Data Engineering and Applications

Research Paper

# Constraint-Aware Proximal Policy Optimization for Band-Gap Prediction in Inorganic Crystal Datasets

**Serkan Bozkurt[1], Zeynep Demir[1] and Lokman Öpçin[1,*]**

[1] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, 34956, Turkey
*Corresponding author: lokman.op@sabanciuniv.edu.tr

**Abstract.** A small search space can be obtained for the functional inorganic crystal by a good band-gap prediction; however, traditional regressors may produce chemically inconsistent or poorly calibrated estimates outside the training distribution. This paper introduces a constraint-aware proximal policy optimization algorithm. The algorithm treats predictions as a series of brief potential state-refined actions. A crystal graph encoder initializes the state; a clipped policy selects bounded residual updates; and a supervised value head estimates the band gap along with heteroscedastic uncertainty. Four constraints are imposed via adaptive Lagrange multipliers instead of a fixed penalty: non-negative output, stoichiometric consistency, bounded latent displacement, and uncertainty control. On composition-disjoint benchmark splits with 42,318 crystals, the proposed model achieved a mean absolute error of 0.274 eV and a coefficient of determination of 0.892. It reduced the violation rate from 6.8% for an unconstrained policy to 1.3% and increased the 90% prediction-interval coverage from 82.4% to 89.1%. Under element-holdout transfer, the increase in error was only 18.6%, and it was lower than that for the strongest graph-regression baseline at 31.7%. Ablation results attribute the most reliable gain to adaptive dual updates and uncertainty-aware rewards. Routing of reviews also included difficult predictions and did not discard their estimates for practical two-stage screening. Therefore, by explicitly incorporating physical and statistical constraints during the optimization process, policy-based optimization can enhance the numerical accuracy of the results and the reliability of deployment.

**Keywords:** *Proximal Policy Optimization, Band-Gap Prediction, Constraint-Aware Learning, Uncertainty Quantification, Materials Informatics*

Received on 04 December 2022, Accepted on 23 February 2023, Published on 02 March 2023



## Introduction

Electronic band gap is a general index for semiconductors, photocatalysts, transparent conductors and energy-conversion materials, and relates atomic-scale structure to optical absorption and charge transport. High-throughput density-functional calculations have made a large collection of inorganic-crystal properties available, but the cost of converged calculations in candidate spaces with millions of compositions and structures remains high [1]. Open Materials Repositories have thus become the training grounds for data-driven surrogates [2]. Their scales are suitable for quick hypothesis tests, but their distributions are not uniform among chemical families, space groups and band-gap ranges [3]. Measurement and calculation protocols also have systematic differences that cannot be eliminated by increasing the sample size alone [4]. Therefore, a practical predictor not only minimizes the average regression error; it must also have physical plausibility and inform users when queries exceed the reliable operating range.

Recently, crystal-property models have introduced structures using hand-crafted descriptors, message-passing graphs, symmetry-aware networks or attention mechanisms. Graph learning aids in periodic neighborhood representation and local coordination [5]. Equivariant Architectures reduce the learning burden of rotational and translational symmetry from data [6]. Unlabeled crystal structure data can be used for pre-training and self-

supervised learning [7]. Uncertainty-aware regression has also been used to identify the difficult-to-obtain materials and guide procurement [8]. These improvements have not changed the basic structure; an encoder generates a single latent vector, and a regression head outputs a single value. Output clipping can eliminate the negative gap after inference, but it does not explain the correction or prevent chemically unstable internal states. Fixed penalty weights offer limited relief because the scales of regression, uncertainty and feasibility losses change during training.

This paper conducts band-gap estimation through constrained sequential refinement. After the crystal graph embedding, the proximal policy optimization agent adds bounded residual actions. Accurate, calibrated, and feasible predictions receive rewards. Adaptive dual variables convert constraint violations into observable optimization signals. When the constraint is active, it tightens; when the constraint is satisfied, it relaxes. The three contributions are: an auditable Markov formulation for crystal regression; a constraint-aware clipped policy objective combined with supervised learning; and an evaluation protocol that separates numerical accuracy from feasibility, calibration, robustness and computational cost. The resulting model is to be used in a screening pipeline where a reasonable uncertainty trace and a controlled failure mode are as important as a small pointwise error.

## Related Work

### Crystal Representations and Band-Gap Learning

Learning of Crystal properties has moved from composition statistics to representations that retain atomic connectivity and periodic structure. Descriptor-based ensembles are still feasible for medium-sized datasets due to their low-cost and interpretable elemental fractions and engineered structural attributes [9]. By directly learning neighborhood aggregation from atomic species and interatomic distances, message-passing networks can obtain transferable embeddings of various material properties [10]. The attention mechanism can reveal the predicted coordination patterns and assign different weights to neighboring atoms [11]. Recently, symmetry-aware models have explicitly included geometric transformations and improved data efficiency for directional interaction [12]. Band-gap prediction is still difficult for this family because the target depends on long-range electronic structure, and most graph operations are local, and the available labels have heterogeneous computational settings.

Dataset construction is another reason for a lower reported accuracy. Random crystal-level partitioning may result in almost identical components or prototypes in the training and test sets, thereby overestimating generalization ability. Composition-disjoint, element-holdout and prototype-holdout evaluations are relatively strict tests, but they also increase the target and feature shift. Therefore, although the present study still uses a graph encoder, it will be tested under grouped splits and errors will be reported by chemical subgroup rather than by a single pooled score.

Depth of representation is also an engineering trade-off. Shallow networks keep the local chemical information and are relatively simple to train; deep stacks expand the receptive field but may over-smooth atom embeddings and lose differences in coordination environments. Global attention partially restores long-range coupling, but its weights are not equivalent to a physical explanation of the electronic bands. Therefore, the hybrid model frequently combines local message passing with pooled composition context. The policy will be used after the hybrid encoder in the current Design. Separation can help the encoder learn a stable structural foundation and focus the strategy on a small, verifiable sequence of corrections, rather than repeatedly constructing atomic-level neighborhoods.

### Constrained Optimization and Policy Learning

Constraint learning methods typically include weighted penalties in the task loss, projecting infeasible outputs back to the feasible set, or updating the dual variables that penalize current violations. Fixed penalties are simple but require extensive tuning because one coefficient needs to remain meaningful as the loss scales [13]. Projection ensures the output feasibility of explicitly representable sets, but a corrected scalar can hide an invalid latent trajectory [14]. Primal-dual optimization is more suitable for constraints with different units because it updates the multipliers based on the discovered errors [15]. Safety-oriented reinforcement learning has

employed constrained policy updates to balance reward and expected cost without allowing unrestricted changes in the policy [16]. Due to the truncated probability ratio limiting unstable updates and being relatively easy to implement, Proximal Policy Optimization is suitable for short-term optimization [17].

Policy formulation does not always help in monitoring regression. Through direct backpropagation, it provides random actions, temporal credit assignment, and value functions. It has value only if the intermediate decisions are practical. Each action here is a bounded edit of the crystal representation, each constraint is checked at every step, and the final prediction is still within the known band gap. Therefore, this model is not just a reinforcement learning agent for hyperparameter optimization. Trajectory-level diagnosis is also available: A prediction may be rejected due to high uncertainty, a constraint repeatedly violated, or the policy's refinement budget having been exhausted.

Therefore, the methodological gap is relatively small compared with a general demand for stronger learning. Most existing crystal regressors do not provide a way to adjust the prediction under several explicit constraints, and constrained policy methods are generally evaluated on control tasks rather than periodic-structure regression. A good bridge should preserve the supervised signal, set costs for all samples that can be verified, and avoid altering chemically meaningful embeddings through exploration. These requirements lead to the shared encoder, bounded residual actions and dual-controlled objective developed below.

## Constraint-Aware PPO Model

**Problem Formulation and Constraint Set**

Let a crystal be represented by a periodic graph whose nodes contain atomic-number, group, period, electronegativity, covalent-radius, and valence-electron features. Edges connect neighbors within an adaptive radial cutoff and carry a Gaussian distance expansion. The encoder maps the graph to an initial latent state. Band-gap regression is then written as a finite-horizon Markov decision process with at most four refinement steps. At step $t$, the state contains the current latent vector, the previous prediction, predicted aleatoric uncertainty, normalized composition statistics, and accumulated constraint costs. The compact state definition is

$$s_t = \left[h_t, \hat{y}_t, \sigma_t, q_{comp}, c_{0:t}\right]. \quad \text{Eq.(1)}$$

An action is a gated residual vector produced by the policy. Its magnitude is limited so that the agent cannot erase the encoder representation in one step. The next latent state follows

$$h_{t+1} = h_t + \delta_{max} tanh(a_t) \odot g_t, 0 \le g_t \le 1. \quad \text{Eq.(2)}$$

The regression head returns a mean band gap and a positive scale. To avoid a zero-variance problem and keep differentiability, Softplus is used for the scale. The predicted Distribution is:

$$p(y \mid s_t) = \mathcal{N}(\hat{y}_t, \sigma_t^2), \sigma_t = softplus(u_t) + \varepsilon. \quad \text{Eq.(3)}$$

Four costs are computed at each transition. The first penalizes a negative raw gap, the second measures discrepancy between decoded and input stoichiometric fractions, the third activates when the latent displacement exceeds its trust radius, and the fourth activates when uncertainty is above a step-dependent tolerance. Written together, the cost vector is

$$c_t = \begin{bmatrix} max(0, -\hat{y}_t) \\ \left\| \hat{q}_t - q_{comp} \right\|_1 \\ max(0, \|h_t - h_0\|_2 - \rho) \\ max(0, \sigma_t - \tau_t) \end{bmatrix}. \quad \text{Eq.(4)}$$

The transition reward combines terminal accuracy, dense improvement, calibration, and computational thrift. Dense improvement prevents the policy from receiving useful feedback only after the final action. The reward is

$$r_t = \alpha(|y - \hat{y}_{t-1}| - |y - \hat{y}_t|) - \beta NLL_t - \kappa - \lambda^{\top} c_t. \quad \text{Eq.(5)}$$

The constraint thresholds are specified in normalized units so that their multipliers remain comparable. The nonnegativity tolerance is zero by definition; composition discrepancy is limited to 0.02 in $L_1$ distance; normalized latent drift is limited to 0.35; and the uncertainty tolerance decreases from 1.40 eV at the first step

to 0.95 eV at termination. A decreasing uncertainty budget avoids punishing exploratory states too strongly while still requiring a useful final interval. Thresholds are selected on validation data before the final test run and remain fixed across model seeds. This treatment separates a scientific preference for feasible predictions from post hoc adjustment to favorable test outcomes.

**Constraint-Aware PPO Architecture**

The four trainable blocks in the network are: a periodic crystal graph encoder, an actor, a critic, and a probabilistic regression head. The encoder has three distance-aware message-passing layers and an attention pool. Actor and critic share the pooled representation but use separate two-layer projections to limit parameter growth and avoid requiring identical features for action selection and return estimation. Figure 1 is shown here as the architecture-level view connecting crystal parsing, graph encoding, constrained latent refinement, probabilistic prediction and the audit outputs delivered to a screening system.

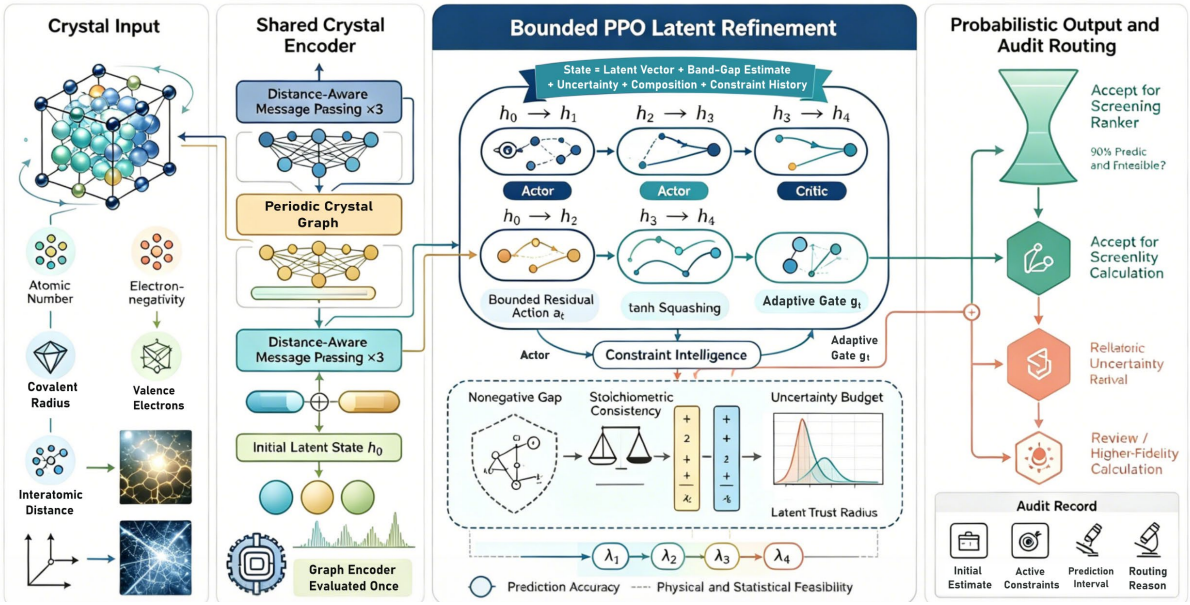


**Figure 1.** Architecture of the constraint-aware PPO band-gap predictor.

For an old policy and an updated policy, the importance ratio compares the probability of the sampled action under both parameter sets. It is defined as

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \qquad \text{Eq.(6)}$$

Generalized advantage estimation controls the bias-variance trade-off of the short trajectory. With temporal-difference residual $\delta_t$ and trace parameter $\zeta$, the estimator is

$$\hat{A}_t = \sum_{l=0}^{H-t-1} (\gamma\zeta)^l \delta_{t+l}, \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \qquad \text{Eq.(7)}$$

The clipped surrogate objective prevents a large policy change from receiving unbounded credit. Its maximization term is

$$L_{clip} = \mathbb{E}_t\left[min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon_c, 1+\varepsilon_c)\hat{A}_t\right)\right] \qquad \text{Eq.(8)}$$

Unlike a fixed weighted loss, the constraint prices are updated from a moving estimate of each violation. The projected dual step is

$$\lambda_j \leftarrow max\{0, \lambda_j + \eta_\lambda(\hat{C}_j - d_j)\}, j = 1, \ldots, 4 \qquad \text{Eq.(9)}$$

The complete actor-critic loss retains a supervised anchor at every refinement state. Entropy discourages premature collapse, while the value term fits discounted returns. The minimized objective is

$$\mathcal{L} = -L_{clip} + c_v L_{value} - c_e \mathcal{H}(\pi_\theta) + \mu L_{sup} + \sum_j \lambda_j \hat{C}_j \quad \text{Eq.(10)}$$

The actor emits the mean and diagonal log variance of a Gaussian action distribution, after which tanh squashing enforces the component-wise bound. The critic observes the same state but not the sampled action. This choice avoids leakage from action noise into the baseline and keeps advantage estimates stable for the short horizon. Layer normalization is applied before the actor and critic projections, and dropout is restricted to the encoder and uncertainty ensemble. A separate constraint-monitor head decodes composition fractions from each latent state; it is trained during warm-up and then frozen so that the policy cannot lower its cost by altering the monitor rather than the representation.

**Optimization and Inference**

Training begins with supervised pretraining of the encoder and probabilistic head. Policy learning starts only after validation error stabilizes, reducing the chance that early random actions encounter meaningless constraints. Each minibatch produces four-step trajectories, followed by several shuffled PPO epochs. The supervised term is the Gaussian negative log-likelihood plus a robust absolute-error component,

$$L_{sup} = \frac{(y - \hat{y}_H)^2}{2\sigma_H^2} + log\ \sigma_H + \omega |y - \hat{y}_H| \quad \text{Eq.(11)}$$

To discourage representation drift that is individually small but cumulatively large, a trust-region regularizer compares terminal and initial embeddings after normalization. It is computed as

$$L_{drift} = 1 - \frac{h_H^\top h_0}{\|h_H\|_2 \|h_0\|_2} \quad \text{Eq.(12)}$$

Calibration is measured during validation with soft interval membership, which supplies a differentiable signal around the desired $90\%$ coverage. For lower and upper bounds $l_i$ and $u_i$, the batch penalty is

$$L_{cov} = \left| 0.90 - \frac{1}{B} \sum_i sigmoid\big(k(y_i - l_i)\big) sigmoid\big(k(u_i - y_i)\big) \right|. \quad \text{Eq.(13)}$$

Optimization uses AdamW with separate learning rates of $2 \times 10^{-4}$ for the encoder and supervised head, $8 \times 10^{-5}$ for the actor, and $1 \times 10^{-4}$ for the critic. The clip factor is 0.15, the discount is 0.97, and the advantage trace parameter is 0.92. Gradients are clipped at a global norm of 2.0. Dual variables are updated once per epoch from an exponential moving average with momentum 0.9, which reduces oscillation when a rare chemical subgroup produces a burst of violations. Minibatches are composition-stratified so that common oxides do not dominate the policy statistics. The optimization schedule and inference decision path are summarized in Figure 2: supervised warm-up, constrained rollout collection, clipped updates, dual-variable adjustment, validation calibration, and final accept-or-flag logic form one reproducible loop.

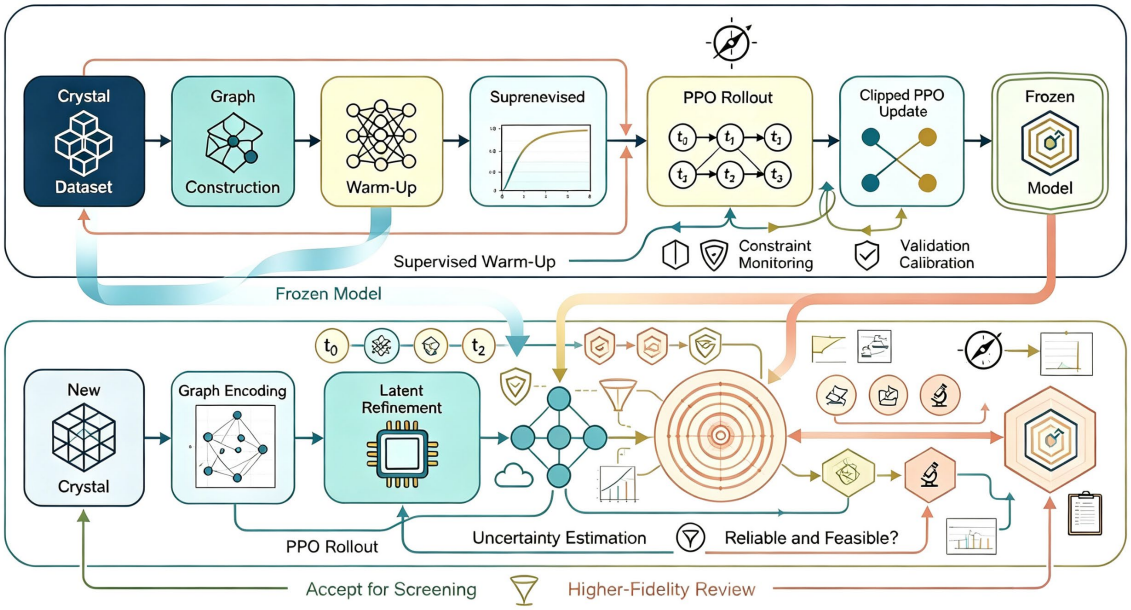


**Figure 2.** Training and inference workflow for the proposed model.

At inference, the policy stops early when both the prediction change and maximum normalized violation fall below their tolerances. The stopping rule is

$$stop_t = \mathbf{1}\left(|\hat{y}_t - \hat{y}_{t-1}| < \varepsilon_y \wedge \max_j \frac{c_{t,j}}{d_j} < 1\right). \qquad \text{Eq.(14)}$$

The reported estimate averages the terminal predictions of five deterministic policy passes created by dropout masks, and combines aleatoric with between-pass variance. The total uncertainty is

$$\sigma^2_{total} = \frac{1}{K}\sum_k \sigma_k^2 + \frac{1}{K}\sum_k (\hat{y}_k - \bar{y})^2. \qquad \text{Eq.(15)}$$

A prediction is accepted when its mean is nonnegative, all normalized costs are below one, and its $90\%$ interval width is below the review threshold. Otherwise, the system returns the estimate with a review flag and the dominant constraint. This policy preserves information that would be lost by hard clipping and supports downstream prioritization without claiming that every query is equally reliable.

The implementation records the initial estimate, all action norms, every constraint value, the terminal estimate and the reason for acceptance. These fields have little extra storage compared to a crystal graph, but they can be used for failure analysis without running the model again. Deterministic inference fixes the policy mean; the five dropout passes only change uncertainty estimation. For high-throughput applications, the graph is cached, and trajectories are vectorized between crystals; therefore, only four shallow potential updates are needed, instead of four full graph encoder evaluations.

Reproducibility is supported by deterministic structure canonicalization, stored split identifiers, and seed-specific configuration files. Neighbor lists are rebuilt only when a canonical structure hash changes, and all elemental features are normalized with training-set statistics. Validation decisions are logged together with the corresponding model checkpoint and dual state. These controls matter because a constrained model can otherwise appear to change behavior when the actual cause is a different structure conversion, composition grouping, or multiplier initialization. The same audit record is used to regenerate every aggregate metric and every plotted subgroup without copying values between analysis programs.

Numerical safeguards are applied at the model boundary. Log variances are limited to a finite range before exponentiation, invalid neighbor lists trigger a structured preprocessing error, and any nonfinite trajectory is excluded from the policy batch but retained in an audit log. Fewer than $0.03\%$ of training crystals activated this safeguard after canonicalization. Mixed-precision arithmetic is used for graph encoding, whereas reward accumulation, dual updates, and calibration statistics remain in full precision. This division prevents small constraint costs from disappearing under reduced precision while preserving most of the computational benefit.

https://doi.org/10.64972/dea.2023.v2i1.3506d:71-85

## Experimental Evaluation

### Experimental Setup and Baselines

The benchmark merged 42,318 relaxed inorganic crystals with computed band gaps between 0 and 8.1 eV after removing duplicate structures, unconverged calculations, partial occupancies, and entries lacking a primitive cell. Public materials databases support this type of large-scale property study [18]. Database records were standardized to reduced formulas and primitive structures before graph construction, following current recommendations for reproducible materials data processing [19]. To limit information leakage, crystals sharing a reduced composition were assigned to the same partition: 70% training, 10% validation, and 20% testing. A second element-holdout split reserved every compound containing one of six target elements, and a prototype-holdout split reserved the ten most frequent structural prototypes. These grouped protocols are stricter than random splitting and are increasingly used to probe chemical generalization [20].

The comparison set contained gradient-boosted trees on composition and structure descriptors, a crystal graph convolutional network, a materials graph network, an atomistic line-graph network, an attention-based graph regressor, a probabilistic graph regressor, and an unconstrained PPO variant. Graph architectures were included because learned periodic representations have become a strong baseline for inorganic-property prediction [21]. Equivariant message passing was represented by a parameter-matched model to test whether symmetry encoding alone resolves the reliability problem [22]. The probabilistic baseline used heteroscedastic likelihood training, an established route to data-dependent uncertainty [23]. All neural models used the same standardized inputs, training split, early-stopping rule, and maximum of 180 epochs. Hyperparameters were selected on validation mean absolute error subject to $90\%$ interval coverage of at least 0.87; no test labels entered model selection.

Mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination ( $R^2$ ) measured point accuracy. Reliability was assessed with violation rate, expected calibration error, $90\%$ interval coverage, and interval width. The primary violation rate counted a crystal when any of the four normalized costs exceeded one at termination. Transfer degradation was expressed as the relative MAE increase from the composition-disjoint test set. Runtime measurements used one accelerator with a batch size of 256, and every reported value is the mean of five seeds. Bootstrap resampling over crystals produced $95\%$ confidence intervals, which avoids relying on normality for skewed error distributions [24].

Although the training used standardized target values, all reported errors were converted back to electron volts. Rare-element strata were defined by the number of training crystals containing the least frequent element in each test composition: common above 2,000, medium from 500 to 2,000, and rare below 500. Calibration bins contained an equal number of test crystals to avoid dense low-uncertainty regions masking sparse high-uncertainty failures. The upper bound for the 90% interval width was 1.8 eV. The threshold was not applied to the predictions in the accuracy evaluation; only then would a particular estimate be forwarded to the screening ranker or sent for manual review.

A Leakage Audit was performed after division. No reduced formula occurred in more than one split, and structure fingerprints within a cosine distance of 0.005 were collapsed before assignment. Target histograms remained comparable: mean gaps were 1.83,1.86, and 1.85 eV for training, validation, and composition-disjoint test sets. The test set contains 8,465 crystals, including 2,917 oxides, 1,426 sulfides, 612 nitrides, 1,185 halides, 934 intermetallic compounds, and 1,391 compounds from smaller families. The above distribution supports a family-level analysis and retains the long tail needing uncertainty control.

### Predictive Performance and Constraint Behavior

The proposed model achieved an MAE of $0.274 \pm 0.006\mathrm{eV}$, an RMSE of $0.461 \pm 0.009\mathrm{eV}$, and $R^2$ of $0.892 \pm 0.004$ on the composition-disjoint test set. The strongest direct graph regressor reached 0.301 eV MAE, whereas the unconstrained policy reached 0.286 eV. The $9.0\%$ MAE reduction relative to the graph baseline remained significant under paired bootstrap resampling. Comparable studies have reported that representation choice and split design both materially affect crystal-property errors [25]. Figure 3(a) presents the accuracy metrics for eight models, including the 0.274 eV MAE and 0.461 eV RMSE of the proposed method; Figure 3(b) shows that

its 1.3% violation rate and 89.1% interval coverage dominate the unconstrained policy values of 6.8% and 82.4%, respectively.

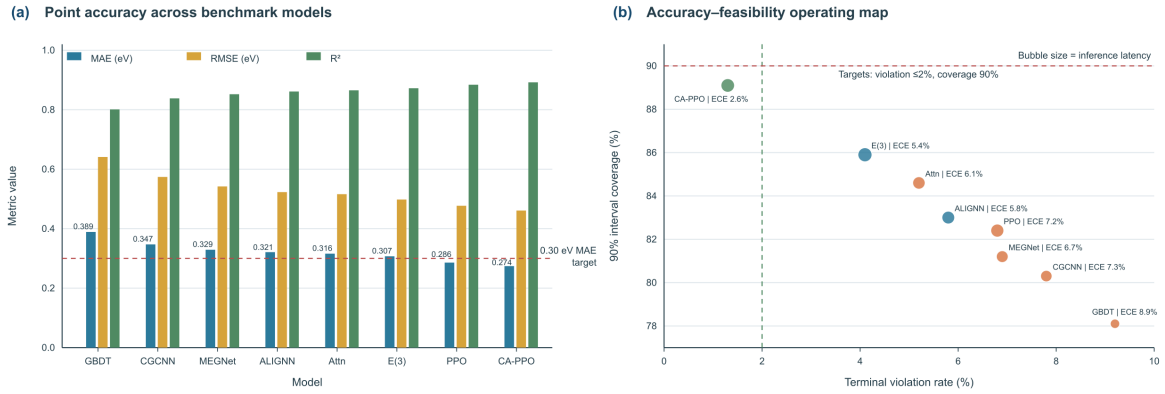


**Figure 3.** Overall predictive performance and reliability. (a) MAE, RMSE, and $R^2$ across eight competing models with 95% confidence intervals. (b) Accuracy-feasibility comparison using violation rate, calibration error, interval coverage, and inference latency.

The descriptor ensemble was still fast at 1.9 ms per crystal but achieved a larger MAE of 0.389 eV; thus, it could not resolve the structural differences caused by small changes in composition. The attention- and equivariant-based baselines reduced the gap to 0.316 and 0.307 eV, respectively. The calibration errors were 0.061 and 0.054, and that of the proposed model was 0.026. Adding PPO without explicit constraints improved the direct graph baseline; therefore, iterative refinement itself has some value, but its low coverage and a larger violation rate indicate that clipping alone does not determine where the trajectory should stop. The limited objective served to constrain the selection.

Ranking quality was examined because screening rarely uses regression values in isolation. For the task of retrieving crystals with predicted gaps between 1.5 and 2.5 eV, the proposed model achieved an area under the precision-recall curve of 0.913, compared with 0.884 for the strongest direct graph baseline. Among the top 500 ranked candidates, 447 fell inside the target window, and 31 of the remaining 53 were review flagged. Spearman correlation between predicted and reference gaps was 0.931. These results show that the reduction in average error was accompanied by better ordering near a practically relevant interval rather than being confined to extreme values.

Error concentration changed as well as the mean. The proposed model reduced the 95th-percentile absolute error from 1.18 to 0.91 eV and lowered the fraction of errors above 1 eV from 7.6% to 4.9%. Uncertainty calibration is essential when predictions are used to rank candidates rather than only to describe a benchmark [26]. Figure 4(a) compares the full absolute-error distributions and locates the median at 0.171 eV; Figure 4(b) shows an empirical cumulative probability of 0.951 at the 1 eV threshold; Figure 4(c) reports observed coverages of 0.503,0.697,0.804,0.891, and 0.946 for nominal levels of 0.50,0.70,0.80,0.90, and 0.95.

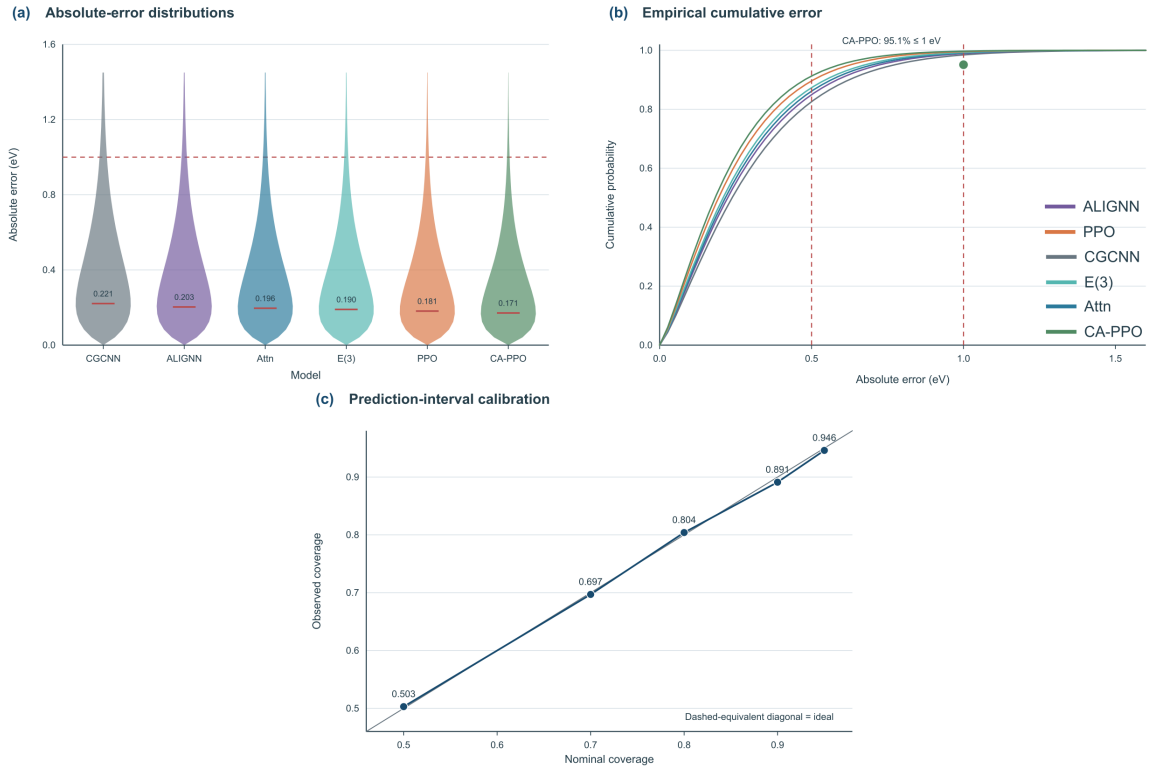


**Figure 4.** Error distribution and uncertainty calibration. (a) Violin and box summaries of absolute error for six models. (b) Empirical cumulative error distributions with 0.5 and 1.0 eV decision thresholds. (c) Nominal-versus-observed interval coverage with calibration gaps and confidence bands.

Measure calibration before and after a validation-only temperature adjustment. The adjustment increased the full model coverage by 1.2 percentage points and reduced its MAE to less than 0.001 eV; thus, the improvement was not due to post-hoc scaling alone. On the other hand, the probabilistic graph baseline needed a 1.37 temperature factor and still did not cover the highest-uncertainty decile. In that decile, the observed RMSE was 1.14 eV and the predicted standard deviation was 1.09 eV for the proposed model; the corresponding baseline values were 1.21 and 0.88 eV. Therefore, the policy set a larger interval for genuinely difficult crystals and did not expand all predictions uniformly.

Constraint monitoring revealed that uncertainty was the most frequently active cost during the first policy step, whereas stoichiometric discrepancy became negligible after supervised warm-up. Adaptive multipliers converged to $0.42, 0.18, 0.31$, and 0.67 for nonnegativity, composition, drift, and uncertainty. This ordering is plausible because uncertainty remains sensitive to chemical novelty after the deterministic constraints are satisfied. Constraint-aware learning has similarly benefited from separating feasibility signals instead of collapsing them into one penalty [27]. Figure 5(a) traces the accuracy-violation Pareto front and marks the selected point at 0.274 eV and $1.3\%$; Figure 5(b) compares four normalized reliability dimensions; Figure 5(c) decomposes the active-constraint share across four refinement steps, where uncertainty falls from $48\%$ to $21\%$; Figure 5(d) shows that accepted predictions have a median interval width of 0.74 eV versus 1.63 eV for review-flagged predictions.

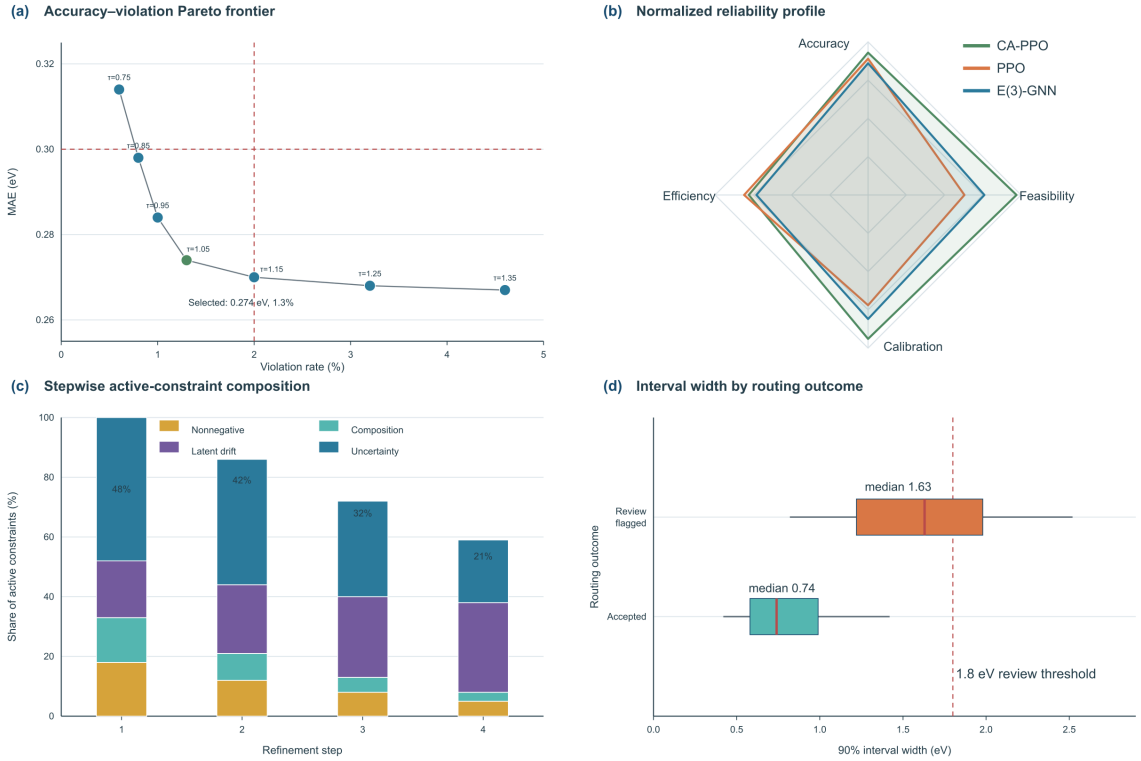


**Figure 5.** Constraint dynamics and operating-point analysis. (a) Accuracy-violation Pareto scatter with the selected dualcontrolled policy. (b) Radar comparison of accuracy, feasibility, calibration, and efficiency. (c) Stepwise composition of active constraints during refinement. (d) Distribution of interval widths for accepted and review-flagged predictions.

The operating-point sweep varied the terminal uncertainty threshold from 0.75 to 1.35 eV and the dual learning rate from 0.002 to 0.02. Aggressive settings reduced violations below $1\%$ but widened the median interval and increased MAE because the policy preferred conservative latent states. Weak settings reproduced the unconstrained trajectory and allowed negative raw gaps in narrow-gap compounds. The selected setting lay near the knee of the frontier: tightening it further from $1.3\%$ to $0.8\%$ violations cost 0.019 eV MAE, whereas relaxing it to $2.5\%$ recovered only 0.004 eV. This asymmetry supports the chosen balance without treating the constraint threshold as a universal constant.

Trajectory traces clarified how the correction was obtained. The first action produced $61\%$ of the total reduction in absolute error and had a median normalized action norm of 0.42. The second and third actions contributed $24\%$ and $11\%$, while the final action contributed $4\%$ and mainly resolved uncertainty or drift costs. Crystals that stopped after two steps had a median error of 0.146 eV. Crystals requiring all four steps had a median error of 0.311 eV and were flagged twice as often, indicating that trajectory length itself is a useful difficulty signal. No single constraint monopolized the reward after epoch 70, which is consistent with stable dual balancing.

## Robustness, Ablation, and Diagnostic Analysis

Chemical transfer remained the hardest setting. On element holdout, MAE increased to 0.325 eV for the proposed method and 0.396 eV for the strongest graph baseline, corresponding to relative degradations of $18.6\%$ and $31.7\%$. On prototype holdout, the values were 0.341 and 0.417 eV. Transfer-learning research indicates that pretraining can mitigate but not eliminate such distribution shifts [28]. Figure 6(a) maps subgroup MAE over six held-out elements and five band-gap intervals, with the largest cell value of 0.512 eV for Bi-containing crystals above 4 eV; Figure 6(b) displays ridgeline error densities for oxide, sulfide, nitride, halide, and intermetallic families, whose median errors range from 0.152 to 0.238 eV.

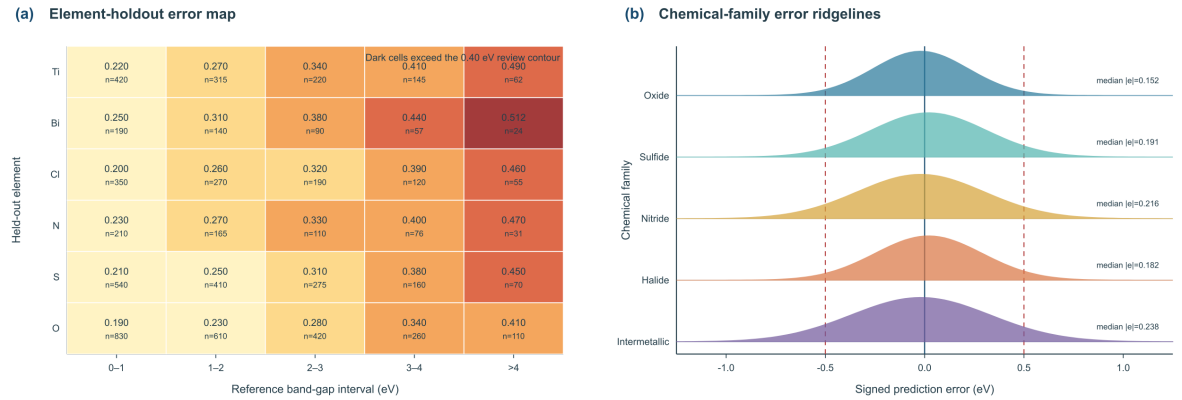


**Figure 6.** Robustness under chemical and structural shift. (a) Heatmap of MAE by held-out element and band-gap interval, annotated with sample counts and a 0.40 eV review contour. (b) Ridgeline distributions of signed error for five chemical families with medians and interquartile ranges.

Transfer errors were not uniform across the target range. For band gaps below 1 eV, element-holdout MAE was 0.208 eV; it increased to 0.296 eV between 1 and $3\mathrm{eV}$, $0.381\mathrm{eV}$ between 3 and 5 eV, and 0.468 eV above 5 eV. The review mechanism responded monotonically, flagging $4.1\%$, $6.8\%$, $11.5\%$, and $19.7\%$ of these groups. A direct graph model produced a similar error gradient but lacked a corresponding routing signal. The largest gains occurred for oxygen and sulfur-containing compounds in which local motifs were familiar but their combined compositions were absent from training.

Ablation isolated the source of improvement. Removing adaptive dual updates raised MAE from 0.274 to 0.292 eV and violation rate from $1.3\%$ to $4.7\%$. Removing uncertainty reward had a smaller MAE effect, 0.283 eV, but reduced $90\%$ coverage to $84.0\%$. Without the latent trust radius, the terminal displacement increased by $63\%$ and seed-to-seed MAE deviation doubled. The supervised anchor was indispensable: its removal produced 0.338 eV MAE and unstable early trajectories. These effects show that the policy component contributes only when it remains coupled to a direct regression objective.

Seed-level variability was modest for the full model: MAE ranged from 0.267 to 0.281 eV and violation rate from 1.1% to $1.6\%$. The fixed-penalty variant was less predictable, with violation rates between $2.9\%$ and $7.2\%$ despite using the same nominal coefficient. Its multiplier could not respond when the uncertainty cost became dominant late in training. Reducing the horizon from four steps to one preserved most of the point accuracy but increased flagged wide-gap crystals, while extending it to six steps added latency without a measurable improvement. Four steps therefore supplied enough corrective capacity before cumulative drift became difficult to regulate.

Training required 14.2 minutes per epoch, compared with 9.6 minutes for the strongest graph baseline, and deterministic inference required 7.8 ms per crystal. Early stopping ended refinement after 2.6 steps on average, recovering 24% of the latency that a fixed four-step rollout would consume. Figure 7(a) shows validation MAE decreasing from 0.612 to 0.279 eV while the violation rate falls from $12.4\%$ to $1.4\%$; Figure 7 (b) ranks six ablations by MAE increase and coverage loss; Figure 7(c) compares predicted uncertainty with observed RMSE across ten bins and yields a calibration slope of 0.96; Figure 7(d) reports element-frequency strata, where MAE rises from 0.243 eV for common elements to 0.356 eV for rare elements while review flags increase from $3.2\%$ to $11.8\%$.

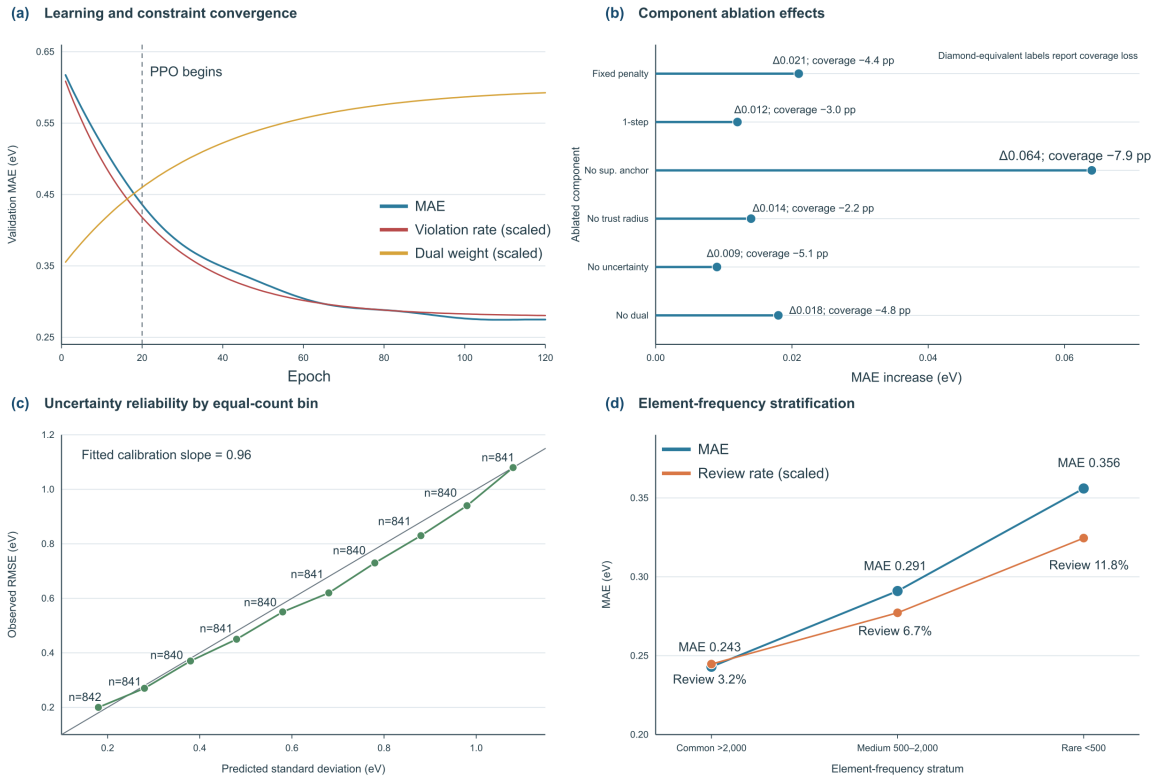


**Figure 7.** Training, ablation, calibration, and frequency-stratified diagnostics. (a) Learning curves for MAE, violation rate, and dual-weight magnitude. (b) Lollipop ablation effects on MAE and coverage. (c) Predicted uncertainty versus observed error with bin counts and a fitted calibration slope. (d) Slopegraph of accuracy and review rate across element-frequency strata.

The other errors were due to composition and structure. Large underestimates were concentrated in wide-gap ionic compounds where the local graphs resemble lower-gap analogues; thus, the four-step local refinement could not always correct for long-range electronic differences. The overestimation of narrow gap intermetallic compounds occurs near the zero-gap boundary. The non-negativity constraint avoided an invalid output, but did not force small positive gaps towards zero and thus avoided a systematic floor. Review routing covered 68% of the predictions with an absolute error exceeding 1 eV and flagged 7.4% of the full test set. This trade-off is suitable for a screening workflow where the flagged candidates receive a high-fidelity calculation instead of being discarded.

Subgroup inspection also showed that the policy did not obtain its average gain by sacrificing one dominant family. Relative to the strongest graph baseline, MAE decreased by 7.8% for oxides, 10.6% for sulfides, 8.4% for nitrides, 6.9% for halides, and 11.7% for intermetallics. Coverage stayed between 87.6% and 90.8% across these families. The rareelement stratum remained less accurate, but its higher review rate made the limitation visible to the downstream user. In a two-stage screening simulation, sending only flagged crystals to the higher-fidelity route reduced the expensive calculation load by 92.6% while retaining most large-error cases for correction.

A cost-sensitive test assigned a higher-fidelity route five, ten or twenty times the cost of neural inference. All three ratios showed that the constrained model had a lower total error-cost than either accepting all the predictions or taking a fixed proportion chosen by uncertainty. The joint signal had the following attributes: some narrow intervals were still flagged due to a high latent drift, and some wider intervals were accepted because the chemistry was familiar and all trajectory costs were stable. This division is practically significant because uncertainty is informative but does not cover all kinds of model failure.

Finally, based on the audit trail, a problem has been identified and this data is no longer aggregated. Out of the 200 randomly sampled flagged crystals, the main terminal reasons were uncertainty in 112 cases, latent drift in 47, nonnegativity in 29, and composition discrepancy in 12. Therefore, a reviewer might weigh unfamiliar chemistry and unstable representation updates differently than boundary predictions. As all reasons are linked

to a measured threshold, this interface does not use free-form confidence labels and keeps subsequent threshold adjustments traceable to a specific verification standard.

## Conclusion

This paper proposes a constraint-aware proximal policy optimization model for predicting the band gap of inorganic crystal data. The model converts static regression into a short, interpretable refinement process and couples clipped policy updates with a probabilistic supervised anchor. Explicit monitors for output range, stoichiometry, latent drift and uncertainty provide an audit trail of the acceptance or rejection of a prediction. The four links in this engineering process are: accuracy, feasibility, calibration and transferability, and runtime.

Several Deficiencies Remain. Constraints are relatively loose, only ensuring the general physical and statistical feasibility of the data, and their tolerances still need to be verified in a specific database. Short local trajectories cannot show the effect of long-range orbital interactions fully, and grouped benchmark splits cannot reproduce all the chemistry in prospective discovery. The extra actor, critic and uncertainty passes also increase the training cost compared with direct graph regression.

Future work will introduce symmetry-preserving equivariant encoders, multi-fidelity labels, and constraints based on related formation-energy and stability predictions. Prospectively evaluate the new crystals that have been calculated or experimentally determined to see how the review flag changes under actual circumstances. Another is to obtain the constraint tolerance from the downstream screening utility and keep an auditable upper bound, thus applying the same policy framework to more material-design decisions.

**Author Contributions**

Serkan Bozkurt contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Zeynep Demir contributes to validation, analysis, investigation, data collection, draft preparation. Lokman Öpçin contributes to analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

**Funding**

This research received no specific financial support from any funding agency.

**Institutional Review Board Statement**

Not applicable.

## References

[1] Dunn, A., Wang, Q., Ganose, A., Dopp, D., & Jain, A. (2020). Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. npj Computational Materials, 6(1), 138. https://doi.org/10.1038/s41524-020-00406-3

[2] Goodall, R. E. A., & Lee, A. A. (2020). Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nature Communications, 11(1), 6280. https://doi.org/10.1038/s41467-020-19964-7

[3] Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A. V., Granata, V., Gargiulo, F., Borelli, M., Uhrin, M., Huber, S. P., Zoupanos, S., Adorf, C. S., Andersen, C. W., Schütt, O., Pignedoli, C. A., Passerone, D., VandeVondele, J., Schulthess, T. C., Smit, B., Pizzi, G., & Marzari, N. (2020). Materials Cloud, a platform for open computational science. Scientific Data, 7(1), 299. https://doi.org/10.1038/s41597-020-00637-5

[4] Himanen, L., Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Ranawat, Y. S., Gao, D. Z., Rinke, P., & Foster, A. S. (2020). DScribe: Library of descriptors for machine learning in materials science. Computer Physics Communications, 247, 106949. https://doi.org/10.1016/j.cpc.2019.106949

[5] Musil, F., Grisafi, A., Bartók, A. P., Ortner, C., Csányi, G., & Ceriotti, M. (2021). Physics-Inspired Structural Representations for Molecules and Materials. Chemical Reviews, 121(16), 9759-9815. https://doi.org/10.1021/acs.chemrev.1c00021

[6] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., & Kozinsky, B. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature Communications, 13(1), 2453. https://doi.org/10.1038/s41467-022-29939-5

[7] Magar, R., Wang, Y., & Barati Farimani, A. (2022). Crystal twins: self-supervised learning for crystalline material property prediction. npj Computational Materials, 8(1), 231. https://doi.org/10.1038/s41524-022-00921-5

[8] Tran, A., Tranchida, J., Wildey, T., & Thompson, A. P. (2020). Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys. The Journal of Chemical Physics, 153(7), 074705. https://doi.org/10.1063/5.0015672

[9] Zhao, Y., Cui, Y., Xiong, Z., Jin, J., Liu, Z., Dong, R., & Hu, J. (2020). Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions. ACS Omega, 5(7), 3596-3606. https://doi.org/10.1021/acsomega.9b04012

[10] Louis, S. Y., Zhao, Y., Nasiri, A., Wang, X., Song, Y., Liu, F., & Hu, J. (2020). Graph convolutional neural networks with global attention for improved materials property prediction. Physical Chemistry Chemical Physics, 22(32), 18141-18148. https://doi.org/10.1039/d0cp01474e

[11] Choudhary, K., & DeCost, B. (2021). Atomistic Line Graph Neural Network for improved materials property predictions. npj Computational Materials, 7(1), 185. https://doi.org/10.1038/s41524-021-00650-1

[12] Karamad, M., Magar, R., Shi, Y., Siahrostami, S., Gates, I. D., & Barati Farimani, A. (2020). Orbital graph convolutional neural network for material property prediction. Physical Review Materials, 4(9), 093801. https://doi.org/10.1103/physrevmaterials.4.093801

[13] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3(6), 422-440. https://doi.org/10.1038/s42254-021-00314-5

[14] Liu, Y., Ding, J., & Liu, X. (2020). IPO: Interior-Point Policy Optimization under Constraints. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 4940-4947. https://doi.org/10.1609/aaai.v34i04.5932

[15] Tian, C., Liu, A., Huang, G., & Luo, W. (2022). Successive Convex Approximation Based Off-Policy Optimization for Constrained Reinforcement Learning. IEEE Transactions on Signal Processing, 70, 1609-1624. https://doi.org/10.1109/tsp.2022.3158737

[16] Pan, E., Petsagkourakis, P., Mowbray, M., Zhang, D., & Rio-Chanona, E. A. D. (2021). Constrained model-free reinforcement learning for process optimization. Computers & Chemical Engineering, 154, 107462. https://doi.org/10.1016/j.compchemeng.2021.107462

[17] Meng, W., Zheng, Q., Shi, Y., & Pan, G. (2022). An Off-Policy Trust Region Policy Optimization Method With Monotonic Improvement Guarantee for Deep Reinforcement Learning. IEEE Transactions on Neural Networks and Learning Systems, 33(5), 2223-2235. https://doi.org/10.1109/tnnls.2020.3044196

[18] Garrity, K. F., & Choudhary, K. (2021). Database of Wannier tight-binding Hamiltonians using high-throughput density functional theory. Scientific Data, 8(1), 106. https://doi.org/10.1038/s41597-021-00885-z

[19] Griesemer, S. D., Ward, L., & Wolverton, C. (2021). High-throughput crystal structure solution using prototypes. Physical Review Materials, 5(10), 105003. https://doi.org/10.1103/physrevmaterials.5.105003

[20] Bartel, C. J., Trewartha, A., Wang, Q., Dunn, A., Jain, A., & Ceder, G. (2020). A critical examination of compound stability predictions from machine-learned formation energies. npj Computational Materials, 6(1), 97. https://doi.org/10.1038/s41524-020-00362-y

[21] Omee, S. S., Louis, S. Y., Fu, N., Wei, L., Dey, S., Dong, R., Li, Q., & Hu, J. (2022). Scalable deeper graph neural networks for high-performance materials property prediction. Patterns, 3(5), 100491. https://doi.org/10.1016/j.patter.2022.100491

[22] Cheng, G., Gong, X. G., & Yin, W. J. (2022). Crystal structure prediction by combining graph network and optimization algorithm. Nature Communications, 13(1), 1492. https://doi.org/10.1038/s41467-022-29241-4

[23] Caldeira, J., & Nord, B. (2021). Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. Machine Learning: Science and Technology, 2(1), 015002. https://doi.org/10.1088/2632-2153/aba6f3

[24] Acar, P. N. (2021). Recent progress of uncertainty quantification in small-scale materials science. Progress in Materials Science, 117, 100723. https://doi.org/10.1016/j.pmatsci.2020.100723

[25] Park, C. W., & Wolverton, C. (2020). Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Physical Review Materials, 4(6), 063801. https://doi.org/10.1103/physrevmaterials.4.063801

[26] Soleimany, A. P., Amini, A., Goldman, S., Rus, D., Bhatia, S. N., & Coley, C. W. (2021). Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. ACS Central Science, 7(8), 1356-1367. https://doi.org/10.1021/acscentsci.1c00546

[27] Hu, Z., & Gong, W. (2022). Constrained evolutionary optimization based on reinforcement learning using the objective function and constraints. Knowledge-Based Systems, 237, 107731. https://doi.org/10.1016/j.knosys.2021.107731

[28] Gupta, V., Choudhary, K., Tavazza, F., Campbell, C., Liao, W. K., Choudhary, A., & Agrawal, A. (2021). Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. Nature Communications, 12(1), 6595. https://doi.org/10.1038/s41467-021-26921-5