TEGO: SYMMETRY-PRESERVING LOCAL EDITING FOR TARGETED INORGANIC MATERIALS DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

The inverse design of inorganic crystalline materials is generally regarded as an end-to-end, property-to-structure generation task, giving rise to a series of *ab initio* methods such as diffusion models. However, such methods overlook an inherent contradiction at the information level: properties can be viewed as a lossy compressed representation of structures, and the property-to-structure mapping follows a one-to-many non-functional relation. Reconstructing complete structures merely from properties constraints essentially attempts to recover high-dimensional full structural information from low-dimensional lossy representations, which inherently suffers from the theoretical bound of information asymmetry and constrains the structural stability and property accuracy of outputs from ab initio methods. To address these challenges, we propose the **Tiny-Edit-Good-Outcome (Tego)** framework. Tego initiates inference from an existing stable structure, thereby performing localized inverse editing to approximate target material properties, circumventing the complexity of full structure reconstruction in *ab initio* methods. Evaluations demonstrate that Tego is capable of generating undiscovered, thermodynamically stable crystalline candidates, while achieving higher property achievement rate and computational efficiency across material inverse design scenarios.

1 INTRODUCTION

Materials underpin a wide range of technologies. Yet despite decades of exploration, only tens of thousands of thermodynamically stable materials have been identified Merchant et al. (2023). Advances in generative AI have opened a new avenue for high-throughput materials discovery, leading to 2 emerging paradigms: Crystal Structure Prediction (CSP) and Crystal Structure Design (CSD)Grunert et al. (2025); Cao & Wang (2026); Liu et al. (2025). CSP targets the prediction of new materials satisfying thermodynamic stability Xie et al. (2022); Jiao et al. (2023a), while CSD further considers target properties on top of this constraint FUJII et al. (2025); Han et al. (2025).

Generally, CSD is regarded as an end-to-end, property-to-structure generation task, giving rise to a series of **Property-to-Structure (P2S) *ab initio*** frameworks Zeni et al. (2025); Zhang et al. (2025). These methods typically take target properties as the only intended input, attempting to construct a direct mapping from property space to high-dimensional structure space.

Such methods overlook an inherent contradiction at the information level: material properties represent low-dimensional projections of intrinsically high-dimensional structures Ortega-Ochoa et al. (2025). Properties can be viewed as a lossy compressed representation of structures, and the P2S mapping follows a one-to-many, non-functional relation Helfrecht et al. (2020). Reconstructing complete structures merely from property constraints essentially attempts to recover high-dimensional full structural information from low-dimensional lossy representations, which inherently suffers from the theoretical bound of information asymmetry.

This lossy mapping poses a challenging task for P2S CSD methods, since P2S methods must simultaneously attend to both structural stability and property accuracy. Existing models are compelled to rely on large-scale datasets Antunes et al. (2024); Zeni et al. (2025) , typically requiring prolonged training to capture latent distribution of structures and properties. The generated structures are also plagued by limited structural validity and modest target property achievement rate.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

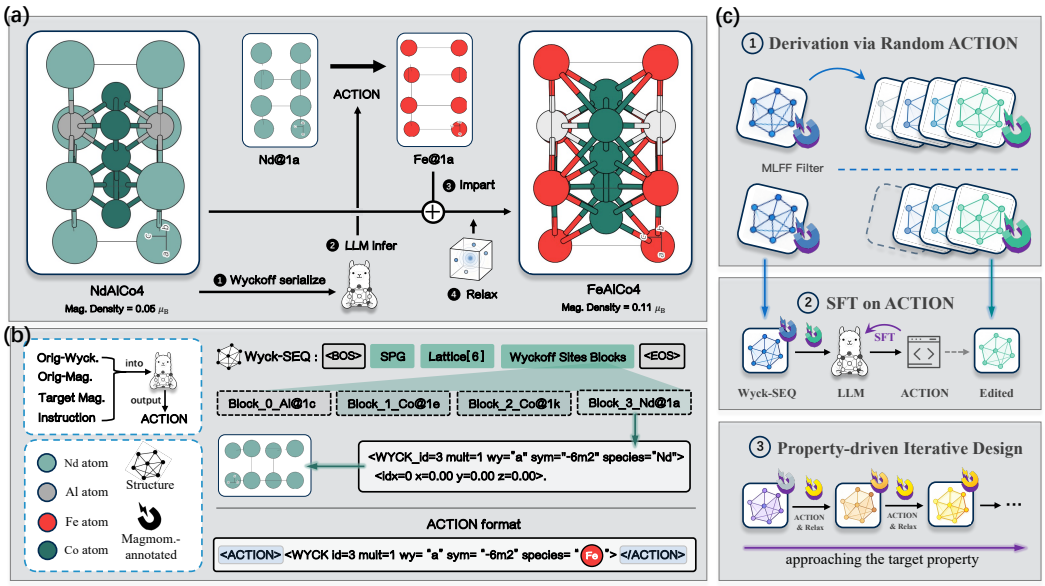

Figure 1: Schematic illustration of Tego. (a) Starting from an input crystal structure, the LLM infers and outputs edit ACTIONs serialized as Wyckoff-SEQ blocks. Each ACTION encodes Wyckoff identifiers, multiplicities, site symmetry, and atomic species to modify crystallographic configurations for target-property pursuit. Edited candidates go through relaxation and MLFF filtering to discard unphysical structures. (c) The model is trained via supervised fine-tuning (SFT) on ACTION-Wyck-SEQ pairs derived from random ACTION sampling. Through iterative editing, the magnetic density evolves from the original $0.05\,\mu_B/\mathring{A}^3$ toward the target value $0.11\,\mu_B/\mathring{A}^3$.

To mitigate these limitations, we propose the **Tiny-Edit-Good-Outcome (Tego)** framework shown in Figure 1. Tego is a **local editing method** and initiates inference from an existing stable structure, which explicitly learns property variations Δ_P and structural variations Δ_S. It transforms the direct P2S mapping in material design into a differentiable, locally iterative process. By fully exploiting priors of structural stability, Tego can place greater emphasis on property approximation. This improves the S.U.N. metric and diversity of generated samples, as well as the success rate for extreme property target. Our contributions are summarized as follows:

- To mitigate the inherent lossiness of the P2S mapping, we introduce Tego, a local editing strategy for targeted material design.
- To model the mutual coupling between Δ_P and Δ_S, we develop a training-inference pipeline that enables their correlated joint learning.
- To realize the above workflow on LLMs, we construct Wyck-SEQ structural representations together with edit-oriented ACTION sequences.

2 Related Work

2.1 Crystal structure prediction.

CSP aims to predict novel materials that have not been discovered. As an early approach, G-SchNet Gebauer et al. (2019a), and its periodic adaptation, P-G-SchNet, introduced autoregressive generation of 3D structures Gebauer et al. (2019b). Thereafter, CDVAE Xie et al. (2022) represented crystals in a variational latent space and reconstructed them through diffusion, providing one of the first general frameworks for de novo periodic material generation. DiffCSPJiao et al. (2023b) moved from latent reconstruction to direct generation by jointly diffusing lattice parameters and fractional coordinates with periodic $E(3)$ equivariance. DiffCSP++ Jiao et al. (2024) further introduced explicit space-group constraints through lattice, making crystallographic symmetry part of the generation process rather than a post-hoc check.

As de novo generation matured, subsequent work focused on better generative distributions, more compact representations, and stronger physical priors. FlowMM Miller et al. (2024) replaced score-based diffusion with Riemannian flow matching. FlowLLM Sriram et al. (2024)further learned a metastable crystal distribution with a language model and used its samples as the base distribution for geometric refinement by flow matching. WyFormer Kazeev et al. (2025a) represented crystals using compact Wyckoff sequences and employed a permutation-invariant autoregressive Transformer for symmetry-conditioned generation. ADiT Joshi et al. (2025) instead mapped molecules and periodic materials into a shared latent space and used a standard diffusion Transformer, demonstrating that crystal generation can also benefit from scalable, relatively weakly biased architectures.

More recently, SymmBFN Ruple et al. (2025) explicitly incorporated asymmetric units, site symmetries, and atomic species into a symmetry-aware Bayesian flow network. PLaID++ Xu et al. (2026) introduced a Wyckoff-based textual representation, iterative DPO-based reinforcement-learning and preference alignment for stable and constrained generation, while Chemeleon2 Park & Walsh (2026) applied reinforcement learning to a latent diffusion model to optimize stability, novelty, diversity, and user-defined material objectives. Crystalite Veljković et al. (2026) further showed that a diffusion Transformer, equipped with chemically structured tokenization and periodic geometric attention, can achieve competitive de novo generation with substantially lower sampling cost.

A complementary direction reduces the difficulty of global generation by reusing structural information that is already known to be plausible. ShotgunCSP Liu et al. (2024) generates large candidate libraries either by element substitution from known structural templates or by symmetry-constrained Wyckoff generation. CHGGen Zhong et al. (2025) moves closer to partial generation: it preserves a predefined symmetrized host structure and uses diffusion inpainting to regenerate only selected atomic positions, followed by relaxation. Such approaches show that preserving a physically meaningful structural prior can substantially reduce the effective search space. However, they do not explicitly **model how a particular structural modification changes a target physical property.**

2.2 Property-guided crystal design.

CSD seeks structures that realize a desired physical property. FTCP Ren et al. (2022) provided an early general inverse-design framework by combining real-space crystal information with reciprocal-space features in an invertible representation and organizing the latent space using material properties. MatterGen Zeni et al. (2025) subsequently established a more general property-conditioned diffusion paradigm. Its model jointly generates atomic species, fractional coordinates, and lattice parameters and can be fine-tuned with property labels to steer generation toward targets including magnetic density ($\mu_B/\mathring{A}^3$), electronic bandgap (eV), and bulk modulus (GPa). UniGenX Zhang et al. (2025) combines autoregressive token prediction with diffusion-based coordinate generation, allowing chemical sequence and continuous structure to be generated within a unified model and supporting multi-property conditional material generation. Chemeleon2 Park & Walsh (2026) fine-tunes a latent diffusion model with group relative policy optimization, using surrogate property predictions to define rewards for targeted bandgap generation.

Most existing property-guided methods must determine composition and geometry while simultaneously satisfying structural validity, thermodynamic stability, and the requested property. Tego takes a complementary view. By retaining most of the initial crystallographic environment and modifying only selected structural degrees of freedom, Tego narrows the search to a physically plausible neighborhood and allows the model to devote more of its capacity to the target property itself. In this sense, Tego complements de novo generation: the latter is well suited to exploring what materials may exist, whereas Tego focuses on how a material can be changed toward a desired property.

3 Method

Tego reformulates crystal structure design as a structured editing problem. Given an initial crystal structure S_0 and a target property p^*, our goal is to iteratively transform S_0 into a new structure S' that satisfies both physical validity and desired property constraints:

$$S' = \arg\max_{S \in \mathcal{S}} \ \mathcal{L}(S; p^*), \tag{1}$$

where $\mathcal{L}$ measures property alignment.

We introduce a symmetry-aware representation and a structured action space. This enables the transformation:

$$(S_0, p^*) \longrightarrow a \longrightarrow S' = a(S_0), \tag{2}$$

where a denotes a discrete, symmetry-preserving editing action. This formulation reduces the complexity of the problem and improves both controllability and physical consistency.

3.1 WYCK-SEQ AND SYMMETRY-PRESERVING ACTIONS

A crystal is represented by its lattice parameters $\boldsymbol{\ell} = (a, b, c, \alpha, \beta, \gamma)$, a set of atomic species and fractional coordinates. Under a space group G, symmetry-equivalent atoms form orbits corresponding to occupied Wyckoff positions Wondratschek & Müller (2004); Kazeev et al. (2025b;a). We organize these orbits into Wyckoff blocks:

$$\begin{aligned} S &= \left(\boldsymbol{\ell}, \{(Z_i, \mathbf{r}_i)\}_{i=1}^{N}\right), \qquad \mathbf{r}_i \in [0,1)^3, \\ B_k &= \left(Z_k, m_k, w_k, s_k, \{\mathbf{r}_i\}_{i\in\mathcal{I}_k}\right), \qquad k = 1, \ldots, K, \\ i, j \in \mathcal{I}_k \implies &\exists\, g \in G : g(\mathbf{r}_i) \equiv \mathbf{r}_j \pmod{\mathbb{Z}^3}. \end{aligned} \tag{3}$$

Each block contains atoms of species Z_k that belong to the same symmetry orbit; $\mathcal{I}_k$ indexes these atoms, $m_k = |\mathcal{I}_k|$ is the multiplicity, w_k is the Wyckoff letter, and s_k is the site symmetry. For a fixed space-group setting, the Wyckoff letter determines the multiplicity and site symmetry. We nevertheless retain these attributes explicitly to expose the symmetry constraints to the model. A block identifies a particular occupied orbit, rather than a Wyckoff letter alone, since multiple occupied orbits may share the same letter. For this reason, each Wyckoff block is assigned a unique identifier to facilitate localization. For concrete examples of Wyck-SEQ, please see the appendix.

To enable the model to better capture symmetry priors of materials and support local-editing operations, we design **Wyck-SEQ**, a formatted sequence that embodies the symmetry properties of materials. Wyck-SEQ serializes the global crystallographic information followed by the occupied Wyckoff blocks:

$$\begin{aligned} \mathcal{T}(S) &= [\texttt{SPG}(G),\ \texttt{LATTICE}(\boldsymbol{\ell}),\ W_1, \ldots, W_K], \\ W_k &= [\texttt{WYCK_id} = k,\ m_k, w_k, s_k, Z_k,\ \{\mathbf{r}_i\}_{i\in\mathcal{I}_k}]. \end{aligned} \tag{4}$$

The SPG field records the space-group number and symbol, the LATTICE field records the six lattice parameters, and each W_k explicitly lists the coordinates of one symmetry-equivalent group. Although CIF can also encode crystallographic symmetry, Wyck-SEQ makes occupied orbits explicit as the units of model input and subsequent editing. Because lattice parameters, species, and atomic coordinates are retained, the structure can be reconstructed up to the numerical precision of serialization. Grouping atoms by orbit removes the need to address equivalent atoms individually during editing; a unique sequence additionally requires consistent ordering of blocks and coordinates.

We define each action as a species substitution applied to one complete Wyckoff block. Let $\mathcal{E}$ denote the allowed set of elements. The action space and its effect on atomic species are

$$\begin{aligned} \mathcal{A}(S) &= \{a_{k,z} \mid k \in \{1, \ldots, K\},\ z \in \mathcal{E} \setminus \{Z_k\}\}, \\ a_{k,z}(S) &= \left(\boldsymbol{\ell}, \{(\widetilde{Z}_i, \mathbf{r}_i)\}_{i=1}^{N}\right), \\ \widetilde{Z}_i &= \begin{cases} z, & i \in \mathcal{I}_k, \\ Z_i, & i \notin \mathcal{I}_k. \end{cases} \end{aligned} \tag{5}$$

Thus, an action replaces the species of all atoms within the selected Wyckoff orbit while largely preserving the symmetry of the entire crystal structure. The model therefore predicts an explicit modification rather than regenerating the full crystal representation. Subsequent relaxation is required to ensure energetic stability, which may alter coordinates, lattice parameters, and even the symmetry of the resulting crystal.

An action is serialized as a short structured output. For example, the following action assigns Fe to the block with identifier :

```
<ACTION>
<WYCK_id=1 mult=1 wy="b" sym="-62m" species="Fe">
</ACTION>
```

3.2 PROPERTY-GUIDED CONTRASTIVE EDITING FINE-TUNING

Our goal is to enable the LLM to learn the correlation between Δp and ΔS. However, obtaining suitable training sample pairs is non-trivial: each pair must consist of two crystals that share a certain degree of similarity. Directly comparing two entirely distinct crystal samples would impose substantial learning difficulties on the model. To address this issue, we design an approach to obtain reasonably a large number of valid similar-structure pairs by deriving and filtering an initial dataset.

Constructing the derived editing dataset. We build the training data before training the editing model. Starting from each initial crystal, we generate a cluster of related structures, assign property labels, and record the actions that produced them. We then turn these records into paired editing examples. This lets us obtain several training examples from one initial structure, each showing how a particular substitution affects a material property.

We first group the atoms in an initial structure S_i into symmetry-equivalent Wyckoff blocks and assign an identifier to each block. We then apply random actions to the initial structure. Each action replaces the species of one randomly selected full block, usually with an element of the same nominal valence. Repeating this procedure gives a set of candidate derivatives. We relax each candidate and retain those that pass the energy filter:

$$\begin{aligned} \widetilde{S}_{ij} &= a_{ij}(S_i), \qquad S_{ij} = \mathrm{Relax}(\widetilde{S}_{ij}), \\ \mathcal{J}_i &= \{j \mid \text{relaxation succeeds and } E_{\mathrm{filt}}(S_{ij}) \le \varepsilon_E\}, \\ \mathcal{C}_i &= \{S_i\} \cup \{S_{ij} : j \in \mathcal{J}_i\}. \end{aligned} \tag{6}$$

Here, a_{ij} is the j-th randomly sampled action, E_{filt} is the energy measure used for filtering, and ε_E is its threshold. The initial structure and its retained derivatives form the structure cluster $\mathcal{C}_i$. We keep the generating action for every retained derivative. Its block identifier always refers to the initial structure, even if relaxation changes the geometry or symmetry of the derivative. This separates the substitution we ask the model to predict from the geometric adjustments made during relaxation.

We convert these records into formatted text. The initial structure is serialized as `Wyckoff-SEQ`, using the recorded block identifiers, and each substitution is serialized in the `ACTION` format.

Property labels and dataset records. Once the clusters are built, we use machine-learning predictors to assign property labels to all structures, including the initial crystals. Let f_q be the predictor for property . For each initial crystal, we store its structure and property labels, together with the property labels and generating action of every retained candidate:

$$\begin{aligned} \widehat{\mathbf{p}}(S) &= \big(f_1(S), \ldots, f_Q(S)\big), \\ R_i &= \Big(S_i,\, \widehat{\mathbf{p}}(S_i),\, \big\{\big(\widehat{\mathbf{p}}(S_{ij}), a_{ij}\big)\big\}_{j\in\mathcal{J}_i}\Big), \\ \mathcal{D}_{\mathrm{cluster}} &= \{R_i\}_{i=1}^{N_{\mathrm{seed}}}. \end{aligned} \tag{7}$$

Here, Q is the number of labeled properties and N_{seed} is the number of initial crystals. Each record therefore connects one initial structure to several observed outcomes, with an explicit action for each outcome. The property labels describe the relaxed candidates. The property labels are included as numerical fields.

Paired training and data augmentation. For training, we take an initial structure and one retained member of its cluster. The input contains the initial structure, its current property value, and the candidate's property value as the target. The output is the action that produced that candidate before relaxation. Writing $\mathcal{F}_{\mathrm{in}}$ for input formatting and $\mathcal{F}_{\mathrm{act}}$ for action serialization, we obtain

$$\begin{aligned} x_{ij}^{(q)} &= \mathcal{F}_{\mathrm{in}}(\mathcal{T}(S_i),\, \widehat{p}_q(S_i),\, \widehat{p}_q(S_{ij})), \\ y_{ij} &= \mathcal{F}_{\mathrm{act}}(a_{ij}), \\ \mathcal{D}_q &= \Big\{\big(x_{ij}^{(q)}, y_{ij}\big) \;\Big|\; i = 1, \ldots, N_{\mathrm{seed}},\; j \in \mathcal{J}_i\Big\}. \end{aligned} \tag{8}$$

Pairing the same initial crystal with different cluster members gives different editing examples. In this way, one initial crystal contributes multiple target–action pairs, providing data augmentation.

We fine-tune the model to predict the formatted action text. For an output sequence $y = (y_1, \ldots, y_{|y|})$, the training objective is

$$\mathcal{L}_{\mathrm{SFT}}^{(q)}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_q}\left[\sum_{t=1}^{|y|} \log P_\theta(y_t \mid x, y_{<t})\right]. \tag{9}$$

We compute the loss only on the action tokens. Here, *contrastive* refers to the comparison between an initial crystal and one of its derivatives, rather than to a separate contrastive loss. The aim is to learn which edits are associated with the requested property change. During training, the target comes from a retained candidate; at inference, it is supplied by the user.

4 Experiments

4.1 S.U.N. and Validity

Although Tego is originally designed for targeted material design, it remains a competitive approach for novel-material prediction.

Since LLMs already possess the capability to generate well-formatted outputs and considerable material knowledge, secondary fine-tuning is not strictly required for local-editing tasks. We therefore demonstrate the unconditional material-editing capability of Tego under **zero post-training cost**.

We select 10,000 samples with the lowest hull energy from MP20 as starting points. The 10000-th sample yields $E_{\text{hull}} = 0.008$ eV/atom. We set the target magnetic-moment density to 0.01 $\mu_B/\text{Å}^3$ to simulate undirected generation. This low value is nearly consistent with the average value over the whole dataset. An un-fine-tuned Llama-3.1-Instruct (8B) is adopted as the ACTION-proposal model with a sampling temperature of 0.7. Since no training is involved, there exists no risk of data cross-contamination.

For ACTION outputs produced by the LLM, regular expressions and Pymatgen utilities are used to apply these ACTIONs onto the original structures. Afterwards, MatterSim-1M performs relaxation with $f_{\max} = 0.01\,\text{eV}/\text{Å}$ and $\text{maxSteps} = 100$ and computes potential energies. Convex-hull energy and novelty are evaluated based on the pymatgen's built-in StructureMatcher tools Ong et al. (2013) and MP2020correction reference library to count unique non-duplicated structures. Structural and compositional validity are examined using Pymatgen and SMACT. Quantitative results are summarized in Table 1. The full prompt and detailed configurations are provided in the appendix.

For baselines, we compare against methods including MatterGen (Nature, 2025) Zeni et al. (2025), PLaID++ (ICML, 2026) Xu et al. (2026), Chemeleon2 (NMI, 2026) Park & Walsh (2026), and other models Gebauer et al. (2019b); Xie et al. (2022); Jiao et al. (2023b). As competing baselines require training, all models use the MP20 dataset split from CDVAE: $\text{train} : \text{val} : \text{test} = 27136 : 9047 : 9046$, with radioactive elements removed. 10,000 CIF outputs of PLaID++ and Chemeleon2 are fetched from their official GitHub repositories, with their energy-above-hull and novelty evaluation settings consistent with Tego. Numerical results of other non-ablation baselines are directly taken from the published GitHub results of MatterGen.

Obs 1. Tego achieves SUN rates comparable to strong baselines with zero secondary training cost, relying only on the local-editing paradigm together with minimal necessary prompts.

4.2 Initial-point Sensitivity and Generated Diversity

We further test Tego using the MP20 test set as initial points, yielding 8 481 samples after removing radioactive elements. We perform editing with a target magnetic moment of $0.01\,\mu_{\text{B}}$and analyse metrics including S.U.N. with respect to the chemical-system complexity and convex-hull energy of starting structures.

Obs 2. As a local-structure editing approach, Tego can be sensitive to the state of starting structures. The results are presented in Figure 2.

We further observe that conventional *ab initio* generation methods may carry inherent empirical biases, which tend to limit the structural diversity of generated materials. As a local editing method, Tego may maintain high structural richness.

Table 1: Performance comparison of different methods across S.U.N., Stable, Unique, and Novel Zeni et al. (2025). Stable is defined as energy above hull < 0.1 eV/atom. In this experiment, Tego performs only one round of structural editing attempt. **Wyckoff Abl.** denotes an ablation study performing Wyck-SEQ ab initio generation based on the same LLM.

Methods	% S.U.N.↑	% Stable↑	% Unique↑	% Novel↑	Validity	
					Structural↑	Composition↑
P-G-SchNet	1.29	3.11	100.0	88.40	99.12	81.90
CDVAE	13.99	19.31	99.89	**92.00**	100.0	86.01
DiffCSP	12.71	36.23	100.0	70.73	100.0	79.98
PLaID++	9.07	**77.27**	58.62	31.87	99.95	**100.0**
Chemeleon2	16.24	51.95	75.20	62.07	99.98	88.82
MatterGen MP	22.27	42.19	100.0	75.44	100.0	88.45
Wyckoff Abl.	2.79	13.34	45.22	57.56	71.43	66.39
Tego	**23.73**	52.90	91.48	69.28	100.0	89.81

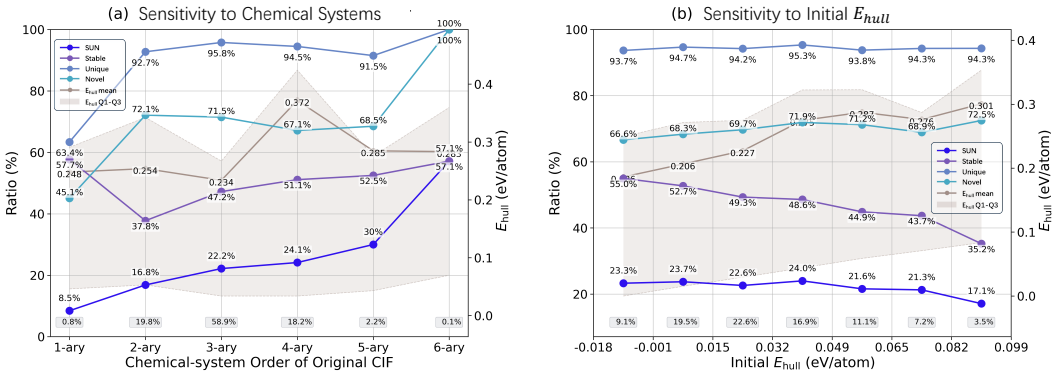

Figure 2: Sensitivity to the characteristics of initial crystal structures. (a) SUN, Stable, Unique and Novel metrics evaluated on initial crystals of different chemical-system orders (1-ary to 6-ary). Low-order chemical systems host abundant already-discovered materials, which restricts the generation of structurally novel candidates and results in very low SUN for 1-ary compounds. Higher-order, more complex chemical systems deliver substantially improved SUN. The right vertical axis presents the mean value and interquartile range (Q1–Q3) of the edited structures' convex hull energy E_{hull}. (b) Performance metrics binned by the initial convex hull energy E_{hull} of input crystals. As the initial E_{hull} increases, the Stable ratio of edited structures gradually declines, indicating that initial thermodynamic status affects the stability of edited outcomes.

We accordingly evaluate absolute counts and Shannon entropy over crystal systems, space groups, and structural prototypes for structures generated by Tego in Table1 and competing baselines. The results are presented in Figure 3. A prototype proxy is defined as the combination of space-group number and anonymous reduced formula. The anonymous formula keeps only atomic stoichiometry and ignores element identities, e.g., $NaCl$ and MgO both correspond to the AB prototype, while Na_2O belongs to the A2B prototype. For baselines, outputs of PLaID++ are obtained from its official repository, whereas the remaining baselines use the 10 000 raw CIF files released by the Xtaledit Benchmark on HuggingFace Negishi & Walsh (2026).

Obs 3. Local-editing frameworks such as Tego largely mitigate structural homogenization caused by model-specific biases, thereby enriching material diversity across crystal systems, space groups, and structural prototypes.

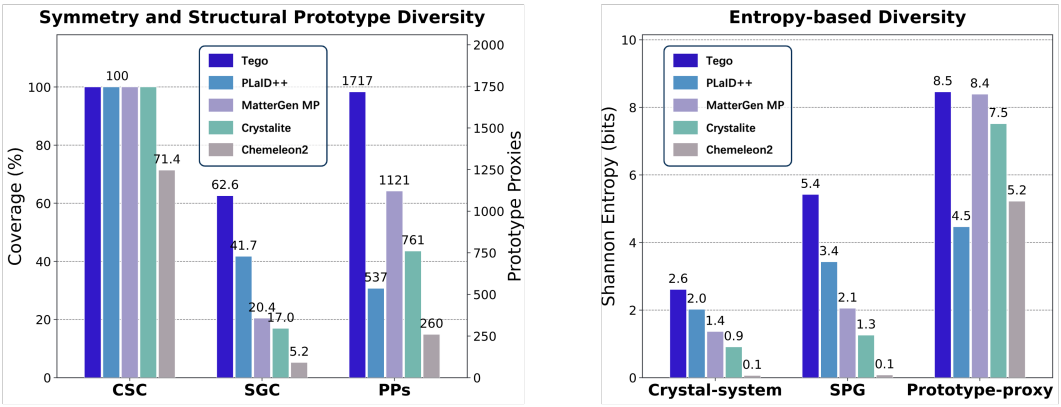

Figure 3: Comparison of structural diversity metrics across different methods. In the left panel, crystal-system coverage (CSC) is the fraction of distinct crystal systems in generated structures out of all 7 crystal systems; space-group coverage (SGC) is the fraction of distinct observed space groups out of all 230 known space groups; and prototype proxies (PPs) denote the count of unique structural prototype proxies. The right panel reports Shannon entropy H computed over crystal systems, space groups and prototype proxies. $H = -\sum_i p_i \log_2 p_i$, where p_i denotes the proportion of generated structures belonging to category i.

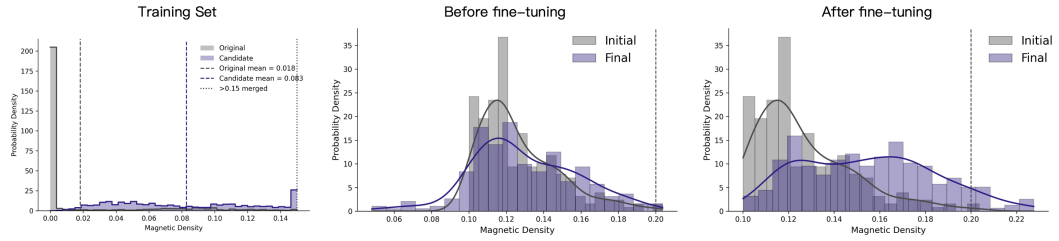

Figure 4: Fine-tuning Ablation.The training set magnetic density averages $0.018\ \mu_B/\text{Å}^3$ for originals and $0.083\ \mu_B/\text{Å}^3$ for candidates.Before fine-tuning, the initial magnetic density distribution peaks around $0.11\ \mu_B/\text{Å}^3$, with the final distribution only slightly shifted; After fine-tuning, the LLM successfully shifts the magnetism of candidates toward higher ranges.

4.3 Property-Guided Structure Design

Dataset Construction From the MP20 train set, we construct a derived data set through random replacement and selection of Wyckoff blocks. For each structure, we **randomly** select a Wyckoff block and replace its atomic species. The modified structures are then relaxed for 300 maxSteps, $f_{max} = 0.05$ eV/Å using MatterSim-1M Yang et al. (2024). We filter out structures whose energy per atom exceeds that of the original structure by more than 0.3 eV/atom. The calculation of potential energy per atom is performed via MatterSim-1M Yang et al. (2024). For the remaining structures, we compute the total magnetic moment using CHGNet Deng et al. (2023) and record the magnetic density $\rho_{\text{mag}} = M_{\text{total}}/V$, where M_{total} is the total magnetic moment and V is the volume of the structure. We denote the resulting dataset as MP20-train-chg. Each sample consists of: the original Wyckoff-SEQ representation, the original magnetic density, the magnetic density after modification, and the corresponding editing action.

From MP20-train-chg, we further select samples exhibiting significant improvement in magnetic density; specifically, we retain samples satisfying either $\rho_{\text{mag}}^{\text{new}} \geq 100 \times \rho_{\text{mag}}^{\text{orig}}$ or $\rho_{\text{mag}}^{\text{new}} > 0.1\ \mu_B/\text{Å}^3$. After excluding unary systems and radioactive elements, we obtain 11,435 training pairs.

Model Fine-tuning We fine-tune Llama-3.1-Instruct (8B) using LoRA with rank 16 on this 11,435 pairs. The training loss decreases from 0.274 to 0.038. The training set distribution and fine-tuning ablation study are presented in Figure 4.

Table 2: Comparison of inference performance and model scale. Metrics: S.U.N.P. denotes the proportion of structures that satisfy Stable, Unique, Novel, and target property criteria simultaneously. Inference Time is the average time cost per generated structure on a single RTX 4090D GPU; Data Cost indicates the size of the training dataset; Task-specific Params. denotes the number of model parameters trained for the CSD task. The 3 successful samples are EuFe ($0.218\ \mu_B/Å^3$), Fe_7Co ($0.226\ \mu_B/Å^3$), and Gd_3Fe ($0.200\ \mu_B/Å^3$), where the values represent their magnetic densities.

Method	S.U.N.P. ↑	Infer. Time ↓	Data Cost ↓	Task-specific Para. ↓
MatterGen Alex MP20	47 / 15360 = 3.06 ‰	67.68 s	600K	76.8 M
Tego MP20	3 / 247 = **12.15 ‰**	2.29 s	**10K**	**41.9 M**
Tego Alex MP20	31 / 4758 = 6.51 ‰	**2.19 s**	**10K**	**41.9 M**

Evaluation We selected 247 structures from the MP20 val set with magnetic density greater than $0.1\,\mu_B/Å^3$ as initial seeds. We perform one-step editing with a target magnetic density of $0.2\,\mu_B/Å^3$. After DFT validation, we obtain 3 novel structures that simultaneously satisfy energy above hull < 0.1 eV/atom and magnetic density $\geq 0.20\,\mu_B/Å^3$. We further selected 4,758 samples with a magnetic moment density greater than $0.10\ \mu_B/Å^3$ from Alex MP20, and targeted a density of $0.20\ \mu_B/Å^3$ for targeted editing. Ultimately, after screening by MatterSim and CHGNet, we obtained a total of 31 SUNP samples. Results are shown in Table2. We conducted the same experiment using the base Llama model without fine-tuning. The model fails to produce any novel structures that satisfy both the stability and the target property constraints.

Obs.4 Compared to MatterGen trained on Alex MP20, Tego achieves a higher success rate in generating novel structures that meet the target property and demonstrates substantially faster inference due to its structured action space and iterative editing framework.

Obs.5 In addition, **we synthesized Fe_7Co (DFT result: $0.226\ \mu_B/Å^3$) and conducted XRD and PPMS measurements**. Although the synthesized samples still contained a small amount of impurity phases, their overall magnetic moment density reached the design target: PPMS gives a measured value of 231 emu/g. Using the mass density of Fe_7Co ($8.08\ \mathrm{g/cm^3}$), the magnetic density converted via the formula is $0.201\ \mu_B/Å^3$, meeting the target of $0.2\ \mu_B/Å^3$.

We conducted additional experiments by increasing the number of editing steps to 3, and obtained 4 stable structures that meet the target: $Fe_{13}Co_3$, Gd_3Fe_2, $Gd_5(FeCo)_2$, and Fe_7Co. Among them, Fe_7Co has already appeared in one-step editing, while the others are newly discovered structures.

Obs.6 Multi-step editing can further expand the search capability of multi-element co-replacement.

5 Limitations and Discussion

In this work, we propose Tego, a local-editing method. Benefiting from the stability and diversity of starting points, Tego can focus more directly on property design and achieve higher success rates for targeted property design on specific CSD tasks. Although reconstructing datasets for distinct target values is unnecessary, Tego currently exhibits limited capacity for fixed-target design. Furthermore, Tego prioritizes property design without explicit stability guarantees, while reinforcement learning and related approaches may further improve stability within this framework.

Nevertheless, Tego exhibits superior robustness under small-data conditions and competitive advantages over baseline models in metrics such as structural stability and S.U.N., while achieving a relatively high target-property achievement rate, holding promise for bridging computational design and practical materials discovery.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing. We used generative AI tools for retrieval and discovery, such as finding related work. We have not used generative AI tools for research ideation or execution. We have not used generative AI tools to draft sections of the paper. We have not used generative AI tools for generating synthetic datasets. We have not used generative AI tools for proving mathematical claims. We have not used generative AI tools for other purposes outside the above list. All AI-assisted work has been reviewed by the authors. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All code and data for dataset construction are available at the anonymous repository: `https://anonymous.4open.science/r/TegoDataset`. Full inference implementation details together with complete prompt templates are documented in the appendix of this paper. We plan to release the model training weights upon paper acceptance. All experimental setups and hyper-parameters can be found in the appendix, enabling full reproduction of our reported results.

ETHICS STATEMENT

No human-subject experiments, sensitive datasets or potentially harmful applications are involved in our study.

REFERENCES

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024. doi: 10.1038/s41467-024-54639-7. URL https://doi.org/10.1038/s41467-024-54639-7.

Zhendong Cao and Lei Wang. Reinforcement fine-tuning for materials design. *Physical Review B*, 113(2), January 2026. ISSN 2469-9969. doi: 10.1103/45zh-44bg. URL http://dx.doi.org/10.1103/45zh-44bg.

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, Sep 2023. doi: 10.1038/s42256-023-00716-3. URL http://dx.doi.org/10.1038/s42256-023-00716-3.

Akihiro FUJII, Yoshitaka Ushiku, Koji Shimizu, Anh Khoa Augustin Lu, and Satoshi Watanabe. Adaptive constraint integration for simultaneously optimizing crystal structures with multiple targeted properties, 2025. URL https://openreview.net/forum?id=NVKwjCIAAX.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a4d8e2a7e0d0c102339f97716d2fdfb6-Paper.pdf.

Niklas W. A. Gebauer, Michael Gastegger, and Kristof T. Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *CoRR*, abs/1906.00957, 2019b. URL http://arxiv.org/abs/1906.00957.

Malte Grunert, Max Großmann, and Erich Runge. Discovery of sustainable energy materials via the machine-learned material space. *Small*, May 2025. ISSN 1613-6829. doi: 10.1002/smll.202412519. URL http://dx.doi.org/10.1002/smll.202412519.

Xiao-Qi Han, Xin-De Wang, Meng-Yuan Xu, Zhen Feng, Bo-Wen Yao, Peng-Jie Guo, Ze-Feng Gao, and Zhong-Yi Lu. AI-driven inverse design of materials: Past, present and future, 2025. URL https://arxiv.org/abs/2411.09429.

Benjamin A. Helfrecht, Rose K. Cersonsky, Guillaume Fraux, and Michele Ceriotti. Structure-property maps with Kernel Principal covariates regression, 2020. URL https://arxiv.org/abs/2002.05076.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *NeurIPS*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/38b787fc530d0b31825827e2cc306656-Abstract-Conference.html.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *CoRR*, abs/2309.04475, 2023b. URL https://doi.org/10.48550/arXiv.2309.04475.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation, 2024. URL https://arxiv.org/abs/2402.03992.

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

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals, 2025a. URL https://arxiv.org/abs/2503.02407.

Nikita Kazeev, Ruiming Zhu, Ignat Romanov, Wei Nong, Andrey E Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff Transformer: Generation of symmetric crystals, 2025b. URL https://openreview.net/forum?id=ursX3k1rTO.

Chang Liu, Hiromasa Tamaki, Tomoyasu Yokoyama, Kensuke Wakasugi, Satoshi Yotsuhashi, Minoru Kusaba, Artem R. Oganov, and Ryo Yoshida. Shotgun crystal structure prediction using machine-learned formation energies, 2024. URL https://arxiv.org/abs/2305.02158.

Shengchao Liu, Hannan Xu, Yan Ai, Huanxin Li, Yoshua Bengio, and Harry Guo. Expert-guided LLM reasoning for battery discovery: From AI-driven hypothesis to synthesis and characterization, 2025. URL https://arxiv.org/abs/2507.16110.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, Nov 2023. doi: 10.1038/s41586-023-06735-9. URL http://dx.doi.org/10.1038/s41586-023-06735-9.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching, 2024. URL https://arxiv.org/abs/2406.04713.

Masahiro Negishi and Aron Walsh. Substitution-based analysis of structural novelty for generative models of materials, 2026. URL https://arxiv.org/abs/2606.23166.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2012.10.028. URL https://www.sciencedirect.com/science/article/pii/S0927025612006295.

Raul Ortega-Ochoa, Alán Aspuru-Guzik, Tejs Vegge, and Tonio Buonassisi. A tomographic interpretation of structure-property relations for materials discovery, 2025. URL https://arxiv.org/abs/2501.18163.

H. Park and A. Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *Nature Machine Intelligence*, 8:1087–1099, 2026. doi: 10.1038/s42256-026-01262-4. URL https://doi.org/10.1038/s42256-026-01262-4.

Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G. Aberle, Shijing Sun, Xiaonan Wang, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, Kedar Hippalgaonkar, Yousung Jung, and Tonio Buonassisi. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, January 2022. ISSN 2590-2385. doi: 10.1016/j.matt.2021.11.032. URL http://dx.doi.org/10.1016/j.matt.2021.11.032.

Laura Ruple, Luca Torresi, Henrik Schopmans, and Pascal Friederich. Symmetry-aware bayesian flow networks for crystal generation, 2025. URL https://arxiv.org/abs/2502.03146.

Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M. Wood. Flowllm: Flow matching for material generation with large language models as base distributions, 2024. URL https://arxiv.org/abs/2410.23405.

Tin Hadži Veljković, Joshua Rosenthal, Ivor Lončarić, and Jan-Willem van de Meent. Crystalite: A lightweight transformer for efficient crystal modeling, 2026. URL https://arxiv.org/abs/2604.02270.

H. Wondratschek and U. Müller (eds.). *International Tables for Crystallography, Volume A1: Symmetry Relations Between Space Groups*. International Tables for Crystallography. Springer, Dordrecht, 1 edition, 2004. doi: 10.1107/97809553602060000101. Published in cooperation with the International Union of Crystallography (IUCr).

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

Andy Xu, Rohan Desai, Larry Wang, Ethan Ritz, and Gabriel Hope. Plaid++: A preference aligned language model for targeted inorganic materials design, 2026. URL https://arxiv.org/abs/2509.07150.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. Mattersim: A deep learning atomistic model across elements, temperatures and pressures, 2024. URL https://arxiv.org/abs/2405.04967.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 2025. doi: 10.1038/s41586-025-08628-5.

Gongbo Zhang, Yanting Li, Renqian Luo, Pipi Hu, Yang Yang, Zeru Zhao, Lingbo Li, Guoqing Liu, Zun Wang, Ran Bi, Kaiyuan Gao, Liya Guo, Yu Xie, Chang Liu, Jia Zhang, Tian Xie, Robert Pinsler, Claudio Zeni, Ziheng Lu, Hongxia Hao, Yingce Xia, Marwin Segler, Maik Riechert, Wei Yang, Hao Jiang, Wen-Bin Zhang, Zhijun Zeng, Yi Zhu, Li Dong, Xiuyuan Hu, Li Yuan, Lei Chen, Haiguang Liu, and Tao Qin. UniGenX: a unified generative foundation model that couples sequence, structure and function to accelerate scientific design across proteins, molecules and materials, 2025. URL https://arxiv.org/abs/2503.06687.

Peichen Zhong, Xinzhe Dai, Bowen Deng, Gerbrand Ceder, and Kristin A. Persson. Crystal structure prediction with host-guided inpainting generation and foundation potentials. *Materials Horizons*, 12(22):9669–9678, 11 2025. ISSN 2051-6347. doi: 10.1039/d5mh00774g. URL https://doi.org/10.1039/d5mh00774g.

APPENDIX CONTENTS

A Dataset Construction Details **14**

A.1 Symmetry-aware substitution. 14

A.2 Chemical substitution space. 14

A.3 Relaxation and energetic filtering. 15

A.4 Computational cost. 15

B Initial E_{hull} Distribution **16**

C Inference Details of Tego **17**

C.1 Model and Inference Configuration 17

C.2 Prompting Protocol 19

C.3 Wyck-SEQ Representation and ACTION 20

C.4 Inference Procedure 21

C.5 Inference Efficiency and Runtime Breakdown 21

D Training Configuration **22**

E Reliability of MLFF-Based Evaluation **23**

Table 3: Main configuration used for dataset construction.

Stage	Parameter	Value
Random sampling	Random seed	42
	Wyckoff blocks sampled	100%
	Max. substitutions per block	3
Symmetry analysis	Symmetry tolerance (`symprec`)	10^{-2}
	Angle tolerance	5°
Relaxation	Interatomic potential	MatterSim-1M
	Optimizer	FIRE
	Force threshold ($f_{\max}$)	0.01 eV/Å
	Maximum relaxation steps	100
	Atomic positions	Relaxed
	Unit cell	Relaxed

A Dataset Construction Details

The reproduction guide and released data are available in the anonymous repository:

`https://anonymous.4open.science/r/Tego-Dataset`.

We construct the training dataset by applying symmetry-aware local substitutions to source crystal structures, followed by structural relaxation and energetic filtering. The goal of this procedure is not to exhaustively enumerate all chemically possible substitutions, but rather to construct a controlled neighborhood around each source structure, where every candidate differs through a simple and explicitly traceable local edit.

A.1 Symmetry-aware substitution.

For each input crystal, we first identify symmetry-equivalent atomic sites using `SpacegroupAnalyzer`. Sites are grouped into Wyckoff blocks according to their chemical species, Wyckoff letter, site-symmetry symbol, and symmetry-equivalent atom index. A substitution is subsequently applied to the entire block rather than to an individual atom, such that the edit remains consistent with the symmetry representation used by our model.

The main parameters that affect candidate construction are summarized in Table 3.

A.2 Chemical substitution space.

We use a predefined same-valence substitution pool to prevent unrestricted element replacement from producing an excessive number of chemically implausible candidates. For an element A, its nominal oxidation state is obtained from the common oxidation states provided by `pymatgen`. Candidate replacements are then sampled from the corresponding valence group, excluding A itself.

The substitution space used in all experiments is reported in Table 4.

A small set of radioactive or otherwise undesirable elements is explicitly excluded from substitution. If more than three admissible replacements are available for a Wyckoff block, the candidate list is shuffled and truncated to three elements. We use a fixed Python pseudo-random seed of 42 throughout dataset construction. Since all Wyckoff blocks are retained (`block_sample_ratio`=1.0), the random seed mainly affects the choice of replacement elements when the corresponding same-valence pool contains more than three alternatives.

For each selected edit, all symmetry-equivalent atoms belonging to the selected Wyckoff block are replaced simultaneously. The resulting operation is stored explicitly as

$\texttt{ACTION} = (\texttt{Wyck.ID}, \texttt{multiplicity}, \texttt{Wyck.letter}, \texttt{site symmetry}, \texttt{new species})$,

Table 4: Element pools used for same-valence substitution. For each Wyckoff block, the candidate pool is randomly shuffled and at most three replacement elements are retained.

Nominal valence	Candidate elements
$+1$	Li, Na, K, Rb, Cs, Ag
$+2$	Be, Mg, Ca, Sr, Ba, Zn, Cd, Mn, Fe, Co, Ni, Cu
$+3$	Al, Ga, In, Sc, Y, La, Ce, Pr, Nd, Sm, Gd, Dy, Er, Yb, Bi, Cr, Fe
$+4$	Ti, Zr, Hf, Si, Ge, Sn, Ce, Mn, Ir, Pt
$+5$	V, Nb, Ta, P, As, Sb
$+6$	Cr, Mo, W, S, Se, Te
-1	F, Cl, Br, I
-2	O, S, Se, Te
-3	N, P, As

which provides a direct correspondence between each candidate crystal and the local structural modification from which it was obtained.

A.3 Relaxation and energetic filtering.

The source structure and all substituted candidates are evaluated using the same MatterSim potential. Each structure is relaxed using the FIRE optimizer with both atomic coordinates and lattice parameters allowed to change. We use a maximum force criterion of

$$f_{\max} = 0.05 \text{ eV/Å}$$

and a maximum of 300 optimization steps.

Let E_{src} denote the MatterSim energy per atom of the relaxed source structure and E_{cand} the corresponding value for a relaxed substituted candidate. We define

$$\Delta E = E_{\text{cand}} - E_{\text{src}}.$$

A candidate is retained when

$$\boxed{\Delta E \leq 0.30 \text{ eV/atom}}$$

and is otherwise discarded.

We emphasize that this quantity is *not* the energy above the convex hull, E_{hull}. It is a local energy difference relative to the corresponding source crystal under the same MatterSim potential. We use this criterion as a lightweight filter for removing strongly unfavorable substitutions while retaining a reasonably diverse neighborhood around each source crystal.

If relaxation of a source structure fails, its single-point MatterSim energy on the original geometry is used as the reference energy. In contrast, a substituted candidate is retained only when its own relaxation and energy evaluation complete successfully.

A.4 Computational cost.

Dataset construction is performed with MatterSim inference on GPU. In our implementation, the complete preprocessing and candidate-screening pipeline requires less than 5 hours on 4 NVIDIA RTX 4090D GPUs.

B INITIAL E_{hull} DISTRIBUTION

To clarify the initial stability of the structures used in different experiments, we report the corresponding E_{hull} distributions in Figure 5. We supplement this information because different reference potentials yield distinct convex hull energy results, for your reference.

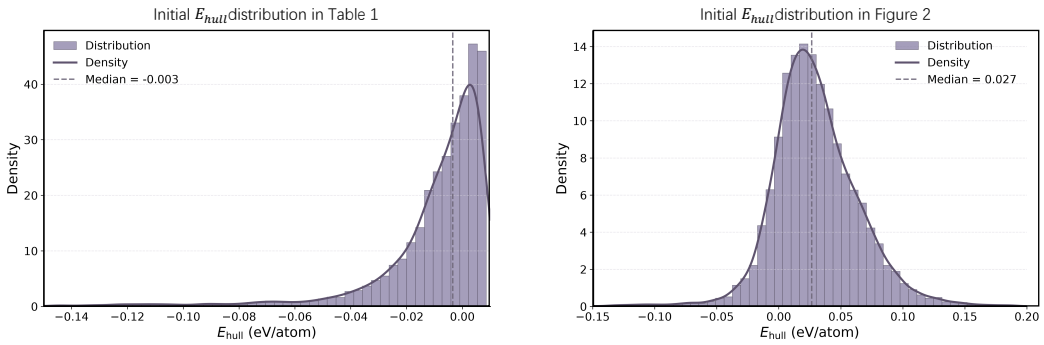

Figure 5: Initial E_{hull} distributions of the structures used in Table 1 (left) and Figure 2 (right). The dashed lines indicate the median values.

In Table 1 and Figure3, the starting points of Tego are the 10,000 samples with the lowest convex hull energies selected from MP20. Among them, the N-th e_{hull} is 0.0085 eV/atom. For the sensitivity experiment in Figure2, to obtain starting points with a broader energy distribution, we adopt the MP20 test set as initial points, where the maximum value is 0.1922 eV/atom. We report these distributions explicitly because the two experiments do not start from identical energetic distributions.

C Inference Details of TEGO

This section provides additional implementation details for the inference pipeline, including the model configuration, prompting protocol, structural representation, action format, and runtime breakdown. We focus on parameters that can materially affect the generated structures or the reported properties, while omitting routine engineering details such as file organization and intermediate logging.

C.1 Model and Inference Configuration

We use Llama-3.1-8B-Instruct as the base language model. For the fine-tuned variant, the corresponding LoRA adapter is loaded on top of the same base model at inference time. Unless otherwise specified, inference is performed in BF16 precision without 4-bit quantization. The tokenizer associated with the base model is used together with its native chat template.

Table 5 summarizes the main inference parameters. We distinguish between the defaults implemented in the inference script and the configuration used in the reported single-step evaluation. In particular, the evaluation runs use one generated action per structure ($k = 1$) and one editing round. The more general implementation also supports multiple candidates and iterative multi-round editing.

The magnetic density used for candidate comparison is computed from the CHGNet-predicted atomic magnetic moments as

$$\rho_{\mathrm{mag}} = \frac{\sum_i m_i}{V}, \tag{10}$$

where m_i denotes the predicted magnetic moment of atom i and V is the volume of the relaxed structure. Therefore, candidate ranking is performed only after structural relaxation.

MatterSim relaxation uses the FIRE optimizer for at most 100 optimization steps with a force convergence threshold of 0.05 eV/Å. Cell relaxation is enabled; the implementation uses a `FrechetCellFilter` when available and falls back to a `UnitCellFilter` otherwise.

For action generation, sampling is enabled whenever the temperature is non-zero. The default generation setting uses a temperature of 0.7 and top-$p = 0.95$. For distributed inference, the random seed of process r is set to

$$s_r = 42 + r, \tag{11}$$

for both PyTorch and NumPy. Consequently, exact reproduction of stochastic generation also requires keeping the number of distributed processes and the sample-to-rank assignment unchanged.

Table 5: Main configuration of the inference pipeline. Parameters listed here are those that can directly affect generation, structure reconstruction, relaxation, or property evaluation.

Component	**Parameter**	**Setting**
Language model	Base model	Llama-3.1-8B-Instruct
Language model	Adapter	LoRA / none for base-model evaluation
Language model	Numerical precision	BF16
Language model	4-bit quantization	Disabled
Generation	Maximum input length	1300 tokens
Generation	Maximum new tokens	96
Generation	Temperature	0.7
Generation	Top-p	0.95
Generation	Candidates per round k	1 in reported evaluation; 8 by default
Generation	Editing rounds	1 in reported evaluation; 5 by default
Symmetry	Symmetry tolerance (`symprec`)	10^{-2}
Symmetry	Angular tolerance	5.0°
Representation	Coordinate precision	3 decimal places
Action validation	Inconsistent Wyckoff metadata	Rejected
Relaxation	Potential	MatterSim
Relaxation	Optimizer	FIRE
Relaxation	Maximum steps	100
Relaxation	Force threshold	0.05 eV/Å
Relaxation	Cell relaxation	Enabled
Property evaluation	Predictor	CHGNet
Property evaluation	Magnetic moment aggregation	Absolute sum
Candidate selection	Criterion	Highest magnetic density
Early stopping	Stop upon reaching target	Disabled
Reproducibility	Base random seed	42

C.2 PROMPTING PROTOCOL

The model is prompted as an action generator rather than as a free-form crystal generator. The prompt explicitly defines the Wyck-SEQ representation and constrains the model to modify the chemical species of exactly one Wyckoff-equivalent group. Lattice parameters, fractional coordinates, multiplicity, Wyckoff letter, and site symmetry are not allowed to change directly in the generated action.

The system prompt used during inference is reproduced below.

```
You are a crystal-material inverse-design action generator.
Your task is to read an initial crystal structure, its magnetic density,
and a target magnetic density. Then you must output exactly one local
    structure edit action, so that the structure obtained after this
    editing operation attains the target magnetic density.
The initial structure input will be represented in the following Wyck-SEQ
     format:
1. <SPG=number@symbol> gives the space group number and Hermann-Mauguin
    symbol.
2. <LATTICE ...> gives the lattice parameters a, b, c, alpha, beta, gamma
    .
3. Each <WYCK_id=...> block describes one Wyckoff-equivalent atomic group
    .
4. In each Wyckoff block:
   - WYCK_id is the group index.
   - mult is the Wyckoff multiplicity.
   - wy is the Wyckoff letter.
   - sym is the site symmetry.
   - species is the chemical element occupying this Wyckoff group.
   - The following <idx=... x=... y=... z=...> lines are fractional
       coordinates
     of all equivalent sites in this Wyckoff group.
5. The structure begins with <BOS> and ends with <EOS>.
Action format:
You must output only one action block in the following format:
<ACTION>
<WYCK_id=... mult=... wy="..." sym="..." species="...">
</ACTION>
- A valid action changes the species of exactly one whole Wyckoff block.
- The coordinates, lattice, multiplicity, Wyckoff letter, and site
    symmetry must not be changed by the action.
Strict output rules:
1. Output only one ACTION block. Do not output explanations.
2. Choose an existing WYCK_id from the input structure.
3. Prefer substitutions that are likely to preserve charge balance and
    structural stability.
4. For magnetic-density enhancement, transition metals and rare-earth
    elements may be useful, but they should still be selected under
    symmetry and stability constraints.
5. Do NOT use radioactive elements including Tc, Pm, Po, At, Rn, Fr, Ra,
    Ac, Th, Pa, U.
```

For each structure, this system instruction is combined with a user message containing the current structure, its current magnetic density, and the target magnetic density:

```
Input:

<ORIGINAL_WYCK_SEQ>
{current Wyck-SEQ}
</ORIGINAL_WYCK_SEQ>

<ORIGINAL_MAG_DENSITY>
{current magnetic density}
</ORIGINAL_MAG_DENSITY>

<TARGET_MAG_DENSITY>
{target magnetic density}
</TARGET_MAG_DENSITY>

Now output the single best action.
```

The current magnetic density is formatted to four decimal places, whereas the target value is formatted to three decimal places before being inserted into the prompt. The complete conversation is then converted using the native chat template of the tokenizer.

C.3 WYCK-SEQ REPRESENTATION AND ACTION

Instead of directly asking the language model to regenerate a complete CIF, we first convert each crystal into a compact symmetry-aware text representation, denoted as Wyck-SEQ. Space-group information is determined using a symmetry tolerance of 10^{-2} and an angular tolerance of 5°. Symmetry-equivalent sites are grouped into Wyckoff blocks, with each block assigned a deterministic `WYCK_id`.

A schematic example of the representation is shown below. The example is intended to illustrate the serialization format rather than to represent a specific sample from the evaluation set.

```
<BOS>
[SPG=229@Im-3m]
[LATTICE a=2.866, b=2.866, c=2.866,
 alpha=90.000, beta=90.000, gamma=90.000]
<WYCK_id=0 mult=2 wy="a" sym="m-3m" species="Fe">
  <idx=0 x=0.000 y=0.000 z=0.000>
  <idx=1 x=0.500 y=0.500 z=0.500>
<EOS>
```

Rather than predicting new atomic coordinates, the model outputs a single local editing operation. For example,

```
<ACTION>
<WYCK_id=0 mult=2 wy="a" sym="m-3m" species="Co">
</ACTION>
```

replaces all atoms belonging to `WYCK_id=0` by Co while leaving the lattice and fractional coordinates unchanged before relaxation. The `mult`, `wy`, and `sym` fields generated by the model must agree with those of the selected block. Actions with inconsistent metadata are discarded.

This design substantially reduces the output space of the language model: the model only decides *where* to edit and *which element* to use, while crystallographic geometry is inherited from the current structure. The edited structure is subsequently relaxed by MatterSim, so atomic coordinates and lattice parameters are allowed to respond physically after the discrete edit has been applied.

Table 6: Average inference time per sample and the relative contribution of each stage. The stage-level values are reported independently and rounded, and therefore their displayed sum may differ slightly from the separately measured end-to-end runtime.

Model	Total	Data loading	Initial eval.	LM inference	Action	MatterSim	CHGNet	Saving
Base	1.567	0.011	0.068	0.708	0.000	0.692	0.063	0.054
Base (%)	100.00	0.71	4.34	45.15	0.03	44.14	4.02	3.44
LoRA	2.186	0.011	0.069	1.192	0.000	0.814	0.065	0.057
LoRA (%)	100.00	0.50	3.17	54.70	0.02	37.36	2.97	2.62

C.4 Inference Procedure

For completeness, the inference procedure for one editing round can be summarized as follows:

1. Parse the initial crystal and construct its Wyck-SEQ representation.
2. Evaluate the initial MatterSim energy and CHGNet magnetic density.
3. Provide the Wyck-SEQ, current magnetic density, and target magnetic density to the language model.
4. Generate k candidate ACTIONs and remove syntactically invalid or duplicated actions.
5. Apply each valid action to the complete corresponding Wyckoff group.
6. Relax the resulting structure using MatterSim.
7. Evaluate the relaxed structure using CHGNet.
8. Among the valid candidates, select the structure with the highest magnetic density as the input to the next round.

When multiple rounds are enabled, the selected relaxed structure is re-encoded into a new Wyck-SEQ before the next action is generated. Thus, the model observes the updated geometry and property value rather than repeatedly editing the original structure.

C.5 Inference Efficiency and Runtime Breakdown

We additionally profile the end-to-end inference pipeline on a per-sample basis. Model initialization is performed once before processing the dataset and is therefore excluded from the per-sample timing. The reported runtime includes structure preparation and initial evaluation, language-model generation, action application, MatterSim relaxation, CHGNet prediction, and result serialization.

Table 6 reports the average runtime for the base model and the LoRA-adapted model. Both variants use the same downstream relaxation and property-evaluation pipeline.

D TRAINING CONFIGURATION

We fine-tune Llama-3.1-8B-Instruct using supervised fine-tuning (SFT) with LoRA. The main training configuration is summarized in Table 7.

Table 7: Main configuration used for LoRA fine-tuning.

Category	**Parameter**	**Setting**
Model	Base model	Llama-3.1-8B-Instruct
Model	Precision	BF16
Model	Quantization	None
LoRA	Rank r	16
LoRA	α	32
LoRA	Dropout	0.05
LoRA	Target modules	`q,k,v,o,gate,up,down proj.`
Optimization	Optimizer	AdamW
Optimization	Learning rate	2×10^{-4}
Optimization	Weight decay	0.01
Optimization	LR scheduler	Cosine
Optimization	Warmup ratio	0.03
Optimization	Max. gradient norm	1.0
Training	Batch size / GPU	2
Training	Gradient accumulation	8
Training	Number of GPUs	4
Training	Effective batch size	64
Training	Maximum sequence length	1300 tokens
Training	Gradient checkpointing	Enabled
Training	Random seed	42

The training sample consists of the original Wyck-SEQ, its original magnetic density, the target magnetic density, and the corresponding editing action. Only the assistant ACTION tokens contribute to the causal language-modeling loss; all prompt tokens are masked with `label` $= -100$. The original and target magnetic densities are formatted to four and three decimal places, respectively.

E RELIABILITY OF MLFF-BASED EVALUATION

Since our evaluation pipeline relies on machine-learning-based surrogate models, we further examine their prediction errors.

MatterSim. MatterSim reported high prediction accuracy in its paper. It achieves an energy MAE of approximately 0.029 eV on the MPF-TP dataset. Considering the number of atoms in the structures, this corresponds to an approximate per-atom error on the order of $0.0015 - 0.005$ eV/atom. Although this value cannot be directly interpreted as an exact E_{hull} error, it is substantially smaller than the 0.1 eV/atom stability threshold adopted in our evaluation. We therefore consider MatterSim sufficiently accurate for distinguishing structures at the stability scale considered in this work.

CHGNet. We directly compared the magnetic densities predicted by CHGNet with the DFT values recorded in MP20. The resulting mean absolute errors are summarized in Table 8.

Table 8: MAE of CHGNet magnetic-density predictions against MP20 DFT values.

Split	**MAE** ($\mu_B/\mathring{A}^3$)
Train	0.001054
Validation	0.001049
Test	0.001072

Compared with the target magnetic density of 0.2 $\mu_B/\mathring{A}^3$, these errors correspond to a relative scale of approximately 0.005, i.e., about 0.5%. The errors are also highly consistent across the train, validation, and test splits. These results suggest that, despite being a learned predictor, CHGNet provides sufficiently accurate and stable magnetic-density estimates for the property-guided evaluation used in this work.