GENERATING MATERIALS WITH EXCELLENT PROPERTIES VIA ACTOR-CRITIC DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models, as the basis for state-of-the-art foundation models for materials discovery, potentially signal a paradigm shift from the expensive, manual trial-and-error industrial practice. Technology across multiple fields is powered by materials with desirable and often rarely seen properties. Classifier-free (or guided) diffusion has been the main approach in generating materials with these conditioned properties, however, these approaches tend to fall back to existing materials when conditioned on extreme property values. This poses significant hurdles in discovering novel, stable materials with excellent properties. Instead of formulating the problem as conditional diffusion, we define a novel Reinforcement Learning (RL) process for the reverse diffusion to maximize property values as rewards. RL explores the composition space freely, hence not bounded by existing materials in the training dataset. We present CrystRL, a reinforcement learning-based method that can generate novel, stable materials with excellent properties. CrystRL fine-tunes the unconditional reverse diffusion using an actor network, which suggests denoising moves to good materials. It is supplemented by a critic network, which predicts the best property values from a noisy state. Experimental results show that CrystRL discovers novel and stable crystals with out-of-training-distribution ML-predicted bulk moduli of up to 408 GPa and DFT-verified bulk moduli of up to 397 GPa more consistently than classifier-free guided generation.

1 INTRODUCTION

Recent advances in generative models made them powerful foundation models for AI-driven materials discovery (Zeni et al., 2025). By directly generating stable candidate materials, scientists are able to overcome the low throughput of traditional manual, trial-and-error methods for discovering new materials with interesting and useful properties. These models effectively learn a complex distribution over thermodynamically stable crystal structures from datasets such as Materials Project (Horton et al., 2025), and sample from that distribution following reverse diffusion.

Classifier-free guidance (Ho & Salimans, 2021) for diffusion models has been the state of the art approach in generating materials with specific property values. However, these approaches tend to fall back to existing materials when conditioned on extreme property values. Standard conditional diffusion models inherently function as interpolators. They struggle significantly with out-of-distribution (OOD) generation, frequently failing to extrapolate into the extreme regions of the chemical property space—such as discovering materials with an exceptionally high bulk modulus—required for breakthrough technological applications. Consider Figure 1 (top left). As the model is conditioned on progressively more extreme properties, the model is not able to find materials beyond those in the training distribution. Particularly, the model struggles to generate materials with a bulk modulus exceeding 400 GPa (ML-predicted). In addition, the novelty of generated materials decreases as it falls back to existing materials. This is a direct consequence of classifier-free guidance, where a model relies on learning the complete conditional probability distribution over the target property. When prompted with extreme target values that lack representation in the pre-training data, the conditioning signal becomes uninformative and saturates towards materials in the dataset. To avoid generating physically unstable structures in these unmapped regions, the guidance mechanism safely defaults to reconstructing the nearest known stable materials it memorized during training, effectively sacrificing novelty to act as an interpolator rather than discovering truly novel, out-of-distribution configurations.

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.

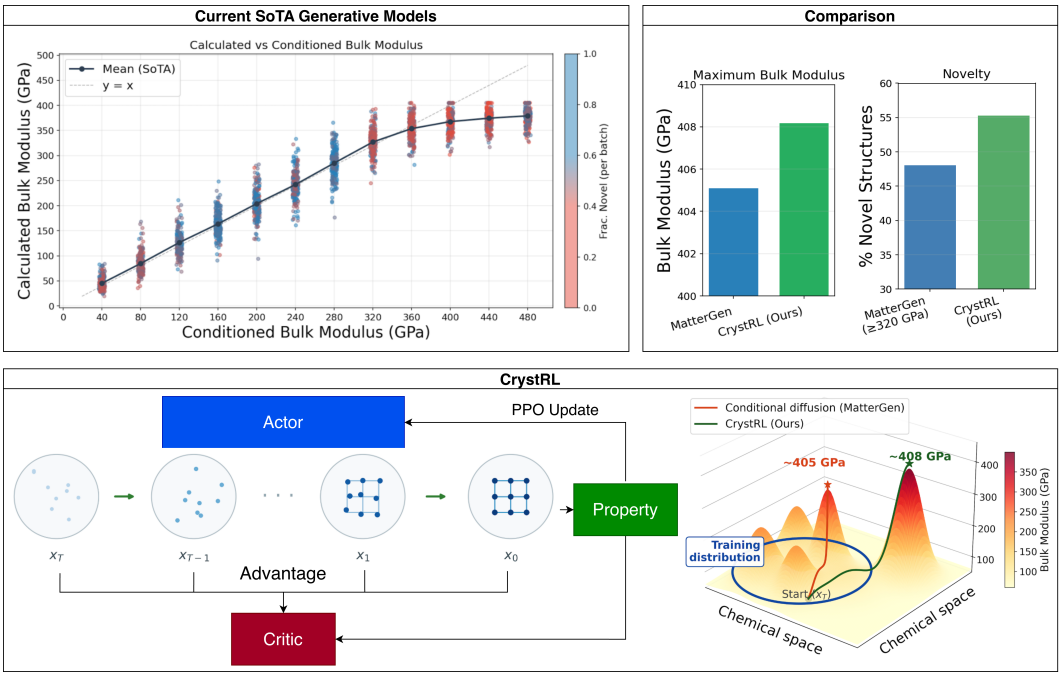


Figure 1: CrystRL is able to generate novel, stable materials with excellent properties. ***(Top Left)*** The materials generated by the state-of-the-art classifier-free guided approach (MatterGen) saturate in bulk modulus as we condition on more and more extreme property values. Moreover, the generated materials fall back to existing ones (denoted by red color) as the conditioned values increase. ***(Top Right)*** CrystRL outperforms MatterGen in both the bulk modulus and the percentage of new materials. ***(Bottom)*** CrystRL explores beyond the training distribution via an integrated actor-critic approach and discovers materials with an ML-predicted bulk modulus higher than that seen in both SoTA models and the Alex-MP-20 dataset (Höllmer et al., 2025) .

This nature of conditional diffusion poses significant hurdles in discovering novel, stable materials with excellent properties, as classifier-free guidance fundamentally has difficulties extrapolating to out-of-distribution properties.

Instead of formulating the problem as conditional diffusion, we define a novel Reinforcement Learning (RL) process within reverse diffusion that maximizes property values as rewards. RL explores the composition space freely, as the support of the distribution that we sample from changes from that learned from the dataset to that of all materials that we can analyze with RL techniques.

We present CrystRL, a reinforcement learning-based method that can generate novel, stable materials with excellent properties. CrystRL searches for both continuous variables (fractional atomic coordinates and lattice parameters) and discrete variables (chemical species) using separate actor networks that point out reverse diffusion steps towards materials with excellent property values. Because physical simulations only provide a sparse terminal reward, we train an intermediate time-conditioned surrogate model that serves as a critic. This critic predicts the expected property values at the end of each denoising trajectory from intermediate noisy states, providing the RL agent with step-by-step guidance. To ensure training stability and prevent policy collapse, we freeze the base diffusion model and apply our RL updates exclusively through Low-Rank Adaptation layers.

Experimental results show that with a fixed budget of 512 structures, CrystRL shifts the distribution of generated bulk moduli upward relative to MatterGen, creating over twice as many structures above 350 GPa (52.5% vs. 20.1%) and above 370 GPa (14.26% vs. 6.25%). In addition, a cumulative mean of 24.18% during training of the materials CrystRL generates are simultaneously stable, novel, and unique. CrystRL is able to achieve these benchmarks without relying on additional, expensive DFT calculations within the training loop.

Our contribution is as follows:

- We observe that current classifier-free (guided) conditional diffusion models for materials discovery fall back to existing materials when prompted to generate materials with excellent properties. Motivated by this, we formulate materials discovery as a reinforcement learning problem.
- We outline a novel physics-guided actor-critic framework capable of fine-tuning score-based graph diffusion models to discover crystalline materials with extreme, out-of-distribution properties. In this process, we develop a surrogate time-conditioned model that accurately predicts the relaxed bulk modulus value with an MAE as low as 11 GPa given a noisy intermediate state.
- We show that, under a fixed generation budget, CrystRL shifts the distribution of generated bulk moduli well above that of classifier-free guided generation, and analyze its top candidates in the C–Re system, which reach bulk moduli of up to 408 GPa (ML-predicted) and 397 GPa (DFT-verified).

2 Background

2.1 Diffusion Models

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019) are a class of generative models that learn to synthesize data by inverting a gradual noising process. Rather than explicitly computing complex probability densities, they operate by systematically corrupting data over a series of timesteps until it reaches a known prior distribution, typically pure noise. A neural network is then trained to approximate the reverse trajectory, learning to iteratively denoise the state step-by-step to recover the original data distribution.

Recent works have applied diffusion across a variety of data types, leading to two primary branches of diffusion frameworks: continuous and discrete. Continuous diffusion models, such as Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) and score-based generative models (Song et al., 2021) have become standard architectures, driving state-of-the-art results in image and video synthesis. Conversely, discrete diffusion models have been developed to handle categorical or relational data such as text, or atomic species. Instead of Gaussian noise, these models utilize categorical transition matrices to model state changes over finite alphabets. Prominent examples include Discrete Denoising Diffusion Probabilistic Models (D3PMs) (Austin et al., 2021) and Score Entropy Discrete Diffusion models (SEDD) (Lou et al., 2023) that allow diffusion to operate directly on discrete tokens.

While standard diffusion models excel at unconditionally capturing the underlying data distribution and conditionally steering towards pre-defined areas of the distribution, they still face limitations in steering the generative process toward new, out-of-distribution target properties.

2.2 Generative Models for Materials

Unlike image generation, where the state space consists of fixed grids of continuous pixels, generating crystalline materials requires modeling complex periodic graphs. We define a crystal as a tuple $\mathcal{M} = (\mathbf{A}, \mathbf{X}, \mathbf{L})$, where $\mathbf{A}$ represents discrete atomic species, $\mathbf{X}$ represents continuous fractional coordinates, and $\mathbf{L}$ represents continuous lattice parameters.

Recently, deep learning approaches have been effective in accelerating materials discovery (Merchant et al., 2023; Xie et al., 2021; Dan et al., 2020; Jiao et al., 2023; Yang et al., 2024b; Zeni et al., 2025; Mistal et al., 2023; Antunes et al., 2024; Zheng et al., 2023). Early approaches utilizing Variational Autoencoders (VAEs) (Xie et al., 2021) and Generative Adversarial Networks (GANs) (Dan et al., 2020) successfully generated graphs but struggled to produce thermodynamically stable structures, with high energy-above-hull (E_{hull}) penalties. Recent advancements in generative models such as CDVAE (Xie et al., 2021), DiffCSP (Jiao et al., 2023), UniMat (Yang et al., 2024b), and MatterGen (Zeni et al., 2025) have been able to generate stable materials by enforcing strict physical symmetries during the denoising process, but still struggle with generation at extreme property ranges.

These models typically augment a probability distribution to generate materials with specific properties; they create an alternative model $p(x|c)$ (Ho & Salimans, 2021), where c is a property that

the model is conditioned on. While this has been very effective in generating samples with a given property, these models inherently act as interpolators of their training distributions e.g. the Materials Project (Horton et al., 2025). Because of this, discovering novel materials with extreme properties such as an extremely high bulk modulus requires pushing the model beyond its learned supports, which is difficult for these techniques.

2.3 Reinforcement Learning for Diffusion Models

The intersection of reinforcement learning and diffusion models has recently emerged as an active area of research across two primary lines of work: using diffusion models to generate RL policies (diffusion for RL)(Fang et al., 2025), and using RL to fine-tune pre-trained generative diffusion models (RL for diffusion)(Black et al., 2023; Zhang et al., 2024). Within the latter, to optimize diffusion models for downstream reward functions, recent literature has proposed casting the reverse denoising process as a Markov Decision Process (MDP). In representative frameworks such as Denoising Diffusion Policy Optimization (DDPO) (Black et al., 2023), the noisy state x_t acts as the MDP state, and the denoising prediction to x_{t-1} acts as the policy action $a_t \sim \pi_\theta(a_t \mid x_t)$.

However, we cannot directly apply this framework to existing materials generation methods. Existing DDPO implementations primarily focus on continuous domains, such as image generation. Crystalline materials require a hybrid action space: continuous actions for atomic coordinates and lattice parameters, and discrete actions for selecting chemical species. The standard continuous policy gradient methods do not directly translate to the categorical distributions governing these discrete atomic types.

This MDP perspective highlights a difference in objectives between our approach and existing RL-for-diffusion methods. The traditional goal of diffusion is to generate samples from a specific probability distribution; many RL fine-tuning methods simply attempt to augment that distribution and sample from the modified space. On the other hand, our objective is to find the maximum point of the target property, searching for the peak reward rather than sampling from a reweighted distribution. Through reinforcement learning, we are no longer constrained by the supports of the learned probability distribution. We are instead able to explore beyond the learned support from the original training data to discover novel materials with exceptional target properties. Our formulation treats generation strictly as a sequence of targeted actions designed to maximize a physical reward, without needing to evaluate about the statistical likelihood of the original distribution.

3 Model Architecture

3.1 Problem Statement

The fundamental challenge in computational materials discovery is generating novel, physically stable structures that maximize specific target properties, such as the bulk modulus. Because the search space of continuous geometric configurations and discrete chemical species is massive, and the physical evaluation of these target properties relies on computationally intensive simulations, direct property optimization is intractable.

To solve this, we frame the materials generation process as a sequential optimization task aimed at maximizing a physical reward. We assume a pre-trained mixed diffusion model parameterized by θ. Suppose this diffusion model produces a sample distribution $x_0 \sim p_\theta(x_0)$, where x_0 represents a fully denoised material. To optimize this within an Actor-Critic framework, we formulate the reverse diffusion process as a Markov Decision Process (MDP). Let the state x_t be the noisy material at timestep t within the reverse diffusion process, represented by the aforementioned tuple of properties $(\mathbf{A}, \mathbf{X}, \mathbf{L})$ (atomic species, fractional coordinates, and lattice). The action $a_t = \Delta x_t = x_{t-1} - x_t$ represents the denoising step sampled from the policy $\pi_\theta(a_t \mid x_t)$. A generation trajectory is thus defined as $\tau = (x_T, a_T, x_{T-1}, a_{T-1}, \ldots, x_1, a_1)$.

In standard diffusion, the model is trained to minimize a variational bound, effectively forcing the reverse transition process to match the true data distribution $q(x)$. We reframe the diffusion process as reinforcement learning when we discard this data-matching objective and interpret the reverse process as a trainable stochastic policy instead, where $\pi_\theta(a_t \mid x_t) := p_\theta(x_{t-1} \mid x_t)$, a "step" moving from one noisy material to another. Under this framework, the denoising steps no longer sample from

a data distribution and instead become a series of actions on a landscape optimized to construct a final material x_0 with maximal physical properties.

Our RL objective then is to fine-tune the policy parameters θ to maximize the expected reward over these trajectories:

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[R(\tau)\right]$$

In this MDP, the agent receives a sparse terminal reward at the end of the trajectory, $R(\tau) = r(x_0)$ (e.g., the calculated bulk modulus), and zero intermediate rewards for all $t > 0$. Because this terminal reward is difficult to backpropagate through both the relaxation and generation processes, we cannot optimize $\mathcal{J}(\theta)$ via standard backpropagation through the entire pipeline, necessitating our reinforcement learning approach.

Algorithm 1 Actor-Critic DDPO with LoRA for MatterGen

Require: Base MatterGen W_0 (frozen), LoRA weights θ_{LoRA}, Critic weights ϕ, PPO epochs K, clip parameter ϵ, learning rates $\alpha_{actor}, \alpha_{critic}$

repeat
  $\theta_{old} \leftarrow \theta_{LoRA}$
  Sample trajectories $\tau = \{x_T, x_{T-1}, \ldots, x_0\}$ from $\pi_{\theta_{old}}(x_{t-1}|x_t, t)$
  for each step $t \in \{T, \ldots, 1\}$ in sampled trajectories **do**
    if $t > 1$ **then**
      $y_t = \hat{V}_\phi^\pi(x_{t-1}, t-1)$
      $\hat{A}_t^\pi = \hat{V}_\phi^\pi(x_{t-1}, t-1) - \hat{V}_\phi^\pi(x_t, t)$
    else
      $y_1 = R(x_0)$
      $\hat{A}_1^\pi = R(x_0) - \hat{V}_\phi^\pi(x_1, 1)$
    end if
  end for
  for $k = 1$ to K **do**
    Let $\rho_t = \frac{\pi_{\theta_{LoRA}}(x_{t-1}|x_t,t)}{\pi_{\theta_{old}}(x_{t-1}|x_t,t)}$

$$L_{actor}(\theta_{LoRA}) = -\frac{1}{T}\sum_{t=1}^{T}\min\left(\rho_t\hat{A}_t^\pi, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\hat{A}_t^\pi\right)$$

$$L_{critic}(\phi) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{V}_\phi^\pi(x_t, t) - y_t\right)^2$$

    Update LoRA Actor: $\theta_{LoRA} \leftarrow \theta_{LoRA} - \alpha_{actor}\nabla_{\theta_{LoRA}} L_{actor}(\theta_{LoRA})$
    Update Critic: $\phi \leftarrow \phi - \alpha_{critic}\nabla_\phi L_{critic}(\phi)$
  end for
until convergence

3.2 Actor-Critic Policy Optimization

To optimize the policy π_θ, we adapt Denoising Diffusion Policy Optimization (DDPO) (Black et al., 2023) into an Actor-Critic framework (Konda & Tsitsiklis, 1999). Moving beyond standard policy optimization is crucial for materials generation for two primary reasons. First, standard policy gradients (Williams, 1992) suffer from high variance, especially in this context which has long trajectories and extremely sparse terminal rewards. Second, they often fail to determine which specific intermediate denoising steps contributed to a high-reward final structure. By introducing a parameterized critic $V_\phi(x_t, t)$ that provides a prediction of the end reward, we establish a learned baseline that provides step-by-step evaluations of intermediate noisy states, greatly improving sample efficiency and reducing training variance/risk of catastrophic forgetting.

Since the reward is purely terminal (all intermediate rewards are zero), the return for every step in a trajectory is simply $r(x_0)$. The advantage (Mnih et al., 2016) reduces to $A_t = r(x_0) - V_\phi(x_t, t)$, which is equivalent to generalized advantage estimation (GAE) (Schulman et al., 2016) with $\gamma = \lambda =$

1. The advantage function A_t centers the learning signal to mean 0, reducing variance by evaluating whether a specific denoising action yielded a better trajectory than expected by the critic. A challenge in optimizing materials generation is the hybrid continuous-discrete action space. Recalling that our state x_t is the intermediate noisy material and our action a_t is the targeted denoising step, we decouple the surrogate objective into separate clipped losses for each action type. The actor loss is:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\mathcal{L}_{\epsilon_c}\big(\rho_t^{\text{cont}}, A_t\big) + \mathcal{L}_{\epsilon_d}\big(\rho_t^{\text{disc}}, A_t\big) \right]$$

where $\mathcal{L}_\epsilon(\rho, A) = -\min(\rho A,\ \text{clip}(\rho,\ 1{-}\epsilon,\ 1{+}\epsilon)\, A)$, and

$$\rho_t^{\text{cont}} = \frac{\pi_\theta(a_t^{\text{cont}} \mid x_t)}{\pi_{\theta_{\text{old}}}(a_t^{\text{cont}} \mid x_t)}, \qquad \rho_t^{\text{disc}} = \frac{\pi_\theta(a_t^{\text{disc}} \mid x_t)}{\pi_{\theta_{\text{old}}}(a_t^{\text{disc}} \mid x_t)}$$

are the separate probability ratios for spatial and species actions respectively. We apply a wider clipping threshold ($\epsilon_c = 0.2$) for the continuous spatial actions to allow sufficient geometric relaxation, while imposing a stricter threshold ($\epsilon_d = 0.1$) on the discrete atomic species classification to prevent forgetting of chemical valences. The full training objective also includes a KL divergence anchor $\beta \cdot D_{\text{KL}}(\pi_{\text{ref}} \| \pi_\theta)$ on the discrete policy and an entropy bonus to encourage exploration.

3.3 Base Architecture and Low-Rank Adaptation

Our foundation model is MatterGen (Zeni et al., 2025), a score-based diffusion model designed for crystalline materials. The model jointly denoises continuous geometric variables (fractional atomic coordinates $\mathbf{X}$ and lattice parameters $\mathbf{L}$) and discrete chemical variables (atom types $\mathbf{A}$). Performing full-parameter reinforcement learning on large-scale diffusion models is unstable and prone to policy collapse. To maintain the strong physical priors learned during pre-training while allowing for target-property exploration, we freeze the pre-trained diffusion backbone and apply Low-Rank Adaptation (LoRA) only to the final output projection heads of the GemNet denoiser. For a frozen weight matrix W, the updated parameters are defined as $W' = W + \frac{\alpha}{r} BA$, where A and B are trainable low-rank matrices, r is the intrinsic rank (set to 16), and $\alpha = 16$. This produces 16 adapted layers and approximately 133K trainable parameters.

3.4 Reward Formulation

The terminal reward $r(x_0)$ evaluates the scalar property value (e.g. bulk modulus, band gap, etc) of the generated crystal. Computing this property via physics-based simulation such as DFT is too computationally intensive for efficient, high-throughput training for the RL model. We resolve this by designing a reward system consisting of a physics-based reward model and a surrogate critic.

We take bulk modulus as a sample application. For high-fidelity ground truth evaluation, we utilize a MatterSim interatomic force field (Yang et al., 2024a). We apply volumetric strains to the generated lattice and perform internal coordinate relaxations to obtain an energy-volume curve. The bulk modulus B is analytically extracted via a quadratic fit to the Birch-Murnaghan equation of state (Birch, 1947):

$$B = V_0 \frac{\partial^2 E}{\partial V^2}$$

where V_0 is the equilibrium volume. To provide fast feedback during the updates, we train a time-conditioned surrogate critic that can predict the bulk modulus of a crystal in an intermediate denoising state. This critic utilizes a frozen GemNet backbone (Gasteiger et al., 2024) with LoRA adapters (Hu et al., 2021) and a MLP head on top. The critic is time-conditioned, taking the noisy material state x_t and timestep t to output a predicted bulk modulus μ. This allows the RL agent to approximate the expected terminal reward (and therefore advantage) directly from intermediate diffusion steps.

3.5 Implementation Details

We initialize the actor policy π_θ from the pre-trained MatterGen base model with injected LoRA adapters. The value function V_ϕ is initialized from a time-conditioned bulk modulus predictor built on MatterGen's pre-trained GemNet score network which is trained prior to RL fine-tuning to predict the relaxed bulk modulus from noisy intermediate states (see Appendix B).

During the policy optimization phase, we update the networks using 3 PPO epochs per rollout batch. To reduce computational overhead, we also implement a 20% strided timestep subsampling mechanism during these trajectory rollouts. We train using a batch size of 64, employing the Adam optimizer with an actor learning rate of 1×10^{-4} and a critic learning rate of 3×10^{-4}. The discrete KL penalty coefficient β is fixed at 0.01.

Our primary model is trained on a single A30 GPU until a mean reward plateau of 30 epochs.

4 Experimental Results

All comparative analyses between our fine-tuned model and the baseline generative models are evaluated on the physical properties of the materials after structural relaxation. Each generated structure undergoes variable-cell relaxation using the FIRE (Bitzek et al., 2006) optimizer guided by the MatterSim surrogate force field, converging when the maximum force falls below 0.05 eV/Å. For transparency, we use the official MatterGen evaluation tool to determine stability and novelty of all results generated from both CrystRL and baselines.

All models were trained on the Alex-MP-20 dataset which contains 675,204 entries: 1,632 structures with ML-predicted bulk modulus over 300 GPa, 247 over 350 GPa, and none over 400 GPa, with a dataset maximum of 398.77 GPa. We utilize the official pre-trained MatterGen model equipped with property-conditioning adapter modules. The baseline uses classifier-free guidance fine-tuned on the Alex-MP-20 dataset to steer the denoising process toward high bulk modulus values.

The primary task is to steer generation toward superhard materials, shifting the distribution of generated structures to higher bulk moduli. We measure this shift by generating 512 candidate structures for each method and count those exceeding 320, 350, 370, and 400 GPa; the last of which exceeds every entry in the training data (ML-predicted). As a baseline, we condition MatterGen on a bulk modulus of 320 GPa which is done for two main reasons. (1) As shown in Figure 1, MatterGen has diminishing returns and a narrowing distribution of bulk moduli when conditioned higher than this. (2) Conditioning on higher bulk moduli would cause falling back to known materials (Figure 1), which would penalize the baseline's novelty.

We evaluate generative performance using the strict S.U.N. (Stable, Unique, Novel) metric (Zeni et al., 2025). Thermodynamic stability is assessed by structurally relaxing the generated candidates using the MatterSim (Yang et al., 2024a) interatomic force field and calculating the final energy above the convex hull (E_{hull}) of MatterGen's reference dataset which combines Alexandria (Schmidt et al., 2022) and MP-20 (Xie et al., 2021) structures with a strict threshold of $E_{\text{hull}} < 0.1$ eV/atom. Uniqueness is assessed within each generated batch and novelty is verified against this reference dataset to ensure the models are not memorizing known structures. Finally, the bulk modulus is derived by applying volumetric strains and fitting the resulting energy-volume curve to the Birch-Murnaghan equation of state.

4.1 Budget-Matched Comparison

Model	**Mean**	**Median**	**≥320**	**≥350**	**≥370**	**≥400**
MatterGen (cond. 320 GPa)	328.0	329.5	63.09%	20.12%	6.25%	0.00%
CrystRL (Ours)	343.9	350.4	92.97%	52.54%	14.26%	1.18%

Table 1: Budget-matched comparison of CrystRL and MatterGen conditioned on a bulk modulus of 320 GPa. Each method generates 512 structures and computes measurements after structural relaxation. CrystRL samples are drawn from the final checkpoint of the main training run with LoRA adapters merged into the base weights. Unlike Table 2, which reports cumulative means over training, this table evaluates a single generation pass.

Table 1 compares the bulk moduli of 512 structures generated by each method. MatterGen (conditioned 320 GPa) produces a distribution centered around its target (mean 328.0, median 329.5), as expected of classifier-free guidance. CrystRL's distribution is shifted upward: 92.97% of its structures exceed 320 GPa compared to 63.09% for MatterGen, and CrystRL produces over twice as

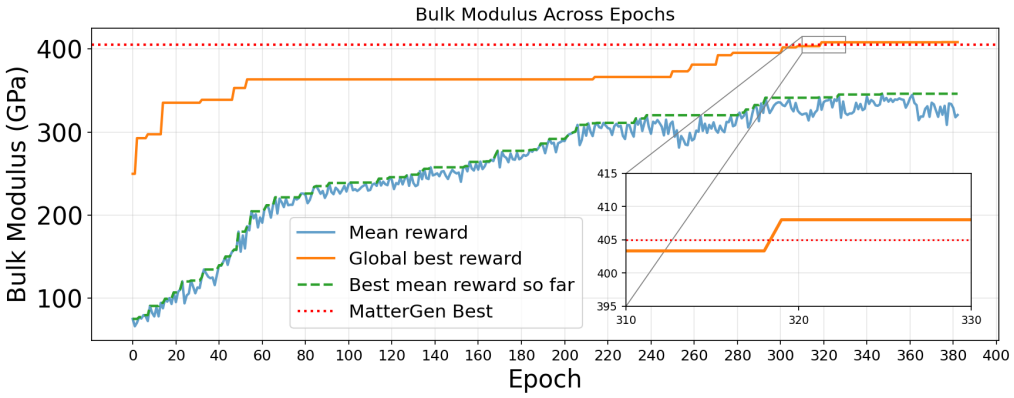


Figure 2: Bulk modulus of generated materials throughout training, showing the mean bulk modulus of each generation batch, the best batch mean so far, and the global best structure. The red dotted line indicates the MatterGen fine-tuned baseline of 405.09 GPa.

many structures above 350 GPa (52.54% vs. 20.12%) and above 370 GPa (14.26% vs. 6.25%). Only CrystRL generates structures beyond 400 GPa (ML-predicted) with 1.18% of its structures (6 of 512), all in the C-Re system. CrystRL's top C-Re candidates from training reach DFT-verified bulk moduli of 386.70-397.03 GPa (see Appendix D).

4.2 Training Dynamics

Whereas Table 1 evaluates a single generation pass from the final checkpoint, we now examine the batches generated during training. Figure 2 tracks the bulk modulus of generated structures over the course of training. Starting from the base MatterGen policy, whose generation batches average about 75 GPa, the mean bulk modulus of each batch rises steadily, reaching a best batch mean of 346.08 GPa. This level is consistent with the mean of 343.9 GPa obtained when sampling from the final checkpoint in Table 1, indicating that the learned policy reliably reproduces the high-modulus distribution reached during training.

The highest bulk modulus structure improves throughout training and reaches an ML-predicted bulk modulus of 408.21 GPa at epoch 375. This exceeds both the best structure generated by conditioned MatterGen (405.09 GPa) and the maximum ML-predicted bulk modulus in Alex-MP-20 (398.77 GPa). With that being said, the 3.12 GPa difference between MatterGen and CrystRL is within the possible error of the surrogate model so this mostly serves to show CrystRL's ability to reach the ≥ 400 GPa range.

4.3 Novelty and Stability

We demonstrate that our framework maintains competitive stability, uniqueness, and novelty throughout training. Across training epochs, our model achieves a cumulative mean of 24.18% for the fraction of generated structures that are simultaneously stable, unique, and novel (Table 2). For individual metrics, CrystRL achieves a stability of 66.61%, a novelty of 55.26%, and a uniqueness of 72.53%. Figure 3 shows how these fractions evolve over training: stability remains close to the baseline and novelty stays above it throughout, while uniqueness declines gradually as the policy concentrates on high-reward regions.

5 Conclusions, Limitations, and Future Work

In summary, we provide a reinforcement learning-based framework, CrystRL, for discovering novel and stable crystal structures with extreme target properties. With a fixed budget of 512 structures, CrystRL shifts the distribution of generated bulk moduli upward relative to MatterGen conditioned at 320 GPa, beyond which MatterGen shows diminishing returns from further increases in the

Model	% Stable	% Unique	% Novel	% S.U.N.
MatterGen (cond. 320 GPa)	64.91%	**87.05%**	48.04%	20.71%
CrystRL (Ours)	**66.61%**	72.53%	**55.26%**	**24.18%**

Table 2: Stability and novelty of generated structures after structural relaxation (cumulative over training). % S.U.N. denotes the fraction of structures that are simultaneously stable, unique, and novel.

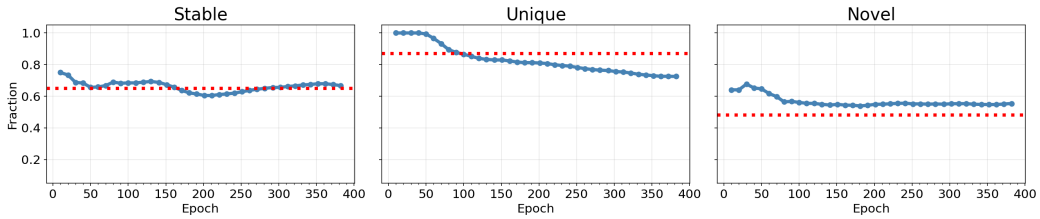


Figure 3: Individual cumulative mean fractions of novel, unique, and stable structures throughout training. The red dotted line shows the fraction achieved by the MatterGen model conditioned on a bulk modulus of at least 320 GPa.

conditioning value (Figure 1), producing over twice as many structures above 350 GPa and above 370 GPa. Its top candidates in the C-Re system reach DFT-verified bulk moduli of up to 397 GPa. Over training, CrystRL also maintains a higher cumulative rate of stable, unique, and novel structures than existing state-of-the-art models. Although our framework successfully steers diffusion models toward extreme properties while maintaining stability and novelty, several limitations still remain that can serve as avenues for future research.

First, our model does not directly generate thermodynamically stable materials in their absolute ground state. Instead, it generates pre-relaxed structures that achieve stability and target properties only after subsequent structural relaxation (e.g., via FIRE optimization guided by a surrogate force field such as Mattersim). Future work could integrate an explicit geometric relaxation penalty directly into the continuous spatial diffusion process to natively generate fully stable, relaxed structures without requiring post-hoc optimization.

Second, our reliance on a surrogate critic introduces potential approximation errors. If the surrogate critic's predictions diverge from the true physical value in highly extrapolated out-of-distribution regions, the policy gradient may be inaccurate. Future research could address this by implementing an online, iterative update mechanism for the critic using DFT rewards in parallel during policy rollouts.

Finally, the current empirical evaluation is focused on maximizing a single scalar property. Real-world materials discovery typically requires balancing multiple objectives such as maximizing bulk modulus while still ensuring a specific band gap and minimizing the energy above the convex hull. Extending our framework to handle multi-objective reinforcement learning will make it more applicable in real-world use cases.

AI USE STATEMENT

In this work, we used generative AI tools for help developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for generating synthetic data sets, formulating mathematical claims, assistance in the writing of proofs, providing critical ingredients for proving mathematical claims, and cleaning and reformatting datasets, and assistance with translation is not applicable to this work. We have reviewed all AI-assisted work. Conceptual work assisted by AI was all manually verified for correctness and is properly cited. No claims were copied verbatim from LLMs. Any LLM-generated code was verified and tested for correctness by at minimum 1

author. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The full training procedure, including the actor-critic framework, PPO hyperparameters, LoRA rank, and optimizer settings, is described in Section 3. The MatterSim-based relaxation procedure and the Birch-Murnaghan fit used to extract bulk modulus, is detailed in Sections 3.4 and 4. Our critic network and its accuracy across noise levels, is analyzed in Appendix B. Full derivations of the probability ratio decomposition for hybrid action spaces and the LoRA adapters are provided in Appendix E. We release the code repository in Appendix A which includes the training and evaluation scripts as well as instructions for reproducing our main results. All baselines used for comparison were generated using the official pre-trained MatterGen checkpoint and its evaluation tool, as described in Section 4.

ETHICS STATEMENT

This work is a methodological contribution to computational materials discovery and does not involve human subjects, crowdsourcing, or sensitive personal data. All datasets used in this work are publicly accessible. We acknowledge that the application of materials generation carries the risk that generated material can be used for harmful application, however, the focus of our model is on a relatively benign property (bulk modulus), and we are not aware of any direct ethical concerns arising from this work.

REFERENCES

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling, 2024. URL https://arxiv.org/abs/2307.04340.

Jacob Austin, Daniel D. Johnson, Jonathan C. Ho, Danny Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Neural Information Processing Systems*, 2021.

Francis Birch. Finite elastic strain of cubic crystals. *Physical Review*, 71(11):809–824, 1947. doi: 10.1103/PhysRev.71.809. URL https://aps.org.

Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97(17):170201, 2006. doi: 10.1103/PhysRevLett.97.170201. URL https://link.aps.org/doi/10.1103/PhysRevLett.97.170201.

Kevin Black, Michael Jänner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2305.13301.

Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj computational materials*, 2020. doi: 10.1038/S41524-020-00352-0.

Linjiajie Fang, Ruoxue Liu, Jing Zhang, Wenjia Wang, and Bing-Yi Jing. Diffusion actor-critic: Formulating constrained policy iteration as diffusion noise regression for offline reinforcement learning, 2025. URL https://arxiv.org/abs/2405.20555.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules, 2024. URL https://arxiv.org/abs/2106.08903.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Philipp Höllmer, Thomas Egg, Maya Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, Mark Transtrum, Richard Hennig, Ellad B. Tadmor, and Stefano Martiniani. Open materials generation with stochastic interpolants. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=gHGrzxFujU.

Matthew K. Horton, Patrick Huck, Ruoxi Yang, Jason M. Munro, Shyam Dwaraknath, Alex M. Ganose, Ryan Kingsbury, Mingjian Wen, Jianxin Shen, Tyler S. Mathis, Aaron D. Kaplan, Karlo Berket, Janosh Riebesell, Janine George, Andrew Rosen, Evan Walter Clark Spotte-Smith, Matthew J. McDermott, Orion A. Cohen, Alexander Dunn, Matthew C. Kuner, Gian-Marco Rignanese, Guido Petretto, David Waroquiers, Sinéad M. Griffin, Jeffrey B. Neaton, D. C. Chrzan, Mark Asta, Geoffroy Hautier, Shreyas Cholia, Gerbrand Ceder, Shyue Ping Ong, Anubhav Jain, and Kristin A. Persson. Accelerated data-driven materials science with the materials project. *Nature Materials*, 2025. doi: 10.1038/S41563-025-02272-0.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *International Conference on Machine Learning*, 2023.

A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023. doi: 10.1038/s41586-023-06735-9. URL https://www.nature.com/articles/s41586-023-06735-9.

Mistal, Alex Hernández-García, Alexandra Volokhova, Alexandre AGM Duval, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Michał Koziarski, and Victor Schmidt. Crystal-GFN: sampling materials with desirable properties and constraints. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL https://openreview.net/forum?id=l167FjdPOv.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016. URL http://arxiv.org/abs/1602.01783.

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. Large-scale machine-learning-assisted exploration of the whole materials space. Materials Cloud Archive 2022.126, 2022.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. 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., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *CoRR*, abs/2110.06197, 2021. URL https://arxiv.org/abs/2110.06197.

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 Zugner, Tianyidan 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. 2024a.

Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation, 2024b. URL https://arxiv.org/abs/2311.09235.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, 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.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models, 2024. URL https://arxiv.org/abs/2401.12244.

Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, He Zhang, Shidi Tang, Hongxia Hao, Peiran Jin, Chi Chen, Frank Noé, Haiguang Liu, and Tie-Yan Liu. Towards predicting equilibrium distributions for molecular systems with deep learning, 2023. URL https://arxiv.org/abs/2306.05445.

A CODEBASE

We provide the full code, pretrained models, and instructions to run experiments through the following anonymized repository: https://anonymous.4open.science/r/cryst-rl

B ANALYSIS OF CRITIC NETWORK

In this section we provide an analysis of the accuracy of the critic network across various timesteps in the overall diffusion trajectory. We evaluate the critic on 100 samples from the MP-20 test set across 16 random seeds, reporting mean MAE and RMSE with two-sigma error bars. At each timestep t, structures are corrupted using the same diffusion corruption process used during training, and the critic predicts the bulk modulus from the resulting noisy state.

The critic achieves its lowest error at small noise levels, with a mean MAE of 11.0 ± 0.7 GPa and RMSE of 15.7 ± 2.9 GPa at $t = 0.02$. Performance degrades gradually as noise levels increase as indicated in Figure 5. At full noise ($t = 1.0$) the MAE reaches 61.9 ± 3.8 GPa. This degradation is expected, as at high noise levels the crystal structure is heavily corrupted. Crucially, the critic visibly starts to provide useful information at noise step ~ 700, showing its utility even at high noise levels.

C S.U.N. FRACTION OVER TRAINING

Figure 6 shows the cumulative fraction of generated structures that are simultaneously stable, unique, and novel throughout training. CrystRL's S.U.N. fraction remains above that of MatterGen's at every point in training.

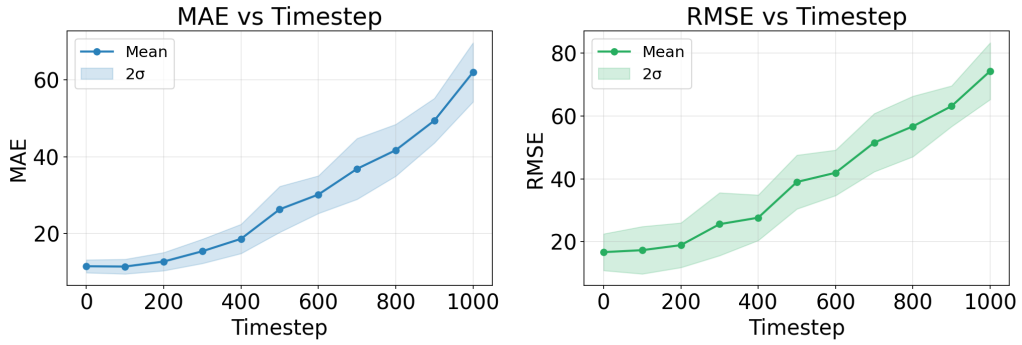


Figure 4: MAE and RMSE for the critic network prediction for noise levels from 0-1000 timesteps with 2-sigma error bars

D Commentary on Generated Materials

Figure 7 shows selected C-Re structures that were generated by CrystRL. We also provide ground-truth density functional theory (DFT) analysis (Kresse & Furthmüller, 1996) for these materials which were performed to independently validate the physical fidelity of the generated structures. All calculations were carried out using VASP with PAW-PBE potentials under consistent computational settings. Across the full set of candidate structures, the total energies per atom are all negative and fall within a relatively narrow range (approximately -10.99 to -11.47 eV/atom), indicating well-relaxed and physically reasonable configurations. The formation energies are consistently negative (approximately -0.95 to -1.31 eV/atom), confirming thermodynamic stability with respect to elemental references. The fitted bulk moduli are uniformly high, spanning roughly 383–397 GPa, with the $R\bar{3}m$ subgroup approaching the upper bound (around 397 GPa). Importantly, these DFT-validated bulk moduli remain comparable to the values predicted by the surrogate model used during training. The consistency between surrogate predictions and DFT evaluations suggests that the learned reward signal is physically meaningful and well-aligned with first-principles calculations. Overall, these results demonstrate that the structures generated by our model are not only energetically stable but also mechanically robust. This provides strong evidence that the proposed generative framework is capable of producing physically valid and high-quality materials.

E Full Training Objective

E.1 Probability Ratio Decomposition for Hybrid Action Spaces

MatterGen jointly denoises continuous geometric variables (fractional coordinates $\mathbf{X}$, lattice parameters $\mathbf{L}$) and discrete atomic species $\mathbf{A}$ via separate score networks. Because these two modalities obey fundamentally different noise processes, the joint policy factorizes as:

$$\pi_\theta(a_t \mid x_t) = \pi_\theta(a_t^{\text{cont}} \mid x_t) \cdot \pi_\theta(a_t^{\text{disc}} \mid x_t), \tag{1}$$

and the log-probability ratio decomposes additively:

$$\log \frac{\pi_\theta(a_t \mid x_t)}{\pi_{\theta_{\text{old}}}(a_t \mid x_t)} = \log \frac{\pi_\theta(a_t^{\text{cont}} \mid x_t)}{\pi_{\theta_{\text{old}}}(a_t^{\text{cont}} \mid x_t)} + \log \frac{\pi_\theta(a_t^{\text{disc}} \mid x_t)}{\pi_{\theta_{\text{old}}}(a_t^{\text{disc}} \mid x_t)}. \tag{2}$$

Rather than clipping this combined ratio — which would conflate geometrically large updates with chemically large updates — we apply separate clipped surrogate losses with independent thresholds $\epsilon_c = 0.2$ and $\epsilon_d = 0.1$:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t\left[\mathcal{L}_{\epsilon_c}(\rho_t^{\text{cont}}, A_t) + \mathcal{L}_{\epsilon_d}(\rho_t^{\text{disc}}, A_t)\right], \quad \mathcal{L}_\epsilon(\rho, A) = -\min(\rho A,\ \text{clip}(\rho, 1-\epsilon, 1+\epsilon)\, A). \tag{3}$$

This decoupling is necessary because continuous coordinate updates require wider policy steps to allow geometric relaxation, while tighter clipping on the discrete species distribution prevents chemical valence collapse.

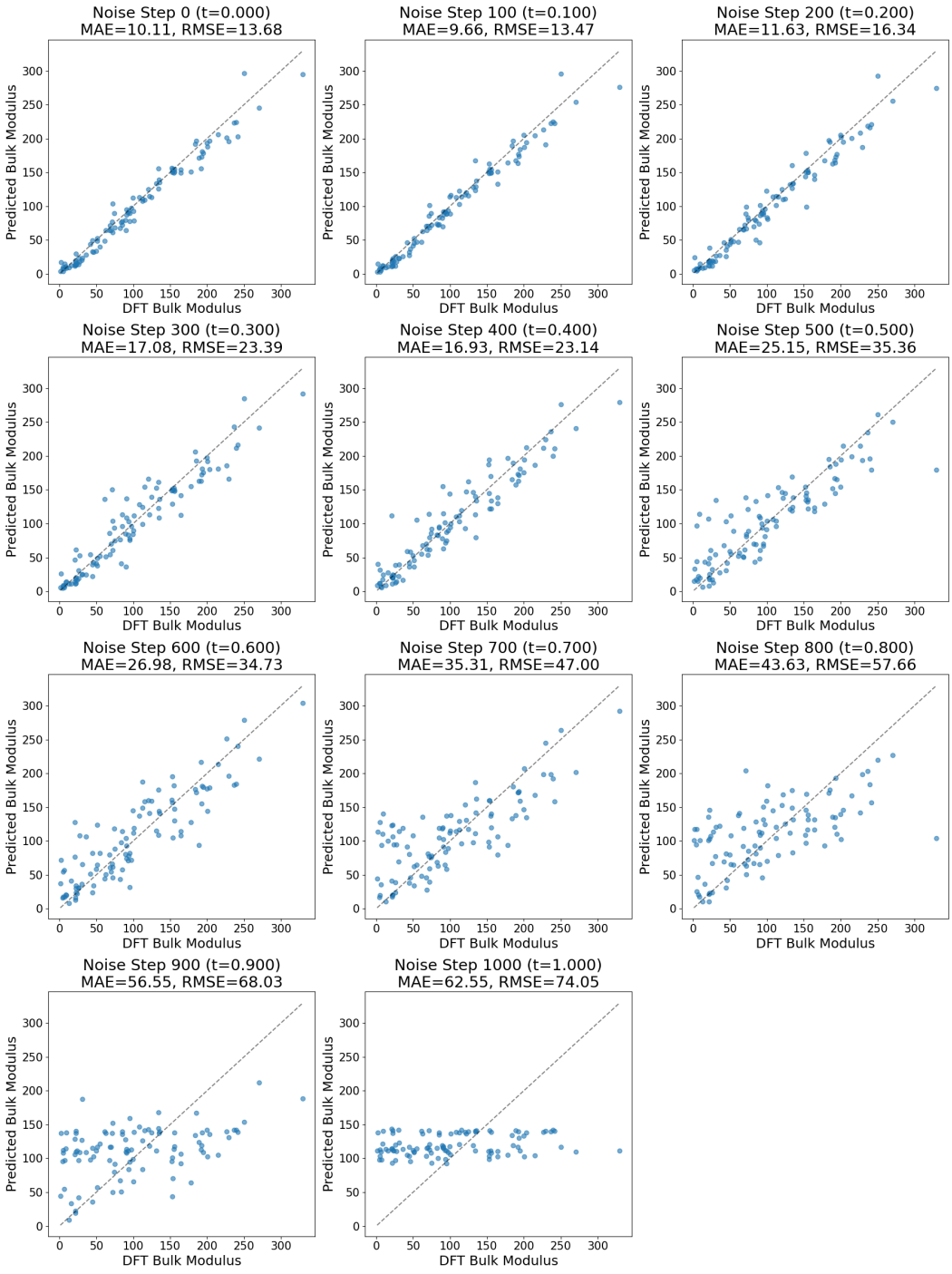


Figure 5: Critic-predicted bulk modulus vs DFT labelled bulk modulus for noise levels from 0-1000 timesteps

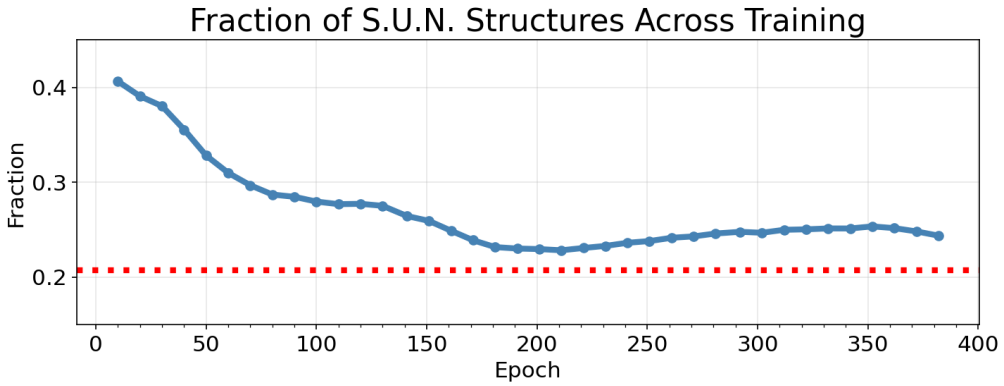
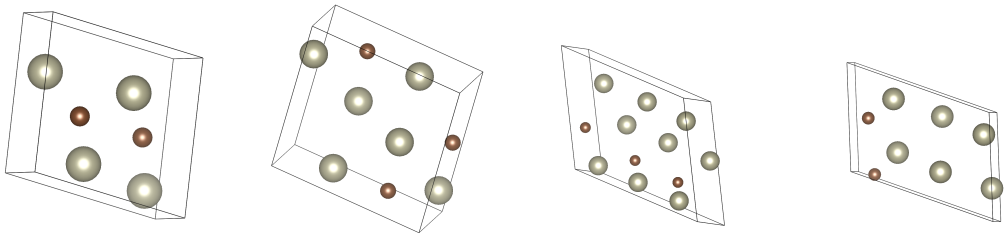


Figure 6: Cumulative mean fraction of S.U.N structures throughout training. The red dotted line shows the fraction achieved by the MatterGen model conditioned on a bulk modulus of at least 320 GPa.



(a) C2Re4 has a machine-learning predicted bulk modulus of 408.02 GPa and a DFT-verified bulk modulus of 397.03

(b) C3Re6 has a machine-learning predicted bulk modulus of 403.51 GPa and a DFT-verified bulk modulus of 394.13

(c) C3Re9 has a machine-learning predicted bulk modulus of 400.49 GPa and a DFT-verified bulk modulus of 386.70 GPa

(d) C2Re6 has a machine-learning predicted bulk modulus of 401.96 GPa and a DFT-verified bulk modulus of 389.36 GPa

Figure 7: Selected generated materials with a ML-predicted bulk modulus over 400 GPa.

E.2 Computing the Continuous Probability Ratio

The continuous reverse process in MatterGen is a Gaussian denoiser. Given the score network prediction $\hat{x}_0 = f_\theta(x_t, t)$, the reverse transition is:

$$\pi_\theta(a_t^{\text{cont}} \mid x_t) = \mathcal{N}\big(a_t^{\text{cont}};\ \mu_\theta(x_t, t),\ \sigma_t^2 I\big)\,, \tag{4}$$

where μ_θ is the DDPM posterior mean derived from $\hat{x}_0$ and σ_t^2 is the fixed posterior variance schedule. Because σ_t^2 is shared between old and new policies, the log-ratio reduces to a squared Mahalanobis distance:

$$\log \rho_t^{\text{cont}} = \frac{1}{2\sigma_t^2}\left(\|a_t^{\text{cont}} - \mu_{\theta_{\text{old}}}\|^2 - \|a_t^{\text{cont}} - \mu_\theta\|^2\right). \tag{5}$$

E.3 Computing the Discrete Probability Ratio

The discrete reverse process operates over categorical distributions over the atomic species vocabulary $\mathcal{A}$. The denoiser outputs a categorical logit vector $\ell_\theta(x_t, t) \in \mathbb{R}^{|\mathcal{A}|}$ per atom site, giving:

$$\pi_\theta(a_t^{\text{disc}} \mid x_t) = \prod_{i=1}^{N} \text{softmax}(\ell_\theta(x_t, t))_{a_t^{(i)}}\,, \tag{6}$$

where $a_t^{(i)}$ is the sampled species for atom i. The log-ratio is then:

$$\log \rho_t^{\text{disc}} = \sum_{i=1}^{N} \left[\log \pi_\theta(a_t^{(i)} \mid x_t) - \log \pi_{\theta_{\text{old}}}(a_t^{(i)} \mid x_t) \right]. \tag{7}$$

The KL anchor $\beta \cdot D_{\text{KL}}(\pi_{\text{ref}} \| \pi_\theta)$ is applied to this discrete head only, penalizing drift from the pre-trained species distribution and preventing mode collapse to a single heavy element.

E.4 Time-Conditioned Critic Architecture

The critic $V_\phi(x_t, t)$ must accurately estimate the expected terminal bulk modulus from a noisy intermediate state x_t at arbitrary diffusion timestep t. We initialize V_ϕ from MatterGen's pre-trained GemNet score network and augment it with:

1. **Timestep embedding:** A sinusoidal embedding of t projected to the model hidden dimension via a two-layer MLP, added to the node feature representations at each GemNet message-passing layer.
2. **Scalar regression head:** A three-layer MLP applied to the graph-level pooled representation, producing a scalar prediction $\hat{\mu} = V_\phi(x_t, t)$.

The critic is trained via mean squared error against the observed terminal reward:

$$\mathcal{L}^{\text{critic}}(\phi) = \hat{\mathbb{E}}_t \left[\left(V_\phi(x_t, t) - r(x_0) \right)^2 \right]. \tag{8}$$

The critic backbone's primary weights are also frozen and adapted via LoRA, preserving message-passing structure while specializing the readout toward bulk modulus prediction at arbitrary noise levels.

E.5 LoRA Injection into the GemNet Score Network

Policy updates are applied exclusively through Low-Rank Adaptation (LoRA) injected into the output projection heads of the GemNet denoiser (`fc_atom`, `out_energy`, and `out_forces`) in each output block, including the lattice output blocks. For a frozen weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the adapted forward pass is:

$$h = Wx + \frac{\alpha}{r} BAx, \tag{9}$$

where $A \in \mathbb{R}^{r \times d_{\text{in}}}$ is initialized from $\mathcal{N}(0, \sigma^2)$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$ is initialized to zero (so adaptation begins from the pre-trained policy), and $r = 16$ is the intrinsic rank, with $\alpha = 16$. Freezing W preserves the crystallographic symmetry priors and thermodynamic stability knowledge encoded during MatterGen's pre-training, while the low-rank perturbation $\frac{\alpha}{r} BA$ steers denoising trajectories toward high-reward regions of composition space.