CRYSTALGYM: A NEW BENCHMARK FOR MATERIALS DISCOVERY USING REINFORCEMENT LEARNING

Prashant Govindarajan[*1,2,4], **Mathieu Reymond**[1,2,5], **Antoine Clavaud**[1,2,4], **Mariano Phielipp**[3]
Santiago Miret[3], **Sarath Chandar**[1,2,4]
[1]Chandar Research Lab, [2]Mila, [3]Intel, [4]Polytechnique Montréal, [5]Université de Montréal

ABSTRACT

In silico design and optimization of new materials primarily relies on high-accuracy atomic simulators that perform density functional theory (DFT) calculations. While recent works showcase the strong potential of machine learning to accelerate the material design process, they mostly consist of generative approaches that do not use direct DFT signals as feedback to improve training and generation mainly due to DFT's high computational cost. To aid the adoption of direct DFT signals in the materials design loop through online reinforcement learning (RL), we propose **CrystalGym**, an open-source RL environment for crystalline material discovery. Using CrystalGym, we benchmark common value- and policy-based reinforcement learning algorithms for designing various crystals conditioned on target properties. Concretely, we optimize for challenging properties like the band gap, bulk modulus, and density, which are directly calculated from DFT in the environment. While none of the algorithms we benchmark solve all CrystalGym tasks, our extensive experiments and ablations show different sample efficiencies and ease of convergence to optimality for different algorithms and environment settings. Our goal is for CrystalGym to serve as a test bed for reinforcement learning researchers and material scientists to address these real-world design problems with practical applications. Furthermore, we introduce a novel class of challenges for *reinforcement learning methods dealing with time-consuming reward signals*, paving the way for future interdisciplinary research for machine learning motivated by real-world applications.

1 INTRODUCTION

Reinforcement learning (RL) methods have demonstrated immense success for complex decision-making problems, robotics (Khan et al., 2020; Xu et al., 2024), autonomous driving systems, and language models (Liu et al., 2023). Recently, the scope of RL has expanded to a variety of scientific areas including energy optimization, quantum systems Martín-Guerrero & Lamata (2021), scientific discovery (Vinuesa et al., 2024), biology, and neuroscience. RL applications in chemistry have been studied for tasks such as molecular design, geometry optimization, and retrosynthesis Sridharan et al. (2024). Yet, RL has been investigated on such applications only on a limited scale because of four main reasons.

First, the diversity of the chemical applications means there is no standardized way of formulating the problem from a RL perspective. Every practioner models their specific problem differently, and proposes solutions tailored towards said problem. It is thus hard to assess if results from one problem apply to another one. Second, next to the required RL expertise, domain expertise is also required to benchmark and evaluate performances on chemistry domains. Both these reasons result in a high barrier of entry for investigating RL-based methods on chemistry applications. Third, the synthetic nature of formulating chemistry applications as sequential decision-making problems, and the immensely diversified nature of the chemical space produce policies that are hard for humans to understand and interpret, especially compared to games and robotics. Fourth, these chemical applications offer unique challenges that have been less studied in the RL literature, such as noisy and time-consuming reward-signals.

[*]Corresponding author: prashant.govindarajan@mila.quebec

Material discovery is one of the applications affected by these challenges. Accelerating material discovery is an important avenue within scientific discovery, the applications of which include designing sustainable and industrially useful materials (Miret et al., 2024). This often involves optimizing for a set of desired properties, computed using physical simulators based on first principles. Density Functional Theory (DFT) (Jones, 2015) is a modeling method that simulates atomic-level properties of molecules and materials using quantum mechanical laws. Considering the time-consuming nature of DFT calculations and the expertise required to operate them, most of the existing generative and language models for material generation do not incorporate them as feedback for material optimization (Gruver et al., 2024; Ding et al., 2024; Levy et al., 2025; AI4Science et al., 2023). However, the predictions made by ML models are very different from DFT computations. For example, the generative models from Gruver et al. (2024), Ding et al. (2024) produce crystals of which 50% are stable according to M3GNet, the state-of-the-art predictor for stability, while only 11% are actually stable according to DFT calculations. Reinforcement learning offers a complementary approach to this problem by directly learning from signals obtained from DFT, potentially without relying on large datasets. Given the dearth of dedicated RL environments and benchmarks for material discovery problems and the domain expertise involved in them, it is difficult for practitioners to investigate RL approaches for these tasks. To this end, we propose **CrystalGym**, a novel and open-source RL environment for crystalline material discovery that offers a way to learn optimal policies from rewards obtained directly from DFT. We focus particularly on optimizing the composition of a crystal by framing it as a sequential decision-making problem and formulating a deterministic Markov Decision Process (MDP), as initially proposed by Govindarajan et al. (2024) – the agent sequentially places a chemical element at a given atomic site in a crystal. We design reward functions based on DFT outputs, individually targeting three challenging properties – band gap, bulk modulus, and density. As the methods and parameters for DFT calculations are preset, they need not be modified unless necessary, hence making it easier for RL practitioners to adopt the environment without explicitly focusing on the correctness of domain-related aspects.

The goal of this environment and benchmark is to design and test RL algorithms for optimizing DFT-based rewards, and to promote future research in this new class of tasks involving optimization of time-consuming reward signals. We provide tasks where the RL agent is expected to explore the exponentially large chemical search space and drive the policy toward designing high-reward crystals. Our work considerably differs from previous works on crystal generation that used generative models (Zeni et al., 2025; Levy et al., 2025; Jiao et al., 2023) without involving DFT in the training loop or active learning works that do not attempt to optimize for functional materials properties(Merchant et al., 2023). The electronic and elastic properties we optimize for have plenty of practical and industrial applications, including efficient semiconductor and battery design, photovoltaics, and hydraulic and aerospace materials. Overall, material discovery directly influences sustainability and climate change mitigation.

Our unique contributions to this work are as follows.

1. Open-source RL environment (http://github.com/chandar-lab/crystal-gym) for crystal discovery based on the `Gymnasium` framework (Towers et al., 2024), that is ready to be adopted and customized by the RL and material science community.
2. Extensive analysis on performance and sample efficiency with different RL algorithms including proximal policy optimization (PPO), soft actor-critic (SAC), Rainbow, and deep Q-networks (DQN) (Schulman et al., 2017; Haarnoja et al., 2018; Hessel et al., 2018) with appropriate graph networks for the policy.
3. We highlight several domain-related challenges in applying RL for material discovery and in general problems that involve time-consuming and noisy reward signals, leading to potentially interesting future directions.

2 Background

2.1 Related Work

Crystalline material generation has gained significant attention in recent years, with generative and language models being more prominent in this space. Diffusion-based models have been proposed to learn a generative distribution from a dataset of crystals. CDVAE (Xie et al., 2022) was one of

the first approaches in this area, which follows an encoder-decoder model with a denoising diffusion process, generating both the structure and composition of crystals. This was followed by models that incorporate symmetric inductive biases, such as DiffCSP (Jiao et al., 2023) and SymmCD Levy et al. (2025). MatterGen (Zeni et al., 2025) also used a diffusion model and performed post-training optimization for properties like band gap, bulk modulus, and magnetic density. Large language models such as Crystal-LLM Antunes et al., Crystal-Text-LLM (Gruver et al., 2024) and MatExpert (Ding et al., 2024) are autoregressive approaches that used text-based representation of crystals in the 3D space. While most of these approaches evaluated the generated samples with DFT, none of them optimized for properties directly computed with DFT or used it as feedback for improving learning. Further, while GNOME Merchant et al. (2023) used an active learning approach for material generation by optimizing for stability with DFT calculations in the loop, it did not focus on other important electronic and mechanical properties. Govindarajan et al. (2024) introduced a new way to optimize crystal composition for properties like formation energy and band gap using a reinforcement learning setup, where offline learning was done to mitigate the time-consuming nature of DFT calculations. In our work, we adopt the MDP framework proposed by Govindarajan et al. (2024) to build an environment and test bed for online RL algorithms. Our work is also loosely related to ChemGymRL (Beeler et al., 2024), the first interactive RL environment focusing on chemical discovery based on a simulated laboratory framework.

2.2 Crystalline Materials

Crystalline materials are everywhere, from the photovoltaic cells of a solar panel to the semiconductors in every chip. They are characterized by a periodic arrangement of atoms in the 3-dimensional space. They are usually described by a lattice, represented by a unit cell with vectors $\mathbf{l_1}, \mathbf{l_2}, \mathbf{l_3} \in \mathbb{R}^3$ of length $a, b, c \in \mathbb{R}$, such that for any atom u at position $\mathbf{x_u}$ in the unit cell, u appears again at every position $\{\mathbf{x_u} + n_1\mathbf{l_1} + n_2\mathbf{l_2} + n_3\mathbf{l_3} \, \forall \, n_1, n_2, n_3 \in \mathbb{Z}\}$ in the lattice. The lattice displays various degrees of symetry encompassed in the space group of the crystal, ranging from 1 to 230, where higher space group means higher level of symetry. While recent works focused on generative and language models for generating a crystal's lattice, atomic positions, and compositions together, we simplify the problem such that the agent only predicts the identities of the atoms given fixed lattice and atom positions. This also aligns with the goal of high-throughput virtual screening (HTVS) (Jain et al., 2011), where atoms are combinatorially substituted in known crystal structures and validated using DFT to design new materials computationally. Hence, we formulate the RL problem with discrete action space and deterministic policies. For simplicity, the scope of this study is limited to only cubic crystals (space groups 200-230) with 4-8 atoms, in which case DFT calculations are faster and certain properties are easier to compute.

2.3 Reinforcement learning

In *reinforcement learning (RL)*, an agent learns to optimize its behaviour by interacting with the environment. Such a setting is modeled as a *Markov decision process (MDP)*, a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \gamma \rangle$, with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probabilities $\mathcal{T}(\boldsymbol{s'}|\boldsymbol{s}, \boldsymbol{a}) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, reward function $R(\boldsymbol{s}, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and discount factor $\gamma \in [0, 1]$. At timestep t, the agent is in state s_t, and selects action a_t using a policy of the form $\pi(a_t \mid s_t)$. Under the policy π, we call $V^\pi(s) = \mathbb{E} \sum_t [\gamma^t r_t \mid \pi, s_t = s]$ the *value*, i.e., the expected sum of discounted rewards (or return). The policy that maximizes the value is said to be the optimal policy $\pi^* = \max_\pi V^\pi$. Closely related to the value is the Q-value, $Q^\pi(s, a) = \mathbb{E} \sum_t [\gamma^t r_t \mid \pi, s_t = s, a_t = a]$.

One of the common approaches to learn Q^* is through the *Bellman equation*, $Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha\delta$, where $\delta = r_t + \gamma \max_a Q(s_{t+1}, a)$ is often referred to as the *temporal-difference target*. Deep Q-networks (DQN) approximate Q with a neural network Q_θ parametrized by θ, by minimizing $(Q_\theta(s_t, a_t) - \delta_{\theta'})^2$, where $Q_{\theta'}$ is a periodically updated copy of Q_θ used to stabilize learning (Mnih et al., 2015). Many recent value-based algorithms still use DQN as their foundation, which is why we use it throughout our experiments to compare the different settings we introduce. Notably, Rainbow (Hessel et al., 2018) integrates the improvements of multiple DQN extensions, such as Dueling DQN (Wang et al., 2016), Double DQN (Van Hasselt et al., 2016), and prioritized experience replay (Schaul et al., 2016) into a single algorithm.

Next to value-based algorithms, we also evaluate alternative approaches, such as actor-critc methods, that learn a policy explicitly. Soft actor-critic (SAC) (Haarnoja et al., 2018) is an off-policy

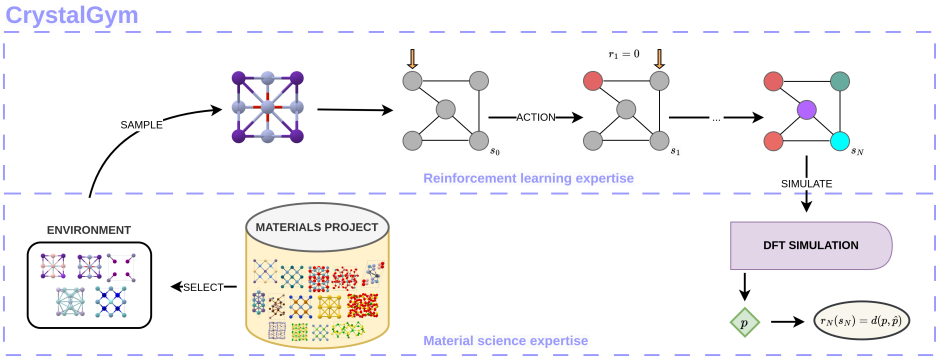

Figure 1: The CrystalGym environment. We select crystals from the Material Project (Jain et al., 2013) database for which DFT calculations can be performed in reasonable time. An episode starts by sampling a crystal structure from this selection. At each timestep, the agent selects an atom to fill a specified position. The episode ends once all positions are filled with atoms, at which point the crystal is evaluated on the DFT simulator. The parameters of the simulator are pre-set, such that they converge in reasonable time for a wide range of compositions. The reward function is computed based on a distance metric with a target value.

method that learns both a policy and a Q-function, optimizing for maximum entropy to encourage exploration. Proximal policy optimization (PPO) (Schulman et al., 2017), on the other hand, is an approximation of trust region policy optimization methods that constrain the size of the policy update – the loss function is based on a clipped surrogate objective.

3 The CrystalGym environment

Our aim is to promote the use of RL for material discovery. Since deep learning methods generate samples in the same distribution as their training data (Levy et al., 2025; Zeni et al., 2025), RL offers an alternative perspective, where it freely explores the chemical space in search for completely new structures. For material science researchers, the ability to explore ouside of known distributions accelerates the discovery process, as it allows for automatic exploration of the exponentially large chemical space, while also directly optimizing for properties obtained from DFT calculations, instead of machine learning proxy models that have proven to be unreliable and insufficiently accurate (Ghugare et al., 2024; Lee et al., 2023; Bihani et al., 2024; Miret et al., 2023). For an RL researcher, CrystalGym allows to focus on challenging aspects for scientific discovery processes, where the transition function is synthetic (meaning that interacting with the environment is a virtually free operation), but the reward function is time-consuming, and noisy or inaccurate. Moreover, much of the required domain expertise is baked in the environment, allowing for RL researchers to focus on algorithmic improvements for scientific discovery.

3.1 Crystal generation as a Markov decision process

In CrystalGym, the agent should optimize the composition of a crystal structure for a desired property value. Starting from an empty structure, the agent iterates over each position, and selects an atom to place in that position. Once each position has been filled, the episode ends and the ensuing crystal is evaluated with a DFT calculator. By training on a pool of different crystals, sampled randomly at the start of each episode, CrystalGym aims to provide a generalizable RL agent, that accurately fills atomic positions even on unseen crystals. However, as we will see in Section 5, this is currently an unattainable goal, as the number of DFT executions required for such a policy to converge requires weeks of consecutive training. As an intermediate step, the RL agent can specialize on a single crystal, by always sampling it at each episode.

Concretely, we adopt the deterministic MDP formulation initially proposed by Govindarajan et al. (2024). We represent crystals using graphs, with atoms as nodes and edges connecting neighboring or bonded atoms. Consider a graph $\mathcal{G}(V, E)$, with nodes (atom positions) $V = \{v_0, \ldots, v_{N-1}\}$ and

edges (connections to other atom positions) E. Each atom position v_i has a label that is either empty ($a_\varnothing$) or set to an atom-type a_i, where i is the index of the i-th element of the periodic table. We consider the state-space $\mathcal{S}$ the empty, partially or fully filled graphs $\mathcal{G}$, with the initial state s_0 the empty crystal $\mathcal{G}_0$, where $v_i = a_\varnothing, \forall i \in \{0..N-1\}$. The action-space $\mathcal{A}$ is defined as the atomic elements a_i of the periodic table. Finally, we transition from state s_t to s_{t+1} by setting v_t to the selected atom a_t. The environment inherits the `Gymnasium` framework and can be easily imported for testing RL algorithms.

Once all positions have been filled, DFT calculations are performed to evaluate properties of interest. These evaluations are then used to compute the reward, which we detail in Section 3.2. The sequence of steps and parameters that DFT calculation requires are preset for each of the properties of interest (i.e., bulk modulus, density, and band gap). Hence, the user just needs to provide the choice of property and the desired target value, without modifying the internal DFT workflow. The overall MDP is illustrated in Figure 1.

3.2 Crystal Properties and Rewards

We focus on individually optimizing the composition of one or more crystals for three different crystal properties – band gap, bulk modulus, and density, each used for various industrial applications, ranging from aerospace engineering to semiconductor design. The aforementioned applications require the different properties to have specific target values (as opposed to being maximized, as is typically the case for RL reward functions). Thus, for each property, we design a reward function based on the magnitude and range of the values the property can have, that encourages to be closer to a target value, $\hat{p}$. We also incorporate a penalty in the reward function if the DFT computation fails due to technical or convergence issues. We use Quantum Espresso (Giannozzi et al., 2009), an open-source software suite for DFT calculations.

Bulk Modulus The bulk modulus is an elastic property of a solid-state material that measures its resistance to change in volume due to bulk compression. This property is useful in many applications involving aerospace engineering and structural design. One of the popular units for the bulk modulus is Gigapascals (GPa), and the magnitude can theoretically be a value between 0 and infinity. We compute the bulk modulus by performing multiple DFT simulations after introducing small volume changes to the original crystal. We then fit a Murnaghan equation of state (Murnaghan, 1944) with the obtained energy values and corresponding volumes. Since we are mostly interested in values between 100 GPa and 1000 GPa, we choose a scaled linear function based on the absolute distance of the computed value p_{BM} from the target, $\hat{p}_{BM}$.

$$r(\boldsymbol{s_N}) = \begin{cases} -5 & \text{if DFT fails} \\ \max\left(-\frac{|p_{BM}-\hat{p}_{BM}|}{\hat{p}_{BM}}, -5\right) & \text{otherwise} \end{cases} \tag{1}$$

Density We calculate the volumetric mass density of a crystalline material, which is the amount of mass per unit volume (g/cm^3). We use the density value obtained using a single-point DFT calculation. As per the Materials Project Database, the density values range from 0 to 28 g/cm^3. Hence, we use an exponential distance function.

$$r(\boldsymbol{s_N}) = \begin{cases} -1 & \text{if DFT fails} \\ \exp(-\frac{(p_D-\hat{p}_D)^2}{\hat{p}_D}) & \text{otherwise} \end{cases} \tag{2}$$

Band Gap Band gap refers to the energy gap between the valence and conduction bands in solids, and the values of interest are usually in the semiconductor range, i.e., 0 eV to 5 eV, given its applications in electronics. Given a desired target value $\hat{p}_{BG}$ and computed value p_{BG}, we choose an exponential reward formulation.

$$r(\boldsymbol{s_N}) = \begin{cases} -1 & \text{if DFT fails} \\ \exp(-(p_{BG}-\hat{p}_{BG})^2) & \text{otherwise} \end{cases} \tag{3}$$

4 THE CRYSTALGYM BENCHMARK

In Section 3, we have presented how we can frame crystal composition as a sequential decision process, and multiple properties of interest to optimize on. Ideally, an RL agent trained on certain crystal structures and optimized for a specific target property should be able to effectively predict the atomic identities for any relevant crystal structure, such that the resulting (filled) crystal's property value matches the target. This, however, is an extremely hard task, not only due to the diversity of the potential crystal structures and chemical space, but also due to the potentially prohibitive computation time required to execute the many DFT calculations encountered during training.

To make measurable progress on this problem, we pair our proposed CrystalGym environment with an associated benchmark. We select 7 different cubic crystals from the Materials Project database (Jain et al., 2013). These crystals have a different number of atoms (4-8), and belong to 5 different space groups, which ensures the genericity of the learning algorithm. We envision multiple degrees of increasing difficulty, spread across 3 different axes of the design problem. First, agents need to be able to optimize crystals for both in-distribution and out-of-distribution property values. For the 3 properties of interest (bulk modulus, density and band gap), we thus specify *in-d. targets* and *o.o.d. targets* (we specify their concrete values in Appendix C.2). Second, agents should not only learn the optimal composition for the crystal structure they have been trained on, but also for novel, unseen crystals. Thus, we devise a *single structure* setting, aimed to assess the feasibility of the desired target property, and a more complex *mixed structure* setting – where the policy is trained on 5 crystals, and evaluated on the 2 remaining ones – to measure how generalizable policies are. Third, agents should be able to freely select any atom from the periodic table, regardless of how unlikely it is to result in an optimal crystal composition. However, in practice, this results in a stark increase of failure rates of DFT calculations. To alleviate this and, consequently, speed up the learning process, we select subsets of atoms that are less prone to failure. The *small action-space* consists of 18 elements of the periodic table, primarily metals and some nonmetals of group 1 and 2, and no transition elements. The more flexible *medium action-space* contains 30 elements, and is a superset of the small action-space with additionally certain metalloids and frequently occuring transition elements, according to the Materials Project database. The full list of selected atoms is available in Appendix C.1.

We believe that, by progressing on the ***mixed structure* with *o.o.d. targets* and *small action-space* setting**, we will also make progress towards the overall goal: designing RL algorithms for material discovery, that can reliably fill any crystal structure for a desired property value. This setting, for which we share initial findings in Section 5.4, strikes a balance between the complexity of the tackled problem and the feasibility of the training procedure in terms of walltime. Notwithstanding, one could use simpler settings that focus on a specific difficulty axis while designing new RL algorithms (e.g., *single structure* with *in-d. targets* and *small action-space* in an active learning scenario designed to minimize the number of DFT calculations).

5 BENCHMARK PERFORMANCE AND RESULTS

Having defined the CrystalGym benchmark, we now perform a set of experiments and ablations to better understand its properties and characteristics. We focus on two important aspects. First, our goal is to test the feasibilty of using RL for the crystal composition completion task (Section 5.2). RL has been understudied for material generation, and DFT signals are known to be complex, so it is important to validate that they can be used as a reward signal. Second, we aim to investigate the evolution of the learning ability when the difficulty of the task is increased (Section 5.3). Following a comprehensive analysis on these variations, we finally evaluate RL-based methods on the proposed benchmark setting, providing the current state of RL for crystal generation (Section 5.4).

5.1 EXPERIMENTAL SETUP

For all our experiments, we compare the performance of some of the most popular value- and policy-based RL algorithms, namely proximal policy optimization (PPO) (Schulman et al., 2017), soft actor-critic (SAC) (Haarnoja et al., 2018), Rainbow (Hessel et al., 2018), and deep Q-networks (DQN) (Mnih et al., 2015) agents.

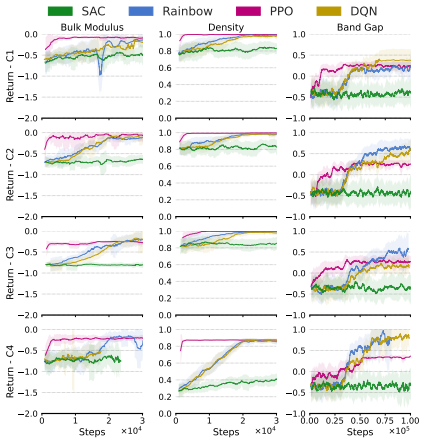

Figure 2: Learning curves for 4 of the 7 crystals structures, on the simplest variation of the CrystalGym benchmark (single structure, in-distribution targets, small action-space).

Additionally, since our crystals are represented as graphs, it is convenient to adopt graph neural networks (GNN) for representation learning (Duval et al., 2023). We leverage MEGNet (Chen et al., 2019), a state-of-the-art GNN architecture for materials. We follow Govindarajan et al. (2024) for creating the graphs and crystal skeletons for the MEGNet architecture. Consequently, the environment can be easily customized to incorporate other graph- and non-graph-based policy networks. For each agent, in each setting, we train 3 different seeds.

5.2 Feasibility of RL-based approaches

To assert that RL-based methods can indeed generate high-quality crystals in terms of desired properties, we select the simplest variation of the different benchmark settings, where the agent trains on the same crystal structure, optimizes for target values that are in-distribution, and uses the small action-space of 18 elements. This allows us to compare the performance of different RL approaches and identify the properties that are difficult to optimize. We also intend to determine if there exists at least one solution, i.e., composition for each of the seven structures (shown in Figure 4) that correspond to a property value close to the desired target. The performance comparison of PPO, Rainbow, DQN and SAC for each structure and all properties is shown in Figure 2 and Figure 6.

PPO In general, PPO quickly finds an optimal or suboptimal solution after a short period of exploration and converges at that point. This is particularly helpful for mechanical properties like density and bulk modulus that have a less complex reward landscape. However, for band gap, considering the large failure rate and the tendency of DFT to produce near-zero values, PPO performs poorly – while it learns to avoid failure states, it converges to a value corresponding to a zero band gap, and does not improve thereafter. While PPO observes high-reward solutions during training, the inherent complexity of the property does not direct the agent toward those useful states.

DQN & Rainbow The exploration in purely value-based methods like DQN and Rainbow follows a $\epsilon-$greedy scheme, unlike PPO. Therefore, the agent starts with a uniform random exploration and gradually exploits the strength of the policy as it learns from more samples from the environment. The samples are temporarily stored in a replay memory during training, which helps the agent process past information in batches and stabilizes learning. For all properties, the learning curve indicates a steady improvement and convergence close to the optimal solution. As expected, band gap, which is the hardest property to optimize, requires additional exploration, resulting in slower learning, and high returns are reached only in some structures. DQN and Rainbow demonstrate similar learning behavior and performance in all cases for bulk modulus and density. However, for band gap, with certain crystals, one of the algorithms performs better in terms of reaching close to optimality – in structure `C1`, DQN performs better, while Rainbow fails to escape the failure state. The opposite true for structures `C2` and `C4`. Hence, at least in this set of experiments, we do not observe the additional benefit of Rainbow having Dueling and Double DQN and multi-step updates.

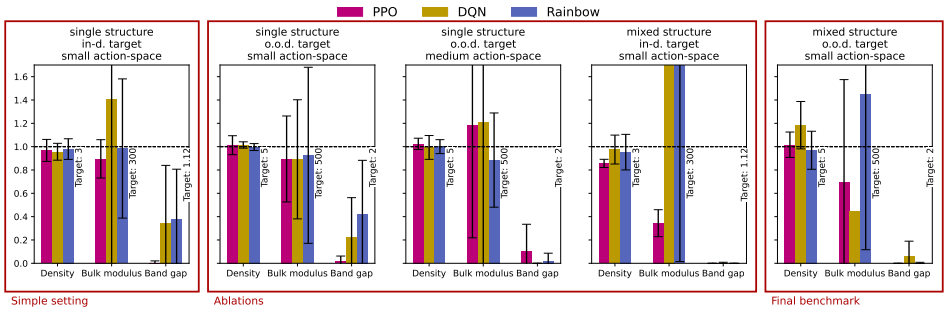

Figure 3: Final performance of each algorithm on each property, for each experiment. After training, each agent is evaluated on 5 trajectories. We report the average achieved property as well as the best performing algorithms for each setting.

SAC In the case of SAC, we notice that none of the experiments for any structure or property indicated a positive learning curve. The agent is unable to eventually escape the exploration phase. Further investigation is required to determine the exact cause of the learning issue with SAC, despite using the same hyperparameters as DQN and Rainbow for the value-based components, i.e., buffer size and target network update frequency.

5.3 Increasing the difficulty of individual settings

We now analyse the impact of increasing the difficulty of each of the 3 axes of the design problem: using out-of-distribution targets, a larger action-space, or optimizing on mixed crystal structures. We summarize all results in Figure 3, while all learning curves are available in Appendix D. Notice the high standard deviation for bulk modulus and band gap results, due to multiple DFT failures or noisy results, as well as significant differences in crystal characteristics.

5.3.1 Harder Target Values

With harder target values, we notice the same behavioral trends in terms of the comparison of different algorithms. The plots for structure `C1` is shown in Figure 7. While it is equally easy to reach the harder target values in the case of bulk modulus and density in most cases, achieving a band bap of close to 2 eV is shown to be much more difficult than 1.12 eV. In `C1`, only Rainbow has managed to show a favorable learning curve, but does not reach close to optimality even after 2×10^5 training steps. As mentioned in Section 3.2, DFT is known to systematically underestimate the band gap energy, which makes it more likely to output lower band gap values (Lejaeghere et al., 2016). As seen in the plots, it is extremely rare that the agent explores the higher band gap regions. Hence, amidst the high failure rate of DFT calculations, i.e., negative reward, and frequent occurrence of near-zero band gap states, the agent fails to learn in a sample efficient way from the very few high-reward states it encounters. Therefore, choosing target values in the rarer regions in the property distribution adds additional complexity to the learning algorithm.

5.3.2 Larger Action Space

Following the previous analysis, we aim to see if increasing the action space by including more frequently seen elements and transition metals like Iron (Fe) and Cobalt (Co) drive the agent towards different and diverse solutions, where the focus is again on harder target values. However, this also increases the complexity of the problem and makes exploration harder particularly due to the higher chance of DFT failures – the presence of transition metals and heavier elements is likely to cause convergence or charge-related issues in DFT calculations. As seen in the results (Figure 3, middle plot and Figure 8), high returns are easily reached in the case of bulk modulus and density with PPO, DQN, and Rainbow. Band gap computation experiences a significantly higher number of failures, thereby making it harder for the reward to even cross 0.0 for all algorithms. Density optimization again appears to be the easiest of the three tasks. In Figure 13, we show examples of policy-generated crystals (structure `C1`) for hard targets when trained with both small and medium action spaces.

5.3.3 TRAINING ON MIXED CRYSTAL STRUCTURES

In the next set of experiments, we increase the difficulty of the task, where the goal is to optimize the properties of 5 crystal structures together. As shown in the results in Figure 9, we notice that the algorithms do not reach close to the optimal solution as quickly as in the previous experiment, where the same crystal structure was sampled in every episode. *PPO's exploration and learning strategy seems to remain the same*, but the returns indicate that it reaches a suboptimal policy for all properties besides density. On the other hand, Rainbow and DQN converge to a higher return, indicating that these value-based methods encourage more exploration and learn better even in this difficult task. In the case of band gap, Rainbow and DQN appear to gradually reach high returns, indicating the possibility of reaching optimality with further training despite the complexity of the property and the task of optimizing multiple structures. Finally, similar to the results for the easier tasks in Section 5.2, SAC again demonstrates the poorest learning performance with all three properties.

5.4 RESULTS ON THE FINAL BENCHMARK

After analysing the different components and settings of CrystalGym, we train all 4 RL algorithms on our proposed CrystalGym benchmark, which uses the mixed structure, o.o.d targets, and small action-space. First, we notice that, just like for the simpler settings, PPO, DQN and Rainbow can generate crystals with desired density values. The density property serves thus more as a sanity check, as it is relatively simple to achieve, and for which DFT computations are fast. However, the algorithms perform poorly on bulk modulus and band gap. Only a few seeds, on a few crystals result in non-zero DFT computations, with many of the DFT calculations resulting in failure. This shows the complexity of optimizing crystals for accurate properties, and makes a stark contrast with optimization through ML property predictors.

6 OVERALL ANALYSIS

We show that varying the RL algorithm, property of interest, and task complexity provides an insightful set of analyses and multiple avenues for future directions. The nature of the calculation differs depending on the property. While bulk modulus and density are mechanical properties and are directly influenced by the atomic weights, the former requires multiple single-point calculations for different volume perturbations. These single-point calculations focus on total energy estimation, and can also provide the estimate of the mass density. Moreover, the distribution of these properties suggests a good range of values and even a randomly sampled composition could result in a value within this distribution. Band gap, which is an electronic property, follows a different set of methods in the single-point calculations that resolve the electronic structure of the crystal and estimate the energies of the highest occupied and lowest unoccupied electronic states. For these types of calculations, DFT is known to have significant underestimation issues and can result in inaccurate estimates. As seen in the results, DFT is highly likely to output near-zero values. The unconventional nature of this property makes it harder for RL algorithms to effectively reach a better solution. While higher-order methods exist for more accurate estimation of the band gap values, they are far more time-consuming than regular DFT computations.

7 CONCLUSION

This research aims to take a step in the direction of accelerating the material discovery process, for which performing atomic simulations is inevitable. We show that crystal design is an interesting and useful set of problems dealing with reinforcement learning with expensive reward signals obtained from expensive atomic simulations. Our new environment is modular and allows the addition of different levels of complexities in the tasks, including the choice of the DFT calculator. From an RL perspective, this environment and benchmark boosts research in the direction of learning with expensive and noisy rewards (Wang et al., 2020), and could influence other domains in scientific discovery and beyond. An important limitation of this analysis is not taking into account the diversity of generated materials. Classical reinforcement learning aims to maximize the expected return and converge to a single behavior. Further investigation of entropy-based RL methods and GFlowNets (Bengio et al., 2021) on these environments is a promising future direction.

REFERENCES

Mila AI4Science, Alex Hernandez-Garcia, Alexandre Duval, Alexandra Volokhova, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Yasmine Benabed, Michał Koziarski, and Victor Schmidt. Crystal-gfn: sampling crystals with desirable properties and constraints. *arXiv preprint arXiv:2310.04925*, 2023.

LM Antunes, KT Butler, and R Grau-Crespo. Crystal structure generation with autoregressive large language modeling (2023). *URL https://arxiv. org/abs/2307.04340*, 2307.

Chris Beeler, Sriram Ganapathi Subramanian, Kyle Sprague, Mark Baula, Nouha Chatti, Amanuel Dawit, Xinkai Li, Nicholas Paquin, Mitchell Shahen, Zihan Yang, et al. Chemgymrl: A customizable interactive framework for reinforcement learning for digital chemistry. *Digital Discovery*, 3 (4):742–758, 2024.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27381–27394. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e614f646836aaed9f89ce58e837e2310-Paper.pdf.

Vaibhav Bihani, Sajid Mannan, Utkarsh Pratiush, Tao Du, Zhimin Chen, Santiago Miret, Matthieu Micoulaut, Morten M Smedskjaer, Sayan Ranu, and NM Anoop Krishnan. Egraffbench: evaluation of equivariant graph neural network force fields for atomistic simulations. *Digital Discovery*, 3(4):759–768, 2024.

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

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

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

Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

Raj Ghugare, Santiago Miret, Adriana Hugessen, Mariano Phielipp, and Glen Berseth. Searching for high-value molecules using reinforcement learning and transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nqlymMx42E.

Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter*, 21(39):395502, 2009.

Prashant Govindarajan, Santiago Miret, Jarrid Rector-Brooks, Mariano Phielipp, Janarthanan Rajendran, and Sarath Chandar. Learning conditional policies for crystal design using offline reinforcement learning. *Digital Discovery*, 3:769–785, 2024. doi: 10.1039/D4DD00024B. URL http://dx.doi.org/10.1039/D4DD00024B.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vN9fpfqoP1.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Anubhav Jain, Geoffroy Hautier, Charles J Moore, Shyue Ping Ong, Christopher C Fischer, Tim Mueller, Kristin A Persson, and Gerbrand Ceder. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science*, 50(8):2295–2310, 2011.

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

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

R. O. Jones. Density functional theory: Its origins, rise to prominence, and future. *Reviews of Modern Physics*, 87(3):897–923, August 2015. doi: 10.1103/RevModPhys.87.897.

Md Al-Masrur Khan, Md Rashed Jaowad Khan, Abul Tooshil, Niloy Sikder, MA Parvez Mahmud, Abbas Z Kouzani, and Abdullah-Al Nahid. A systematic review on reinforcement learning-based robotics within the last decade. *IEEE Access*, 8:176598–176623, 2020.

Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv:2309.05934*, 2023.

Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, Peter Blaha, Stefan Blügel, Volker Blum, Damien Caliste, Ivano E Castelli, Stewart J Clark, Andrea Dal Corso, et al. Reproducibility in density functional theory calculations of solids. *Science*, 351(6280):aad3000, 2016.

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

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*, 2023.

José D Martín-Guerrero and Lucas Lamata. Reinforcement learning and physics. *Applied Sciences*, 11(18):8589, 2021.

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, 2023.

Santiago Miret, Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, and Matthew Spellings. The open matsci ML toolkit: A flexible framework for machine learning in materials science. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QBMyDZsPMd.

Santiago Miret, NM Anoop Krishnan, Benjamin Sanchez-Lengeling, Marta Skreta, Vineeth Venugopal, and Jennifer N Wei. Perspective on ai for accelerated materials design at the ai4mat-2023 workshop at neurips 2023. *Digital Discovery*, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Francis Dominic Murnaghan. The compressibility of media under extreme pressures. *Proceedings of the National Academy of Sciences*, 30(9):244–247, 1944.

Hillary Pan, Alex M Ganose, Matthew Horton, Muratahan Aykol, Kristin A Persson, Nils ER Zimmermann, and Anubhav Jain. Benchmarking coordination number prediction algorithms on inorganic crystal structures. *Inorganic chemistry*, 60(3):1590–1603, 2021.

Gianluca Prandini, Antimo Marrazzo, Ivano E Castelli, Nicolas Mounet, and Nicola Marzari. Precision and efficiency in solid-state pseudopotential calculations. npj computational materials, 4 (1): 72, 2018.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2016.

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

Bhuvanesh Sridharan, Animesh Sinha, Jai Bardhan, Rohit Modee, Masahiro Ehara, and U Deva Priyakumar. Deep reinforcement learning in chemistry: A review. *Journal of Computational Chemistry*, 2024.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulao, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016.

Ricardo Vinuesa, Jean Rabault, Hossein Azizpour, Stefan Bauer, Bingni W Brunton, Arne Elofsson, Elias Jarlebring, Hedvig Kjellstrom, Stefano Markidis, David Marlevi, et al. Opportunities for machine learning in scientific discovery. *arXiv preprint arXiv:2405.04161*, 2024.

Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6202–6209, 2020.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1995–2003. PMLR, 2016.

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_.

Charles Xu, Qiyang Li, Jianlan Luo, and Sergey Levine. Rldg: Robotic generalist policy distillation via reinforcement learning. *arXiv preprint arXiv:2412.09858*, 2024.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, pp. 1–3, 2025.

A Appendix

B Assumptions

Feasibility of a solution In all the experiments for single structure optimzation, we assume that, given a target property $\hat{p}$, there is at least one composition for the structure that can be achieved with the given action space, and results in a property value p, where $|p - \hat{p}| < \delta$. Here δ is assumed to be practically small, and can vary depending on the structure chosen to be optimized.

Fidelity of reward function The current version of the CrystalGym environment uses Quantum Espresso, a software suite for performing DFT calculations, and the signals obtained from DFT are used to compute the rewards for the RL agent. There are several approaches to improve the accuracy of DFT calculations or use higher-order methods for estimating challenging properties like band gap. However, such approaches are expected to take orders of magnitude more time than the current configuration of DFT we rely on for our experiments, that works for most practical cases. Hence, during policy learning, we assume that the signals obtained from the chosen DFT configuration is functionally the highest fidelity we can observe. Consequently, the reinforcement learning workflow aims to directly optimize for the scalar values obtained from DFT calculations.

Crystal Validity In standard chemical discovery tasks, it is a common practice to report the percentage of valid candidates generated by the model. The criteria for validity for small molecule discovery is usually molecules following appropriate valency rules. For crystals, the structural and compositional validities are generally measured, where the former deals with the closeness of two atomic sites in a crystal unit cell, and the latter checks if the total charge adds to zero. However, we directly rely on the outputs of DFT, which is expected to fail to simulate or converge for theoretically infeasible crystals, or estimate the energy value to be higher.

Structure Relaxation For practical reasons, we do not perform structure relaxation on policy-generated crystals. Although DFT relaxation optimizes the crystal structure to minimize system energy, which benefits downstream applications, it requires multiple single-point DFT calculations per sample, significantly increasing computational complexity. Hence, the backbone structure and lattice of each crystal candidate in the enironment is unchanged during training and evaluation.

C Experimental Details

C.1 Action Space

The CrystalGym environment allows the possiblity of using different action spaces. The scope of this benchmark is limited to action spaces corresponding to two sets of elements from the periodic table. The smallest action space contains 18 elements, which are mostly Group-1 and Group-2 metals and some nonmetals, but no transition elements. This **small** action space simplifies DFT calculations, resulting in lesser number of failures in simulations, but vastly reduces the exploration space compared to the ideal action space (118 elements in the periodic table). The **medium**-sized action space consists of some of the frequently ocurring transition metals, in addition to the elements in the smaller action space. We also propose a **larger** action space that includes rarer transition elements, which we aim to test in the future.

- **Small**: Li, Na, K, Rb, Be, Ca, Mg, Sr, H, C, N, O, P, S, Se, F, Cl, Br
- **Medium**: Li, Na, K, Rb, Be, Ca, Mg, Sr, H, C, N, O, P, S, Se, F, Cl, Br, B, Si, Ge, Fe, Cu, Co, Ni, Mn, Al, Zn, Sn, Cr
- **Large**: Li, Na, K, Rb, Be, Ca, Mg, Sr, H, C, N, O, P, S, Se, F, Cl, Br, B, Si, Ge, Fe, Cu, Co, Ni, Mn, Al, Zn, Sn, Cr, In, Sb, V, Mo, Ga, Ag, Ti, Ba, Y, Te, I, Pd, Rh, As, Pt, Cs, Au, Bi, Zr, La

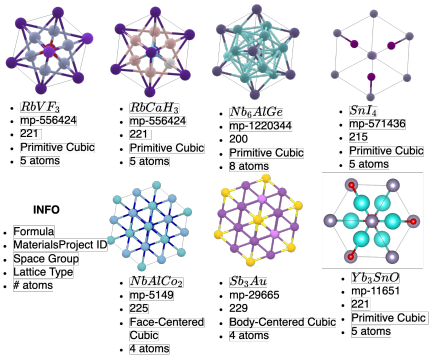

Figure 4: Every possible starting crystal structure considered for our experiments. Each structure has been picked from existing crystals in the Material Project database and their corresponding geometric properties are displayed below their representation. From left to right and top to bottom, the crystals have been refered to as C1, C2, C3, C4, C5, C6 and C7 in the paper.

C.2 Environment Variations

The CrystalGym environment allows testing RL algorithms on a variety of tasks with customizable levels of difficulties. The list of variations supported in the current version of the environment is shown in Table 1.

Table 1: List of all the variations of experimental components. Each experiment is designed to study the impact of specific variations accross different configurations of experimental components.

Experiment	Variations
RL Algorithm	PPO SAC DQN Rainbow
Properties	Density Bulk Modulus Band Gap
Structures	Single Mixed
Mode	Completion Substitution
Policy Net	MEGNet CHGNet
Action Space	18 30

C.3 Target Properties

For our experiments, we use two sets of target values, one being *in-distribution* and other being *out-of-distribution*. The values are listed in Table 3. For each of the three properties, distributions of the values for **all cubic crystals** in the Materials Project database are shown in Figure 5.

Bulk Modulus The bulk modulus distribution shows that the mode falls between 100-150 GPa. Our chosen easy target of 300 GPa is in the rarer regions in the distribution, but there is a reasonable number of crystals that have a bulk modulus of close to this target. However, the target of 500 GPa exists outside this distribution, indicating that it could be a hard value to reach through exploration.

Density The distribution of densities shows that both the chosen target values lie well within the distribution. However, density is directly related to the total mass of the crystal, which is dependent

Table 2: Experimental setup detail. Experiments 1-5 generate crystals from scratch while experiment 6 replaces atoms in fully completed crystals. Each experiment has a unique combination of action space size, target value of the property and size of the pool of starting (empty) crystal structures.

Exp No.	Mode	Target	Action Space
Completion			
1	Single	Easy	Small
2	Single	Hard	Small
3	Single	Hard	Medium
4	Mixed	Easy	Small
5	Mixed	Hard	Small
Substitution (CHGNet)			
6	Mixed	Easy	Small

Table 3: List of the different properties values for the **easy** and **hard** settings.

	Bulk Modulus (GPa)	**Density** (g/cm^3)	**Band Gap** (eV)
Easy	300.0	3.0	1.12
Hard	500.0	5.0	2

on the atomic weights. Hence, it is mostly the choice of the action space that determines how easily the agent can reach higher density values. While in most cases the agents could reach 5 g/cm^3 easily, our separate analysis of PPO's performance with a target of $8g/cm^3$ highlighted failure of the agent to reach densities close to optimality for all crystal structures (Figure 12).

Band Gap The plot of band gap frequency shows a highly skewed distribution, where majority of the crystals have a band gap value close to zero. Both the easy (1.12 eV) and hard (2 eV) targets lie in the rarer regions of the distribution. However, with the type of DFT calculations we perform with Quantum Espresso, it can be noticed that it is extremely rare that the agent experiences states with band gaps higher than 1.5 eV during training. This makes 2 eV much harder as a target than 1.12 eV.

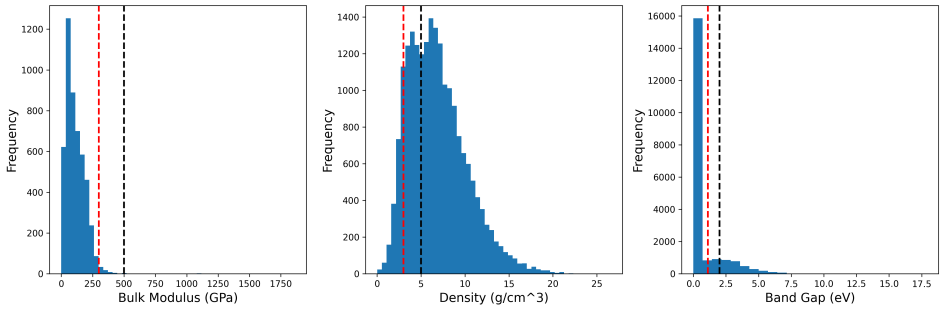

Figure 5: Histograms of the distribution of Bulk Modulus, Density and Band Gap for crystals in the Material Project database. The dashed red line represents the value chosen for the **easy** target and the black for the **hard** one.

C.4 Reward Functions

The reward functions were chosen based on the type and range of the properties of interest. For instance, bulk modulus can take a wide range of values, and since exponential distance would hugely amplify small deviations, we chose to use the absolute distance function – the reward is therefore the negative absolute distance. Since the reward is always negative for bulk modulus, we decided to clip it to a minimum value of -5 to avoid large negative rewards. Table 4 shows further details of the reward functions for each property including the bounds and computation times.

Table 4: Properties of the reward functions used for each property. In addition to the mathematical details of the normalization used we provide some important DFT-specific characteristics.

	Bulk Modulus	**Density**	**Band Gap**
Reward Formulation	Absolute distance	Exponential distance	Exponential distance
Max. Reward	0.0	1.0	1.0
Min. Reward (failure)	-5.0	-1.0	-1.0
Range	[-5,0]	$\{-1\} \cup (0, 1]$	$\{-1\} \cup (0, 1]$
Time (s)	≈ 130	≈ 20	≈ 20
Failure Rate (%)	≤ 0.1	≤ 0.01	≈ 20

C.5 DFT SETTINGS (QUANTUM ESPRESSO)

We performed DFT single-point SCF simulations using Quantum Espresso v7.1 (Giannozzi et al., 2009), which is fully open-source. Solid-state pseudopotentials from SSSP version 1.3.0 (Prandini et al., 2018) were used for the calculations. The settings used are listed below.

1. `calculation`
   - `scf` for band gap and bulk modulus
   - `vc-relax` for density
2. `nstep`: 1
3. `nbnd`: $\left\lceil \left(\left(\sum_i^N Z_i \right) \text{ div } 2 \right) * 1.12 \right\rceil$
4. `ecutwfc`: 50
5. `ecutrho`: 400
6. `occupations`
   - `smearing` for bulk modulus and density
   - `fixed` for band gap
7. `degauss`: 0.001
8. `nspin`: 1
9. `electron_maxstep`: 300
10. `mixing_mode`: plain
11. `mixing_beta`: 0.7
12. `diagonalization`: david
13. `kpoints`: Chosen automatically from Kpoint density.

C.6 GNN DETAILS

In order to extract meaningful representations from crystal structures, we chose to use graph neural networks conditionned on the target property in every algorithm. These representations are then fed to projection layers to compute each algorithm's relevant quantities. DQN and Rainbow use this arcihtecture as their Q-networks and PPO and SAC use it for both their value and policy networks. In each case, we only need to adapt the MLP's output shape.

C.6.1 MEGNET

MEGNet (Chen et al., 2019) is a universal graph machine learning framework for molecules and crystals that provides expressive graph representations through a message passing scheme specifically tailored for crystals and molecules. We used MEGNet as the default GNN architecture in our experiments. It takes as input a graph $\tilde{\mathcal{G}}(\tilde{\mathbf{H}}, \tilde{\mathcal{I}}, \tilde{\mathbf{y}}; \hat{p})$, where $\tilde{\mathbf{H}}$, $\tilde{\mathcal{I}}$, $\tilde{\mathbf{y}}$ and $\hat{p}$ are respectively the embeded node features, the embeded edge features, the embeded graph-level features and the target property the model is conditioned on. The categorical node features $\mathbf{H}$ are defined as the one-hot encoding of the atom type for each node of the graph, with an additional dimension indicating whether

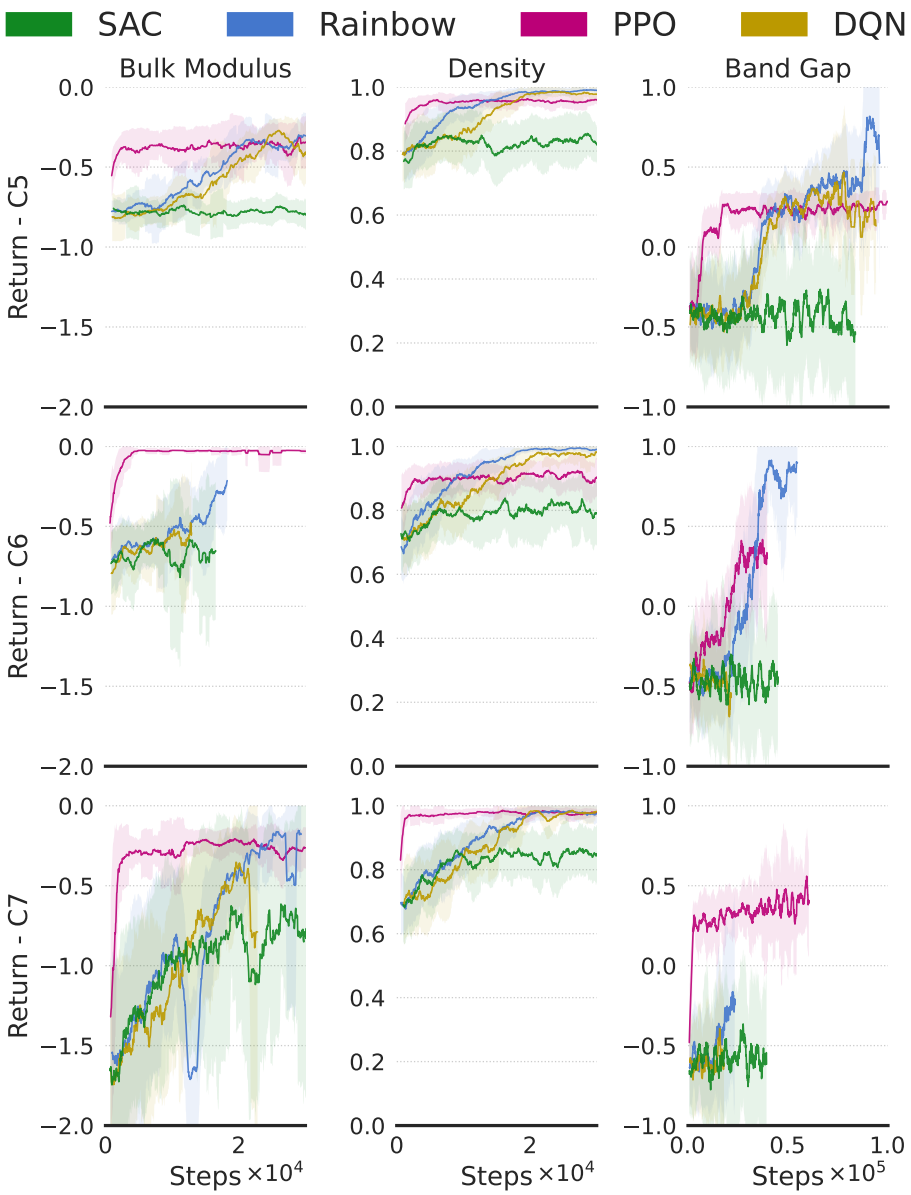

Figure 6: Training curves for crystals C5, C6 and C7 for Experiment 1, with a **single** starting structure, a **small** action space and **easy** targets.

the node is filled with an atom or not. They are then passed through embeding layers to obtain $\tilde{\mathbf{H}}$. Edges connect neighbouring atoms based on the CrystalNN scheme (Pan et al., 2021) for determining their type and presence. We derive the edge features $\mathcal{I}$ as the set $\{t_{uv,(c_1,c_2,c_3)}\}$ of gaussian distances between atoms u in the reference unit cell and v in a unit cell shifted by $c_1\mathbf{l_1}+c_2\mathbf{l_2}+c_3\mathbf{l_3}$.

$$t_{uv,(c_1,c_2,c_3)} = \exp\left[-\frac{d_{uv,(c_1,c_2,c_3)}{}^2}{\rho}\right] \tag{4}$$

$$d_{uv,(c_1,c_2,c_3)} = \sqrt{(\mathbf{x_v}+c_1\mathbf{l_1}+c_2\mathbf{l_2}+c_3\mathbf{l_3}-\mathbf{x_u})^2} \tag{5}$$

where $\mathbf{x_v}, \mathbf{x_v} \in \mathbb{R}^3$ are the 3D coordinates of atoms u and v respectively in the reference unit cell. These edge features are then passed through MLP layers to obtain $\tilde{\mathcal{I}}$. Finally, the graph-level

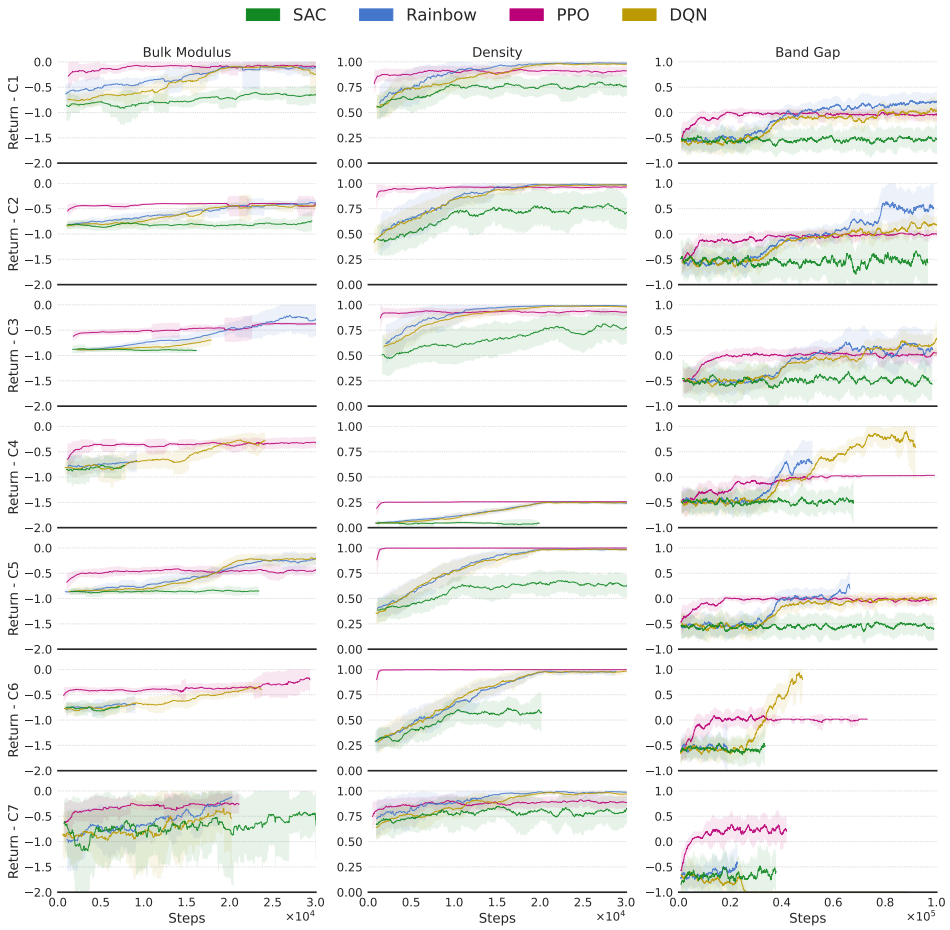

Figure 7: Training curves for all crystals for Experiment 2, with a **single** starting structure, a **small** action space and **hard** targets.

features are defined as $\mathbf{y} = [a, b, c, \phi_1, \phi_2, \phi_3, S, \hat{p}, \tilde{\mathbf{f}}]$ where a, b and c are the lengths of the edges of the lattice ($a = \|\mathbf{l_1}\|$, $b = \|\mathbf{l_2}\|$ and $c = \|\mathbf{l_3}\|$), ϕ_1, ϕ_2 and ϕ_3 are the angles of the lattice, S is the space-group number, $\hat{p}$ is the target property the model is conditionned on and $\tilde{\mathbf{f}}$ is the embeding of the one-hot vector of the categorical feature $\mathbf{f}$, called focus, which instructs the policy which node is to be filled next. $\mathbf{y}$ is then passed through MLP layers to obtain $\tilde{\mathbf{y}}$.

The MEGNet architecture consists of taking as input the graph $\tilde{\mathcal{G}}^{(0)} = \tilde{\mathcal{G}}$ of embeded node, edge and graph-level features and applying K MEGNet layers to it, followed by a readout layer designed to obtain graph-level representations. The Q-values (or values or logits) are obtained by feeding these representations to a MLP layer.

$$\tilde{\mathcal{G}}^{(k+1)} = \text{MEGNet}\left(\tilde{\mathcal{G}}^{(k)}\right) \ \forall\ 0 \leq k \leq K-1 \tag{6}$$

$$\psi\left(\tilde{\mathcal{G}}^{(K)}\right) = \text{Readout}\left(\tilde{\mathcal{G}}^{(K)}\right) \tag{7}$$

$$Q_\theta\left(\mathbf{s} = \mathcal{G}; \hat{p}\right) = \text{MLP}\left(\psi\left(\tilde{\mathcal{G}}^{(K)}\right)\right) \tag{8}$$

C.7 CHGNET

CHGNet (Deng et al., 2023) is a state of the art graph neural network for modeling a universal potential surface. It is a pretrained model designed to provide rich representations for molecules and

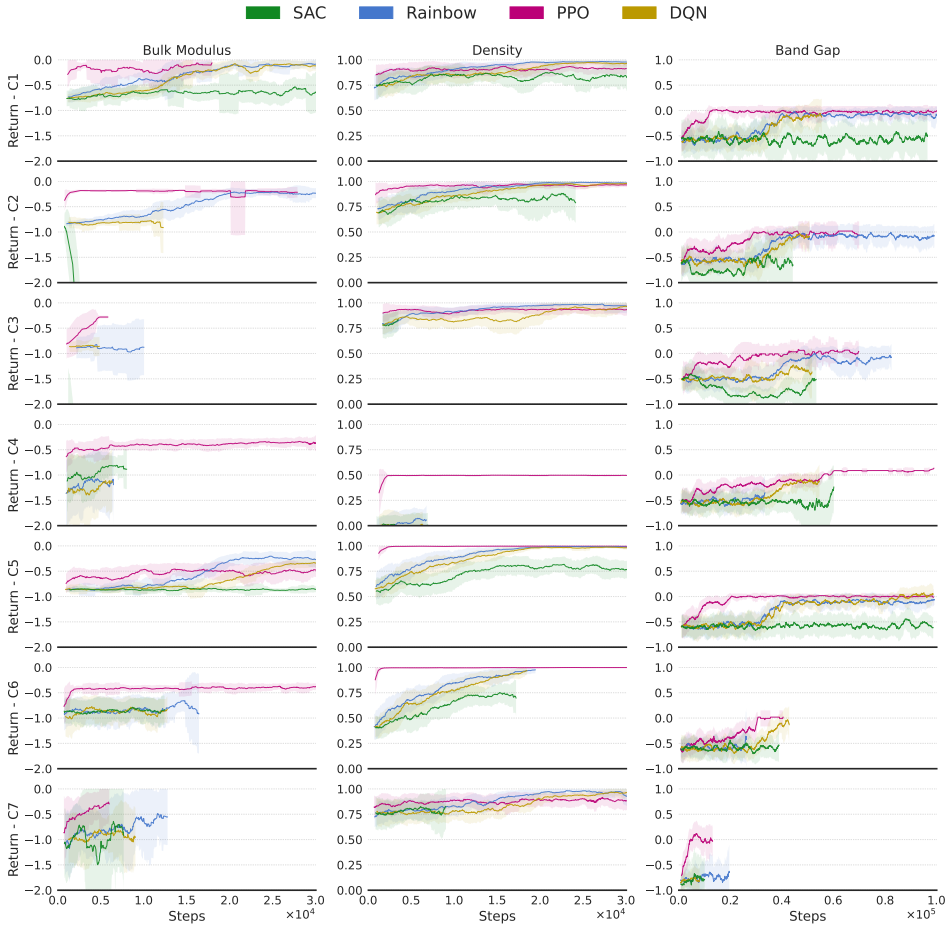

Figure 8: Training curves for all crystals for Experiment 3 with **single** starting structures, a **medium** action space and **hard** targets.

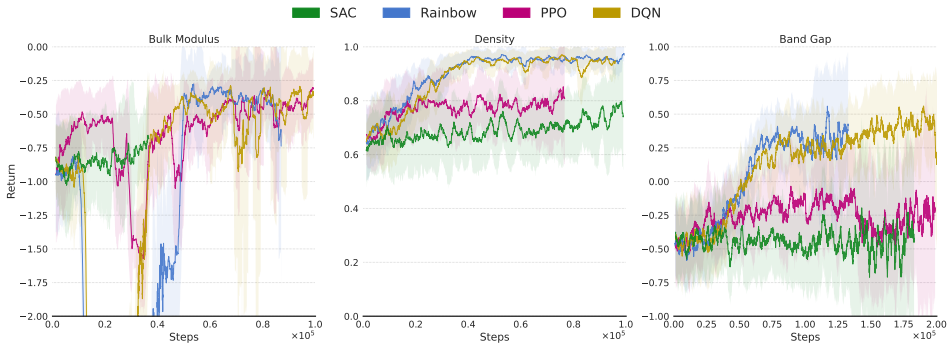

Figure 9: Training curves for Experiment 4, with **mixed** starting structures, a **small** action space and **easy** targets.

crystals. It preserves translation, rotation and permutation invariance of its inputs and has a more complicated process to generate its inputs from the graph of the crystal, namely it considers the graph of atoms and their different bonds as edges, as well as the graph of bonds and their relative angles as edges. Its inputs features include a Fourier representation of the angle information in addition to the regular edge (bonds) and node (atoms) features. The CHGNet layer is applied K times just like for MEGNet, but its message passing function is more complex, allowing for deeper

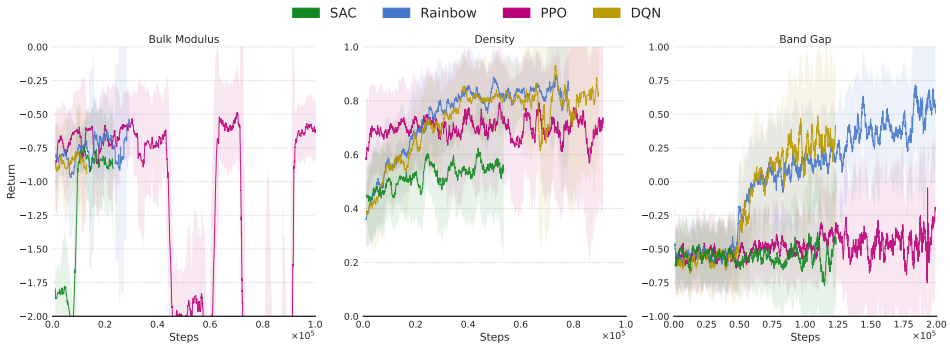

Figure 10: Training curves for Experiment 5, with **mixed** starting structures, a **small** action space and **hard** targets.

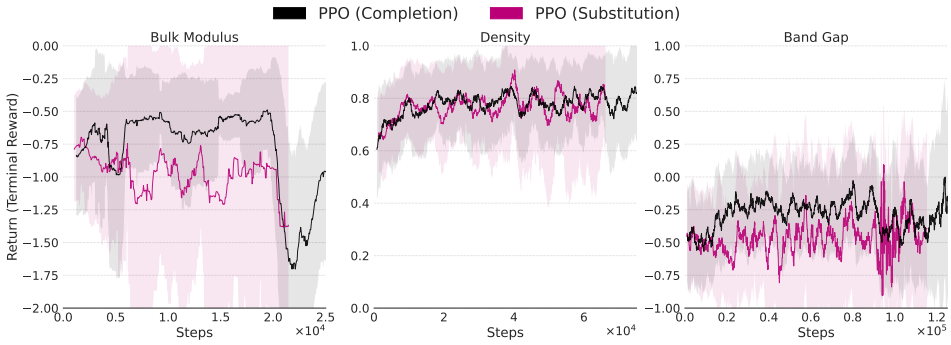

Figure 11: Training curves for Experiment 6. This experiment compares the differences between the completion and substitution approaches on crystal C1 with PPO. The experimental configuration has **mixed** starting structures, a **small** action space and **easy** targets.

interactions between the node, edge and angle informations. We replaced the energy prediction layer by an uninitialized MLP to output Q-values, state values or logits depending on the algorithm used. We froze the weights of the CHGNet as the network is pretrained and provides good representations and only trained the MLP layer we added.

D Additional Results

D.0.1 Substitution

In all the previous experiments, we focused on completing the backbone of a crystal structure, where the initial state does not have any atoms filled, and the intermediate states are partially filled crystals. In this experiment, we intend to determine if using a large pre-trained physics-based graph neural network (GNN) trained on crystalline materials could serve as an effective initial policy. However, with completion, the intermediate states are invalid crystals, and cannot be directly used with these GNNs. We instead focus on substitution, where the agent substitutes an atom in a given atomic site at each step. In such a case, the initial state would be a potentially valid crystal with randomly placed atoms in all the atomic sites. The intermediate states can also be rendered into a valid crystal, making it easier to pass them as inputs to state-of-the-art pre-trained GNNs. These networks are then subsequently fine-tuned with the RL objective. As the scope of this analysis is limited to the policy network architecture, we only investigate the performance of PPO for optimizing each of the properties with a pre-trained CHGNet backbone model as the initial policy. The results indicate no favorable performance with the larger and pre-trained policy. This further, suggests that the complexity of the problem is primarily tied to the nature of the reward signals.

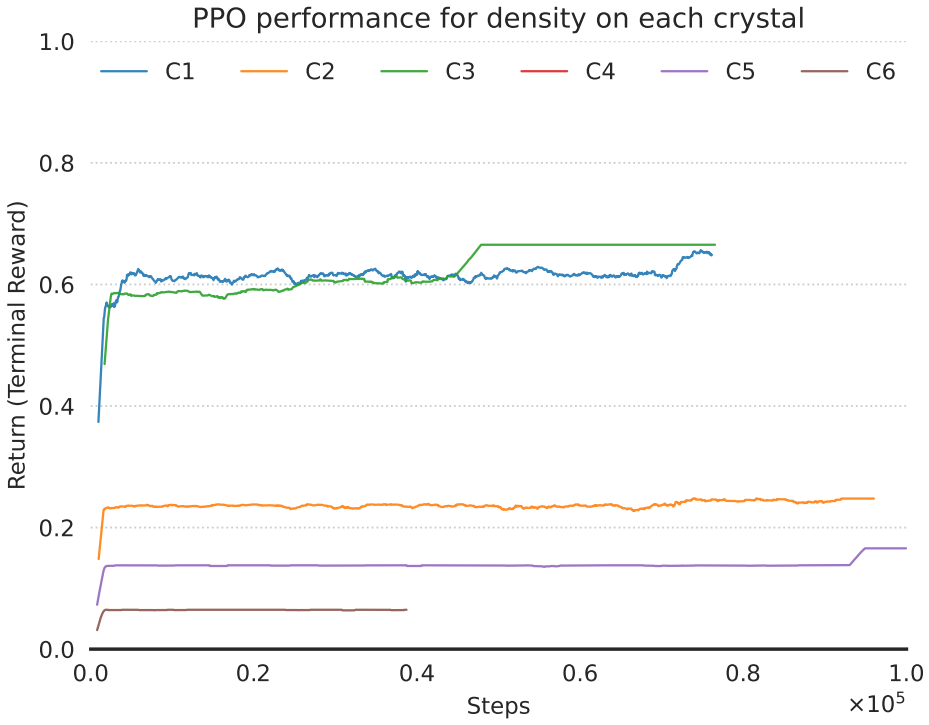

Figure 12: Training curves of PPO for crystals C1, C2, C3, C4, C5, C6, C7. This experiment studies the ability of the agent to generate crystal with a **very hard** target density of $8g/cm^3$, a **single** starting structure and a **small** action space.

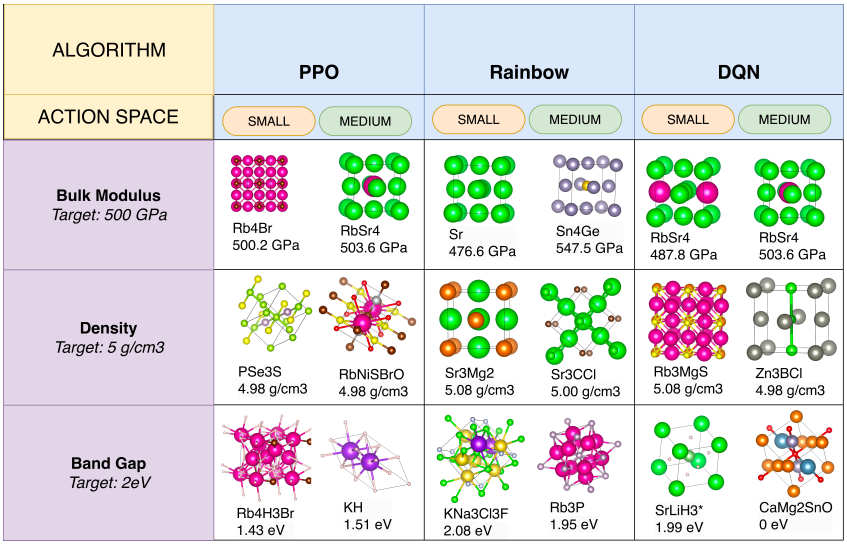

Figure 13: Visualisation of the best performing crystal generated for all algorithms and **hard** targets accross multiple seeds. All algorithms manage to generate good candidates for hard density and bulk modulus targets. * The composition $SrLiH_3$ matches an existing experimentally observed crystal in the Materials Project (mp-24419).