focus of this work.

To fine-tune different Qwen-2.5 7B models on both text-based and Wyckoff-based crystal representation, we followed the steps outlined by Gruver et al. (2024) to reproduce their results. We use an AdamW optimizer with batch size of 16 samples, a learning rate of 10^{-5}, fp-4 mixed precision alongside LoRA adapters (with a LoRA rank of 8, LoRA alpha of 32, and LoRA dropout of 0.05) to fine-tune over the MP-20 dataset for 10 epochs.

Iterative SFT Ablation. To directly compare rejection-sampled supervised finetuning (SFT) against Direct Preference Optimization (DPO), we trained a baseline that repeatedly fine-tuned only on metastable structures, matching our DPO iteration count, learning rate, and sample size. Results are summarized below.

We find that although iterative SFT improves upon vanilla SFT, it underperforms DPO by a large margin, underscoring the benefit of fine-grained preference learning for stability and novelty.

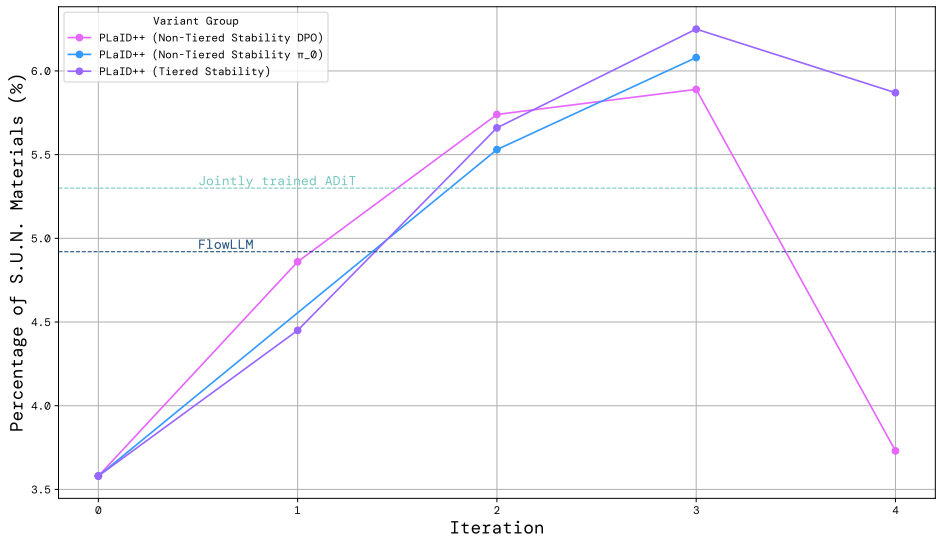

Figure 6: DPO Ablation, highlighting percentage of S.U.N. Materials across iterations of PLaID++ variants with different DPO pairs

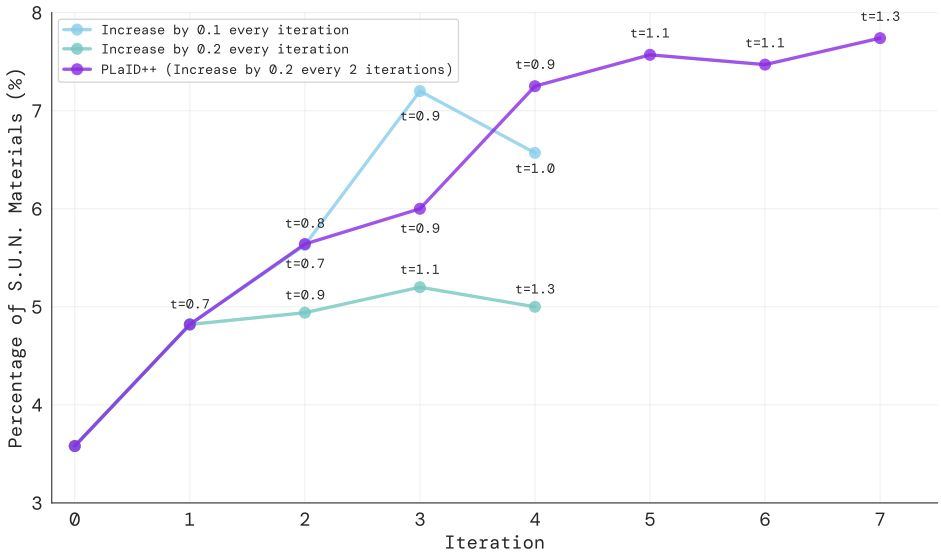

Figure 7: Comparison of temperature schedules during DPO. Fast temperature ramps (0.1/iter, 0.2/iter) degrade S.U.N. performance, while the PLaID++ schedule (0.2 every two iterations) balances the exploration-exploitating tradeoff, yielding the best results.

DPO. After supervised finetuning, we apply Direct Preference Optimization (DPO) to align the model's probability distribution to favor selected structures. We use the DPO Trainer from TRL (Wolf et al., 2020) using the Adam optimizer with a batch size of 16 samples, a learning rate of 10^{-6}, fp-4/bfloat-16 precision, and a β of 0.1 on one epoch of our dataset.

To fine-tune our model using Direct Preference Optimization (DPO), we curate a dataset of preference pairs drawn from two sources: (1) unconditional generation outputs and (2) space group–conditioned generation outputs. These datasets are merged and used jointly in the DPO training objective.

For the 10,000 samples generated from our unconditional generation task, we classify each structure using energy above the hull as predicted by eqV2. We use the following stability thresholds: stable ($E^{\text{hull}} \leq 0$), metastable ($0 < E^{\text{hull}} \leq 0.08$), and unstable ($E^{\text{hull}} > 0.08$). We form preference pairs using the following procedure. For each stable crystal, we sample one metastable and two unstable crystals to create three pairs: one $(\text{stable}, \text{metastable})$ and two $(\text{stable}, \text{unstable})$. For each metastable crystal, we sample two unstable crystals to create $(\text{metastable}, \text{unstable})$ pairs.

This tiered pairing strategy provides a more fine-grained reward signal than binary stable/unstable classification and avoids overfitting to exact energy values while increasing the diversity of preference samples.

We also construct a preference dataset from crystals generated using prompts that specify a desired space group number as described in Figure 2. For each of the seven selected space groups, we sample 1,000 structures and compute both the eqV2-predicted energy and the actual space group using Pyxtal with a symmetry tolerance of 0.01 (Fredericks et al., 2021). Crystals with $E^{\text{hull}} \leq 0.08$ are considered acceptable. Within this set, we further divide crystals into those that match the target space group and those that do not. We construct preference pairs by sampling $(\text{matching SG}, \text{non-matching SG})$ pairs among stable/metastable crystals and sampling $(\text{stable/metastable and unstable})$ pairs in a 1:2 accept reject ratio to enforce both symmetry and stability constraints.

By combining these unconditional and conditional preference pairs into a single dataset, we enable the model to jointly learn to generate stable crystals and adhere to symmetry constraints within a unified DPO training pipeline.

This creates between 10,000 to 20,000 DPO pairs depending on the iteration number and metastability and symmetry rates. Interestingly, the model incursworse performance when we compile our dataset using a 1:1 and 1:10 ratio. We hypothesize that smaller ratios perform worse as our training dataset becomes noticeably smaller, and larger ratios perform worse as the model overfits to each positive sample and consequently generates less diverse structures and lower S.U.N. rates.

DPO Ablation. We analyze the S.U.N. performance for different DPO variants in Figure 6. The PLaID++ (Non-Tiered Stability DPO) variant applies DPO using stability-based preference pairs, where accepted crystals are stable or metastable and rejected crystals are unstable. The PLaID++ (Non-Tiered Stability π_0) variant keeps the reference model fixed as the supervised fine-tuned Qwen Wyckoff base model instead of updating the reference to the previous iteration's model ($\pi_{\text{ref}} = \pi_{\theta-1}$) as in the flagship PLaID++ method. The PLaID++ (Tiered Stability) variant is the model which uses tiering on (stable, metastable), and (metastable, unstable) pairs. The results demonstrate the efficacy of both using a tiered DPO stability objective and using the previous iteration model as the reference model for KL divergence calculation.

eSEN-30M-OAM and eqV2 MLIPs. We employ eqV2 to curate our preference dataset (assessing crystal structures above and below the 0.08 eV/atom E^{hull} threshold) and eSEN-30M-OAM (eSEN) for our stability evaluations by computing relaxed formation energies. For eqV2 and eSEN, we used their reported settings with 500 relaxation steps and a max force of 0.02 (Fu et al., 2025; Barroso-Luque et al., 2024).

Density Functional Theory Setup. We used the Vienna Ab initio Simulation Package (VASP) version 6.3.2 software. All unit cells were relaxed using the default VASP parameters used for calculations in the Materials Project database as described in Pymatgen v2025.5.1 (Ong et al., 2013). These are collinear spin-polarized calculations which include a plane wave energy cutoff of 520 eV, a maximum force tolerance of 0.5×10^{-5} eV/Å, a Γ centered k-point mesh of 64 k-points per Å^3.

Experiments – Server Details. We conducted training on a high performance computing (HPC) cluster equipped with NVIDIA H100 GPUs. We used a single NVIDIA H100 GPU for Qwen-2.5 7B SFT and DPO as well as inference. We used an AMD Epyc 7443P for S.U.N. rate evaluations and DFT calculations.

7.3 Space Group Distribution Analysis

We compare visualizations between the space group distribution of the MP-20 dataset we used to perform SFT on our base models, a sample of 10,000 crystals from Qwen SFT'd on the 3D coordinate representation, and a sample of 10,000 crystals we from our PLaID++ model trained on iterative

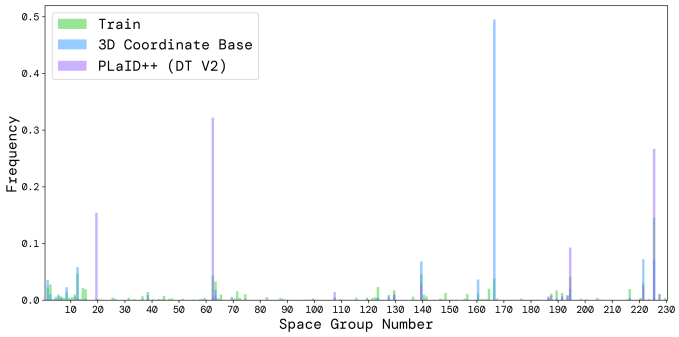

Figure 8: A comparison between the three space group distributions allows us to extract valuable insights on how our PLaID variants affect model behavior.

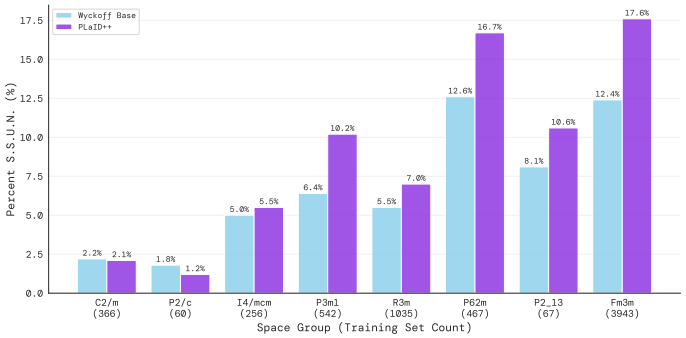

Figure 9: S.S.U.N. rates for space-group–conditioned generation on space groups not directly optimized for in our RLIP pipeline. PLaID++ improves stability, uniqueness, and novelty across most space groups, when compared to the Wyckoff base model.

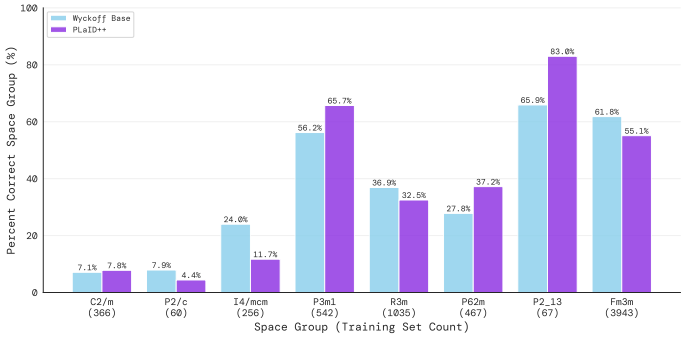

Figure 10: Space-group accuracy for conditioned generation on space groups not directly optimized for in our RLIP pipeline. PLaID++ generates more crystals in the correct space group for most groups.

DPO finetuning with our Wyckoff representation in Figure 8. We note that the flagship model favors generating out-of-train-distribution space groups such as $P2_12_12_1$ (19) and $Pnma$ (62), a consequence of our RLIP framework's novelty incentive.

Space groups Cm (8) and $P1$ (1) are prominent in the base 3D coordinate representation PLaID++ model's crystals, but not in the flagship model. Space group *Cm* has a single symmetry operation (a glide plane), while space group $P1$ only has the identity operation. In addition, space groups $P6_3/mmc$ (194) and $I4mmm$ (139) are among the five most frequently generated space groups in the flagship model, even though they are not in the top five of the training set. This supports the

hypothesis that the Wyckoff representation helps PLaID++ generate systems with complex symmetry requirements, skewing the distribution towards generating crystals with higher symmetries and ensuring a more even distribution across all space groups.

7.4 Validation of MLIPs as a Surrogate for DFT Energy Above Hull

To justify using eSEN and eqV2 MLIP energies for evaluations in place of expensive DFT calculations, we compared eSEN/eqV2-predicted and DFT-computed energies above the convex hull (E_{hull}) on a sample of 1000 generated structures. All structures were sampled from PLaID++, then filtered to those whose DFT energies are between $[-0.2, 0.2]$ eV/atom. This "near-hull" window was chosen because (i) our stability metric depends on correctly classifying structures around the zero-energy threshold, (ii) crystals with $|E_{\text{hull}}| > 0.2$ eV/atom either fail catastrophically in DFT or are unambiguously unstable/stable, making the precise MLIP E_{hull}'s accuracy less relevant for our evaluation metrics like S.U.N, and (iii) correctly classifying crystals as stable, metastable, or unstable is essential for creating accurate preference pairs used during the iterative DPO process.

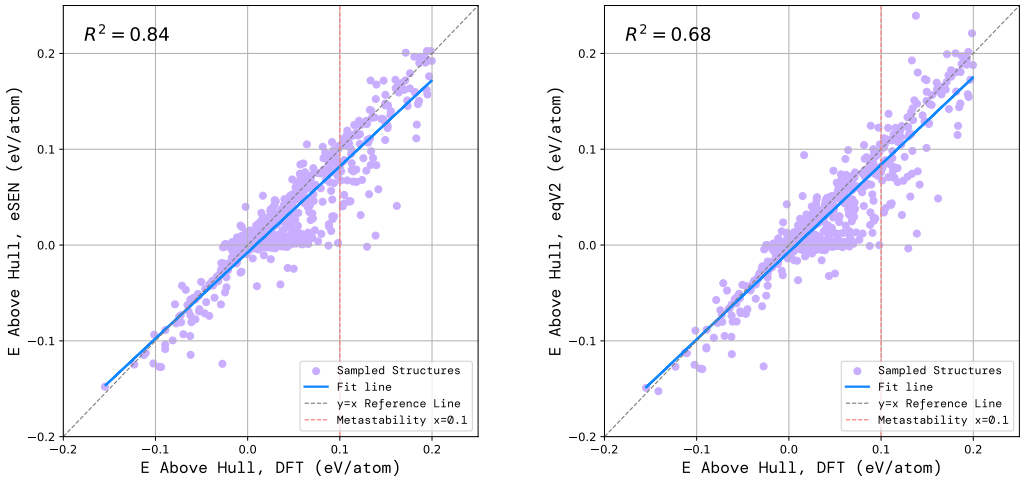

Figure 11: Scatter of MLIP vs. DFT energies above hull for 1000 sampled structures each with a least-squares fit line. **Left:** eSEN predictions ($R^2 = 0.84$). **Right:** eqV2 predictions ($R^2 = 0.68$).

eSEN. Figure 11 on the left shows a strong linear correlation ($R^2 = 0.84$) between eSEN and DFT energies in the critical near-hull region. Based of this high coefficient of determination, eSEN serves as a reasonable proxy for DFT in our evaluation pipeline: it reproduces stability classifications (stable vs. unstable) and catastrophic failure cases outside $[-0.2, 0.2]$ eV/atom while enabling a large reduction in computational cost. Experimentally, eSEN relaxes 10,000 crystals in approximately 5 hours, whereas DFT requires 12 hours for 1,000 crystals—yielding a $\sim$24× speedup with only a minimal accuracy trade-off.

We calculate eSEN's accuracy as a proxy for DFT in evaluation of the stability and metastability of crystals. When classifying the stability of structures at 0.0 eV eSEN achieves a precision of 0.696, a recall of 0.688 and $F_1 = 0.692$ (accuracy = 0.90), highlighting its accuracy in stability classification. Allowing for a small metastability margin (cutoff = 0.1 eV) boosts precision to 0.933 and recall to 0.992, giving $F_1 = 0.962$ (accuracy 0.95). Although imperfect, eSEN's relatively high F1 and accuracy scores demonstrate that it can be useful to approximate metrics like stability rate and S.U.N. without needing the prohibitive cost of DFT on a full 10,000 samples.

eqV2. Figure 11 demonstrates a strong linear correlation ($R^2 = 0.68$) between eqV2 and DFT energies near the stability threshold. Although noisier than eSEN, its high coefficient of determination demonstrates that eqV2 is a reliable surrogate for DFT in the critical near-hull region: replacing DFT with eqV2 in reward evaluation allows the pipeline to accurately create preference pairs for the iterative DPO process at a significantly lower computational cost compared to DFT. Experimentally, eqV2 performed at similar speeds to eSEN, yielding a $\sim$25× speedup over DFT.

Algorithm 1: PLAID++ Pipeline

Input: Crystal corpus $\mathcal{D} = \{\hat{y}_i\}_{i=1}^{D}$ ▷ In Wyckoff format
Initial policy $\pi_{\theta_{\text{ref}}}$ ▷ QWEN-2.5 7B
Stability predictor f_{MLIP} ▷ eqV2 MLIP
Batch size B, DPO Regularization β, Prompt x
Output: PLAID++ policy π_{θ_i}

Supervised fine-tuning
$\theta_0 \leftarrow \theta_{ref}$ ▷ Initialize from reference
for mini-batch $\mathcal{B} = \{\hat{y}_b\}_{b=1}^{B} \sim \mathcal{D}$ **do**
  $\mathcal{L}_{SFT} = -\sum_{b=1}^{B} \log \pi_0(\hat{y}_b \mid x)$
  $\theta_0 \leftarrow \texttt{OptimizerStep}(\mathcal{L}_{SFT};\ \theta_0)$

RLIP fine-tuning (Stability-only)
for $i = 1, \ldots, R$ **do**
  $\theta_i \leftarrow \theta_{i-1}$ ▷ Initialize from previous round
  $\{y_j\}_{j=1}^{K} \sim \pi_{\theta_{i-1}}(y \mid x)$ ▷ Generate candidates
  $s_j \leftarrow f_{\text{MLIP}}(y_j)$ ▷ $\approx$ e above hull
  Create preference pairs $\mathcal{P} = \{(y_w, y_l)_j\}_{j=1}^{S}$ where $s_w \succ s_l$ according to stability/metastability thresholds
  for mini-batch $\mathcal{B} = \{(y_w, y_l)_b\}_{b=1}^{B} \sim \mathcal{P}$ **do**
    $\mathcal{L}_{\text{DPO}} = -\frac{1}{B}\sum_{b=1}^{B}\left[\sigma\Big(\beta \log \frac{\pi_{\theta_i}(y_w|x)}{\pi_{\theta_{i-1}}(y_w|x)} - \beta \log \frac{\pi_{\theta_i}(y_l|x)}{\pi_{\theta_{i-1}}(y_l|x)}\Big)\right]$
    $\theta_i \leftarrow \texttt{OptimizerStep}(\mathcal{L}_{DPO};\ \theta_i)$

return π_{θ_i} ▷ Optimized policy

We further justify eqV2's ability to act as an accurate and efficient reward model proxy for the crystal stability and metastability classification tasks. At the stability cutoff (0.0 eV/atom), eqV2 attains a precision of 0.591, a recall of 0.818, and an F_1 score of 0.686 (accuracy = 0.87), reflecting strong sensitivity but moderate specificity around the exact hull threshold. Increasing the cutoff to metastability (0.08 eV) boosts precision to 0.915 and recall to 0.992, yielding an F_1 score of 0.952 (accuracy = 0.94). This justifies our tiered DPO reward formulation, whereby we provide the model with useful but noisy stability information and more accurate but less useful metastability information.

7.5 PLAID++ Overview

Algorithm 1 provides an overview of the full PLaID++ pipeline for the unconditional case (without space-group conditioning). Note that $s_w \succ s_l$ denotes preference (stable $\succ$ metastable $\succ$ unstable).

7.6 Wyckoff Theory: Mathematical Background

A Wyckoff position defines a set of equivalent atomic sites in a given space group (Evarestov and Smirnov, 1997). These positions are characterized by symmetry constraints, which reduce the degrees of freedom in atomic placement.

For a crystal structure belonging to a space group G, the Wyckoff positions can be defined as:

$$W = \{gx | g \in G\} \tag{4}$$

where x is a fractional atomic coordinate, and g is an element of the space group.

Each Wyckoff site follows symmetry constraints that ensure atoms occupy positions dictated by the underlying crystallographic symmetry:

$$x' = Rx + t \tag{5}$$

where R is a symmetry operation (such as a rotation, reflection, or inversion) that can be represented as a matrix, and t is a translation vector.

In this notation, x' is considered to be part of the same crystallographic orbit as x, where x is referred to as the generating point. As any one arbitrary point in the crystallographic orbit can be used to generate all other points in the orbit, the choice of x is not unique.

All symmetry operations (R, t) which map a point in the crystallographic orbit to itself form a finite subgroup of G, which can be identified as one of the 32 crystallographic point groups. Each Wyckoff site in each space group is associated with one such point group, and is often labeled according to the multiplicity of W (the number of points in the orbit generated by the space group in a single unit cell) and a letter encoding the point group (i.e. '4a').

7.7 Miscellaneous Examples of Generated Crystals

We present additional qualitative examples of crystal structures generated by our model. The prompts and generated Wyckoff representations from PLaID for both unconditional generation and two examples of space group conditioned generation are shown in Table 3. Seperately, we visualize four crystals using the ASE and PyMatGen toolkits. Figure 12 show several samples from different space groups.

Prompt	**Model Output**

`Below is a description of a bulk material. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice:`	`Ho4Al4Ni4`
	`Spacegroup: Pnma`
	`abc: 6.91 4.32 7.44`
	`angles: 90.00 90.00 90.00`
	`Sites (12)`
	`Ho 0.005 0.250 0.195 4c`
	`Al 0.190 0.250 0.590 4c`
	`Ni 0.789 0.250 0.591 4c`

`Below is a description of a bulk material. The spacegroup number is 15. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice:`	`Mn8O12F4`
	`Spacegroup: C2/c`
	`abc: 9.55 4.98 6.95`
	`angles: 90.00 96.49 90.00`
	`Sites (24)`
	`Mn 0.826 0.495 0.074 8f`
	`O 0.832 0.906 0.599 8f`
	`O 0.000 0.489 0.250 4e`
	`F 0.000 0.000 0.000 4a`

`Below is a description of a bulk material. The spacegroup number is 216. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice:`	`Ti4Ni4Sn4`
	`Spacegroup: F-43m`
	`abc: 5.89 5.89 5.89`
	`angles: 90.00 90.00 90.00`
	`Sites (12)`
	`Ti 0.500 0.500 0.500 4b`
	`Ni 0.750 0.750 0.750 4d`
	`Sn 0.000 0.000 0.000 4a`

Table 3: Example PLaID++ generations in response to unconditional and space group–conditioned prompts using the Wyckoff-based representation.

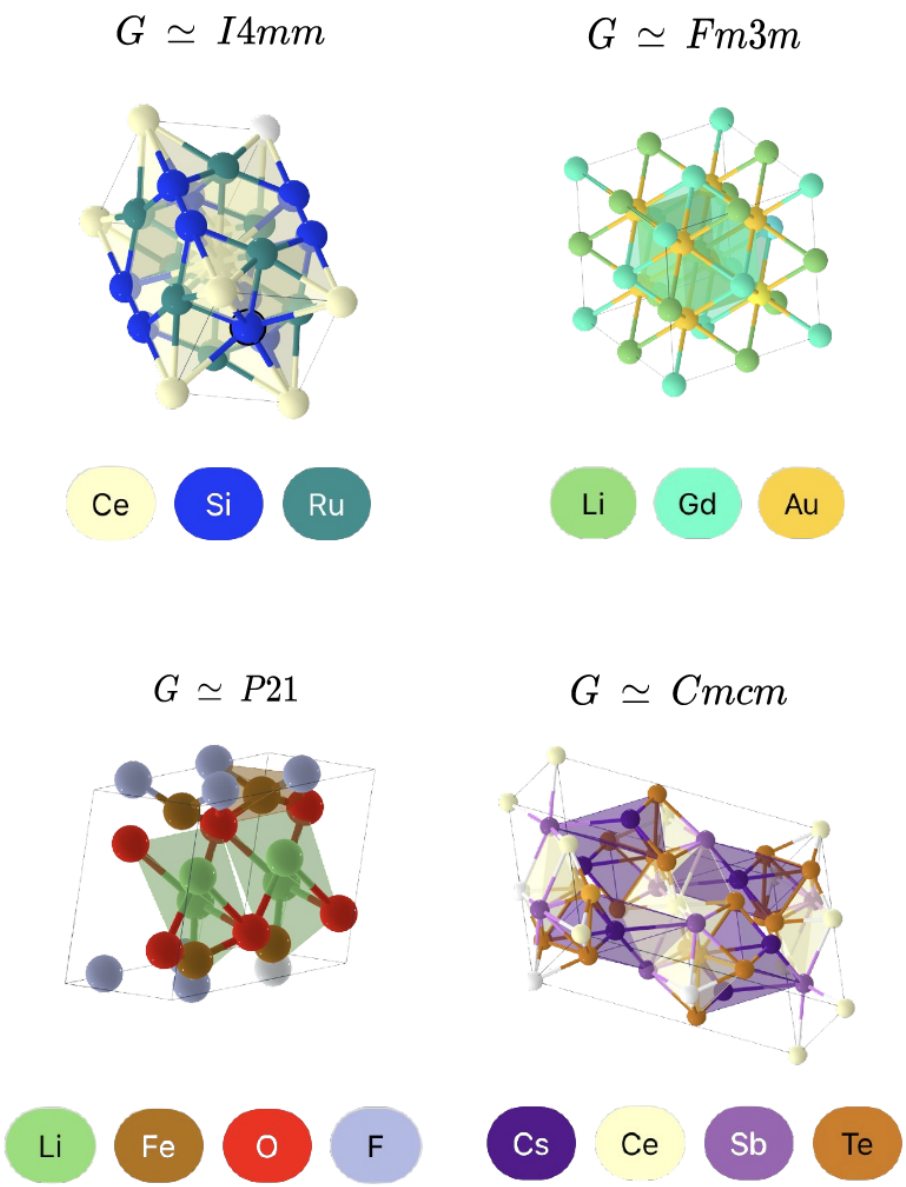

Figure 12: Examples of crystals generated by our Qwen-2.5-7B model fine-tuned with Wyckoff representation and Direct Preference Optimization (DPO). The crystals' corresponding chemical formulas are from top left to bottom right respectively, $Ce_2Si_4Ru_4$, $Li_4Gd_4Au_8$, $Cs_4Ce_4Sb_4Te_{12}$, and $Li_4Fe_4O_4F_4$, respectively.