GENERATIVE ADVERSARIAL NETWORKS FOR DATA AUGMENTATION AND INVERSE DESIGN OF SYNTHESIS CONDITIONS IN PEROVSKITE SOLAR CELLS

Daniel Cerro-Ramos, Mónica Botero-Londoño & F. Alexander Sepúlveda
School of Electrical, Electronics and Telecommunications Engineering
Universidad Industrial de Santander (www.uis.edu.co),
680006 Bucaramanga-Colombia,
daniel2258050@correo.uis.edu.co,
franklin@e3t.uis.edu.co

ABSTRACT

Modeling the complex relationships among synthesis parameters, material compositions, and performance metrics is essential for accelerating the development of perovskite solar cells (PSCs). While common approaches utilize discriminative models, this study adopts Generative Adversarial Networks (GANs) for modeling the underlying data distribution. In this work, we evaluate this generative framework on two tasks. First, we utilize an unconditional GAN for data augmentation to densify the experimental manifold. Second, to enable targeted inverse design, we implement a Conditional GAN (cGAN) based on a Weighted AC-GAN architecture with an inverse frequency-based loss weighting strategy. Results show that, regarding data augmentation, our method reduces the root mean square error (RMSE) in predictive tasks by 7.1%. Concerning inverse design, our proposed model enables the generation of synthesis recipes, even for high-efficiency targets, offering a new method to accelerate the discovery of perovskite-based photovoltaic devices.

1 INTRODUCTION

Perovskites have become one of the best options for next-generation photovoltaics due to their low cost and high performance, making them an alternative to silicon-based technology. Perovskite solar cells (PSCs) could become a serious competitor in the PV market now that single-junction certified efficiencies have recently reached 27%[1]. However, optimizing the structure, materials, and fabrication configuration is a complex task, making traditional methods time-consuming and expensive (Tao et al., 2021).

To tackle this problem, the field has been adopting and using machine learning (ML) techniques that are able to perform tasks such as performance prediction, feature importance analysis, data generation, and inverse design (Chen et al., 2023; Shrivastav et al., 2024). However, having access to high-quality and well-organized data is a key requirement for successfully implementing machine learning models. The Perovskite Database Project Jacobsson et al. (2022); Unger & Jacobsson (2022) has the most complete dataset on real experimental perovskite devices. It has device-level data for more than 42,000 metal-halide perovskite solar cells taken from peer-reviewed literature (Jacobsson et al., 2022).

Among ML tasks using the data in Jacobsson et al. (2022), the most common are: performance and stability prediction (Yang et al., 2023a; 2024; Hu et al., 2024; Roberts et al., 2024; Ping et al., 2025; Shi et al., 2025; Chen et al., 2025), detecting devices with high power conversion efficiency (Hu et al., 2024; Ping et al., 2025; Kusuma et al., 2025), and feature importance analysis (Vélez et al., 2024; Hu et al., 2024; Liu et al., 2024; Chen et al., 2025; Ping et al., 2025; Shi et al., 2025). In contrast, PSCs data generation has been scarcely reported. Iranipour et al. (2025) generated synthetic

[1] https://www.nrel.gov/pv/cell-efficiency

data using autoencoders and cGANs, which were then incorporated into a small existing dataset from research articles in order to improve the accuracy of deep learning models.

Regarding the use of ML for PSC inverse design, a few works have been reported. Chenebuah et al. (2024) propose a generative model, based on Variational Autoencoders (VAE) and GANs, for designing perovskite materials in accordance with geometrical and thermodynamic stability constraints. The resulting material candidates were validated using simulated data from Density Functional Theory (DFT). The study in Li et al. (2024) was also based on a numerical simulation dataset. However, simulations may not be able to capture the intricacies in the synthesis process of PSC present in the fabrication process; thus, experimental data are preferred when the interest is on determining optimal synthesis conditions.

The model presented in Lu et al. (2022) was trained on a dataset of about 1200 real experimental values of band gap taken from reported research articles. Bayesian Optimization (BO) is another very promising approach that is used to run a "closed-loop" experimental workflow (Wu et al., 2024). Finally, the work in Sepúlveda et al. (2026) applies a combination of probability density function estimation, based on mixtures of Gaussians, with optimization strategies to carry out the inverse modeling of experimental synthesis conditions.

In this study, we propose using GANs for data augmentation and the inverse design of synthesis conditions. Unlike Gaussian Mixture Models (GMMs) or Variational Autoencoders (VAEs), which often impose restrictive statistical assumptions or suffer from distribution smoothing (Bond-Taylor et al., 2021; Xu et al., 2019), GANs implicitly learn the complex, non-linear manifolds inherent to experimental fabrication data. Furthermore, while Diffusion Models suffer from slow iterative denoising (Yang et al., 2023b), GANs provide rapid, single-step sampling and a structured latent space, making them highly efficient for the fast querying required in inverse design loops. Because standard generative models often overlook rare, high-performance samples, we implement an Auxiliary-Classifier GAN (AC-GAN) with an inverse frequency weighting mechanism. This transforms the model into a prescriptive tool capable of proposing novel recipes for high-efficiency targets. Empirically, our approach reduces regression error—outperforming baseline and GMM-augmented models—and successfully generates promising new synthesis configurations.

2 METHOD

2.1 DATA

The original experimental data was taken from *Perovskite Database Project*(2022) Jacobsson et al. (2022); and, the curated dataset was previously introduced in Sepúlveda et al. (2026), which consists of $N = 5252$ samples. In this study, $N_{train} = 4201$ and $N_{test} = 1051$ samples are used for training and testing sets, respectively. Each sample vector $x \in \mathbb{R}^{10}$ from the data contains 9 synthesis parameters and the Power Conversion Efficiency (PCE) as the target variable. These parameters are described in §A. In the present work, we focus on perovskite solar cell (PSC) devices with power conversion efficiencies (PCE) greater than 10%, as in Lu et al. (2023).

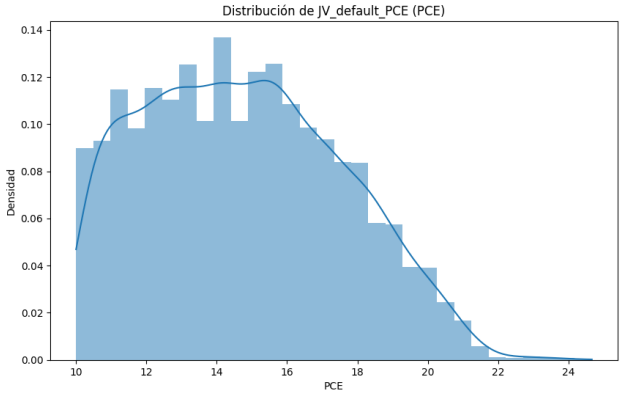


Figure 1: **Experimental Efficiency Distribution.** Histogram and KDE for PCE $> 10\%$.

A histogram with Kernel Density Estimation (KDE) of the PCE from the experimental dataset is shown in Figure 1, which reveals a distribution centered at approximately $14 - 15\%$. The bin width was automatically determined via the Freedman-Diaconis estimator ($\approx 0.5\%$). In addition, it can be observed a data scarcity in the high-performance region ($PCE > 18\%$), highlighting the difficulty of modeling optimal devices. Thus, a weighting strategy is required. This observation is important: in standard training (without weighting), the loss function would be dominated by the mode of the distribution (14-16%), leading the generator to ignore chemical configurations yielding high efficiencies. Consequently, we implemented an *Inverse Frequency Weighting*.

2.2 GANs AND C-GANs

In this work, we first implemented a standard Generative Adversarial Network (GAN) for generating synthetic-but-physically-plausible samples. It could enhance the generalization capability of predictive models without the cost of additional laboratory experiments. The objective is to learn the distribution $P_{data}(x)$ by minimizing the Adversarial Loss (Goodfellow et al., 2014):

$$\min_G \max_D L_{Adv}(D, G) = \mathbb{E}_{x \sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim P_z(z)}[\log(1 - D(G(z)))] \tag{1}$$

For inverse design, we leverage the *Conditional GANs (cGANs)* concept (Mirza & Osindero, 2014); that is, assuming a target PCE, we use the estimated conditional probability to infer experimental configurations. However, standard conditional models are frequently prone to *mode collapse* (Salimans et al., 2016). This phenomenon occurs when the generator fails to capture the full diversity of the data distribution, converging instead to a limited set of samples typically located around the statistical mean. Thus, to improve stability and correct this bias, we propose to use the *Auxiliary Classifier GAN (AC-GAN)* architecture (Odena et al., 2017). In this scheme, the discriminator includes an auxiliary output that predicts the material's efficiency, transforming the problem into a multi-task assignment (real/fake discrimination + performance regression).

In addition, we propose to apply *Inverse Frequency Weighting* to overcome data imbalance. In particular, we introduce a re-weighting mechanism in the auxiliary loss function assigning a scalar weight $w_i \in \mathbb{R}$ to each sample i, inversely proportional to the probability density of its corresponding target c_i (PCE value):

$$w_i = \frac{1}{P(c_i) + \epsilon} \tag{2}$$

Where $P(c_i)$ is the density estimated by KDE and ϵ is a smoothing factor. This weight modifies the global objective function (L_{Total}), defined as:

$$L_{Total} = \underbrace{L_{Adv}}_{\text{Eq. 1}} + \lambda_{aux} \frac{1}{N} \sum_{i=1}^{N} w_i \cdot L_{aux}(c_i, D_{aux}(G(z, c_i))) \tag{3}$$

where, c_i represents the continuous condition label (PCE value) associated with sample i; L_{aux} is defined as the *Mean Squared Error (MSE)*, $L_{aux}(y, \hat{y}) = ||y - \hat{y}||^2$, suitable for the efficiency regression task; and, λ_{aux} is the hyperparameter that balances the importance between visual fidelity and condition accuracy.

2.3 ARCHITECTURE AND TRAINING DETAILS

To ensure reproducibility and provide a comprehensive description of the generative framework, this section details the network configurations and stability strategies. Both the unconditional GAN and the Weighted AC-GAN utilize deep fully connected architectures designed to capture the underlying topology of the 10-dimensional experimental manifold. These configurations are summarized in Table 1, with full implementation details available in Section 5. Input features were pre-processed using a *Data Scaling* strategy where synthesis parameters were normalized to the range $[-1, 1]$ via a MinMaxScaler to maintain consistency with the *Tanh* activation in the Generators' output layer. For the AC-GAN, the target PCE was normalized to $[0, 1]$ to facilitate stable convergence during the regression task. During training, we implemented the following stability mechanisms to prevent common failure modes such as mode collapse and vanishing gradients: **(1) Label Smoothing:**

Instead of hard binary labels, we employed soft labels (0.95 for real and 0.05 for fake). This prevents the Discriminator from becoming overly dominant, ensuring a continuous gradient flow for the Generator. **(2) Dropout:** A Dropout rate of 0.4 was applied across all layers of the AC-GAN Discriminator. This acts as a regularizer, preventing the model from memorizing the training set and enhancing the generalization of generated recipes during inverse design. **(3) Weight Rescaling:** To accommodate the high auxiliary loss weight ($\lambda_{aux} = 150$), the importance weights w_i (defined in Eq. 2) were normalized to a mean of 1.0. This prevents gradient explosion while allowing the model to prioritize high-efficiency, low-density regions. Finally, both models were optimized using the Adam algorithm ($\alpha = 0.0002, \beta_1 = 0.5, \beta_2 = 0.999$). The unconditional GAN was trained for 3000 epochs, while the Weighted AC-GAN required 5000 epochs to ensure precise conditioning in the high-performance regime.

Table 1: Architectural Hyperparameters and Network Sizes.

Parameter	**Unconditional GAN**	**Weighted AC-GAN**
Latent Dimension (z)	128	128
Conditioning (c)	None	Concatenation ($z + c$)
Generator Layers	256, 512, 1024	256, 512, 1024
Discriminator Layers	1024, 512, 256	1024, 512, 256 (Backbone)
Normalization	None	Batch Norm (Generator)
Regularization	None	Dropout 0.4 (Discriminator)
Stability	Soft Labels (0.95/0.05)	Weight Rescaling ($\bar{w} = 1$)
Batch Size	64	64

2.4 Baseline Model

Related works regarding the use of generative models for real experimental data of perovskite solar cells are scarce. Among these few studies is the work reported by Sepúlveda et al. (2026), which utilized Gaussian Mixture Models (GMMs) for various tasks such as cluster discovery, missing data analysis, data generation, and inverse experiment design. Although GMMs suffer from several limitations compared to deep generative models, the aforementioned study demonstrated their capability for cluster discovery and, in conjunction with optimization, for inverse experiment design. Regarding data generation, in the present work, we compare their utility against GANs. GMM is a generative probabilistic model that explicitly represents the joint probability density function of the experimental data. In addition, it represents a pragmatic and interpretable approach specifically tailored for the scarce, low-dimensional tabular data characteristic of perovskite synthesis.

The probability density function (pdf) $f_{\mathbf{Z}}(\mathbf{z}) = f(\mathbf{z})$ can be explicitly represented by a linear combination of J Gaussian components of dimension d as follows,

$$\mathcal{P}(\mathbf{z}) = \sum_{j=1}^{J} \alpha_j \cdot \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}^{(j)}, \mathbb{C}^{(j)}) \tag{4}$$

where, $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}^{(j)}, \mathbb{C}^{(j)})$ is a normal joint probability density function with mean $\boldsymbol{\mu}^{(j)}$ and covariance matrix $\mathbb{C}^{(j)}$. $\boldsymbol{\mu}^{(j)}$ is the j_{th} d-dimensional vector of the j_{th} Gaussian component; and, each $\mathbb{C}^{(j)}$ is a matrix of dimension $d \times d$. In addition, $0 \leq \alpha_j \leq 1$ with $\sum_{j=1}^{J} \alpha_j = 1$.

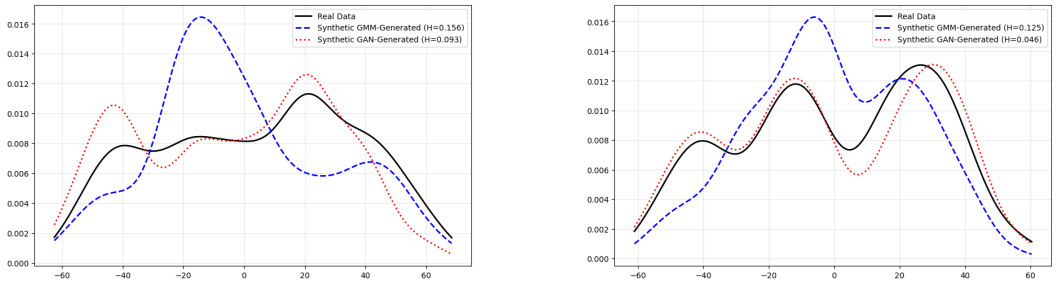
3 Results

3.1 Data Generation Performance Evaluation

In this study, we have 10-dimensional tabular data; consequently, the evaluation strategy for our proposed generative model should take into account this fact. In particular, the multimodal nature of our data Sepúlveda et al. (2026) affects the use of the Fréchet Distance, which is standard in image analysis but relies on Gaussian assumptions. Therefore, we propose a dual approach: employing manifold learning techniques for qualitative visual inspection and Maximum Mean Discrepancy (MMD) for quantitative analysis.

3.1.1 QUALITATIVE INSPECTION.

t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008) was employed to project the high-dimensional feature space ($\mathbb{R}^{10}$) into an observable 2D plane. As illustrated in Figure 2, we compare the marginal distributions of the resulting embedding for the proposed GAN versus the GMM baseline. It is evident that the GAN exhibits a distribution density that closely adheres to the real data; in contrast, the GMM tends to concentrate density excessively around the modes, failing to capture the dispersed structure of the manifold. The alignment in these marginals provides preliminary qualitative evidence that the GAN has correctly learned the underlying topology of the data.



(a) t-SNE Dimension 1

(b) t-SNE Dimension 2

Figure 2: **t-SNE Manifold Visualization.** Marginal KDEs demonstrate that the GAN (red/dotted) captures the real data distribution (black/solid) with higher fidelity than the GMM (blue/dashed).

3.1.2 QUANTITATIVE ASSESSMENT.

We calculated the *Hellinger Distance* (Gibbs & Su, 2002) to quantify density similarity. Having the distributions P and Q of the real and synthetic datasets, respectively; the *Hellinger Distance* is defined as (Basu & Lindsay, 1994),

$$H^2(P,Q) = \frac{1}{2}\int \left(\sqrt{dP(x)} - \sqrt{dQ(x)}\right)^2 \tag{5}$$

An important property that helps for interpretation is $0 \leq H(P,Q) \leq 1$, where a value of 0 implies identical distributions, indicating optimal performance of the generative model. Given that this metric is highly sensitive to the number of bins and dimensionality, we validated it across multiple configurations using Principal Component Analysis (PCA) to reduce the original space to 2, 4, and 6 dimensions, and discretizing with 5, 8, and 12 bins (see Table 2). The results show that while the GMM performs better in a low-complexity space (PCA=2), the GAN consistently outperforms the GMM as dimensionality increases, indicating a superior ability to capture complex, multi-dimensional correlations. For our baseline configuration (8 bins, 4 PCA components), we incorporated a bootstrap analysis ($N = 50$ iterations) to estimate 95% confidence intervals (CI). We obtained a mean Hellinger distance of 0.4741 (95% CI: $[0.3586, 0.5672]$) for the GMM and 0.3327 (95% CI: $[0.2790, 0.3837]$) for the GAN.

Table 2: Hellinger distance validation across different PCA dimensions and histogram bins.

	PCA = 2		**PCA = 4**		**PCA = 6**	
Bins	**GMM**	**GAN**	**GMM**	**GAN**	**GMM**	**GAN**
5	0.0887	0.1548	0.3012	0.2012	0.3522	0.2740
8	0.1303	0.1731	0.4123	0.2170	0.4770	0.3028
12	0.1859	0.1927	0.4534	0.2820	0.5925	0.3797

However, a better measure corresponds to the Maximum Mean Discrepancy (MMD) Statistic, which does not assume the data follows any specific distribution, and compares the statistical moments (not

just mean and variance) directly in the 10-dimensional space, without estimating any probability density. This metric quantifies the distance between the mean embeddings of distributions P and Q in a Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}$, as follows,

$$\text{MMD}^2(P, Q) = \|\mathbb{E}_{x\sim P}[\phi(x)] - \mathbb{E}_{y\sim Q}[\phi(y)]\|^2_{\mathcal{H}} \tag{6}$$

Regarding implementation, instead of relying on a single kernel, we employed a Multi-Kernel approach to capture discrepancies across various resolution scales. We used a linear combination of isotropic Radial Basis Function (RBF) kernels, $k(x, y) = \exp(-\gamma||x - y||^2)$. Crucially, the determination of the base bandwidth hyperparameter γ is vital for the validity of the metric. To ensure an unbiased resolution scale, a base $\gamma_{base} \approx 1.94 \times 10^{-3}$ was calculated using the *median heuristic* (Gretton et al., 2012) over the aggregated dataset ($X_{real} \cup X_{synth}$). We then utilized a set of gamma γ values corresponding to various multipliers applied to γ_{base}, as detailed in Table 3, to evaluate both global structures and fine-grained local details. To report uncertainty, a bootstrap analysis revealed a Multi-Kernel MMD of 0.0542 (95% CI: $[0.0422, 0.0695]$) for the GMM, and a significantly lower 0.0082 (95% CI: $[0.0039, 0.0147]$) for the GAN. The complete lack of overlap between these intervals provides strong statistical evidence that favors GANs.

Table 3: Bandwidth parameters (γ) for the Multi-Kernel MMD, derived from the base median heuristic ($\gamma_{base} \approx 1.94 \times 10^{-3}$).

Kernel	**Multiplier** (c)	$\gamma = c \times \gamma_{base}$	**Resolution Scale Captured**
1	0.1	0.0002	Global structure (broad differences)
2	0.5	0.0010	Intermediate-global features
3	1.0	0.0019	Base scale (median pairwise distance)
4	2.0	0.0039	Intermediate-local features
5	10.0	0.0194	Local structure (fine-grained details)

In addition, we applied the *energy distance test of homogeneity* (Ramos et al., 2023), using the same Multi-Kernel MMD statistic, in order to analyze if the two samples (e.g., real vs. generated data) are drawn from the same underlying probability distribution. The null hypothesis corresponds to $\text{H}_0 : P = Q$ (*random vectors have the same distribution*), for which high values in its statistic are evidence that favors the alternative hypothesis $\text{H}_1 : P \neq Q$ (*random vectors are drawn from different distributions*). For both models, H_0 was rejected at a significance level of $\alpha = 0.05$, yielding the minimum possible p-value $= 0.0010$ after $N = 1000$ permutations. Consequently, we can conclude that the synthetic and real distributions are statistically different. Nevertheless, we can utilize the empirical test statistic to quantify and compare the relative distributional discrepancy: the MMD statistic of the GAN (0.0125) is approximately four times lower than that of the GMM (0.0464), reinforcing the superiority of the GAN.

3.1.3 Diversity, Memorization, and Correlation Preservation.

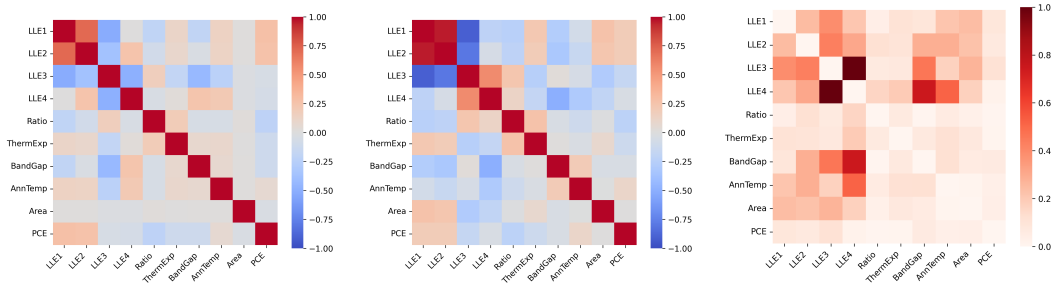
Beyond marginal distributions, evaluating a generative model requires assessing its ability to produce diverse, novel samples while preserving multi-dimensional feature dynamics. To this end, we computed the Distance to Closest Record (DCR), the 1-Nearest Neighbor Adversarial Accuracy (1-NN AA), and the pairwise correlation Mean Absolute Error (MAE) between the training dataset ($N = 4201$) and the synthetic GAN dataset ($N = 1501$).

First, to ensure the model acts as a true generator rather than a database memorizer, we calculated the Euclidean distance from each synthetic record to its nearest real neighbor in a standardized space. The GAN yielded zero exact copies (0.00%), with a mean DCR of 0.2641. This confirms that the model successfully safeguards data novelty, generating entirely new samples rather than merely overfitting and regurgitating the training instances.

Second, we evaluated the 1-NN AA using a 5-fold cross-validation setup to determine if a classifier could distinguish between real and synthetic records. The GAN achieved an accuracy of 0.7788. This result indicates that while the synthetic samples possess detectable artificial signatures in the high-dimensional space, the model remains safely above the overfitting threshold, demonstrating a robust balance between data fidelity and generation diversity.

Finally, we assessed how well the synthetic data maintains the underlying relationships between the 10 variables by calculating the absolute difference between the Pearson correlation matrices of

the real and synthetic sets (see Figures 3a-c). The model achieved a Correlation MAE of 0.1851. This low error margin, visually represented by the difference heatmap (Figure 3c), indicates that the GAN effectively captured the linear inter-variable dependencies, ensuring that the synthetic dataset respects the physical and statistical constraints of the original perovskite domain.



(a) Real Data (b) GAN Synthetic Data (c) Absolute Difference

Figure 3: **Pearson Correlation Matrices.** Comparison between (a) real training data, (b) GAN synthetic data, and (c) their absolute difference. Key variables: Ratio (DMF/DMSO), ThermExp (thermal exposure), AnnTemp (1st annealing temp.), BandGap, Area, and PCE.

3.2 Data Augmentation on Regression

We analyze the utility of generating synthetic PSCs samples in a regression task. In particular, an XGBoost model was trained to predict PCE using a 5-Fold Cross-Validation scheme. This validation was performed on the 5252 available real samples, allocating 80% of the data for training and 20% for testing in each fold. To ensure a rigorous evaluation and prevent data leakage, the synthetic samples (generated by both the GAN and GMM models) were introduced exclusively into the training partitions strictly after the cross-validation splits were defined. XGBoost is regarded among the best accuracy in PCE prediction tasks. We compared the baseline experiment (training a regression model with only real data) versus real plus augmented data. The results, presented in Table 4, reveal a systematic improvement across evaluation metrics when incorporating data augmented by our GAN model. A decrease in Root Mean Square Error from 2.064 to 1.918 is observed. This represents a relative error reduction of 7.1%. The Mean Absolute Percentage Error decreased from 0.116 to 0.106, that is, an 8.6% reduction in the model's relative error. Conversely, augmenting the training set with samples generated by a Gaussian Mixture Model (GMM) resulted in a performance degradation compared to the baseline. The GMM-augmented model yielded an R^2 of 0.403, an RMSE of 2.169, and a MAPE of 0.122. This contrast highlights the GAN's superior ability to capture and reproduce the underlying physical distributions required for accurate PCE prediction.

Table 4: *Performance on Regression Analysis.* Comparison of predictive metrics (R^2, RMSE, MAPE) across the baseline model (real data only) and models augmented with GMM and GAN synthetic data.

Metric	Baseline (XGBoost)	GMM Augmented	GAN Augmented	Improvement (GAN)
R^2	0.457	0.403	0.504	10.3%
RMSE	2.064	2.169	1.918	7.1%
MAPE	0.116	0.122	0.106	8.6%

3.3 Targeted Inverse Design of Synthesis Parameters

We now evaluate the Weighted AC-GAN architecture for the inverse design of synthesis conditions. To estimate the Power Conversion Efficiency (PCE) of the generated recipes, two independent "Virtual Laboratories" based on XGBoost and Random Forest (RF) models were employed. The predictive reliability of these baseline laboratories was first characterized: the XGBoost model achieved a maximum coefficient of determination (R^2) of 0.5416 with an associated Root Mean Square Error (RMSE) of 1.90, while the RF model achieved an R^2 of 0.5716 and an RMSE of 1.85.

The generative model was then conditioned to generate recipes given the target PCE ranging from 10% to 21%. A Monte Carlo simulation was performed where $N = 1000$ samples were obtained for each PCE target value, and subsequently evaluated by both Virtual Laboratories. The continuous validation yielded a Global Mean Absolute Error (MAE) of 1.04% with a target-prediction correlation (R) of 0.98 using the XGBoost evaluator. Similarly, a Global MAE of 1.21 with an R of 0.987 was obtained using the RF evaluator. Considering the intrinsic experimental variability of perovskites, an error of about 1% in PCE validates the practical utility of the model for synthesis screening. The scatter plot of these two estimations is shown in Figure 4.

The overall behavior can be divided into three distinct zones:

$12 - 16\%$**:** In this interval, the expected performance of generated devices (blue line for XGBoost and green line for RF) simulated by our "Virtual Laboratories" closely aligns with the targeted PCEs (dashed gray line). This indicates that the Weighted AC-GAN, on average, is able to provide the targeted PCE values.

$> 17\%$**:** Above 18%, the predicted PCEs begin to "saturate" and fall off the target line; subsequently, the response flattens at about $18.2\% - 18.5\%$. Far from being an error, this behavior indicates that the model has internalized the physical limits of the explored materials space; rather, it provides the best feasible configurations within the explored space.

$\leq 12\%$**:** In this zone, the generated devices tend to exhibit an optimistic bias, predicting baseline values around 12.2% to 12.6% even for a 10% target.

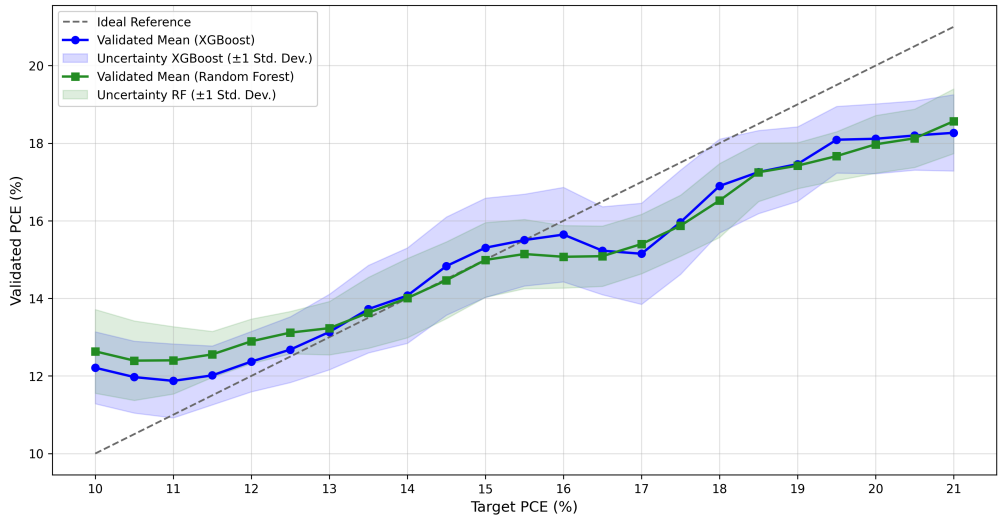


Figure 4: **Continuous Validation of Inverse Design.** Trajectory of validated PCE versus requested target for both XGBoost and RF models. The shaded region represents the standard deviation of the predicted PCE across the 1,000 distinct synthetic recipes generated for each target PCE ($\pm 1\sigma$)

On the other hand, at extreme targets such as 21.0%, the model generates recipes with a predicted mean of approximately 18.27% (XGBoost) and 18.57% (RF). It is important to remark that this target value lies in the top percentile ($> 95\%$) of the original distribution, confirming the positive effect of applying the weighted loss function (Eq. 3) in counteracting mode collapse, thereby forcing the generator to explore and exploit low-density regions (high efficiency) rather than converging towards the statistical mode of the dataset.

An internal ablation study further validated this architecture. Without inverse frequency weighting, the generator exhibited a strong bias towards the dataset's statistical mode, predominantly yielding recipes around 14.5% PCE regardless of the target. Similarly, removing the auxiliary classifier caused a complete loss of conditional control; lacking feedback to map synthesis parameters to target efficiencies, the generator again reverted to mean performance. These results underscore that integrating both components is essential to explore and exploit low-density, high-efficiency regions.

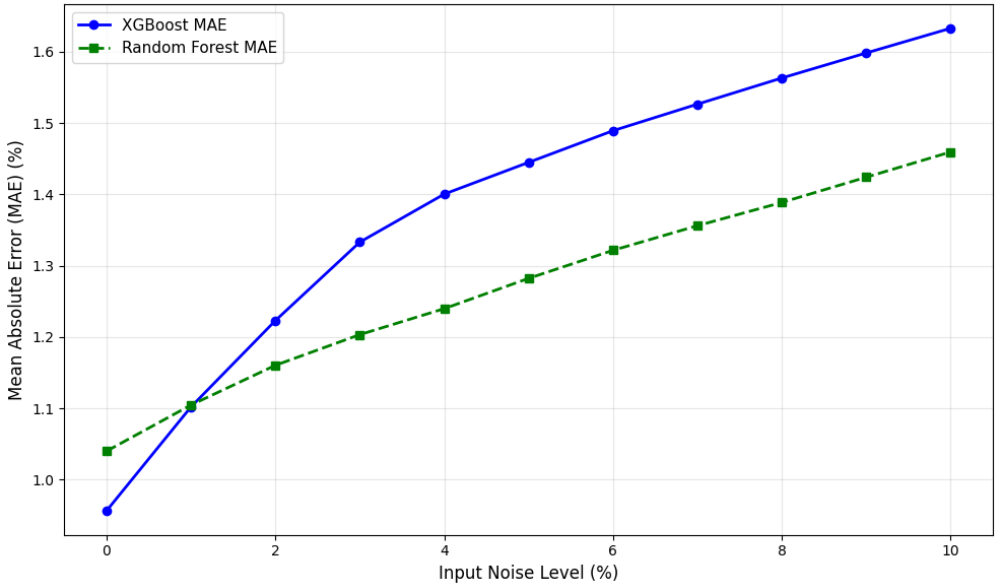


Figure 5: **Robustness Analysis.** Model degradation under increasing input noise levels ($0 - 10\%$). The plot tracks the increase in Mean Absolute Error (MAE) when evaluating the noisy synthetic recipes using both the XGBoost and Random Forest Virtual Laboratories.

Finally, to assess the resilience of the inverse design framework against inevitable experimental uncertainties and synthesis deviations found in laboratories, we conducted the following experiment: Random noise ranging from 0% to 10% was introduced into the synthesis parameters during a Monte Carlo simulation ($N = 1000$ iterations per target). This range was selected to emulate the inherent uncertainties and cumulative errors encountered in physical fabrication processes—such as measurement tolerances and precursor concentration variations. By testing this interval, we provide a representative scenario to evaluate how such experimental deviations affect the model's predictive reliability. Results are depicted in Figure 5. As expected, the Mean Absolute Error (MAE) increases with the noise level for both evaluation models. The XGBoost simulator (blue solid line) exhibits an MAE ranging from approximately 0.96% at baseline to 1.63% under maximum noise conditions (10%). Similarly, the Random Forest simulator (green dashed line) shows a steady degradation, with its MAE increasing from 1.04% to 1.46%. Despite this loss in precision due to perturbations, the error margins remain relatively constrained, demonstrating that the recipes proposed by the Weighted AC-GAN maintain a reasonable degree of robustness against experimental variations.

4 Conclusions and Future Work

We introduced a robust generative framework that effectively overcomes the small data constraints found in perovskite solar cell research based on experimental data. Our results show that synthetic data augmentation improves the generalization and predictive accuracy of regression tasks that come after it. However, the most important contribution is that our weighted AC-GAN architecture gets rid of statistical bias when estimating the conditional probability distribution for the case of inverse design, which makes it possible to explore high-efficiency regimes ($> 18.5\%$) in a targeted way. While the physical fabrication and experimental validation of the proposed solar cells remain the ultimate goal, such laboratory implementation is beyond the scope of this computational study.

Regarding future directions, we identify a critical open question concerning uncertainty estimation. Our results showed constant variance throughout the validation range—a counterintuitive finding, given that uncertainty should theoretically increase in high-efficiency regions where information is inherently scarcer. On the other hand, Local Linear Embeddings (LLE) were used for the perovskite material representation, but it is not strictly invertible. This makes it harder to include the materials in the inverse design process.

5 Data and Software Availability

The complete dataset has been downloaded from the *Perovskite Database Project* at `https://www.perovskitedatabase.com/`; and, the curated dataset used for the analysis is available at `https://doi.org/10.5281/zenodo.16809654`. The code used for running the analysis and for generating the figures are available at `https://github.com/danielcerro7/GAN-Augmentation-Inverse-PSC`.

6 Acknowledgment

We gratefully acknowledge the financial support of *Ministerio de Ciencia, Tecnología e Innovación* (*MinCiencias*-Colombia, `https://minciencias.gov.co/`) under contract 2022-0724 (Call 890), with resources administered by *ICETEX*-Colombia.

References

Ayanendranath Basu and Bruce G Lindsay. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46: 683–705, 1994.

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7327–7347, 2021.

Chen Chen, Ayman Maqsood, and T Jesper Jacobsson. The role of machine learning in perovskite solar cell research. *Journal of Alloys and Compounds*, 960:170824, 2023.

Jiacheng Chen, Yaohui Zhan, Zhenhai Yang, Yue Zang, Wensheng Yan, and Xiaofeng Li. Predicting and analyzing stability in perovskite solar cells: Insights from machine learning models and SHAP analysis. *Materials Today Energy*, 48:101769, 2025.

Ericsson Chenebuah, Michel Nganbe, and Alain Tchagang. A deep generative modeling architecture for designing lattice-constrained perovskite materials. *npj Computational Materials*, 10, 2024.

Alison L Gibbs and Francis E Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Jinghao Hu, Zhengxin Chen, Yuzhi Chen, Hongyu Liu, Wenhao Li, Yanan Wang, Lin Peng, Xiaolin Liu, Jia Lin, Xianfeng Chen, et al. Interpretable machine learning predictions for efficient perovskite solar cell development. *Solar Energy Materials and Solar Cells*, 271:112826, 2024.

Behzad Iranipour, Mohammadreza Sadeghian, and Ezeddin Mohajerani. Artificial data generation: A strategy to improve efficiency predictions in mixed Sn-Pb perovskite solar cells. *Materials Today Communications*, 43:111625, 2025.

T Jesper Jacobsson, Adam Hultqvist, Alberto García-Fernández, et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nature Energy*, 7: 107–115, 2022.

Frendy Jaya Kusuma, Eri Widianto, Wahyono, Iman Santoso, Sholihun, Moh Adhib Ulil Absor, Setyawan Purnomo Sakti, and Kuwat Triyana. Optimizing novel device configurations for perovskite solar cells: Enhancing stability and efficiency through machine learning on a large dataset. *Renewable Energy*, 247:122947, 2025.

Zong-Zheng Li, Chaorong Guo, Wenlei Lv, Peng Huang, and Yongyou Zhang. Machine learning-enabled optical architecture design of perovskite solar cells. *The Journal of Physical Chemistry Letters*, 15:3835–3842, 2024.

Yiming Liu, Xinyu Tan, Peng Xiang, Yibo Tu, Tianxiang Shao, Yue Zang, Xiong Li, and Wensheng Yan. Machine learning as a characterization method for analysis and design of perovskite solar cells. *Materials Today Physics*, 42:101370, 2024.

Tian Lu, Hongyu Li, Min Li, Shenghao Wang, and Wencong Lu. Inverse design of hybrid organic–inorganic perovskites with suitable bandgaps via proactive searching progress. *ACS Omega*, 7, 2022.

Y Lu, X Chen, Y Wang, and Y Li. Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results. *Journal of Energy Chemistry*, 77:200–208, 2023.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning (ICML)*, pp. 2642–2651. PMLR, 2017.

Jiayi Ping, Pei Wang, Bowen Liu, Hairui Zhou, Yuanhua Li, and Jia Lin. A hierarchical HGBT-based machine learning framework for predicting the power conversion efficiency of perovskite solar cells. *Materials Today Communications*, 46:112747, 2025.

D Ramos et al. The energy distance test of homogeneity. *SoftwareX*, 22:101375, 2023.

Nicholas Roberts, Dylan Jones, Alex Schuy, Shi-Chieh Hsu, and Lih Y Lin. Machine Learning for Perovskite Solar Cells: An Open-Source Pipeline. *Advanced Physics Research*, 3(11), 2024.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29, 2016.

F Alexander Sepúlveda, Daniel Cerro-Ramos, and T Jesper Jacobsson. Density estimation based on mixtures of Gaussians for perovskite solar cells modeling. *Journal of Chemical Information and Modeling*, 0, 2026. In Press.

Yudong Shi, Jiansen Wen, Cuilian Wen, Linqin Jiang, Bo Wu, Yu Qiu, and Baisheng Sa. Interpretable machine learning insights of power conversion efficiency for hybrid perovskites solar cells. *Solar Energy*, 290:113373, 2025.

Nikhil Shrivastav, Jaya Madan, and Rahul Pandey. Predicting photovoltaic efficiency in Cs-based perovskite solar cells: A comprehensive study integrating SCAPS simulation and machine learning models. *Solid State Communications*, 380:115437, 2024.

Qiuling Tao, Pengcheng Xu, Minjie Li, and Wencong Lu. Machine learning for perovskite materials design and discovery. *npj Computational Materials*, 7(23), 2021.

Eva Unger and T Jesper Jacobsson. The Perovskite Database Project: A Perspective on Collective Data Sharing. *ACS Energy Letters*, 7(3):1240–1245, 2022.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

Jeison Vélez, Mónica A Botero L, and Alexander Sepulveda. Measurement of information content of Perovskite solar cell's synthesis descriptors related to performance parameters. *Emergent Materials*, 7(5):1961–1968, 2024.

Jianchang Wu, Luca Torresi, ManMan Hu, et al. Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells. *Science*, 386(6727):1256–1264, 2024.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chao Yang, Xiaoyu Chong, Mingyu Hu, Wei Yu, Jingjin He, Yalan Zhang, Jing Feng, Yuanyuan Zhou, and Lin-Wang Wang. Accelerating the Discovery of Hybrid Perovskites with Targeted Band Gaps via Interpretable Machine Learning. *ACS Applied Materials & Interfaces*, 15(34): 40419–40427, 2023a.

Jiaqi Yang, Panayotis Manganaris, and Arun Mannodi-Kanakkithodi. Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm. *The Journal of Chemical Physics*, 160(6):064114, 2024.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023b.

A DESCRIPTORS

We adopt the descriptors described in (Sepúlveda et al., 2026):

- $L_1, \cdots, L_4$, component of *Local Linear Embedding* transformation representing the perovskite material.
- $\chi_{\text{sol.}}$. DMSO:DMF ratio, expressed in logarithmic scale, where DMSO and DMF (along with other solvents reported in the Perovskite Project Database) are used in the deposition of the perovskite layer.
- T_1. First temperature during thermal annealing process.
- TB. Thermal budget.
- $\mathcal{A}$. Cell area measured.
- E_g, Perovskite band gap.