Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: X. Yu, *Digital Discovery*, 2025, DOI: 10.1039/D6DD00153J.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D6DD00153J

Uncertainty-Aware Active Learning Reveals Reliability Limits in Lead-Free Halide Perovskite Screening

Xiyao Yu*

School of Chemistry, Chemical Engineering and Biotechnology (CCEB), Nanyang Technological University, Singapore

Email: xyu029@e.ntu.edu.sg

Abstract

The discovery of stable, lead-free perovskite materials for photovoltaic applications is challenged by the vast chemical space of possible compositions and by the systematic inaccuracies inherent in high-throughput density functional theory (DFT) calculations. In particular, widely used semi-local functionals such as PBE are known to underestimate band gaps, while data-driven screening workflows often treat all machine learning predictions as equally reliable.

In this work, we present an uncertainty-aware active learning framework for the screening of lead-free halide perovskites that explicitly distinguishes between reliable predictions and regions of limited model knowledge. By expanding the search space beyond ideal cubic perovskites to include distorted, vacancy-ordered, and mixed-anion structures, we intentionally address a more realistic and challenging materials landscape. An ensemble regression model is employed to predict DFT band gaps while quantifying epistemic uncertainty arising from data sparsity and model disagreement.

To correct the systematic bias of PBE-calculated band gaps, we introduce a statistically validated, stratified PBE-to-experiment calibration scheme based on experimentally characterized benchmark compounds. This calibration aligns theoretical predictions with experimental trends without artificially improving predictive accuracy. The resulting screening reveals recurring patterns in candidate selection, including the frequent emergence of heavy d-electron halides, which we identify as potential false positives arising from functional limitations and feature-level abstractions.

Rather than claiming definitive material discoveries, this study demonstrates how uncertainty quantification and active learning can be used to expose blind spots in conventional screening pipelines and to prioritize materials for higher-fidelity

View Article Online
DOI: 10.1039/D6DD00153J

electronic structure calculations. The proposed framework provides a principled strategy for allocating computational and experimental resources in the search for lead-free perovskite photovoltaics.

1. Introduction

1.1 Lead-Free Perovskites and the Limits of Conventional Screening

Metal halide perovskites have rapidly emerged as one of the most promising classes of photovoltaic materials, achieving power conversion efficiencies exceeding 26% within a decade of development.[1] Their exceptional optoelectronic properties, combined with low-cost solution processability, have positioned them as strong contenders to complement or replace conventional silicon-based technologies. However, the most efficient perovskite absorbers rely almost exclusively on lead-based compositions, raising significant concerns regarding toxicity, environmental impact, and long-term regulatory compliance.[2]

Efforts to replace lead have led to the exploration of a broad range of alternative chemistries, including tin-, germanium-, bismuth-, and antimony-based compounds, as well as double perovskites and vacancy-ordered derivatives.[3] While these substitutions alleviate toxicity concerns, they dramatically expand the accessible chemical and structural space.[4] Beyond ideal cubic phases, real lead-free perovskites often exhibit octahedral distortions, reduced dimensionality, mixed anions, or ordered vacancies, all of which complicate both theoretical modeling and experimental validation.[5]

High-throughput DFT calculations have become a cornerstone of materials discovery in this context, enabling rapid evaluation of thousands of candidate compounds. However, the sheer size of the search space, combined with the computational cost of accurate electronic structure methods, necessitates the use of approximate functionals and simplified screening criteria. As a result, many screening pipelines operate at the edge of their physical validity.

1.2 Systematic Errors and Overconfidence in Data-Driven Workflows

A critical limitation of high-throughput screening is the systematic underestimation of band gaps by semi-local DFT functionals such as PBE.[6] This bias is not random but highly structured, depending on chemical composition, orbital character, and the presence of heavy elements where spin–orbit coupling plays a dominant role.[7] While this issue is widely recognized, many screening studies implicitly assume that relative trends remain meaningful, or that machine learning models trained on DFT data can compensate for these deficiencies.

View Article Online
DOI: 10.1039/D6DD00153J

Machine learning has indeed demonstrated remarkable success in accelerating materials discovery by learning nonlinear relationships between composition and target properties. However, most composition-based models lack explicit information about local coordination, orbital hybridization, or symmetry-driven selection rules.[8] Consequently, these models may produce predictions that appear numerically reasonable while being physically misleading. More importantly, conventional screening workflows often lack mechanisms to identify when a prediction is unreliable due to data sparsity or extrapolation beyond the training domain.

In such cases, failure does not manifest as large random errors but as confidently wrong predictions. This overconfidence is particularly problematic in expanded chemical spaces, where novel bonding motifs or electronic structures are poorly represented in existing datasets. Without explicit uncertainty quantification, these blind spots remain hidden and can systematically bias candidate selection.

1.3 Motivation and Scope of This Work

The central motivation of this study is not to identify a single "optimal" lead-free perovskite absorber, but to develop a screening framework that explicitly acknowledges and manages uncertainty. We argue that, in realistic materials discovery workflows, the ability to distinguish between reliable predictions and speculative regions of chemical space is as important as predictive accuracy itself.[9]

To this end, we construct an uncertainty-aware active learning pipeline that combines ensemble machine learning, systematic DFT-to-experiment calibration, and physics-informed screening criteria. By deliberately expanding the search space beyond high-symmetry cubic perovskites, we expose the limits of conventional composition-based models and investigate how these limits manifest in large-scale screening results.

The contributions of this work are threefold. First, we demonstrate how ensemble-derived epistemic uncertainty can be used as a diagnostic tool to identify regions of chemical space where model predictions are inherently unreliable. Second, we introduce a data-driven calibration strategy that corrects systematic DFT bias without conflating calibration with improved predictive power. Third, we show how active learning can guide the prioritization of high-fidelity electronic structure calculations, transforming apparent screening failures into actionable insight.

2. Methodology

2.1 Data Acquisition and Physics-Based Filtering

The initial materials dataset was obtained from the Materials Project database using the Materials Project API (v2024). To focus on lead-free halide perovskites while

View Article Online
DOI: 10.1039/D6DD00153J

avoiding overly restrictive assumptions, we applied a set of chemistry- and stoichiometry-based filters rather than enforcing idealized crystallographic symmetry.

Candidate compounds were required to contain halide anions (Cl, Br, I) and to satisfy general perovskite-derived stoichiometries, including single perovskites (ABX_3), double perovskites ($A_2BB'X_6$), vacancy-ordered phases, and related distorted or cluster-based structures.[10] This relaxed definition intentionally expands the search space beyond high-symmetry cubic phases, reflecting the structural diversity commonly observed in experimentally synthesized lead-free perovskites.

All structures and computed properties were programmatically retrieved using the Materials Project next-generation API via the `mp-api` Python client (MPRester). An API key was supplied through an environment variable (`MATERIALS_PROJECT_API_KEY`) loaded from a local `.env` file. The query was restricted to lead-free, halide, perovskite-like compositions, and the resulting dataset was exported as formulas paired with DFT properties (e.g., PBE band gap and energy above hull) for downstream modeling.

Additional physical constraints were applied to exclude chemically unreasonable compositions, including excessive elemental complexity, non-halide anions, or extreme stoichiometric imbalance. Thermodynamic stability was assessed using the energy above hull (E_hull) provided by Materials Project, and compounds with E_hull ≤ 0.05 eV/atom were retained for downstream screening. After filtering, a total of 1,117 lead-free halide perovskite–like materials were included in the final dataset.

The dataset is exclusively inorganic; organic–inorganic hybrids (methylammonium and formamidinium variants) are not represented in the Materials Project database under the query used and were therefore excluded. The composition of the final dataset is as follows: 84.2% of compounds contain a single B-site metal (single perovskite), while 15.8% contain two distinct B-site metals (double perovskite). By halide anion, 44.9% are chlorides, 31.0% bromides, and 24.1% iodides. By B-site chemistry, the dominant families are coinage metals (Cu/Ag/Au, 24.6%), post-transition metals (Ga/In/Tl, 21.8%), Group-15 metals (Bi/Sb, 17.1%), transition metals (14.9%), and Group-14 metals (Sn/Ge, 4.8%). Of the 1,117 compounds, 678 (60.7%) satisfy the thermodynamic stability criterion (E_hull ≤ 0.05 eV/atom). All 1,117 compounds have non-zero PBE band gaps: the query applied a lower band gap filter of 0.3 eV, so all materials in the dataset are semiconducting at the PBE level, with PBE gaps ranging from 0.30 to 3.98 eV (mean 2.09 eV). Note that the 20 calibration benchmark compounds (Table S2) were curated independently from primary literature and are not subject to this band gap threshold; the benchmark set therefore includes compounds such as $CsSnI_3$ with PBE gaps slightly below 0.30 eV (PBE gap = 0.258 eV), which are excluded from the ML training set but included in the calibration dataset.

View Article Online
DOI: 10.1039/D6DD00153J

2.2 Feature Engineering

Each compound was represented using a composition-based feature vector derived from elemental properties, without explicit structural or orbital descriptors. Elemental attributes were collected from the periodic table and included atomic number, atomic mass, Pauling electronegativity, and valence electron characteristics. For each property, statistical descriptors such as the weighted mean, standard deviation, and range were computed based on the compound's stoichiometry.

This approach resulted in a total of 29 compositional features per material. While composition-based representations enable efficient large-scale screening, they inherently abstract away local coordination geometry, orbital hybridization, and symmetry-driven selection rules.[11] Rather than treating this limitation as a defect, we explicitly acknowledge it as a defining characteristic of the screening regime investigated in this work, and its implications are examined in the subsequent analysis.

In the accompanying codebase, we also evaluate an expanded feature set (35+ descriptors) that augments basic composition statistics with simple perovskite heuristics. In particular, we introduce three low-cost "proxy" features (proxy tolerance factor, proxy octahedral factor, and an electronegativity mismatch term) to help flag chemically implausible candidates that otherwise appear attractive under purely compositional models.

The composition-only feature set is used here as a deliberate stress test of a widely adopted screening regime, rather than as an attempt to maximise predictive accuracy. Our goal is to map where this commonly used abstraction remains informative and where it becomes unreliable; in this framing, model failures and uncertainty spikes are diagnostic signals that delimit the domain of credible composition-driven screening.

2.3 Ensemble Learning and Uncertainty Quantification

The ML surrogate is not intended to improve on existing DFT values, but to extend predictions to compositions absent from the Materials Project, to quantify compound-level epistemic uncertainty via ensemble disagreement, and to provide a uniform feature-space representation that supports active learning acquisition.

To predict DFT-calculated band gaps while quantifying model uncertainty, we employed an ensemble regression framework composed of multiple independently trained models.[12] Each ensemble member was trained on a bootstrapped subset of

View Article Online
DOI: 10.1039/D6DD00153J

the training data, using tree-based regressors selected for their robustness to nonlinear relationships and heterogeneous feature scales.

Implementation details: the final pipeline uses a heterogeneous ensemble of 15 tree-based regressors. Every third member is a GradientBoostingRegressor, and the remaining members are RandomForestRegressors. Each model is trained on a bootstrap resample of the training set with independent random seeds; input features are standardized using a StandardScaler fit on the training data. Predictive uncertainty is estimated as the standard deviation of the ensemble predictions (epistemic uncertainty).

For a given material, the ensemble prediction is defined as the mean of the individual model outputs, while the epistemic uncertainty is estimated from the standard deviation of these predictions.[13] This uncertainty reflects model disagreement arising from limited or uneven data coverage and serves as an indicator of predictive reliability rather than numerical noise.

The dataset was randomly split into training (80%) and test (20%) subsets, and five-fold cross-validation was used to assess model stability. All reported uncertainties correspond to epistemic uncertainty captured by the ensemble and do not include aleatoric uncertainty inherent to experimental measurements.

Although PBE band gaps for all 1,117 compounds in the dataset are already available from the Materials Project, the ensemble uncertainty estimates serve purposes beyond simple lookup. First, the trained surrogate is intended for prospective deployment on novel compositions absent from the Materials Project, where no DFT values exist; the present dataset validates the framework before such application. Second, even within the known dataset, high ensemble disagreement flags compounds where composition-only features inadequately represent the underlying electronic structure—precisely the systems (heavy d-electron, strong SOC) where PBE labels are themselves least trustworthy. Third, uncertainty on known compounds identifies which materials would benefit most from re-evaluation with higher-fidelity methods (hybrid functionals, GW), guiding resource allocation for validation rather than re-prediction.

2.4 Screening Criteria and Active Learning Strategy

Virtual screening was conducted using a multi-criteria filtering strategy designed to balance photovoltaic relevance, thermodynamic stability, and predictive confidence. Candidate materials were required to satisfy the following conditions:

1. A calibrated experimental band gap within 1.34 ± 0.25 eV, targeting the optimal range for single-junction photovoltaic absorbers.
2. Thermodynamic stability with $E_{hull} \leq 0.05$ eV/atom.

View Article Online
DOI: 10.1039/D6DD00153J

3. Ensemble-predicted uncertainty within a predefined window (0.05–0.50 eV), excluding both trivially well-known regions and extreme extrapolation.

In the implementation, these thresholds are enforced through a configurable `Config` object (target_bandgap=1.34 eV, bandgap_tolerance=0.25 eV, max_e_above_hull=0.05 eV/atom, and an uncertainty window of $\sigma \in [0.05, 0.50]$ eV). Candidate lists are then refined using simple physics-aware heuristics that reject clearly implausible compositions (e.g., wide-gap LiF/CuF motifs, polyatomic anions, unphysical predicted gaps outside 0.1–4.0 eV, or $\sigma > 0.5$ eV).

Rather than selecting final materials, this framework defines an active learning region that prioritizes compounds for higher-fidelity electronic structure calculations. Materials exhibiting moderate to high uncertainty but physically reasonable properties are treated as high-value targets for subsequent validation, maximizing information gain per computational cost.[14]

All components of the pipeline are unit-tested, including feature extraction, calibration, uncertainty estimation, and end-to-end execution.

3. Model Performance and Calibration

3.1 Predictive Performance on DFT Band Gaps

The ensemble model achieved a coefficient of determination (R^2) of 0.620 on the independent test set, with a root mean squared error (RMSE) of 0.625 eV and a mean absolute error (MAE) of 0.494 eV. The observed error scale should be interpreted as an intrinsic limitation imposed by compositional abstraction and chemical diversity, rather than insufficient model optimization.[15]

While these metrics are modest compared to models trained on narrowly defined, high-symmetry perovskite datasets, they reflect the increased complexity of the present screening space. The inclusion of distorted structures, mixed-anion compounds, and heavy-element chemistries substantially broadens the range of bonding motifs and electronic structures, imposing an intrinsic limit on the predictive accuracy achievable with composition-only descriptors.

Importantly, the consistency between cross-validation and test performance indicates that the model generalizes within the defined chemical space and does not rely on favorable data partitioning or overfitting.

View Article Online
DOI: 10.1039/D6DD00153J

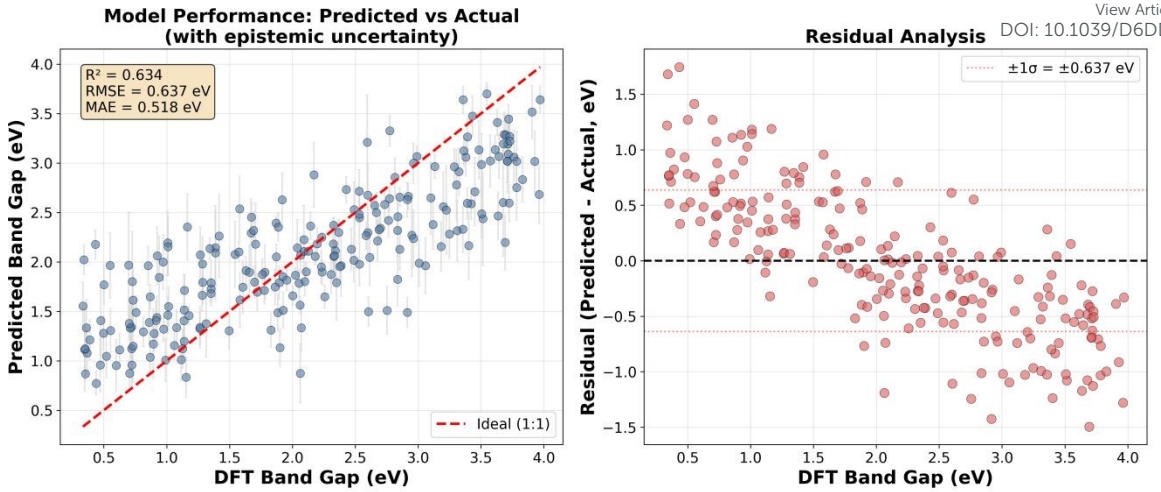

Figure 1. Ensemble model performance on the held-out test set (20% of 1,117 compounds, n = 224). (a) Predicted vs. DFT-calculated PBE band gap; shaded bands indicate ±1σ epistemic uncertainty from the 15-member ensemble. (b) Residual distribution with annotated RMSE and R^2 values. All predictions are generated from composition-only features (29 descriptors).

3.2 Systematic Bias in PBE Band Gaps

Semi-local DFT functionals such as PBE are well known to underestimate band gaps, particularly in systems containing heavy elements, localized d or f electrons, or strong spin–orbit coupling effects.[16] This underestimation is systematic rather than random, resulting in a structured bias that propagates directly into machine learning models trained on PBE-level data.

Consequently, raw ML predictions of DFT band gaps cannot be directly interpreted as experimental values. Instead of attempting to implicitly correct this bias within the regression model, we treat PBE underestimation as a separate, physically motivated problem.

3.3 Data-Driven Calibration to Experimental Band Gaps

The calibration relates PBE-calculated band gaps to reported experimental values for the 20-compound benchmark set:

$E_{\text{exp}} = 0.885\, E_{\text{PBE}} + 0.966$ (in-sample $R^2 = 0.655$, $p = 1.6 \times 10^{-5}$, n = 20; bootstrap 95% CI on slope: [0.635, 1.249])

View Article Online
DOI: 10.1039/D6DD00153J

We note that this $R^2 = 0.655$ is the in-sample fit quality of the linear calibration on 20 reference compounds — it should not be confused with the ML ensemble's test-set $R^2 = 0.620$ (Section 3.1), which evaluates predictive accuracy on PBE-level band gaps of held-out materials. The calibration R^2 is a measure of how well the PBE-to-experiment linear trend is captured across diverse chemical families; the relatively modest value reflects genuine scatter arising from chemistry-dependent DFT errors rather than poor data quality.

To account for chemistry-dependent biases, additional stratified calibrations were performed for specific material classes, including tin-based, germanium-based, and double perovskite compounds. These subgroup fits reveal quantitatively distinct behaviors: while the global model suggests a scaling factor of ~0.885, tin-based (Sn) and germanium-based (Ge) perovskites exhibit slopes closer to unity (1.049 and 1.078, respectively), showing the role of specific electronic structures and relativistic effects in governing functional-dependent errors, confirming that systematic bias is not uniform across the entire chemical space. Calibration is applied post hoc and does not alter model training or uncertainty estimation; therefore, predictive uncertainty remains referenced to the original PBE learning task.

It is important to state explicitly what the calibration step is and is not intended to do. First, calibration is not used to improve the intrinsic predictive performance of the machine learning model, which is trained and evaluated solely on PBE-level targets. Instead, calibration serves as a scale-alignment layer that enables physically meaningful interpretation of screening thresholds (e.g., the photovoltaic-relevant band gap window) in experimental units. Second, we do not interpret the calibrated values as high-confidence experimental predictions for individual compounds; rather, they provide a bias-corrected reference scale for population-level screening patterns. Third, calibration does not alter the epistemic uncertainty estimated by the ensemble, which remains anchored to the original learning task (PBE band gaps). In other words, calibration changes the meaning of the x-axis in downstream screening plots, but it does not "sharpen" the model nor convert ensemble dispersion into a calibrated probabilistic error bar.

Robustness of the global calibration parameters was further assessed via LOOCV and bootstrap resampling (Fig. S4).

In practice, we assess calibration robustness by (i) leave-one-out cross-validation (LOOCV) over the calibration set to quantify how individual reference points influence the fitted slope/intercept, and (ii) bootstrap resampling (10,000 iterations) to estimate 95% confidence intervals for the calibration parameters. We additionally evaluate how different calibration strategies (uncalibrated, global, and stratified) change the screened candidate set using Jaccard overlap metrics.

View Article Online
DOI: 10.1039/D6DD00153J

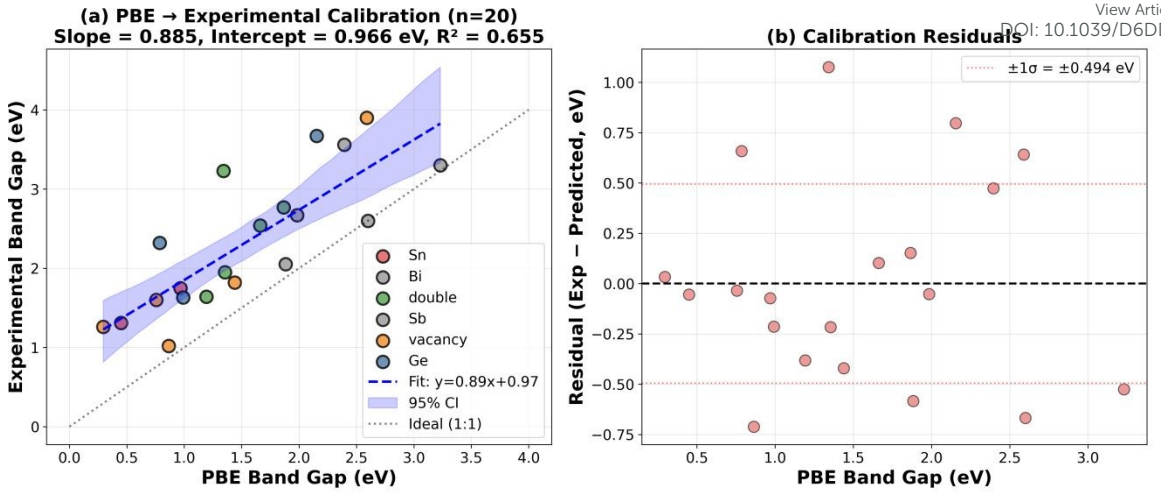

Figure 2. PBE-to-experimental band gap calibration. (a) Linear fit for 20 inorganic benchmark compounds (Table S2), coloured by material class (Sn-based, Ge-based, double perovskite, vacancy-ordered); dashed line = fitted calibration ($E_{exp} = 0.885 \times E_{PBE} + 0.966$ eV, $R^2 = 0.655$), shaded region = bootstrap 95% CI. (b) Calibration residuals vs. PBE band gap; dotted lines mark ±1σ.

3.4 Calibration of Ensemble Uncertainty

To evaluate whether ensemble dispersion is quantitatively meaningful, we performed post hoc uncertainty calibration diagnostics on the held-out test set (Fig. S1). Absolute error shows a positive but limited association with predicted uncertainty (Spearman $\rho \approx 0.40$; Pearson $r \approx 0.35$), indicating utility for relative risk ranking rather than pointwise error prediction.[19] Consistent with this, nominal prediction intervals constructed as $\mu \pm k\sigma$ substantially under-cover the observed targets (e.g., ~50% empirical coverage for a nominal 95% interval), and the expected calibration error is large (ECE ≈ 0.30 eV). Therefore, in this work uncertainty is interpreted primarily as a marker of epistemic limitation rather than as a calibrated probabilistic error bar, motivating its use for prioritizing validation rather than making coverage claims.

We quantify uncertainty reliability using several complementary diagnostics: (i) correlation between predicted σ and absolute residuals (Pearson and Spearman); (ii) empirical coverage of nominal predictive intervals assuming a normal error model; (iii) calibration error metrics such as Expected Calibration Error (ECE) and RMSCE; and (iv) proper scoring rules including negative log-likelihood (NLL) and continuous ranked probability score (CRPS). These metrics help distinguish "useful" uncertainty signals from purely relative rankings.

Accordingly, throughout this work we use uncertainty primarily for relative risk ranking and for identifying regions of chemical space where the screening abstraction is likely to break down. We avoid making coverage or confidence-interval claims based on the raw ensemble dispersion. This conservative interpretation is consistent

View Article Online
DOI: 10.1039/D6DD00153J

with the broader theme of the study: when label fidelity is limited (PBE-level systematic error) and representation is intentionally coarse (composition-only), the most reliable output of the pipeline is not a precise error bar for each compound, but a reliability-aware partition of chemical space into "well-supported" versus "speculative" regions.

3.5 Generalization Beyond Random Splits

Because random train–test splits can overestimate performance in chemically correlated datasets, we evaluated generalization under structured partitions (Fig. S2; Table S3). While random splits yield moderate accuracy (MAE ≈ 0.50 eV), performance degrades markedly under B-site grouped splits (MAE ≈ 0.72 eV) and can fail catastrophically in leave-one-element-out tests for specific elements. Cluster-based splits show similar degradation.[20] These results indicate that the model primarily interpolates within well-represented chemistries and that extrapolation across sparsely sampled elemental subspaces constitutes a dominant reliability limit. Accordingly, screening outcomes should be interpreted as a reliability-aware map rather than a definitive candidate ranking.

In addition to a standard random split baseline, we benchmark generalization under: (a) GroupKFold splits that hold out entire B-site element groups, (b) a leave-one-element-out (LOEO) protocol where the model is tested on each B-site element with at least 10 samples, and (c) an unsupervised cluster split (KMeans with 5 clusters on standardized features) that tests transfer across chemically distinct regions of feature space.

4. Screening Results: Patterns, Not Winners

4.1 Distribution of Predicted Band Gaps and Uncertainty

Applying the calibrated band gap window (1.34 ± 0.25 eV) and thermodynamic stability criterion ($E_{hull} \leq 0.05$ eV/atom) yields 1 strictly screened candidate (with 124 ML-priority compounds for active learning prioritisation) from the initial pool of 1,117 lead-free halide perovskite–like compounds.[21] Rather than interpreting this subset as a list of definitive photovoltaic absorbers, we analyze the statistical and chemical patterns underlying its composition.

View Article Online
DOI: 10.1039/D6DD00153J

Figure 3 illustrates the joint distribution of calibrated band gaps and ensemble-derived uncertainty. Two distinct regimes emerge. Materials with low predictive uncertainty are clustered within well-sampled regions of chemical space, dominated by conventional halide perovskites and closely related chemistries. In contrast, materials located near the target band gap window but exhibiting elevated uncertainty typically belong to less explored chemical families, including mixed-anion systems, vacancy-ordered structures, and compounds containing late transition metals.

This separation highlights an important distinction between numerical agreement with screening criteria and predictive reliability. While both low- and high-uncertainty materials may satisfy band gap and stability thresholds, their physical credibility differs substantially. Consequently, uncertainty serves as a critical contextual variable rather than a secondary metric.

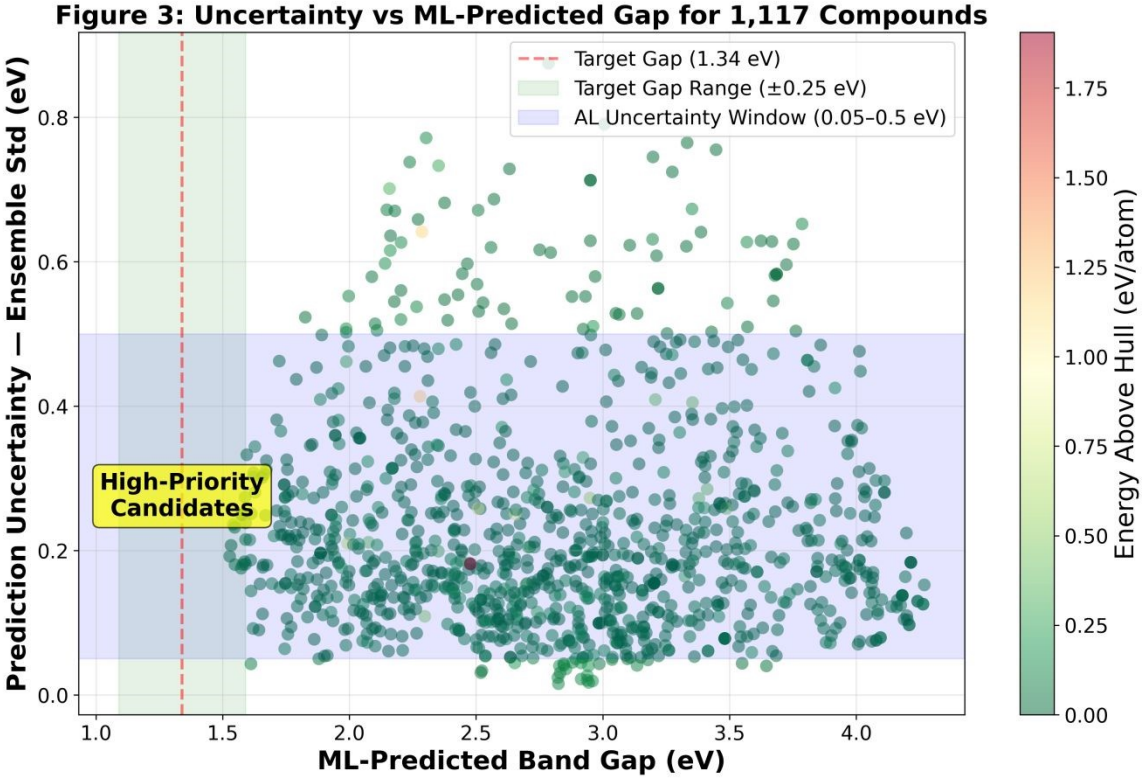

Figure 3. Joint distribution of calibrated band gap and ensemble epistemic uncertainty for 1,117 lead-free halide perovskites. Each point represents one compound; the dashed rectangle marks the photovoltaic target window (1.09–1.59 eV). Colour encodes thermodynamic stability (E_hull). High-uncertainty compounds within the target window (upper-right quadrant) define the active learning region prioritised for higher-fidelity validation.

4.2 Recurring Chemical Motifs in Screened Candidates

Inspection of the screened candidates reveals the recurrent appearance of compounds containing heavy d-electron elements such as Au, Pd, Pt, and Hg.[22] These materials consistently exhibit low PBE-calculated band gaps that, after calibration, fall within the photovoltaic target range.[23]

This trend is not interpreted here as evidence of superior optoelectronic performance. Instead, it reflects a confluence of known methodological limitations. Semi-local DFT functionals systematically underestimate band gaps in systems with localized d states and strong relativistic effects, while composition-based machine learning models lack explicit descriptors for orbital localization, symmetry-forbidden transitions, and spin–orbit coupling. When combined with linear calibration, these effects can amplify apparent agreement with target band gap criteria.

As a result, such materials represent a class of *potential false positives by construction*. Their frequent emergence in the screened set underscores the importance of uncertainty-aware interpretation and cautions against treating calibrated predictions as definitive indicators of photovoltaic suitability.

From a physical standpoint, these heavy d-electron halides are also a particularly challenging class for gap-based screening because their band-edge character and optical activity can be strongly affected by relativistic effects, d-state localization, and symmetry-dependent transition matrix elements. In such systems, an apparently "good" band gap can coincide with weak optical absorption near the onset (e.g., symmetry-forbidden or low-oscillator-strength transitions) or with unfavorable transport characteristics arising from heavy effective masses and flat bands. These effects are not explicitly represented in a composition-only model and are only partially treated at the PBE level. Therefore, the repeated selection of heavy d-electron chemistries should be interpreted less as an endorsement of photovoltaic promise and more as an indicator of a reliability boundary where higher-fidelity electronic-structure analysis (e.g., SOC-aware hybrid functionals or GW) becomes decisive.

To probe whether this heavy d-electron enrichment is an artifact of feature-level abstraction, we introduced additional proxy descriptors and repeated the screening. In the baseline screened set, heavy d-electron compounds account for 21/78 (26.9%), whereas under the proxy-augmented feature set they account for 23/91 (25.3%). While the proxy features slightly reduce the heavy d fraction, the effect is modest, suggesting that the observed enrichment is driven primarily by underlying functional and representation limitations rather than a single missing descriptor. (Fig. S5)

5. Active Learning: Where Computation Should Go Next

View Article Online
DOI: 10.1039/D6DD00153J

5.1 Defining the Active Learning Region

Building on the uncertainty-aware screening results, we define an active learning region characterized by three simultaneous conditions: (i) calibrated band gaps within the photovoltaic target range, (ii) thermodynamic stability consistent with experimental feasibility, and (iii) moderate to high epistemic uncertainty.

High uncertainty is not interpreted as low quality, but as high expected information gain under constrained computational budgets.

Applying these criteria yields 113 materials prioritized for further investigation. Notably, this set includes compounds that would be excluded by conventional confidence-based filtering, as well as materials that are numerically promising but chemically unconventional. The purpose of this selection is not to expand the candidate list indiscriminately, but to identify materials whose validation would most effectively reduce model uncertainty and improve future screening reliability.

This set is distinct from the broader 124-compound ML-priority pool referenced in Section 4 (Fig. S5), which was defined by band gap and stability criteria alone without the strict uncertainty window. The 113-compound active learning region applies the additional constraint $\sigma \in [0.05, 0.50]$ eV, excluding the 11 compounds that either fall in highly confident ($\sigma < 0.05$ eV) or maximally uncertain ($\sigma > 0.50$ eV) regimes.

Table 1. Summary of candidate sets defined throughout the screening workflow.

Candidate Set	Screening Criteria	n
Full dataset	Lead-free halide perovskites from Materials Project (no stability or gap filter)	1,117
Stable subset	Full dataset + $E_{hull} \leq 0.05$ eV/atom	678
Strictly screened	Band gap (1.34 ± 0.25 eV after calibration) + stability + $\sigma < 0.05$ eV	1
ML-priority pool (baseline features)	Band gap + stability, any uncertainty level; 29 compositional features	124
ML-priority pool (proxy features)	Band gap + stability, any uncertainty level; 35+ features with proxy descriptors	110
Baseline screened (heavy-d analysis)	Band gap + stability; subset used in Section 4.3 heavy d-electron fraction analysis	78
Proxy-augmented screened (heavy-d analysis)	Same as above with proxy features	91
Active learning region	Band gap + stability + $\sigma \in [0.05, 0.50]$ eV (moderate-to-high uncertainty)	113

All band gap criteria are applied after global PBE-to-experiment calibration ($E_{exp} = 0.885 \times E_{PBE} + 0.966$ eV). Uncertainty thresholds refer to ensemble standard deviation σ on PBE-level predictions.

5.2 Prioritization for High-Fidelity Electronic Structure Calculations

Within the active learning region, materials are ranked using a composite priority score that balances uncertainty magnitude, proximity to the target band gap, and thermodynamic stability. High-priority candidates typically exhibit band gaps near the center of the target window while residing in sparsely sampled regions of chemical space.

Formally, each of the three components is min–max normalised to [0, 1] over the active learning region, and the final score is computed as:

$$\text{priority score} = 0.4 \times \bar{u} + 0.4 \times \tilde{g} + 0.2 \times \tilde{s}$$

where $\bar{u}$ = MinMaxScaler(σ) rewards high epistemic uncertainty (maximising expected information gain), $\tilde{g}$ = MinMaxScaler(−|E_cal − E_target|) rewards proximity to the photovoltaic target band gap (E_target = 1.34 eV), and $\tilde{s}$ = MinMaxScaler(−E_hull) rewards thermodynamic stability. The 0.4/0.4/0.2 weighting reflects the primary objective of uncertainty-guided prioritisation, with band gap proximity as an equally weighted photovoltaic relevance criterion and stability as a secondary feasibility filter.

For these materials, higher-level electronic structure methods such as hybrid functionals (HSE06), many-body perturbation theory (GW), and explicit inclusion of spin–orbit coupling are expected to provide decisive corrections to both band gap values and qualitative electronic character.[25] Importantly, the goal of such calculations is not merely to confirm or refute individual candidates, but to generate informative data points that refine the model's understanding of previously underrepresented chemistries.

5.3 Retrospective Active Learning Simulation

The active learning framework functions as a resource allocation strategy rather than a material selection tool: by identifying where predictions are uncertain yet physically plausible, it transforms uncertainty into an actionable signal for prioritising validation.[26] To quantify the practical impact of uncertainty-guided acquisition, we performed a retrospective pool-based active learning simulation with one acquisition round under a fixed group-based test split (Fig. S3). Starting from $n_0 = 200$ labeled

View Article Online
DOI: 10.1039/D6DD00153J

samples and querying K = 50 additional points, target-oriented acquisition yields a larger average MAE reduction (~4.4%) than random querying (~1.9%), whereas pure uncertainty querying shows a smaller improvement (~2.1%). However, across 10 random seeds the advantage does not reach statistical significance (paired test $p \approx 0.12$).

The simulation code evaluates multiple acquisition functions, including pure uncertainty sampling (selecting the highest-σ points), a target-aware rule that trades off σ against distance to a target band gap, and expected improvement. Performance is aggregated over repeated random seeds, and the final-round gains relative to random acquisition are assessed using paired t-tests.

Across six acquisition rounds, the active learning strategies did not outperform random sampling (best relative change: −1.9% at Round 6), supporting the conclusion that model limitations are dominated by representation and out-of-distribution effects rather than data acquisition heuristics.[24]

These results support our framing of active learning as a resource allocation strategy under uncertainty, with measurable gains likely requiring additional rounds and/or higher-fidelity labels. We view this outcome as consistent with the role of active learning in the present setting rather than as a contradiction. When the labeling function is itself approximate and systematically biased (PBE band gaps) and the representation is intentionally coarse (composition-only), the achievable improvement from a single acquisition round is expected to be modest and may exhibit strong dependence on the particular split and seed. In this regime, the primary value of active learning is not to deliver dramatic accuracy gains, but to prioritize informative validation targets that reduce epistemic blind spots under a constrained computational budget. This supports our framing of active learning as a resource allocation strategy: it helps decide where high-fidelity calculations are most valuable, even when aggregate test-set metrics improve only marginally in short simulations.

To benchmark against GP-based active learning, which represents the dominant approach in the literature, we additionally evaluated a Gaussian Process regressor (Matérn-5/2 kernel, scikit-learn GaussianProcessRegressor, 3 restarts) with Upper Confidence Bound (UCB, $\kappa = 2$) and Expected Improvement (EI) acquisition functions under the identical retrospective simulation setup ($n_0 = 200$, K = 50, 1 acquisition round, 10 random seeds, group-based test split). GP + UCB achieved a mean MAE reduction of +1.12% ± 3.36% relative to the initial model (paired t-test vs random sampling: $p = 0.944$); GP + EI achieved +1.14% ± 3.37% ($p = 0.959$ vs random). Neither GP-based acquisition strategy significantly outperformed random sampling, consistent with our ensemble-based results ($p \approx 0.12$). This concordance across fundamentally different uncertainty models — ensemble disagreement versus GP posterior variance — confirms that the limitation is inherent to the chemical space coverage and dataset size rather than to the specific choice of surrogate model. The GP baseline results are summarised in Fig. S8.

View Article Online
DOI:10.1039/D6DD00153J

6. Limitations and Physical Blind Spots

Despite the methodological rigor of the uncertainty-aware screening framework, several important limitations remain. These limitations are not incidental but arise from fundamental trade-offs inherent to large-scale, composition-based materials discovery workflows. Explicitly identifying these blind spots is essential for the correct interpretation of the screening results and for guiding future methodological improvements.

6.1 Composition-Based Representation and Electronic Structure Effects

The present model relies exclusively on composition-derived features and does not explicitly encode crystal symmetry, local coordination geometry, or orbital hybridization.[27] While this representation enables efficient screening across thousands of compounds, it inherently limits the model's ability to capture electronic structure phenomena that are sensitive to local bonding environments.

In particular, systems with strong d- or f-electron localization, parity-forbidden optical transitions, or symmetry-protected band edges may exhibit optical and transport properties that are poorly correlated with composition alone.[28] As observed in the screening results, such effects can lead to the recurrent selection of materials that satisfy numerical band gap criteria while remaining physically ambiguous. These cases highlight the boundary beyond which structural or orbital descriptors become indispensable.

The dominance of electronegativity dispersion and tolerance-factor-like descriptors highlights the extent to which the model relies on indirect geometric and bonding proxies, rather than explicit electronic structure.

6.2 Functional-Dependent Errors and Relativistic Effects

The screening workflow is ultimately grounded in PBE-level DFT data, which introduces systematic biases that cannot be fully eliminated through machine learning or calibration.[29] Although linear calibration effectively corrects the average underestimation of band gaps, it does not account for chemistry-specific deviations arising from spin–orbit coupling, self-interaction errors, or exchange–correlation functional inadequacies.

This limitation is particularly relevant for compounds containing heavy elements, where relativistic effects can significantly reshape band dispersion and orbital

View Article Online
DOI: 10.1039/D6DD00153J

character. In such cases, calibrated predictions should be interpreted as first-order estimates rather than quantitative substitutes for hybrid-functional or many-body calculations.

6.3 Absence of Defect Physics and Kinetic Stability

Another key blind spot of the present screening framework is the absence of explicit defect physics. Photovoltaic performance is strongly influenced by defect formation energies, charge-state transition levels, and defect tolerance, none of which are captured by bulk band gap predictions or thermodynamic stability metrics.[30]

Similarly, kinetic stability and degradation pathways are not addressed. Materials that are thermodynamically stable at zero temperature may nevertheless undergo rapid degradation under illumination, moisture, or thermal stress.[31] These effects are especially prominent in tin- and germanium-based perovskites and cannot be inferred from the present screening criteria.[32]

6.4 Interpretation of Uncertainty as a Boundary Marker

While ensemble-derived uncertainty provides a powerful diagnostic signal, it does not distinguish between different physical origins of uncertainty.[20] High uncertainty may arise from sparse training data, unconventional chemistry, or genuine breakdowns of the underlying modeling assumptions. Consequently, uncertainty should be interpreted as a marker of epistemic limitation rather than as a direct measure of prediction quality.

Recognizing this distinction is critical to avoiding misinterpretation of uncertainty as a ranking criterion. In the present framework, uncertainty is used to guide validation efforts rather than to suppress or promote specific materials.

Alternative uncertainty quantification strategies (Monte Carlo dropout, quantile regression, conformal prediction) may be appropriate in other settings; here, ensemble variance is chosen because it directly reflects model disagreement under limited data coverage, aligning with the goal of identifying reliability boundaries rather than producing strict probabilistic confidence intervals.

A key implication of these results is that screening outcomes should be interpreted as patterns and boundary markers rather than as definitive winners. In particular, candidates that satisfy numerical band gap and stability criteria while exhibiting elevated uncertainty represent high-leverage opportunities for targeted validation. Such compounds are not “better” candidates by default; instead, they are candidates for which the current abstraction (PBE-level labels + composition-only features)

View Article Online
DOI: 10.1039/D6DD00153J

provides insufficient physical resolution. Importantly, this includes chemically unconventional regions where known failure modes are plausible, such as heavy-element halides with strong SOC or systems where optical activity depends sensitively on symmetry and orbital character. In this sense, the most actionable output of the workflow is a reliability-aware map that highlights where higher-fidelity calculations are most likely to change screening conclusions and improve future model coverage.

7. Conclusion

The primary output of this study is not a ranked candidate list, but a reliability-aware map of chemical space. In this work, we have developed an uncertainty-aware active learning framework for the screening of lead-free halide perovskites that explicitly confronts the limitations of conventional high-throughput materials discovery pipelines. By expanding the search space beyond idealized cubic structures and incorporating ensemble-based uncertainty quantification, we demonstrate how data-driven screening can move beyond static candidate selection toward informed exploration of chemical space.

Rather than claiming definitive material discoveries, our results reveal systematic patterns in screening outcomes that reflect both the strengths and blind spots of PBE-based machine learning models. In particular, the recurrent emergence of certain chemical motifs underscores the necessity of treating calibrated predictions with caution and of contextualizing numerical agreement with physical plausibility.

The active learning strategy presented here reframes uncertainty as an actionable signal that guides the allocation of computational resources. By prioritizing materials located in sparsely sampled and uncertain regions of chemical space, the workflow enables targeted deployment of higher-fidelity electronic structure methods, accelerating knowledge acquisition while avoiding overconfident conclusions.

More broadly, this study illustrates that the success of data-driven materials discovery depends not only on predictive accuracy, but on the ability to recognize and manage uncertainty. As materials challenges increasingly involve complex chemistries and competing physical mechanisms, uncertainty-aware frameworks such as the one presented here will be essential for integrating machine learning with first-principles theory and experiment in a principled and transparent manner.

View Article Online
DOI: 10.1039/D6DD00153J

Data and Code Availability

The code supporting this study (data acquisition from the Materials Project, feature engineering, ensemble training with uncertainty estimation, calibration analyses, active-learning simulations, and GP baseline comparison) is openly available on GitHub at:

https://github.com/PillowSoprano/Lead-Free-Halide-Perovskite-Screening-with-Active-Learning-Public-

The exact version used in this work (v1.0.0, commit 9a76319) is permanently archived on Zenodo:

PillowSoprano. (2026). Lead-Free Halide Perovskite Screening with Active Learning: Revised submission – Digital Discovery (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.19249710

The repository contains all scripts, configuration files, and a pinned requirements.txt. Running python improved_perovskite_screening.py reproduces the end-to-end screening workflow and writes all processed datasets, figures, and tables to ./outputs as described in the repository documentation.

Source materials data were retrieved from the Materials Project database via its public API (https://next-gen.materialsproject.org/api; https://materialsproject.org), which is freely accessible under a Creative Commons Attribution 4.0 International licence. A Materials Project API key (available free of charge upon registration) is required and should be provided via the MATERIALS_PROJECT_API_KEY environment variable..

Reference

[1] Toward Lead-Free Perovskite Solar Cells (ACS Energy Letters, 2016)
[2] Lead-Free Halide Perovskite Materials and Optoelectronic Devices (Advanced Functional Materials, 2023)
[3] Challenges and Strategies Toward Long-Term Stability of Lead-Free Tin Perovskites (Communications Materials, 2022)
[4] Recent Developments of Lead-Free Halide Double Perovskites (Materials Advances, 2022)
[5] Lead-Free Perovskites for Next-Generation Applications (Materials Advances, 2025)
[6] Predicting Band Gaps of Oxide Perovskites Using DFT and ML (Physical Review B, 2022)

View Article Online
DOI: 10.1039/D6DD00153J

[7] DFT-PBE Band Gap Correction Using Machine Learning (Computational Materials Science, 2024)
[8] Methods for Comparing Uncertainty Quantifications for Material Properties (Machine Learning: Science and Technology, 2020)
[9] Machine Learning in Materials Design and Discovery: Examples from Perovskites (Physical Review Materials, 2018)
[10] Prediction and Screening of Lead-Free Double Perovskite Optoelectronic Materials (Journal of Physical Chemistry Letters, 2024)
[11] Limitations of Machine Learning Models When Predicting Compounds (npj Computational Materials, 2022)
[12] Neural Network Ensembles for Band Gap Prediction (Computational Materials Science, 2024)
[13] Band-Gap Regression with Architecture-Optimized Message-Passing Neural Networks (Chemistry of Materials, 2024)
[14] Active Meta-Learning for Predicting and Selecting Perovskite Crystallization (Journal of Chemical Physics, 2022)
[15] Band Gap Predictions of Double Perovskite Oxides Using Machine Learning (Communications Materials, 2023)
[16] Putting Error Bars on Density Functional Theory (Scientific Reports, 2024)
[17] Accurate and Efficient Band Gap Predictions of Metal Halide Perovskites (Physical Chemistry Chemical Physics, 2017)
[18] Uncertainty Prediction for Machine Learning Models of Material Properties (ACS Omega, 2021)
[19] Machine Learning Materials Properties with Accurate Predictions and Uncertainty (OSTI, 2024)
[20] Benefits and Limits of Using ML for Materials Discovery (SemiEngineering, 2023)
[21] A High-Throughput Computational Dataset of Halide Perovskite Alloys (Digital Discovery, 2023)
[22] Apparent Defect Densities in Halide Perovskite Thin Films (ACS Energy Letters, 2021)
[23] Electron Perovskite Oxides on Magnetism (University of Colorado, 2022)
[24] Opportunities for Machine Learning to Accelerate Halide-Perovskite Commercialization (Matter, 2022)
[25] Efficient Dataset Generation for Machine Learning Perovskite Alloys (arXiv, 2025)
[26] Active Learning Training Strategy for Predicting O Adsorption Free Energy (ChemCatChem, 2021)
[27] Improving Machine-Learning Models in Materials Science Through Better Data (Materials Today, 2024)
[28] Composition and Structure Analyzer/Featurizer for Explainable Materials Discovery (Digital Discovery, 2025)
[29] Electronic Structure of (Organic-)Inorganic Metal Halide Perovskites (Advanced Theory and Simulations, 2021)
[30] Intrinsic Defect Physics in Indium-Based Lead-Free Halide Double Perovskites (Journal of Physical Chemistry Letters, 2017)
[31] Atomic-Scale Understanding on the Physics and Control of Intrinsic Point Defects in

Lead-Free Perovskite Solar Cells (Applied Physics Reviews, 2021)
[32] Defect Passivation in Lead-Free CsSnI3 Perovskite Nanowires (Nano-Micro Letters, 2022)

View Article Online
DOI: 10.1039/D6DD00153J

View Article Online
DOI: 10.1039/D6DD00153J

Data Availability Statement

The code supporting this study (data acquisition from the Materials Project, feature engineering, ensemble training with uncertainty estimation, calibration analyses, active-learning simulations, and GP baseline comparison) is openly available on GitHub at:

https://github.com/PillowSoprano/Lead-Free-Halide-Perovskite-Screening-with-Active-Learning-Public-

The exact version used in this work (v1.0.0, commit 9a76319) is permanently archived on Zenodo:

PillowSoprano. (2026). Lead-Free Halide Perovskite Screening with Active Learning: Revised submission – Digital Discovery (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.19249710

The repository contains all scripts, configuration files, and a pinned requirements.txt. Running python improved_perovskite_screening.py reproduces the end-to-end screening workflow and writes all processed datasets, figures, and tables to ./outputs as described in the repository documentation.

Source materials data were retrieved from the Materials Project database via its public API (https://next-gen.materialsproject.org/api; https://materialsproject.org), which is freely accessible under a Creative Commons Attribution 4.0 International licence. A Materials Project API key (available free of charge upon registration) is required and should be provided via the MATERIALS_PROJECT_API_KEY environment variable.