A REINFORCEMENT LEARNING PIPELINE FOR BAND GAP-DIRECTED CRYSTAL GENERATION

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

ABSTRACT

Property-driven AI-automated material discovery presents unique challenges owing to the complex nature of the chemical structural space and computationally expensive simulations. For crystalline solids, the band gap is an important property for designing semiconductors and batteries. However, optimizing crystals for a target band gap is difficult and not well-explored. Reinforcement learning (RL) shows promise towards optimizing crystals, as it can freely explore the chemical space. However, it relies on regular band gap evaluations, which can only be accurately computed through expensive Density Functional Theory (DFT) simulations. In this study, we propose an active learning-inspired pipeline that combines RL and DFT simulations for optimizing crystal compositions given a target band gap. The pipeline includes an RL policy for predicting atom types and a band gap network that is fine-tuned with DFT data. Preliminary results indicate the need for furthering the state-of-the-art to address the inherent challenges of the problem.

1 INTRODUCTION

Discovering new materials with desired properties is a long and cumbersome process even with accurate simulations and sufficient computational resources. The computational materials design process often involves optimization in the exponentially large chemical space amidst complex atomic simulations for computing energies and properties. Recently, there has been a lot of enthusiasm to discover novel organic and inorganic materials for various industrial applications using machine learning (Miret et al., 2024; Duval et al., 2023; Musielewicz et al., 2022). Crystals are of particular interest due to their distinct properties that are useful in modern electronics, optics, photovoltaics, and nanotechnology (Govindarajan et al., 2024). Crystals are characterized by ordered and periodic arrangement of atoms in the 3D space governed by symmetry groups. One of the useful properties of solid-state crystals is the band gap, which is the energy difference between the lowest unoccupied and highest occupied electronic states. It relates to the conductivity of the material, with conductors having zero band gap, insulators having a value greater than 5 eV, and semiconductors in the intermediate range. The band gap is generally estimated using density functional theory (DFT) simulations. However, accurate estimation of the band gap is difficult and time-consuming even for simple crystal systems (Perdew, 1985). Most recent studies that focused on crystal discovery do not optimize for this property.

From a reinforcement learning (RL) perspective, learning a crystal generation policy with a fully online feedback scheme for band gap optimization is implausible considering the high computation times of DFT. Meanwhile, offline learning (i.e., learning on a static dataset of samples and properties) was explored by Govindarajan et al. (2024), which highlighted the advantages and shortcomings of a fully offline approach for energy and band gap optimization. This study aims to extend the work by incorporating DFT simulations in an online RL training pipeline such that the number of DFT calls is reduced. We implement Deep Q Networks (DQN) (Mnih et al., 2015) with a pretrained neural network reward function that acts as a proxy for the band gap output of DFT. Calls to DFT are made at a fixed frequency, allowing fine-tuning of the reward model after every successful DFT simulation. The pipeline also includes structure relaxation by a state-of-the-art

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

machine learning interatomic potential (MLIP) model prior to simulation (Deng et al., 2023). Overall, through simple experiments, we demonstrate the performance of an online RL approach for band gap-conditioned crystal generation and highlight the challenges for future work.

2 METHODS

We adapt the formulation by Govindarajan et al. (2024) for the RL problem 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 fully 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. For simplicity, our experiments are limited to optimizing actions for a single crystal skeleton (appendix A.2 and A.3) with 10 atoms. 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. Hence, $|\mathcal{A}| = 21$, $|\mathcal{S}| = 10^{21}$. For all our experiments, 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 $\hat{p}$. Additionally, it penalizes DFT failures and crystals with more than 5 atom types, as these are unlikely to result in successful DFT computations.

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

The objective of this RL problem is to learn a policy π_θ to generate optimal crystals (i.e., terminal state $\boldsymbol{s_N}$) with high rewards, i.e., band gap value 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 fine-tuned for band gap prediction, or DFT simulation. For the former, we fine-tuned a pre-trained CHGNet model (Deng et al., 2023) by replacing the final layers with a network that predicts the band gap. We used the MP-20 dataset for fine-tuning (a subset of the Materials Project database containing crystals with less than 20 atoms, previously used by Xie et al. (2022)). For all our experiments, we choose a target of $\hat{p} = 1.12$ eV, which is the band gap of Silicon at room temperature (Klimm, 2014). Our pipeline consists of four components: 1) RL policy learning (DQN), 2) structure relaxation using MLIP, 3) DFT simulation, and 4) reward model fine-tuning. In our experiments, we aim to see if the online agent converges to an optimal solution with a parameterized reward model that is dynamically trained in the loop based on DFT outputs. The pipeline is illustrated in fig. 1a. We use Quantum Espresso v7.1 (Giannozzi et al., 2009), an open-source software suite for DFT calculations, with PBE functional (Perdew et al., 1996) and CUDA support. Prior to simulation, we relax the generated crystal using CHGNet (Deng et al., 2023), a state-of-the-art MLIP for crystal energies and forces with the FIRE (Bitzek et al., 2006) optimizer.

3 EXPERIMENTS

To assess the importance of DFT computations in the RL pipeline, we design a set of 4 different experiments. Our first online RL experiment deals with training DQN to optimize the composition of a single crystal skeleton for the band gap, which is fully based on an MLP model with no DFT involved. In our second experiment, we train another online DQN model with purely DFT-based rewards. This experiment involves querying DFT after every episode and is hence extremely slow. Our next experiments fine-tune MLP-BG, the MLP model that predicts the band gap with values obtained from DFT simulations. While training the RL algorithm, we query DFT at a given frequency of once in 30 episodes, and a successful simulation allows the reward model to be fine-tuned in a supervised manner. For the third experiment, the initial policy is a freshly initialized graph neural

EXPERIMENT	Reward	Initial Policy
Exp. 1	MLP	Random
Exp. 2	DFT	Random
Exp. 3	MLP & DFT	Random
Exp. 4	MLP & DFT	Pretrained

Table 1: Online DQN experiments for band gap optimization.

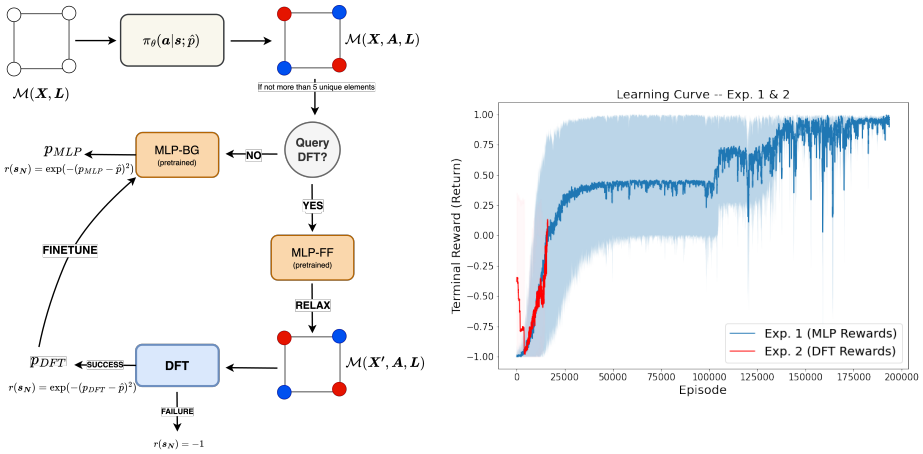


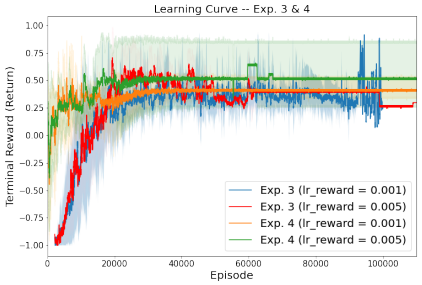
Figure 1: (a) Online RL pipeline with DFT in the loop for automated material design. The policy generates a composition given a crystal skeleton, which is evaluated by a band gap model dynamically trained with DFT outputs. (b) Learning curve for experiments 1 (blue) and 2 (red). With a fully MLP-based reward model, the agent learns the optimal policy given sufficient episodes. The trend is observable in the ongoing experiment 2 (fully DFT-based), but is extremely time-consuming.

network and the exploration scheme follows the first two experiments. For the fourth experiment, we use a pre-trained initial policy obtained from the first experiment. We fine-tune MLP-BG by taking 10 stochastic gradient steps against a mean-squared error (MSE) loss objective on a single sample with a reasonably higher learning rate.

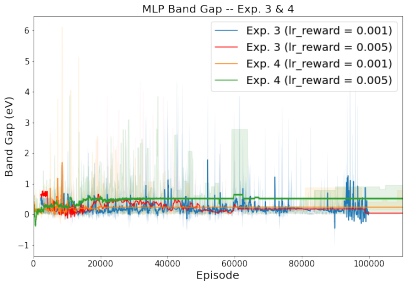
3.1 Results

We first discuss the experiments without reward fine-tuning. In experiment 1, we show that by training a DQN model with a fully MLP-based reward model, it is possible to reach an optimal policy (fig. 1b), thereby the desired band gap of 1.12 eV. Note that we do not explicitly ensure the validity or stability of the generated crystals for this experiment, and they are not relaxed during policy learning. In experiment 2, where the rewards are purely based on DFT calculations, the learning curve shows a similar trend, i.e. increasing rewards with more training. For this experiment, we relax the crystal structure with CHGNet prior to DFT simulation. However, since the computation times of rewards were orders of magnitudes higher, training the model for 1 million steps was impractical. While each episode demands a DFT calculation, many failed (table 2) resulting in a reward of -1. Nevertheless, there were close to 4000 successful DFT calls.

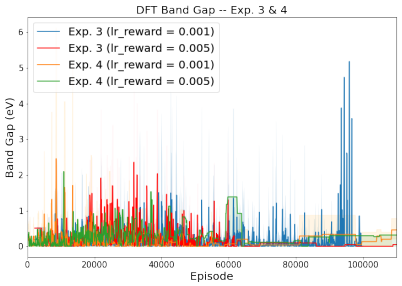
For the third experiment, the agent gets rewards from both MLP-BG and DFT, while the former is fine-tuned with values obtained from the latter. This is a very hard problem for the agent for two main reasons. First, the reward model is dynamic and this might lead to instability. Secondly, the policy can converge to a suboptimal version, leading to the same/similar output crystals that prevent both the improvement of the policy and the reward model, and thereby lead to wastage of DFT computations. Moreover, since most of the band gap values obtained from DFT have lower magnitudes or zero, rigorously fine-tuning the reward model with these samples resulted in the policy producing crystals of lower (near-zero) band gaps, which is an undesired behavior given our target is 1.12 eV. This is evident from the learning curves and band gap plots shown in fig. 2 – with more training, the MLP's band gap converges to less than 0.5 eV and does not improve from there. The band gaps obtained from DFT during training is highly noisy, with only a small fraction of simulations resulting in a band gap of close to 1.12 eV. Similar results are also observed in experiment 4, where we start training with a policy pretrained with MLP-based rewards in experiment 1. This indicates that starting with an MLP-optimal policy does not mitigate the issues and challenges discussed above.



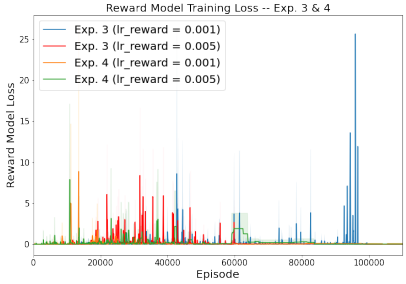
(a) Learning curve for experiments 3 (blue, red), i.e., random initial policy and experiment 4 (orange, green), i.e., pretrained initial policy. They appear to converge to a suboptimal policy.



(b) MLP-predicted band gap plot for experiments 3 and 4. In all models, the band gap appears to converge to a value of less than 1.



(c) DFT-estimated band gap plot for experiments 3 and 4. In all models, the DFT outputs appear to be highly noisy and have lower values.



(d) Fine-tuning loss curve for the band gap model (i.e., which is used in reward calculation). The curves do not show a clear trend of decreasing loss.

Figure 2: Results from experiments 3 and 4. They have 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	**% DFT Calls**	**% DFT Success**	**Band Gap**	**Avg. Simulation Time (s)**
Exp. 1	0	N/A	N/A	N/A
Exp. 2*	100	21.21	0.046	40.54
Exp. $3^{(1)}$	3.33	38.68	0.162	51.02
Exp. $4^{(1)}$	3.33	9.01	0.176	64.77

Table 2: Insights from online RL experiments – 1) % of DFT calls made, 2) % successful DFT simulations, 3) average DFT-computed band gap of the last 100 successful DFT simulations, and 4) average DFT simulation time. The models seldom generate materials with band gaps close to the target, indicating the need to improve the pipeline further (* – ongoing, (1) – lr_reward = 0.001).

4 Discussion and Conclusion

In automated material design, considering the difficulty of using offline RL approaches and the impracticality of fully DFT-based online approaches, we emphasize the need for a middle ground that uses both machine learning property predictors and DFT. In this work, we attempt to address the band gap optimization problem by integrating RL and DFT simulations – the agent receives rewards from both a machine learning model and DFT simulations. We highlight the issues in training a DFT-in-the-loop pipeline that is open for future work. First, we rely on an MLIP for crystal structure relaxation. Relaxing with DFT could significantly reduce the noisy nature of band gap calculations. However, DFT relaxation increases simulation time by manifold and is infeasible for large-scale training. Second, further investigation is required to determine the appropriate way to fine-tune the reward model such that the learning is stabilized and the policy does not converge to a suboptimal

solution. Finally, the pipeline must be scalable in terms of learning from large datasets. We also do not consider diversity as a factor for performance which is important in scientific discovery, because it cannot be characterized with deterministic RL policies. In conclusion, we expect our pipeline to be modular in terms of substituting RL and fine-tuning components with other approaches like language and generative models, and more accurate DFT simulations.

REFERENCES

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

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

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.

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.

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`.

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

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.

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.

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

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

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

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.

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.

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_.

A Appendix

A.1 Compute

We used NVIDIA A100-SXM4-80GB GPUs for training models and performing DFT simulations. For faster data loading, we used 64 CPUs.

A.2 Graph Representation

Following Xie & Grossman (2018), we use multigraphs to represent crystals in 3 dimensions. In multigraphs, a pair of nodes can be connected by more than one edge. In graph $\mathcal{G} = (V, E)$ with nodes (atoms) $V = \{v_0, \cdots, v_{N-1}\}$ and edges, $E = \{e_{uv,(c_1,c_2,c_3)} | 0 \leq u \leq N-1, 0 \leq v \leq N-1, c_1, c_2, c_3 \in \mathbb{Z}, u, v \in V\}$, $e_{uv,(c_1,c_2,c_3)}$ is a edge from atom u to atom v in a unit cell translated by the vector $c_1\boldsymbol{l_1} + c_2\boldsymbol{l_2} + c_3\boldsymbol{l_3}$. Following Govindarajan et al. (2024), we define a crystal skeleton as that with all the structural information (e.g. lattice parameters $\boldsymbol{L}$, coordinates of atomic sites $\boldsymbol{X}$, space group, and graph connectivity) but with hidden/masked atomic elements. This formulation makes it easier for the RL agent to focus on optimizing only the composition in the discrete chemical space, thereby simplifying the search problem. Likewise, we also use MEGNet architecture, which is a GNN suitable for materials and molecules (Chen et al., 2019).

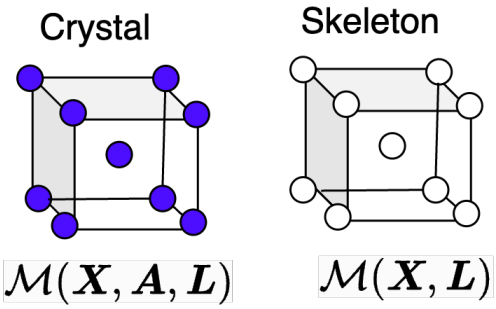


Figure 3: Crystal Skeleton

A.3 Experimental Details

For all our experiments, we attempt to determine the atomic composition of a single crystal skeleton. This skeleton is obtained from an existing crystal in the validation set of MP-20 (ID: mp-1114693). It contains 10 atoms, and has the chemical formula Rb3ScF6. The space group is 225, and is hence cubic. The original band gap is 6.2281, making it an insulator. The action space consists of 21 elements in the vocabulary – 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-Network: MEGNeT (Chen et al., 2019) architecture that predicts Q-values for all actions given a state.
- Discount factor: 0.99
- Target update frequency: 1000 (steps)
- Batch size: 64
- ϵ_{start} (initial exploration rate): 1.0 for Exp 1, 2, and 3. 0.2 for Exp. 4.
- ϵ_{min} (minimum exploration rate): 0.001
- Decay method: Exponential (rate: 10^{-5})
- Replay buffer size: 200,000

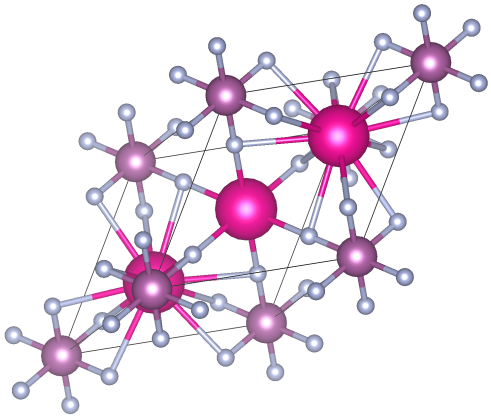


Figure 4: Rb3ScF6 crystal structure.

A.5 DFT SETTINGS (QUANTUM ESPERESSO)

DFT single-point SCF calculations were performed using the open-source Quantum Espresso v7.1 (Giannozzi et al., 2009). Our simulation protocol was uniform for all experiments that involve DFT. We obtained solid-state pseudopotentials (SSSP) version 1.3.0 (Prandini et al., 2018). We used (3,3,3) k-points and David diagonalization method. Simulations were performed for at most 200 iterations. We acknowledge that our DFT setup for calculating band gaps, while being simpler and comparatively faster, is far less accurate than other methods (e.g. B3LYP functional (Lee et al., 1988) and GW (Aryasetiawan & Gunnarsson, 1998)).

A.6 BAND GAP MODEL (MLP-BG)

For training a band gap model, we used a state-of-the-art crystal graph neural network (CHGNet), proposed by Deng et al. (2023) with initial pre-trained weights for force/energy estimation. We performed supervised learning for 1000 epochs with the training set of the MP-20 dataset and evaluated it against the validation set. We also evaluate the model's performance in predicting the band gap values from Quantum Espresso (QE). We notice that the validation loss reaches a stable value fairly soon, indicating overfitting. This is reflected while assessing the scatter plots in fig. 6. Moreover, the band gaps from QE are not always close to those from the MP-20 dataset, and this leads to worse performance when compared with QE band gaps. However, at this point, we do not perform any other experiments to further reduce the generalization error.

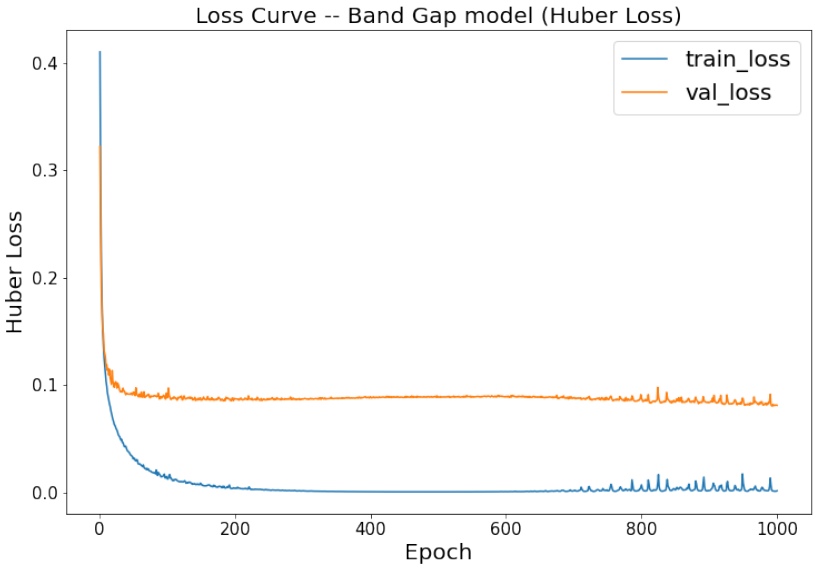
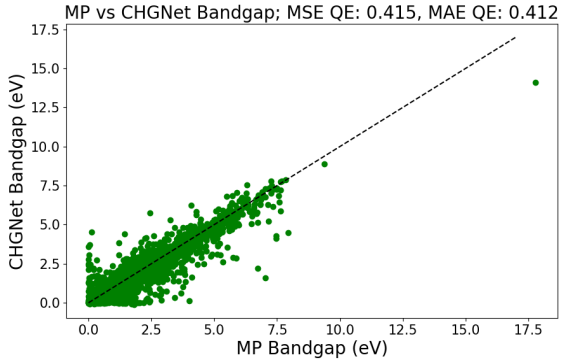
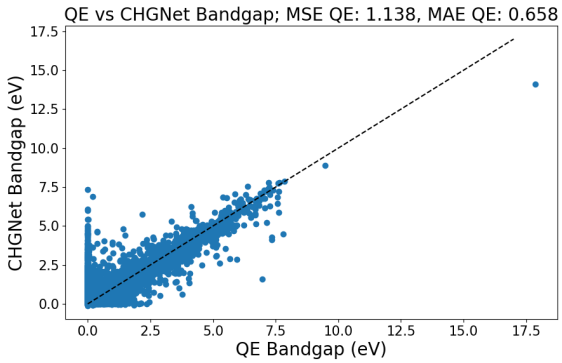


Figure 5: Training curves of band gap prediction model.



(a) % Comparing band gap predictions from the trained CHGNet model with ground truth values from MP-20 validation set.



(b) Comparing band gap predictions from the trained CHGNet model with Quantum Espresso simulation outputs for MP-20 validation set.

Figure 6: Examining the predictive performance of the trained band gap model.