MatLLMSearch: Crystal Structure Discovery with Evolution-Guided Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Crystal structure generation is fundamental to materials science, enabling the discovery of novel materials with desired properties. While existing approaches leverage Large Language Models (LLMs) through extensive fine-tuning on materials databases, we show that pre-trained LLMs can inherently generate novel and stable crystal structures without additional fine-tuning. Our framework employs LLMs as intelligent proposal agents within an evolutionary pipeline that guides them to perform implicit crossover and mutation operations while maintaining chemical validity. We demonstrate that MatLLMSearch achieves a 78.38% metastable rate validated by machine learning interatomic potentials and 31.7% DFT-verified stability, outperforming specialized models such as CrystalTextLLM. Beyond crystal structure generation, we further demonstrate that our framework adapts to diverse materials design tasks, including crystal structure prediction and multi-objective optimization of properties such as deformation energy and bulk modulus, all without fine-tuning. These results establish our framework as a versatile and effective framework for consistent high-quality materials discovery, offering training-free generation of novel stable structures with reduced overhead and broader accessibility.

1 Introduction

Crystal Structure Generation (CSG) and Prediction (CSP) represent critical bottlenecks in materials discovery, requiring both chemical validity and thermodynamic stability to determine whether a material can be synthesized (Bagayoko, 2014). These tasks demand navigating an expansive chemical space while satisfying multiple constraints: three-dimensional periodicity, proper atomic coordination, charge balance, and minimized formation energy. While computational approaches have emerged as indispensable tools for accelerating materials discovery (Dunn et al., 2020; Eremin et al., 2023), developing reliable systems that effectively explore this vast and complex space remains challenging.

Recent advances in deep learning have introduced various approaches for structure prediction, from variational autoencoders to diffusion models (Flam-Shepherd and Aspuru-Guzik, 2023; Gruver et al., 2024; Jiao et al., 2024; Xie et al., 2022; Zeni et al., 2025). Meanwhile, Large Language Models (LLMs) have emerged as powerful tools for materials discovery (Achiam et al., 2023; Antunes et al., 2023; Fu et al., 2023). Prior work (Flam-Shepherd and Aspuru-Guzik, 2023) demonstrated that autoregressive models using character-level tokenization can generate valid crystal structures, and Gruver et al. (2024) showing that fine-tuning pre-trained language models like Llama (Grattafiori et al., 2024) can produce physically stable structures.

Current approaches often fine-tune LLMs on materials databases such as the Materials Project (Gruver et al., 2024), which contains only tens of thousands of structures compared to the vast space of possible stable compounds. While we propose a fundamentally different perspective: Recognizing pre-trained LLMs not as tools requiring domain-specific fine-tuning, but as intelligent agents already possessing rich embedded knowledge from vast scientific corpora. This perspective raises the question: *How can we exploit the comprehensive scientific knowledge already embedded in pre-trained LLMs to build a system that can consistently produce valid stable crystal structures?*

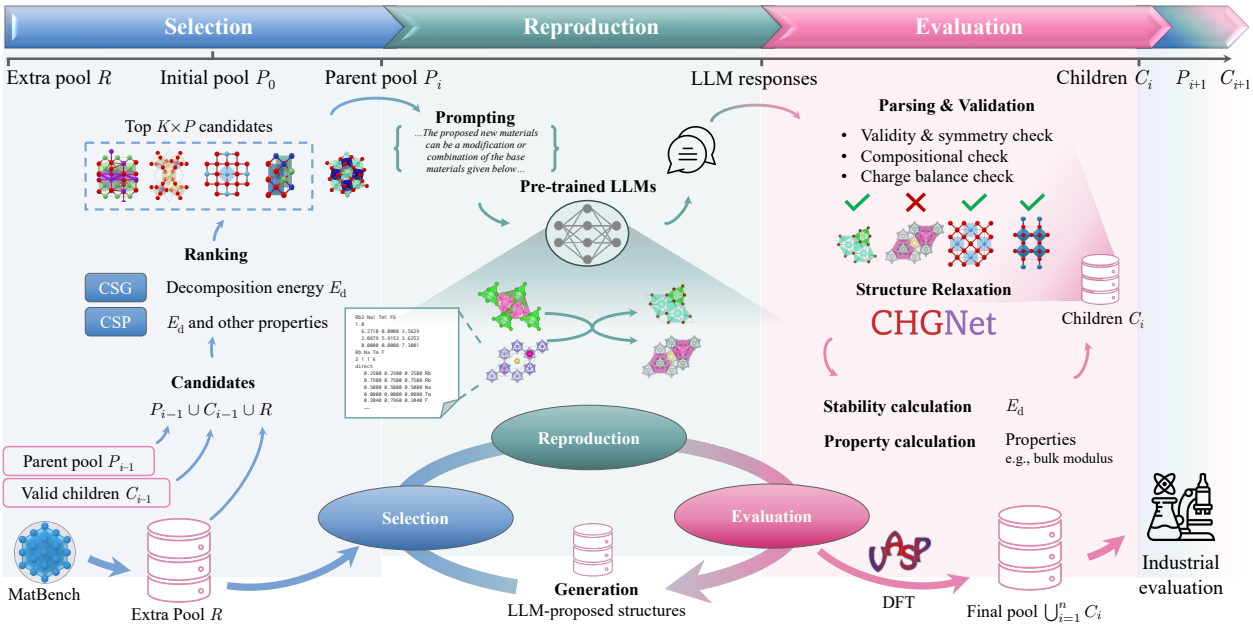


Figure 1: The workflow of MatLLMSearch for crystal structure generation. Starting from an initial population of known structures, our framework iteratively evolves new crystal structures through LLM-guided reproduction, evaluation, and selection.

Intuitively, we may directly prompt a commercial LLM to generate crystal structures. However, our ablational experiments in Section 4.4 and Appendix G across multiple configurations reveal that simple prompting fails to consistently generate valid crystal structures that are both stable and novel. These attempts often produce either copies of known structures, chemically invalid configurations, or thermodynamically unstable structures. The failure suggests that LLMs struggle to simultaneously satisfy the multiple constraints of CSG, indicating the need for a more sophisticated approach to exploit the potential of LLMs for materials discovery.

Evolutionary algorithms provide an effective framework for exploring the vast chemical space (Allahyari and Oganov, 2020; Oganov and Glass, 2006; Wang et al., 2024). By mimicking biological evolution through iterative selection, reproduction, and mutation operations, these algorithms can gradually improve the candidates, enabling automated property-guided materials optimization. Previous evolutionary approaches to CSG and CSP rely on explicit crossover and mutation operators, such as swapping structural motifs or introducing atomic displacements. While effective, these traditional operators lack the chemical intuition to efficiently navigate the complex constraints of crystal structures, often resulting in physically implausible candidates.

Our work advances this paradigm by leveraging the rich scientific knowledge embedded in LLMs to perform chemically-informed operations within the evolutionary algorithm framework. Unlike traditional operators that manipulate structures based on predefined rules, LLMs can implicitly reason about chemical bonding patterns, structural motifs, and stability principles learned from vast scientific literature (Boiko et al., 2023; Bran et al., 2024; Guo et al., 2023). This knowledge-guided approach enables more intelligent exploration of the chemical space, potentially discovering novel structures that traditional evolutionary methods might miss due to their limited chemical knowledge.

In this work, we introduce MatLLMSearch, a novel framework that integrates the rich scientific knowledge of LLMs into the evolutionary framework for crystal structure discovery. In our proposed framework, LLMs function as intelligent proposal agents that analyze parent structures to perform implicit crossover and mutation operations, while Machine Learning Interatomic Potentials (MLIPs) evaluate the physicochemical validity of generated structures. As illustrated in Figure 1, through iterative selection, reproduction, and evaluation, MatLLMSearch progressively discovers crystal structures with desired properties.

Our experiments demonstrate that MatLLMSearch successfully generates diverse, thermodynamically stable crystal structures while maintaining crystallographic validity. The framework achieves a 76.8% metastable structure generation rate, with 31.7% of structures verified as stable through DFT calculations, surpassing

the state-of-the-art fine-tuned model CrystalTextLLM (Gruver et al., 2024). Notably, this performance is achieved with minimal computational overhead, requiring only LLM inference and stability evaluation with MLIPs rather than extensive model training.

Beyond crystal structure generation, our framework demonstrates remarkable flexibility across various materials discovery tasks. Through simple modifications in prompting and reference seed structures selection, our method extends to CSP, which we validate by discovering several metastable Na_3AlCl_6 polymorphs with significantly higher stability than existing structures in the Materials Project database. Furthermore, the framework enables multi-objective optimization of properties such as bulk modulus, without requiring specialized fine-tuning. While we demonstrate results using general-purpose pre-trained LLMs, the framework could also incorporate domain-specialized fine-tuned models or alternative search algorithms, offering a computationally efficient approach to materials discovery with reduced overhead and broader accessibility.

2 Background: Computational Materials Discovery with Machine Learning

2.1 Problem Definition

Crystal Structure Generation (CSG). The objective of CSG is to learn a probability distribution $p(c, l, s)$ over crystalline materials, where $c \in \mathbb{R}^{N \times K}$ represents the chemical composition matrix for N atoms of K distinct chemical species, $l \in \mathbb{R}^6$ denotes the lattice parameters (lengths and angles), and $s \in \mathbb{R}^{N \times 3}$ defines the spatial coordinates of atoms within a periodic unit cell. Samples drawn from this distribution should ideally satisfy fundamental thermodynamic stability criteria.

Crystal Structure Prediction (CSP). CSP addresses a more constrained problem of determining stable crystal structures for a specified chemical composition. Formally, it learns a conditional probability distribution $p(s, l \mid c)$ to identify thermodynamically favorable atomic arrangements and lattice parameters given a fixed composition c. This formulation addresses the practical scenario of discovering stable polymorphs for a specified chemical formula.

Crystal Structure Design (CSD). CSD extends beyond structure prediction by incorporating property optimization and conditional generation. An example objective is finding the optimal crystal structure that maximizes a target property $h(c, l, s)$: $m^* = \text{argmax}_{c,l,s \sim p(c,l,s)} h(c, l, s)$, where $h : \mathbb{R}^{N \times K} \times \mathbb{R}^6 \times \mathbb{R}^{N \times 3} \to \mathbb{R}$ represents an oracle function evaluating the desired materials property. It can also be formulated as sampling from a tilted distribution $p(c, l, s) \exp(h(c, l, s))$ (Rafailov et al., 2024). Additional constraints can be integrated into the design process, allowing for flexible tasks such as compositional substitution (learning $p(c \mid l, s)$) and composition/structure completion (inpainting generation, learning $p(c^{\text{unknown}}, s^{\text{unknown}} \mid c^{\text{known}}, l, s^{\text{known}})$) (Dai et al., 2024).

2.2 (Meta)Stability of Materials

Among computational approaches for evaluating crystal structure stability, Density Functional Theory (DFT) is the most reliable method for predicting formation energies in solid-state materials, showing close alignment with experimental measurements (Jain et al., 2011; Sun et al., 2016). The thermodynamic stability of a structure is quantified through its decomposition energy (E_d) with respect to the convex hull of known stable phases: $E_\text{d} = E_\text{s} - \sum_i x_i E_i$, where E_s represents the total energy per atom, x_i denotes the molar fraction of the i-th competing phase, and E_i corresponds to its ground-state energy per atom. While the convex hull serves as a fixed reference, the evaluated structure s need not be part of this hull. A negative decomposition energy ($E_\text{d} < 0$) indicates a thermodynamically stable state below the convex hull, while $E_\text{d} > 0$ suggests a metastable phase with a driving force for decomposition into more stable compounds. Our main objective for CSG is to identify stable crystal structures where $E_\text{d} \leq 0$.

Given the computational intensity of DFT calculations, universal Machine Learning Interatomic Potentials (MLIPs), trained on millions of DFT calculations, have emerged as efficient and reliable proxies for structure stability assessment. Notable among these is CHGNet (Deng et al., 2023), a Graph Neural Network (GNN)-based MLIP that uniquely incorporates magnetic moments to capture both atomic and electronic interactions. M3GNet (Chen and Ong, 2022) offers an alternative approach, implementing three-body interactions in

its graph architecture for accurate structural predictions across diverse chemical spaces. Recent advances in universal MLIPs include MACE (Batatia et al., 2023), DPA-1 (Zhang et al., 2024), and JMP (Shoghi et al., 2024), which demonstrate high accuracy in predicting crystal thermodynamic stability, particularly when trained on industrial-scale datasets comprising millions of compounds and non-equilibrium atomic configurations (Barroso-Luque et al., 2024; Merchant et al., 2023; Yang et al., 2024a). In this work, we employ the pre-trained CHGNet as our universal MLIP due to its closer alignment with DFT results, using a fixed phase diagram derived from the Materials Project 2023 DFT calculations (Jain et al., 2011; Wang et al., 2021).

3 MatLLMSearch

We propose MatLLMSearch, an evolutionary workflow that leverages pre-trained LLMs to search for stable and optimized crystal structures with. In this section, we introduce three key stages of the workflow as illustrated in Figure 1: (1) **Selection**, which identifies promising candidate structures from existing pools based on stability and property metrics; (2) **Reproduction**, where the LLM generates new candidates through implicit crossover and mutations of parent structures; and (3) **Evaluation**, which assesses proposed structures for validity, stability, and target properties. The overall workflow, outlined in Appendix L, iteratively evolves a population of crystal structures while maintaining physical constraints and optimizing desired properties.

3.1 Initialization

Our evolutionary search begins by sampling $(K \times P)$ structures from a database of known stable structures $\mathcal{D}$ to form our initial parent pool $\mathcal{P}_0$. These structures are organized into K groups of P parents each to serve as reference examples in LLM prompts. We optionally retrieve an extra pool $\mathcal{R}$ from $\mathcal{D}$ to expand the candidate space during selection. $\mathcal{R}$ can be customized to suit various design objectives, with more details and ablation studies provided in Section 4.4.

3.2 Reproduction

Genetic algorithms traditionally mimic biological evolution through explicit crossover and mutation operations (Heiles and Johnston, 2013; Johnston, 2003). In crystal structure prediction, crossover typically involves combining structural fragments from parent structures (e.g., swapping atomic positions or structural motifs), while mutation introduces random variations through predefined operations like atomic displacement, lattice transformation, or element substitution (Curtis et al., 2018; Kadan et al., 2023). While effective, these rigid operators can limit the exploration of the complex crystal structure space. In MatLLMSearch, we explore the flexibility of LLMs for structure reproduction. Through prompt-based guidance, we ask LLMs to perform implicit crossover and mutation by analyzing and combining structural information from parent materials. Specifically, LLMs are instructed to "modify or combine the base materials", while maintaining chemical validity and enhancing target properties. This approach allows LLMs to freely and simultaneously introduce variations across multiple structural aspects, including atomic positions, lattice parameters, and element substitutions, or even generate completely new structures functionally relevant to parent structures.

3.3 Evaluation

Our evaluation pipeline consists of two stages:

- **Rule-based validation** ensures structural integrity by verifying three-dimensional periodicity, physical connectivity (interatomic distances between 0.6-1.3 times the sum of atomic radii), and chemical validity through charge balance analysis.
- **Stability and property evaluation** begins with structure relaxation using CHGNet. We quantify thermodynamic stability through decomposition energy E_d calculated as the distance to the Materials Project convex hull. Notably, we observe that LLM-proposed structures typically require minimal relaxation, with 61.1% of structures exhibiting small energy changes ($|\Delta E| < 0.5$ eV/atom) during this

process (detailed in Appendix H), indicating their initial stability. For stability-focused optimization, we quantify thermodynamic stability through the decomposition energy E_{d} using CHGNet, calculated as the distance to the convex hull from the Materials Project database (version `2023-02-07-ppd-mp`). For multi-objective optimization, additional properties such as bulk modulus can be evaluated. These quantitative scores then guide the selection process for subsequent generations, allowing our framework to flexibly adapt to different design goals.

3.4 Selection

Last, the selection stage evolves a population of candidate structures that meet the optimization objectives, such as thermodynamic stability or other desired physical properties. For each iteration i, we construct a new parent pool $\mathcal{P}_{i+1}$ of the same size ($K \times P$) by selecting top-ranked candidates from three sources: the current parent pool ($\mathcal{P}_i$), newly generated children structures ($\mathcal{C}_i$), and an optional extra pool ($\mathcal{R}$) to improve diversity. Candidates in $\mathcal{P}_i \cup \mathcal{C}_i \cup \mathcal{R}$ are ranked according to optimization objectives, either single-objective (e.g., E_{d} for stability) or multi-objective criteria (e.g., alternating between different properties).

3.5 Final DFT Verification

After completing all evolutionary iterations, we collect the cumulated offspring structures $\mathcal{S} = \bigcup_i \mathcal{C}_i$ for final validation using DFT. To save computational cost, we focus on meta-stable structures with CHGNet-predicted decomposition energy $E_d < 0.1$ eV/atom. DFT calculations are performed using VASP 6 in the Generalized Gradient Approximation (GGA) with PBE functional (Perdew et al., 1996), using the projector-augmented wave method (Kresse and Furthmüller, 1996; Kresse and Joubert, 1999). We employed a plane-wave basis set with an energy cutoff of 520 eV and a k-point mesh of 1,000 per reciprocal atom (Jain et al., 2013). The calculations converged to 10^{-6} eV in total energy for electronic self-consistent field cycles and 0.02 eV/Å in interatomic forces for the ionic steps. The computational settings are consistent with MPGGARelaxSet and MPGGAStaticSet (Jain et al., 2011).

4 Experiments

4.1 Experimental Settings

We use Llama 3.1 (70B) (Grattafiori et al., 2024) as the base LLM. We set temperature to 0.95 to balance creativity and reliability. All experiments use parent size $P = 2$, reproduction size $C = 5$, and $N = 10$ iterations, with population size $K = 100$ unless otherwise specified. Crystal structures are represented in POSCAR format with 12 decimal digits.

For initialization, we use the MatBench dataset (Dunn et al., 2020) as the known stable structure set $\mathcal{D}$. We sample some known stable structures from $\mathcal{D}$ to form the initial generation and fill the parents during each iteration. These samples are chosen based on their CHGNet-predicted decomposition energy or the properties to be optimized. Detailed ablation studies regarding this selection policy are provided in Appendix A.

4.2 Main Experimental Results

In this section, we evaluate our proposed pipeline on progressively more challenging tasks, from crystal structure generation through design to crystal structure prediction.

Crystal structure generation. We first evaluate the ability of our framework to generate stable crystal structures by optimizing decomposition energy E_{d} as the sole objective. The LLM prompting template is detailed in Appendix C.

The generation results are reported in Table 1. Following previous work (Gruver et al., 2024; Xie et al., 2022), we report structural and compositional validity, which assess non-overlapping atomic radii and charge neutrality respectively. Metastability is evaluated using both CHGNet and M3GNet as surrogate models, measuring the percentage of structures with decomposition energies below 0.1 eV/atom and 0.03 eV/atom

Model	f-ele in Parents[†]	Validity		Metastability			Stability[‡]	
		Structural	Composition	M3GNet	CHGNet		DFT	
				$E_d < 0.1$	$E_d < 0.1$	$E_d < 0.03$	w/ f-ele	w/o f-ele[§]
CDVAE*	—	100.0%	86.7%	28.8%	—	—	5.4%	—
CrystalTextLLM-7B*	—	96.4%	93.3%	35.0%	—	—	8.7%	—
CrystalTextLLM-13B*	—	95.5%	92.4%	38.0%	—	—	14.4%	—
CrystalTextLLM-70B*	—	99.6%	95.4%	49.8%	—	—	10.6%	—
MatLLMSearch (Llama 3.1-70B)	✓	100.0%	79.4%	81.1%	76.8%	56.5%	31.7%	14.0%
	✗	100.0%	89.0%	81.9%	78.4%	54.8%	27.0%	24.6%

Table 1: Performance comparison of crystal structure generation. Metastability is first assessed using surrogate models, where we report both M3GNet and CHGNet results for fair comparison with baselines CDVAE and CrystalTextLLM (which use M3GNet). *Results taken from the original papers. [†]Indicates whether f-electron elements are excluded in parent structures (not applicable to CDVAE and CrystalTextLLM as they are trained on data including f-electron elements). [‡]The stable fraction represents the percentage of DFT-verified stable structures ($E_d < 0.0$ eV/atom) over structures predicted to be metastable ($E_d < 0.1$ eV/atom) by respective surrogate models (M3GNet for CDVAE and CrystalTextLLM, CHGNet for ours, with CHGNet being more rigorous as evidenced by lower metastability rates). [§]We exclude structures containing f-electron in DFT verification while keeping the denominator as all metastable structures.

thresholds. Structures identified as metastable ($E_d < 0.1$ eV/atom) by CHGNet undergo further DFT calculations for stability assessment.

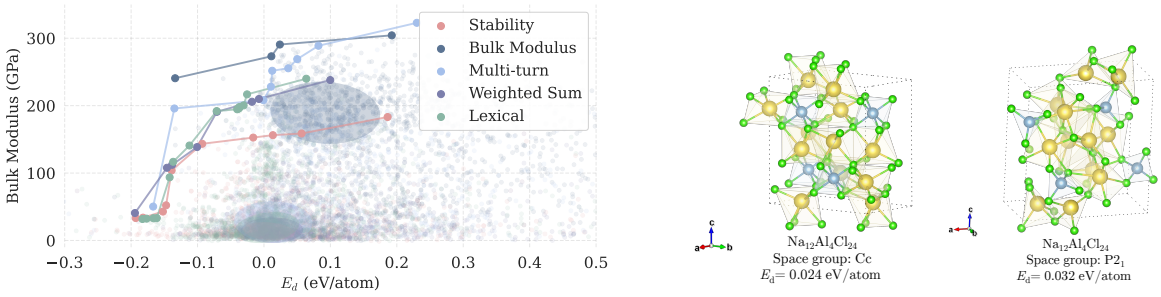
We compare our model against two baseline models CDVAE (Xie et al., 2022) and CrystalTextLLM (Gruver et al., 2024). Among 1,479 generated structures, 76.8% and 81.1% are metastable based on CHGNet and M3GNet evaluations respectively, outperforming the 49.8% metastability rate by M3GNet of the state-of-the-art CrystalTextLLM 70B model, which has a comparable model size to our base model. Under rigorous DFT validation, 31.7% of the metastable structures remain stable, substantially improving the 10.6% stability rate from CrystalTextLLM 70B.

However, structures containing f-electron elements (actinides and lanthanides, abbreviated as f-ele) challenge stability prediction with their strongly correlated electron interactions, which may not be adequately captured by DFT approaches under GGA and Hubbard U corrections (Ani). These structures consistently yield lower decomposition energies (E_d), creating a potential computational shortcut. To assess this effect, we report the percentage of stable structures without f-ele (denoted as "w/o f-ele") among the metastable structures. By excluding f-electron structures from parent selection (marked with ✗), we improved metastability rates to 78.4% and increased stable non-f-electron structures from 14.0% to 24.6%. This simple intervention demonstrates our framework's ability to effectively guide exploration toward diverse stable configurations, which remain unaddressed by existing methods.

While achieving better performance, our method also offers significant computational advantages. Compared to CrystalTextLLM which requires extensive fine-tuning on more than 120K structures, we achieve higher stability rates using only a few reference structures and direct LLM inference. The computational cost is primarily from structure evaluation rather than model training or fine-tuning, making our approach more accessible.

Crystal structure design. We also explore multi-objective optimization by extending our framework to balance stability with desired material properties. We demonstrate this capability by alternating between optimizing stability (E_d) and bulk modulus in each iteration. While this multi-objective setting naturally yields lower stability rates (57.1% metastable with $E_d < 0.1$ eV/atom and 15.6% DFT-verified stable structures with f-electron elements) compared to stability-only optimization, it enables the discovery of structures with favorable property-stability trade-offs.

As shown in Figure 2(a), the Pareto frontiers under various optimization strategies converge in regions with high bulk modulus (> 200 GPa) and metastability ($E_d \leq 0.1$ eV/atom) in the stability-property space, indicating successful discovery of potentially valuable structures that balance both objectives. The



(a) Pareto frontiers

(b) CSP Examples

Figure 2: (a) Pareto frontiers of bulk modulus versus decomposition energy (E_d) for structures optimized towards stability, bulk modulus and multi-objective (multi-turn). Ellipses indicate regions of highest structure density. (b) Examples of predicted crystal structures with composition Na_3AlCl_6.

regions of highest structure density, estimated using Gaussian KDE and visualized as ellipses, reveal how optimization goals affect the distribution. Prioritizing bulk modulus shifts the density distribution toward higher mechanical strength at the cost of increased decomposition energy. We provide additional discussions of multi-objective optimization strategies in Appendix D.

Crystal structure prediction. We next evaluate our framework on crystal structure prediction tasks, which aim to predict stable structure (i.e. lattice and atomic coordinates) for a given composition. As a case study, we prompt the LLM to predict polymorphs of Na_3AlCl_6. For context, the Materials Project database currently contains only one structure for this composition (`mp-1111450`, $Fm\bar{3}m$, $E_\mathrm{d} = 0.142$ eV/atom), which is significantly unstable.

During the prompting process, we apply specific structural filters to select seed structures containing only three distinct elements in a 3:1:6 ratio, matching the stoichiometry of Na_3AlCl_6. From MatBench, we identified 820 structures meeting the criteria to build the initial population. Example structures proposed by the LLM for this composition are visualized in Figure 2(b), with DFT-verified decomposition energies of 0.024 and 0.032 eV/atom respectively. Although these predicted polymorphs remain metastable, their decomposition energies E_d are significantly lower than the previously reported structure in MatBench (E_d reduced by up to 83%), exemplifying the potential of our evolutionary pipeline for CSP applications. We provide more successful structure prediction demonstrations including Ag_6O_2, Bi_2F_8, etc. in the Appendix K.

4.3 CSG Quality Evaluation Metrics

4.3.1 Structural and Compositional Diversity

To better understand the effectiveness of our MatLLMSearch, we analyze the diversity of generated structures and evaluate different prompting strategies, with and without fine-tuning. Additional ablation studies on factors affecting generation performance are presented in Appendices I and J.

To evaluate the diversity of our generated structures, we analyzed their compositional and structural characteristics by comparing LLM-proposed structures and with the $K \times P$ most stable structures from MatBench that forms the initial generation. Our element co-occurrence analysis reveals high compositional diversity in the LLM-proposed structures, with even the most frequent compositions appearing only twice (approximately 0.14% of total structures). Examination of element co-occurrences with F in Figure 3 highlights the effectiveness of our evolutionary method in guiding structure generation toward stable F-based compounds particularly with alkali metals and transition metals.

The structural diversity is further evidenced by space group distribution and stability analysis for our **Stability** configuration. The generated structures demonstrates broad structural diversity with high metastability rates across multiple space groups, confirming that our evolutionary method successfully navigates toward

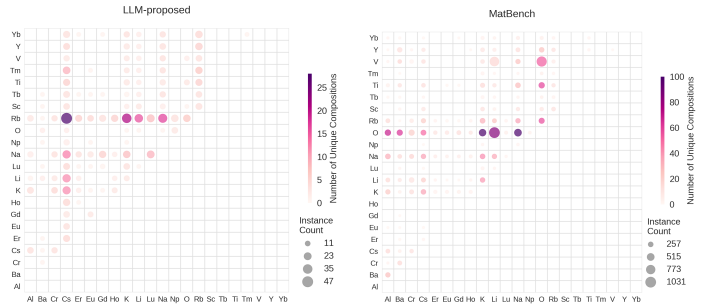


Figure 3: Element co-occurrence patterns with fluorine (F) in LLM-proposed structures (left) versus MatBench structures (right). Bubble size indicates frequency of occurrence for each element pair, while color intensity represents compositional diversity (darker indicates more unique compositions with that element pair).

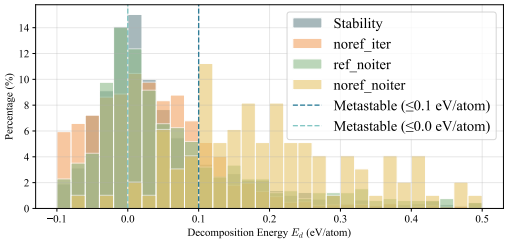


Figure 4: Decomposition energy (E_{d}) distribution comparison across experimental configurations. Vertical lines indicate metastable thresholds at 0.0 eV/atom (stable) and 0.1 eV/atom (metastable). Reference-guided approaches show more balanced distributions.

stable regions of chemical space while maintaining diverse structural motifs across different crystallographic symmetries. Additional diversity and novelty evaluations and analyses are provided in Appendix E.

4.3.2 Comparison with Baseline Methods

To ensure fair comparison with existing methods, we follow the same evaluation protocol as FlowMM (Zeni et al., 2025). We pre-relax all structures using CHGNet followed by density functional theory relaxation for stability verification. The results in Table 2 demonstrate that our framework maintains higher stability rate and S.U.N. (Stable, Unique, Novel) rate than the two baselines, achieving 24.34% stability rate and 23.60% S.U.N. rate compared to DiffCSP and FlowMM. As S.U.N. depends on chosen reference sets and collapses multi-dimensional quality into counts, we treat it as a supplementary signal alongside multi-MLIP metastability and DFT verification. In addition, we provide comprehensive evaluation across all S.U.N. dimensions in Appendix E: metastability rates with multiple MLIPs (CHGNet, M3GNet and Orb-v3 for CSP) and DFT-calculated stability rates for Stability; space group distributions and crystal system diversity for Uniqueness; compositional and structural novelties, elemental co-occurrence pattern shifts for Novelty. Further discussions on comparison fairness provided in Appendix G and additional generation time overhead analysis is provided in Appendix F.

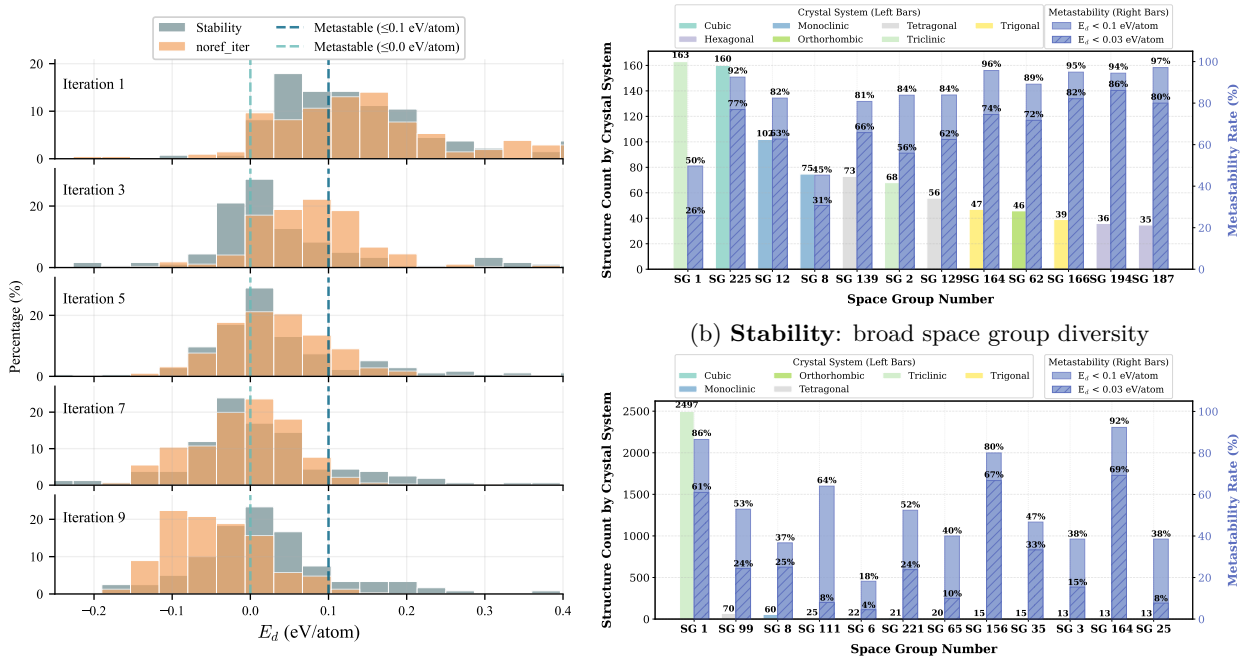
Method	**Stability Rate** ($E_d < 0.0$ eV/atom, %)	**S.U.N. Rate** (%)
MatLLMSearch (Llama 3.1-70B)	24.34	23.60
DiffCSP	5.06	3.34
FlowMM	4.65	2.34

Table 2: S.U.N. rate of generated structures. S.U.N. computed against MatBench reference structures used in our workflow.

4.4 Ablation Analysis

We evaluate four configurations to isolate the impact of reference structures and evolutionary iterations: **Stability** (full framework), **noref_iter** (evolutionary search without references), **ref_noiter** (single-iteration with references), and **noref_noiter** (zero-shot). Decomposition energy distributions (Figure 4) reveal that reference-guided approaches (**Stability**, **ref_noiter**) tightly concentrate structures near metastable thresholds, while evolutionary search alone (**noref_iter**) shows a similar but slightly dispersed distribution. In contrast, **noref_noiter** generates structures with substantially inferior metastability.

Impact of evolutionary iterations. Evolutionary iterations enable practical generation volumes: **Stability** produces 1,479 structures versus 741 for **ref_noiter**, showing 2× higher productivity under equivalent



(a) E_{d} distribution evolution

(b) **Stability**: broad space group diversity

(c) **noref_iter**: concentrated in space group 1

Figure 5: Ablation analysis comparing reference-guided (**Stability**) vs. reference-free (**noref_iter**) generation. (a) E_{d} distributions across iterations. (b,c) Space group diversity: reference structures enable broad space group diversity with high metastability (b); w/o references collapse to space group 1 (c).

computational budget (1,000 LLM inferences). As we investigate the distribution shift over iterations (Figure 5a), we observe that iterative optimization progressively shifts Ed distributions toward stability.

Impact of reference structures. While evolutionary search alone (**noref_iter**) achieves high metastability in later iterations, it suffers structural collapse with 88% of generated structures concentrate in triclinic space group 1 and limited exploration of other crystallographic symmetries (Figure 5c), indicating convergence to a single crystallographic motif. In contrast, reference-guided **Stability** configuration maintains comparable high metastability while distributing structures across 12 major space groups (Figure 5b). This demonstrates that reference structures prevent premature convergence while preserving thermodynamic quality.

In general, **noref_iter** achieves high metastability but diversity collapse, **ref_noiter** achieves structural diversity but limited volume, while **noref_noiter** fails on both dimensions (20.7% metastability, minimal volume). By combining reference structures with evolutionary search, **Stability** simultaneously achieves high metastability, balanced structural diversity, and practical generation scale. Additional analysis including novelty metrics, model scaling effects, and fine-tuning comparisons appears in Appendix G.

5 Related Work

5.1 Language Models for Materials Science

The increasing capabilities of LLMs have prompted materials science community to explore their potential for understanding and predicting material properties (Jablonka et al., 2023). However, benchmarking studies suggest fine-tuning LLMs over specific materials datasets is necessary to achieve performance comparable to or better than specialized graph neural networks (Rubungo et al., 2024). Research in crystal structure generation has developed along two main paths. Flam-Shepherd and Aspuru-Guzik (2023) demonstrate that autoregressive models trained from scratch with character-level tokenization can generate chemically valid crystal structures by directly tokenizing CIF files into string sequences. Secondly, CrystalTextLLM (Gruver

et al., 2024) fine-tunes a pre-trained LLM (over massive texts) on generating crystalline structures with task-specific prompts. Mat2Seq (Yan et al., 2024) converts 3D crystal structures into unique 1D sequences that preserve $SE(3)$ and periodic invariance for language model training. While these approaches produce valid structures, they sacrifice the general conversation capabilities of LLMs due to specialized training or fine-tuning on crystallographic data. In parallel developments within molecular chemistry, MolLEO (Wang et al., 2025) successfully employs pre-trained LLMs without domain-specific fine-tuning to search for small molecules. Subsequent work (Lu et al., 2024) extended this evolutionary optimization approach to more complex transition metal chemistry using advanced base LLMs with enhanced reasoning capabilities. However, these applications benefit from natural string representations for molecules (e.g., SMILES or SELFIES), which are considerably simpler than the three-dimensional representations required for crystal structures. Our work bridges this gap by adapting the evolutionary approach to the more complex domain of crystal structures without requiring fine-tuning.

5.2 Generative Models for Materials Discovery

Besides autoregressive language models, various generative models including variational autoencoders, diffusion models, and flow models have emerged as promising solutions for crystal structure generation. Early work proposes generative crystal structures using variational autoencoders that represent crystal structures as 3D voxels (Court et al., 2020; Noh et al., 2019). CDVAE first proposes to generate crystal structures with a score-based generative (diffusion) model and optimize crystal structure properties through gradient-based optimization in the latent space (Xie et al., 2022). This approach has been extended in several directions: Jiao et al. (2024) developed Riemannian diffusion models to better handle periodic coordinates, Zeni et al. (2025) scaled the approach to encompass elements across the entire periodic table with various design criteria, and Dai et al. (2024) applied it to crystal inpainting tasks. Most recently, Sriram et al. (2024) introduced Riemannian flow matching models to better address periodic boundary conditions with improved performance. Yang et al. (2024b) explore the synergy between language and generative models by leveraging LLMs to propose chemical formulae under design constraints before feeding them to a diffusion model.

5.3 Classical Substitution Methods

Classical substitution techniques generate candidates by replacing elements within known prototypes according to heuristic rules or ionic radii/valence compatibility. Such methods are efficient and ensure validity within predefined chemical spaces, but they are limited by the expressivity of the rule set and rarely propose nuanced edits such as partial substitutions or motif-level reorganizations. Our training-free LLM-based framework differs in goal and mechanism: by conditioning on multiple parents and free-form text structure strings, it proposes chemically coherent edits beyond strict substitution, while evolutionary selection guides validity and stability. We regard substitution as a complementary initialization strategy rather than a direct baseline to our broader search capability.

6 Conclusion

In this paper, we present an evolutionary workflow for computational materials discovery, encompassing crystal structure generation, prediction, and objective-based optimization. We demonstrate that a pre-trained LLM trained on general text can identify a higher proportion of (meta)stable materials compared to state-of-the-art generative models specifically trained on materials datasets. These findings suggest that LLMs inherently function as effective crystal structure generators, with both compositional and structural information naturally embedded within their text inference capabilities. In conclusion, our method complements existing structure discovery techniques by providing refined optimization capabilities while maintaining versatility in addressing various optimization objectives, offering an efficient approach for high-throughput materials discovery.

Looking forward, a natural extension of this work would be synthesis prediction based on the evolutionary method. Improved machine learning interatomic potentials will complement this process, as discussed in Appendix M. Such development would benefit from integration with high-quality experimental data

from automated, high-throughput experiments, bridging the gap between computational predictions and experimental synthesis, which would accelerate high-throughput materials discovery.

References

Hubbard-corrected dft energy functionals: The lda + u description of correlated systems. 6

Josh Achiam, Steven Adler, Sandhini Agarwal, et al. GPT-4 Technical Report. *arXiv.org*, 2023. 1

Zahed Allahyari and Artem R. Oganov. Coevolutionary Search for Optimal Materials in the Space of All Possible Compounds. *npj Comput. Mater.*, 2020. 2

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal Structure Generation with Autoregressive Large Language Modeling. *Nat. Commun.*, 2023. 1

Diola Bagayoko. Understanding Density Functional Theory (DFT) and Completing It in Practice. *AIP Adv.*, 2014. 1

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, et al. Open Materials 2024 (OMAT24) Inorganic Materials Dataset and Models. *arXiv.org*, 2024. 4, 24

Ilyes Batatia, Philipp Benner, Yuan Chiang, et al. A Foundation Model for Atomistic Materials Chemistry. *arXiv.org*, 2023. 4, 24

Simon Batzner, Albert Musaelian, Lixin Sun, et al. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.*, 2022. 24

Erik Bitzek, Pekka Koskinen, Franz Gähler, et al. Structural Relaxation Made Simple. *Phys. Rev. Lett.*, 2006. 24

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous Chemical Research with Large Language Models. *Nature*, 624(7992):570–578, December 2023. 2

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting Large Language Models with Chemistry Tools. *Nat. Mach. Intell.*, 2024. 2

Chi Chen and Shyue Ong. A Universal Graph Deep Learning Interatomic Potential for the Periodic Table. *Nat. Comput. Sci.*, 2022. 3, 24

Bingqing Cheng. Cartesian Atomic Cluster Expansion for Machine Learning Interatomic Potentials. *npj Comput. Mater.*, 2024a. 24

Bingqing Cheng. Response Matching for Generating Materials and Molecules. *J. Chem. Theory Comput.*, 2024b. 25

Callum J. Court, Batuhan Yildirim, Apoorv Jain, et al. 3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning. *J. Chem. Inf. Model.*, 2020. 10

Farren Curtis, Xiayue Li, Timothy Rose, et al. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *J. Chem. Theory Comput.*, 2018. 4

Xinzhe Dai, Peichen Zhong, Bowen Deng, et al. Inpainting Crystal Structure Generations with Score-Based Denoising. In *ICML Workshop on AI for Science*, 2024. 3, 10

Bowen Deng, Peichen Zhong, KyuJung Jun, et al. CHGNet as a Pretrained Universal Neural Network Potential for Charge-Informed Atomistic Modelling. *Nat. Mach. Intell.*, 2023. 3, 24, 25

Yuanqi Du, Limei Wang, Dieqiao Feng, et al. A New Perspective on Building Efficient and Expressive 3D Equivariant Graph Neural Networks. *NeurIPS*, 2023a. 24

Yuanqi Du, Yingheng Wang, Yining Huang, et al. M^2Hub: Unlocking the Potential of Machine Learning for Materials Discovery. *NeurIPS*, 2023b. 24

Alexander Dunn, Qi Wang, Alex Ganose, et al. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *npj Comput. Mater.*, 2020. 1, 5, 15

Roman Eremin, Innokentiy Humonen, Alexey Kazakov, et al. Graph Neural Networks for Predicting Structural Stability of Cd- and Zn-doped λ-$CsPbI_3$. *Comput. Mater. Sci.*, 2023. 1

Daniel Flam-Shepherd and Al'an Aspuru-Guzik. Language Models Can Generate Molecules, Materials, and Protein Binding Sites Directly in Three Dimensions as XYZ, CIF, and PDB Files. *arXiv.org*, 2023. 1, 9, 20

Nihang Fu, Lai Wei, Yuqi Song, et al. Material Transformers: Deep Learning Language Models for Generative Materials Design. *Mach. Learn.: Sci. Technol.*, 2023. 1

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. *arXiv.org*, 2024. 1, 5

Nate Gruver, Anuroop Sriram, Andrea Madotto, et al. Fine-Tuned Language Models Generate Stable Inorganic Materials as Text. In *ICLR*, 2024. 1, 3, 5, 6, 9, 18, 20

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. What Can Large Language Models Do in Chemistry? A Comprehensive Benchmark on Eight Tasks. In *NeurIPS*, 2023. 2

Sven Heiles and Roy L. Johnston. Global Optimization of Clusters Using Electronic Structure Methods. *Int. J. Quantum Chem.*, 2013. 4

Kevin M. Jablonka, Qianxiang Ai, Alexander Al-Feghali, et al. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digit. Discov.*, 2023. 9

Anubhav Jain, Geoffroy Hautier, Shyue P. Ong, et al. Formation Enthalpies by Mixing GGA and GGA + U Calculations. *Phys. Rev. B*, 2011. 3, 4, 5

Anubhav Jain, Shyue P. Ong, Geoffroy Hautier, et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.*, 2013. 5

Rui Jiao, Wenbing Huang, Peijia Lin, et al. Crystal Structure Prediction by Joint Equivariant Diffusion. *NeurIPS*, 2024. 1, 10, 20

Roy L. Johnston. Evolving Better Nanoparticles: Genetic Algorithms for Optimising Cluster Geometries. *Dalton Trans.*, 2003. 4

Amit Kadan, Kevin Ryczko, Andrew Wildman, et al. Accelerated Organic Crystal Structure Prediction with Genetic Algorithms and Machine Learning. *J. Chem. Theory Comput.*, 2023. 4

G. Kresse and J. Furthmüller. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B*, 1996. 5

G. Kresse and D. Joubert. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B*, 1999. 5

Yi-Lun Liao, Brandon Wood, Abhishek Das, et al. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *ICLR*, 2024. 24

Jon L'opez-Zorrilla, Xabier M. Aretxabaleta, In Won Yeu, et al. ænet-PyTorch: A GPU-Supported Implementation for Machine Learning Atomic Potentials Training. *J. Chem. Phys.*, 2023. 24

Jieyu Lu, Zhangde Song, Qiyuan Zhao, et al. Generative Design of Functional Metal Complexes Utilizing the Internal Knowledge of Large Language Models. *arXiv.org*, 2024. 10

Amil Merchant, Simon Batzner, Samuel Schoenholz, et al. Scaling Deep Learning for Materials Discovery. *Nature*, 2023. 4, 24

Juhwan Noh, Jaehoon Kim, Helge S. Stein, et al. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter*, 2019. 10

Artem Oganov and Colin Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics*, 124:244704, 07 2006. doi: 10.1063/1.2210932. 2

Yutack Park, Jaesun Kim, Seungwoo Hwang, et al. Scalable Parallel Algorithm for Graph Neural Network Interatomic Potentials in Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 2024. 24

Max Peeperkorn, Tom Kouwenhoven, Daniel G. Brown, et al. Is Temperature the Creativity Parameter of Large Language Models? In *ICCC*, 2024. 22

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 1996. 5

Chris J. Pickard and R. J. Needs. Ab Initio Random Structure Searching. *J. Phys.: Condens. Matter*, 2011. 25

Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. *NeurIPS*, 2024. 3

Zekun Ren, Siyu I. P. Tian, Juhwan Noh, et al. An Invertible Crystallographic Representation for General Inverse Design of Inorganic Crystals with Targeted Properties. *Matter*, 2022. 25

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale, 2025. URL https://arxiv.org/abs/2504.06231. 23

Andre N. Rubungo, Kangming Li, Jason Hattrick-Simpers, et al. LLM4Mat-Bench: Benchmarking Large Language Models for Materials Property Prediction. *arXiv.org*, 2024. 9

Nima Shoghi, Adeesh Kolluru, John R. Kitchin, et al. From Molecules to Materials: Pre-Training Large Generalizable Models for Atomic Property Prediction. In *ICLR*, 2024. 4, 24

Anuroop Sriram, Benjamin K. Miller, Ricky T. Q. Chen, et al. FlowLLM: Flow Matching for Material Generation with Large Language Models as Base Distributions. In *NeurIPS*, 2024. 10, 20

Wenhao Sun, Stephen T. Dacek, Shyue P. Ong, et al. The Thermodynamic Scale of Inorganic Crystalline Metastability. *Sci. Adv.*, 2016. 3

Nathan J. Szymanski and Christopher J. Bartel. Establishing baselines for generative discovery of inorganic crystals. *Materials Horizons*, 2025. 25

Amanda Wang, Ryan Kingsbury, Matthew McDermott, et al. A Framework for Quantifying Uncertainty in DFT Energy Corrections. *Sci. Rep.*, 2021. 4, 25

Haorui Wang, Marta Skreta, Cher-Tian Ser, et al. Efficient evolutionary search over chemical space with large language models. 2024. URL https://arxiv.org/abs/2406.16976. 2

Haorui Wang, Marta Skreta, Cher-Tian Ser, et al. Efficient Evolutionary Search over Chemical Space with Large Language Models. In *ICLR*, 2025. 10

Mingjian Wen, Matthew K. Horton, Jason M. Munro, et al. An Equivariant Graph Neural Network for the Elasticity Tensors of All Seven Crystal Systems. *Digit. Discov.*, 2024. 16

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, et al. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. In *ICLR*, 2022. 1, 5, 6, 10, 18, 20

Keqiang Yan, Xiner Li, Hongyi Ling, Kenna Ashen, Carl Edwards, Raymundo Arroyave, Marinka Zitnik, Heng Ji, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Invariant tokenization of crystalline materials for language model enabled generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=18FGRNd0wZ. 10

Han Yang, Chenxi Hu, Yichi Zhou, et al. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures. *arXiv.org*, 2024a. 4, 24

Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander L Gaunt, Brendan McMorrow, Danilo Jimenez Rezende, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Generative hierarchical materials search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=PsPR4NOiRC. 10

Bangchen Yin, Jiaao Wang, Weitao Du, et al. AlphaNet: Scaling Up Local Frame-Based Atomistic Foundation Model. *arXiv.org*, 2025. 24

Claudio Zeni, Robert Pinsler, Daniel Z"ugner, et al. A Generative Model for Inorganic Materials Design. *Nature*, 2025. 1, 8, 10, 18, 20, 25

Duo Zhang, Hangrui Bi, Fu-Zhi Dai, et al. Pretraining of Attention-Based Deep Learning Potential Model for Molecular Simulation. *npj Comput. Math.*, 2024. 4, 24

Linfeng Zhang, Han Wang, Roberto Car, et al. Phase Diagram of a Deep Potential Water Model. *Phys. Rev. Lett.*, 2021. 24

Ruiming Zhu, Wei Nong, Shuya Yamazaki, et al. WyCryst: Wyckoff Inorganic Crystal Generator Framework. *Matter*, 2024. 25

Supplementary Material for MatLLMSearch

A Experimental Details of Population Initialization

The retrieval set $\mathcal{R}$ of stable structures are sampled from known stable structures–Matbench-bandgap dataset (Dunn et al., 2020), which consists of 106,113 crystal structures in total. To initialize the parent structures for the first iteration, we calculated the decomposition energy for each structure with CHGNet. For CSG task, we removed binary compounds and structures with high-order compositions, i.e., retaining candidate structures with 3 to 6 elements. For CSP tasks, we filtered the seed structures to find those matching a desired compositional pattern. In addition, we applied de-duplication by composition to the candidate structures. The analysis of how sampling rule of the extra pool affect the performance is provided in Section 4.4. To further enhance the structure generation, we envision future work that could explore how structures can be ensembled to form a larger candidate pool for parent selection.

B Reproducibility

The crystal structures generated by MatLLMSearch can be downloaded here.

The implementation of our evolutionary search pipeline is available here.

C Prompt for CSG

You are an expert material scientist. Your task is to propose hypotheses for {reproduction_size} new materials with valid stable structures and compositions. No isolated or overlapped atoms are allowed.

The proposed new materials can be a modification or combination of the base materials given below.

Format requirements:

1. *Each proposed structure must be formatted in JSON with the following structure:*

```
{{
    "i":  {{
        "formula":  "composition_formula",
        "POSCAR":  "POSCAR_format_string"
    }}
}}
```

2. *Use proper JSON escaping for newlines (\n) and other special characters*

Base material structure for reference:
{reference_structures}

Your task:
1. *Generate {reproduction_size} new structure hypotheses*
2. *Each structure should be stable and physically reasonable*
3. *Format each structure exactly as shown in the input*

Output your hypotheses below:

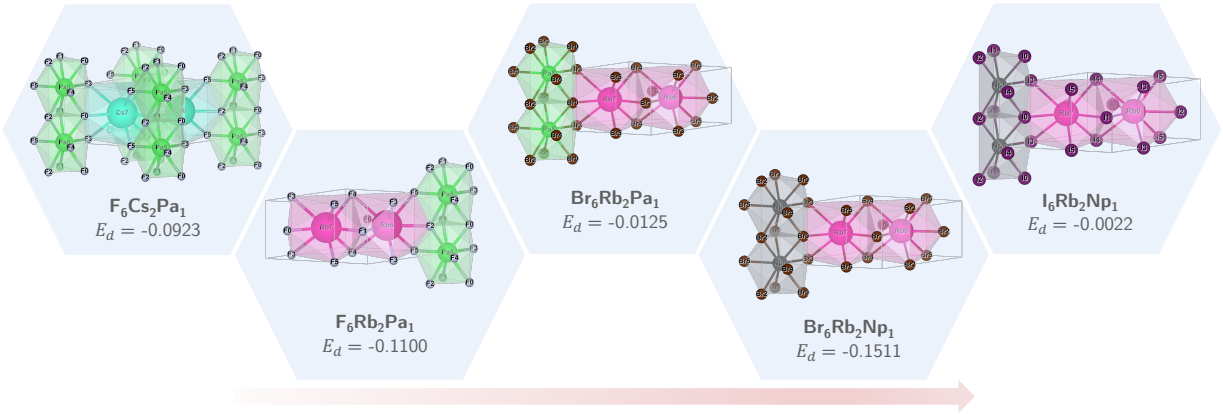


Figure S1: Example trajectory of CSG.

D Additional Experiments of Crystal Structure Generation and Design

Model	f-ele in Parents	Validity		Metastability		
		Structural	Composition	M3GNet	CHGNet	
				$E_\mathrm{d} < 0.1$	$E_\mathrm{d} < 0.1$	$E_\mathrm{d} < 0.03$
CDVAE	—	100.0%	86.7%	28.8%	—	—
CrystalTextLLM-7B	—	96.4%	93.3%	35.0%	—	—
CrystalTextLLM-13B	—	95.5%	92.4%	38.0%	—	—
CrystalTextLLM-70B	—	99.6%	95.4%	49.8%	—	—
MatLLMSearch (Llama 3.1-70B)	Stability	100.0%	79.4%	81.1%	76.8%	56.5%
	Bulk Modulus	100.0%	82.9%	27.0%	43.3%	8.3%
	Multi-turn	100.0%	84.1%	70.9%	57.1%	29.8%
	Weighted Sum	100.0%	88.1%	74.0%	59.8%	36.5%
	Lexical	100.0%	89.5%	84.7%	78.0%	59.9%

Table S1: Compare experimental results under various optimization goals. We explored multi-objective optimization for stability and bulk modulus in two different ways.

The flexibility of our evolutionary pipeline is demonstrated by its ability to guide LLMs in proposing novel crystal structures with diverse mechanical characteristics. We evaluate model performance under five distinct optimization strategies: (1) stability-oriented optimization ("Stability"), (2) property-oriented optimization ("Bulk Modulus"), (3) alternating multi-objective optimization ("Multi-turn"), (4) normalized weighted-sum optimization ("Weighted Sum"), and (5) lexicographic optimization ("Lexical"). As shown in Table S1, all optimization strategies maintain high metastability rates for the proposed structures, demonstrating that our algorithm can optimize specific properties while maintaining high validity and stability.

Stability optimization. We visualize an example stability-oriented optimization trajectory in Figure S1.

Bulk modulus optimization. To validate the capability of MatLLMSearch for property-guided generation, we conduct single-property optimization by modifying the selection criteria from decomposition energy (E_d) to bulk modulus. In crystalline solids, bulk modulus serves as a key indicator for designing materials with enhanced mechanical hardness. Our experiments used bulk modulus values derived from the Birch-Murnaghan equation of state as a proof of concept. For more comprehensive materials design applications, this approach can be extended to include elastic tensors from DFT calculations predictions using equivariant graph neural networks (Wen et al., 2024).

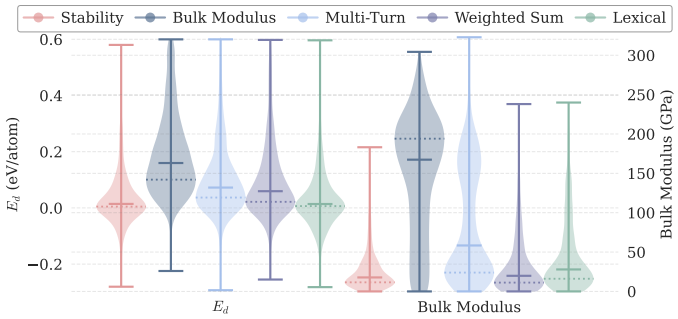


Figure S2: Comparison of optimization strategies targeting different objectives evaluated based on thermodynamic stability (decomposition energy E_d) and mechanical property (bulk modulus).

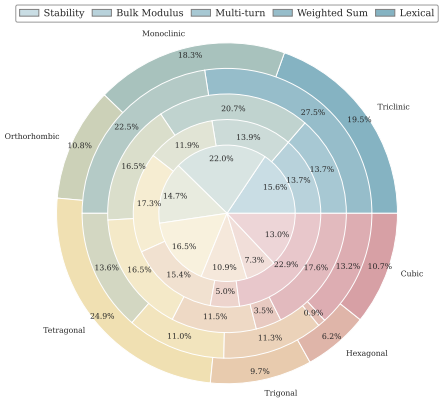


Figure S3: Crystal systems distribution under varied objectives.

Figure S2 presents the distribution comparison of decomposition energy (E_d) and bulk modulus for structures generated under varied optimization strategies, revealing distinct performance trade-offs. The bulk modulus optimization generated more structures with larger bulk modulus values, reaching a peak density at 194 GPa compared to only 19 GPa in stability-oriented optimization. However, this enhancement comes at the cost of increased decomposition energy, with the E_d density peaks shifting from 0.0 eV/atom in stability-oriented optimization to 0.1 eV/atom in bulk modulus optimization, indicating reduced thermodynamically stability across iterations.

Multi-objective optimization. Beyond single-objective optimization, we explored multi-objective optimization approaches to simultaneously target both thermodynamic stability and mechanical properties using two different multi-objective optimization strategies.

First we implement an alternating optimization strategy ("Multi-turn"), where the algorithm alternates between optimizing stability and property in successive iterations. Stability is optimized in the first iteration to set a foundation for property optimization. For customized multi-objective optimization, the number of iterations for each optimization goal can be adjusted. As shown in Figure S2, this method achieves balanced performance in optimizing stability and bulk modulus, with E_d centered around 0.037 eV/atom. We observe that bulk modulus distribution separates structures into groups with high mechanical strength at moderate stability versus high stability with lower mechanical strength, suggesting the inherent trade-off in crystal structure generation.

Then we consider a normalized weighted-sum approach that combines both objectives into a single scalar function. We apply min-max normalization to both E_d and bulk modulus values, then compute the objective as $\mathcal{J} = w_e \cdot \hat{E}_\mathrm{d} + w_b \cdot (1 - \hat{B})$, where $\hat{E}_\mathrm{d}$ and $\hat{B}$ are the normalized values, and $w_e = 0.7$, $w_b = 0.3$ are the weights. This strategy produces crystal structures with bulk modulus centered around 141 GPa and E_d centered at 0.034 eV/atom.

Then, the "Lexical" method prioritizes stability as the primary criterion, only considering bulk modulus for metastable structures ($E_\mathrm{d} < 0.03$ eV/atom). For stable structures, the weighted sum of the two objectives are then being optimized. Low stability structures are penalized to ensure that it remains the dominant factor. While single-objective stability optimization achieves the highest metastability rate of 76.8%, all multi-objective approaches maintain metastability rates above 50% while enhancing mechanical properties.

In addition, the analysis of crystal system distributions in Figure S3 indicates that our framework preserves structural diversity regardless of the optimization objective.

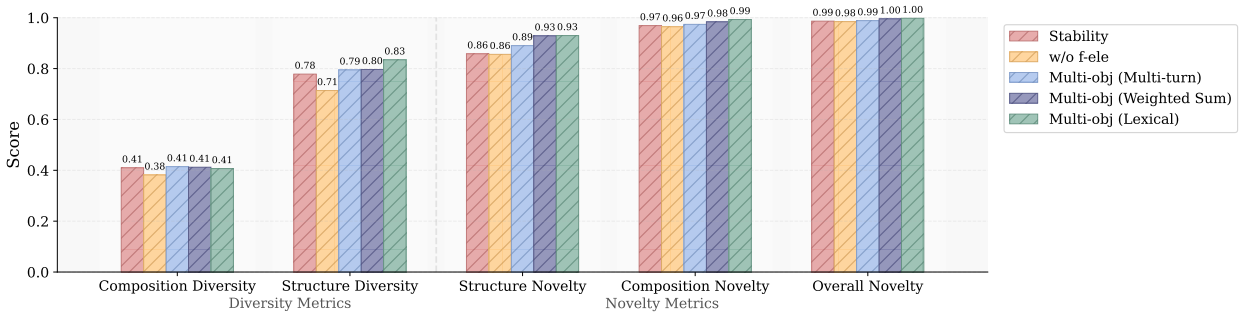


Figure S4: Diversity and novelty evaluation results for structures proposed under different experimental settings.

E Additional Evaluation on Diversity and Novelty of Generated Structures

We quantitatively evaluate the diversity and novelty of structures generated by our framework across configurations using established metrics from prior work (Gruver et al., 2024; Xie et al., 2022). Crystal diversity is measured by computing pairwise distances between their structural and compositional fingerprints. Additionally, we apply log normalization to composition diversity for 0-1 scale standardization. The novelty measures the distance between generated samples and their closest neighbors in the extra pool of reference structures. The structural distance cutoff and composition distance cutoff used for novelty calculation are 0.1 and 2 respectively. To align with previous work, all metrics are computed on structures predicted to be metastable.

The results are summarized in Figure S4. Across different optimization goals, we observe an interesting trade-off between property-specific optimization and novelty, balancing targeted enhancement against chemical space exploration. When optimizing beyond stability alone, such as targeting bulk modulus or performing multi-objective crystal structure design, we observe decreased novelty while diversity remains consistently high across all optimization goals.

While LLMs do tend to favor known stable configurations, our framework's evolutionary approach encourages the exploration of diverse structural motifs, as evidenced by the relatively uniform crystal system distributions.

On S.U.N. metric fairness. The S.U.N. (Stable, Unique, Novel) metric (Zeni et al., 2025) is convenient but limited: it collapses multi-dimensional quality into counts and depends on the chosen reference/training sets. Our MatLLMSearch is training-free; therefore, we compute S.U.N. against the MatBench structures used as parents, following DiffCSP/FlowMM protocols. Direct cross-paper comparison is inherently challenging. We therefore position S.U.N. as a supplementary signal alongside multi-MLIP metastability and DFT verification, and we report additional diversity/novelty analyses to capture broader scientific value.

To further assess structural diversity, we extended our evaluation with additional metrics, including the S.U.N. Rate (Stable, Unique, Novel) aligning to MatterGen and diversity and novelty regarding composition and structure in Table 2. High diversity and novelty are consistently achieved by our approach under various settings.

We provide comprehensive evaluation across all S.U.N. dimensions: For Stability, we report metastability rates computed with multiple MLIPs (CHGNet and M3GNet for CSG and additional Orb-v3 for CSP) and DFT-calculated stability rates, providing more rigorous thermodynamic assessment than S.U.N.'s single stability threshold. For Uniqueness, we provide detailed analysis including space group distributions and crystal system diversity, offering deeper insights into the true diversity of generated structures. For Novelty, we measure compositional and structural novelties and demonstrate elemental co-occurrence pattern shifts that indicate genuine exploration of novel composition space regions.

Methodological note. In Table 2, the "Stable" criterion follows FlowMM: we first relax with CHGNet and then verify stability with DFT at the 0.0 eV/atom threshold on the convex hull. This aligns the protocol for

fair comparison while acknowledging that our training-free setup necessitates computing uniqueness/novelty against MatBench parents.

F Time Overhead.

The efficiency of our method is primarily determined by the hardware resources and model size used. As reported in Table S2, with Llama-3.1-70B-Instruct running on 4 A6000 GPUs and a population size of 100, the average time to propose one valid unique structure is 62.35 seconds. In comparison, Crystal-TextLLM takes 51.6 seconds on average to propose one valid structure. Our time overhead can be further optimized with improved computational resources.

Model	Avg. Generation Time (s)	Avg. Evaluation Time (s)
70B	55.99	6.36
8B	53.40	10.80

Table S2: Time overhead for different model sizes

Inference cost rationale. We report average time per successfully generated valid structure, rather than raw FLOPs, because larger models achieve higher success rates per attempt. Although a 70B model has higher per-step compute than an 8B model, its greater validity rate often yields lower cost per successful sample.

G Additional Discussion on Ablation Analysis.

G.1 Base LLM Analysis

Model Scale Effects. We evaluate the framework across different LLM configurations, including pre-trained and fine-tuned Llama 3.1 models (8B and 70B). Table S3 demonstrates that model capability significantly impacts CSG metastability. The 70B model achieves 76.8% metastability in our full framework compared to 27.7% for the 8B model, indicating that crystallographic knowledge emerges at sufficient model scale.

Base Model	Fine-Tuning?	Prompting Strategy	Metastable(%) (CHGNet)	
			$E_d < 0.1$ eV/atom	$E_d < 0.03$ eV/atom
Llama 3.1 8B	✗	Zero-shot	0.0	0.0
		MatLLMSearch	27.7	10.0
	✓(8 bit)	Zero-shot	0.0	0.0
		MatLLMSearch	45.5	22.7
Llama 3.1 70B	✗	Zero-shot	25.8	12.9
		MatLLMSearch	76.8	56.5
	✓(4 bit)	Zero-shot	13.9	2.8
		MatLLMSearch	66.0	48.0

Table S3: Meta-stability comparison of prompting strategy across models with and without fine-tuning.

Fine-tuning Integration. Fine-tuned models show substantial improvements when integrated with our evolutionary framework. The 8B fine-tuned model achieves 45.5% metastability (from 27.7% baseline), while the 4-bit quantized 70B model maintains 66.0% metastability despite compression constraints. See Table S3 for detailed results and prompting strategies. Importantly, our information value metric demonstrates that both fine-tuned and pre-trained models integrate seamlessly into the evolutionary framework, with performance scaling according to base model capability.

G.2 Generation Strategy Comparison.

We extend the comparison to two generation strategies: zero-shot (w/o reference structures nor evolution) and our evolutionary framework. Zero-shot approaches fail for 8B models and achieve only 25.8% metastability for 70B models. Our evolutionary framework overcomes these limitations by systematically exploring chemical space while maintaining high stability rates.

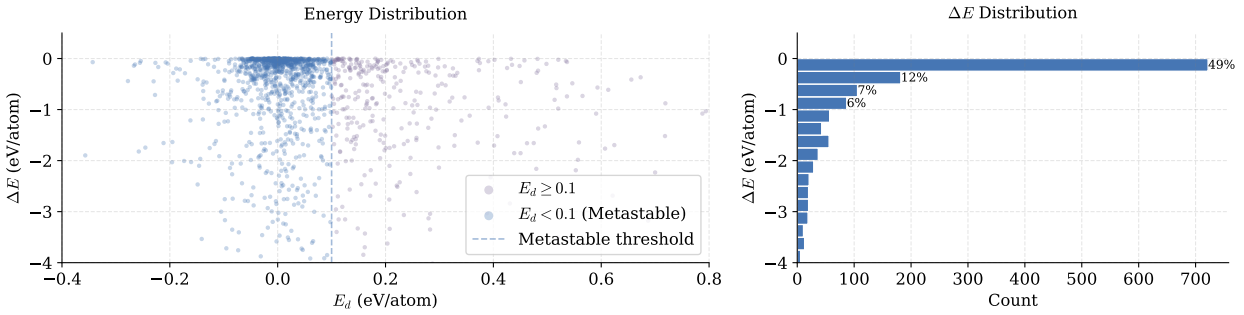


Figure S5: Distribution of energy change ΔE before/after structural relaxation and decomposition energy (E_{d}) for structures proposed by LLM, evaluated using the pretrained CHGNet.

Method	Primary Format	Generative	Model	Training
CDVAE (Xie et al., 2022)	3D	Diffusion	GNN	Training
MatterGen (Zeni et al., 2025)	3D	Diffusion	GNN	Training
Flam-Shepherd and Aspuru-Guzik (2023)	3D	AR	Transformer	Training
DiffCSP (Jiao et al., 2024)	3D	Diffusion	GNN	Training
CrystalTextLLM (Gruver et al., 2024)	Text/CIF	LLM	Transformer	Fine-tuning
FlowMM (Sriram et al., 2024)	3D	Flow	GNN	Training
MatLLMSearch (Ours)	Text/CIF/POSCAR	LLM	Llama 3.1	N/A

Table S4: A collection of generative models on computational materials discovery. Training denotes if training/fine-tuning is required on materials databases.

The comprehensive analysis reveals several critical insights: (1) Reference structures simultaneously accelerate convergence and promote structural diversity exploration across multiple dimensions, (2) Evolutionary iterations are essential for practical generation volumes and sustained quality optimization, (3) The synergistic combination of both components achieves optimal performance that exceeds individual contributions, (4) Our information value metric successfully captures complex trade-offs that single-dimensional metrics miss, enabling objective comparison across methodologies. These findings establish a robust evaluation framework for crystal structure generation while demonstrating the effectiveness of our evolutionary approach.

H Impact of Structure Relaxation

To measure the contribution of structural relaxation in our framework, we introduce a quantity ΔE to represent the energy difference after and before structural relaxation using CHGNet. Figure S5 reveals that the majority of the proposed structures proposed by LLMs exhibit a relatively small ΔE, with 61.2% showing minimal energy changes ($|\Delta E| < 0.5$ eV/atom) during relaxation. This distribution indicates that our framework generates physically meaningful structures that are already close to their local energy minima, requiring only modest refinements through relaxation.

I Impact of Structural Representation

A number of computational methods has emerged for crystal structure generation using machine learning approaches, as shown in Table S4. Most methods represent crystal structures using 3D information processed through either Graph Neural Networks (GNN) or Transformer architectures, employing various generative strategies like diffusion models or autoregressive approaches. More recently, text-based formats and Large Language Models (LLMs) have emerged as an alternative approach, signaling a promising shift in crystal structure generation and analysis techniques.

Format	# Unique / # Total generated	$E_d < 0.1$ eV/atom	$E_d < 0.03$ eV/atom
POSCAR (4)	76.7%	75.4%	55.3%
POSCAR (12)	72.3%	76.8%	56.5%
CIF	75.1%	68.9%	49.5%

Table S5: Proportion of unique structures and their CHGNet-predicted metastability using different structure formats.

The encoding of crystallographic structures into text-based format is essential for LLM processing. We investigated the impact of different structural representation strategies on generation efficiency and performance: CIF format and POSCAR format with either 4 or 12 decimal places of precision. See Figure S6 for examples.

First, we examine the token efficiency by analyzing the MatBench dataset for token length distribution as shown in Figure S7. The distribution indicates that the POSCAR format with 4 decimal places offers the most token-efficient representation while maintaining reasonable precision, followed by the POSCAR with 12 digits and CIF format. CIF format requires more tokens than POSCAR format, given that CIF uses a more verbose structure and additional metadata.

Performance evaluation shown in Table S5 suggests that POSCAR formatting in 12 decimal places demonstrates slightly better overall performance in the rate of (meta)stability of generated structures under different criteria ($E_d < 0.03$ or 0.1 eV/atom). Therefore, we employ POSCAR of 12 decimal places as a trade-off results of token efficiency and informativeness. The marginal difference across format may be attributed to the crystallographic data exposed to the LLMs during pre-training. However, it is noteworthy that performance differences across formats remain modest, suggesting the resilience of our approach across different structural representations.

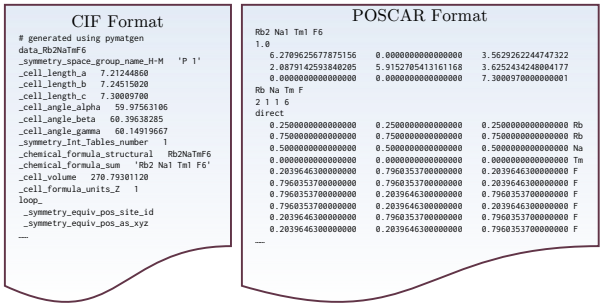


Figure S6: Structure string examples of CIF format and POSCAR format.

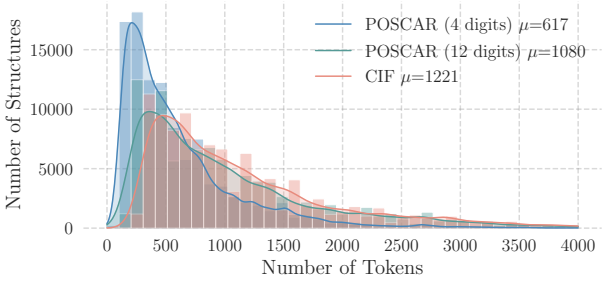


Figure S7: Token efficiency comparison under CIF formatting and POSCAR formatting for the precision of 4 and 12 decimal. μ indicate the mean of token lengths.

J Hyper-Parameter Studies

Reproduction parameters. Our training-free evolutionary framework significantly reduces hyperparameter sensitivity compared to traditional machine learning methods. The reproduction phase introduces several key hyper-parameters that influence LLMs' generation behavior and efficiency, including population size (K), context size (C), and children size (c). Our baseline configuration ($C = 2$, $c = 5$) leverages the Llama 3.1 (70B) model to achieve balanced performance, generating 72.29% unique structures while maintaining high stability rates.

Analysis of parent-to-children ratios reveals that increasing parent diversity ($C = 5$, $c = 2$) can enhance composition uniqueness of generated structures to 95.49%, though at the price of slight decrease in stability, as presented in Table S6. Conversely, results with single parent demonstrates that crossover between multiple parent structures is beneficial for maintaining structural diversity and stability in the generation process.

Reproduction Configuration	# Unique / # Total generated	$E_d < 0.1$eV/atom	$E_d < 0.03$ eV/atom
$1 \rightarrow 5$	56.5%	79.8%	56.4%
$2 \rightarrow 5$	72.3%	76.8%	56.5%
$2 \rightarrow 2$	86.3%	74.8%	54.3%
$5 \rightarrow 5$	92.7%	72.3%	47.3%
$5 \rightarrow 2$	95.5%	68.3%	46.1%

Table S6: Proportion of unique structures and their CHGNet-predicted metastability under varying reproduction configurations.

LLM Temperature	# Unique / # Total generated	$E_d < 0.1$ eV/atom	$E_d < 0.03$ eV/atom
0.95	72.3%	76.8%	56.5%
0.7	70.7%	75.4%	56.6%
0.5	70.7%	71.2%	51.4%
0.2	69.8%	70.3%	50.2%

Table S7: Proportion of unique structures and their CHGNet-predicted metastability with different LLM temperatures.

Overall, we believe that higher parent-to-children ratios can lead to better overall quality in generated structures.

Our analysis also reveals that larger population sizes K can maintain high stability and validity rates comparable to smaller populations. One potential benefit of increasing population size is the diversity introduced in the iteration process, which can alleviate the overpopulation of f-ele structures but also lead to higher compositional diversity. However, the increased diversity is offset by higher rates of structural duplication across iterations, suggesting earlier convergence may be needed. Our findings above enable application-specific optimization of the framework's parameters.

Model temperature. The temperature hyper-parameter controls sampling randomness in language models by scaling the logits before softmax transformation. Higher temperatures flatten the probability distribution, increasing sampling diversity, while lower temperatures concentrate probability mass on the most likely tokens. While temperature is commonly associated with model creativity, with higher temperatures generally producing slightly more novel outputs (Peeperkorn et al., 2024), this relationship remains an active area of research.

Crystal structure generation is a creative task that requires exploring diverse structural possibilities while maintaining physical validity. We employed an LLM inference temperature of 0.95 in our baseline experiments to facilitate broader structural exploration while maintaining reasonable generation stability. In Table S7, we present the metastability evaluated by CHGNet for structures generated with different LLM temperatures. At the temperature of 0.95, the LLM generated 76.81% metastable structures with $E_d < 0.1$ eV/atom as evaluated by CHGNet. Reducing the temperature to 0.7 maintained robust performance, producing 75.38% metastable structures. Further lowering the temperature to 0.5 yields 71.18% metastable structures. If we choose $E_d < 0.03$ eV/atom as the stability criterion, the percentage of qualifying structures at temperatures 0.95, 0.7, 0.5 and 0.2 are 56.5%, 56.6%, 51.4% and 50.2% respectively. The consistent high stability rates across temperature settings demonstrate the robustness of our pipeline to LLM hyper-parameter variations.

K Crystal Structure Prediction Examples.

We further evaluate MatLLMSearch on crystal structure prediction across multiple compositions. Crystal structure prediction fundamentally involves searching for the optimal atomic arrangement that minimizes the system's energy for a given composition.

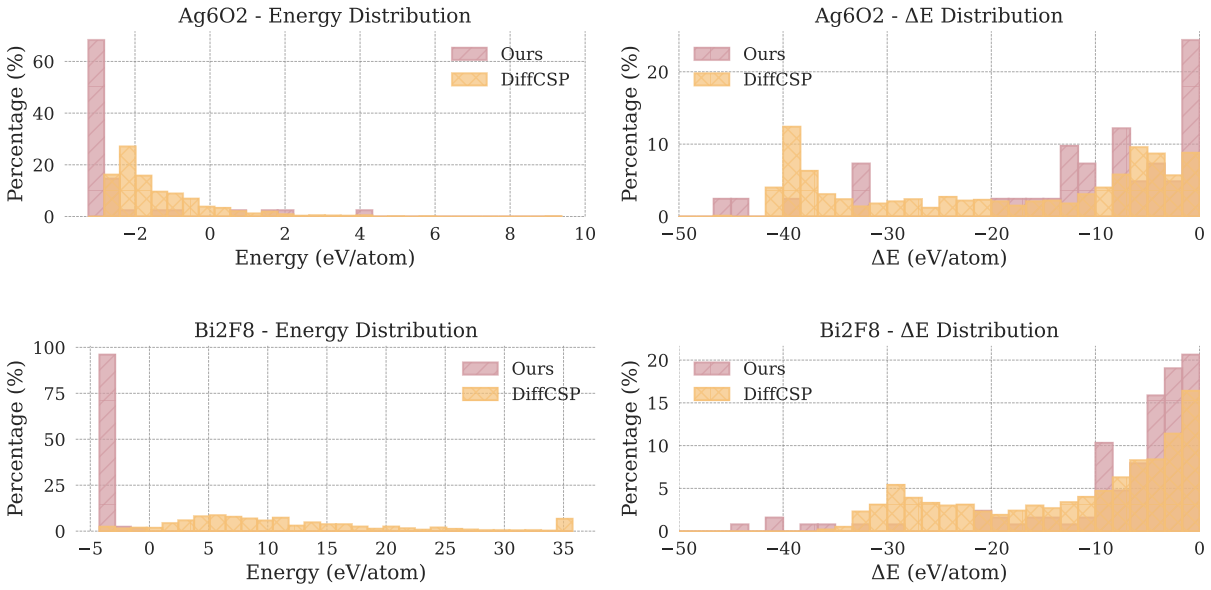


Figure S8: Energy and ΔE distribution of CSP results for Ag_6O_2 and Bi_2F_8.

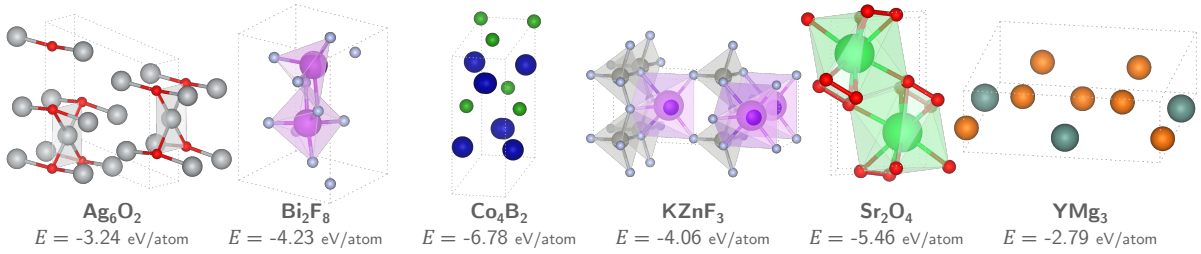


Figure S9: CSP examples.

For CSP task, we first apply compositional constraints to filter seed structures, and then execute MatLLMSearch for 10 iterations. This process enables the LLM to predict crystal structures by referencing from optimized structures with similar compositions.

To benchmark the effectiveness of our framework, we compare our results against structures generated and optimized by DiffCSP. We first used DiffCSP to sample 100 candidate structures for each composition. We then trained an energy predictor model on the MP-20 dataset for 1000 epochs to learn formation energy predictions. Then we applied the energy-guided optimization procedure to refine the structures with the trained energy model, generating 10 optimized variants for each initial structure. Orb-v3 (Rhodes et al., 2025) are employed to evaluate the energy of the relaxed structure from both sources.

MatLLMSearch successfully predicts structures for various compositions, including Ag_6O_2, Bi_2F_8, Co_4B_2, $KZnF_3$, Sr_2O_4, and YMg_3. Figure S9 presents representative examples of these predicted crystal structures, which achieve lower energies than the best structures predicted by DiffCSP.

To further evaluate the two approaches, we analyzed the energy and ΔE distributions of the generated structures. As shown in Figure S8, a large portion of structures predicted by DiffCSP experience substantial structural changes during relaxation, as indicated by large $|\Delta E|$ values. In contrast, our LLM-generated structures demonstrate superior initial stability, requiring minimal relaxation. This suggests that our initial configurations are already close to local energy minima.

L Algorithm Details

Algorithm 1 The MatLLMSearch Framework

Require: Population size K, parent size P, reproduction size C, number of iterations N, known stable structures $\mathcal{D}$, oracle function O, extra pool $\mathcal{R}$

```
▷ Initialization
Form population P_0 by sampling K groups of P structures from D
Initialize structure collection S ← ∅
for i ← 0, 1, ···, (N − 1) do
    ▷ LLM-guided reproduction
    Generate prompts from parent structures in P_i
    Obtain offspring structures C_i via LLM inference and parsing
    ▷ Structure evaluation
    Relax structures C_i ← CHGNetRelax(C_i)
    Calculate decomposition energy E_d and properties
    Evaluate objective scores using oracle function O
    Update structure collection S ← S ∪ C_i
    ▷ Selection
    Form candidate pool from parents P_i, offspring C_i, and extra pool R
    Select top-(K × P) structures based on objective scores from the candidate pool
    Construct next parent groups P_{i+1}
Validate final structures via DFT
return cumulated structures S
```

M Details of Machine Learning Interatomic Potentials

A significant breakthrough in addressing computational cost challenges has emerged through the development of machine learning interatomic potentials (MLIPs) trained based on high-fidelity quantum mechanical calculations (e.g., DFT) (Batzner et al., 2022; Cheng, 2024a; Du et al., 2023a;b; Liao et al., 2024; L'opez-Zorrilla et al., 2023; Yin et al., 2025; Zhang et al., 2021). In MLIPs, the total energy is expressed as a sum of atomic contributions, where each atom's energy depends on its local environment including the atomic coordinates and chemical species of neighboring atoms within a cutoff radius:

$$\hat{E} = \sum_{i}^{n} \phi(\{\vec{r}_j\}_i, \{C_j\}_i), \quad \hat{\boldsymbol{f}}_i = -\frac{\partial \hat{E}}{\partial \boldsymbol{r}_i}, \quad \boldsymbol{\sigma} = \frac{1}{V}\frac{\partial \hat{E}}{\partial \boldsymbol{\varepsilon}}. \tag{S1}$$

Here, ϕ is a learnable function that maps the set of position vectors $\{\vec{r}_j\}_i$ and chemical species $\{C_j\}_i$ of the neighboring atoms j to the energy contribution of atom i. The forces $\boldsymbol{f}_i$ and stress $\boldsymbol{\sigma}$ are calculated via auto-differentiation of the total energy with respect to the atomic Cartesian coordinates and strain. Recent advances have demonstrated that MLIPs, trained on extensive density functional theory (DFT) calculations accumulated over the past decade across diverse materials systems, exhibit remarkable transferability in performing atomistic simulations across various material and chemical systems. These broadly applicable potentials are known as universal MLIPs (uMLIPs) (Batatia et al., 2023; Chen and Ong, 2022; Deng et al., 2023; Park et al., 2024). By leveraging uMLIPs as surrogate energy models, researchers can rapidly optimize crystal structures and obtain structure-energy relationships for assessing thermodynamic stability. By leveraging uMLIPs as surrogate energy models, one can rapidly optimize crystal structure and obtain the structure-energy relationships for assessing thermodynamic stability. Recent benchmark studies, including MACE (Batatia et al., 2023), DPA-1 (Zhang et al., 2024) and JMP (joint multi-domain pretraining) (Shoghi et al., 2024), have demonstrated the high accuracy of these uMLIPs in predicting crystal thermodynamical stability, particularly for industrial-scale implementations trained on millions of compounds and non-equilibrium atomic configurations (Barroso-Luque et al., 2024; Merchant et al., 2023; Yang et al., 2024a).

To accelerate the oracle function evaluation in the evolutionary iterations, we performed all structure relaxations with the `FIRE` optimizer (Bitzek et al., 2006) over the potential energy surface provided by CHGNet, where the atom positions, cell shape, and cell volume were optimized to reach converged interatomic

forces of 0.1 eV/atom (Deng et al., 2023). The output energy prediction is directly compatible with the Materials Project phase diagrams with the `MaterialsProject2020Compatibility` (Wang et al., 2021).

Limitations. Our study serves as a proof-of-concept and requires further validation in real-world materials discovery workflows. While we have demonstrated that LLM inference is a powerful tool for searching materials under thermodynamic stability guidance, the practical realization of new materials remains challenging, particularly in terms of successful synthesis.

One limitation observed in our CSP tasks is that the generated structures exhibit similarities to the provided reference structures. The evolutionary nature of the genetic algorithm naturally favors incremental modifications over radical structural changes. Additionally, LLMs exhibit an inductive bias toward known stable structures, often resorting to their pre-trained knowledge and simple atomic substitutions. The challenge of AI for material sciences discovery are further discussed in Szymanski and Bartel (2025). Nevertheless, our approach can serve as an effective optimization tool in addition to the suggestion of novel structural prototypes, which can be more readily obtained through alternative methods, including variational autoencoders (Ren et al., 2022; Zhu et al., 2024), diffusion models (Zeni et al., 2025), random structure searching (Pickard and Needs, 2011), or response-matching approaches (Cheng, 2024b). However, the capability of these methods for comprehensive materials discovery across diverse chemical spaces remains under-explored. In addition, it is an open question that whether the LLM-proposed materials design hypotheses are free of intellectual property issues.

Broader Impact.

This work aims to advance machine learning and computational materials discovery by making crystal structure generation more accessible and efficient. This advancement will particularly benefit researchers who have limited computational resources, enabling scientific discovery without the need to train large ML models. Additionally, the oracle functions can be further adapted to incorporate experimental data or high-fidelity property predictors for the generated crystal structures when applying this pipeline to practical materials discovery.