Crystal Design Amidst Noisy DFT Signals: A Reinforcement Learning Approach

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

Abstract

In-silico design of novel materials demands a large number of atom-level calculations for optimizing the desired properties. In practice, it is extremely time-consuming and cumbersome to perform density functional theory calculations at an exponential scale. In the hope of accelerating material discovery, we investigate the feasibility of an active learning-inspired reinforcement learning approach based on online reward model fine-tuning to learn a policy that can generate compositions of crystalline materials optimized for a specific band gap. Through an extensive set of online learning experiments, we show that while RL policies can be effectively trained using machine learning-based proxy reward functions, they fail to converge for DFT-based rewards. This failure of convergence could be related to the inherently noisy nature of DFT in resolving the electronic band structure, which severely affects policy learning. To this end, we emphasize the need for more specialized and domain-driven methods for band gap optimization.

1 Introduction

The material discovery process involves computational validation of properties with atomic simulation frameworks like density functional theory (DFT), which are time-consuming and difficult to perform [1–6]. However, many other means of faster computational evaluation of materials do not fully resemble first-principles calculations, and hence their estimation accuracy might be lower [7–9]. This is particularly true for electronic properties like the band gap, which involves computing the energy of the highest occupied and the lowest unoccupied electronic states [10]. An important aspect of automating the material discovery process with machine learning while still relying on first-principles calculations involves intelligently selecting the appropriate material candidates to perform more costly simulations. Much recent progress in AI-automated material discovery primarily deals with crystalline materials, owing to their well-defined structural properties, abundance in public databases, and practical and industrial applications [11–15]. However, most of them do not directly incorporate DFT simulations in the learning pipeline or condition on desired properties like the band gap [16, 17].

In this work, we adopt reinforcement learning (RL) to sequentially construct crystals aimed for band gap optimization and study the effect of different reward models, including ML-based proxies based on established property estimators [18] and DFT simulation, on the optimal convergence of the RL policy. Inspired by active learning-based approaches [19], we incorporate an uncertainty-based strategy for querying DFT by ensembling reward models, which are fine-tuned from band gaps computed from DFT simulations. Our results demonstrate the difficult nature of considering the band gap, computed by DFT, as a property of interest – it severely slows down learning and does not allow the policy to converge to an optimal solution. We therefore hypothesize that the issue could be relevant to the inaccurate nature of DFT-based band gap estimation[20, 21], including the

*Corresponding author: prashant.govindarajan@mila.quebec

AI for Accelerated Materials Design Workshop at NeurIPS 2024.

underestimation problem. Further domain-driven investigation and additional RL experiments are hence required to confirm this hypothesis and to deduce strategies to mitigate the effect of noisy and potentially inaccurate DFT outputs.

2 Methods

RL Formulation We adapt the general framework by Govindarajan et al. [22] for the RL formulation and environment. It follows an MDP $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}(\boldsymbol{s'}|\boldsymbol{s}, \boldsymbol{a}) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the environment transition probability function, $R(\boldsymbol{s}, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1]$ is the discount factor. The state space consists of empty, partially or filled multigraphs ($\mathcal{G}(V, E)$) of crystal structures. The action space $\mathcal{A}$ consists of atomic elements from which the agent assigns an atom at a given site in a crystal. The action space consists of 21 elements (appendix A.3) in the periodic table that do not include transition metals, lanthanides, actinides, and rare elements whose presence results in inaccurate and slow DFT calculations. Intermediate rewards are zero, and the final reward aims to minimize the distance between the crystal's estimated band gap p and the target band gap $\hat{p}$. It also penalizes DFT failures and crystals with more than 5 atom types, as these are likely to result in unsuccessful DFT computations. Also, 99.9% of crystals in the training dataset consist of at most 5 unique atom types.

$$r(\boldsymbol{s_N}) = \begin{cases} -1 & \text{if more than 5 unique elements (or) DFT fails} \\ \exp(-(p - \hat{p})^2) & \text{otherwise} \end{cases} \tag{1}$$

The RL objective is to learn a policy π_θ to generate optimal crystals (i.e., terminal state $\boldsymbol{s_N}$) with band gap values closer to the target. The estimated band gap p for a given crystal could either be obtained from a computationally cheaper and less accurate ML model or DFT simulation.

Pipeline For the band gap prediction model (referred to as reward model or **MLP-BG** in the figures), we fine-tuned a pre-trained backbone of CHGNet [18] in a supervised manner. To facilitate uncertainty-based querying, we trained 5 of those models on disjoint subsets of the MP-20 dataset (crystals from the Materials Project database containing crystals with at most 20 atoms, used by [16]). We choose a target of $\hat{p} = 1.12$ eV, the band gap of Silicon at room temperature [23]. The components of our pipeline (fig. 1) are 1) policy learning (DQN), 2) MLIP[2]-based structure relaxation, 3) DFT simulation, and 4) reward model fine-tuning. We aim to see if the online agent converges to an optimal solution using an ensemble reward model that is dynamically fine-tuned with DFT outputs. For DFT calculations, we use Quantum Espresso v7.1 [24] with PBE functional [21] and CUDA support. We relax the generated crystal before simulation using CHGNet [18] with the FIRE [25] optimizer. DFT is queried either 1) once in 50 episodes or 2) when the standard deviation (uncertainty) of the band gap predictions from the 5 models is greater than a threshold of 0.2. For policy learning and reward model fine-tuning, the average of the ensemble is considered. This way, all the models in the ensemble get fine-tuned when DFT is queried.

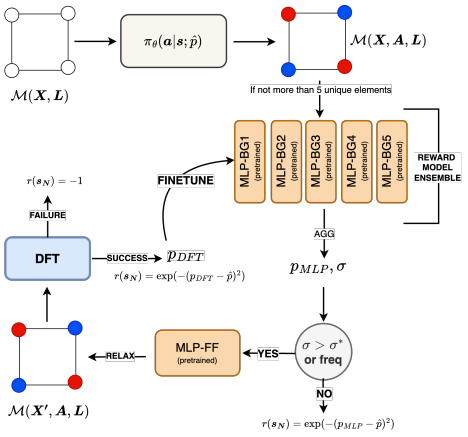


Figure 1: RL pipeline with DFT in the loop for automated material design. The policy generates a composition given a crystal skeleton, which is evaluated by the reward model dynamically trained with DFT outputs.

3 Experiments

We design four online RL experiments – **1)** training DQN to optimize the composition of a single crystal skeleton for the band gap, which is fully based on the ensemble reward model with no DFT

[2]Machine Learning Interatomic Potential

involved, **2)** training DQN with purely DFT-based rewards, by querying DFT after every episode, **3)** training DQN by dynamically fine-tuning MLP-BG (ensemble model), with values obtained from DFT simulations, and **4)** replace DFT with a suitable proxy in experiment 3. For all experiments, the initial policy is a randomly initialized graph neural network [26]. In experiment 4, the proxy is a simple model that explains the relation between the reward model's predictions and the ground truth band gap values. This experiment is useful under the assumption that the proxy approximates DFT's outputs, and to determine if performing real DFT calculations while training could affect the policy learning. We run experiments with 2 different proxies in place of DFT simulation – 1) a linear model that maps the MLP-BG's (which is trained on the entire training data of MP-20) band gap predictions to the ground truth values from the Materials Project, 2) use the table containing MLP-BG and Quantum Espresso (QE) predicted band gaps (including failures) of policy generated crystals from experiment 2 to perform k-nearest neighbor sampling based on MLP-BG's online prediction – this way, we choose the corresponding QE band gap of the sampled nearest neighbor (further details provided in Appendix A.6). Note that the former is only exposed to the stable crystals and their band gaps present in the MP-20 dataset, which might not reflect the signals from Quantum Espresso's calculations with policy-generated crystals (they are not fully relaxed with DFT). With the latter, there were several failure cases (close to 50%) encountered during simulation in experiment 2, which were also included while sampling from the k-nearest neighbors, in which case the reward is -1. This way, we not only have a proxy that is based on policy-generated crystals but also one that mimics QE's noisy behavior. Lastly, all the experiments deal with optimizing the composition of a fixed crystal skeleton consisting of 10 atomic sites, obtained from the Materials Project (`mp-1209282`).

EXPERIMENT	Reward	% DFT/Proxy Calls	% DFT Success	Band Gap
Exp. 1	MLP-BG	0	N/A	N/A
Exp. 2	DFT	100	30.61	0.156
Exp. 3	MLP-BG & DFT	2	38.68	0.241
Exp. 4_{lin}	MLP-BG & Proxy	2	N/A	0.742^{+}
Exp. 4_{knn}	MLP-BG & Proxy	2	N/A	0.396^{+}

Table 1: Insights from online RL experiments – 1) % of DFT calls made, 2) % successful DFT simulations, 3) average DFT-computed (proxy-computed for Exp. 4) band gap of the last 500 successful DFT simulations, and 4) average DFT simulation time. $^{+}$ proxy band gap.

4 Results

We first discuss the experiments without reward fine-tuning, i.e., experiments 1 & 2. In experiment 1, we show that by training a DQN model with a fully MLP-based reward model ensemble, it is possible to reach an optimal policy (fig. 2), thereby the desired band gap of 1.12 eV. In experiment 2, where the rewards are purely based on DFT calculations, the reward increases with more training but eventually converges to a suboptimal policy. For this experiment, we relax the crystal structure with CHGNet before DFT simulation. However, since the computation times of rewards were orders of magnitude higher, training the model for 1 million steps took approximately 9 days. While each episode demands a DFT calculation, many failed (table 1) resulting in a reward of -1. Nevertheless, there were more than 5000 successful DFT calls. If the nature of two reward signals, i.e., DFT and MLP-BG is the same (e.g. process and noise), one would expect both policies to demonstrate a similar learning behavior and sample efficiency. However, the results indicate much slower learning and convergence to near-zero band gap. This is not only because of penalizing DFT failures but also due to the underestimation and inaccuracies of DFT-based band gap calculations.

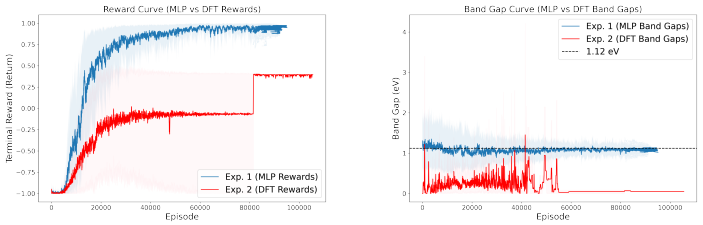
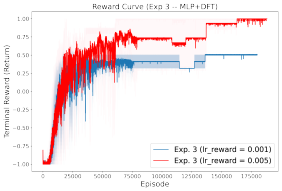
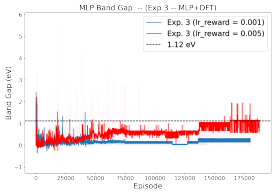


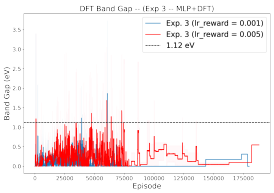
Figure 2: **(a)** Learning curve for experiments 1 (blue) and 2 (red). With a fully MLP-based reward model, the agent learns the optimal policy. With DFT rewards, it converges to a suboptimal policy. (b) Band gap curves for experiments 1 (MLP values) and 2 (DFT values). The former converged to the target, while the latter converged to a near-zero value.



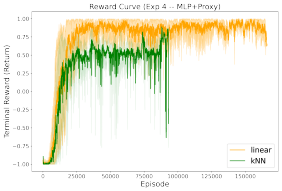
(a) Exp 3. Reward learning curve



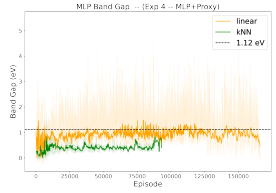
(b) Exp 3. Average of MLP-predicted band gaps of the ensemble.



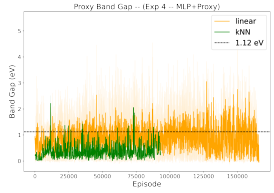
(c) Exp 3. DFT-computed band gaps



(d) Exp 4. Reward learning curve



(e) Exp 4. Average of MLP-predicted band gaps of the ensemble.



(f) Exp 4. Proxy-computed band gaps. NaN values are present in the kNN version.

Figure 3: Results from experiments 3 and 4. Experiment 3 has two models with a different learning rate for fine-tuning the band gap model – 0.001 (less rigorous, orange) and 0.005 (more rigorous, green). Experiment 4 (where we set the fine-tuning learning rate as 0.005) has two cases – 1) linear proxy and 2) kNN proxy.

In experiment 3, the agent gets rewards from both the MLP-BG ensemble and DFT, and the former is fine-tuned after a successful simulation. This is a hard problem because of the additional level of non-stationarity due to the dynamic reward model. From fig. 3a, we see that while the reward improves and the MLP-predicted band gap reaches close to the target (fig. 3b), DFT's band gap values do not seem to converge near the target. Hence, the policy did not align with DFT's preferences (fig. 3c). It has to be investigated further if more training is required for DFT values to converge to the target. Given the issues concerning the incorporation of DFT calculations in the loop, we conducted further experiments to determine if replacing DFT with a simple proxy approximator results in similar issues, in which case we can say that the solution to this problem is likely unrelated to DFT calculations in particular. In the first case where the proxy is a linear model that maps the MLP model's (pre-trained CHGNet originally fine-tuned on all of MP-20 training crystals) outputs to the ground truth band gap values in the dataset. Note that these are not QE-predicted band gaps. As seen in fig. 3f, the proxy band gap reaches the target value on average, and the reward reaches the optimal value (fig. 3d), indicating that it works in this case. In the second case, where the proxy is a value sampled using a kNN approach, the outputs are noisy due to the stochasticity and multiple failures being sampled, and hence struggle to reach optimality. However, in 100k episodes, we do not observe convergence to a lower or near-zero value band gap.

5 Conclusion

Given the impracticality of fully DFT-based online approaches, we emphasize the need for methods that use both machine learning property predictors and DFT. Here, we integrate RL and DFT simulations to address the band gap optimization problem – the agent receives rewards from both a machine learning model and DFT simulations. We highlight important issues in training algorithms from DFT signals due to their noisy and inaccurate nature. This could be mitigated by relaxing with DFT or improving the resolution but at the cost of increased simulation time. Given our deterministic policy learning framework and small-scale experiments, we do not evaluate the diversity and scalability of our approach, which are important in scientific discovery. We expect that our pipeline can be tested with language or generative models as backbones with an appropriate feedback scheme. To conclude, we emphasize the need for domain-driven methods to address property-driven material design.

References

[1] Janice Lan, Aini Palizhati, Muhammed Shuaibi, Brandon M Wood, Brook Wander, Abhishek Das, Matt Uyttendaele, C Lawrence Zitnick, and Zachary W Ulissi. Adsorbml: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Computational Materials*, 9(1):172, 2023.

[2] Alexandre Duval, Victor Schmidt, Santiago Miret, Yoshua Bengio, Alex Hernández-García, and David Rolnick. Phast: Physics-aware, scalable, and task-specific gnns for accelerated catalyst design. *Journal of Machine Learning Research*, 25(106):1–26, 2024. URL `http://jmlr.org/papers/v25/23-0680.html`.

[3] Janosh Riebesell, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*, 2023.

[4] 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`.

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

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

[7] 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.

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

[9] Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. Mattext: Do language models need more than text & scale for materials modeling? In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

[10] Pedro Borlido, Jonathan Schmidt, Ahmad W Huran, Fabien Tran, Miguel AL Marques, and Silvana Botti. Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *npj Computational Materials*, 6(1):1–17, 2020.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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_.

[17] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] 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.

[19] Joseph Musielewicz, Xiaoxiao Wang, Tian Tian, and Zachary Ulissi. Finetuna: fine-tuning accelerated molecular simulations. *Machine Learning: Science and Technology*, 3(3):03LT01, 2022.

[20] John P Perdew. Density functional theory and the band gap problem. *International Journal of Quantum Chemistry*, 28(S19):497–523, 1985.

[21] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[22] 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.

[23] Detlef Klimm. Electronic materials with a wide band gap: recent developments. *IUCrJ*, 1(5): 281–290, 2014.

[24] 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.

[25] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical review letters*, 97(17):170201, 2006.

[26] 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.

[27] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14), apr 2018. doi: 10.1103/physrevlett.120.145301. URL https://doi.org/10.1103%2Fphysrevlett.120.145301.

[28] 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.

[29] Chengteh Lee, Weitao Yang, and Robert G Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical review B*, 37(2):785, 1988.

[30] Ferdi Aryasetiawan and Olle Gunnarsson. The gw method. *Reports on Progress in Physics*, 61 (3):237, 1998.

A Appendix / supplemental material

A.1 Compute

For our experiments, we used NVIDIA A100-SXM4-80GB GPUs model training and performing DFT simulations with Quantum Espresso.

A.2 Graph Representation

We represent crystals in 3 dimensions as multigraphs, following [27]. We follow the exact procedures of Govindarajan et al. [22] for creating the graphs and crystal skeletons. We also use MEGNet architecture for the policy, which is a GNN suitable for materials and molecules [26].

A.3 Experimental Details

In our experiments, we aim to optimize the composition of a fixed skeleton of a crystal that is present in the validation set of MP-20 (ID: mp-1114693). It has 10 atomic sites, and a chemical formula of Rb3ScF6. The space group number is 225, and is hence cubic. The original band gap of the crystal is 6.2281, which makes it and insulator. The action space consists of 21 elements in the periodic table – Li, Na, K, Rb, Be, Ca, Mg, Sr, H, C, N, O, P, S, Se, F, Cl, Br, He, Ne, Kr.

A.4 DQN Hyperparameters

- Q-Net: MEGNeT [26]
- Discount factor: 0.99
- Target update frequency: 1000 (steps)
- Sample Batch size: 64
- ϵ_{start} (initial exploration rate): 1.0
- ϵ_{min} (minimum exploration rate): 0.001
- Decay method: Exponential (rate: 10^{-5})
- Replay buffer size: 200,000

A.5 DFT Settings (Quantum Espresso)

We conducted DFT single-point SCF calculations using the open-source Quantum Espresso v7.1 [24], applying a consistent simulation protocol across all DFT experiments. We utilized solid-state pseudopotentials from SSSP version 1.3.0 [28], employed (3,3,3) k-points, and used the David diagonalization method. Simulations were limited to a maximum of 200 iterations. While our DFT setup for band gap calculations is simpler and faster, it is less accurate compared to methods such as the B3LYP functional [29] and GW [30].

A.6 Band Gap Model (MLP-BG)

We used a pretrained crystal graph neural network (CHGNet), proposed by Deng et al. [18] with initial pre-trained weights for force and energy estimation. We performed supervised learning with the training set of the MP-20 dataset and evaluated it against the validation set.

Ensembling For training the 5 MLP-BG models, we divided the training dataset into five disjoint subsets and trained them individually. In the RL experiments, the mean of these models was considered to be the predicted band gap, and the standard deviation was used for querying. In experiments 3 and 4, all 5 models were fine-tuned after a successful simulation.

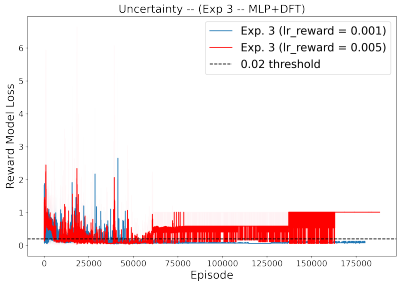
Proxy Model

- **Case 1: Linear Proxy** – This model is obtained by performing a linear regression to map MLP-BG's band gap predictions with the ground truth values present in the validation set of the MP-20 dataset. While querying the linear proxy in experiment 4, we first compute

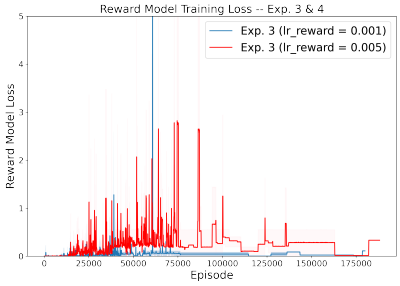
MLP-BG's prediction and use the linear model to map it to the corresponding ground truth estimate.

- **Case 2: kNN Proxy** In this case, we leverage the simulations from experiment 2 (i.e., online RL with purely DFT-based rewards). In a total of 3 seeds of the experiment, we obtained around 30k examples of crystals with band bap values calculated by QE. However, around 70% were failed simulations (denoted by NaN). From this data, we used kNN-based sampling while querying the proxy. In other words, during query time, we first compute MLP-BG's prediction and compute the indices of the k-nearest neighbors from the MLP's band gap estimates in the data obtained from experiment 2. We use those indices to select k corresponding values in the list of Quantum Espresso's estimates, which include NaN values too. From these k entries, we randomly choose one value to represent the proxy version of the true band gap. If the sampled value is NaN, the reward is assigned as -1, and the algorithm continues without reward model fine-tuning.

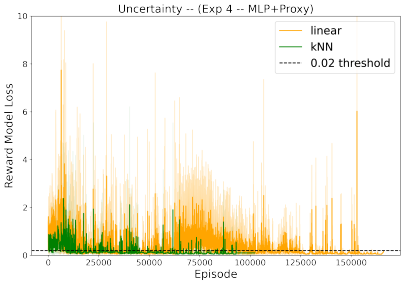
A.7 Additional Results



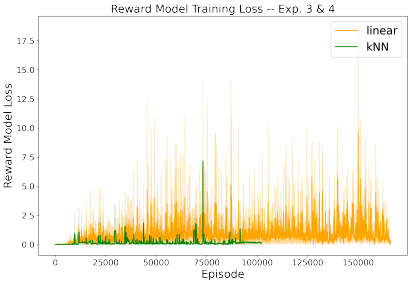
(a) Standard deviation for Exp. 3



(b) Reward model loss for Exp. 3



(c) Standard deviation for Exp. 4



(d) Reward model loss for Exp. 4

Figure 4: Additional results from experiment 3 and 4.