Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

Published on 28 September 2026
Licensed under CC-BY 4.0

Check for updates

Cite this: *Digital Discovery*, 2026, **5**, 4882

Multi-axis quality evaluation of data augmentation for perovskite solar cell datasets

Minseong Kim,[a] Jinho Lee,[a] Hye W. Chun,[b] Eun Seo Shin,[c] Hajra Yousf,[a] Shujuan Huang,[a] Dong-Won Kang,[d] Hyosung Choi,[c] Jong H. Kim[b] and Jincheol Kim*[a]

Experimental data in perovskite solar cell (PSC) research are often costly and time-intensive to obtain, making data augmentation an attractive strategy for expanding limited datasets for machine learning-based prediction and optimisation. However, augmentation quality has typically been judged only by downstream predictive performance, which can obscure structural distortions when high-capacity models mask degraded data quality. Here, we propose a multi-axis evaluation framework that explicitly separates augmentation quality assessment from predictive modelling, evaluating each condition along three independent axes of predictive response, data compatibility, and structural preservation. Applying this framework to 374 laboratory-measured PSC samples across four augmentation methods and three ratios, we found that the three axes ranked conditions in different orders, revealing quality differences that no single metric could detect. The framework successfully distinguished augmentation conditions that preserved the physical feature relationships in the original data from those that severely distorted them, and this distinction remained consistent across 50 repeated train-test splits. These findings provide a new evaluation standard for data augmentation in scientific datasets, ensuring that domain-specific feature relationships are preserved rather than silently compromized.

Received 16th April 2026
Accepted 29th August 2026

DOI: 10.1039/d6dd00198j

rsc.li/digitaldiscovery

1. Introduction

Perovskite solar cell (PSC) manufacturing data have become a central target for data-driven modeling,[1–3] as device performance emerges from the complex interplay of composition, processing conditions, and device architecture.[4] In PSC research, machine learning is increasingly used to support performance prediction,[5–7] process optimization,[8] and experimental exploration.[9] However, experimentally accumulated PSC datasets are often limited in size[10] and can exhibit substantial heterogeneity.[11,12] Features in PSC datasets are often interdependent due to underlying physical and processing constraints,[13,14] and heterogeneity across laboratories and measurement conditions can further complicate the observed data.[15,16] These characteristics make limited data availability a continuing challenge for predictive modeling,[15,17] and strategies that expand dataset size without disrupting their underlying structure are therefore needed.

Data augmentation represents a practical strategy for small experimental materials datasets, where additional data collection through fabrication and characterization is costly and time-intensive.[18] Data augmentation approaches can be broadly categorized as statistical, simulation-guided, and physics-constrained. Statistical approaches are useful for enriching patterns already represented in observed data, simulation-guided approaches for exploring conditions supported by validated computational models, and physics-constrained approaches for enforcing known physical or processing rules during sample generation. The four methods examined in this study—noise injection, interpolation, KDE, and CTGAN—belong to the statistical category, as they generate synthetic samples directly from the observed experimental dataset. To date, however, data augmentation for experimental PSC tabular datasets has not been systematically evaluated, as augmentation studies in the perovskite field have focused primarily on image-based or computationally generated data rather than laboratory-measured device datasets. In PSC data, augmentation quality is determined not by the number of generated samples, but by whether synthetic data remain physically consistent and structurally compatible with experimental observations. PSC features are subject to physical and processing constraints: the accessible composition space is governed by

[a]*Macquarie University, School of Engineering, Sydney, NSW 2109, Australia. E-mail: jincheol.kim@mq.edu.au*

[b]*Department of Molecular Science and Technology, Ajou University, Suwon 16499, Republic of Korea*

[c]*Department of Chemistry, Research Institute for Convergence of Basic Science, and Hanyang Institute for Quantum Technology, Hanyang University, Seoul, Republic of Korea*

[d]*School of Energy Systems Engineering, Chung-Ang University, Seoul, 06974, Republic of Korea*

© 2026 The Author(s). Published by the Royal Society of Chemistry

crystallographic and phase-related limitations, while fabrication variables must be selected within feasible process windows defined by the manufacturing context.[19] Under these conditions, unconstrained augmentation may generate unrealistic samples or distort correlation structures that are important for preserving the underlying data structure.[9,20–22] Thus, the objective is not to increase data volume, but to expand datasets in a way that preserves their physical consistency and structural integrity.[23,24]

Despite these concerns, augmentation assessment has frequently centered on downstream predictive utility.[25,26] Such reliance is problematic, however, because predictive models can absorb or mask the effects of data corruption, maintaining acceptable performance even when structural inconsistencies or distributional distortions are present.[27,28] These limitations indicate that augmentation assessment requires a broader evaluation framework that considers multiple dimensions of synthetic data quality beyond predictive performance alone.[29,30] In particular, evaluation strategies are needed that distinguish downstream predictive utility from structural consistency, while also considering whether augmented data remain compatible with the original dataset and preserve its structural relationships.[31,32]

Recent studies have proposed multi-dimensional evaluation frameworks for synthetic tabular data, using multiple quality dimensions, including fidelity, utility, privacy, and distributional or application-level preservation.[29,30] However, these frameworks were developed primarily for general tabular settings and do not account for the domain-specific physical and processing constraints that characterize scientific experimental datasets, such as the crystallographic boundaries and fabrication windows that govern PSC feature relationships. In particular, these frameworks do not include an explicit axis for evaluating the preservation of physically constrained feature-pair relationships, which are a defining characteristic of experimental materials datasets. In PSC data, feature relationships are not merely statistical patterns but can reflect physically grounded dependencies among composition, band gap, and fabrication conditions. As a result, augmentation methods that appear acceptable under general synthetic data metrics may still distort domain-relevant structural relationships. This distinction motivates the need for an evaluation framework that explicitly considers the preservation of physically meaningful feature-pair relationships in PSC datasets.

To address this gap, this study proposes a multi-axis evaluation framework in which the roles of data quality assessment and predictive modeling are explicitly separated. First, support vector regression (SVR),[33,34] a regularization-based and capacity-controlled model, is used as a probe to detect augmentation-induced changes in data quality, while XGBoost,[35] a high-capacity ensemble model, serves as a predictive reference model. Second, augmentation strategies are evaluated along three independent axes, rather than through a single performance metric: performance response, compatibility, and structural deviation.[29,30] Third, the results from these axes are integrated into a unified decision workflow. Rather than recommending a specific augmentation method, this framework establishes a generalizable validation paradigm that enables dataset-specific selection of augmentation strategies, reducing the risk of misleading performance gains arising from structurally distorted data and the unreliable model conclusions that follow.

2. Results

2.1 PSC dataset overview and evaluation framework

Table 1 summarizes the dataset structure and augmentation design used in this study. The analysis was conducted on 374 PSC experimental samples provided by three independent laboratories in Korea, located at Hanyang University, Ajou University, and Chung Ang University. All samples were collected as part of the same collaborative project and recorded from the outset using a shared, predefined data template. The three laboratories used the same column structure, variable definitions, measurement units, and data-entry conventions. The prediction target is the power conversion efficiency (PCE) measured from the reverse scan, in which the voltage is swept from open-circuit to short-circuit conditions. The input consists of 18 features capturing composition, processing conditions, and device structure. Among these, 12 are continuous numerical variables, including compositional fractions, bandgap, solvent composition, annealing conditions, and process and geometric parameters such as thickness and area. The remaining six are categorical variables that describe structural information, including cell architecture and the stack sequence of the substrate, electron transport layer (ETL), hole transport layer (HTL), and back contact layers. Open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), and fill factor (FF) were excluded from the inputs because they are outcome variables.

Fig. 1 outlines the overall role separated evaluation framework used in this study. In this framework, augmentation quality was evaluated across three complementary axes: SVR probe response, train-on-real compatibility, and structural deviation. Each axis captures a different aspect of augmentation quality, so augmentation conditions were not judged by a single metric alone.

2.2 SVR probe response (R^2) across augmentation ratios

Fig. 2 shows how R^2 changes across augmentation ratios (1×, 2×, 5×, and 10×) for the SVR and XGBoost models. This reveals whether augmentation-induced differences are equally visible across models with different capacity characteristics. The critical observation is not the absolute R^2 level, but how clearly each model separates the responses of different augmentation methods and ratios. Full model and evaluation details are provided in Method (Section 5.2.1).

In SVR, noise injection and interpolation show a tendency for R^2 to increase as the augmentation ratio increases. In contrast, KDE and GAN show a decreasing trend in R^2 with increasing augmentation ratio, with particularly pronounced performance degradation at 10×. Thus, even under the same starting dataset and the same augmentation ratios, the pattern of R^2 response differs substantially depending on the

© 2026 The Author(s). Published by the Royal Society of Chemistry

Table 1 Summary of the PSC dataset, feature composition, and augmentation design used in this study

Category	Item	Value
Dataset overview	Original dataset size	374
	Train-test split	8 : 2
	Target variable	Perovskite solar cell efficiency (%)
Feature composition	Numeric features	12
	Categorical features	6
	Composition	A-, B-, X-site composition fractions, perovskite bandgap
	Solvent	DMF[a], DMSO[a] ratio
	Process conditions	Annealing temperature, time
	Device geometry	Active area, back contact thickness, ETL[a] thickness
	Device structure	Cell architecture, ETL/HTL[a] stack sequence, back contact stack sequence
Augmentation design	Augmentation methods	Noise injection, interpolation, KDE[a], GAN[a]
	Augmentation ratio	2×, 5×, 10×

[a] ETL, electron transport layer; HTL, hole transport layer; DMF, dimethylformamide; DMSO, dimethyl sulfoxide; KDE, kernel density estimation; GAN, generative adversarial network.

augmentation method. SVR therefore appears sensitive to augmentation-induced changes in data quality, capturing differences that high-capacity models do not reflect.

In XGBoost, R^2 remains generally high across all methods, and the magnitude of change with increasing augmentation ratio is more limited than in SVR. Although some methods show a decreasing trend as the augmentation ratio increases, the method specific responses are not separated as clearly as in SVR, suggesting that high-capacity models tend to mask augmentation-induced differences. The results separated by augmentation method are shown in Fig. S1 of the SI. However, probe response alone does not capture whether augmented data remain compatible with the original data distribution.

2.3 Train on real, test on augmented: compatibility (ΔR^2) across augmentation ratios

To examine this, the compatibility axis fixes the baseline model trained on the original non augmented training set and evaluates it on augmented data generated under each condition. Here, ΔR^2 denotes the change in R^2 relative to the same model evaluated on the original test set. In this setting, ΔR^2 is interpreted not as a direct measure of performance improvement, but as an indicator of how well augmented data remain compatible with the original data distribution. Full model and evaluation details are provided in Method (Section 5.2.2).

Under the 10× condition, noise injection and KDE showed ΔR^2 values greater than or equal to 0, whereas interpolation showed a negative ΔR^2 (Fig. 3). GAN exhibited a strongly negative ΔR^2, indicating a much more pronounced performance decrease than the other methods. Even when augmentation starts from the same real dataset, therefore, the degree to which performance is preserved differs substantially across methods.

The SVR R^2 response shown in Fig. 2 and the ΔR^2 results shown in Fig. 3 are also not aligned in the same order. For example, interpolation shows an increasing trend in the R^2 response as the augmentation ratio increases, yet its ΔR^2 is negative, indicating performance loss. In contrast, KDE shows a decreasing trend in the R^2 response, but its ΔR^2 remains greater than or equal to 0. Even for the same augmentation condition, the relative pattern observed can differ depending on the evaluation metric. Compatibility therefore captures a distinct aspect of augmentation quality that probe-based performance evaluation alone cannot reflect.

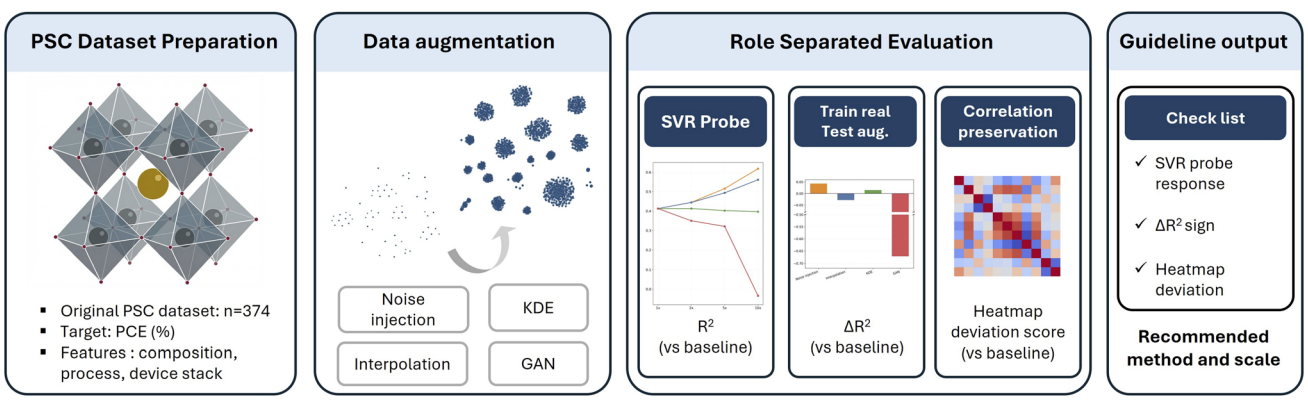

Fig. 1 Workflow of the proposed augmentation evaluation framework for PSC datasets, showing dataset preparation, data augmentation, role-separated evaluation across three axes (SVR probe response, train-on-real compatibility, and structural deviation), and guideline output for selecting an appropriate and reliable augmentation method and scale.

© 2026 The Author(s). Published by the Royal Society of Chemistry

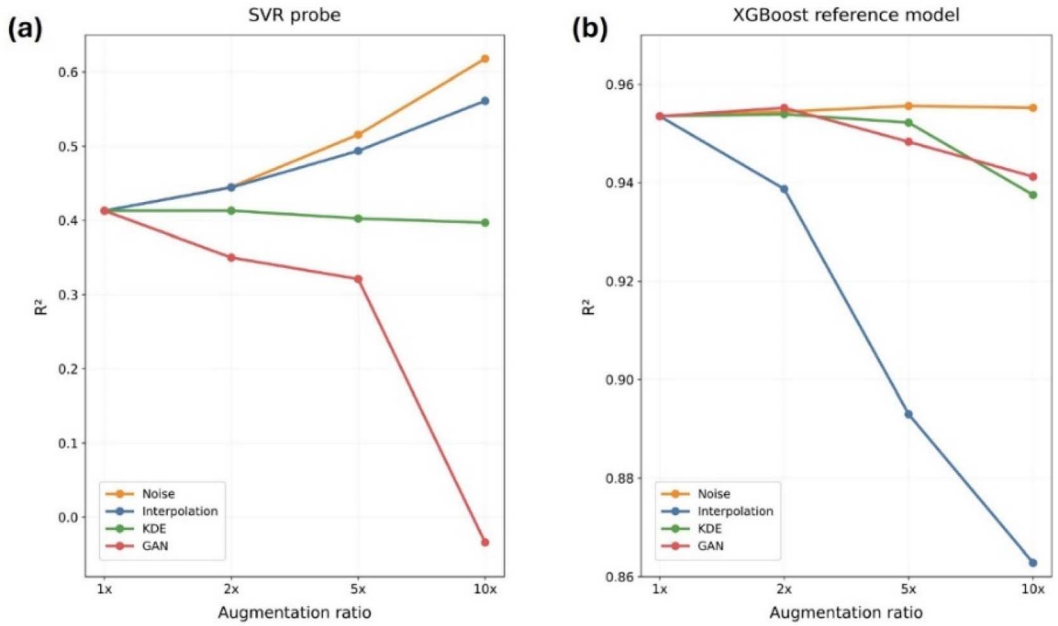

Fig. 2 R^2 responses across augmentation ratios for each augmentation method, comparing (a) SVR (probe model) and (b) XGBoost (reference model). Different y-axis ranges are used to preserve the visibility of within-model variations. Apparent fluctuation amplitudes should therefore not be compared directly across panels.

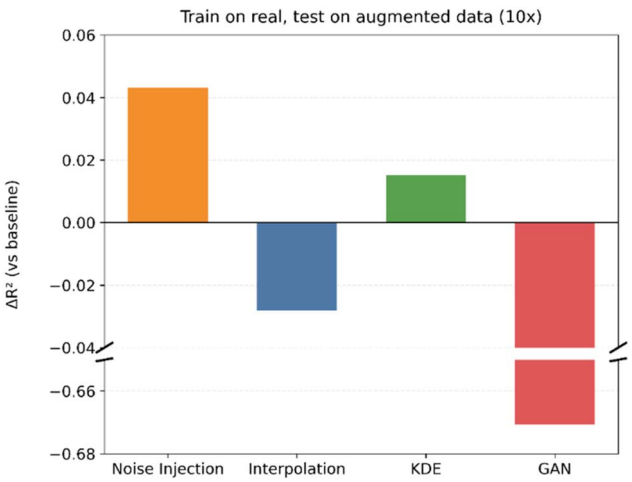

Fig. 3 ΔR^2 relative to the non-augmented (1×) baseline when a model trained on real data is evaluated on 10× augmented data for each augmentation method.

The ΔR^2 results for the other augmentation ratios, 2× and 5×, are provided in Fig. S2 of the SI. However, ΔR^2 alone does not capture whether structural relationships are preserved. This is further examined through correlation-based structural analysis.

2.4 Correlation preservation at 10× augmentation

Fig. 4 examines structural preservation under the 10× augmentation condition using Pearson r, Spearman ρ, and distance correlation (d_{Cor}). Fig. 4a–c compare the baseline heatmaps with those of each augmentation method, while Fig. 4d summarizes the corresponding mean absolute changes relative to the baseline. For compact visual comparison, all heatmaps in Fig. 4a–c are displayed using an identical feature order and metric-specific color scale. Full calculation details are provided in Method (Section 5.2.3).

In noise injection and interpolation, patterns similar to the baseline are largely preserved. In contrast, KDE and GAN show substantial deviations in the heatmap patterns. In particular, KDE and GAN exhibit weakened off-diagonal patterns in r and ρ, while d_{Cor} shows broad color changes, indicating that relationship structures distinct from those of the baseline were formed.

Fig. 4d provides a quantitative comparison based on the mean absolute change relative to the baseline heatmap for each metric. Across Pearson r, Spearman ρ, and d_{Cor}, noise injection and interpolation show relatively small changes, whereas KDE and GAN show much larger changes. Thus, under the 10× condition, the extent of change in feature pair relationships relative to the baseline differs clearly across augmentation methods. These quantitative results are consistent with the qualitative patterns observed in the heatmaps above. KDE and GAN therefore alter the underlying relational structure of the data, rather than merely introducing variability around existing observations.

Fully labelled heatmaps with individual values for the baseline and the 2×, 5×, and 10× augmentation conditions are provided in Fig. S3–S5 of the SI. In addition, the baseline subsampling variation distributions used as the reference for the heatmap changes, together with the procedure for defining T95 as the reference threshold, are presented in Fig. S6 of the SI.

© 2026 The Author(s). Published by the Royal Society of Chemistry

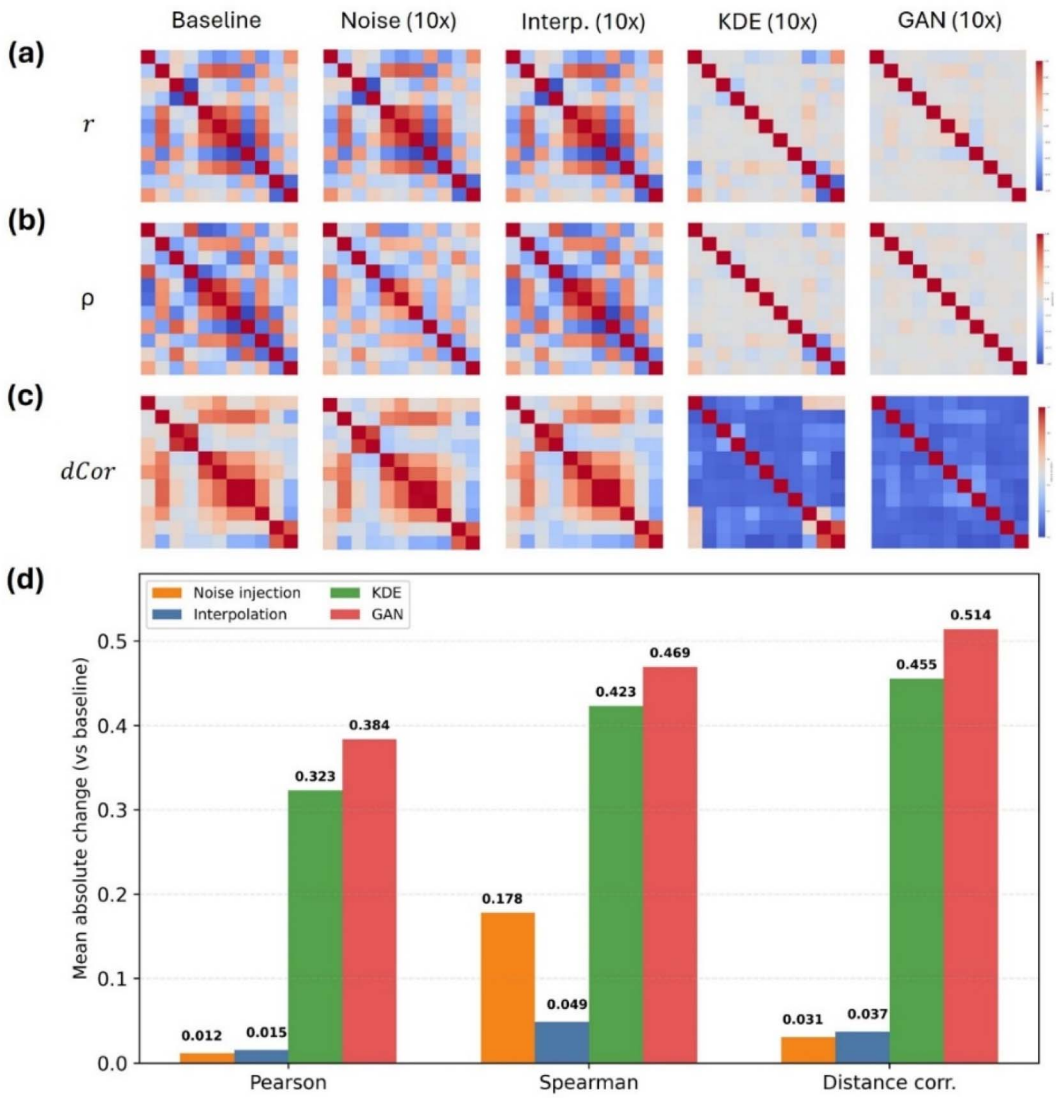

Fig. 4 Structure preservation analysis at 10× augmentation across different augmentation methods. (a–c) Baseline and augmented feature pair heatmaps for Pearson *r*, Spearman ρ, and distance correlation (d_{Cor}) across augmentation methods. (d) Mean absolute heatmap change relative to the baseline for each metric (lower is better).

These structural results are integrated with the probe response and compatibility axes in the following section to support a unified augmentation selection decision.

2.5 Multi-axis candidate map, decision rule, and repeated split validation

Fig. 5 integrates the three evaluation axes into a unified selection framework and demonstrates the robustness of the selection across 50 repeated random train-test splits. Fig. 5a shows a candidate map in which SVR R^2 is plotted on the vertical axis and the structural deviation score C is plotted on the horizontal axis for all augmentation methods and ratios, while the sign of ΔR^2 is simultaneously indicated by marker shape. Here, C summarizes the extent of structural deviation from the baseline, with smaller values indicating better preservation. The dashed line represents the structural deviation threshold, $C = 1$, defined from the baseline variability criterion.

Based on Fig. 5a, the selection rule proposed in this study was used to identify the final selected condition and the worst condition. SVR R^2 was first used to assess whether each augmentation condition maintained or exceeded the baseline response, followed by ΔR^2 to evaluate compatibility by checking for the absence of performance loss ($\Delta R^2 \geq 0$). Structural deviation was then compared using C, a normalized summary score derived from the heatmap changes in Pearson *r*, Spearman ρ, and d_{Cor} and referenced to the 95th percentile of baseline subsampling variation (T95), with full calculation details provided in Method (Section 5.2.3). When multiple ratios within the same augmentation method satisfied these criteria without meaningful differences in compatibility or structural deviation, the SVR R^2 response was additionally used to select the final

© 2026 The Author(s). Published by the Royal Society of Chemistry

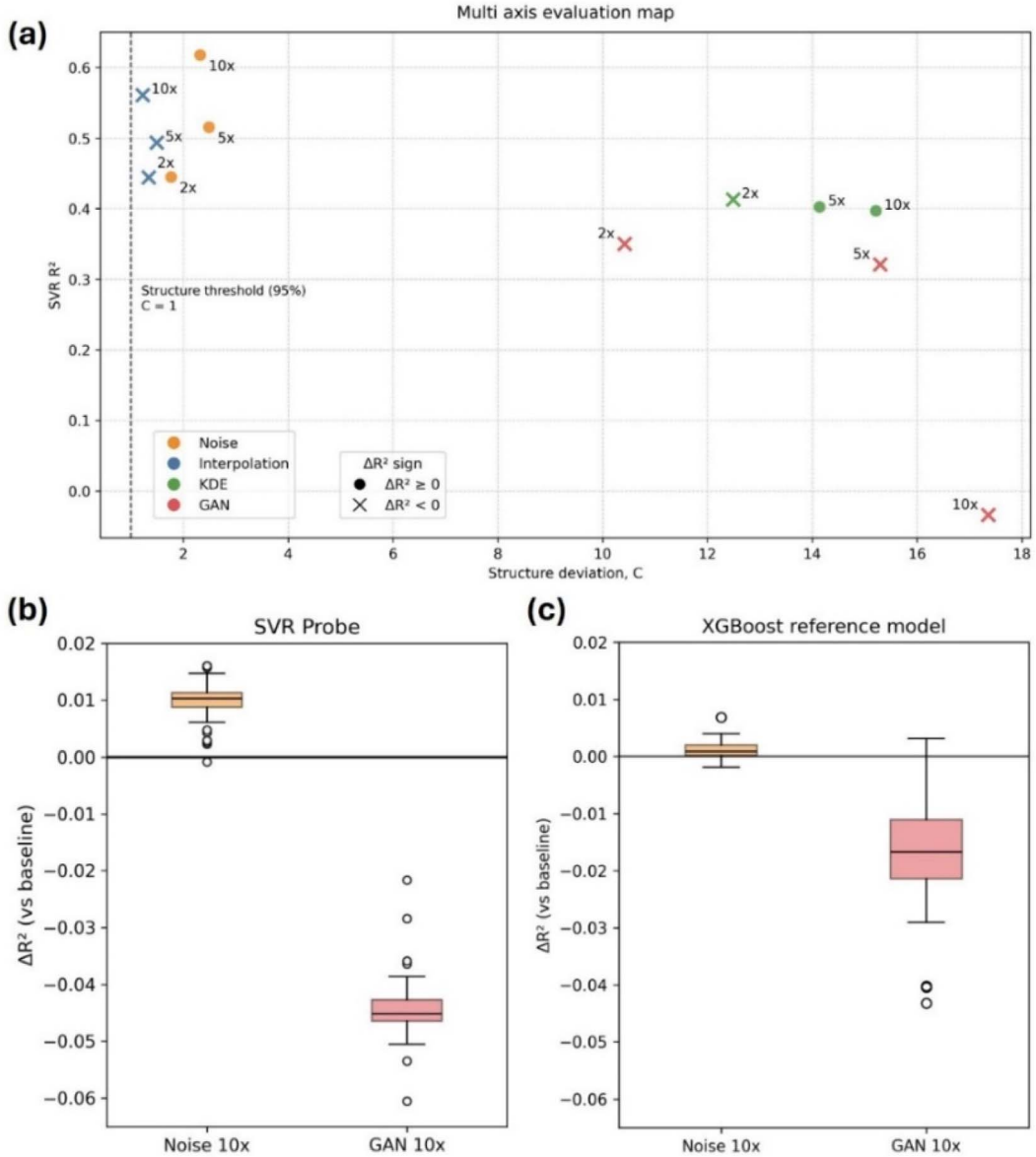

Fig. 5 Integrated multi-axis map and repeated split validation for augmentation selection. (a) Candidate map across methods and ratios showing SVR probe R^2 *versus* structural deviation C. Marker shape indicates the sign of ΔR^2 in the train-on-real, test-on-augmented setting relative to the non-augmented (1×) baseline. The dashed line marks $C = 1$. (b) ΔR^2 *versus* the non-augmented (1×) baseline over 50 random train-test splits for the selected best condition (noise 10×) and worst condition (GAN 10×) using the SVR probe. (c) The same repeated split comparison using the XGBoost reference model.

ratio. The full selection workflow is described in Method (Section 5.3).

Following these criteria, Noise 10× was selected as the best condition, whereas GAN 10× was selected as the worst condition. The SVR R^2, ΔR^2, and C values for all augmentation conditions are summarized in Table S1 of the SI.

Fig. 5b and c compare the distributions of ΔR^2 across 50 repeated splits for the two selected conditions, Noise 10× and GAN 10×. The repeated split validation procedure is described in Method (Section 5.4). In SVR, the ΔR^2 distribution for Noise 10× is located near 0 or in the positive region, whereas the ΔR^2 distribution for GAN 10× is located in the negative region (Fig. 5b). When the same comparison is performed using XGBoost, Noise 10× again shows a distribution near 0, whereas GAN 10× shows a distribution in the negative region (Fig. 5c). The proposed multi-axis framework therefore enables consistent identification of reliable and unreliable augmentation strategies across varied data splits. A detailed procedural workflow for this selection strategy is provided in Fig. S7 of the SI.

3. Discussion

The contrasting responses of SVR and XGBoost under identical augmentation conditions highlight why role separation was necessary in this study. XGBoost showed relatively compressed

© 2026 The Author(s). Published by the Royal Society of Chemistry

performance differences across augmentation methods and ratios, indicating that its predictive robustness can mask augmentation-induced changes in data quality. By contrast, SVR exhibited clearer performance separation, making it more suitable as a probe model for detecting changes introduced by augmentation. In this context, the robustness of XGBoost is advantageous for prediction, but less informative when augmentation quality is judged from predictive performance alone. These results suggest that a model used to evaluate augmentation quality and a model used for final prediction should not necessarily be assumed to serve the same role.

A key finding of this study is that augmentation quality could not be determined reliably from any single metric. The SVR R^2 response and the compatibility metric ΔR^2 did not always rank conditions in the same way, and structural preservation provided an additional perspective that could not be captured by performance-based evaluation alone. In particular, conditions that appeared acceptable from one metric were not always supported by the others, showing that predictive performance, directional compatibility, and structural consistency reflect different aspects of augmentation quality. By summarizing Pearson r, Spearman ρ, and d_{Cor}-based relationship changes into the normalized score C, this study introduced a quantitative scale for assessing how far augmented data deviated from the natural variation of the baseline. Taken together, these observations support the use of multiple complementary axes rather than any single criterion when selecting augmentation conditions in small sample scientific datasets.

The superior structural preservation of Noise Injection can be interpreted in relation to its local perturbation mechanism. Because noise is added independently to each variable at a relatively small scale, the augmented samples remain close to the original observations and are less likely to disrupt the feature-pair relationships already present in the real data. In PSC datasets, where process variables, composition, and device properties are jointly constrained by physical and fabrication conditions, such conservative perturbation is better suited to preserving the local structure of the original distribution. This interpretation is consistent with the present results, in which Noise Injection maintained relatively favorable structural deviation while also satisfying the screening criteria used in the selection workflow.

By contrast, the poor performance of CTGAN in this study may reflect the difficulty of learning a reliable joint distribution from a small and physically constrained tabular dataset. The dataset used here contained only 374 samples spanning mixed continuous and categorical variables, while the relationships among variables were governed not only by statistical association but also by domain-specific physical constraints. Under these conditions, a generative model can produce samples that are numerically plausible at the individual feature level but inconsistent with the relationship structure observed in real PSC data. This interpretation is consistent with the observation that the GAN 10× condition showed the largest structural deviation and was identified as the worst condition in the integrated evaluation. The present findings do not imply that CTGAN-based augmentation is inherently unsuitable for scientific tabular data. Rather, the observed behavior should be interpreted in the context of the dataset and implementation examined here, as alternative hyperparameter settings and architectural choices may yield different outcomes. Physics-informed GAN variants and explicit regularization of feature relationships were beyond the scope of the present comparison and remain relevant directions for future investigation.

Interpolation showed a structurally conservative pattern similar to Noise Injection, yet its compatibility was negative under the 10× condition, leading to its exclusion from the feasible candidate set. This apparent contradiction can be interpreted in relation to the nature of the interpolation mechanism. By linearly mixing pairs of real samples, interpolation generates synthetic points that lie between existing observations, keeping individual feature values within the observed range and thereby preserving pairwise correlation structures. However, the resulting samples occupy regions of the feature space that are not directly represented in the original training data. A model trained exclusively on real data may therefore encounter distributional mismatch when evaluated on interpolated samples, which can manifest as a drop in compatibility even when structural metrics remain favorable. In PSC datasets, where the feasible regions of the feature space are physically constrained, interpolated points may fall in areas that are statistically plausible but physically underrepresented in the original distribution. This distinction between structural preservation and distributional compatibility highlights why the two axes can diverge, and why both are needed as independent evaluation criteria.

KDE showed a pattern distinct from both Noise Injection and CTGAN-based augmentation. Although it did not fail as severely as the GAN 10× condition in directional compatibility, its structural deviation remained elevated and its SVR response weakened progressively as the augmentation ratio increased. This pattern suggests that KDE appears to have generated samples that broadly followed the marginal distribution of the original data while smoothing feature-pair relationships that were important for preserving the internal structure of the dataset. Unlike Noise Injection, which perturbs samples locally and conservatively, KDE fits a global density model and samples from it, a process that can flatten interdependencies that are unevenly distributed across the feature space. In PSC data, where correlations among composition, bandgap, and processing variables reflect domain-specific constraints rather than uniform statistical patterns, this global smoothing behavior may be particularly disruptive. In this sense, KDE occupied an intermediate position between conservative local perturbation and fully generative modeling, and its suitability may depend on whether the dataset exhibits strong localized structure.

Taken together, these method-specific patterns reinforce the value of the role-separated multi-axis framework proposed in this study. The results show that predictive response, compatibility, and structural preservation capture distinct aspects of augmentation quality and therefore need to be considered jointly when selecting augmentation conditions for scientific datasets where feature interdependence and domain-specific constraints are present. The practical value demonstrated

© 2026 The Author(s). Published by the Royal Society of Chemistry

here lies in comparing and selecting among statistical augmentation methods or ratios for small experimental tabular datasets, particularly when predictive performance alone does not reveal whether synthetic data remain compatible with the original observations and preserve scientifically meaningful relationships. However, this dataset-level assessment does not directly certify the physical feasibility of individual synthetic samples, which may require separate screening against explicit compositional, processing, or device-architecture constraints before use in model training. The framework may also be applicable to simulation-guided or physics-constrained synthetic data, but its utility in those settings requires independent validation.

We note several limitations of this study. First, the present dataset comprises 374 PSC samples collected from three Korean laboratories, and its composition space, device architectures, and PCE distribution have not been established as representative of the broader PSC literature. Accordingly, the relative performance of the augmentation methods observed here should be regarded as dataset-specific, and the framework should be reapplied and independently evaluated when transferred to datasets with different compositional, architectural, or performance distributions. Moreover, when the available dataset does not adequately cover the composition, processing, or device-architecture region of interest, additional experimental data collection is more appropriate because augmentation cannot reliably substitute for information absent from the original observations. Second, structural preservation was evaluated at the feature pair level, and future work could extend this analysis to higher-order multivariate structure. Third, the behavior of methods such as KDE and GAN may vary substantially with implementation details and hyperparameter settings, making broader sensitivity analysis an important next step. That said, sensitivity analysis using alternative reference thresholds showed that the relative ordering of augmentation conditions by C was unchanged across T90, T95, and T99, and the final selected and worst conditions also remained unchanged, indicating that the main comparative conclusions were not sensitive to the threshold choice. Finally, although SVR functioned effectively as the probe model in this study, the concept of a probe is not limited to SVR and could be extended to other regularization-based and capacity-controlled models.

4. Conclusions

This study establishes a systematic framework for evaluating augmentation quality under consistent criteria, rather than aiming to identify a universally best augmentation method. The framework separates the roles of probe and reference models, combines screening by SVR R^2 and ΔR^2 with relative comparison by the structural deviation score C, and enables augmentation quality to be judged simultaneously from predictive response, compatibility, and structural preservation. Applied to the PSC dataset examined here, the framework identified Noise Injection 10× as the most reliable condition and GAN 10× as the worst condition. The proposed framework is specifically motivated by scientific experimental datasets, in which variable interdependence and physical constraints must be preserved rather than treated as incidental statistical patterns. More broadly, this work provides an interpretable evaluation basis that can be applied to other scientific datasets for selecting reliable augmentation conditions beyond single-metric assessment, provided that the framework is re-evaluated within the composition, process, architecture, and performance distributions of each target dataset.

5. Methods

5.1 Data augmentation strategies

After random shuffling, the data were split into a training set and a test set at a ratio of 8 : 2, and augmentation was applied only to the training set after the split. The augmentation ratios were set to 2×, 5×, and 10×. For each ratio, synthetic samples equal in size to the original training set were generated and combined iteratively to reach total dataset sizes of two, five, and ten times the original training set, respectively. All augmentation methods generated synthetic samples based exclusively on the training set, including both input variables and the target used for model training. Categorical variables were handled according to the requirements of each augmentation method. For noise injection, the original categorical values were preserved. For interpolation, each synthetic sample inherited the categorical values of one of the two selected parent samples. For KDE, categorical values were sampled with replacement from the original training set. For GAN, categorical variables were modeled as discrete columns in conditional tabular generative adversarial networks (CTGAN).

5.1.1 Noise injection-based augmentation. Noise injection generated synthetic samples by adding Gaussian noise to each numerical variable in the original training set, with the noise magnitude scaled by the standard deviation of the corresponding variable.[36] For numerical input variables, Gaussian noise with a standard deviation equal to 0.1 times the standard deviation of each variable was added to introduce local perturbations while keeping augmented samples close to the original observations and avoiding unrealistic deviations. For the target variable, PCE, Gaussian noise with a standard deviation equal to 0.05 times the standard deviation of the target distribution was added separately to limit excessive perturbation of the target variable.

5.1.2 Interpolation-based augmentation. Interpolation generated synthetic samples by randomly selecting two distinct samples from the original training set and linearly mixing their numerical variables and the target variable, PCE, using a weight α randomly drawn from a uniform distribution between 0 and 1.[37] For categorical variables, the synthetic sample inherited the value from one of the two selected samples.

5.1.3 Kernel density estimation (KDE). KDE-based augmentation first fitted a kernel density model with a Gaussian kernel to the multivariate space defined by the numerical variables and the target variable, PCE. Synthetic samples were then generated by sampling from the learned distribution.[38,39] The bandwidth parameter was fixed at 0.3 for all KDE augmentation ratios. Categorical variables were

© 2026 The Author(s). Published by the Royal Society of Chemistry

assigned by sampling with replacement from the original training set and then combined with the sampled numerical variables and target.

5.1.4 GAN-based augmentation. GAN-based augmentation used CTGAN. The input variables and the target were organized as a single tabular dataset from the original training set, and categorical variables were specified as discrete columns within CTGAN and learned jointly with the continuous variables.[40] After training, the generative model was used to generate synthetic samples in the same number as the original training set. Default CTGAN settings were used for all parameters except the number of training epochs (300).

5.2 Role-separated evaluation framework

5.2.1 SVR probe response (R^2). For each augmentation condition, an SVR model was trained on the corresponding augmented training set, and the coefficient of determination, R^2, was computed on the non-augmented original test set to compare performance responses across augmentation conditions.[33,34] SVR used a radial basis function (RBF) kernel with C set to 100, ε set to 0.1, and γ set to scale. These hyperparameters were fixed across all augmentation conditions. Because SVR was used as a probe model for relative comparison rather than for absolute performance optimization, maintaining a consistent hyperparameter setting was considered more important to ensure fair comparison across augmentation conditions. Categorical variables were converted using one-hot encoding and included as model inputs.

5.2.2 Train on real, test on augmented: compatibility (ΔR^2). This evaluation was designed to complement Section 5.2.1 by reversing the training-testing configuration to assess how a model trained only on the original training set, without using augmented data for training, performs when evaluated on augmented data. The baseline was defined as the model trained on the original training set without augmentation and evaluated on the original test set. For each augmentation condition, the baseline model trained on the original training set was evaluated using the corresponding augmented dataset as the test data, and R^2 was computed. The change relative to the baseline was summarized as ΔR^2. The model configuration and the input preprocessing pipeline were identical to those used in Section 5.2.1.

5.2.3 Correlation preservation and structural deviation score C. Composition, band gap, solvent composition, and process variables in PSC data are often observed in interdependent patterns because they are jointly constrained by physical relationships and fabrication mechanisms. Therefore, independently of predictive performance, this study evaluated correlation-based structural preservation to assess how well augmentation preserves feature pair relationships.

First, feature pair matrices were computed using only numerical variables based on the Pearson correlation coefficient r,[41] the Spearman rank correlation coefficient ρ,[42] and distance correlation d_{Cor}.[43] Pearson r captures linear relationships, Spearman ρ summarizes monotonic relationships, and d_{Cor} captures more general dependence, including nonlinear associations. Pearson r and Spearman ρ range from minus 1 to plus 1, whereas d_{Cor} ranges from 0 to 1.

For each augmentation condition, defined by method and ratio, the resulting matrices were compared with the baseline non-augmented matrix, and the mean absolute change was calculated. The mean absolute change was defined as the average absolute difference over the upper triangular elements of the correlation matrix, excluding the diagonal. The resulting changes for each metric were denoted as Δr, $\Delta\rho$, and Δd_{Cor}.

Next, using only the baseline data, random subsampling was repeated 100 times to construct the distribution of variation for each metric, and the 95th percentile of each distribution was defined as the reference threshold T95. The subsample size was set to 80 percent of the baseline data. In each repetition, the mean absolute change between the matrix computed from the subsample and that computed from the full baseline data was recorded.

Finally, the structural deviation score C was defined as the average of the three metric-specific changes after normalization by their corresponding T95 values:

$$C = \frac{1}{3}\left[\frac{\Delta r}{\mathrm{T95}(r)} + \frac{\Delta\rho}{\mathrm{T95}(\rho)} + \frac{\Delta d_{\mathrm{Cor}}}{\mathrm{T95}(d_{\mathrm{Cor}})}\right] \quad (1)$$

As defined in eqn (1), C is a dimensionless normalized score that enables direct comparison across different correlation metrics, where smaller values indicate that the feature pair relationships in the augmented data remain closer to the natural variation range observed in the baseline data.

5.3 Multi-axis candidate map and decision rule

To integrate the results of the three evaluation axes into a single interpretable representation, a candidate map was constructed. This representation enables simultaneous assessment of performance, compatibility, and structural preservation for each augmentation condition. Each augmentation condition, defined by method and ratio, was represented as one point. The structural deviation score C was placed on the horizontal axis, and the SVR R^2 value was placed on the vertical axis. In addition, the sign of ΔR^2 from the compatibility evaluation was indicated by marker shape so that the presence or absence of performance loss could be assessed simultaneously. The line at $C = 1$ was used as a reference line corresponding to the 95th percentile criterion of the baseline subsampling variation, providing a visual reference for assessing the extent to which each condition deviates from natural structural variation.

The selection of augmentation conditions was performed according to the following rule. First, conditions with SVR R^2 values equal to or greater than the baseline (defined as the non-augmented training condition) were treated as eligible candidates. Next, conditions with ΔR^2 values equal to or greater than 0 were defined as feasible conditions. Finally, among the feasible conditions, the condition with the smallest structural deviation score C was selected as the final candidate. When multiple ratios within the same augmentation method satisfied the feasible condition, the SVR R^2 response across

© 2026 The Author(s). Published by the Royal Society of Chemistry

augmentation ratios was additionally considered, and the ratio with the highest SVR R^2 was selected as the final condition. This map provides an interpretable basis for selecting reliable augmentation strategies.

5.4 Repeated split validation and implementation details

The main analysis was conducted using the same hyperparameter settings across all augmentation methods and ratios, so that the augmentation condition was the only analysis factor varied and condition-specific model tuning was not introduced into the comparison. To assess whether the observed contrast between the selected and worst conditions depended on a specific train-test split configuration, 50 train-test splits were generated using predefined random states from 1000 to 1049 with an 80 : 20 train-test ratio. In each repetition, the same split was used for the baseline and augmented datasets and shared across SVR and XGBoost for direct comparison. The baseline model was trained on the original training set without augmentation and evaluated on the corresponding original test set. The models were then separately trained on the augmented training sets constructed using Noise 10× and GAN 10× and evaluated on the same original test set. In the repeated validation, the change was defined as the difference in R^2 values computed on the original test set, calculated as the R^2 of the model trained with augmentation minus that of the baseline model.

The same procedure was applied separately to SVR and XGBoost. For XGBoost, the hyperparameters were set as follows: n_estimators = 500, learning_rate = 0.05, max_depth = 5, subsample = 0.8, colsample_bytree = 0.8, and objective = reg:squarederror. For both models, input preprocessing was performed in the same manner as described in Section 5.2.1.

Author contributions

M. K. conducted the study, developed the methodology, performed the data analysis and augmentation experiments, prepared the figures, and wrote the original draft of the manuscript. J. L. contributed to discussion on the analytical approach and manuscript revision. H. W. C. and E. S. S. contributed to experimental data collection, coordination, and data interpretation and discussion. H. Y. contributed to early-stage conceptual design and the development of the machine learning backbone. S. H. contributed to manuscript revision, data interpretation, and funding acquisition. D. W. K., H. C., and J. H. K. provided the experimental data, contributed to data interpretation and discussion, and acquired funding for the project. J. K. conceived and supervised the project, interpreted the results, revised the manuscript, and acquired funding for the project. All authors reviewed and approved the final manuscript.

Conflicts of interest

The authors declare no competing interests.

Data availability

The perovskite solar cell datasets used in this study were obtained from Hanyang University, Ajou University, and Chung Ang University and are subject to third party data sharing restrictions. The data are therefore not publicly available. Access requests may be directed to the corresponding author and will be considered in consultation with the original data providers. The code used for data preprocessing, augmentation, model training, and evaluation in this study is available in the GitHub repository (https://github.com/MINTATO/Data-augmentation) and permanently archived on Zenodo. The archived release (v1.0.0) is available at https://doi.org/10.5281/zenodo.21981292, while https://doi.org/10.5281/zenodo.21981291 represents all versions and resolves to the latest release. Additional data supporting the findings of this study are available within the article and its supplementary information (SI). Supplementary information: additional model evaluation results, feature correlation heatmaps, structural deviation analyses, and guidelines for selecting appropriate data augmentation methods and scales. See DOI: https://doi.org/10.1039/d6dd00198j.

Acknowledgements

This work was supported by the Australian Government under grant GSTDS000001-7. This work was also supported by the National Research Foundation of Korea funded by the Korea government (MSIT) under grant RS-2024-00436187, and by 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 (IG240100331).

References

1 D. Sadhu, D. Dattatreya, A. Deo, K. Tarafder and D. De, Performance prediction and analysis of perovskite solar cells using machine learning, *J. Alloys Compd. Commun.*, 2024, **3**, 100022.
2 Y.-M. Liu, X.-Y. Tan, J. Liang, H.-W. Han, P. Xiang and W.-S. Yan, Machine Learning for Perovskite Solar Cells and Component Materials: Key Technologies and Prospects, *Adv. Funct. Mater.*, 2023, **33**, 2214271.
3 S. Valsalakumar, S. Bhandari, A. Roy, T. K. Mallick, J. Hinshelwood and S. Sundaram, Machine learning driven performance for hole transport layer free carbon-based perovskite solar cells, *npj Comput. Mater.*, 2024, **10**, 212.
4 J. Kim, B. W. Park, J. Baek, J. S. Yun, H. W. Kwon, J. Seidel, *et al.*, Unveiling the Relationship between the Perovskite Precursor Solution and the Resulting Device Performance, *J. Am. Chem. Soc.*, 2020, **142**, 6251–6260.
5 F. J. Kusuma, E. Widianto, Wahyono, I. Santoso, Sholihun, M. A. U. Absor, *et al.*, Optimizing novel device configurations for perovskite solar cells: Enhancing

© 2026 The Author(s). Published by the Royal Society of Chemistry

stability and efficiency through machine learning on a large dataset, *Renewable Energy*, 2025, **247**, 122947.
6 A. Toprak, High-accuracy machine learning approach for predicting J-V characteristics of perovskite solar cells under variable irradiance, *Sci. Rep.*, 2025, **15**, 41304.
7 J. Kumar, B. Nath, T. Bhattacharjee and P. C. Ramamurthy, Machine learning aided optoelectric characterization modelling and prediction of the IV parameters of perovskite solar cells with > 90% accuracy, *Sci. Rep.*, 2025, **15**, 30660.
8 B. Yilmaz and R. Yildirim, Critical review of machine learning applications in perovskite solar research, *Nano Energy*, 2021, **80**, 105546.
9 M. Ahmadi, M. Ziatdinov, Y. Y. Zhou, E. A. Lass and S. Kalinin, Machine learning for high-throughput experimental exploration of metal halide perovskites, *Joule*, 2021, **5**, 2797–2822.
10 M. Salah, Z. Ismail and S. Abdellatif, Selecting an appropriate machine-learning model for perovskite solar cell datasets, *Mater. Renew. Sustain. Energy.*, 2023, **12**, 187–198.
11 E. M. Tennyson, T. A. S. Doherty and S. D. Stranks, Heterogeneity at multiple length scales in halide perovskite semiconductors, *Nat. Rev. Mater.*, 2019, **4**, 573–587.
12 E. Unger and T. J. Jacobsson, The Perovskite Database Project: A Perspective on Collective Data Sharing, *ACS Energy Lett.*, 2022, **7**, 1240–1245.
13 S. Subba, P. Rai and S. Chatterjee, Machine Learning Approaches in Advancing Perovskite Solar Cells Research, *Adv. Theory Simul.*, 2025, **8**, 2400652.
14 M. A. Rahman and M. J. Alam, Optimization of Lead-Free Perovskite Solar Cell Architecture Using Machine Learning and Numerical Simulations, *Adv. Theory Simul.*, 2025, e01590.
15 V. de la Asunción-Nadal, C. Iliffe Sprague, B. Guijarro-Berdiñas, U. B. Cappel and A. García-Fernández, Machine learning for perovskite solar cells: a comprehensive review on opportunities and challenges for materials scientists, *EES Sol.*, 2025, **1**, 927–957.
16 J. S. Yun, J. Seidel, J. Kim, A. M. Soufiani, S. J. Huang, J. Lau, *et al.*, Critical Role of Grain Boundaries for Ion Migration in Formamidinium and Methylammonium Lead Halide Perovskite Solar Cells, *Adv. Energy Mater.*, 2016, **6**, 1600330.
17 S. Cho, S. C. Cho, P. Pandey, S. Lee, H. Ahn, I. W. Hwang, *et al.*, Achieving Seasonal Reproducibility in Wide-Bandgap Sn-Based Perovskite Solar Cells *via* Proton-Locking Interface Engineering, *Adv. Energy Mater.*, 2026, **16**, e05598.
18 F. Oviedo, Z. Ren, S. J. Sun, C. Settens, Z. Liu, N. T. P. Hartono, *et al.*, Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks, *npj Comput. Mater.*, 2019, **5**, 60.
19 S. Jain, G. Seth, A. Paruthi, U. Soni and G. Kumar, Synthetic data augmentation for surface defect detection and classification using deep learning, *J. Intell. Manuf.*, 2022, **33**, 1007–1020.
20 Z. Liu, N. Rolston, A. C. Flick, T. W. Colburn, Z. K. Ren and R. H. Dauskardt, Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing, *Joule*, 2022, **6**, 834–849.
21 S. Moon, P. G. Park, N. Choi, J. H. Kim, S. H. Lim, J. H. Lee, *et al.*, Dual knowledge-guided data augmentation for robust clinical prediction models, *Sci. Rep.*, 2026, 18588.
22 D. H. Won, K. S. Shin and S. Youm, Synthetic Data Augmentation for Imbalanced Tabular Data: A Comparative Study of Generation Methods, *Electronics*, 2026, **15**, 883.
23 J. H. Ding, X. C. Li, X. J. Kang and V. N. Gudivada, A Case Study of the Augmentation and Evaluation of Training Data for Deep Learning, *ACM J. Data Inf. Qual.*, 2019, **11**, 20.
24 X. Tan, S. Wang, H. Wu and F. Chen, Enhancing pumping unit diagnosis with similarity splicing data augmentation and wavelet denoising, *Sci. Rep.*, 2025, **15**, 37045.
25 S. Chang, X. Wang, Y. Luo and L. Jia, Data augmentation alters feature importance in XGBoost for CVD prediction, *Sci. Rep.*, 2025, **15**, 41754.
26 A. Figueira and B. Vaz, Survey on Synthetic Data Generation, Evaluation Methods and GANs, *Mathematics*, 2022, **10**, 2733.
27 S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, *et al.*, The effects of data quality on machine learning performance on tabular data, *Inf. Syst.*, 2025, **132**, 102549.
28 N. George, S. Gurumurthi, V. Sridharan, H. D. Dixit, E. Goksu, B. Parthasarathy, *et al.*, Silent Data Corruption in Artificial Intelligence: A Growing Challenge for Large-Scale Machine Learning, *IEEE Micro.*, 2026, **46**, 66–72.
29 M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde and N. Aginako, Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees, *Front. Digit. Health.*, 2025, **7**, 1576290.
30 B. van Breugel, T. N. S. Liu, D. Oglic and M. van der Schaar, Synthetic data in biomedicine *via* generative artificial intelligence, *Nat. Rev. Bioeng.*, 2024, **2**, 991–1004.
31 Synthetic data generation: A comparative study, *Proceedings of the 26th International Database Engineered Applications Symposium*, ed. M. Endres, A. Mannarapotta Venugopal and T. S. Tran, 2022.
32 L. R. DaCosta, K. Sytwu, C. K. Groschner and M. C. Scott, A robust synthetic data generation framework for machine learning in high-resolution transmission electron microscopy (HRTEM), *npj Comput. Mater.*, 2024, **10**, 165.
33 D. Basak, S. Pal and D. C. Patranabis, Support vector regression, *Neural Information Process.-Lett. Rev.*, 2007, **11**, 203–224.
34 D. Reeves and G. Jacyna, Support vector machine regularization, *Wiley Interdiscip. Rev. Comput. Stat.*, 2011, **3**, 204–215.
35 Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. T. Chen and C. Guestrin, 2016.

© 2026 The Author(s). Published by the Royal Society of Chemistry

36 *SMOTE and gaussian noise based sensor data augmentation. 2019 4th International Conference on Computer Science and Engineering (UBMK)*, ed. M. Arslan, M. Guzel, M. Demirci and S. Ozdemir, IEEE, 2019.
37 F. S. Melícias, T. F. Ribeiro, C. Rabadão, L. Santos and R. L. D. C. Costa, GPT and Interpolation-Based Data Augmentation for Multiclass Intrusion Detection in IIoT, *IEEE Access*, 2024, **12**, 17945–17965.
38 A. Gramacki, Kernel Density Estimation, *Nonparametric Kernel Density Estimation and Its Computational Aspects*, ed. Gramacki A., Cham, Springer International Publishing, 2018, p. 25–62.
39 Y.-C. Chen, A tutorial on kernel density estimation and recent advances, *Biostat. Epidemiol.*, 2017, **1**, 161–187.
40 B. He, D. J. Armaghani, S. H. Lai, P. Samui and E. T. Mohamad, Applying data augmentation technique on blast-induced overbreak prediction: Resolving the problem of data shortage and data imbalance, *Expert Syst. Appl.*, 2024, **237**, 121616.
41 P. Sedgwick, Pearson's correlation coefficient, *BMJ Br. Med. J.*, 2012, **345**, e4483.
42 J. Hauke and T. Kossowski, Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data, *Quaest. Geogr.*, 2011, **30**, 87–93.
43 G. J. Székely and M. L. Rizzo, Partial Distance Correlation with Methods for Dissimilarities, *Ann. Math. Stat.*, 2014, **42**, 2382–2412.

© 2026 The Author(s). Published by the Royal Society of Chemistry