

Aligning LLMs for Crystal Material Design: Test-Time Reinforcement Learning and Model Merging

Pengyu Wang^{1,4} Yang Han^{1,4} Xin Chen⁴ Kai Yu^{1,3,4}
Tong Zhu^{2,5} Lu Chen^{1,2,3,4} Bo Chen^{4,†}

¹X-LANCE Lab, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

²Shanghai Innovation Institute, Shanghai, China

³Jiangsu Key Lab of Language Computing, Suzhou, China

⁴Suzhou Laboratory, Suzhou, China

⁵East China Normal University, Shanghai, China

chenb@szlab.ac.cn

Abstract

Conditional crystal generation requires Large Language Models (LLMs) to produce Crystal Strings with discrete and continuous variables while satisfying composition, symmetry, and property constraints. Desired regions are often sparse in supervised data, while retraining for each objective preference is costly. We present a verifier-grounded framework combining label-free Test-Time Reinforcement Learning (TTRL) and Model Merging (MM). TTRL performs one batch-level, partially transductive adaptation per property using composition, symmetry, and learned band-gap rewards, without target reference structures. Property-specific experts are interpolated to navigate bounded symmetry–band-gap trade-offs without retraining intermediate policies. On MP-20, TTRL improves symmetry control, band-gap accuracy, and compositional fidelity over High-Precision SFT and Standard RL, and is competitive with diffusion and preference-aligned baselines. In a held-out high-band-gap region, TTRL generalizes to test prompts unseen during adaptation, reducing the mean squared error between surrogate-predicted and target band gaps by 52% relative to standard reinforcement learning under best-of-ten sampling. An independent band-gap predictor shows consistent gains, while density functional theory calculations further assess selected candidates. Similar gains with Qwen3.5-9B provide cross-backbone evidence. Model Merging recovers competitive two-objective trade-offs from two endpoint experts without additional RL, enabling efficient adaptation and preference control for structured scientific generation.

1 Introduction

Designing materials with targeted properties is critical for fields like batteries (Zhao et al., 2020), catalysis (Zhao et al., 2019), and carbon capture (Osman et al., 2021). While high-throughput screening has identified many novel materials from extensive databases (Curtarolo et al., 2013; Zhong et al., 2020; Merchant et al., 2023), the theoretical space of stable inorganic compounds far exceeds known catalogs. Generative approaches present a compelling opportunity to unlock this latent potential by actively exploring the vast, continuous chemical space far beyond the rigid boundaries of existing datasets.

Generating crystalline materials is a formidable challenge, requiring the concurrent production of continuous attributes (lattice geometry, atomic coordinates) and discrete components (atomic types) while simultaneously capturing intricate structure-property relationships. Recently, denoising models (Cornet et al., 2025; Zeni et al., 2025; Xie et al., 2021; Miller et al., 2024; Jiao et al., 2023, 2024; Yang et al., 2024b; Levy et al., 2025; Höllmer et al., 2025; Yang et al., 2023; Kelvinius et al., 2025) and autoregressive architectures (Antunes et al., 2024; Cao et al., 2024) have achieved notable success in material design. Inspired by their reasoning capabilities in scientific domains (White, 2023; Zhang et al., 2024; Zhao et al., 2025), by representing complex 3D structural topologies as sequential text strings, researchers are increasingly leveraging large language models (LLMs) for this task with remarkable outcomes (Gruver et al., 2024; Miret and Krishnan, 2025; Gan et al., 2025).

Practical material science extends beyond ba-

sic stability, demanding functional materials with targeted properties (e.g., specific band gaps or symmetries). Pre-trained on vast chemical corpora, LLMs utilize natural language to flexibly specify these complex constraints (Ding et al., 2024; Sri-ram et al., 2024). Nonetheless, conditional LLM-based design faces three primary challenges: (1) **Data Scarcity for Functional Targets:** Desired property regions are often underrepresented, making it difficult for supervised learning to generalize to out-of-distribution (OOD) spaces (Zeni et al., 2025). (2) **Physical Validation:** Relying solely on textual patterns without explicit physical feedback often induces generative “hallucinations” of thermodynamically invalid structures. (3) **Dynamic Multi-Objective Adaptation:** Practical design involves constantly shifting priorities. Existing methods require costly re-training for each new objective trade-off, lacking the critical flexibility needed for diverse applications.

To address these challenges, we propose a verifier-grounded alignment framework combining label-free test-time reinforcement learning (TTRL) with model merging. TTRL adapts a shared policy over target prompts using structure-derived rewards for composition, symmetry, and band gap. We define “label-free” adaptation as updating the model without access to ground-truth Crystal Strings or reference structures for the target prompts. We then interpolate the weights of independently adapted symmetry and band-gap experts to support changing objective trade-offs within the evaluated range, without training a separate multi-objective RL (MORL) policy for each interpolation point. Our main MP-20 evaluation follows a partially transductive protocol, with one shared batch-level adaptation run per property rather than independent per-prompt adaptation. A separate study in a held-out upper-tail band-gap region evaluates transfer from OOD-Adapt prompts to disjoint OOD-Test prompts.

Our main contributions are as follows:

- **Verifier-grounded target-prompt adaptation.** We couple autoregressive generation of mixed discrete–continuous Crystal Strings with structure-based rewards, enabling policy adaptation without target reference structures or pre-collected preference pairs.
- **Modular objective alignment.** We combine property-specific TTRL experts through

weight interpolation, providing an alternative to retraining MORL policies for different symmetry–band-gap trade-offs within the evaluated range.

- **Improved conditional generation and OOD transfer.** On MP-20, TTRL improves symmetry control, band-gap accuracy, and compositional fidelity over SFT and training-prompt RL, and generalizes to disjoint upper-tail band-gap test prompts. Cross-backbone comparisons, independent-surrogate evaluation, and statistical analyses support these gains.

2 Related Work

RL for Crystal Material Design Reinforcement Learning (RL) has emerged as a powerful paradigm to effectively steer generative models toward complex, goal-oriented material design. Prior works have applied offline RL to various architectures, including GNNs and diffusion models, to optimize properties (Karpovich et al., 2024; Chen et al., 2025; Govindarajan et al., 2025) or enhance stability (Cao and Wang, 2025). More recently, Xu et al. (2025) employed Direct Preference Optimization (DPO) to align generative frameworks for targeted crystal discovery. Unlike these approaches, which rely on static datasets, supervised targets, or pre-collected preference pairs, our framework evaluates generated scientific sequences with external tools and updates the policy on target-condition prompts.

Test-Time RL TTRL boosts inference performance by updating parameters using verifiable signals rather than ground-truth supervision (Sun et al., 2020; Phan et al., 2025; Du et al., 2025). While successful in mathematical reasoning and code generation via logical verifiers (Hübotter et al., 2024; Zuo et al., 2026; Wen et al., 2026), its application in natural sciences remains underexplored. Our setting differs from ordinary natural-language generation because Crystal Strings mix discrete symbols and continuous numerical values, and fluent outputs may still violate structural or property constraints. We connect this representation to scientific verification and define label-free adaptation specifically as using no target reference structure during policy updates. More broadly, the NLP community has studied LLM-based molecular optimization and feedback-aligned model merging for structured chemical generation (Dey et al., 2025; Gkoumas and Liakata, 2025). Our work extends

this line toward tool-grounded alignment of crystal sequences whose correctness depends on external scientific evaluation.

Multi-Objective Alignment via Model Merging

Aligning LLMs with multifaceted preferences is challenging. Traditional MORL methods, often based on scalarization (Van Moffaert and Nowé, 2014), suffer from severe computational inefficiency, as adapting to any new combination of preference weights necessitates a complete re-training cycle from scratch (Yang et al., 2024a; Wang et al., 2024). Conversely, MM offers a computationally efficient alternative by fusing specialized models in weight space (Ilharco et al., 2022; Rame et al., 2023; Wortsman et al., 2022). This paradigm is further supported by recent theoretical findings on shared weight subspaces (Kaushik et al., 2025). We adopt linear model merging as an established component and study whether independently adapted scientific property experts can recover useful trade-offs. Our evidence is limited to bounded interpolation between symmetry and band-gap experts, rather than general high-dimensional preference composition.

3 Methodology

Our proposed framework is illustrated in Figure 1. This section details our methodology, beginning with the representation used for crystal materials. We then describe our TTRL strategy and the design of verifiable reward functions for property optimization. We conclude by presenting our MM technique for achieving scalable multi-objective material design.

3.1 Material representation

Crystal materials are typically represented using the Crystallographic Information Files (CIFs) format, a standardized text description for crystal structures. However, CIFs often contain extraneous and redundant fields that can hinder language models from capturing key structural information. To mitigate the redundancy of standard CIFs, we adopt the *Crystal String* representation in Crystal-Text-LLM (Gruver et al., 2024), which serializes lattice and atomic data into concise fixed-precision tokens. Although efficient, the default precision settings in Crystal-Text-LLM proved inadequate for capturing fine-grained structural properties (see Appendix B.2). Consequently, we increased the numerical precision to enhance generation fidelity.

For inputs, we utilize the flexible text-conditional template from Crystal-Text-LLM (cf. Table 5).

3.2 Label-Free TTRL

LLMs have demonstrated the capability to generate stable crystal materials (Gruver et al., 2024; Sriram et al., 2024). However, designing functional materials with specific intrinsic attributes remains a challenge (Zeni et al., 2025). To address this, we introduce a Label-Free TTRL strategy. Unlike traditional fine-tuning that relies on static ground-truth labels, TTRL uses verifier feedback to update the policy on target-condition prompts (Phan et al., 2025; Du et al., 2025). Crucially, our approach constructs rewards by extracting properties from generated materials using scientific computing tools. This removes the need for a reference structure for each target prompt. The main protocol adapts once over a batch-level prompt stream; Section 4.4 separately evaluates generalization from OOD-Adapt to instance-disjoint OOD-Test prompts. The overall flow is illustrated in Figure 1 (b).

3.2.1 Policy Optimization Algorithm

We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to fine-tune the LLM. For a conditional prompt x , the policy π_θ samples a group of G outputs $\{y_1, y_2, \dots, y_G\}$. Instead of using a value function critic, GRPO estimates the advantage of each output by normalizing its reward against the group average: $\hat{A}_i = (R(y_i) - \text{mean}(R_{\text{group}})) / \text{std}(R_{\text{group}})$. This critic-free mechanism significantly reduces memory overhead, making it highly efficient for optimizing LLMs. We provide the detailed objective function and hyperparameters in Appendix A.2.

3.2.2 Verifiable Reward Formulation

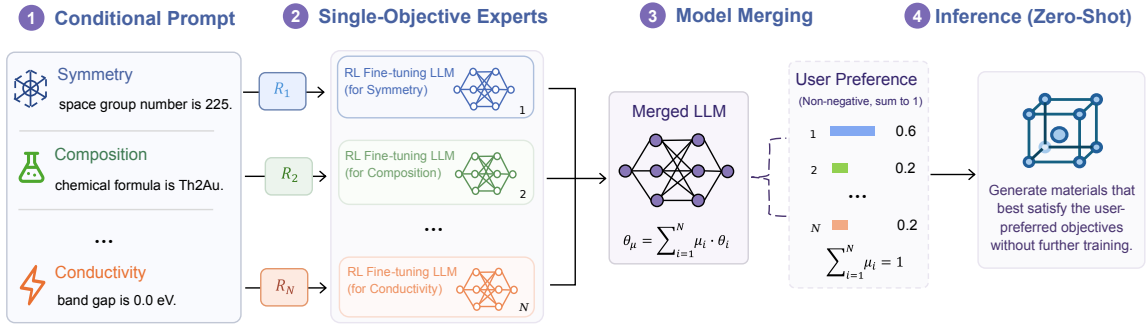
To enable label-free optimization, we design reward functions that quantify the alignment between the generated material’s properties and the target constraints, computed by deterministic analysis tools or learned property surrogates.

Reward for Chemical Composition We use the Jaccard Coefficient (Niwattanakul et al., 2013) to quantify compositional similarity. The reward compares the reduced chemical formula of the generated material (C_{gen}) with the target (C_{tgt}):

$$R_{\text{COMP}}(C_{\text{tgt}}, C_{\text{gen}}) = \frac{\sum_{e \in \mathcal{E}} \min(C_{\text{tgt}}[e], C_{\text{gen}}[e])}{\sum_{e \in \mathcal{E}} \max(C_{\text{tgt}}[e], C_{\text{gen}}[e])}, \quad (1)$$

where $\mathcal{E} = C_{\text{tgt}} \cup C_{\text{gen}}$ represents the set of all unique elements involved. The term $C[e]$ denotes

(a) Multi-objective Crystal Material Design via Model Merging



(b) Label-Free Test-Time RL (TTRL) for Conditional Crystal Design

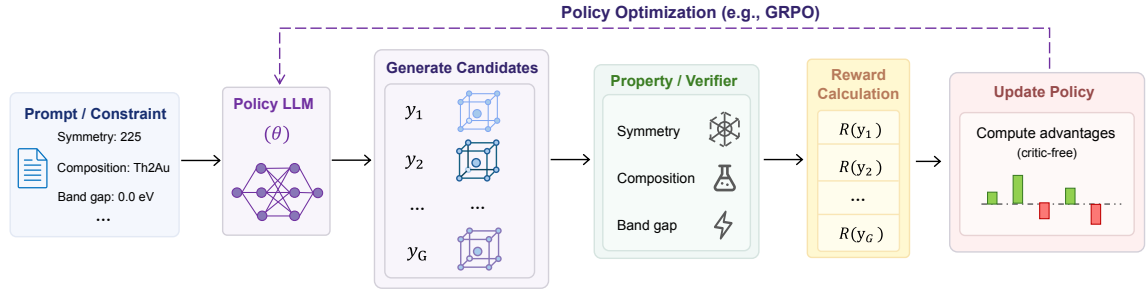


Figure 1: Overview of our proposed framework integrating MM and TTRL. **(a) Multi-Objective Adaptation via Model Merging.** Given a conditional prompt, a Supervised Fine-Tuning (SFT) LLM is separately optimized for symmetry and band gap to create property experts. These experts are composed through bounded weight interpolation ($0 \leq \mu \leq 1$), providing preference points within the evaluated two-objective range without additional RL training. **(b) Label-Free TTRL.** The LLM generates candidate materials, which are parsed and evaluated by external scientific tools. The adaptation is label-free in the specific sense that it does not access a target Crystal String or reference structure for each prompt.

the stoichiometric count of element e in the chemical formula C (e.g., for SiO_2 , $C[\text{O}] = 2$).

Reward for Symmetry Using Pymatgen (Ong et al., 2013), we compute the space group number (spg) of the generated structure. To mitigate reward sparsity, we use synergy (syn) agreement as a training-time shaping signal:

$$R_{\text{SYM}} = \begin{cases} 1 & \text{if } \text{spg} = \text{spg}_{\text{target}} \\ 0.5 & \text{if } \text{syn} = \text{syn}_{\text{target}} \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where spg and syn denote the target constraints. Final evaluation always uses exact space-group matching. An ablation in Appendix D.4 identifies 0.5 as the best empirical partial reward among the tested values, not a theoretically optimal constant.

Reward for Band-Gap Control We prioritize computational speed for on-policy RL by employing MEGNet (Chen et al., 2019)¹, a learned property surrogate, to predict the band gap (bg). The

¹Available from the [MatGL model repository](#).

reward is the clipped negative squared error relative to the target bg:

$$R_{\text{BG}} = \text{clip}(-[\text{bg} - \hat{\text{bg}}]^2, -1, 0). \quad (3)$$

3.3 Bounded Multi-Objective Design via Model Merging

Preventing Reward Hacking with Composite Rewards Optimizing a property reward alone can collapse the policy to a small set of repeated structures. We incorporate the prompt-dependent composition reward (R_{COMP}) as a regularizer that preserves conditional diversity. The composite rewards are:

$$\begin{aligned} R_{\text{symmetry}} &= \lambda_1 R_{\text{SYM}} + R_{\text{COMP}}, \\ R_{\text{bandgap}} &= \lambda_2 R_{\text{BG}} + R_{\text{COMP}}. \end{aligned} \quad (4)$$

MM vs. MORL Material design requires balancing competing objectives. Traditional MORL typically addresses this via linear scalarization ($R = \sum w_i R_i$). However, this approach is inherently rigid as it binds the model to fixed weights

$\{w_i\}$, necessitating costly re-training whenever preferences change. While recent MORL methods such as EMORL (Kong et al., 2025) and GAPO (Li et al., 2025) address efficiency, MM (Rame et al., 2023) provides a modular alternative based on independently trained experts.

Specifically, we independently fine-tune a symmetry expert θ_{sym} and a band-gap expert θ_{bg} using Eq. 4. At inference time, we interpolate them as $\theta_\mu = \mu\theta_{\text{sym}} + (1 - \mu)\theta_{\text{bg}}$, where $0 \leq \mu \leq 1$. This bounded two-expert construction provides intermediate preference points without retraining. We do not evaluate extrapolation or general N -way merging.

4 Experiments

4.1 Experimental Setup

Dataset and Backbones We utilize the MP-20 dataset (Jain et al., 2013), containing 45,231 inorganic crystalline materials, split into 27,136 training, 9,045 validation, and 9,050 test structures following standard protocols. Unless otherwise noted, our SFT, Standard RL, and TTRL results in the main text use LLaMA-2-7B (Touvron et al., 2023). We additionally repeat the High-Precision SFT and TTRL pipeline with Qwen3.5-9B-Base (Qwen Team, 2026); complete cross-backbone results are reported in Appendix D.5. Detailed hyperparameters for SFT, RL, and sampling are provided in Appendix A.2.2.

Main Conditional-Generation Baselines The in-distribution conditional-generation experiments compare the following four models or training configurations:

- **Crystal-Text-LLM (Gruver et al., 2024):** The original LLaMA-2 based model fine-tuned on text-serialized crystal structures which encodes lattice parameters and atomic coordinates as standard text strings, serving as the primary SFT baseline.
- **High-Precision SFT:** An enhanced version of Crystal-Text-LLM trained with our proposed *high-fidelity crystal string representation* (see Section 3.1) to ensure structural precision during generation.
- **MatterGen (Zeni et al., 2025):** A diffusion-based generative model. For an MP-20 comparison, we adopt the official mp_20_base check-

point² and fine-tune it for our conditional tasks using the same data splits.

- **PLaID++ (3D Coord) (Xu et al., 2025):** A preference-aligned generative framework that combines iterative Direct Preference Optimization (DPO) with a symmetry-informed Wyckoff representation. To decouple the alignment algorithm’s performance from the underlying structural representation, we adopt the variant of PLaID++ utilizing 3D coordinates (see Appendix C.2).

Additional Comparisons The held-out adaptation study compares High-Precision SFT, Standard RL, and TTRL to isolate the effect of target-prompt adaptation. The cross-backbone study repeats High-Precision SFT and TTRL with Qwen3.5-9B-Base. The wide-bandgap screening study additionally includes CDVAE (Xie et al., 2021) and LLEMA (Abhyankar et al., 2025); its systems use different backbones and search procedures and are therefore interpreted as screening-pipeline comparisons rather than controlled algorithmic rankings.

RL Evaluation Settings **Standard RL** optimizes on training prompts and evaluates a fixed checkpoint on the test set. For **TTRL**, we train separate symmetry and band-gap policies. Each policy receives one shuffled stream drawn from the 9,050 MP-20 test prompts and is updated for 400 GRPO steps with a batch size of 16, corresponding to 6,400 adaptation prompts. We then evaluate the checkpoint on all 9,050 prompts. Thus, the main protocol is batch-level and partially transductive: evaluation includes prompts both encountered and not encountered during adaptation. No target Crystal Strings or reference structures are used for rewards.

Attribution Controls The comparisons isolate three effects: Crystal-Text-LLM $\rightarrow$ High-Precision SFT measures the end-to-end effect of representation precision under SFT; High-Precision SFT $\rightarrow$ Standard RL measures verifier-guided optimization on training prompts; and High-Precision SFT or Standard RL $\rightarrow$ TTRL measures adaptation on target prompts. The 49.28% to 96.17% result in Appendix B.2 is an encoding–decoding fidelity test, not an end-to-end generation improvement.

²Publicly available at <https://github.com/microsoft/mattergen>

Evaluation Protocol For each prompt, we sample $N = 10$ candidates from the same fixed checkpoint. Average candidate performance (AVE) averages the evaluation metric over all N candidates, whereas Best-of- N (BoN-10; Appendix A.3) uses the task verifier to select one candidate and reports the metric of that selected candidate. We then aggregate the prompt-level results over the evaluation set. The reported mean $\pm$ sample standard deviation is computed over three independent inference seeds. Neither AVE nor BoN-10 performs parameter updates during evaluation.

4.2 Evaluation Metrics and Verification Protocol

We evaluate the generated crystals based on their adherence to three key property constraints:

- **Band Gap:** Measured by the Mean Squared Error (MSE) relative to the target value and the **Hit Rate** (percentage of structures within $[0, 0.01]$ eV for metallic targets).
- **Symmetry:** Measured by the **Match Ratio**, defined as the percentage of generated structures whose space group number exactly matches the target.
- **Chemical Composition:** Measured by the Jaccard **Similarity** between the generated and target reduced chemical formulas.

We additionally report structural and compositional validity, stability, and the Stable–Unique–Novel (SUN) rate to check whether improved property control degrades general crystal quality (Appendix D.7).

Verification Hierarchy Band-gap evaluation follows three levels. First, MEGNet provides the efficient on-policy reward and the primary band-gap metric. Second, ALIGNN, which is not used for policy optimization, independently re-evaluates the held-out experiment. Third, DFT calculations assess ALIGNN-selected candidate pools at higher fidelity. MEGNet’s correlation with DFT ($r = 0.79$; Appendix D.6) supports its use as an informative surrogate but does not guarantee candidate-level ranking or eliminate proxy bias.

4.3 In-Distribution Conditional Generation

We first evaluate single-property symmetry and band-gap control, then examine whether these

gains persist when chemical composition is imposed as an additional constraint. Section 4.4 subsequently tests transfer from OOD-Adapt prompts to instance-disjoint OOD-Test prompts under independent evaluators.

Symmetry Control Table 1 reports symmetry-control results; λ denotes the symmetry reward weight (λ_1 in Eq. 4). At $\lambda = 2.0$, TTRL raises AVE exact matching to 35.77%, versus 26.01% for High-Precision SFT and 29.81% for Standard RL. With composition constraints, it raises BoN matching to 48.91%, versus 42.01% and 42.35%, respectively. Its 1.48-point BoN gain over PLaID++ is modest but statistically supported, while the compound-constraint gains are larger (Appendix D.2). Qwen3.5-9B shows the same trend (Table 17).

Band Gap Optimization Table 2 evaluates band-gap control using MSE and a binary metallic Hit Rate. At $\lambda = 2.0$, TTRL reduces single-condition BoN MSE to 1.074, from 2.462 for High-Precision SFT and 1.840 for Standard RL; its difference from PLaID++ (1.083) is not statistically significant. With composition constraints, TTRL achieves 0.339 MSE and over 99% similarity. Its difference from Standard RL (0.357) is likewise not statistically significant, whereas its gains over High-Precision SFT and PLaID++ are significant (Appendix D.2). Hit Rate emphasizes the dominant near-zero-band-gap region, while MSE captures continuous accuracy. Qwen3.5-9B shows the same trend (Table 18).

Compound Constraints TTRL retains strong composition adherence while improving the target property under compound constraints. At $\lambda = 2.0$, BoN composition similarity reaches 99.40% for Symmetry + Composition and 99.11% for Band Gap + Composition, alongside the best corresponding property results in Tables 1 and 2. This indicates that the target-property gains do not require sacrificing the prompt-specified chemical composition.

4.4 Held-Out Adaptation and Independent Verification

Instance-Disjoint Upper-Tail Split To test generalization beyond both the supervised property range and the prompts used for adaptation, we exclude every MP-20 structure with reference band gap $E_g > 2.5$ eV from SFT, Standard RL, vali-

Table 1: Results for conditional crystal generation with symmetry constraints. **Standard RL** denotes optimization on training prompts, while **TTRL-Symmetry** denotes target-prompt adaptation. The super-columns **Symmetry** and **Symmetry + Composition** denote the constraints in the input prompts: single-condition and multi-condition generation, respectively. λ denotes the weight of the symmetry reward. All results are reported as mean $\pm$ sample standard deviation over three inference seeds. Results of **BoN sampling** are highlighted with **blue background**.

Method	RL Params (λ)	Symmetry		Symmetry + Composition			
		Match Ratio (%) $\uparrow$		Match Ratio (%) $\uparrow$		Similarity (%) $\uparrow$	
Crystal-Text-LLM	—	23.08 $\pm$ 0.41	48.92 $\pm$ 0.46	25.17 $\pm$ 0.44	39.38 $\pm$ 0.61	96.87 $\pm$ 0.17	98.95 $\pm$ 0.13
High-Precision SFT	—	26.01 $\pm$ 0.38	51.95 $\pm$ 0.44	29.92 $\pm$ 0.45	42.01 $\pm$ 0.68	95.38 $\pm$ 0.15	98.07 $\pm$ 0.12
MatterGen (Fine-tuned)	—	24.92 $\pm$ 0.34	35.98 $\pm$ 0.39	24.33 $\pm$ 0.40	29.45 $\pm$ 0.48	95.85 $\pm$ 0.14	96.87 $\pm$ 0.12
PLaID++ (3D Coord-SG DPO)	—	32.40 $\pm$ 0.29	53.19 $\pm$ 0.35	31.70 $\pm$ 0.38	39.82 $\pm$ 0.52	96.17 $\pm$ 0.10	96.99 $\pm$ 0.11
Standard RL-Symmetry	1.0	26.53 $\pm$ 0.35	51.06 $\pm$ 0.37	29.76 $\pm$ 0.41	41.16 $\pm$ 0.50	97.23 $\pm$ 0.13	99.20 $\pm$ 0.09
	2.0	29.81 $\pm$ 0.31	52.43 $\pm$ 0.34	31.17 $\pm$ 0.42	42.35 $\pm$ 0.55	94.77 $\pm$ 0.16	99.04 $\pm$ 0.12
	4.0	17.66 $\pm$ 0.62	29.13 $\pm$ 0.71	28.71 $\pm$ 0.74	29.15 $\pm$ 0.83	90.09 $\pm$ 0.42	26.50 $\pm$ 1.15
TTRL-Symmetry	1.0	28.78 $\pm$ 0.28	51.88 $\pm$ 0.31	30.22 $\pm$ 0.37	44.72 $\pm$ 0.43	98.69 $\pm$ 0.09	99.44 $\pm$ 0.07
	2.0	35.77 $\pm$ 0.20	54.67 $\pm$ 0.27	35.58 $\pm$ 0.31	48.91 $\pm$ 0.36	95.40 $\pm$ 0.18	99.40 $\pm$ 0.11
	4.0	19.47 $\pm$ 0.58	32.03 $\pm$ 0.66	28.03 $\pm$ 0.69	28.93 $\pm$ 0.91	91.01 $\pm$ 0.38	27.38 $\pm$ 1.02

Table 2: Results for conditional crystal generation with band-gap constraints. **Standard RL** denotes optimization on training prompts, while **TTRL-Bandgap** denotes target-prompt adaptation. The super-columns **Band Gap** and **Band Gap + Composition** denote the constraints in the input prompts: single-condition and multi-condition generation, respectively. λ denotes the weight of the band gap reward. All results are reported as mean $\pm$ sample standard deviation over three inference seeds. Results of **BoN sampling** are highlighted with **blue background**.

Method	RL Params	Sampling	Band Gap		Band Gap + Composition		
			MSE $\downarrow$	Hit Rate (%) $\uparrow$	MSE $\downarrow$	Hit Rate (%) $\uparrow$	Similarity (%) $\uparrow$
Crystal-Text-LLM	—	AVE	6.335 $\pm$ 0.091	81.70 $\pm$ 0.58	3.539 $\pm$ 0.067	77.10 $\pm$ 0.73	96.37 $\pm$ 0.21
		BoN	2.785 $\pm$ 0.073	99.98 $\pm$ 0.02	0.596 $\pm$ 0.014	95.38 $\pm$ 0.41	95.71 $\pm$ 0.28
High-Precision SFT	—	AVE	6.182 $\pm$ 0.084	79.85 $\pm$ 0.62	3.587 $\pm$ 0.061	76.23 $\pm$ 0.71	94.65 $\pm$ 0.24
		BoN	2.462 $\pm$ 0.061	99.94 $\pm$ 0.03	0.411 $\pm$ 0.009	96.01 $\pm$ 0.38	93.91 $\pm$ 0.31
MatterGen (Fine-tuned)	—	AVE	5.459 $\pm$ 0.079	79.17 $\pm$ 0.64	2.507 $\pm$ 0.052	75.44 $\pm$ 0.78	96.02 $\pm$ 0.20
		BoN	1.682 $\pm$ 0.036	97.03 $\pm$ 0.21	0.409 $\pm$ 0.010	89.85 $\pm$ 0.55	97.32 $\pm$ 0.18
PLaID++ (3D Coord-BG DPO)	—	AVE	6.621 $\pm$ 0.102	85.23 $\pm$ 0.55	4.125 $\pm$ 0.074	80.72 $\pm$ 0.66	96.88 $\pm$ 0.19
		BoN	1.083 $\pm$ 0.018	99.12 $\pm$ 0.17	0.917 $\pm$ 0.028	97.43 $\pm$ 0.32	95.46 $\pm$ 0.26
Standard RL-Bandgap	1.0	AVE	6.443 $\pm$ 0.088	62.07 $\pm$ 0.79	3.628 $\pm$ 0.058	57.16 $\pm$ 0.93	95.07 $\pm$ 0.22
		BoN	2.501 $\pm$ 0.056	69.56 $\pm$ 0.71	0.420 $\pm$ 0.011	80.20 $\pm$ 0.58	99.06 $\pm$ 0.09
	2.0	AVE	5.132 $\pm$ 0.071	31.07 $\pm$ 0.91	2.499 $\pm$ 0.048	50.77 $\pm$ 1.05	98.22 $\pm$ 0.16
		BoN	1.840 $\pm$ 0.047	96.03 $\pm$ 0.28	0.357 $\pm$ 0.008	93.04 $\pm$ 0.47	99.05 $\pm$ 0.08
TTRL-Bandgap	1.0	AVE	6.135 $\pm$ 0.076	64.41 $\pm$ 0.72	3.895 $\pm$ 0.063	55.30 $\pm$ 0.88	99.07 $\pm$ 0.14
		BoN	2.653 $\pm$ 0.049	75.40 $\pm$ 0.64	0.482 $\pm$ 0.012	86.12 $\pm$ 0.50	97.34 $\pm$ 0.19
	2.0	AVE	4.797 $\pm$ 0.065	30.68 $\pm$ 0.84	2.285 $\pm$ 0.041	46.59 $\pm$ 0.96	99.12 $\pm$ 0.12
		BoN	1.074 $\pm$ 0.015	97.39 $\pm$ 0.24	0.339 $\pm$ 0.007	94.83 $\pm$ 0.41	99.11 $\pm$ 0.07

dation, hyperparameter tuning, and checkpoint selection. The resulting upper-tail region is split by material instance into OOD-Adapt (4,017 structures) and OOD-Test (1,722 structures). TTRL observes only conditional prompts from OOD-Adapt, without their target structures, and is frozen before evaluation on OOD-Test. Reduced formulas may overlap, so this experiment establishes transfer to unseen structures and prompts within a held-out property region, not to unseen compositions or arbitrary distribution shifts.

Independent-Evaluator and DFT Evidence Table 3 shows that TTRL transfers from OOD-Adapt to disjoint OOD-Test prompts and preserves the method ordering under ALIGNN, which is not used as the reward. TTRL without BoN also improves over High-Precision SFT with BoN-10 (MEGNet

MSE 1.92 vs. 2.40), separating parameter adaptation from candidate selection. In an ALIGNN-guided DFT screening pipeline, 100 selected valid and unique candidates per BoN method yield DFT MAE values of 1.28, 1.10, and 0.72 eV for High-Precision SFT, Standard RL, and TTRL, respectively. These selected pools provide higher-fidelity corroboration but are not unbiased samples of the complete generation distributions (Appendix D.1).

4.5 Robustness and Generality

Failure Modes and Reward Design A first-failure analysis identifies exact space-group mismatch, rather than parsing, validity, or composition errors, as the dominant failure mode (Appendix D.3). Controlled ablations further show that the composition reward mitigates reward-driven mode collapse, while moderate synergy shaping

Table 3: Target-property performance on OOD-Test. Target tolerance denotes $|\hat{E}_g - E_g| \leq 0.5$ eV. All results are reported as mean $\pm$ sample standard deviation over three inference seeds. Results of **BoN sampling** are highlighted with **blue background**.

Method	Evaluation	Band-Gap Error		Target Satisfaction	
		MEGNet MSE $\downarrow$	ALIGNN MSE $\downarrow$	Tolerance (%) $\uparrow$	$E_g > 2.5$ eV (%) $\uparrow$
High-Precision SFT	AVE	6.10 $\pm$ 0.14	6.55 $\pm$ 0.17	20.8 $\pm$ 0.76	38.5 $\pm$ 0.93
	BoN	2.40 $\pm$ 0.08	2.75 $\pm$ 0.10	44.6 $\pm$ 1.04	66.2 $\pm$ 1.18
Standard RL	AVE	4.85 $\pm$ 0.11	5.20 $\pm$ 0.14	27.5 $\pm$ 0.82	45.8 $\pm$ 1.06
	BoN	1.85 $\pm$ 0.06	2.10 $\pm$ 0.08	52.7 $\pm$ 0.94	72.9 $\pm$ 0.97
TTRL	AVE	1.92 $\pm$ 0.05	2.25 $\pm$ 0.07	49.8 $\pm$ 0.88	68.6 $\pm$ 0.84
	BoN	0.88 $\pm$ 0.03	1.08 $\pm$ 0.04	71.4 $\pm$ 0.73	86.3 $\pm$ 0.61

improves exact space-group control over exact-only training (Appendix D.4).

Structural Quality On LLaMA-2-7B, TTRL preserves 100% structural validity and a Stability Rate close to High-Precision SFT (6.18% vs. 6.24%), while increasing the SUN rate from 4.03% to 4.48%. It also reduces the Wasserstein distances for density and number of elements from 0.101 and 0.025 to 0.033 and 0.019, respectively (Table 20). These proxy results show unchanged structural validity and similar stability, although the absolute stability rates remain limited and require higher-fidelity validation.

Cross-Backbone Evidence With Qwen3.5-9B, TTRL improves Symmetry + Composition BoN matching from 43.26% to 52.38%, reduces Band Gap + Composition BoN MSE from 0.386 to 0.286, and raises the SUN rate from 4.28% to 5.08% (Tables 17–19). This supports effectiveness beyond LLaMA-2-7B, while results from two model families do not establish universal architecture independence.

4.6 Bounded Multi-Objective Design via Model Merging

We evaluate bounded interpolation between symmetry and band-gap experts while retaining composition as a regularizing constraint. We compare MM against separately trained MORL baselines at the same evaluated weights. Detailed numerical results are provided in Appendix D.8. It is worth noting that across all settings, the chemical composition similarity remains consistently high ($> 95\%$), acting as a satisfied constraint. Thus, we focus our analysis on the trade-off between symmetry and band gap.

Pareto Optimality: MM vs. MORL Figure 2 visualizes the Pareto fronts formed by these two methods. The Optimization Direction arrow indicates that the ideal model should achieve low MSE

and high Match Ratio (bottom-right corner). The MM front is competitive with the MORL front and is better at two evaluated transition points. For instance, at the balanced interpolation point ($\mu = 0.5$), MM achieves a lower Band Gap MSE (2.440 vs. 2.508) and a higher Symmetry Match Ratio (27.31% vs. 26.83%) compared to the MORL counterpart ($\lambda = 0.5$). This provides a proof of concept that the two adapted endpoints admit useful bounded interpolation.

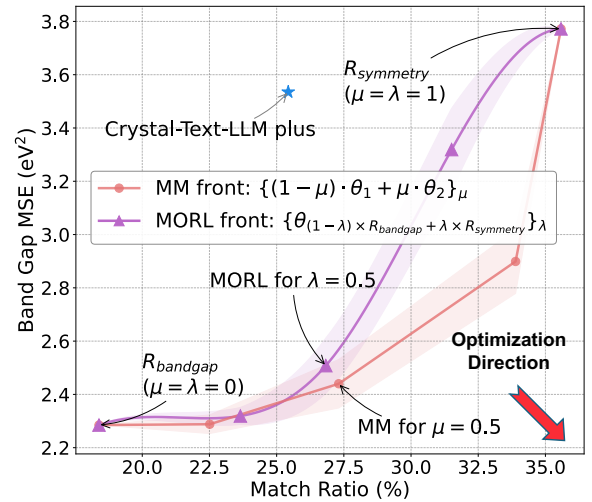


Figure 2: Comparison of Pareto fronts for simultaneous symmetry and band gap optimization. The **blue star** represents the initial SFT baseline. The **purple triangles** represent discrete MORL models trained with varying reward weights (λ). The **red circles** represent MM models obtained by interpolating specialist model weights (μ). MM is competitive with MORL in the evaluated bounded two-objective interpolation range while using only two trained endpoints.

Zero-Shot Adaptation Efficiency MM offers a distinct advantage in computational cost. Generating the MORL Pareto front requires re-training a separate model for every specific weight combination (λ). For instance, fine-tuning a single policy requires approximately 3 hours on 8 NVIDIA A800

Table 4: Surrogate-predicted hit rate in wide-bandgap screening. A hit is not an experimentally or DFT-certified material. LLEMA uses GPT-4o-mini with iterative refinement.

Method	Hit Rate (%) $\uparrow$
CDVAE	0.04
MatterGen	6.45
Crystal-Text-LLM	1.21
High-Precision SFT	1.28
TTRL-Bandgap (Ours)	8.67
LLEMA (GPT-4o-mini)	21.62

GPUs. Constructing a fine-grained Pareto front (e.g., 10 distinct preference points) with MORL would thus demand roughly 240 GPU-hours. In contrast, MM involves training single-property experts only once (the endpoints in Figure 2). Any tested intermediate preference ($\mu \in (0, 1)$) is obtained by linear weight interpolation without an additional RL run. These experiments do not evaluate extrapolation or higher-dimensional N -way merging.

4.7 Wide-Bandgap Screening

We evaluate the framework in a realistic computational screening scenario for wide-bandgap semiconductors, which are important for power and optoelectronic devices (Mashiko et al., 2016). This study uses learned property predictors and does not constitute experimental synthesis or first-principles certification.

Experimental Setup and Criteria Following Abhyankar et al. (2025), a generated candidate is counted as a screening “Hit” when the predictors jointly indicate: (1) band gap ≥ 2.5 eV via CGCNN (Xie and Grossman, 2018); and (2) formation energy ≤ -1.0 eV/atom and energy above hull ≤ 0.1 eV/atom via ALIGNN (Choudhary and DeCost, 2021). We compare *TTRL-Bandgap* with CDVAE (Xie et al., 2021), MatterGen (Zeni et al., 2025), and LLEMA (Abhyankar et al., 2025).

Comparison with Baselines As shown in Table 4, TTRL-Bandgap yields an 8.67% surrogate-predicted Hit Rate, compared with 0.04% for CDVAE, 6.45% for MatterGen, and approximately 1.2% for the SFT baselines. The result indicates improved candidate-pool quality under this screening pipeline, rather than certified stability or synthesizability.

Efficiency vs. Closed-Source Agents LLEMA obtains 21.62% using GPT-4o-mini with iterative refinement, whereas our 7B checkpoint performs one-shot candidate generation. Because the systems use different backbones and search procedures, this comparison primarily contextualizes screening throughput rather than establishing a controlled algorithmic ranking.

Overall, TTRL improves symmetry and band-gap control on the in-distribution tasks while retaining composition adherence and general structural quality. The instance-disjoint upper-tail experiment further shows transfer from OOD-Adapt to OOD-Test prompts, with the improvement preserved by ALIGNN and supported by DFT screening of selected candidates. Finally, model merging recovers competitive bounded two-objective trade-offs from two endpoint experts without additional RL training for intermediate preferences, while the wide-bandgap study illustrates the resulting candidate-screening utility.

5 Conclusions

We presented a verifier-grounded alignment framework for autoregressive crystal generation. TTRL updates an LLM on target-prompt streams using feedback computed from generated structures, without target reference structures, while bounded interpolation between symmetry and band-gap experts provides preference control without retraining every intermediate MORL policy. Controlled MP-20 experiments separate representation, training-prompt RL, and test-prompt adaptation effects. A held-out upper-tail band-gap study further shows transfer from OOD-Adapt to disjoint OOD-Test prompts, and the method ordering persists under an independent ALIGNN evaluator. These results support verifier-guided adaptation for the evaluated discrete and continuous constraints, but do not establish arbitrary distribution-shift generalization, high-dimensional model merging, or experimental realizability. The wide-bandgap study should therefore be viewed as surrogate-based computational screening that produces candidates for subsequent DFT calculations or laboratory validation.

Limitations

Adaptation and Generalization The main MP-20 protocol adapts once on a test-prompt stream and evaluates on the full test set, making it partially transductive rather than independent per-request

adaptation. Although the OOD-Adapt/OOD-Test split removes prompt-instance overlap, it covers only an upper-tail band-gap shift, may contain overlapping reduced formulas, and does not establish transfer to unseen compositions, other shifts, or datasets beyond MP-20.

Verification MEGNet and ALIGNN are learned surrogates rather than DFT-level or experimental oracles. Their correlations provide useful signals but do not guarantee candidate ranking, and our DFT analyses use selected candidate pools rather than unbiased distribution samples. Generated candidates therefore require further first-principles and experimental validation.

Methodological Scope Model Merging is evaluated only for bounded linear interpolation between symmetry and band-gap experts; extrapolation, nonlinear preferences, and high-dimensional merging remain untested. Results from LLaMA-2-7B and Qwen3.5-9B also do not establish architecture- or scale-independent effectiveness. Moreover, uncertainty over three inference seeds excludes variance across independently trained checkpoints.

Compute Cost Each property-specific adaptation requires approximately 24 A800 GPU-hours, excluding verifier and BoN inference. Although amortized over 6,400 prompts, this remains costly for small or latency-sensitive deployments. Future work should investigate lighter updates, efficient high-fidelity rewards, broader datasets and objectives, and independently trained checkpoints.

Acknowledgments

This work was supported by the National Science and Technology Major Project (2023ZD0120703), the China NSFC Projects (62120106006, 92370206, 62576212, U23B2057), and the Frontier Technologies R&D Program of Jiangsu (BF2025061).

References

Nikhil Abhyankar, Sanchit Kabra, Saaketh Desai, and Chandan K Reddy. 2025. Accelerating materials design via llm-guided evolutionary search. *arXiv e-prints*, pages arXiv–2510.

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

Siddharth Betala, Samuel P Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Cl  mentine Fourrier, Nikita Kazeev, Chaitanya K Joshi, S  kou-Oumar Kaba, and 1 others. 2025. Lemat-genbench: A unified evaluation framework for crystal generative models. *arXiv preprint arXiv:2512.04562*.

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

Zhendong Cao and Lei Wang. 2025. Crystalformer-rl: Reinforcement fine-tuning for materials design. *arXiv preprint arXiv:2504.02367*.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. 2019. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572.

Junwu Chen, Jeff Guo, and Philippe Schwaller. 2025. Matinvent: Reinforcement learning for 3d crystal diffusion generation. In *AI for Accelerated Materials Design-ICLR 2025*.

Kamal Choudhary and Brian DeCost. 2021. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185.

Fran  ois Cornet, Federico Bergamin, Arghya Bhowmik, Juan Maria Garcia Lastra, Jes Frellsen, and Mikkel N Schmidt. 2025. Kinetic langevin diffusion for crystalline materials generation. *arXiv preprint arXiv:2507.03602*.

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

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. 2023. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041.

Vishal Dey, Xiao Hu, and Xia Ning. 2025. [GeLLM³O: Generalizing large language models for multi-property molecule optimization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25192–25221. Association for Computational Linguistics.

Qianggang Ding, Santiago Miret, and Bang Liu. 2024. Matexpert: Decomposing materials discovery by mimicking human experts. *arXiv preprint arXiv:2410.21317*.

Yong Du, Yuchen Yan, Fei Tang, Zhengxi Lu, Chang Zong, Weiming Lu, Shengpei Jiang, and Yongliang Shen. 2025. Test-time reinforcement learning for gui

grounding via region consistency. *arXiv preprint arXiv:2508.05615*.

Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Daniel Schwalbe-Koda, Carla P Gomes, Kristin Persson, and Wei Wang. 2025. Large language models are innate crystal structure generators. In *AI for Accelerated Materials Design-ICLR 2025*.

Dimitris Gkoumas and Maria Liakata. 2025. [Less for more: Enhanced feedback-aligned mixed LLMs for molecule caption generation and fine-grained NLI evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2882–2902. Association for Computational Linguistics.

Prashant Govindarajan, Mathieu Reymond, Antoine Clavaud, Mariano Phielipp, Santiago Miret, and Sarath Chandar. 2025. Crystalgym: A new benchmark for materials discovery using reinforcement learning. In *AI for Accelerated Materials Design-ICLR 2025*.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. 2024. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*.

Philipp Höllmer, Thomas Egg, Maya M Martirosyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, and 1 others. 2025. Open materials generation with stochastic interpolants. *arXiv preprint arXiv:2502.02582*.

Jonas Hübötter, Sascha Bongni, Ido Hakimi, and Andreas Krause. 2024. Efficiently learning at test-time: Active fine-tuning of llms. *arXiv preprint arXiv:2410.08020*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

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

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

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. 2024. Space group constrained crystal generation. In *International Conference on Learning Representations*, volume 2024, pages 6836–6853.

Christopher Karpovich, Elton Pan, and Elsa A Olivetti. 2024. Deep reinforcement learning for inverse inorganic materials design. *npj Computational Materials*, 10(1):287.

Prakhar Kaushik, Shravan Chaudhari, Ankit Vaidya, Rama Chellappa, and Alan Yuille. 2025. The universal weight subspace hypothesis. *arXiv preprint arXiv:2512.05117*.

Filip Ekström Kelvinius, Oskar B Andersson, Abhijith S Parackal, Dong Qian, Rickard Armiento, and Fredrik Lindsten. 2025. Wyckoffdiff—a generative diffusion model for crystal symmetry. *arXiv preprint arXiv:2502.06485*.

Scott Kirklin, Bryce Meredig, and Chris Wolverton. 2013. High-throughput computational screening of new li-ion battery anode materials. *Advanced Energy Materials*, 3(2):252–262.

Lingxiao Kong, Cong Yang, Susanne Neufang, Oya Deniz Beyan, and Zeyd Boukhers. 2025. Emorl: Ensemble multi-objective reinforcement learning for efficient and flexible llm fine-tuning. *arXiv preprint arXiv:2505.02579*.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. 2025. Symmcd: Symmetry-preserving crystal generation with diffusion models. *arXiv preprint arXiv:2502.03638*.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. 2025. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. *arXiv preprint arXiv:2507.01915*.

Hiroki Mashiko, Katsuya Oguri, Tomohiko Yamaguchi, Akira Suda, and Hideki Gotoh. 2016. Petahertz optical drive with wide-bandgap semiconductor. *Nature physics*, 12(8):741–745.

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

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

Santiago Miret and NM Anoop Krishnan. 2025. Enabling large language models for real-world materials discovery. *Nature Machine Intelligence*, 7(7):991–998.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multicongress of engineers and computer scientists*, volume 1, pages 380–384.

- 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. 2013. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319.
- Ahmed I Osman, Mahmoud Hefny, MIA Abdel Maksoud, Ahmed M Elgarahy, and David W Rooney. 2021. Recent advances in carbon capture storage and utilisation technologies: a review. *Environmental Chemistry Letters*, 19(2):797–849.
- John P Perdew, Kieron Burke, and Matthias Ernzerhof. 1996. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865.
- Peter Phan, Dhruv Agarwal, Kavitha Srinivas, Horst Samulowitz, Pavan Kapanipathi, and Andrew McCallum. 2025. Migrate: Mixed-policy grpo for adaptation at test-time. *arXiv preprint arXiv:2508.08641*.
- Qwen Team. 2026. Qwen3.5-9B-Base. <https://huggingface.co/Qwen/Qwen3.5-9B-Base>. Model card, accessed 2026-08-27.
- Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Anuroop Sriram, Benjamin K Miller, Ricky T Chen, and Brandon M Wood. 2024. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046.
- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Kristof Van Moffaert and Ann Nowé. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512.
- Haoliang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*.
- Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, and 1 others. 2026. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. In *International Conference on Learning Representations*, volume 2026, pages 49450–49483.
- Andrew D White. 2023. The future of chemistry is language. *Nature Reviews Chemistry*, 7(7):457–458.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. 2021. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*.
- Tian Xie and Jeffrey C Grossman. 2018. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301.
- Andy Xu, Rohan Desai, Larry Wang, Gabriel Hope, and Ethan Ritz. 2025. Plaid++: A preference aligned language model for targeted inorganic materials design. *arXiv preprint arXiv:2509.07150*.
- Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024a. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.
- Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander Gaunt, Brendan C McMorro, Danilo Jimenez Rezende, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. 2024b. Generative hierarchical materials search. *Advances in Neural Information Processing Systems*, 37:38799–38819.
- Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. 2023. Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*.
- Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien H. Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, and 7 others. 2025. A

generative model for inorganic materials design. *Nature*, 639(8055):624–632.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, and 1 others. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

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

Zhi-Jian Zhao, Sihang Liu, Shenjun Zha, Dongfang Cheng, Felix Studt, Graeme Henkelman, and Jinlong Gong. 2019. Theory-guided design of catalytic materials using scaling relationships and reactivity descriptors. *Nature Reviews Materials*, 4(12):792–804.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, and 1 others. 2025. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4).

Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, and 1 others. 2020. Accelerated discovery of co2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, and 1 others. 2026. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483.

Table of Contents for Appendices

Below is a summary of the contents provided in the appendices:

- **Appendix A: Methodology and Implementation Details**

- A.1 Prompt Template and Structure Representation

- A.2 GRPO Formulation and Experimental Details

- A.3 Best-of-N (BoN) Sampling Strategy

- **Appendix B: Data Analysis and Representation**

- B.1 Statistics of Band Gap in MP-20

- B.2 Representation Fidelity

- **Appendix C: Experimental Setup and Evaluation Protocols**

- C.1 Pre-Relaxation Evaluation Protocol

- C.2 Configuration of PLAIID++ Baselines

- **Appendix D: Additional Results and Verifications**

- D.1 Held-Out Upper-Tail Adaptation

- D.2 Statistical Robustness

- D.3 Failure Analysis

- D.4 Reward Ablations

- D.5 Cross-Backbone Evaluation

- D.6 Validation against Ground-Truth DFT

- D.7 Basic Evaluation: Proxy Metrics and Stability

- D.8 Experimental Results of Multi-Objective Design

- D.9 Sensitivity Analysis of Reward Coefficients

A Methodology and Implementation Details

A.1 Prompt template and structure representation

Table 5 illustrates the prompt template used for conditional generation and the corresponding output structure (Crystal String). The input prompt consists of a task instruction, conditional constraints (e.g., chemical formula, space group, band gap), and a generation trigger. The output represents the crystal structure as a sequence of lattice parameters followed by atomic species and their fractional coordinates.

A.2 GRPO Formulation and Experimental Details

This appendix provides the mathematical formulation of the Group Relative Policy Optimization (GRPO) algorithm omitted from the main text for brevity, followed by the detailed hyperparameter settings used in our experiments.

A.2.1 GRPO Objective Formulation

Unlike standard PPO, which utilizes a value function critic, GRPO estimates the baseline directly from the group average of rewards. Given a conditional generation prompt x , the LLM π_θ autoregressively generates a crystal string y , consisting of T tokens $\{y_1, y_2, \dots, y_T\}$. We use RL to optimize the policy to maximize the expected reward from a verifier:

$$\max_{\theta} J(\theta) := \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(x)} [R(y)], \quad (5)$$

Generation Prompt	Crystal string
<s> Below is a description of a bulk material.	6.09 6.09 6.09
The chemical formula is	104.3 104.3 120.3
Th2Au.	Th
The spacegroup number is 225.	0.1574 0.6574 0.8148
The band gap is 0.0 eV.	Th
...	0.8426 0.3426 0.1852
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:	Th
	0.3426 0.1574 0.5000
	Th
	0.6574 0.8426 0.5000
	Au
[Crystal string] </s>	0.2500 0.2500 0.0000
	Au
	0.7500 0.7500 0.0000

Table 5: Prompt template and output representation. The **blue text** indicates the conditional constraints (omitted for unconditional generation). The **purple text** represents the target crystal string, comprising lattice parameters (lengths and angles) and atomic sites (element type and fractional coordinates).

where $\mathcal{D}$ represents the data distribution and $R(y)$ is the reward function.

For each prompt x , we sample a group of G responses $\{y^1, y^2, \dots, y^G\}$ and estimate the advantage for the i -th response as:

$$\hat{A}_i = \frac{R(y^i) - \frac{1}{G} \sum_{j=1}^G R(y^j)}{\text{std}\left(\{R(y^j)\}_{j=1}^G\right)}. \quad (6)$$

Note that the advantage is computed at the sequence level and assigned to each token. The GRPO objective is then formulated as:

$$J(\theta) = \mathbb{E}_{x \sim \mathcal{D}} \frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=1}^T \left\{ \min \left[\rho_{i,t} \hat{A}_i, \text{clip}(\rho_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_i \right] - \beta \mathbb{D}_{KL}(\pi_\theta || \pi_{ref}) \right\}, \quad (7)$$

where $\rho_{i,t} = \frac{\pi_\theta(y_t^i | x, y_{<t}^i)}{\pi_{old}(y_t^i | x, y_{<t}^i)}$ is the importance sampling ratio, ϵ is the clipping threshold, and β is the KL penalty coefficient.

To ensure stability during training, the KL divergence term is approximated using the estimator:

$$\mathbb{D}_{KL}(\pi_\theta || \pi_{ref}) \approx r_t - \log r_t - 1, \quad (8)$$

where $r_t = \frac{\pi_{ref}(y_t^i | x, y_{<t}^i)}{\pi_\theta(y_t^i | x, y_{<t}^i)}$ represents the probability ratio between the reference model and the current policy.

A.2.2 Implementation Details and Computational Budget

We implemented our framework using the `verl` library. All experiments were conducted on a compute node equipped with NVIDIA A800 (80GB) GPUs. Each Standard RL or TTRL run uses 400 optimization steps and takes approximately 3 hours on 8 GPUs, or 24 GPU-hours per property-specific model. For TTRL, a batch size of 16 corresponds to 6,400 adaptation prompts. This is not independent single-prompt latency because prompts and rollouts are processed in parallel. Symmetry and band-gap runs have similar end-to-end time; MEG-Net rewards are evaluated asynchronously with a 40-thread pool.

TTRL is not trained independently for each test prompt. For each objective, we shuffle the complete set of 9,050 MP-20 test prompts and perform one batch-level adaptation run. The resulting checkpoint is evaluated on all 9,050 prompts, including prompts both seen and unseen during adaptation. We therefore describe this setting as partially transductive test-time adaptation. The explicitly held-out protocol in Appendix D.1 removes this overlap by adapting on OOD-Adapt and freezing the checkpoint before evaluation on OOD-Test. Detailed hyperparameters for the supervised fine-tuning (SFT), RL, and inference stages are summarized in Table 6. To ensure a fair and rigorous comparison, we strictly maintained identical RL hyperparameters (e.g., learning rate, KL penalty coefficient) across both the Standard RL (training set) and the Label-Free TTRL (test set) experiments.

A.3 Best-of-N (BoN) Sampling Strategy

Drawing inspiration from high-throughput materials screening (Kirklin et al., 2013), which discovers functional materials by filtering large candidate pools, we adapt the Best-of-N (BoN) strategy to our generative framework. Recent studies have demonstrated that LLMs can generate superior structural distributions when coupled with iterative selection mechanisms (Sriram et al., 2024). In our conditional crystal material design, BoN serves as a versatile inference-time optimization method that allows users to customize selection criteria without retraining the model.

The procedure is defined as follows:

1. **Candidate Generation:** For a given prompt x , the LLM policy π_θ generates N candidate structures $\{y_1, y_2, \dots, y_N\}$.

Table 6: Detailed experimental settings and hyperparameters.

Parameter	Value
<i>SFT Training</i>	
Base Model	LLaMA-2-7B
Epochs	65
Batch Size	32
Initial Learning Rate	1e-4
LR Scheduler	Cosine
LoRA Rank	8
LoRA Alpha	32
LoRA Dropout	0.05
<i>RL Training (Both Standard RL & TTRL)</i>	
Implementation	ver1 v0.5
KL Penalty Coeff. (β)	1e-3
Initial Learning Rate	1e-5
LoRA Rank	32
LoRA Alpha	32
Training Batch Size	16
Target Modules	All Linear
Training Steps	400
Adaptation Prompts (TTRL)	6,400
Group Size (G)	40
Rollout Batch Size	64
Rollout Temperature	0.7
Rollout Nucleus Sampling p	0.9
<i>Inference & Sampling</i>	
Temperature	0.7
Nucleus Sampling p	0.9
Best-of-N (N)	10
Evaluation Prompts	9,050
Hardware / Wall Time	8 A800 / ~ 3 hours

2. **Verifiable Evaluation:** Each candidate y_i is evaluated using a task reward $R(y_i, x)$ (e.g., a symmetry checker or learned property surrogate).
3. **Selection:** The optimal candidate y^* is selected to maximize the reward:

$$y^* = \arg \max_{y_i \in \{y_1, \dots, y_N\}} R(y_i, x) \quad (9)$$

Specific Settings in Experiments We apply BoN with specific reward definitions for different tasks:

- **Symmetry Optimization (Table 1):** The reward is binary, indicating whether the structure’s space group (verified by PyMatgen) matches the target. In cases where multiple candidates satisfy the constraint (multiple matches), we select the first valid candidate to report performance.

- **Band Gap Optimization (Table 2):** The reward is defined as the negative Mean Squared Error (MSE) between the target band gap and the value predicted by the proxy model (MEG-Net). The candidate with the minimum MSE is selected as the final output.

B Data Analysis and Representation

B.1 Statistics of Band Gap in MP-20

Upon analyzing the band gap distribution within the MP-20 dataset, we observed a significant data imbalance: more than half of the entries possess a band gap at or near 0 eV. Specifically, 62.08% of the training set and 62.24% of the test set consist of materials with a band gap below 0.01 eV (dashed red line in Figure 3). We attribute this to the high prevalence of metallic compounds in the dataset. This long-tail distribution poses a challenge for Supervised Fine-Tuning, as models tend to overfit to the dominant zero-bandgap mode, necessitating the exploration capabilities of Reinforcement Learning.

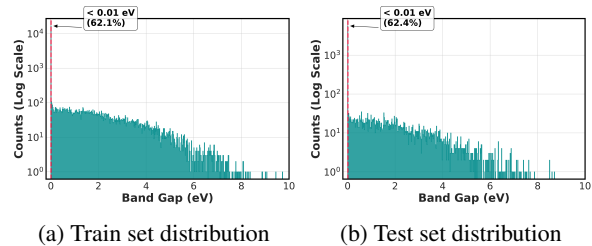


Figure 3: Histograms of band gap values in MP-20, plotted on a logarithmic scale. The red dashed line marks the 0.01 eV threshold.

B.2 Representation Fidelity

We conducted representation fidelity experiments to evaluate the recovery of structural properties under different crystal representation precisions. The experimental pipeline is illustrated in Figure 4. Our process begins with raw CIFs from the MP-20 dataset. First, we encode these CIFs into crystal strings using representation methods with varying levels of precision. Subsequently, these strings are parsed to reconstruct the decompressed CIFs. We then employ symmetry analysis and learned property models to compute key properties for both the original CIF (C) and the decompressed CIF (C'). Finally, we quantify the fidelity of the representation by calculating the consistency between these two sets of properties.

Our fidelity experiments focus on the following three properties:

- **Symmetry:** We use PyMatgen to extract the space group number from the CIFs and calculate the Match Ratio between C and C' .
- **Formation Energy:** We utilize CHGNet (Deng et al., 2023) to calculate the formation energy and quantify the fidelity using the Mean Squared Error (MSE).
- **Band Gap:** We leverage MEGNet to predict the band gap (as detailed in Section 3.2.2) and use MSE to measure fidelity.

The results indicate that our high-precision representation significantly better preserves material properties against information loss during the compression-encoding process (cf. Table 7). We refrain from further increasing the precision as the current level already ensures sufficient fidelity; further increases would incur unnecessary computational costs and training difficulties.

Table 7: Results of representation fidelity experiments. **Low** denotes the original precision used in Crystal-TextLLM. **High** denotes the improved precision used in our method (as shown in Table 5).

Metric	Low Precision	High Precision
Match Ratio (%) $\uparrow$	49.28	96.17
MSE (E_{form}) $\downarrow$	3.44×10^{-4}	0.00
MSE (Band Gap) $\downarrow$	9.42×10^{-2}	2.83×10^{-3}

C Experimental Setup and Evaluation Protocols

C.1 Pre-Relaxation Evaluation Protocol

We adopt a strict pre-relaxation evaluation protocol to rigorously assess the intrinsic generative capabilities of our model. As emphasized by a recent unified evaluation framework (Betala et al., 2025), pre- and post-relaxation metrics measure different aspects of a generator. We evaluate raw outputs before relaxation so that improvements are not conflated with corrections introduced by an external DFT or potential-based optimizer. This protocol measures one-shot structural quality; it does not establish synthesizability or experimental stability.

C.2 Configuration of PLaID++ Baselines

To decouple the alignment algorithm’s performance from the underlying structural representation, we

Table 8: Upper-tail band-gap split.

Original Split	Band-Gap Regime		Total
	$E_g \leq 2.5$ eV	$E_g > 2.5$ eV	
Training	23,647	3,489	27,136
Validation	7,923	1,122	9,045
Test	7,922	1,128	9,050
Total	39,492	5,739	45,231

adopt the variant of PLaID++ utilizing standard 3D coordinates. Specifically, we establish two distinct Direct Preference Optimization (DPO) variants tailored for symmetry and band gap optimization:

- **Symmetry:** We build the **PLaID++ (3D Coord-SG DPO)** variant upon the supervised fine-tuned PLaID++ base model with 3D coordinate representations. Following Xu et al. (2025), we conduct DPO using preference samples drawn from seven typical space groups ($P1$, $C2/c$, $Amm2$, $I4m2$, $P3$, $P6_3/mmc$, and $F43m$).
- **Band Gap:** Utilizing the same SFT base model, we construct preference pairs by comparing the Mean Squared Error (MSE) between the target band gap and the predicted band gap of the generated structures, as evaluated by MEGNet. We then acquire the **PLaID++ (3D Coord-BG DPO)** variant by applying the identical DPO training pipeline used for the symmetry model.

D Additional Results and Verifications

D.1 Held-Out Upper-Tail Adaptation

Controlled Split For the controlled OOD experiment, all structures with reference band gap $E_g > 2.5$ eV are excluded from SFT, Standard RL, validation, hyperparameter tuning, and checkpoint selection. The 45,231 MP-20 entries are partitioned as shown in Table 8. The 5,739 OOD structures are further divided with seed 2026 into instance-disjoint OOD-Adapt (4,017) and OOD-Test (1,722) subsets. No material identifier or exact duplicate structure occurs in both subsets, although reduced formulas may overlap. TTRL observes only OOD-Adapt conditional prompts and never their target Crystal Strings or CIF structures.

Target-property results are reported in Table 3. Table 9 shows that the improvement does not materially reduce validity, composition fidelity, uniqueness, novelty, or the Stable–Unique–Novel (SUN)

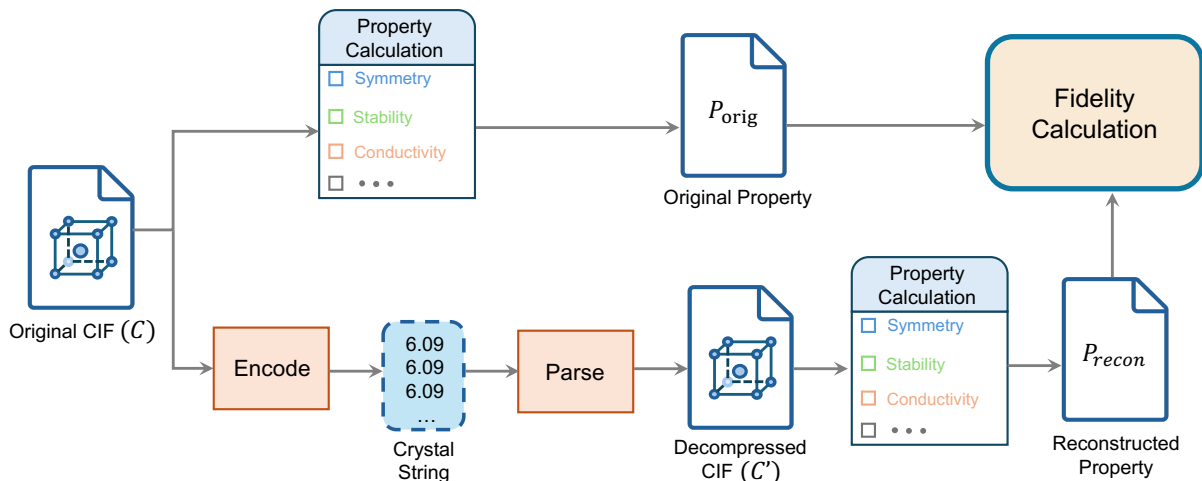


Figure 4: Illustration of the representation fidelity experiment pipeline. We measure the consistency of properties (Symmetry, Formation Energy, Band Gap) derived from the original CIF (C) and the reconstructed CIF (C') after the encoding-decoding process. High consistency indicates minimal information loss during representation.

rate. Structural quality is reported under AVE because BoN ranks candidates by band-gap error rather than these quality metrics.

Paired OOD Comparisons For each prompt, metrics are averaged over three inference seeds. We compute 95% intervals with 10,000 paired-bootstrap resamples over prompts and two-sided p -values with 10,000 paired sign-flip permutations. Table 10 shows that TTRL improves over the non-adapted controls, including when compared with High-Precision SFT using BoN-10.

In-Distribution Retention After OOD adaptation, we evaluate the policy on the 7,922 original test structures with $E_g \leq 2.5$ eV. Table 11 indicates limited forgetting relative to the ID-trained controls.

DFT Screening For each BoN method, we rank valid and unique candidates with ALIGNN rather than the optimized MEGNet reward and run DFT on the top 100. Candidate-level intervals use 10,000 bootstrap resamples. Table 12 evaluates selected-pool quality under this screening pipeline; it is not an unbiased estimate of the complete generation distribution, and overlapping intervals should not be interpreted as significant pairwise differences.

D.2 Statistical Robustness

Tables 1 and 2 report sample standard deviations over three independent inference seeds for the principal comparisons. These estimates capture stochastic decoding and candidate sampling, not

variance across independently trained RL checkpoints. For prompt-level inference, we average each prompt over the three seeds, use 10,000 paired-bootstrap resamples for 95% intervals, and use 10,000 two-sided paired sign-flip permutations for p -values. Prompts, rather than individual BoN candidates, are the independent units.

The single-condition band-gap comparison with PLaid++ and the compound-condition comparison with Standard RL are statistically comparable. We consequently avoid claiming significant improvement for these two cases.

D.3 Failure Analysis

We analyze the final BoN-selected output for every Symmetry + Composition prompt. Each output is assigned to the first failed stage: parsing, pre-relaxation structural validity, exact reduced-composition matching, or exact space-group matching. The mutually exclusive categories in Table 14 show that TTRL primarily reduces symmetry mismatch rather than formatting or lower-level validity errors.

D.4 Reward Ablations

Syngony-Level Shaping The partial syngony reward is used only during RL training; all reported Match Ratios require exact space-group agreement. Under High-Precision SFT, 26.0% of candidates exactly match the target and an additional 35.1% match only its syngony, so partial credit provides a denser training signal. We vary the same-syngony reward α while fixing initialization, prompt stream,

Table 9: Structural quality on OOD-Test under AVE. All results are reported as mean $\pm$ sample standard deviation over three inference seeds.

Method	Validity (%) $\uparrow$	Similarity (%) $\uparrow$	Stability (%) $\uparrow$	Uniqueness (%) $\uparrow$	Novelty (%) $\uparrow$	SUN (%) $\uparrow$
High-Precision SFT	96.8 ± 0.33	96.1 ± 0.26	6.25 ± 0.62	92.4 ± 0.58	89.6 ± 0.54	4.32 ± 0.77
Standard RL	96.4 ± 0.32	96.8 ± 0.25	6.94 ± 0.65	91.9 ± 0.56	89.2 ± 0.53	4.46 ± 0.73
TTRL	96.7 ± 0.35	97.3 ± 0.27	6.08 ± 0.61	92.0 ± 0.47	89.4 ± 0.46	4.84 ± 0.66

Table 10: Paired comparisons on OOD-Test. Positive Δ MSE denotes a reduction in prompt-level MEGNet MSE in favor of TTRL.

Comparison	Protocol	Improvement		Statistical Significance	
		Δ MSE $\uparrow$	Relative Reduction (%) $\uparrow$	95% CI	p
TTRL vs. High-Precision SFT	AVE vs. AVE	4.18	68.5	[3.76, 4.61]	< 0.001
TTRL vs. Standard RL	AVE vs. AVE	2.93	60.4	[2.55, 3.31]	< 0.001
TTRL vs. High-Precision SFT + BoN	AVE vs. BoN-10	0.48	20.0	[0.22, 0.74]	0.008
TTRL + BoN vs. High-Precision SFT + BoN	BoN-10 vs. BoN-10	1.52	63.3	[1.34, 1.70]	< 0.001
TTRL + BoN vs. Standard RL + BoN	BoN-10 vs. BoN-10	0.97	52.4	[0.80, 1.15]	< 0.001

Table 11: In-distribution retention after OOD adaptation under AVE. All results are reported as mean $\pm$ sample standard deviation over three inference seeds.

Method	Target-Property Performance		Structural Quality		
	MSE $\downarrow$	Tolerance (%) $\uparrow$	Validity (%) $\uparrow$	Similarity (%) $\uparrow$	SUN (%) $\uparrow$
High-Precision SFT	2.35 ± 0.06	72.4 ± 0.63	97.1 ± 0.27	96.6 ± 0.24	6.60 ± 0.55
Standard RL	2.20 ± 0.05	73.8 ± 0.52	96.9 ± 0.25	97.0 ± 0.28	6.63 ± 0.57
TTRL	2.39 ± 0.06	72.0 ± 0.68	96.8 ± 0.22	97.2 ± 0.21	6.58 ± 0.53

Table 12: DFT evaluation of selected OOD candidate pools. Results are reported as point estimates with 95% bootstrap confidence intervals.

Method	DFT Property Error	Target Satisfaction	
	DFT MAE $\downarrow$	Tolerance Hit (%) $\uparrow$	Wide-Gap Hit (%) $\uparrow$
High-Precision SFT + BoN	1.28 [1.12, 1.45]	32.0 [23.0, 41.0]	58.0 [48.0, 68.0]
Standard RL + BoN	1.10 [0.95, 1.26]	39.0 [30.0, 49.0]	64.0 [54.0, 73.0]
TTRL + BoN	0.72 [0.61, 0.84]	61.0 [51.0, 70.0]	79.0 [70.0, 87.0]

Table 13: Prompt-level paired statistical comparisons. A positive improvement favors TTRL.

Task / Metric	Comparison	Improvement	95% CI	p
Symmetry BoN Match	TTRL vs. High-Precision SFT	+2.72 pp	[1.80, 3.63]	< 0.001
	TTRL vs. PLaiD++	+1.48 pp	[0.55, 2.40]	0.0024
	TTRL vs. Standard RL	+2.24 pp	[1.31, 3.17]	< 0.001
Symmetry + Composition BoN Match	TTRL vs. High-Precision SFT	+6.90 pp	[5.83, 7.95]	< 0.001
	TTRL vs. PLaiD++	+9.09 pp	[8.03, 10.14]	< 0.001
	TTRL vs. Standard RL	+6.56 pp	[5.48, 7.63]	< 0.001
Band Gap BoN MSE	TTRL vs. High-Precision SFT	1.388	[1.21, 1.57]	< 0.001
	TTRL vs. PLaiD++	0.009	[-0.018, 0.036]	0.516
	TTRL vs. Standard RL	0.766	[0.61, 0.92]	< 0.001
Band Gap + Composition BoN MSE	TTRL vs. High-Precision SFT	0.072	[0.048, 0.096]	< 0.001
	TTRL vs. PLaiD++	0.578	[0.49, 0.67]	< 0.001
	TTRL vs. Standard RL	0.018	[-0.004, 0.038]	0.114

400-step budget, $\lambda = 2.0$, and BoN-10 evaluation. Table 15 shows that exact-only training improves over SFT, moderate shaping gives the best exact-match results in the tested range, and stronger shaping overemphasizes coarse symmetry. We therefore treat $\alpha = 0.5$ as an empirical default rather than a

theoretically optimal value.

Composition Reward and Mode Collapse We compare property-only rewards with the composite rewards in Eq. 4, holding the prompt order, random seed, optimization settings, and evaluation prompts fixed. At each saved checkpoint, ten fresh

Table 14: First-failure decomposition for Symmetry + Composition generation. Percentages are computed over test prompts and sum to 100% within each method.

Method	First-Failure Rate (%) ↓				Success (%) ↑
	Parse	Invalid	Composition	Symmetry	
High-Precision SFT	0.3	1.1	1.8	54.8	42.0
Standard RL	0.3	1.3	1.5	54.5	42.4
TTRL	0.2	0.9	1.1	48.9	48.9

Table 15: Sensitivity to the same-synony partial reward. All results are reported as mean $\pm$ sample standard deviation over three inference seeds.

Method	α	Exact Symmetry		Symmetry + Composition		Synony
		AVE (%) ↑	BoN (%) ↑	AVE (%) ↑	BoN (%) ↑	AVE (%) ↑
High-Precision SFT	—	26.01 $\pm$ 0.38	51.95 $\pm$ 0.44	29.92 $\pm$ 0.45	42.01 $\pm$ 0.68	61.12 $\pm$ 0.41
TTRL exact-only	0.00	32.16 $\pm$ 0.29	52.74 $\pm$ 0.31	32.71 $\pm$ 0.37	45.86 $\pm$ 0.44	64.02 $\pm$ 0.33
TTRL weak shaping	0.25	34.61 $\pm$ 0.24	54.02 $\pm$ 0.29	34.71 $\pm$ 0.33	47.96 $\pm$ 0.40	66.87 $\pm$ 0.30
TTRL default	0.50	35.77 $\pm$ 0.20	54.67 $\pm$ 0.27	35.58 $\pm$ 0.31	48.91 $\pm$ 0.36	68.15 $\pm$ 0.28
TTRL strong shaping	0.75	34.88 $\pm$ 0.23	53.79 $\pm$ 0.30	34.69 $\pm$ 0.34	47.72 $\pm$ 0.41	70.92 $\pm$ 0.27

candidates are sampled per prompt and grouped by structural equivalence rather than raw-string equality. Table 16 shows that property-only optimization rapidly concentrates on a few structures. The prompt-dependent composition reward preserves approximately 90% uniqueness and limits the largest cluster to about 3%. This supports the specific conclusion that R_{COMP} mitigates reward-driven structural mode collapse; it does not by itself prove physical validity.

D.5 Cross-Backbone Evaluation

We repeat the High-Precision SFT and TTRL pipeline using Qwen3.5-9B-Base (Qwen Team, 2026). Both backbones use the same MP-20 splits, high-precision representation, prompts, rewards, optimization procedure, and inference protocol. The experiment changes the underlying LLM but does not establish universal cross-architecture generality.

D.6 Validation against Ground-Truth DFT

To assess the scope of MEGNet as a learned surrogate reward, we compare its predictions with Density Functional Theory (DFT) calculations. This experiment provides partial evidence about surrogate fidelity but does not treat MEGNet as a physical oracle.

Experimental Setup We utilized the model optimized via Label-Free TTRL to perform conditional generation targeting band gap constraints across the MP-20 test set. From the generated pool, we selected 100 structures with low MEGNet-predicted

error and computed their band gaps using the Vienna Ab initio Simulation Package (VASP) with the PBE functional (Perdew et al., 1996). Because the sample is selected by the optimized surrogate, it is not an unbiased estimate of the full generated distribution.

Correlation Analysis Figure 5 shows a positive MEGNet–DFT association, with Pearson $r = 0.79$ and MAE 1.21 eV. GRPO normalizes rewards within each sampled group, so systematic calibration bias may be less consequential when local candidate ordering is preserved. Pearson correlation, however, does not prove correct ordering for every group, and the observed MAE leaves meaningful room for proxy bias. The independent ALIGNN and ALIGNN-guided DFT results in Appendix D.1 provide complementary evaluation beyond the optimized reward model.

D.7 Basic Evaluation: Proxy Metrics and Stability

We evaluated the comprehensive generative performance on the MP-20 dataset, focusing on standard proxy metrics and the SUN (Stable, Unique, and Novel) rate. Detailed results are presented in Table 20. Stability and SUN rates were calculated following the methodology established by Miller et al. (2024). For the RL evaluation, we use the LLaMA-2-7B TTRL checkpoint reported in the main conditional-generation experiments. In order to avoid the effect of representations of crystal structures, we compared with PLaid++ (3D Co-ord) (Xu et al., 2025) where the crystal structures

Table 16: Training dynamics with and without the composition reward. Unique Structure Ratio measures structural diversity during training, while Dominant Mode reports the frequency of the most common structure at step 400.

Objective	Reward	Unique Structure Ratio (%) $\uparrow$						Dominant Mode (%) $\downarrow$
		Step 0	25	50	100	200	400	
Symmetry	λR_{SYM}	94.8	76.3	43.7	14.5	9.8	8.9	70.6
	$\lambda R_{\text{SYM}} + R_{\text{COMP}}$	94.8	94.1	93.5	92.8	92.0	91.4	2.7
Band Gap	λR_{BG}	94.8	60.5	21.8	8.4	6.9	6.3	78.2
	$\lambda R_{\text{BG}} + R_{\text{COMP}}$	94.8	93.7	92.9	91.8	90.7	89.9	3.2

Table 17: Cross-backbone results for symmetry-controlled crystal generation. All results are reported as mean $\pm$ sample standard deviation over three inference seeds. Results of **BoN sampling** are highlighted with **blue background**.

Backbone	Optimization	Symmetry		Symmetry + Composition			
		Match Ratio (%) $\uparrow$		Match Ratio (%) $\uparrow$		Similarity (%) $\uparrow$	
		AVE	BoN	AVE	BoN	AVE	BoN
LLaMA-2-7B	High-Precision SFT	26.01 $\pm$ 0.38	51.95 $\pm$ 0.44	29.92 $\pm$ 0.45	42.01 $\pm$ 0.68	95.38 $\pm$ 0.15	98.07 $\pm$ 0.12
	TTRL	35.77 $\pm$ 0.20	54.67 $\pm$ 0.27	35.58 $\pm$ 0.31	48.91 $\pm$ 0.36	95.40 $\pm$ 0.18	99.40 $\pm$ 0.11
Qwen3.5-9B	High-Precision SFT	27.36 $\pm$ 0.34	53.08 $\pm$ 0.40	31.08 $\pm$ 0.41	43.26 $\pm$ 0.57	95.72 $\pm$ 0.14	98.31 $\pm$ 0.10
	TTRL	39.12 $\pm$ 0.23	57.46 $\pm$ 0.29	39.01 $\pm$ 0.30	52.38 $\pm$ 0.33	96.04 $\pm$ 0.16	99.55 $\pm$ 0.08

Table 18: Cross-backbone results for band-gap-controlled crystal generation. All results are reported as mean $\pm$ sample standard deviation over three inference seeds. Results of **BoN sampling** are highlighted with **blue background**.

Backbone	Optimization	Band Gap		Band Gap + Composition			
		MSE $\downarrow$		MSE $\downarrow$		Similarity (%) $\uparrow$	
		AVE	BoN	AVE	BoN	AVE	BoN
LLaMA-2-7B	High-Precision SFT	6.182 $\pm$ 0.084	2.462 $\pm$ 0.061	3.587 $\pm$ 0.061	0.411 $\pm$ 0.009	94.65 $\pm$ 0.24	93.91 $\pm$ 0.31
	TTRL	4.797 $\pm$ 0.065	1.074 $\pm$ 0.015	2.285 $\pm$ 0.041	0.339 $\pm$ 0.007	99.12 $\pm$ 0.12	99.11 $\pm$ 0.07
Qwen3.5-9B	High-Precision SFT	5.934 $\pm$ 0.078	2.298 $\pm$ 0.052	3.421 $\pm$ 0.054	0.386 $\pm$ 0.008	95.04 $\pm$ 0.21	94.26 $\pm$ 0.24
	TTRL	4.312 $\pm$ 0.057	0.901 $\pm$ 0.012	2.018 $\pm$ 0.035	0.286 $\pm$ 0.006	99.27 $\pm$ 0.10	99.34 $\pm$ 0.06

Table 19: Cross-backbone structural quality. $W(\rho)$ and $W(N_{\text{el}})$ denote Wasserstein distances for density and number of elements, respectively.

Backbone	Optimization	Validity (%)		Coverage (%)		Distribution Distance		Stability (%)	
		Struct. $\uparrow$	Comp. $\uparrow$	Recall $\uparrow$	Precision $\uparrow$	$W(\rho)$ $\downarrow$	$W(N_{\text{el}})$ $\downarrow$	Stable $\uparrow$	SUN $\uparrow$
LLaMA-2-7B	High-Precision SFT	100.00	91.17	99.88	100.00	0.101	0.025	6.24	4.03
	TTRL	100.00	91.05	99.98	100.00	0.033	0.019	6.18	4.48
Qwen3.5-9B	High-Precision SFT	100.00	92.05	99.91	100.00	0.088	0.022	6.38	4.28
	TTRL	100.00	92.20	99.99	100.00	0.025	0.015	6.47	5.08

are represented as text-based 3D coordinates.

Analysis Validity and coverage are high for most methods. TTRL obtains the lowest reported Wasserstein distances for density (ρ) and number of elements (N_{el}), while its Stability Rate remains close to High-Precision SFT (6.18% vs. 6.24%) and its SUN rate increases from 4.03% to 4.48%. PLaid++ (3D Coord-DPO) has a higher raw Stability Rate but a lower SUN rate, illustrating why stability alone is insufficient to characterize useful candidate diversity. These proxy metrics complement, rather than replace, direct property-control and high-fidelity evaluations.

D.8 Experimental Results of Multi-Objective Design

To simulate varying preference weights, we conduct a two-objective optimization experiment targeting symmetry and band gap. We assigned varying weights to each property to reflect different trade-offs. Specifically, the control parameters λ (for MORL) and μ (for Model Merging) were varied across the set $\{0, 0.25, 0.5, 0.75, 1.0\}$. Detailed quantitative results are presented in Table 21.

Within the tested range, MM is competitive with the separately trained MORL baselines and is better on both reported objectives at $\mu = 0.50$ and 0.75 . Adjusting μ provides intermediate symmetry–band-gap trade-offs without another RL run. This is evidence for bounded two-objective interpolation

Table 20: Results of proxy metrics and stability evaluation on MP-20. **wdist** denotes the Wasserstein distance between the generated and ground-truth distributions (lower is better). **SUN Rate** measures the percentage of structures that are simultaneously Stable, Unique, and Novel.

Methods	Validity (%) $\uparrow$		Coverage (%) $\uparrow$		Property Distribution $\downarrow$		Stability Rate (%) $\uparrow$	SUN Rate (%) $\uparrow$
	Structural	Composition	Recall	Precision	wdist (ρ)	wdist (N_{el})		
CDVAE	100.00	86.70	99.10	99.50	0.688	1.430	5.40	—
DiffCSP	100.00	83.25	99.71	99.76	0.350	0.125	5.06	3.34
FlowMM	96.85	83.19	99.49	99.58	0.239	0.083	4.65	2.34
Crystal-Text-LLM	96.40	93.30	91.10	94.90	3.610	1.060	6.20	4.32
High-Precision SFT	100.00	91.17	99.88	100.00	0.101	0.025	6.24	4.03
PLaID++ (3D Coord)	98.93	89.94	99.12	98.03	1.819	0.106	7.20	2.81
PLaID++ (3D Coord-DPO)	96.62	89.94	99.03	99.18	1.332	0.122	13.94	1.69
TTRL (Ours)	100.00	91.05	99.98	100.00	0.033	0.019	6.18	4.48

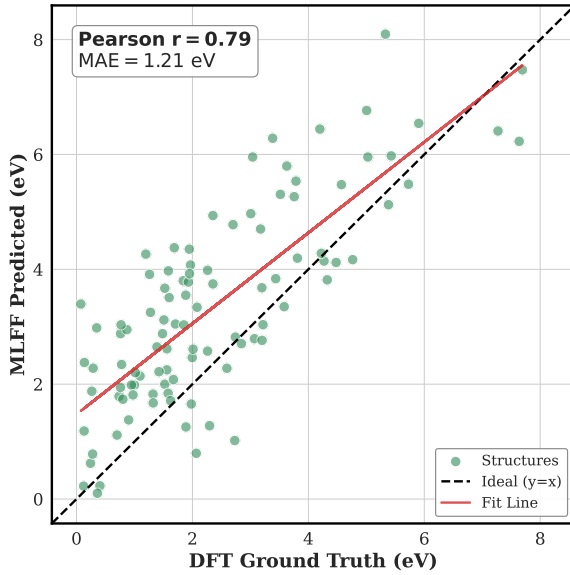


Figure 5: Parity plot comparing MEGNet-predicted and DFT-calculated band gaps for 100 candidates selected using MEGNet. The positive correlation ($r = 0.79$) indicates an informative association, while the 1.21 eV MAE and selection procedure limit stronger conclusions about calibration or the complete generated distribution.

only; it does not establish extrapolation or general N -way preference composition.

D.9 Sensitivity Analysis of Reward Coefficients

To justify our hyperparameter choice, we conducted a fine-grained parameter sweep for reward coefficients $\lambda \in [0.0, 4.0]$, balancing target properties against composition constraints. As illustrated in Figure 6, both Symmetry (blue line) and Band Gap (red line) objectives exhibit a consistent trend: performance peaks at $\lambda = 2.0$, with a Match Ratio of 35.77% and an MSE of 4.797, and remains

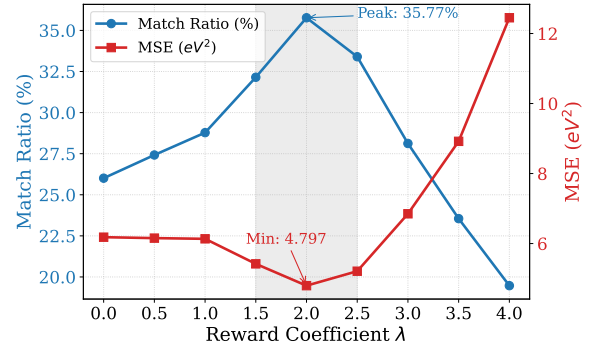


Figure 6: Sensitivity analysis of reward coefficients (λ). The **blue line** (left axis) represents the Match Ratio for symmetry optimization, while the **red line** (right axis) represents the MSE for band gap optimization. The **shaded gray region** ($\lambda \in [1.5, 2.5]$) indicates the robust operating range where the model achieves optimal trade-offs between target properties and structural validity.

robust within the window $\lambda \in [1.5, 2.5]$ (shaded region). Beyond this range ($\lambda \geq 3.0$), performance degrades sharply due to reward hacking, where excessive weights compromise structural validity. Consequently, we adopt $\lambda = 2.0$ as the default robust setting for all reported experiments.

Table 21: Quantitative comparison for multi-objective conditional crystal material design. **Match Ratio** and **Similarity** measure constraint alignment, while **MSE** and **Hit Rate** measure band-gap performance. The weights λ (for MORL) and μ (for MM) control the trade-off between objectives. Weights 0.00 and 1.00 correspond to the single-objective **TTRL-Bandgap** and **TTRL-Symmetry** experts, respectively, serving as the interpolation endpoints. All results are reported as mean $\pm$ sample standard deviation over three independent random seeds. **MM** results are highlighted with **gray background**. Best mean results are in **bold**.

Method	Weight (λ/μ)	Constraint Alignment		Band-Gap Performance	
		Match Ratio (%) $\uparrow$	Similarity (%) $\uparrow$	MSE $\downarrow$	Hit Rate (%) $\uparrow$
Crystal-Text-LLM	—	25.43 $\pm$ 0.39	97.00 $\pm$ 0.17	3.524 $\pm$ 0.066	74.67 $\pm$ 0.76
High-Precision SFT	—	26.57 $\pm$ 0.42	95.60 $\pm$ 0.21	3.535 $\pm$ 0.059	73.58 $\pm$ 0.71
TTRL-Bandgap	0.00	18.39 $\pm$ 0.32	98.01 $\pm$ 0.14	2.285 $\pm$ 0.041	50.37 $\pm$ 0.93
MORL	0.25	23.65 $\pm$ 0.45	97.98 $\pm$ 0.15	2.319 $\pm$ 0.032	49.26 $\pm$ 1.12
MM	0.25	22.51 $\pm$ 0.68	97.34 $\pm$ 0.18	2.288 $\pm$ 0.035	50.02 $\pm$ 0.98
MORL	0.50	26.83 $\pm$ 0.52	98.60 $\pm$ 0.11	2.508 $\pm$ 0.104	64.51 $\pm$ 1.45
MM	0.50	27.31 $\pm$ 0.41	98.12 $\pm$ 0.14	2.440 $\pm$ 0.091	57.33 $\pm$ 1.20
MORL	0.75	31.51 $\pm$ 0.65	97.54 $\pm$ 0.22	3.319 $\pm$ 0.155	59.37 $\pm$ 1.62
MM	0.75	33.89 $\pm$ 0.50	98.10 $\pm$ 0.19	2.899 $\pm$ 0.120	66.42 $\pm$ 1.35
TTRL-Symmetry	1.00	35.58 $\pm$ 0.31	96.88 $\pm$ 0.16	3.772 $\pm$ 0.142	62.68 $\pm$ 0.88