Towards Generating Stable Materials via Large Language Models with Reinforcement Learning Finetuning

Zhang-Wei Hong* IBM zwhong@mit.edu

Nofit Segal* DMSE, MIT nofit@mit.edu

Aviv Netanyahu EECS, MIT avivn@mit.edu

Hoje Chun DMSE, MIT hojechun@mit.edu

Rafael Gómez-Bombarelli DMSE, MIT rafagb@mit.edu

Pulkit Agrawal EECS, MIT pulkit@mit.edu

Abstract

Discovering novel materials is essential for advancing technology, yet generating thermodynamically stable crystal structures remains a significant challenge due to the difficulty of directly steering generative models toward physically realistic structures. We investigate the impact of reinforcement learning (RL) finetuning of Large Language Models (LLMs) for crystal structure generation using energy-based rewards. Our results show that RL-finetuning improves the rate of generating metastable crystals compared to supervised finetuning (SFT) and performs comparably to established diffusion-based baselines. Notably, the RL-steered model produces structures significantly closer to their relaxed states, which potentially reduces the computational overhead of downstream structural optimization. Future efforts may build upon these results by investigating reward formulations better aligned with thermodynamic stability and exploring methods to maintain structural variety during optimization.

1 Introduction

The discovery of new materials drives technological innovation, enabling advances in fields ranging from energy storage [1] to electronics [2] and medicine [3]. Traditionally, the process of identifying and synthesizing novel materials has been slow and resource-intensive, relying on iterative experimentation and domain expertise. Machine learning (ML) offers an alternative, data-driven approach that can accelerate this process by guiding exploration within vast chemical and structural spaces. While predictive models can identify promising candidates from existing materials datasets [4], generative models provide the additional capability of proposing entirely new compositions and structures, thereby expanding the search space beyond existing materials and opening opportunities for breakthrough discoveries.

Modeling materials is particularly challenging because it requires capturing joint distributions over variables of different types: atomic species, which are discrete, and lattice parameters and atomic positions, which are continuous. A further challenge in generative materials design is ensuring that the proposed candidates are **thermodynamically stable**, meaning they can exist without spontaneously breaking down into other material phases. Among the vast number of ways atoms can be arranged, only a tiny fraction corresponds to such stable structures. Despite progress in generative modeling, re-

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Science.

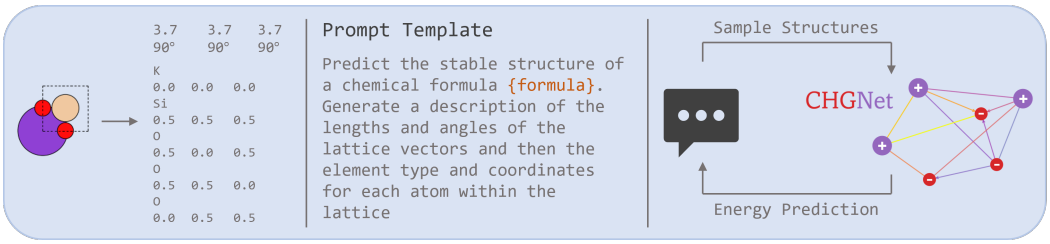


Figure 1: **Left:** Example representation of a crystal structure, showing lattice parameters (lengths and angles) along with atom types and positions, encoded as a string. **Middle:** Text prompt provided to the model, with a chemical composition given as input. **Right:** Overview of our reinforcement learning (RL) finetuning setup, where an energy predictor (CHGNet [9]) guides the model toward lower-energy, more stable structures.

liably producing stable candidates under high-fidelity quantum-mechanical (QM) evaluation remains a difficult task [5–8].

Large language models (LLMs) offer a promising avenue for generative materials design due to their ability to incorporate natural language prompting, enabling straightforward and flexible conditioning on desired properties or constraints. Pretrained on broad corpora that include chemistry and materials science knowledge from the scientific literature, these models possess strong prior knowledge of chemical rules and patterns. Finetuning an LLM for crystallographic information file (CIF) generation further aligns it with the structural and compositional distributions of known materials, improving its capacity to generate valid candidates.

In this work, we explicitly steer generation toward thermodynamically stable materials by integrating an LLM-based generative model with a reinforcement learning (RL) framework, as shown in 1. The RL component provides feedback based on stability evaluations, enabling the model to iteratively refine its outputs and improve the likelihood of producing candidates with low energy, thereby shifting the distribution. This integration bridges the flexibility of language-based conditioning with targeted optimization for stability.

2 Related Work

Early work on **crystal structure prediction** (CSP), the task of generating a crystal structure given a chemical composition, relies on producing candidate materials using atomistic simulations and high-throughput quantum-mechanical calculations, followed by high-throughput quantum-mechanical calculations [10] to estimate their energies and identify stable structures [11–13]. This screening process can be accelerated using ML interatomic potentials (MLIPs), which are used to relax crystal structures via potential energy minimization, such as CHGNet [9], M3GNet [14], and UMA [15].

More recent efforts have focused on **generative modeling** to accelerate the discovery of stable materials. Trained on large databases of QM-verified stable structures, these models aim to generate new materials that follow similar distributions. Approaches include combining variational autoencoders (VAEs) with diffusion decoders [7, 16], using diffusion or flow-matching models that jointly model lattices and atomic positions [17, 18, 6], and applying classifier-free guidance to enable conditional generation and improved alignment with target properties [5].

Language models based approaches include work focusing on domain-specific tokenization schemes [19], supervised finetuned (SFT) LLMs [20, 21], and hybrid methods that combine LLMs with flow matching [22], predictive models and diffusion [23] or evolutionary search algorithms [24]. Concurrent work includes Cao and Wang [25], which use RL to finetune a transformer-based model with explicit knowledge of symmetry, using energy or other property values as a reward. Xu et al. [26] finetune a language model on symmetry-informed textual representations of crystals, and further apply Direct Preference Optimization (DPO) [27] using stability labels approximated with an MLIP.

Recent work has explored **RL** as a general framework for enhancing the abilities of **LLMs** by providing more reliable training signals. Verifiable rewards, such as correctness of intermediate steps or final answers, have proven especially effective for guiding models toward consistent reasoning

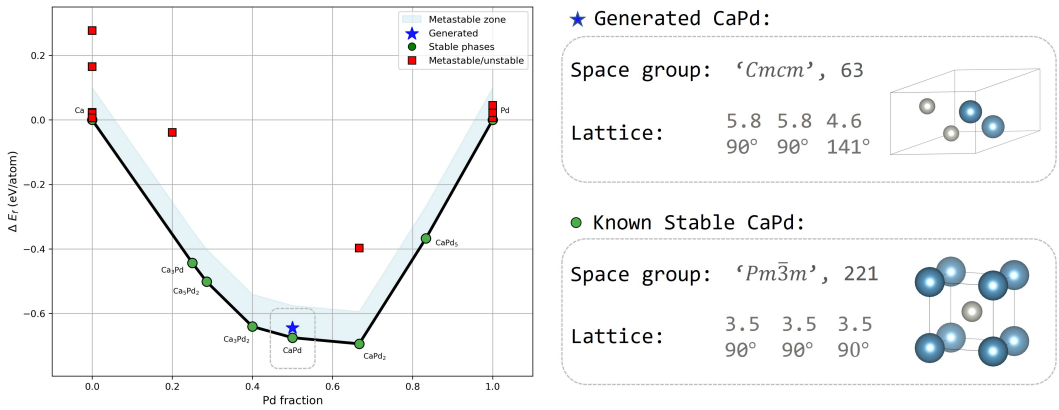


Figure 2: **Left:** Binary phase diagram for the Ca–Pd system. The black line shows the convex hull, representing the most stable phases. All phases shown have $E_{\text{hull}} < 0.4\ eV/\text{atom}$. Stable and metastable/unstable structures are indicated, with the metastable zone $E_{\text{hull}} < 0.1\ eV/\text{atom}$ highlighted in blue. The blue star highlights our generated material. At this composition, the stable phase CaPd lies on the hull, while the generated structure sits slightly above it with $E_{\text{hull}} = 0.001\ eV/\text{atom}$, a value considered effectively stable. **Right:** Comparison of the crystal symmetries of the two polymorphs (different structural forms of the same composition), including their space group symbols and numbers, and lattice parameters (lengths and angles).

[28, 29]. In parallel, separate efforts on preference alignment leverage human or proxy feedback to better match model outputs with user expectations [27, 30].

3 Preliminaries

Crystal Representation A crystal structure can be described mathematically as $\mathcal{C}(L, A, X)$, where $L \in \mathbb{R}^{2\times 3}$ is the lattice parameters $a, b, c, \alpha, \beta, \gamma$ defining the periodic unit cell through its lengths and the angles, $A_i \in \mathbb{Z}^+$ are the atomic species, and X_i are their fractional coordinates $\mathbf{X}_i \in [0, 1)^3$ within the unit cell. This representation uniquely specifies a periodic arrangement of atoms in three dimensions. Crystal structures can be stored in CIFs, which encode the lattice parameters, atomic species, and atomic coordinates in a standardized text format, as shown in Figure 1.

Stability of Materials Stability is commonly evaluated using the *convex hull* of formation energies, which defines the phases that are energetically favorable relative to all competing phases. The stability of a material is quantified by its *energy above the hull*, E_{hull}, defined as

$$E_{\text{hull}} = E_{\text{tot}} - \Sigma_i x_i E_i,$$

Where E_{tot} is the total energy per atom of the material under consideration, x_i is the fraction of the i-th competing phase, and E_i is the ground-state energy per atom of that phase. Intuitively, E_{hull} measures how much higher in energy a material is, compared to the most favorable mixture of competing phases. In this work, we employ the pre-trained UMA model for energy evaluation due to its close alignment with QM Density Functional Theory (DFT) results, and following [24], use fixed phase diagrams derived from the Materials Project 2023 DFT calculations for reference [15, 31, 32].

Materials with $E_{\text{hull}} = 0.0$ eV/atom lie exactly on the hull and are considered the most stable phase at a given composition. Small positive values (e.g., < 0.1 eV/atom) indicate metastable materials, which can be considered experimentally synthesizable under nonequilibrium synthesis conditions [33]. Larger E_{hull} values indicate a stronger tendency to decompose to the competing, more energy-favorable phases. Figure 2 illustrates this concept for the Ca–Pd binary system.

Crystal Symmetries The *space group* G of a crystal is the group of Euclidean transformations that leave the crystal invariant. Therefore, in crystallography, space groups represent a description of the crystal's symmetry. These transformations include translations, rotations, inversions, and reflections. In three dimensions, space groups are classified into 230 types (when chiral copies are considered

distinct), and are numbered in order of increasing symmetry, starting with the least symmetric groups (triclinic system) and ending with the most symmetric (cubic system).

Symmetry and stability are closely related, but not in a one-to-one manner. The lowest-energy crystal structure arises from a complex optimization process that involves multiple factors. Empirically, symmetry is often correlated with more stable atomic configurations [34–36]. Stability can drive the emergence of symmetry, as it depends on how atoms arrange to minimize overlap of electron clouds, and such configurations may exhibit different levels of symmetry depending on the electron density distributions of the constituent atoms [37, 38].

4 Method

Tokenization. Tokenization of the CIF string is done by *byte pair encoding* (BPE) as suggested in [20], a compression method that assigns tokens to common substrings, making overall sequence lengths shorter [39].

Model. We build on the Qwen family of large language models (LLMs) [40–42], using Qwen2.5-7B-Instruct as our base model. Qwen2.5 is a general-purpose transformer trained on web-scale corpora for natural language and programming tasks, with the instruct variant further tuned to follow natural language prompts via sueprvised finetuning (SFT). In our setup, the model is prompted with a chemical formula and CIF examples (see A.4) and generates candidate crystal structures in text form.

We study the use of *reinforcement learning (RL) finetuning* to bias the LLM generation toward physically stable crystal structures. As a complementary variant, we also consider *supervised finetuning* (SFT), in which the model is trained to reproduce ground-truth CIFs conditioned on the input prompt using a standard maximum-likelihood objective. This stage adapts the model to crystallographic syntax and structural conventions and can be used as an initialization for RL finetuning.

Reinforcement Learning Finetuning. To bias the model toward generating stable materials, we apply RL finetuning using Group Relative Policy Optimization (GRPO) [43], a variant of Proximal Policy Optimization (PPO) [44]. We define the state (i.e., model inputs) x as the prompt containing a chemical formula of atom composition, and output y, the string formatted in bulk representation of the crystal structure of that formula. RL learns a policy π that takes in x and generates y that maximizes a given reward R. We finetune the full weight of the model with the reward function defined below. PPO maximizes a clipped objective:

$$L_{PPO} = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)], \tag{1}$$

where $r(\theta) = \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}$ is the policy ratio that measures how likely the new policy π_θ is to take the same action compared to the old policy $\pi_{\theta_{\text{old}}}$, and A_t is the advantage estimated by a value function. The clipping ensures $r_t(\theta)$ does not stray far from 1, thereby constraining updates so that the policy improves steadily without collapsing. GRPO eliminates the need for a value function by defining relative advantages within a group of candidates sampled from the model at the same input:

$$A_i = \frac{R_i - \mu}{\sigma} \tag{2}$$

where R_i is candidate i's reward, $\mu = \frac{1}{n}\sum_{j=1}^{n} R_j$ is the group mean and $\sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(R_j - \mu)^2}$ is the group standard deviation. We set the group size to be 8. This formulation is simpler and more computationally efficient while maintaining stability. In our setting, Qwen2.5-7B-Instruct acts as the policy, candidate crystal structures are sampled, and CHGNet [9] provides the reward signal based on predicted potential energies E_{tot}, directly aligning the LLM with the goal of producing low-energy stable structures. Moreover, we give the model a penalty of -0.1 when it generates an invalid crystal structure due to a parsing error or exceeding 20 atoms.

5 Experiments

5.1 Setup

Dataset. We report results on **MP-20**, a realistic benchmark comprising all materials in the Materials Project database (circa July 2021) with at most 20 atoms per unit cell and an energy above the convex hull of less than 0.08 eV/atom [45]. The dataset contains approximately 45,000 materials.

Baselines. We evaluate our approach against two classes of baselines. **(1) Domain-specific baselines.** We compare against two established crystal structure generation models: FlowMM, a flow-matching–based method [6], and DiffCSP, a diffusion-based generative framework [17]. Both methods are evaluated in the CSP setting, where the chemical formula is provided as input and strictly enforced throughout the generation process. **(2) Language model variants.** We additionally compare against instruction-tuned large language models based on Qwen2.5-7B. These include the off-the-shelf base model (Qwen2.5), a supervised finetuned variant (Qwen2.5 SFT) trained to reproduce domain-specific CIF examples, and a model finetuned with reinforcement learning on top of SFT (Qwen2.5 SFT+RL).

As described in Section 4, the language model is prompted with a chemical formula and example CIFs. Under the SFT objective, the model is trained to reproduce the provided CIF exactly. In contrast, during RL finetuning, the objective does not explicitly encourage reproducing the same stoichiometry. Instead, the model is incentivized to preserve the set of atomic species via a penalty term, while allowing the relative atomic ratios to vary in order to optimize the energy-based reward. In both cases, generated structures are not post-filtered to enforce exact agreement with the input formula. This behavior contrasts with the diffusion- and flow-based baselines, which strictly constrain the formula during generation and is an important distinction in our experimental setting.

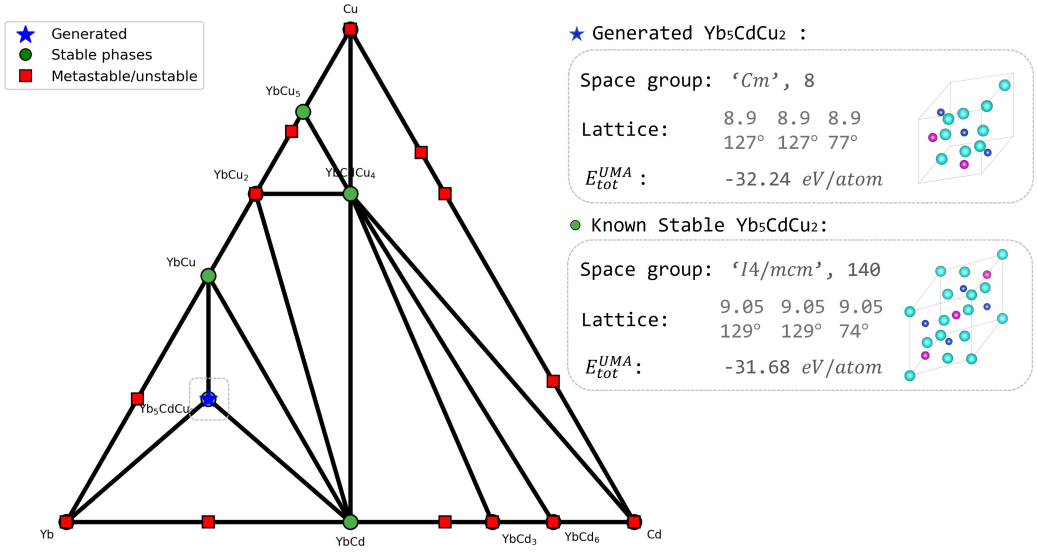


Figure 3: **Ternary phase diagram of the Yb-Cd-Cu system.** Phases with DFT $E_{\text{hull}} < 0.4\,eV/\text{atom}$ are shown. Stable phases are shown as green circles, and metastable/unstable phases as red squares. The blue star marks the composition of our generated material, Yb_5CdCu_2. Insets compare the two polymorphs of Yb_5CdCu_2, including space group (symbol and number), lattice parameters (lengths and angles), and UMA-predicted total energies. The UMA-predicted energy of the generated structure is lower, suggesting that it could be a novel ground state.

5.2 Stability

Metric. To assess thermodynamic stability, we first relax the generated structures using UMA-s-1p1 [15] (see A.2 for details), a large pretrained MLIP, and obtain their total energies. The energy above the convex hull, $\boldsymbol{E}_{\text{hull}}$, is then calculated by comparing these energies to reference values from the

Materials Project database [45]. Based on this metric, we define the **Metastability Rate** as the fraction of generated (and subsequently relaxed) structures with $E_{\text{hull}} < 0.1$ eV/atom. Additionally, we report the **Match Rate**, which quantifies how often a generated CIF preserves its structure after relaxation, and the **RMSD**, which measures the root-mean-square deviation in atomic positions between each generated structure and its relaxed counterpart. Both Match rate and RMSD are calculated using pymatgen `StructureMatcher` with default settings. Since most current generative models require post-generation structural relaxation, achieving a high Match Rate and low RMSD is desirable as it can reduce the associated computational cost. While QM simulations would provide the most accurate relaxations, they are very expensive; we therefore use UMA as a computationally efficient proxy.

Results. Table 1 reports the stability of generated materials after structural relaxation across all methods. DiffCSP achieves the highest metastability rate (30.84%) and the lowest average E_{hull}. Qwen2.5 SFT+RL, Qwen2.5 SFT, and FlowMM perform comparably, with Qwen2.5 SFT+RL attaining a slightly higher metastability rate of 27.29%.

Figure 2 shows a binary metastable phase generated by the SFT+RL finetuned model, while Figures 3 and 4 present representative ternary phases. In Figure 2, the convex hull is drawn from known ground-state phases, and a newly generated structure is highlighted. At this composition, the known stable phase CaPd lies on the hull, while our generated structure is predicted by UMA [15] to have $E_{\text{hull}} = 0.001$ eV/atom relative to quantum-mechanically computed reference energies. This value is well within the metastability threshold and can therefore be considered metastable. The two phases are illustrated alongside their structural information, with their distinct crystal symmetries highlighting that they are separate polymorphs of the same composition. In Figure 3, the generated material has a UMA-predicted energy lower than that of the known phase, suggesting a potential novel ground state. Figure 4 shows a generated material with a UMA-predicted energy between two experimentally observed polymorphs, indicating that it may correspond to a synthesizable metastable phase.

RL finetuning alone does not substantially improve the metastability rate, increasing it only from 5.03% (Qwen2.5) to 7.6% (Qwen2.5 RL). One likely explanation is that the model exploits the total energy reward E_{tot} by favoring compositions with heavier elements, which can lower the absolute total energy without improving thermodynamic stability. As defined in Section 3, stability is determined by the energy above the convex hull, E_{hull}, which measures stability relative to competing phases. Minimizing E_{tot} alone therefore does not guarantee a reduction in E_{hull}, as the reference energies of competing phases can similarly decrease. In contrast, when RL is applied on top of supervised finetuning (SFT+RL), metastability slightly improves relative to SFT alone (from 26.5% to 27.29%), suggesting that supervised pre-alignment helps constrain the RL optimization toward physically meaningful solutions.

SFT finetuning yields generated structures with a match rate of 56.92% to their relaxed states. In contrast, RL finetuning leads to a high match rate of 84.40%, meaning structures are generated closer to relaxed geometries. Examples in Figures 5 and 6 illustrate this contrast, as the RL-finetuned model produces highly symmetric structures that change little upon relaxation, while SFT-generated structures deform significantly in order to reach an energy minimum.

Table 1: Stability Evaluation

Model	E_{hull} [eV/atom] ↓	Metastability Rate (%)† ↑	Match Rate (%) ↑	RMSD (Å) ↓
FlowMM [6]	0.25 ± 0.94	25.66	46.78	0.053
DiffCSP [17]	$\mathbf{0.22 \pm 0.54}$	**30.84**	84.31	0.028
Qwen2.5	3.43 ± 10.96	5.03	31.02	0.052
Qwen2.5 SFT	0.37 ± 1.79	26.5	56.92	0.044
$\mathbf{E_{tot}}$ **Reward**				
Qwen2.5 RL	0.63 ± 0.49	7.6	**84.40**	**0.025**
Qwen2.5 SFT+RL	0.35 ± 1.26	27.29	77.59	0.037

† $E_{\text{hull}} < 0.1$ (eV/atom), predicted by UMA [15]

5.3 Validity, diversity and Novelty

Metric. We evaluate generated crystal structures using complementary metrics. **Validity** is assessed along two dimensions: structural validity, which requires that no two atoms overlap in positions (atoms must be farther apart than half the sum of their covalent radii), and compositional validity, which requires zero net charge. **Diversity** measures variability across the generated set by computing pairwise distances in the Matminer feature space [46]. **Novelty** quantifies how distinct a generated structure is from the training data, defined by the distance to its nearest neighbor in the training set Matminer feature space. All metrics are calculated directly on raw model outputs, without relaxation or post-processing.

Results. Table 2 summarizes the basic evaluation of generated structures. All methods achieve near-perfect structural validity, while compositional validity varies more widely: SFT ensures 100% validity, whereas RL tends to reduce compositional validity, which strengthens the possibility for reward hacking as mentioned. RL, however, improves diversity, both in composition and structural spaces. Novelty shows a trade-off, where RL seems to reduce structural novelty but increase compositional novelty.

Table 2: Validity, Diversity and Novelty Evaluation

Method	Validity (%) ↑		Diversity ↑		Novelty ↑	
	Structure	Composition	Structure	Composition	Structure	Composition
FlowMM [6]	0.99	0.90	0.78	15.64	071	0.54
DiffCSP [17]	0.99	0.82	0.89	15.64	0.81	0.52
Qwen2.5 [41]	0.99	0.85	0.92	15.66	0.73	0.49
Qwen2.5 SFT	1.00	1.00	0.912	15.64	0.56	0.33
$\mathbf{E_{tot}}$ **Reward**						
Qwen2.5 RL	1.00	0.74	1.2	16.25	0.40	0.61
Qwen2.5 SFT+RL	1.00	0.75	0.97	16.68	0.57	0.67

6 Future Work

A promising direction is to redefine the reward function. Instead of optimizing only for total potential energy, incorporating the energy above hull directly would better align training with thermodynamic stability. To avoid reduced novelty, future work would also explore mechanisms that encourage structural diversity during finetuning, balancing stability with exploration and thereby enhancing the likelihood of discovering novel stable materials.

Furthermore, rigorous evaluation of generated structures requires ground-truth assessments of formation energies, motivating the use of DFT calculations on a subset of generated candidates. Such evaluations would provide reliable validation beyond MLIPs.

References

[1] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

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

[3] Usman Shareef, Aisha Altaf, Madiha Ahmed, Nosheen Akhtar, Mohammed S Almuhayawi, Soad K Al Jaouni, Samy Selim, Mohamed A Abdelgawad, and Mohammed K Nagshabandi. A comprehensive review of discovery and development of drugs discovered from 2020–2022. *Saudi Pharmaceutical Journal*, 32(1):101913, 2024.

[4] Nofit Segal, Aviv Netanyahu, Kevin P Greenman, Pulkit Agrawal, and Rafael Gomez-Bombarelli. Known unknowns: Out-of-distribution property prediction in materials and molecules. *arXiv preprint arXiv:2502.05970*, 2025.

[5] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

[6] 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.

[7] 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.

[8] Yong Zhao, Edirisuriya M Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):38, 2023.

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

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

[11] Anubhav Jain, Yongwoo Shin, and Kristin A Persson. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials*, 1(1):1–13, 2016.

[12] Mercedes Boronat, Antonio Leyva-Perez, and Avelino Corma. Theoretical and experimental insights into the origin of the catalytic activity of subnanometric gold clusters: attempts to predict reactivity with clusters and nanoparticles of gold. *Accounts of chemical research*, 47(3): 834–844, 2014.

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

[14] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

[15] Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

[16] 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. *arXiv preprint arXiv:2503.03965*, 2025.

[17] 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.

[18] Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*, 2023.

[19] Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*, 2023.

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

[21] Nihang Fu, Lai Wei, Yuqi Song, Qinyang Li, Rui Xin, Sadman Sadeed Omee, Rongzhi Dong, Edirisuriya M Dilanga Siriwardane, and Jianjun Hu. Material transformers: deep learning language models for generative materials design. *Machine Learning: Science and Technology*, 4(1):015001, 2023.

[22] Anuroop Sriram, Benjamin Miller, Ricky TQ Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

[23] Izumi Takahara, Teruyasu Mizoguchi, and Bang Liu. Accelerated inorganic materials design with generative ai agents. *arXiv preprint arXiv:2504.00741*, 2025.

[24] Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Carla P Gomes, Kristin A Persson, Daniel Schwalbe-Koda, and Wei Wang. Large language models are innate crystal structure generators. *arXiv preprint arXiv:2502.20933*, 2025.

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

[26] Andy Xu, Rohan Desai, Larry Wang, Gabriel Hope, and Ethan Ritz. Plaid++: A preference aligned language model for targeted inorganic materials design. *arXiv preprint arXiv:2509.07150*, 2025.

[27] 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*, 36:53728–53741, 2023.

[28] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

[29] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[31] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, Charles J Moore, Christopher C Fischer, Kristin A Persson, and Gerbrand Ceder. Formation enthalpies by mixing gga and gga+ u calculations. *Physical Review B—Condensed Matter and Materials Physics*, 84(4):045115, 2011.

[32] Amanda Wang, Ryan Kingsbury, Matthew McDermott, Matthew Horton, Anubhav Jain, Shyue Ping Ong, Shyam Dwaraknath, and Kristin A Persson. A framework for quantifying uncertainty in dft energy corrections. *Scientific reports*, 11(1):15496, 2021.

[33] Yabi Wu, Predrag Lazic, Geoffroy Hautier, Kristin Persson, and Gerbrand Ceder. First principles high throughput screening of oxynitrides for water-splitting photocatalysts. *Energy & environmental science*, 6(1):157–168, 2013.

[34] Emilie Voisin, E Johan Foster, Muriel Rakotomalala, and Vance E Williams. Effects of symmetry on the stability of columnar liquid crystals. *Chemistry of Materials*, 21(14):3251–3261, 2009.

[35] Yao Chen, Pooya Sareh, and Jian Feng. Effective insights into the geometric stability of symmetric skeletal structures under symmetric variations. *International Journal of Solids and Structures*, 69:277–290, 2015.

[36] VS Urusov and TN Nadezhina. Frequency distribution and selection of space groups in inorganic crystal chemistry. *Journal of Structural Chemistry*, 50(Suppl 1):22–37, 2009.

[37] SV Borisov. On the symmetry-stability problem. *Journal of Structural Chemistry*, 36(6): 1061–1062, 1995.

[38] SV Borisov, SA Magarill, and NV Pervukhina. Crystallographic analysis of symmetry-stability relations in atomic structures. *Journal of Structural Chemistry*, 60(8):1191–1218, 2019.

[39] Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

[40] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[41] A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, et al. Qwen2. 5 technical report. *arXiv preprint*, 2024.

[42] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

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

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

[46] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

[47] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

A Supplementary Material

A.1 Implementation details of RL

We used GRPO implemented in `verl`[2]. The training batch size was 1024, with maximum prompt and response lengths of 1024 tokens each; overlong prompts were filtered and truncation errors were enforced to ensure data consistency. The model was initialized from Qwen2.5-7B-Instruct and optimized with a learning rate of (10^{-6}). PPO optimization used mini-batches of 256 and micro-batches of 8 per GPU. A KL regularization loss was applied during optimization (low-variance KL) with coefficient 0.001, while the KL term was not included directly in the reward; entropy regularization was disabled. Gradient checkpointing was enabled for memory efficiency. Rollouts were generated using vLLM with tensor model parallelism of 2, a GPU memory utilization cap of 0.8, and 5 sampled responses per prompt. Log-probability computation for actor and reference models used micro-batches of 8 per GPU. FSDP was used without parameter or optimizer offloading for the actor, while parameter offloading was enabled for the reference model.

A.2 Post Generation Relaxation Details

We perform a post-generation relaxation step using the UMA model as an ASE calculator [47]. Given a CIF string, we reconstruct the structure and evaluate its initial potential energy. We then run an LBFGS optimizer (wrapped with a Frechet cell filter) for up to 100 steps or until the forces fall below a convergence threshold of 0.02. The procedure outputs the initial (for the unprocessed generated structure) and relaxed energies, along with the fully relaxed structure.

A.3 Baselines

We compare with domain-specific baselines: FlowMM [6], DiffCSP [17]. Additionally, we run several variants of our LLM and RL method as an ablation.

For **DiffCSP**, we use the provided checkpoints in the DiffCSP repository of the conditional and non-conditional models and run evaluations. For **FlowMM**, we train ourselves the conditional and non-conditional versions, and run evaluations.

[2]https://github.com/volcengine/verl

A.4 Text Prompt

In all our experiments (both RL and SFT finetuning), LLM prompts are in the following form:

Generation Prompt

You are a material scientist expert in crystal structure prediction. Your task is to predict the stable structure of a given chemical formula {formula}. 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. Format your answer as lattice lengths, lattice angles, then element symbols with coordinates:

Example 1:

```
10.3 6.0 4.7
90 90 90
Li
0.25 0.50 0.75
Fe
0.75 0.50 0.25
P
0.50 0.00 0.50
O
0.10 0.60 0.40
O
0.90 0.40 0.60
O
0.40 0.90 0.10
O
0.60 0.10 0.90
```

Example 2:

```
5.2 5.2 11.8
90 90 120
Mg
0.33 0.67 0.25
Al
0.00 0.00 0.00
O
0.31 0.69 0.38
O
0.69 0.31 0.62
```

Provide ONLY the bulk representation like the example with no additional text.

A.5 Additional Generation results

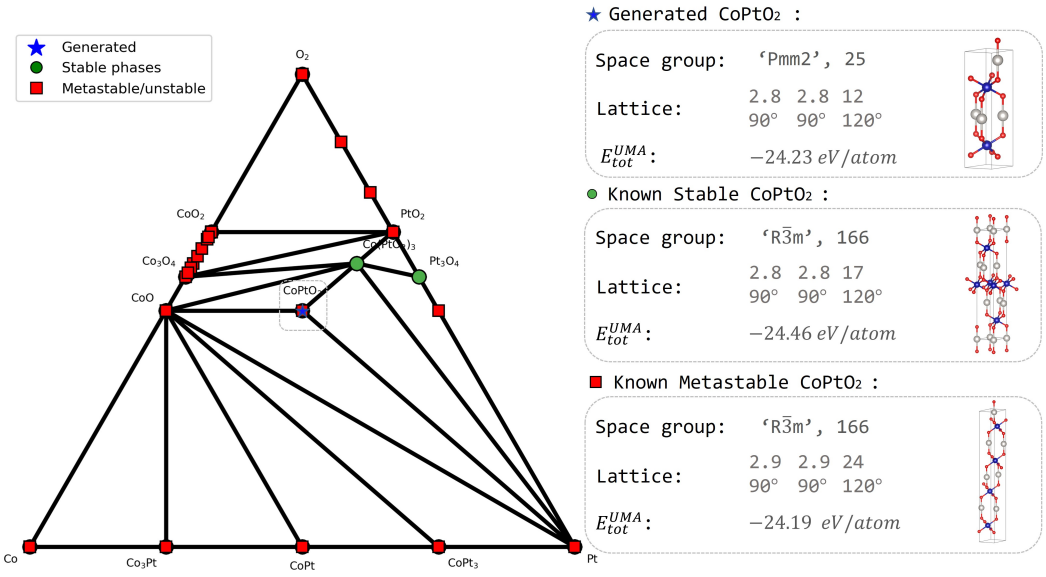
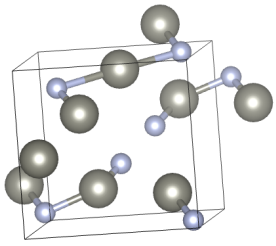
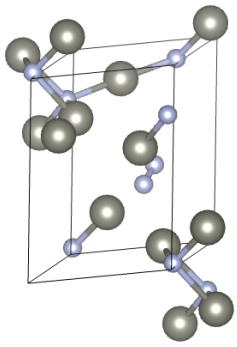


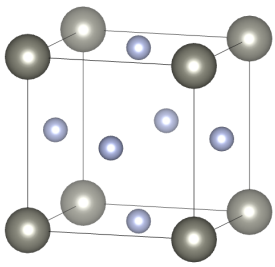
Figure 4: **Ternary phase diagram of the Co–Pt–O system.** Stable phases are shown as green circles and metastable/unstable phases as red squares, with only those satisfying $E_{\text{hull}} < -0.4$ eV/atom displayed. The blue star marks the composition of our generated material, $CoPtO_2$. Insets compare the three polymorphs of $CoPtO_2$, showing space group (symbol and number), lattice parameters (lengths and angles), and UMA-predicted total energies. The two known polymorphs have been experimentally observed. The UMA-predicted energy of the generated structure lies between those of the known polymorphs, suggesting it may also be a synthesizable metastable phase.



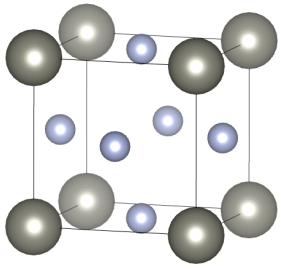
(a) SFT finetuned (Generated CIF)



(b) SFT finetuned (After relaxation)

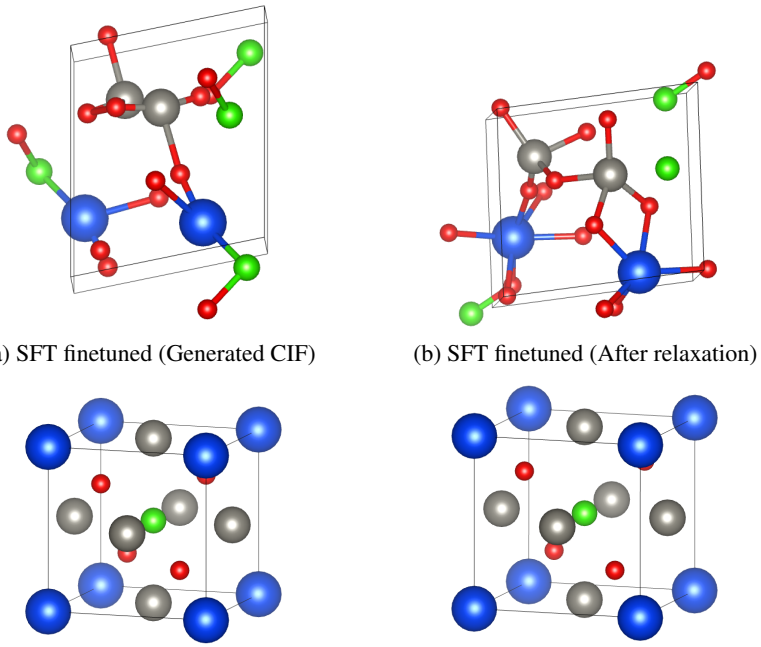


(c) RL finetuned (Generated CIF)



(d) RL finetuned (After relaxation)

Figure 5: Comparison of generated CIF structures with the elements Zn, N (left) and their UMA relaxed counterparts (right) for SFT and RL finetuned models.



(a) SFT finetuned (Generated CIF)

(b) SFT finetuned (After relaxation)

(c) RL finetuned (Generated CIF)

(d) RL finetuned (After relaxation)

Figure 6: Comparison of generated CIF structures with the elements $\mathrm{Ho}, \mathrm{W}, \mathrm{Cl}, \mathrm{O}$ (left) and their UMA relaxed counterparts (right) for the models finetuned via SFT and via RL.