A Experimental Details

A.1 Pre-simulation Metrics

Pre-simulation metrics were computed for crystals designed by the policies prior to performing simulation using DFT – they are 1) Accuracy, 2) Similarity 3) Compositional Validity (referred to as validity for simplicity), and 4) Novelty. Further details on how to calculate them are provided below.

A.1.1 Accuracy

Accuracy is measured as the percentage of predicted atoms that match the ground truth. Note that the accuracy in this case is comptued globally across atoms predicted in all the crystals present in the validation dataset.

$$\text{Accuracy } (\%) = \frac{\#\text{ predicted atoms that exactly match the ground truth}}{\text{Total number of predicted atoms}}$$

We can also measure the fraction of crystals that were reconstructed to match the ground truth exactly. However, this was a very small percentage ($\sim$2-7%) for all the models.

A.1.2 Similarity

While accuracy measures the fraction of exact matches, our similarity metric considers a prediction as a match if the predicted atom and the ground truth atom belong to the same category. The categories are defined as follows.

1. **Group 1**: Li, Na, K, Rb, Cs
2. **Group 2**: Be, Mg, Ca, Sr, Ba
3. **Transition Metals**: Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg
4. **Nonmetals**: H,B, C, N, O, Si, P, S, As, Se, Te
5. **Halogens**: F, Br, Cl, I
6. **Noble**: Xe, Ne, Kr, He
7. **Lanthanides**: La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu
8. **Actinides**: Ac, Th, Pa, U, Np, Pu

$$\text{Similarity } (\%) = \frac{\#\text{ predicted atoms that have same category as ground truth}}{\text{Total number of predicted atoms}}$$

A.1.3 Compositional Validity

We follow [45] and compute the compositional validity of crystals using SMACT [7].

$$\text{Validity } (\%) = \frac{\#\text{ valid crystals}}{\text{Total number of crystals}}$$

A.1.4 Novelty

To assess the novelty aspect of our approach, we compute the fraction of valid generated crystals whose compositions are novel, i.e., when the compositions are not present in the Materials Project Database[14]. We utilised the API of Materials Project (`mpr.summary.search` function) to retrieve crystals with matching compositions. Note that our novelty percentage is conditioned on the valid crystals, and we do not query invalid compositions. Hence, in Table 1, while other metrics are computed by dividing the total number of crystals in the validation set, novelty is computed by dividing the number of valid crystals generated by the model.

$$\text{Novelty } (\%) = \frac{\#\text{ crystals with novel compositions}}{\#\text{ valid crystals}}$$

A.2 Post-simulation Metrics

Post-simulation metrics were computed for crystals designed by the policies after performing simulation using DFT. As indicated in Appendix A.6.1, crystals that failed DFT simulation were not included while computing post-simulation metrics. Details on how to calculate them are provided below.

A.2.1 Average Formation Energy

Following Appendix A.7, the formation energy was calculated for all the generated and valid crystals that successfully underwent DFT simulation. The average formation energy if therefore,

$$\bar{E}_{form} = \sum_{i=1}^{N} \frac{E_{form,i}}{N} \text{ (eV/atom)}$$

where N is the number of valid crystals whose formation energy values were computed successfully using DFT.

A.2.2 EMD (Band Gap)

The Earth Mover's Distance (EMD) was computed to determine the distributional distance between the properties of generated crystals and the ground truth crystals in the validation dataset. For band gap, the Γ^{p}_{true} was calculated as follows.

$$\Gamma^{p}_{true} = EMD(\{p_i\}_{i=1}^{M}, \{\tilde{p}_j\}_{j=1}^{N})$$

where M is the total number of crystals in the validation set, and N is the number of valid generated crystals that successfully underwent DFT simulation. p_i is the property value of the i^{th} crystal in the validation set, and $\tilde{p}_j$ is the property value of the j^{th} valid crystal generated by the model.

A.2.3 EMD (Formation energy)

Similar to Γ^{p}_{true}, EMD between the true and generated formation energy distributions, Γ^{E}_{true} were computed as follows.

$$\Gamma^{E}_{true} = EMD(\{E_{form,i}\}_{i=1}^{M}, \{\tilde{E}_{form,j}\}_{j=1}^{N})$$

$E_{form,i}$ is the property formation energy (ev/atom) of the i^{th} crystal in the validation set, and $\tilde{E}_{form,j}$ is the property value of the j^{th} valid crystal generated by the model.

A.2.4 Desired Range

The desired range metric (ν) is the fraction of generated crystals whose property (here, band gap) lies between $\hat{p} - 0.25$ and $\hat{p} + 0.25$, where $\hat{p}$ is the target property. For simplicity and easier analysis, the denominator of this fraction is the total number of crystals in the validation set. This way, the metric provides a way to quantitatively compare the corresponding percentages across different models.

$$\nu = \frac{\text{\# generated crystals in the property range } (\hat{p} - 0.25, \hat{p} + 0.25)}{M}$$

Here, M is the number of crystals in the validation set.

A.2.5 OOD Design

Through the κ metric, we compared the number of crystals generated (from the validation set) whose property value lies in the desired range, i.e., $(\hat{p} - 0.25, \hat{p} + 0.25)$, but the corresponding ground truth property is outside the desired range (hence, OOD crystals). This indicates that the model has learned to place atoms such that the property shifts from a value outside the desired range to within the range. Similar to ν, the denominator is M, the number of crystals in the validation set.

$$\kappa = \frac{\text{\# OOD crystals}}{M}$$

A.3 Additional Post-simulation Results

As part of the post-simulation analysis, we also investigated the average band gap of the crystals designed by each model.

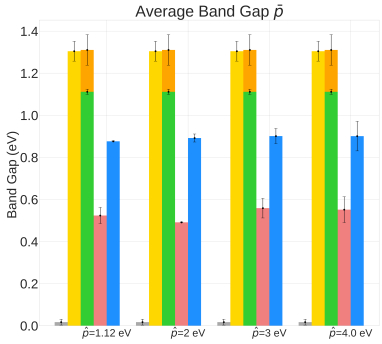


Figure 4: Analysis of average band gap of generated crystals in the validation set. BC and unconditional polcies have an average band gap closer to the ground truth average (1.892 eV). Random policy failes to generate crystals with higher band gap.

A.4 MEGNet

In our work, we adopted the MEGNet [4] model to process crystal graphs and extract state representation, as part of the Q-network Q_θ. The important hyperparameters of the model are listed below.

- Number of MEGNet blocks: 3
- Node embedding dimensions: 16
- Edge embedding dimensions: 1
- State embedding dimensions: 8
- $READOUT$ Function: Order-invariant *set2set* [42]

A.5 Offline RL

We adopt conservating Conservative Q-Learning (CQL) [19] as the offline RL approach. The important hyperparameters of our training process is listed below.

- Number of steps trained: 250000
- Discount factor: 0.99
- Batch size: 1024
- Learning rate: 3×10^{-4}
- Soft target network update rate: 5×10^{-3}
- Optimizer: Adam

A.6 DFT Parameters (Quantum Esperesso)

For performing DFT calculations, we use the Quantum Espresso v7.1 [11] simulation suite. The details of the DFT parameters are given below. For simplicity, this configuration was used for all crystals, and the evaluation is consistent for the training and generated crystals. Note that we do not perform structure relaxation in any of the cases.

- Calculation: SCF
- Pseudopotentials: Solid-state pseudopotentials (SSSP) version 1.3.0 obtained from https://www.materialscloud.org/discover/sssp/table/efficiency

- Tolerance: 10^{-6}
- Number of Bands: 256
- k-points: (3-3-3)
- Occupations: fixed (since our training set consists only of nonmetallic crystals)
- Diagonalization: David
- ecutrho: 245
- ecutwfc: 30
- mixing_beta: 0.7
- degauss: 0.001
- Default charge: 0
- Maximum iterations: 1000

A.6.1 Handling Failures

It is important to note that DFT can be best leveraged once we know certain properties of the crystals – for example, charge, magnetization, and metallicity. Considering the difficulty in determining these properties for completely unknown crystals, we standardized the evaluation procedure by using the same DFT configuration for all crystals (except for the crystal-specifc parameters like number of atoms, species, and pseudopotentials directory). However, this resulted in multiple crystals failing DFT simulation. Some of the errors are explained below.

- *Charge is wrong. Smearing is needed.*: This error mainly occurs because of unpaired electrons in the system, and can be resolved by changing the occupation to 'smearing' instead of 'fixed'. However, doing so will not help in determining the band gap of crystals, as it will only output the Fermi energy. Another way is to set the 'nspin' parameter to 2 and specify the total magnetization value as an additional input to Quantum Espresso. This helped us resolve most of the failures for the MP-20 crystals in the training and validation set because the total magnetization value is retrievable from the Materials Project, but for the newly generated crystals, we had to ignore those that failed because of this error. The error could also occur if the generated crystal is metallic, and this property is also difficult to identify directly from the structure and composition.
- *NOT converged in 1000 iterations*: For some crystals, the DFT simulation did not converge even after 1000 iterations. These crystals were ignored while constructing the offline dataset, and also when evaluating the policy-generated crystals.
- *Time limit exceeded*: For constructing the offline dataset using known crystals, we used a flexible time limit to ensure none of the crystals were discarded because of time restrictions. However, while performing DFT simulation for the policy-generated crystals, due to the high-throughput nature of our evaluation pipeline, we had to ignore crystals that did not converge in 15 minutes.
- *Too few bands*: This error occurs when the number of bands specified, through 'nbnds' parameter is insufficient for the crystal system being simulated. This error was largely resolved by specifying a higher number of bands. In our case, we used 256 bands for all crystals.

Overall, during evaluation of generated crystals, only 50-70% of the valid crystals successfully underwent DFT simulation to output the energy and band gap (Table 2), and the rest failed because of the above errors.

A.6.2 % DFT Success

Table 2 shows the percentage of policy-generated crystals that successfully underwent DFT simulation based on failure handling strategies discussed in Appendix A.6.1.

	% DFT Success	
CQL Weight	$\omega = 1$	$\omega = 5$
$Random$	15.18	
BC	68.25	
u-CQL	68.97	70.29
c-CQL($\hat{p} = 1.12$ eV)	58.59	66.82
c-CQL($\hat{p} = 2$ eV)	56.04	67.99
c-CQL($\hat{p} = 3$ eV)	56.55	68.38
c-CQL($\hat{p} = 4$ eV)	55.64	66.19

Table 2: % Generated valid crystals that successfully underwent DFT simulation, for random policy and each of the trained models. Most of the crystals generated by the random policy failed DFT simulation.

A.7 Formation Energy Calculation

The formation energy per atom was calculated using the total energies of the crystals and their constituent elements. The total energies of the isolated elements (88 in the action space) were calculated by performing SCF calculations on the most stable elemental crystals (i.e., 0 formation energy) present in the Materials Project. For elements that do not have a stable elemental crystal (e.g. Lu) or those that have large number of atoms in the elemental crystal (e.g. P, Se), the total energies were calculated for a single atom inside a primary cubic cell of length 10Å. For a crystal with N atoms, the formation energy (per atom) calculation is defined as follows.

$$E_{form} = \left(\frac{E_{tot} - \sum_i \frac{N_i}{n_i} E^i_{tot}}{N} \right) * 13.6057039763 \ \text{(eV/atom)} \tag{15}$$

Here, N_i is the number of atoms of the constituent element i present in the crystal, n_i is the number of atoms (sites) of i in the elemental crystal, and E^i_{tot} is the total energy of i in the most stable elemental crystal form. 13.6057039763 is the value of 1 Rydberg constant in eV.

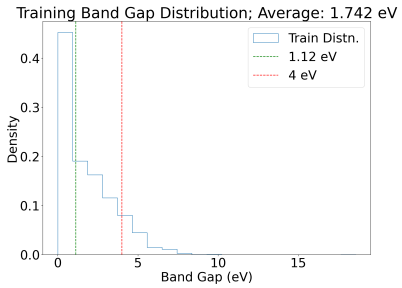
A.8 Algorithm

Algorithm 1 Training Conditional CQL: DQN Version for Crystal Design with Target Property $\hat{p}$

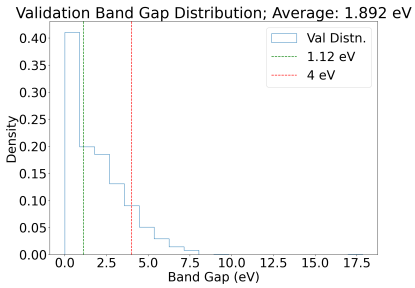
Construct dataset $\mathcal{D}$ of size $N_{\mathcal{D}}$ consisting of transitions $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}', r)$ using known crystals
Load $\mathcal{D}$ in Replay Buffer $\mathcal{B}$
Initialize Q-network Q_θ and target network $Q_{\theta'}$, batch size B
for $j = 1$ to max_steps **do**
  Sample B transitions, $\{(\boldsymbol{s_i}, \boldsymbol{a_i}, \boldsymbol{s'_i}, r_i)\}_{i=1}^B$ from $\mathcal{B}$
  Compute TD loss
  $$L_i^{TD}(\theta) = \begin{cases} \left(Q_\theta(\boldsymbol{s_i}, \boldsymbol{a_i}; \hat{p}) - (r_i + \gamma \max_{\boldsymbol{a}} Q_{\theta'}(\boldsymbol{s'_i}, \boldsymbol{a}; \hat{p}))\right)^2 & \text{if } \boldsymbol{s'_i} \text{ is not terminal} \\ \left(Q_\theta(\boldsymbol{s_i}, \boldsymbol{a_i}; \hat{p}) - r_i\right)^2 & \text{otherwise} \end{cases}$$
  $L^{TD}(\theta) = \frac{1}{B} \sum_{i=1}^B L_i^{TD}(\theta)$
  Compute conservative loss, $L^C(\theta) = \frac{1}{B} \sum_{i=1}^B \left[\log \sum_{\boldsymbol{a}} \exp(Q_\theta(\boldsymbol{s_i}, \boldsymbol{a}; \hat{p})) - Q_\theta(\boldsymbol{s_i}, \boldsymbol{a_i}; \hat{p})\right]$
  Compute total CQL loss $L^{CQL}(\theta) = \omega L^C(\theta) + \frac{1}{2} L^{TD}(\theta)$
  Compute gradients and backpropogate: $\theta \leftarrow \theta - \eta \nabla L^{CQL}(\theta)$, η is the learning rate
  Update target network parameters θ'
end for

B True Distributions of Properties

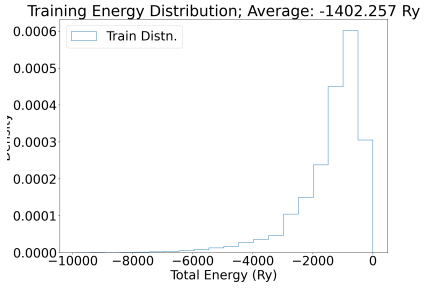
This section shows the true distribution of the band gaps and total energies for both training and validation data.



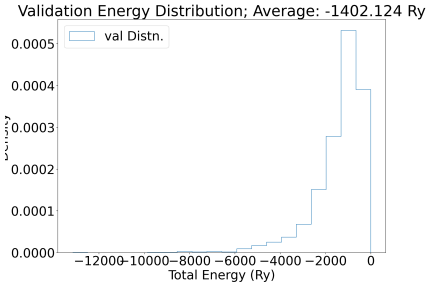
(a) Band Gap Distribution (Training Data)



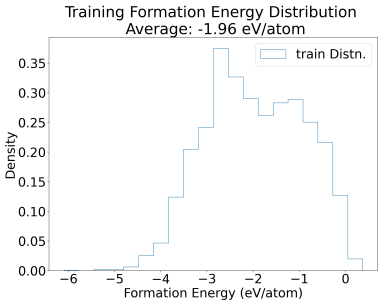
(b) Band Gap Distribution (Validation Data)



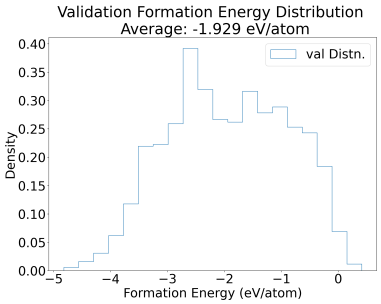
(c) Energy Distribution (Training Data)



(d) Energy Distribution (Validation Data)



(e) Formation Energy Distribution (Training Data)



(f) Formation Energy Distribution (Validation Data)

C Experiments with Total Energy

As part of our initial analysis, we performed the experiments with total energy (E_{tot}) in the reward formulation instead of formation energy, with the aim of designing crystals that are generally considered stable (in an absolute sense), so they can be used for practical purposes. However, total energy is less meaningful when it comes to comparing the stability of different crystals, while energy above hull is the best-known metric to compare thermodynamic stability.

C.1 Reward Formulation

Since the units of total energy are in Rydberg (Ry), our reward function in Equation (7) can be redefined as follows.

$$r_N = \alpha_1 \log_{10}(-E_{tot}) + \alpha_2 \exp\left[-\frac{(p-\hat{p})^2}{\beta}\right]. \tag{16}$$

C.2 Full Experimental Metrics with Total Energy

We provide full experimental for our reward function design parameters for both the 1.12 eV design case (Table 3 and Figure 7 and 4 eV case (Table 4 and Figure 8) below. The tables and figures include evaluation of both the pre-simulation (except Novelty) and post-simulation metrics (except κ)

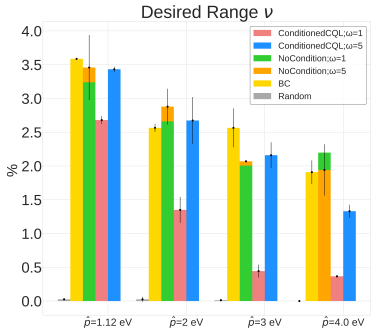
	Accuracy (%)		**Similarity (%)**		**Validity (%)**	
CQL Weight	$\omega = 1$	$\omega = 5$	$\omega = 1$	$\omega = 5$	$\omega = 1$	$\omega = 5$
Random	0.0115		0.1254		NaN	
BC	52.26		71.98		85.00	
uCQL	49.77	51.53	70.85	71.26	81.50	82.54
$(0-5-1)$	38.64	48.85	61.23	69.38	69.99	77.84
$(0-5-3)$	43.02	46.43	65.01	67.04	73.57	78.44
$(0-10-1)$	36.54	43.72	59.3	65.18	73.33	80.81
$(0-10-3)$	35.16	42.42	57.48	64.15	71.20	81.30
$(1-5-1)$	42.11	47.72	64.00	68.12	75.62	80.29
$(1-5-3)$	40.59	47.57	63.70	67.26	72.93	76.51
$(1-10-1)$	35.02	43.18	58.63	65.13	67.82	75.14
$(1-10-3)$	35.38	43.81	57.23	65.58	61.87	77.19

Table 3: Pre-simulation metrics for band gap design case of 1.12 eV with $(\alpha_1 - \alpha_2 - \beta)$ corresponding to the terms of the reward function in Equation (16) with the policy in Figure 2 and best by metric highlighted. Unconditional policies perform better on pre-simulation metrics while conditioned policies produce target designs shown as in Figure 2 and discussed in Section 5.2.

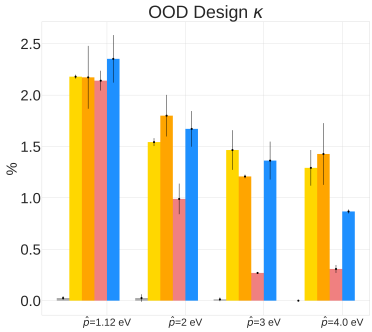
	Accuracy (%)		**Similarity (%)**		**Validity (%)**	
CQL Weight	$\omega = 1$	$\omega = 5$	$\omega = 1$	$\omega = 5$	$\omega = 1$	$\omega = 5$
Random	0.0115		0.1254		NaN	
BC	52.26		71.98		85.00	
uCQL	49.77	51.53	70.85	71.26	81.50	82.54
$(0-5-1)$	41.82	48.09	64.34	68.82	80.21	82.18
$(0-5-3)$	39.46	47.61	61.59	68.24	74.46	80.09
$(0-10-1)$	33.24	39.42	60.78	53.42	62.39	67.82
$(0-10-3)$	35.24	41.47	57.14	64.06	64.40	75.54
$(1-5-1)$	38.80	46.79	60.09	68.77	70.80	80.17
$(1-5-3)$	42.06	47.49	63.36	68.35	78.32	81.0
$(1-10-1)$	36.52	42.21	59.57	65.07	76.55	74.41
$(1-10-3)$	35.94	42.91	56.8	64.2	68.95	77.63

Table 4: Band gap design case of 4 eV with similar nomenclature and conclusions as Table 1.

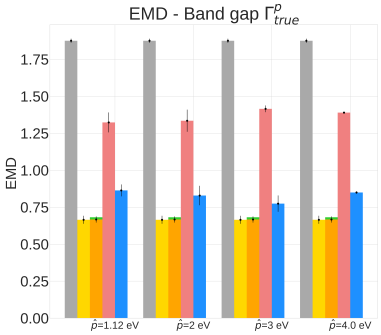
described in Section 5. With $\alpha_1 = 1, \alpha_2 = 5, \beta = 1$, the post-simulation results for all four band gap targets are shown in Figure 6. All models in highlighted in this section were trained for 500000 steps.



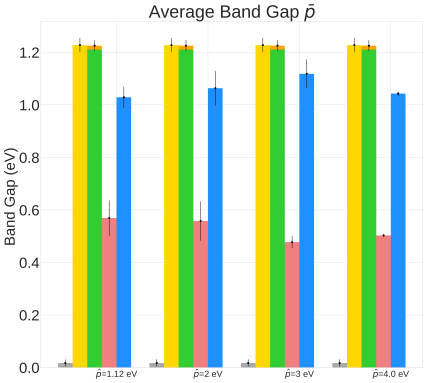
(a) % Desired range for different band gaps targets for various policies. Conditioned policies outperform random policy and compete with unconditional policies in designing crystals in the desired property range.



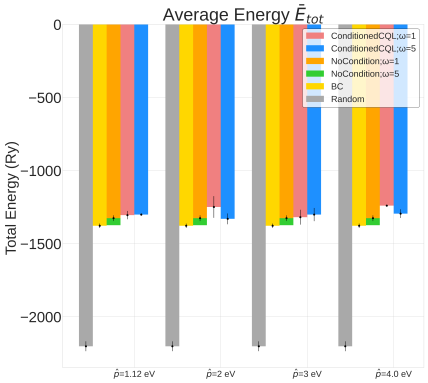
(b) % of generated crystals with property in the desired range with corresponding ground truth crystals outside the desired range.



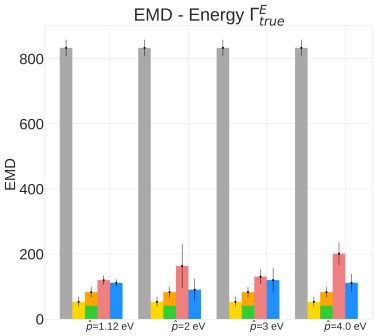
(c) Band gap EMD (generated vs. true) for various policies showing that unconditioned policies reproduce the original dataset better. Lower is better.



(d) Average Band Gap for various policies. Greater CQL conditioning ($\omega = 5$) yields greater alignment to the desired band gap for conditioned policies.

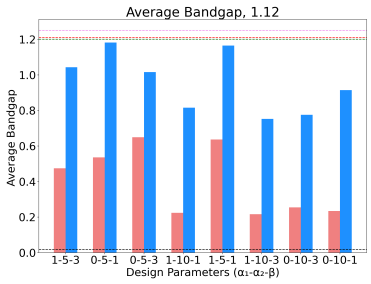


(e) Average total energy for various policies yielding valid crystals with energy below 0. Possible reasons for random policy having the lowest energies are provided in Section 5.2.

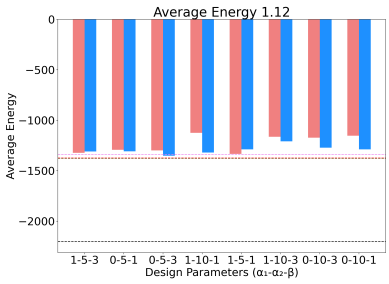


(f) Energy EMD (generated vs. true) for various policies showing that unconditioned policies reproduce the original dataset better. Lower is better.

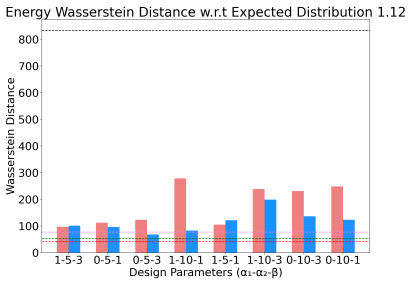
Figure 6: Results for conditioned CQL policies on all band gap design targets. Conditioned and more conservative policies perform better in the κ metric in some cases, while unconditioned policies, including behavioral cloning, perform better at reproducing the original distribution. Random policies fail to reproduce the original distribution and achieve desired properties.



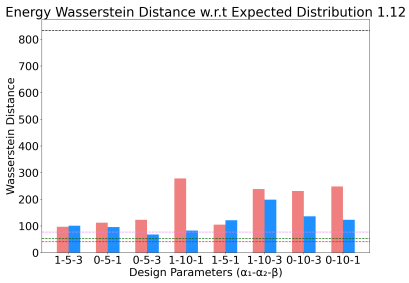
(a) Average band gap



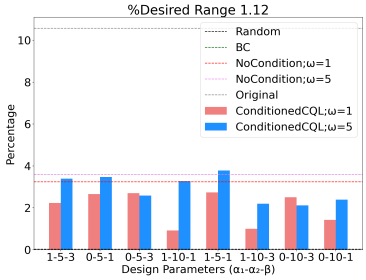
(b) Average energy



(c) Band gap Wasserstein distance (generated vs true)

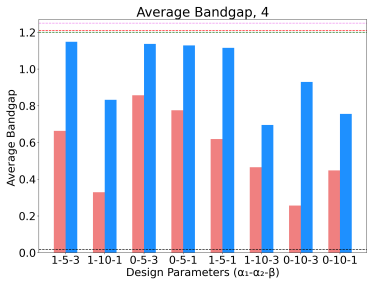


(d) Energy Wasserstein distance (generated vs true)

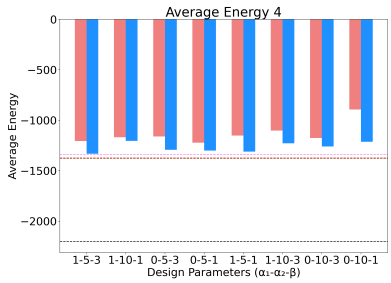


(e) % Desired range (0.87-1.87eV)

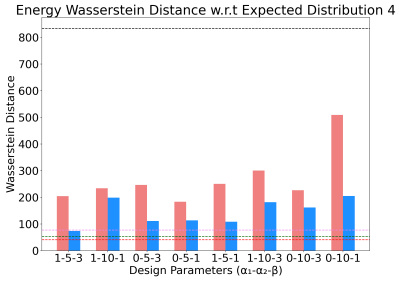
Figure 7: Full design parameter values for all learned policies for the band gap design case of 1.12 eV. Nomenclature of the table is $(\alpha_1 - \alpha_2 - \beta)$ corresponding to the terms of the reward function in Equation (16)



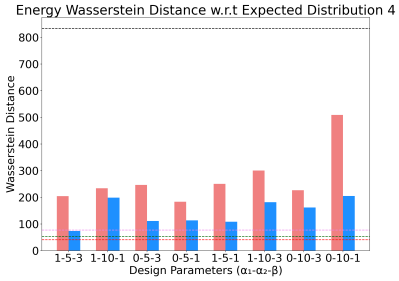
(a) Average band gap



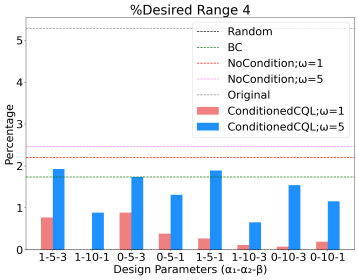
(b) Average energy



(c) Band gap Wasserstein distance (generated vs true)



(d) Energy Wasserstein distance (generated vs true)



(e) % Desired range (3.75-4.25)

Figure 8: Full design parameter values for all learned policies for the band gap design case of 4.0 eV. Nomenclature of the table is $(\alpha_1 - \alpha_2 - \beta)$ corresponding to the terms of the reward function in Equation (16)