

Cross-laboratory generalization failure in perovskite solar cell machine learning: A diagnostic protocol and evaluation threshold

Minseong Kim^a, Jinho Lee^a, Hye W. Chun^b, Eun Seo Shin^c, Hyosung Choi^c,
Dong-Won Kang^d, Jong H. Kim^b, Shujuan Huang^a, Jincheol Kim^{a,*}

^a Macquarie University, School of Engineering, Sydney, NSW 2109, Australia

^b Department of Molecular Science and Technology, Ajou University, Suwon 16499, Republic of Korea

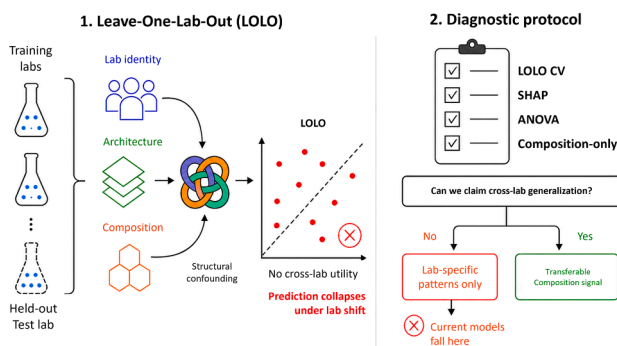
^c Department of Chemistry, Research Institute for Convergence of Basic Sciences, Research Institute for Natural Sciences, Hanyang University, Seoul, Republic of Korea

^d School of Energy Systems Engineering, Chung-Ang University, Seoul, 06974, Republic of Korea

HIGHLIGHTS

- Within-lab R^2 in perovskite machine learning can conceal cross-lab failure.
- Within-lab R^2 of 0.44 collapses to leave-one-lab-out R^2 of -3.04 in 819 cells.
- The collapse is reproducible across five model families and three laboratories.
- A diagnostic protocol with explicit cross-lab threshold (LOLO $R^2 > 0$) is provided.
- Lab-aware evaluation must become the default, not an optional addition.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:

Perovskite solar cells
Machine learning
Cross-laboratory generalization
Leave-one-laboratory-out cross-validation
SHAP analysis

ABSTRACT

Machine learning (ML) models for perovskite solar cell power conversion efficiency are commonly evaluated using within-laboratory cross-validated R^2 , with high values often interpreted as evidence that composition–efficiency relationships have been learned. Here, we show that this metric can diverge sharply from cross-laboratory predictive performance on the same dataset. Using a controlled three-laboratory cohort of 819 p-i-n perovskite solar cells and five model families spanning linear, regularized, and ensemble approaches, within-laboratory R^2 values of up to 0.44 collapse to leave-one-laboratory-out (LOLO) R^2 values as low as -3.05 . Cross-laboratory prediction errors are several times larger than those obtained by predicting the mean, and the collapse persists across all five model classes. Restricting the feature matrix to composition descriptors only partially mitigates the failure (LOLO Random Forest (RF) R^2 improves from -3.05 to -2.10), indicating that the problem is not a feature-engineering artifact but a structural property of multi-laboratory datasets in which composition, device architecture, and laboratory identity are intrinsically confounded. To systematically expose and evaluate this failure mode, we introduce a diagnostic framework combining within-laboratory, full-dataset, and LOLO cross-validation with SHapley Additive exPlanations stability analysis and analysis-of-variance decomposition. Under this framework, meaningful cross-laboratory generalization requires LOLO $R^2 > 0$; the

* Corresponding author.

E-mail address: jincheol.kim@mq.edu.au (J. Kim).

<https://doi.org/10.1016/j.egyai.2026.100856>

Available online 25 July 2026

2666-5468/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

composition-only LOLO RF R^2 of -2.10 reported here indicates that current methods fail to satisfy this criterion in the present cohort. Similar dissociation is likely across materials ML studies trained on pooled multi-laboratory datasets. Lab-aware evaluation should therefore be a standard component of ML reporting, not an optional addition.

1. Introduction

Halide perovskite solar cells have achieved certified single-junction power conversion efficiencies (PCEs) exceeding 27% [1,2], driven largely by systematic composition engineering within the ABX₃ lattice [3]. The accessible chemistry, encompassing A-site (Cs, formamidinium (FA), methylammonium (MA)), B-site (Pb, Sn), and X-site (I, Br, Cl) substitutions, defines a vast compositional space that is increasingly explored using machine learning (ML) as an alternative to one-variable-at-a-time experimentation [4–9]. Reported within-laboratory cross-validated performance in this literature is typically high. Gradient-boosted trees, random forests, and Gaussian process regression frequently achieve R^2 values that, under conventional evaluation protocols, appear to indicate that underlying composition–efficiency relationships have been successfully learned. Recent studies have applied ML to simulation-derived datasets, for example coupling first-principles or SCAPS-1D device simulations with ML to predict efficiency and optimize compositions in lead-free and double-perovskite absorbers [10–12], as well as the first-principles design of related wide-bandgap and tandem photovoltaic materials [13]. These efforts illustrate the growing role of ML in perovskite photovoltaics but rely on computationally generated data for individual material systems, whereas whether models transfer across independent experimental laboratories, the focus of the present work, remains unaddressed.

The interpretation of these metrics, however, rests on an assumption that has received limited empirical examination. In practice, research groups differ systematically in their choice of hole transport layers (HTLs), electron transport layers (ETLs), deposition equipment, processing atmospheres, and aging protocols, and these experimental choices covary with the compositional spaces explored by individual laboratories [14–18]. If ML models encode laboratory-specific process signatures rather than transferable composition–efficiency relationships, they may appear highly predictive within the training environment while failing to generalize to new facilities. To our knowledge, the extent to which this failure mode affects perovskite ML has not been quantitatively benchmarked.

The distinction between composition and device architecture is therefore central to framing this problem. In principle, perovskite composition can be defined solely by ABX₃ chemistry, and a composition–PCE relationship that remains valid across laboratories would represent a transferable physical relationship of practical utility. Device architecture, by contrast, including hole and electron transport layers, electrode materials, and their associated energy-level alignments, is constrained by each laboratory’s equipment, processing workflow, and material inventory, and therefore varies systematically across facilities. Consequently, an ML model that depends strongly on architectural descriptors must extrapolate to combinations absent from the training data, whereas a model based solely on composition descriptors may, in principle, generalize wherever the same composition can be realized. Whether composition descriptors retain sufficient predictive information once architectural variables are removed is the central empirical question addressed in this work.

Related challenges have been *recognized* and addressed in adjacent disciplines. In genomic and clinical ML, batch effects between studies were shown in the mid-2000s to inflate within-study performance estimates, leading to the adoption of leave-one-study-out evaluation and explicit batch-correction procedures as community standards [19–21]. Materials informatics has similarly examined issues of cross-validation

leakage, distribution shift, and applicability domains [22–27]. However, lab-aware evaluation has not yet become standard practice in perovskite ML. The closest community-scale resource, the Perovskite Database Project [28], aggregates devices from hundreds of laboratories and provides the infrastructure necessary for laboratory-stratified evaluation. Nevertheless, the structural consequences of cross-laboratory evaluation in perovskite ML remain insufficiently quantified, and the field currently lacks a benchmark against which transfer-learning or batch-correction approaches can be assessed.

To address this gap, we analyze a controlled three-laboratory perovskite device cohort. The use of an internally curated cohort, rather than a community-scale database, is intentional, as each device is associated with complete experimental provenance, enabling rigorous laboratory-stratified evaluation that is not feasible when laboratory identity is incomplete or unreliable. Emerging autonomous experimentation platforms and self-driving laboratories are increasingly producing multi-campaign datasets with similarly documented provenance [29, 30], making the diagnostic framework developed here directly relevant to those settings and, more broadly, to multi-laboratory materials ML problems in which laboratory-specific processing covaries with the prediction target. The framework integrates leave-one-laboratory-out (LOLO) cross-validation, analysis-of-variance (ANOVA) decomposition, and SHapley Additive exPlanations (SHAP) stability analysis.

Here, we present a three-laboratory diagnostic study of cross-laboratory generalization in perovskite ML. We show quantitatively that within-laboratory cross-validated R^2 does not constitute evidence of cross-laboratory predictive utility in the present cohort. We further identify the structural confounding among laboratory identity, processing architecture, and composition that drives this dissociation, and we propose a diagnostic framework with an explicit performance criterion for evaluating cross-laboratory perovskite ML. Under this framework, LOLO $R^2 > 0$ represents the minimum requirement for any method claiming cross-laboratory generalization, as $R^2 \leq 0$ indicates predictive performance no better than the mean PCE across laboratories. The composition-only LOLO RF R^2 of -2.10 obtained here demonstrates that current feature engineering and model-selection approaches do not satisfy this criterion in the present cohort. More broadly, the proposed framework is intended for direct application to other multi-laboratory datasets to support comparable and reproducible reporting across the field.

2. Results and discussion

2.1. Dataset overview

We assembled a cohort of 819 p-i-n perovskite solar cells from three independent research groups (Lab A, $n = 177$; Lab B, $n = 496$; Lab C, $n = 146$). Fig. 1a shows the PCE distributions: Lab A has a mean PCE of 18.6% (standard deviation (SD) 1.6%), Lab B 17.8% (SD 4.4%), and Lab C 22.2% (SD 1.4%). The higher mean and narrower distribution observed for Lab C reflect its more homogeneous near-iodide compositions and consistent HTL processing, whereas Lab B exhibits a broader distribution containing a subpopulation of low-performing devices absent from the other laboratories. Fig. 1b demonstrates substantial divergence in bandgap distributions across laboratories. Lab B occupies a narrow range of 1.530–1.550 eV, whereas Lab A spans 1.530–1.720 eV, reflecting Br fractions up to 0.298. All Lab C devices in the present cohort report a bandgap of 1.570 eV. Fig. 1c plots Br fraction against PCE and reveals a clear laboratory–composition confound: Lab A

occupies substantially higher Br content (mean $x_{Br} \approx 0.16$), while Labs B and C cluster near zero. Consequently, any model learning a negative Br–PCE relationship from Lab A data would primarily capture a laboratory-specific correlation rather than a transferable composition law. Fig. 1d further shows that preferred top-electrode thicknesses differ systematically between laboratories, introducing additional architectural heterogeneity confounded with laboratory identity. Collectively, the four panels in Fig. 1 demonstrate that the three laboratories constitute compositionally and architecturally distinct subpopulations.

2.2. Within-lab versus cross-lab performance dissociation

Within-laboratory 20-fold cross-validation (CV) on the p-i-n training set ($n = 654$ after holdout exclusion) reveals substantial heterogeneity across laboratories (Table 1, Fig. 2a). Lab B achieves RF $R^2 = 0.443 \pm 0.195$ and Ridge $R^2 = 0.433 \pm 0.192$ ($n = 394$), substantially above chance. In contrast, Lab A and Lab C yield near-zero or negative values: Lab A RF $R^2 = -0.241 \pm 0.515$ and Ridge $R^2 = -0.292 \pm 0.595$ ($n = 142$), while Lab C RF $R^2 = -0.552 \pm 0.818$ and Ridge $R^2 = -0.516 \pm 0.752$ ($n = 118$). The negative within-laboratory R^2 in Lab C likely reflects the combination of limited sample size and low PCE variance (SD 1.4%, range 16–25%), which restricts the discriminative signal available to the models. Fig. 2b shows the fold-level R^2 distributions underlying these means, confirming that Lab B performance is consistently above chance, whereas Lab A and Lab C exhibit large fold-to-fold variability spanning positive and negative values. Fig. 2c provides a diagnostic control for the Lab B result. Appending the *target_derived_quality_flag* covariate to the otherwise identical 45-feature matrix increases RF R^2 from 0.443 to 0.728, demonstrating how a quality flag derived from the regression target can artificially inflate within-laboratory performance. This covariate is excluded from all primary analyses and is discussed

Table 1

Within-laboratory 20-fold cross-validated performance on the post-holdout training set using the 45-feature derived block with *target_derived_quality_flag* excluded.

Lab	Model	N (train)	CV R^2 Mean	CV R^2 Std
A	RF	142	−0.241	0.515
A	Ridge	142	−0.292	0.595
B	RF	394	0.443	0.195
B	Ridge	394	0.433	0.192
C	RF	118	−0.552	0.818
C	Ridge	118	−0.516	0.752

separately in Section 2.3 as a target-leakage diagnostic.

Full-dataset 20-fold CV yields higher R^2 values than any single-laboratory analysis (Table 2): RF achieves 0.54 ± 0.11 and Ridge 0.50 ± 0.16 using the same 45-feature block. This apparent improvement arises because Lab B contributes 60% of training devices (394 of 654), ensuring that standard k-fold splitting includes substantial Lab B representation in every fold. As a result, the models learn Lab B-specific distributions without being required to *generalize* across laboratories. Standard CV therefore treats confounded laboratory structure as though it were a transferable composition–efficiency relationship.

LOLO cross-validation exposes this failure directly. After excluding the locked 165-device holdout set, three-fold GroupKFold LOLO evaluation was performed on the remaining 654 training devices and yielded a mean RF R^2 of -3.05 (Lab B held out: -0.29 ; Lab A: -2.34 ; Lab C: -6.51). Ridge regression produces a median R^2 of -7.05 across the same folds; the mean is dominated by an extreme Lab A fold ($R^2 = -91.7$) caused by linear extrapolation into a previously unseen architectural space, making the median a more robust summary statistic. Full fold-level results are provided in Supplementary Information Table S1.

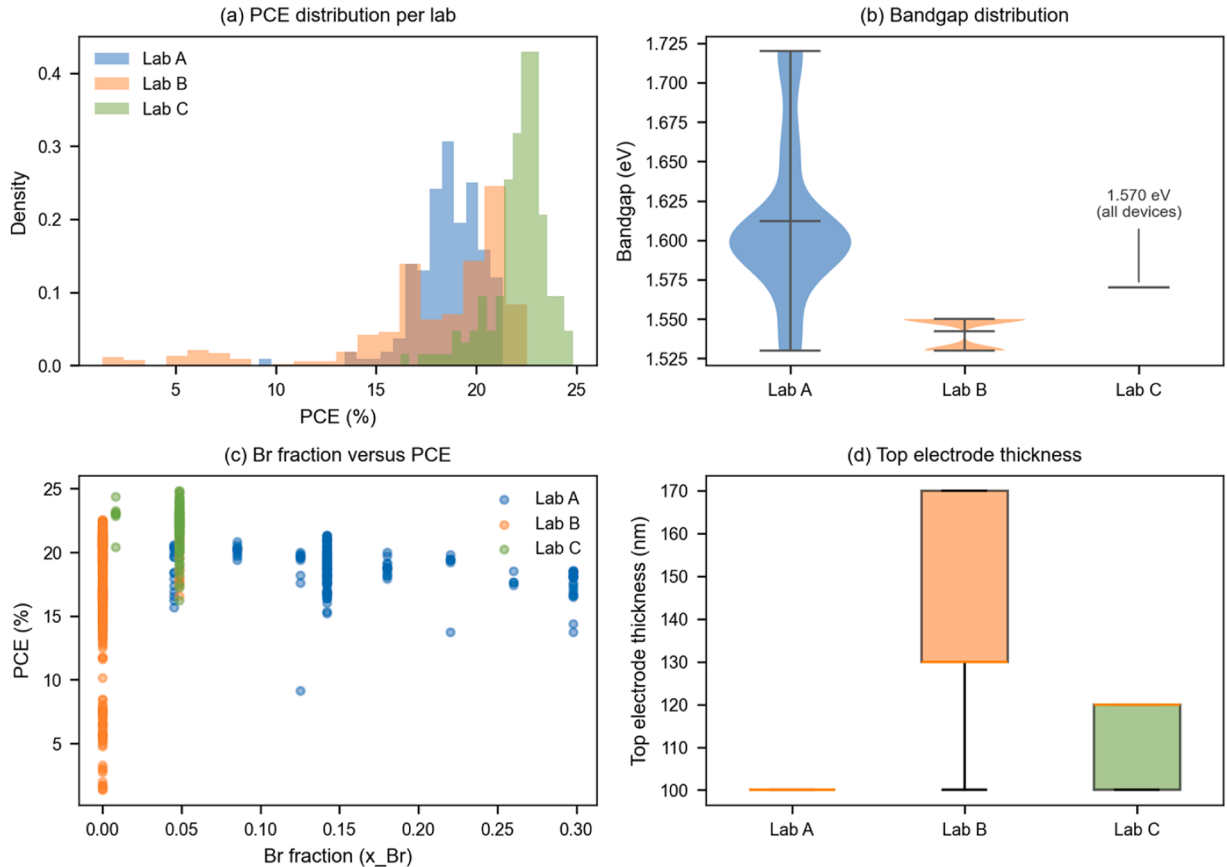


Fig. 1. Dataset overview ($n = 819$ p-i-n devices). (a) PCE distribution by laboratory. (b) Bandgap distribution by laboratory. (c) Br fraction versus PCE, colored by laboratory. (d) Top-electrode thickness distribution by laboratory.

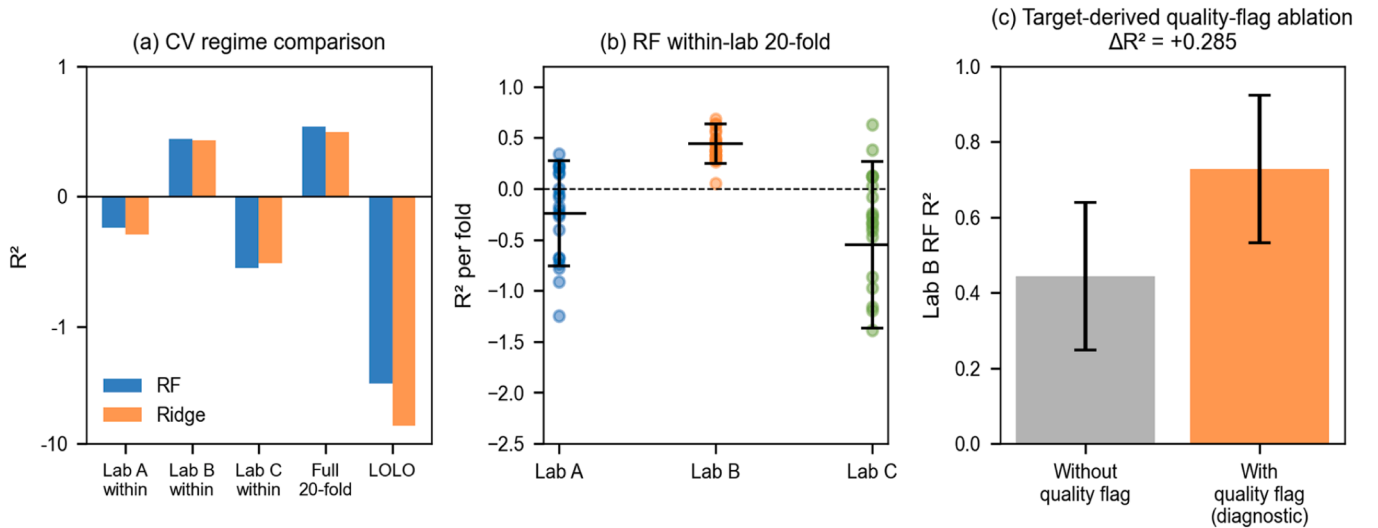


Fig. 2. . Cross-laboratory *generalization failure*. (a) Mean R^2 for RF and Ridge across within-laboratory, full-dataset, and leave-one-laboratory-out (LOLO) evaluation regimes. The y-axis uses a signed-log scale to display positive within-laboratory/full-dataset performance and strongly negative LOLO performance on the same axis. Negative values are shown after applying $\log(|R^2| + 1)$ to the absolute magnitude while preserving the original sign. (b) Fold-level RF R^2 for within-laboratory 20-fold CV, with mean ± 1 SD. (c) Diagnostic *target_derived_quality_flag* ablation in Lab B, showing inflation in RF R^2 when the *target_derived_quality_flag* is added to the primary 45-feature matrix. *target_derived_quality_flag* is excluded from all primary analyses.

Table 2

Full-dataset 20-fold cross-validated performance on the post-holdout training set using the 45-feature derived block with *target_derived_quality_flag* excluded.

Model	CV R^2 Mean	CV R^2 Std	N (train)
RF	0.537	0.108	654
Ridge	0.495	0.165	654

Mean LOLO R^2 is strongly negative for every model class evaluated (RF -3.05 ; Ridge -33.3 , median -7.05 ; Lasso -4.49 ; XGBoost -3.51 ; LightGBM -2.34), and the best single model-fold combination (LightGBM with Lab B held out) reaches only $R^2 = 0.07$, confirming that the failure is not specific to a single algorithm and ruling out model misspecification as the primary cause. Under LOLO evaluation, models must predict devices fabricated with previously unseen material and architectural combinations, requiring systematic extrapolation in energy-level alignment feature space. This behavior is consistent with prior materials informatics studies identifying distributional shift as a major source of generalization failure [22,23,25,26,31,32]. A predicted-versus-observed analysis of the extreme Lab A Ridge fold (Supplementary Figure S1) confirms that its large negative R^2 reflects systematic extrapolation rather than a single anomalous device: 43% of the held-out predictions fall outside the observed Lab A PCE range, and removing the worst-predicted devices does not recover performance. A Random Forest hyperparameter sensitivity analysis further shows the negative LOLO result is robust to model tuning: across 18 hyperparameter configurations the mean LOLO R^2 remained negative, with the best configuration still below zero (-2.26 versus baseline -3.05).

The order-of-magnitude gap between within-laboratory and LOLO performance has important implications for interpreting perovskite ML results. Within-laboratory R^2 values in the 0.4–0.8 range, which are common in the literature, cannot be assumed to demonstrate cross-laboratory predictive utility, since LOLO evaluation on the same dataset and models can yield strongly negative values without altering either the feature set or model class. This result indicates cross-laboratory predictive failure: models trained on the available laboratories do not reliably predict devices from an unseen laboratory. However, the source of this predictive failure is an identifiability problem rather than model choice alone. Because composition, device architecture, and laboratory identity covary in the present cohort, the transferable

composition–efficiency relationship cannot be cleanly separated from laboratory-specific process signatures. If within-laboratory performance instead arises from encoding laboratory-specific process signatures, then feature-importance rankings should diverge systematically across laboratories.

2.3. SHAP feature importance: lab-specific signatures and target leakage

Fig. 3 shows the rank heatmap of the top SHapley Additive exPlanations (SHAP) features for each laboratory. No feature appears in the top-10 rankings across all three laboratories, and pairwise top-10 overlap ranges from 1/10 to 5/10 features (Table S2), indicating that the most influential predictors are not stable across the cohort. The most recurrent high-ranking features are process- and architecture-related parameters, including interlayer concentration and top-electrode thickness, rather than composition descriptors. These features have physically plausible links to photovoltaic performance: interlayer concentration can affect interfacial coverage, defect passivation, charge extraction, and interfacial recombination, whereas top-electrode thickness can influence series resistance, charge collection, optical losses, and deposition-induced damage. Their prominence in individual laboratories is therefore consistent with known photovoltaic loss pathways. However, because these descriptors also encode laboratory-specific processing choices, their SHAP importance should be interpreted as local process–performance correlations rather than universal transferable design rules.

The A–X electronegativity difference (*en_diff_a_minus_x*) is the only composition descriptor appearing in the top-10 SHAP rankings for more than one laboratory: Lab A rank 3 (mean $|SHAP| = 0.035$) and Lab B rank 7 (mean $|SHAP| = 0.245$), while ranking 31st in Lab C. We therefore interpret it as a partial two-laboratory signal rather than a confirmed cross-laboratory descriptor. The feature has a physically plausible interpretation through A–X bond ionicity, which has been linked to carrier mobility and defect tolerance in perovskites [33–35], making it a candidate for validation in larger multi-laboratory cohorts. Such effects may contribute to photovoltaic performance through changes in carrier transport and non-radiative recombination, which ultimately influence voltage, fill factor, and power conversion efficiency. Its absence from the Lab C top-10 nevertheless indicates that the signal is not transferable across all three laboratories in the present

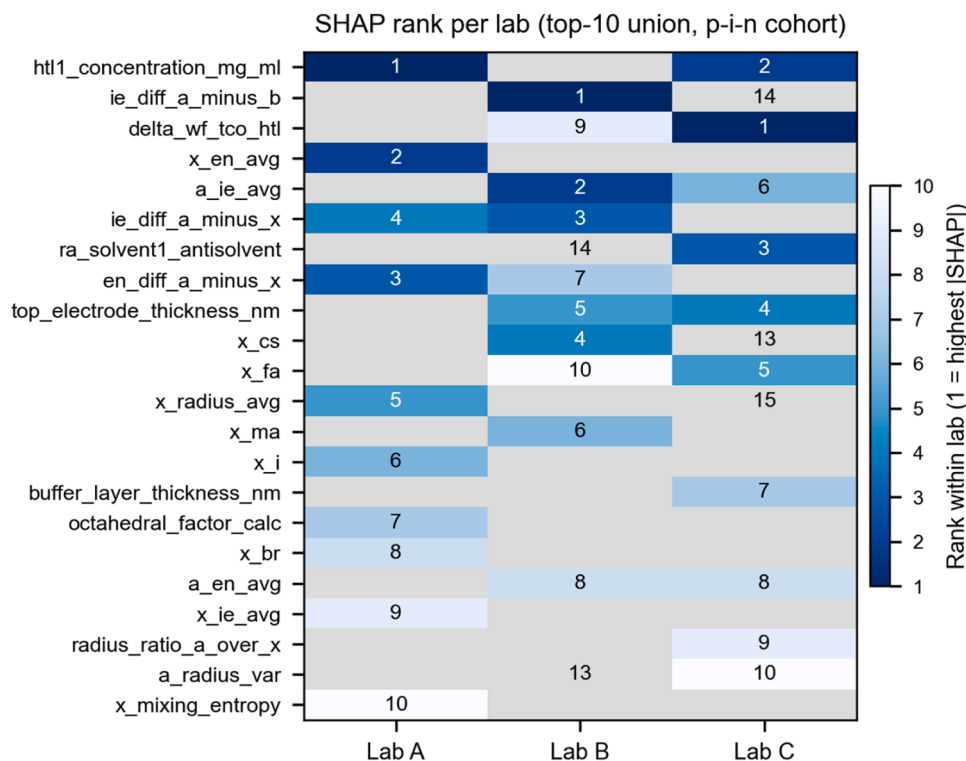


Fig. 3. SHAP feature importance is laboratory-specific. Rank heatmap showing the union of top-10 SHAP features across the three laboratories. Cell color indicates within-laboratory rank, with darker shading representing higher importance; numbers indicate the corresponding rank. Gray cells indicate features outside the top-10 for that laboratory. The A–X electronegativity difference ($en_diff_a_minus_x$) appears in the top-10 for Labs A and B, but not Lab C, illustrating a partial two-laboratory signal rather than a cohort-wide transferable descriptor.

dataset.

We next used the `target_derived_quality_flag` covariate as a diagnostic experiment to quantify how target-derived quality indicators can inflate within-laboratory performance. This binary flag was constructed by assigning `target_derived_quality_flag = 1` to devices with fill factor (FF) < 30% or PCE < 5%. In total, 24 Lab B devices were assigned `target_derived_quality_flag = 1`, corresponding to 4.8% of Lab B p-i-n devices (Table S3). Because the flag is defined directly from PCE and FF, it acts as a near-tautological predictor rather than an independent processing or compositional descriptor. To quantify this effect, `target_derived_quality_flag` was appended as a 46th feature to the Lab B matrix and 20-fold CV was repeated. RF R^2 increased from 0.443 ± 0.195 to 0.728 ± 0.195 ($\Delta R^2 = +0.285$; Table S4), demonstrating that a target-derived quality covariate can inflate within-laboratory performance by a magnitude comparable to the model's entire signal. The `target_derived_quality_flag` covariate is excluded from all primary within-laboratory, full-dataset, and LOLO analyses. We report this diagnostic separately to illustrate the need for studies containing such covariates to report performance both with and without them. Empirical batch-correction methods such as ComBat, which uses empirical Bayes modeling to remove systematic between-batch offsets [21,36], provide a complementary strategy when batch identity cannot be eliminated experimentally. Multi-batch SHAP stability analysis combined with permutation-importance corrections [37] may provide an additional safeguard.

The combination of laboratory-specific SHAP rankings and target-derived covariate inflation has important implications. First, SHAP importance derived from a single laboratory should not be interpreted as evidence of generalizable feature significance, even when within-laboratory performance is high. Second, studies including quality indicators derived from the regression target should report results both with and without those covariates, as the resulting inflation can approach the magnitude of the model's full predictive signal [21,36,37].

Although SHAP analysis identifies the dominant predictors within each laboratory, it does not partition the underlying PCE variance into compositional, architectural, and laboratory-specific components. Determining whether cross-laboratory failure is driven primarily by architecture, composition, or laboratory identity therefore requires a separate variance-decomposition analysis.

2.4. Variance decomposition highlights architecture-associated confounding

Fig. 4a presents the analysis of variance (ANOVA) Type II decomposition of total PCE sum-of-squares across the three factor blocks. Architecture accounts for 45.4% of total variance, compared with 1.5% for composition and 0.7% for laboratory identity, leaving 52.4% residual variance (Table 3). Fig. 4b shows the complementary RF permutation group importance analysis, which assigns 81.5% importance to architecture, 16.6% to composition, and 1.9% to laboratory identity. Although the two methods differ quantitatively, both produce the same qualitative ordering: architecture > composition > laboratory. This quantitative discrepancy reflects the different attribution strategies of the two analyses rather than a contradiction between them. ANOVA partial sum-of-squares partitions shared variance according to the specified model structure and degrees of freedom. RF permutation importance estimates the loss in predictive performance when each feature group is withheld. Therefore, neither value should be interpreted as uniquely correct. The key result is that both analyses independently rank the architecture-associated feature block as the dominant statistical contributor under the predefined grouping. This shared ordering supports the interpretation that composition–efficiency relationships are under-identified in this cohort because architecture, composition, and laboratory identity are not independently sampled. The substantial residual variance (52.4%) likely reflects within-laboratory process variability, including spin-coating conditions, annealing profiles, and

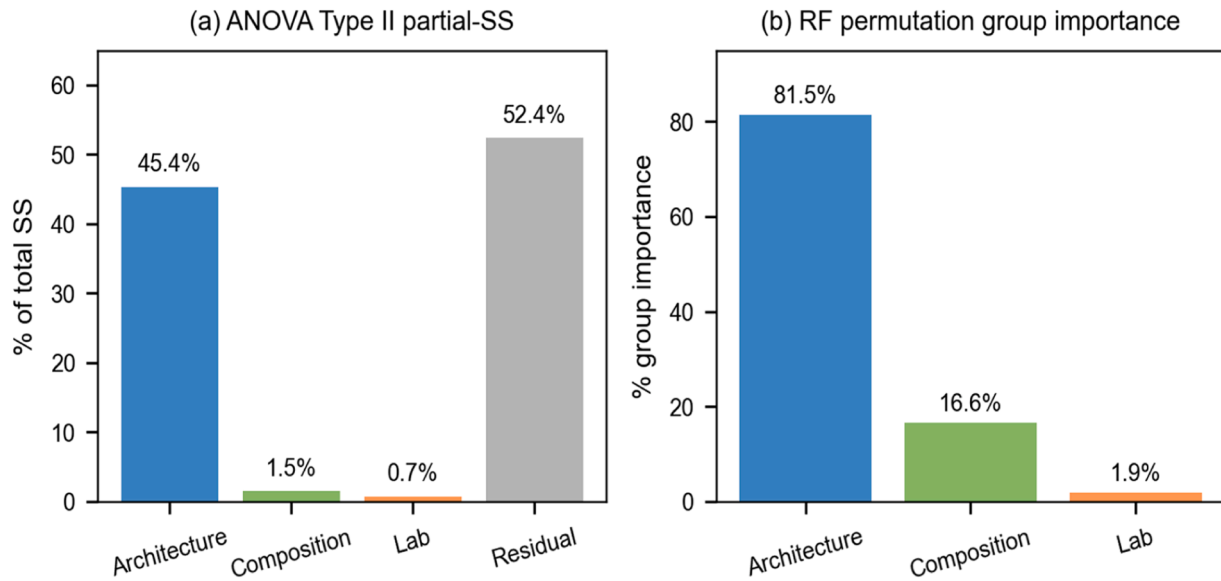


Fig. 4. Variance decomposition of the p-i-n cohort. (a) ANOVA Type II decomposition of total PCE sum-of-squares into architecture, composition, laboratory, and residual components. (b) RF permutation group importance for the same factor blocks. Both analyses identify architecture as the dominant explainable source of PCE variance.

Table 3

Variance decomposition of PCE using ANOVA Type II and RF permutation group importance. Values are reported as percentages of total sum-of-squares for ANOVA and relative group importance for RF permutation analysis. Full model specification is provided in SI Section 4.

Factor	% Total SS (ANOVA)	% Group Imp. (RF perm.)
Architecture	45.4%	81.5%
Composition	1.5%	16.6%
Lab	0.7%	1.9%
Residual	52.4%	-

substrate preparation, that is not captured by the 10-covariate ANOVA model and is consistent with the within-laboratory heterogeneity described in Section 2.2.

These feature blocks should therefore be interpreted as statistical groupings rather than mutually independent causal categories. For example, bandgap is grouped with composition but can also reflect processing-dependent film properties, while HTL/ETL material codes and electrode thickness are grouped with architecture but can also encode laboratory-specific material choices and fabrication conventions. The variance-decomposition results therefore identify the dominant statistical source of explainable variation under the chosen grouping, rather than assigning uniquely causal variance to isolated physical factors.

Cross-lab *generalization* of a composition-centered ML model would require either composition to dominate PCE variance or composition effects to remain sign-consistent across laboratories. Neither condition is satisfied in the present cohort. Architecture explains substantially more variance than composition under both analyses, and the three laboratories occupy distinct regions of architecture space, with systematically different feature combinations and no HTL or ETL material shared uniformly across all laboratories. Under LOLO evaluation, the model must therefore extrapolate into previously unseen energy-level alignment space, creating a structural barrier that cannot be resolved simply by increasing single-laboratory dataset size. Whether composition descriptors such as *en_diff_a_minus_x* (Section 2.3) can function as transferable cross-laboratory predictors remains an open question requiring validation in larger community-scale datasets spanning broader interface-engineering space.

Beyond variance attribution, these findings have direct methodological implications. When architecture accounts for the dominant share of explainable variance and laboratories occupy disjoint regions of architecture space, removing architecture features still leaves the model operating within a composition space that remains laboratory-specific. Consequently, modeling approaches that do not explicitly address between-laboratory structure, including transfer learning, multi-task learning, and domain adaptation [25,26], are likely to inherit this confound regardless of feature selection. Variance decomposition alone, however, does not determine how much architecture contributes to LOLO failure, because variance attribution on the full 819-device cohort and LOLO predictive performance measure different quantities. A direct test therefore requires repeating LOLO evaluation after removing architecture features and assessing whether composition-only models can approach the mean-prediction threshold ($R^2 = 0$).

2.5. Composition-only LOLO performance and the cross-laboratory threshold

Section 2.4 showed that architecture-associated descriptors capture the largest share of explainable PCE variance under the predefined feature grouping, whereas composition descriptors contribute substantially less. We therefore asked whether the cross-laboratory *generalization* failure documented in Section 2.2 arises primarily from architecture-feature confounding or extends to composition descriptors themselves. To test this, we restricted the feature matrix to composition-only descriptors and compared LOLO performance against the full-feature baseline.

The composition-only matrix comprised 34 features describing A-site, B-site, and X-site elemental properties, including ionic radii, ionization energies, electronegativities, composition fractions (x_{cs} , x_{fa} , x_{ma} , x_{Pb} , x_{Sn} , x_i , x_{Br} , x_{Cl}), Goldschmidt tolerance and octahedral factors, site-specific variance terms, pairwise elemental difference and ratio descriptors, and A-site and X-site mixing entropies (Table S5). The 11 architecture and processing features present in the full 45-feature block, including energy-level alignment, layer thickness, and solvent parameters, were excluded (Table S6). LOLO evaluation was then repeated on the same 654-device post-holdout training set using the protocols described in Section 4.5.

Fig. 5a shows the fold-level RF LOLO R^2 values for the full 45-feature

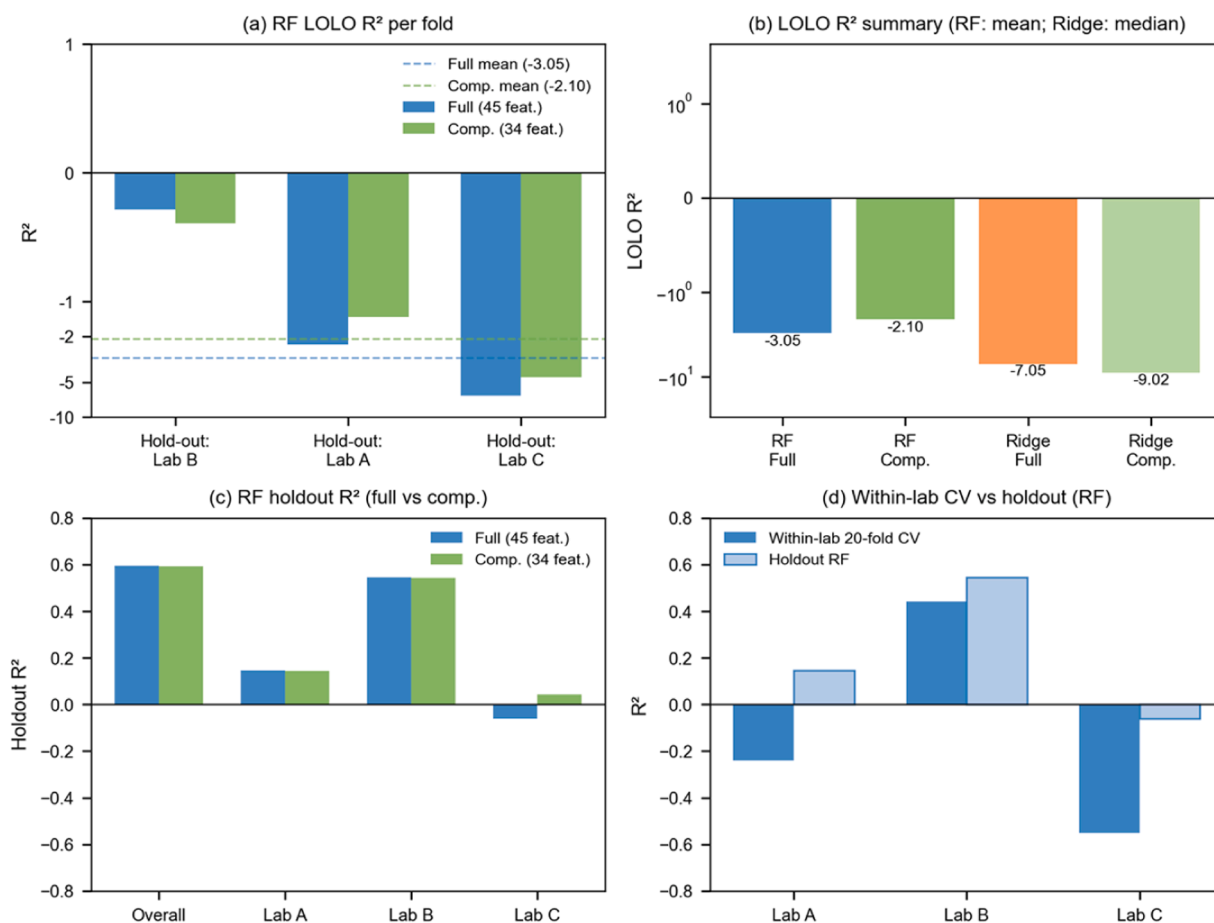


Fig. 5. Composition-only LOLO baseline and holdout validation. (a) Fold-level RF LOLO R² comparing the full derived feature block with the composition-only feature matrix. (b) LOLO R² summary under full and composition-only feature conditions (mean for RF; median for Ridge, whose mean is dominated by the extreme Lab A extrapolation fold, Table S1). (c) RF holdout R² on the locked holdout set, shown overall and by laboratory for both feature sets. (d) Comparison between within-laboratory 20-fold CV and holdout R² by laboratory, showing reproduction of the within-laboratory heterogeneity pattern on unseen data.

block (mean -3.05 ; Lab B held out: -0.29 , Lab A: -2.34 , Lab C: -6.51) and the composition-only matrix (mean -2.10 ; Lab B: -0.39 , Lab A: -1.35 , Lab C: -4.54). Fig. 5b summarizes all four conditions (RF and Ridge with full and composition-only feature sets). Restricting the model to composition descriptors partially attenuates LOLO failure: RF LOLO R² improves from -3.05 to -2.10 ($\Delta R^2 = 0.95$), corresponding to a 23% reduction in residual sum-of-squares. However, performance remains well below the mean-prediction baseline ($R^2 = 0$), indicating that composition-only modeling does not recover meaningful cross-laboratory predictability. Architecture features therefore contribute to the failure, but their removal is insufficient because the composition space itself remains laboratory-specific and partially out of support. We therefore report composition-only LOLO RF R² = -2.10 as the principal result for the present cohort under this protocol, remaining below the threshold ($R^2 > 0$) required for claims of cross-laboratory generalization. Ridge regression remains unstable under both feature conditions. With the full 45-feature block, the median LOLO R² is -7.05 , while the mean is dominated by an extreme Lab A fold caused by linear extrapolation into previously unseen architectural space. Complete fold-level values are provided in Table S1. This behavior highlights the sensitivity of linear models to systematic extrapolation under small-group evaluation.

To test whether this residual failure reflects laboratory-specific composition coverage, we quantified cross-laboratory overlap in composition-descriptor space (Supplementary Figure S2). Principal component analysis of the 34 composition descriptors (PC1 and PC2 capturing 90.5% of the variance) shows that Lab A and Lab B occupy

almost completely separated composition regions (k-nearest-neighbour laboratory purity 1.00 and 0.99 versus a random-mixing baseline of 0.45), whereas Lab C overlaps substantially with Lab B (purity 0.33; smallest inter-laboratory centroid distance). Despite this compositional overlap, Lab C remains the least predictable laboratory under LOLO (composition-only RF R² = -4.54), and the failure is asymmetric: holding out the largest, most heterogeneous laboratory (Lab B) yields the least negative LOLO R² (-0.29), whereas holding out Lab C, which spans the highest and narrowest PCE range and a distinct interface-material region, yields the most negative (-6.51). Cross-laboratory predictability therefore does not track composition-space proximity, reinforcing that architecture- and processing-related confounding, rather than composition coverage alone, governs transferability.

Fig. 5c presents holdout R² values for both feature sets. RF trained on all 654 training devices and evaluated on the locked 165-device holdout achieved an overall R² of 0.59 for both the full and composition-only feature matrices, indicating moderate random-split generalization independent of feature condition. For the full feature block, Lab B achieved R² = 0.55 (95% bootstrap confidence interval (CI): 0.47–0.61; $n = 102$), Lab A 0.15 (-0.08 to 0.37; $n = 35$), and Lab C -0.06 (-0.62 to $+0.09$; $n = 28$) using 1000 percentile bootstrap resamples (Table S7). The composition-only model produced similar estimates for Labs B and A (0.54 and 0.14), while Lab C shifted from -0.06 to $+0.04$. Because the Lab C holdout contains only 28 devices, this apparent improvement should be interpreted cautiously and not considered evidence that composition-only features resolve Lab C predictability. A paired bootstrap comparison confirmed this interpretation: the 95% CI for the Lab C

R^2 difference included zero ($\Delta R^2 = +0.104$, 95% CI: -0.062 to $+0.353$; Table S8). Importantly, random-split holdout evaluation cannot validate LOLO failure because training and test sets share laboratory membership by design. Nevertheless, because the 165-device holdout set was locked before analysis, it provides a partial check of within-laboratory reproducibility under the available retrospective dataset. A stronger test of within-laboratory temporal drift would require devices fabricated in the same laboratory months apart, which was not available in the present cohort.

Fig. 5d compares within-laboratory 20-fold CV R^2 with per-laboratory holdout R^2 . For Lab B, within-laboratory RF CV (0.44 using the full feature block) tracks the holdout estimate (0.55) closely. In contrast, Labs A and C remain weak under both evaluations. Lab A exhibits a holdout CI spanning zero, whereas Lab C remains at or below zero for the full feature block, with wide uncertainty reflecting the small holdout size. These results confirm that the within-laboratory heterogeneity identified in Section 2.2 persists on unseen data.

The composition-only LOLO RF R^2 of -2.10 reported here is specific to the defined protocol conditions: p-i-n perovskite PCE regression, the 34 composition descriptors listed in Table S5, and three-fold GroupK-Fold LOLO evaluation with laboratory identity as the grouping variable. Other cohorts applying the same protocol will produce different values depending on laboratory count, composition coverage, and feature design. LOLO $R^2 > 0$ was used as a minimum requirement because R^2 was computed separately within each LOLO test fold. Since each LOLO test fold consists of one held-out laboratory, the mean-prediction baseline corresponds to the held-out laboratory mean rather than the pooled multi-laboratory mean. Thus, the threshold already tests whether the model improves upon a laboratory-specific mean baseline, although it remains a necessary rather than sufficient condition for practical utility. Absolute error metrics corroborate this interpretation (Supplementary Table S9): the mean LOLO RF error (MAE 3.07%, RMSE 4.01% PCE) exceeds that of predicting the held-out laboratory mean, whereas the random-split holdout error is far smaller (MAE 1.61% PCE), confirming that LOLO $R^2 > 0$ is necessary but not sufficient for practical cross-laboratory utility. In addition to R^2 , we report mean absolute error (MAE) and root-mean-square error (RMSE) for all validation regimes (Supplementary Table S9). R^2 measures variance explained relative to the held-out mean baseline, whereas MAE and RMSE quantify absolute prediction error in PCE percentage points; reporting both avoids overinterpreting R^2 values close to the threshold.

The diagnostic framework therefore provides two directly transferable outputs. First, the protocol itself, which can be applied to any multi-laboratory cohort to generate comparable evaluation results. Second, the explicit threshold for cross-laboratory generalization under this protocol, namely LOLO $R^2 > 0$, which distinguishes models that improve upon the mean baseline from those that do not. In the present cohort, no model or feature configuration reaches this threshold; the composition-only LOLO RF R^2 of -2.10 is the closest among the evaluated conditions. Collectively, these findings point to a broader issue in multi-laboratory materials ML: laboratory-specific processing covaries with composition and jointly shapes the prediction target wherever datasets are aggregated across laboratories.

2.6. Implications for multi-laboratory materials machine learning

The diagnostic protocol and performance threshold defined here were demonstrated using a perovskite p-i-n cohort, but the structural conditions underlying the observed cross-laboratory failure are not unique to perovskites. In multi-laboratory materials ML datasets, three characteristics commonly co-occur. First, each laboratory occupies a distinct region of feature space determined by its equipment, material inventory, and processing conventions. Second, the prediction target depends on interactions between composition and processing rather than on composition alone. Third, the per-laboratory data volume is often insufficient to support reliable independent cross-validation.

These conditions are likely to arise in other materials ML applications, including catalyst screening, battery cathode optimization, and organic semiconductor characterization, where laboratory-specific processing similarly covaries with the target property. The diagnostic procedures developed here, namely LOLO evaluation combined with SHAP stability analysis and ANOVA variance decomposition, can therefore be applied directly to these settings. Likewise, the proposed cross-laboratory threshold (LOLO $R^2 > 0$) generalizes without recalibration.

The adequacy of the present 819-device cohort should be interpreted in this diagnostic context. The dataset is sufficiently large to reveal a reproducible gap between standard validation and laboratory-aware validation across multiple model classes, but it should not be viewed as sufficiently large or representative for training a deployment-ready cross-laboratory predictor. The cohort includes only three laboratories, has imbalanced laboratory sizes, and occupies laboratory-specific composition and architecture spaces. These properties increase the risk that standard cross-validation overestimates model robustness by fitting laboratory-specific correlations. The negative LOLO results therefore do not imply that cross-laboratory prediction is impossible, but rather that broader, more balanced, and more compositionally overlapping datasets are required for robust transferable models.

Two developments are particularly relevant to extending this framework. First, autonomous experimentation platforms and self-driving laboratories increasingly generate multi-campaign datasets with the documented provenance necessary for laboratory-stratified analysis [29,30], providing a natural context in which these diagnostics can be incorporated into closed-loop experimentation workflows. Second, community-scale resources such as the Perovskite Database Project [28], which aggregates $>42,400$ devices from hundreds of laboratories, provide sufficient scale to evaluate how broadly the present findings extend across the perovskite ML literature. Laboratory-aware analysis of such datasets may clarify how closely existing methods approach the LOLO $R^2 > 0$ threshold and identify which composition descriptors, if any, remain transferable across diverse laboratories.

From an engineering perspective, the proposed diagnostic protocol functions as a validation gate for machine-learning models before they are used for perovskite device design, process optimization, or scale-up decisions. Models validated only by random or within-laboratory splits may recommend composition, interface, or processing rules that reflect laboratory-specific fabrication histories rather than transferable device behavior. Therefore, achieving robust cross-laboratory generalization will likely require methods that address the underlying structural confound directly rather than relying solely on feature engineering. Candidate approaches include transfer learning and domain adaptation methods that explicitly model between-laboratory differences [25,26], multi-task learning across laboratories, and batch-correction strategies adapted from genomics [21,36]. Under the framework proposed here, such methods should be evaluated against the LOLO $R^2 > 0$ criterion on the same cohort they are designed to address. The composition-only LOLO RF R^2 of -2.10 reported for the present cohort therefore provides a concrete baseline against which alternative methods can be compared. More broadly, rigorous cross-laboratory validation will require laboratory-stratified holdout datasets in addition to the within-laboratory and LOLO evaluations performed here, which we identify as an important direction for future work.

3. Conclusion

This work establishes a quantitative framework for diagnosing cross-laboratory generalization failure in perovskite solar cell machine learning and defines an explicit performance threshold against which corrective methods can be evaluated. We show that the failure is substantial, reproducible, and structural. Within-laboratory R^2 values of up to 0.44 collapse to leave-one-laboratory-out (LOLO) R^2 values as low as -3.05 in a controlled three-laboratory cohort of 819 p-i-n devices, and

the collapse persists across all five model families evaluated. The underlying mechanism is structural confounding among laboratory identity, processing architecture, and composition. Architecture accounts for 45.4% of total PCE variance under ANOVA and 81.5% under RF permutation importance, whereas composition accounts for only 1.5% and 16.6%, respectively. Removing architecture features only partially mitigates the failure (LOLO RF R2 improving from -3.05 to -2.10 ; $\Delta R2 = 0.95$, corresponding to a 23% reduction in residual sum-of-squares), because composition distributions are themselves laboratory-disjoint by design. This work provides two connected outputs for the perovskite ML community. First, it introduces a diagnostic framework combining LOLO cross-validation, SHAP stability analysis, and ANOVA variance decomposition, applicable to any multi-laboratory dataset in which laboratory-specific processing covaries with the prediction target. Second, it defines an explicit cross-laboratory performance criterion (LOLO R2 > 0) that distinguishes predictions outperforming the mean baseline from those that do not. In the present cohort, the composition-only LOLO RF R2 of -2.10 is the closest result to this threshold among the evaluated conditions and remains substantially below it. More broadly, high within-laboratory cross-validated R2 values should not be interpreted as evidence of cross-laboratory predictive utility unless accompanied by laboratory-aware evaluation. Laboratory-stratified validation in larger multi-laboratory cohorts will be necessary to determine how closely existing methods approach the LOLO R2 > 0 threshold and to identify composition descriptors that remain transferable across the broader perovskite ML landscape.

4. Methodology

4.1. Dataset

The dataset comprised 819 Pb-based p-i-n perovskite solar cells collected from three independent research groups (Lab A, $n = 177$; Lab B, $n = 496$; Lab C, $n = 146$). All devices shared the p-i-n carrier-transport architecture, enabling direct cross-laboratory comparison within a common device class.

The three laboratories differed simultaneously in composition, architecture, and performance. Lab A focused on wide-bandgap, Br-rich compositions ($x_{Br} = 0.045$ – 0.298 ; mean bandgap 1.61 eV; mean PCE 18.6%). Lab B occupied near-pure iodide compositions ($x_{Br} = 0.000$ – 0.049 ; mean bandgap 1.54 eV; mean PCE 17.8%) and exhibited a broad, left-skewed PCE distribution (SD 4.4%) due to a subpopulation of low-performing devices. Lab C also used near-pure iodide compositions (bandgap 1.57 eV) but achieved the highest mean PCE (22.2%; SD 1.4%). Because these differences occurred simultaneously, cross-laboratory performance divergence could not be attributed to any single variable.

Device architectures were also non-overlapping at the material level. Lab A used MeO-2PACz as the HTL and C60 as the ETL on FTO and ITO substrates. Lab B used NiOx or self-assembled monolayers (SAMs) as HTLs and C60 or PCBM as ETLs on ITO. Lab C used MeO-2PACz, Me-4PACz, or NiOx as HTLs and C60 as the ETL. No HTL or ETL material was shared uniformly across all three laboratories.

4.2. Feature engineering

Two feature blocks were constructed. The first was a 45-feature derived descriptor block computed from documented perovskite composition fractions, reference physical properties, and selected architecture and processing descriptors. Of these, 34 were composition-only descriptors, including A-site, B-site, and X-site weighted averages of ionic radius, ionization energy, and Pauling electronegativity; composition fractions (x_{Cs} , x_{Fa} , x_{Ma} , x_{Pb} , x_{Sn} , x_{I} , x_{Br} , x_{Cl}); Goldschmidt tolerance and octahedral factors; site-fractional variance descriptors; pairwise elemental difference and ratio descriptors; and A-site and X-site mixing entropies (Table S5). The remaining 11 features

captured architecture and processing information, including energy-level alignment descriptors, layer thicknesses, HTL precursor concentration, and perovskite solvent ratios (Table S6). The composition-only 34-feature subset used in Section 2.5 was generated by retaining only the composition descriptors from this derived block.

The second block was a 164-feature raw process matrix assembled from fabrication records. This block included directly measured or specified fabrication parameters, including electrode thicknesses, spin-coating speeds, accelerations and durations, annealing temperatures and times, layer concentrations, solvent codes, and material work functions. Some raw process variables overlapped conceptually with derived descriptors, for example where measured layer thicknesses were also used to construct derived architecture descriptors. After concatenation of the raw and derived blocks, zero-variance columns and columns containing >80% missing values were removed, yielding a final 201-feature combined matrix rather than the nominal total of 209 features. The reduction of eight features resulted from post-concatenation filtering.

The primary analyses reported in Sections 2.2–2.5 used the 45-feature derived block. This choice enabled direct comparison between within-laboratory, full-dataset, and LOLO results and the 34-feature composition-only subset examined in Section 2.5. The 201-feature combined matrix was not used for the primary analyses because inclusion of raw process variables increased within-laboratory predictability without resolving the cross-laboratory generalization failure that motivated this study. Restricting the primary analysis to the derived block therefore provided a more controlled comparison between architecture-inclusive and composition-only descriptors.

A binary diagnostic covariate, *target_derived_quality_flag*, was constructed by flagging devices with FF < 30% or PCE < 5%. Of the 819 p-i-n devices, 24 were assigned *target_derived_quality_flag* = 1, corresponding to 2.9% of the full cohort and 4.8% of Lab B. All *target_derived_quality_flag* = 1 devices originated from Lab B. Because this flag was constructed directly from PCE and FF thresholds, it constituted a near-tautological predictor and was excluded from the primary feature matrix used in Sections 2.2, 2.4, and 2.5 and in Tables 1–3. Instead, Section 2.3 reports a separate diagnostic experiment in which *target_derived_quality_flag* was appended to the Lab B feature matrix to quantify the extent to which a target-derived covariate can inflate within-laboratory performance. This separation ensured that the primary within-laboratory, full-dataset, and LOLO analyses were based exclusively on composition and architecture descriptors.

4.3. Data validation

Physical bounds checking was applied before modeling: PCE 0.1–35%, bandgap 1.1–3.2 eV, open-circuit voltage (Voc) 0.1–3.0 V, short-circuit current density (Jsc) 1–40 mA cm⁻², and FF 10–100%. Zero-variance columns and columns containing >80% missing values were removed. Remaining missing values were imputed using the median within a scikit-learn Pipeline consisting of SimpleImputer(strategy='median') followed by the regression model. The imputation statistic was computed separately for each training fold and applied to the corresponding test fold, preventing test-set leakage during pre-processing. Devices with PCE values between 0.1% and 5% passed the physical bounds checks but were independently flagged by the *target_derived_quality_flag* covariate (PCE < 5% or FF < 30%); the two filters were applied separately.

4.4. Models and hyperparameters

Five baseline models were evaluated. Hyperparameters were either optimized through inner cross-validation or fixed at predefined baseline settings. For Ridge regression, alpha was tuned over {0.01, 0.1, 1, 10, 100} using five-fold inner CV nested within each outer fold. Under the LOLO protocol, inner CV used GroupKFold with laboratory identity as

the grouping variable to prevent inner-fold laboratory leakage. Lasso regression used a fixed alpha of 0.1 as a regularized linear baseline and was not separately tuned. For Random Forest, the parameters were $n_estimators = 300$ and $max_features = 'sqrt'$, with all remaining parameters at scikit-learn defaults. XGBoost used $n_estimators = 500$, $learning_rate = 0.05$, and $max_depth = 5$, with remaining parameters at library defaults. LightGBM used $n_estimators = 500$, $learning_rate = 0.05$, and $num_leaves = 31$, with remaining parameters at library defaults. The tree-based ensemble models were not independently optimized. Because the principal finding of this study is the dissociation between within-laboratory and LOLO performance, models were compared using standardized baseline settings to avoid conflating hyperparameter optimization with cross-laboratory generalization. All stochastic operations used $seed = 42$ throughout, including KFold, GroupKFold, RandomForestRegressor, XGBoost, LightGBM, and numpy. random within bootstrap confidence interval calculations. Seed robustness was assessed by repeating within-laboratory 20-fold CV across five seeds; all findings were directionally consistent (Table S10). A Random Forest hyperparameter sensitivity analysis under the LOLO protocol ($n_estimators \in \{100, 300, 500\}$, $max_depth \in \{None, 5, 10\}$, $max_features \in \{sqrt, 0.5\}$; 18 configurations) was performed to assess robustness to hyperparameter choice.

4.5. Cross-validation protocol

A PCE-stratified 20% holdout set (165 devices) was locked before analysis and excluded from all full-dataset and LOLO cross-validation procedures. The primary protocol used 20-fold KFold CV ($shuffle = True$, $seed = 42$) on the remaining 654 training devices. A sensitivity analysis comparing 5-fold and 20-fold KFold CV for RF on the full training set showed a mean R^2 difference of approximately 0.01 (Table S11). A separate LOLO robustness check varying Random Forest hyperparameters was also performed to confirm that the negative LOLO result was not driven by the baseline model setting; the full results are reported in Supplementary Table S12. The 20-fold protocol was adopted for consistency across within-laboratory and full-dataset evaluations and because the smallest laboratory training subset (Lab C, $n = 118$) yielded approximately six samples per fold, making fold-level variance informative for assessing prediction instability. Within-laboratory analysis applied 20-fold KFold CV independently to each laboratory subset after holdout exclusion (Lab A, $n = 142$; Lab B, $n = 394$; Lab C, $n = 118$). LOLO evaluation used three-fold GroupKFold on the 654 training devices, with laboratory identity specified as the grouping variable. For LOLO evaluation, R^2 was computed independently within each held-out laboratory fold using the standard scikit-learn implementation. Therefore, the mean-prediction baseline for each LOLO R^2 value corresponds to the mean PCE of the corresponding held-out laboratory fold rather than the pooled multi-laboratory mean.

4.6. Variance decomposition

Type II ANOVA was performed on the full 819-device dataset using Laboratory, Composition (x_{cs} , x_{fa} , x_{ma} , x_i , x_{br} , $bandgap_{ev}$), and Architecture (tco_code , $htl1_code$, etl_code) as factor blocks. These blocks were predefined for statistical diagnosis and were not assumed to represent mutually independent causal categories, because several descriptors can encode both physical design choices and laboratory-specific processing conventions. Partial sums-of-squares were reported as percentages of total sum-of-squares. RF permutation group importance ($n_repeats = 30$, $seed = 42$) was computed using the same feature groups as a complementary nonlinear measure.

4.7. SHAP analysis

For each laboratory, an RF model was trained on the corresponding post-holdout training subset using the 45-feature derived block with

target_derived_quality_flag excluded. SHAP values were then computed using TreeExplainer [38,39] on $\min(200, n)$ randomly sampled devices ($seed = 42$). Features were ranked within each laboratory by mean absolute SHAP value (mean $|SHAP|$), and the top 10 features were retained. Cross-laboratory stability was assessed by comparing rank positions for all features appearing in the top-10 list of at least one laboratory.

The 819-device experimental dataset analyzed in this study was compiled from real perovskite solar cell data generated by three collaborating Korean universities: Hanyang University, Chung-Ang University, and Ajou University. The raw device-level data are not publicly deposited because the data were provided by the contributing research groups and are subject to institutional and data-provider restrictions. Requests for access to the dataset may be directed to the corresponding author, and data sharing will be considered by the data providers on a case-by-case basis, subject to their approval and any applicable institutional requirements. These restrictions are legal and institutional in nature, arising from an ongoing collaborative project; the complete analysis code is nonetheless publicly available and, together with the feature definitions, fold-level statistics, and validation protocols provided in the Supporting Information, enables full reproduction of the diagnostic workflow on any comparably structured multi-laboratory dataset.

CRedit authorship contribution statement

Minseong Kim: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation. **Jinho Lee:** Writing – review & editing, Data curation, Conceptualization. **Hye W. Chun:** Writing – review & editing, Resources, Methodology. **Eun Seo Shin:** Writing – review & editing, Resources, Methodology. **Hyosung Choi:** Writing – review & editing, Supervision, Funding acquisition. **Dong-Won Kang:** Writing – review & editing, Supervision, Funding acquisition. **Jong H. Kim:** Writing – review & editing, Supervision, Funding acquisition. **Shujuan Huang:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jincheol Kim:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jincheol Kim reports financial support was provided by Australian Government. Jincheol Kim reports financial support was provided by National Research Foundation of Korea. Jincheol Kim reports financial support was provided by Korea Institute of Energy Technology Evaluation and Planning. Jincheol Kim reports financial support was provided by Australia's Economic Accelerator Ignite Grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Australian Government under grant GSTDS000001-7, the National Research Foundation of Korea funded by the Korea government (MSIT) under grant RS-2024-00436187, and the Korea Institute of Energy Technology Evaluation and Planning and the Ministry of Trade, Industry and Energy of the Republic of Korea under grant RS-2024-00437030. This work was also partially supported by Australia's Economic Accelerator Ignite Grants under grant IG240100331.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.egyai.2026.100856](https://doi.org/10.1016/j.egyai.2026.100856).

Data availability

The code used for data preprocessing, feature engineering, lab-aware cross-validation, model training, SHAP analysis, variance decomposition, holdout validation, and figure generation in this study is available in the GitHub repository [<https://github.com/MINTATO/psc-cross-lab-ml-generalization>] (accessed on 15 May 2026). Additional data supporting the findings of this study are available within the article and its supplementary information.

References

- [1] Jeong J, Kim M, Seo J, Lu HZ, Ahlawat P, Mishra A, et al. Pseudo-halide anion engineering for α -FAPbI₃ perovskite solar cells. *Nature* 2021;592:381–5.
- [2] Yoo JJ, Seo G, Chua MR, Park TG, Lu YL, Rotermund F, et al. Efficient perovskite solar cells via improved carrier management. *Nature* 2021;590:587–93.
- [3] Jeon NJ, Noh JH, Yang WS, Kim YC, Ryu S, Seo J, et al. Compositional engineering of perovskite materials for high-performance solar cells. *Nature* 2015;517:476–80.
- [4] Li JX, Pradhan B, Gaur S, Thomas J. Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. *Adv Energy Mater* 2019;9:1901891.
- [5] Liu YM, Tan XY, Liang J, Han HW, Xiang P, Yan WS. Machine learning for perovskite solar cells and component materials: key technologies and prospects. *Adv Funct Mater* 2023;33:2214271.
- [6] Sun SJ, Hartono NTP, Ren ZKD, Oviedo F, Buscemi AM, Layurova M, et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* 2019;3:1437–51.
- [7] Tao QL, Xu PC, Li MJ, Lu WC. Machine learning for perovskite materials design and discovery. *npj Comput Mater* 2021;7:23.
- [8] Odabasi C, Yildirim R. Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools. *Nano Energy* 2019;56:770–91.
- [9] Choe H, Yoon H, Kim I, Kim SY. Machine learning unlocks new directions in halide perovskite research. *ChemElectroChem* 2026;13:e202500282.
- [10] Chargui T, Idrissi R, Bacha A, Lmai F. Investigation of the optical and electronic properties of double perovskite Li₂CuBiX₆ (X = Br, I) for photovoltaic applications using first-principles and machine learning approaches. *Sol Energy* 2025;299:113737.
- [11] El Badraoui A, Chargui T, Elkhoul A, El Mokhtari B, El Idrissi R, El Amri L, et al. Optimization of lead-free CsSnBr₃-based perovskite solar cells via SCAPS-1D simulations and machine learning models. *J Phys Chem Solids* 2026;210:113340.
- [12] Sadouqi A, Chargui T, El Idrissi R, Lmai F. Numerical validation and machine learning-based modeling of energy conversion efficiency in Cs₂AgBi_{0.75}Sb_{0.25}Br₆ double perovskite solar cells. *Sol Energy Mater Sol Cells* 2026;302:114330.
- [13] Chargui T, El Badraoui A, Errouyssi Y, Lmai F, Rahmani K. First-principles study of the optoelectronic properties of Ag₂MgGe(S/Se)₄ chalcogenides for tandem solar cell applications. *J Phys Chem Solids* 2026;216:113772.
- [14] Jeon GG, Lee DS, Choi MJ, Seo YH, Huang SJ, Kim JH, et al. Mitigation of parasitic leakage current in indoor perovskite photovoltaic modules using porous alumina interlayer. *EcoMat* 2024;6:e12455.
- [15] Yousaf T, Lee SM, Shim H, Yun JS, Yun JH, Kim J, et al. Facet-dependent charge dynamics of 2D butylammonium lead bromide perovskite single crystals. *Small Struct* 2026;7:e202500816.
- [16] Jeon GG, Yoon SE, Han J, Chun HW, Shin SJ, Choi MJ, et al. Surfactant-assisted colloidal size and charge control in perovskite solutions for high-performance and stable solar cells. *Chem Eng J* 2025;519:164991.
- [17] Han JH, Kim RH, Huang SJ, Kim J, Yun JS. Green solution processing of halide perovskite solar cells: status and future directions. *Sol RRL* 2024;8:2400262.
- [18] Sakib SN, Payne DNR, Kim J, Huang SJ, Veettil BP. Rapid microwave annealing of perovskite films: exploring the mechanism of heat generation and influence on growth kinetics. *Sol Energy Mater Sol Cells* 2026;295:113967.
- [19] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;11:733–9.
- [20] Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* 2022;23:169–81.
- [21] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;8:118–27.
- [22] Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, Thompson A, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol Syst Eng* 2018;3:819–25.
- [23] Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput Mater* 2020;6:138.
- [24] Li KM, et al. Probing out-of-distribution generalization in machine learning for materials. *Commun Mater* 2025;6:9.
- [25] Sutton C, et al. Identifying domains of applicability of machine learning models for materials science. *Nat Commun* 2020;11:4428.
- [26] Xiong Z, et al. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Comput Mater Sci* 2020;171:109203.
- [27] Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater* 2019;5:83.
- [28] Jacobsson TJ, et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nat Energy* 2022;7:107–15.
- [29] MacLeod BP, et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci Adv* 2020;6:eaz8867.
- [30] Eo P-K, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* 2022;8:84.
- [31] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010;11:2079–107.
- [32] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform* 2006;7:91.
- [33] Bartel CJ, et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci Adv* 2019;5: eaav0693.
- [34] Lee JW, Tan S, Seok SI, Yang Y, Park NG. Rethinking the A cation in halide perovskites. *Science* 2022;375:eabj1186.
- [35] Saidi WA, Shadid W, Castelli IE. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. *npj Comput Mater* 2020;6:36.
- [36] Zhang YQ, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2020;2: lqaa078.
- [37] Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010;26:1340–7.
- [38] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017;30:4765–74.
- [39] Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;2:56–67.