IOSR Journal of Applied Physics (IOSR-JAP)
e-ISSN: 2278-4861.Volume 18, Issue 2 Ser. 1 (Mar – Apr 2026), PP 77-91
www.iosrjournals.org

SCAPS-1D-Guided Machine-Learning Surrogates for Lead-Free $PeDAMA_4Pb_5I_{16}/Cu_2BaSnS_3$ Photovoltaic Performance Prediction

[1]K. C. Dubey, [2]Susheel Kumar Singh, [2*]R. K. Shukla, [3]Shobhit Shukla and [2]Anchal Srivastava

[1]*Department of Physics, Shia P.G. College, Lucknow-226020*
[2]*Department of Physics, University of Lucknow, Lucknow-226007, INDIA*
[3]*Department of Information Technology, Dr. Shakuntala Misra National Rehabilitation University, Lucknow-226017, INDIA*

Abstract
Accurate surrogate models can accelerate photovoltaic device optimization when physics-based simulation data are expensive to generate or difficult to interpret across many design variables. This work presents a complete machine-learning analysis of SCAPS-1D simulated solar-cell outputs for a multilayer lead-free device architecture involving $PeDAMA_4Pb_5I_{16}$ and Cu_2BaSnS_3 absorber-related sweeps. Irregular SCAPS workbook exports were converted into a tidy dataset by pairing V_{oc}/J_{sc} and FF/PCE blocks according to the varied simulation parameter. The final dataset contained 270 complete records covering absorber thickness, absorber donor/acceptor concentration, defect density, ETL and HTL material variation, transport-layer thickness and doping, temperature, series resistance, and shunt resistance. Random Forest regression was selected as the primary nonlinear surrogate, with support vector regression, Ridge regression, and Lasso regression used as comparative baselines. Models were trained with an 80:20 train-test split, K-fold cross-validation, and grid-search hyperparameter optimization. The optimized Random Forest models achieved test-set R^2 values of 0.6677 for V_{oc}, 0.8423 for J_{sc}, 0.6564 for FF, and 0.8417 for PCE. The best SCAPS record reached 33.57% PCE, while the Random Forest prediction was 32.33%, corresponding to a relative error of 3.69%. Feature-importance analysis identified absorber family terms, active parameter magnitude, Cu_2BaSnS_3 defect density, shunt resistance, and layer-thickness variables as dominant performance drivers. The results indicate that machine-learning surrogates can reproduce major SCAPS trends and provide interpretable screening guidance, while also revealing the need for future multi-parameter simulation designs before claiming global device optimization capability.

Keywords: *SCAPS-1D, photovoltaic simulation, machine learning, Random Forest regression, support vector regression, lead-free solar cells, feature importance, power conversion efficiency.*

I. Introduction

Photovoltaic energy conversion remains one of the central engineering routes for decarbonized electricity generation, but the practical value of a candidate solar-cell stack depends on the co-optimization of optical absorption, carrier collection, recombination control, contact selectivity, and manufacturable layer design. The theoretical interpretation of photovoltaic conversion has long been anchored by detailed-balance considerations, which connect bandgap, radiative recombination, and the maximum achievable open-circuit voltage and power output [1]. Modern thin-film and perovskite devices extend this classical framework by introducing heterogeneous interfaces, strongly varying defect populations, and transport layers whose band alignment can determine whether photogenerated carriers are collected or lost to recombination.

Halide and hybrid perovskite photovoltaics have attracted sustained attention since their early use as visible-light absorbers [2]. The broader field has since expanded from lead-halide systems toward low-toxicity absorbers, inorganic alternatives, low-dimensional structures, and tandem-compatible architectures. Reviews of perovskite photovoltaic development emphasize that high performance is not only a materials problem; it is also a device-physics problem involving energy-level alignment, interface passivation, mobility balance, defect energetics, and contact engineering [3]. In this context, the present work focuses on a simulated lead-free multilayer architecture that includes $PeDAMA_4Pb_5I_{16}$ and Cu_2BaSnS_3 absorber-related parameter studies together with ZnO, Cu_2O, and alternative transport-layer screening. The physical question is whether the resulting SCAPS-derived performance landscape can be represented accurately by data-driven surrogates while retaining interpretable engineering insight.

Machine learning has become increasingly important in computational materials science because it can approximate expensive simulations, extract nonlinear relationships from high-dimensional data, and guide experiments toward regions of high expected utility [4]. For photovoltaic devices, the same logic is compelling: once a reliable simulation or measurement dataset exists, regression models can map device parameters to photovoltaic metrics such as open-circuit voltage, short-circuit current density, fill factor, and power conversion efficiency. However, solar-cell performance data are often not produced as conventional machine-learning tables. Numerical simulators are commonly used in one-factor sweeps, where each worksheet or export block varies one device parameter while other conditions are held fixed. This creates a practical data-curation challenge before any robust model can be trained.

SCAPS-1D is widely used for thin-film photovoltaic simulation because it solves the coupled semiconductor transport equations under user-defined layer, defect, and contact conditions [5], [6]. The tool is particularly useful for exploring trends in absorber thickness, doping density, defect density, transport-layer thickness, series resistance, shunt resistance, and temperature before experimental fabrication. Recent discussion in the SCAPS community has also stressed that simulation outputs must be interpreted critically: parameter choices, boundary conditions, and unrealistic optimization claims can lead to misleading efficiencies if the model is not physically constrained [7]. Consequently, a responsible machine-learning surrogate should not hide the simulation assumptions. It should instead preserve traceability to the underlying SCAPS sweeps and communicate the domain over which predictions are justified.

Recent SCAPS studies of lead-free inorganic perovskite and chalcogenide-related devices demonstrate the value of systematic simulation for absorber, interface, and transport-layer optimization [8], [9]. Those studies generally report individual optimized architectures, whereas this manuscript reconstructs a broader data-analysis workflow from a collection of SCAPS outputs. The resulting question is not only which parameter gives the highest simulated PCE, but also whether a machine-learning model can learn the dominant response functions from heterogeneous exported worksheets. This emphasis on end-to-end data extraction, modeling, validation, and scientific interpretation is important for turning simulator output into reproducible engineering knowledge.

The present study uses the scikit-learn ecosystem for reproducible regression modeling [10]. Random Forest regression is treated as the primary surrogate because decision-tree ensembles can represent nonlinear thresholds, saturation effects, and interactions without imposing a parametric form [11]. Support vector regression provides a nonlinear kernel-based comparator [12], while Ridge and Lasso regression provide regularized linear baselines [13], [14]. Hidden Markov Models are discussed only as a methodological boundary condition: they are powerful for sequential state inference [15], but the available SCAPS output consists of independent parameter sweeps rather than time-ordered degradation or process trajectories.

A further motivation for the present analysis is the need to distinguish between optimization claims and optimization evidence. In simulated photovoltaics, a high reported efficiency can arise from a narrow combination of assumptions, some of which may be difficult to realize experimentally. A machine-learning model trained on the full sweep history can reveal whether an optimum sits within a broad performance plateau or at an isolated edge of the explored parameter space. This distinction is crucial for device engineering because broad high-performance regions are usually more robust to fabrication tolerance, whereas sharp optima may disappear when interfacial recombination, mobility imbalance, or contact nonideality is included.

The studied dataset is also representative of a common situation in computational photovoltaic research: the simulator is used productively, but the resulting files are not immediately ready for statistical learning. Worksheet-level exports frequently contain blank rows, repeated headings, separate metric blocks, mixed material labels, and numerical values spanning many orders of magnitude. A publication workflow that moves directly from such files to regression plots risks silent data leakage or target misalignment. The present pipeline therefore treats data reconstruction as a scientific step rather than a clerical step.

The combination of $PeDAMA_4Pb_5I_{16}$- and Cu_2BaSnS_3-related sweeps is especially useful because it creates both high-performance and failure-mode examples. The dataset includes conditions where PCE approaches 33.57%, but it also includes severe degradation down to 0.11% PCE. This range forces the model to learn the difference between benign parameter variation and severe electrical loss. It also allows the discussion to connect numerical accuracy to device physics rather than treating R^2 as the only scientific outcome.

The gap addressed by this work is therefore threefold. First, many simulation papers report optimized SCAPS values but do not provide a reproducible path from irregular simulator exports to machine-learning tables. Second, many machine-learning studies emphasize prediction accuracy without sufficiently linking feature importance back to semiconductor physics. Third, comparisons among nonlinear and regularized regression models are often presented without explicitly describing the limits imposed by one-factor simulation designs. This manuscript addresses these issues by constructing a complete analysis pipeline from the project workbook, generated CSV outputs, figures, model files, and discussion notes.

The contributions of this paper are as follows. A SCAPS-derived workbook is converted into a clean 270-record dataset with complete V_{oc}, J_{sc}, FF, and PCE targets. A publication-ready Random Forest surrogate workflow is evaluated against SVR, Ridge, and Lasso baselines using held-out testing and cross-validation. Feature-importance and correlation analyses are interpreted in terms of absorber recombination, resistive loss, layer-thickness trade-offs, and transport-layer selection. Finally, the study identifies the limitations of one-factor SCAPS sweeps and proposes a path toward multi-parameter simulation campaigns, boosted-tree models, uncertainty-aware regression, and physics-informed neural surrogates.

The manuscript is organized as follows. Section II describes the extraction of the SCAPS outputs, the feature-engineering strategy, the semiconductor-performance quantities, and the regression-validation workflow. Section III reports descriptive statistics, correlation structure, comparative regression metrics, feature importance, residual diagnostics, and the SCAPS-versus-ML comparison for the best device. Section IV summarizes the findings, states the limitations of one-factor sweep data, and outlines future directions for physically constrained simulation-assisted photovoltaic optimization.

II. Methodology

The overall workflow is summarized before the detailed mathematical and computational description. Fig. 1 is cited here because it defines the sequence used throughout the study, from SCAPS simulation to final conclusions.

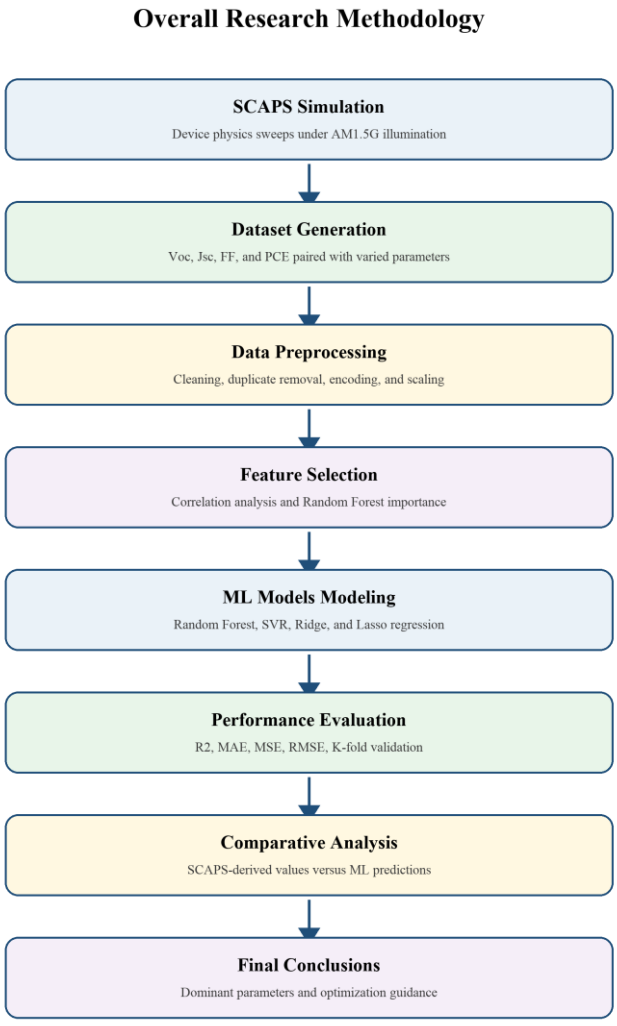

Fig. 1. Overall Research Methodology, The workflow begins with SCAPS-1D simulation, converts simulator outputs into a structured dataset, performs preprocessing and feature selection, trains regression models, evaluates predictive accuracy, compares SCAPS-derived and machine-learning values, and concludes with photovoltaic design insights.

The source data were stored in the workbook $PeDAMA_4Pb_5I_{16}$.xlsx. The worksheets consisted mainly of SCAPS one-factor parameter sweeps rather than a single design matrix. Several sheets contained one block for V_{oc} and J_{sc} and a separate block for FF and PCE. The analysis script therefore searched for metric headers, inferred the corresponding varied parameter label, parsed the numeric rows, and paired the two metric blocks by matching the varied parameter value. After duplicate removal and exclusion of incomplete target rows, the dataset contained 270 records. The code file available in the workspace is solar_scaps_ml_study.py; the requested filename solar_cell_ml_scaps_analysis.py was not present, and the requested ./work/outputs/ directory was likewise absent. All manuscript results were therefore taken from the available outputs/scaps_ml_study directory.

The canonical features created by the preprocessing routine include a categorical parameter identifier, a parameter-family label, the active numeric value, its base-10 logarithm for positive values, and one-hot material indicators for ETL and HTL screening. Separate numeric feature columns were generated for each physically meaningful parameter class, including $PeDAMA_4Pb_5I_{16}$ absorber thickness and doping, Cu_2BaSnS_3 absorber thickness and doping, absorber defect densities, transport-layer thicknesses, ETL donor concentration, HTL acceptor concentration, series resistance, shunt resistance, and temperature. This long-format representation is appropriate for one-factor sweep data because it preserves the identity of the varied variable while also exposing the active magnitude to the regression models.

The SCAPS simulation physics are represented indirectly through the exported performance metrics. For a thin-film solar cell, the one-dimensional electrostatic and transport problem is governed by Poisson's equation and the carrier-continuity equations. In compact notation, the simulator solves the spatially dependent electrostatic potential, carrier concentrations, current densities, generation, and recombination terms under illumination and bias. The photovoltaic outputs are then extracted from the simulated current-voltage curve. Power conversion efficiency is related to the principal device metrics by

The device parameters represented in the workbook can be grouped into five physical classes. Absorber parameters describe the optically active layer and include thickness, donor or acceptor concentration, and bulk defect density. Transport-layer parameters describe the selective extraction layers and include thickness, doping concentration, and material identity for the ETL and HTL. Resistive parameters describe parasitic electrical losses through the series and shunt resistances. Thermal parameters describe temperature-dependent voltage and recombination behavior. Material-screening parameters encode changes in band alignment, carrier selectivity, and contact compatibility. Preserving these classes in the feature table made the model interpretable at both numerical and physical levels.

No attempt was made to infer unreported microscopic quantities such as exact defect energy levels, capture cross sections, or mobility values when they were not explicitly varied in the exported sheets. This conservative decision avoids a common error in simulation-based machine learning: adding plausible but unsupported descriptors can make a dataset appear richer than it actually is. Instead, the model uses only values that were present in the workbook or deterministically derived from those values, such as the logarithm of positive concentration-like variables.

$$\eta = (V_{oc} \times J_{sc} \times FF) / Pin, \quad (1)$$

where η is the power conversion efficiency, V_{oc} is the open-circuit voltage, J_{sc} is the short-circuit current density, FF is the fill factor expressed as a fractional quantity in the physical formula, and Pin is the incident optical power density. The exported dataset reports FF and PCE in percent, while J_{sc} is reported in mA cm^{-2}.

Table I is cited here to define the numerical range of the modeled target variables. The values are derived directly from cleaned_scaps_dataset.csv.

TABLE I: Descriptive statistics of SCAPS-derived photovoltaic target variables.

Target	n	Mean	SD	Min.	Q1	Median	Q3	Max.
V_{oc}	270	1.283	0.1461	0.2351	1.199	1.2	1.443	1.551
J_{sc}	270	20.71	4.058	0.1167	18.63	21.34	23.51	30.2
FF	270	88.23	5.351	25	87.64	88.23	90.82	92.25
PCE	270	23.75	6.371	0.11	19.57	22.58	30.66	33.57

Table II is cited here to identify the range of device parameters represented in the final modeling dataset. It also shows the observed spread in PCE associated with each one-factor sweep.

TABLE II: Extracted parameter-sweep groups and target ranges.

Parameter	Family	Rows	V_{oc} min	V_{oc} max	J_{sc} min	J_{sc} max	FF min	FF max	PCE min	PCE max
cbts_absorber_acceptor_concentration_cm^3	absorber	12	1.061	1.383	21.21	23.68	83.58	90.82	21	26.64
cbts_absorber_donor_concentration_cm^3	absorber	12	1.061	1.383	21.21	23.68	83.17	90.82	21	26.64
cbts_absorber_thickness_nm	absorber	31	1.181	1.253	11.75	21.59	87.59	88.71	12.89	22.95
pedama_absorber_acceptor_concentration_cm^3	absorber	9	1.199	1.199	21.3	21.35	88.2	88.23	22.53	22.58
pedama_absorber_donor_concentration_cm^3	absorber	12	1.199	1.199	21.26	21.35	88.2	88.23	22.49	22.58
pedama_absorber_thickness_nm	absorber	27	1.194	1.2	13.96	21.3	86.6	88.23	14.51	22.53
cbts_absorber_defect_density_cm^3	defect	12	1.051	1.551	0.1167	23.71	87.35	91.43	0.11	33.57
etl_defect_density_cm^3	defect	12	1.199	1.199	18.56	18.63	87.64	87.64	19.49	19.58
htl_defect_density_cm^3	defect	12	1.443	1.443	23.4	23.51	90.64	91.08	30.68	30.9
pedama_absorber_defect_density_cm^3	defect	12	1.142	1.199	7.059	21.41	86.1	88.23	6.94	22.65
etl_material	material	11	1.349	1.443	14.09	23.67	89.78	91.09	17.06	31.11
htl_material	material	15	0.8943	1.443	23.53	30.2	83.93	91.14	19.75	31.11
series_resistance_ohm_cm^2	resistance	11	1.443	1.444	23.51	23.51	75.51	91.08	25.64	30.9
shunt_resistance_ohm_cm^2	resistance	10	0.2351	1.443	23.51	23.51	25	91.08	1.38	30.9
temperature_k	thermal	16	1.389	1.465	23.46	23.58	88.57	92.25	29.02	31.71
etl_donor_concentration_cm^3	transport	12	1.199	1.199	18.17	18.63	87.38	87.64	19.03	19.58
etl_thickness_nm	transport	16	1.19	1.199	18.61	18.62	87.64	87.64	19.55	19.57
htl_acceptor_concentration_cm^3	transport	12	1.431	1.443	23.49	23.51	88.29	91.13	29.67	30.9
htl_thickness_nm	transport	16	1.443	1.443	23.36	23.56	91.12	91.14	30.72	30.99

Model development followed a supervised regression framework. For each target variable, the dataset was split into 80% training and 20% held-out testing using a fixed random state. Numerical predictors were median-imputed, and categorical predictors were imputed with a missing-value category and one-hot encoded. For SVR, Ridge, and Lasso models, numerical predictors were standardized to avoid scale-driven coefficient instability. Random Forest regression did not require feature scaling because tree splits are invariant to monotonic transformations of individual features, although the logarithmic active-value feature was retained to represent the concentration-like variables spanning many orders of magnitude.

Random Forest regression predicts a target by averaging an ensemble of decision trees. If h_b(x) is the prediction of the b-th tree for feature vector x, the ensemble estimate is

The use of a logarithmic active-value descriptor is important because several simulated quantities, including defect and doping concentrations, vary over many decades. In semiconductor devices, the difference between 10^{10} and 10^{11} cm^{-3} may be less consequential than a later order-of-magnitude change after a recombination or space-charge threshold has been crossed. Supplying both the raw value and the log10-transformed value lets the tree ensemble split on either absolute or order-of-magnitude regimes. The scaled baselines also benefit from this representation because a purely linear model operating on raw concentration values would be dominated by the largest numerical ranges.

$$\hat{y}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \quad (2)$$

where B is the number of trees. The hyperparameter grid included n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. GridSearchCV selected the best configuration according to cross-validated R^2.

The evaluation metrics were defined in Eqs. (3-8).

$$\boldsymbol{MAE} = \frac{1}{n}\sum_i |\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i| \quad \textbf{(3)}$$

$$\boldsymbol{MSE} = \frac{1}{n}\sum_i (\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i)^2 \quad \textbf{(4)}$$

$$\textbf{RMSE} = \sqrt{\boldsymbol{MSE}} \quad \textbf{(5)}$$

$$\textbf{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad \textbf{(6)}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (7)$$

and the adjusted coefficient of determination was

$$R^2_{adjusted} = 1 - (1 - R^2)(n - p - 1) \quad (8)$$

Table III is cited here to summarize the implemented model families and validation protocol.

TABLE III: Machine-learning models and validation design.

Model	Principle	Tuned hyperparameters	Role in study
Random Forest	Nonlinear ensemble of decision trees	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	Primary surrogate and feature-importance source
SVR	Kernel-based margin regression	C and epsilon	Nonlinear comparator with scaled inputs
Ridge	L2-regularized linear regression	alpha	Linear baseline with coefficient shrinkage
Lasso	L1-regularized linear regression	alpha	Sparse linear baseline

Statistical reliability was assessed at two levels. First, K-fold cross-validation was used inside the training set to select tuned models and estimate out-of-fold generalization behavior [16]. Second, the held-out Random Forest residuals were tested for mean bias using a one-sample t-test against zero. This test does not prove physical extrapolation outside the SCAPS sweep domain, but it helps determine whether the held-out residuals show a systematic positive or negative shift.

The study design intentionally separates model-selection evidence from final held-out evidence. Cross-validation was used only within the training portion, while the 20% test split remained unavailable during grid search. This separation is essential because parameter sweeps can contain many nearby records with similar baseline values. If hyperparameters were tuned directly on the test set, the resulting metrics would overstate generalization. The final evaluation therefore reports both the best cross-validated R^2 and the independent test-set R^2 for each target and model family.

All generated figures and tables were treated as traceable artifacts. The manuscript includes the correlation heatmap, feature-importance plot, actual-versus-predicted panels, residual diagnostics, error-distribution histograms, and cross-validation plot from the output directory. Additional residual and error grids were assembled only by stitching already generated target-specific plots; no new numerical results were introduced during figure assembly.

III. Results and Discussion

The cleaned dataset exhibits a broad but physically interpretable response range. V_{oc} spans from 0.2351 V to 1.5510 V, J_{sc} spans from 0.1167 to 30.2045 mA cm^{-2}, FF spans from 25.00% to 92.25%, and PCE spans from 0.11% to 33.57%. The low extremes arise primarily from severe shunt leakage or high absorber-defect conditions, whereas the high-PCE region is associated with favorable absorber defect-density conditions and transport-layer choices. The median response values, 1.1998 V, 21.3444 mA cm^{-2}, 88.23%, and 22.58% for V_{oc}, J_{sc}, FF, and PCE respectively, indicate that many sweeps remain near the baseline simulated operating region, while a smaller subset probes highly degraded or enhanced regimes.

The target-target Spearman correlations show that PCE is jointly controlled by voltage, current, and fill factor rather than by a single metric. PCE correlates with V_{oc} at 0.8318, with J_{sc} at 0.7979, and with FF at 0.8000. V_{oc} and FF also show a strong positive correlation of 0.7178, consistent with the fact that leakage and recombination can reduce both quasi-Fermi-level splitting and maximum-power point quality. The J_{sc}-FF correlation is weaker at 0.4412, which is physically reasonable because current collection and resistive loss do not always vary together in one-factor sweeps.

The parameter-sweep ranges provide a useful physical interpretation before regression is considered. The shunt-resistance sweep produces the lowest V_{oc} and FF values in the cleaned dataset, with V_{oc} falling to 0.2351 V and FF falling to 25.00%. This confirms that leakage pathways are catastrophic for voltage retention and maximum-power-point shape. By contrast, the series-resistance sweep leaves V_{oc} and J_{sc} nearly unchanged but reduces FF from 91.08% to 75.51% and PCE from 30.90% to 25.64%, matching the expected loss mechanism of voltage drop under load.

The absorber-defect sweeps contain the widest performance span. The Cu_2BaSnS_3 defect-density sweep covers the full PCE interval from 0.11% to 33.57%, while the $PeDAMA_4Pb_5I_{16}$ absorber-defect sweep ranges from 6.94% to 22.65% PCE. These values make defect density a natural high-impact variable for J_{sc} and PCE. In contrast, ETL thickness and ETL defect-density sweeps produce comparatively narrow PCE windows

around 19.5%, implying that those variables are less decisive within the simulated ranges and baseline architecture.

The material-screening sweeps provide a different kind of evidence. ETL material variation generated PCE values between 17.06% and 31.11%, while HTL material variation generated PCE values between 19.75% and 31.11%. These wide ranges indicate that the simulator captured large changes in band alignment or selective-contact quality across candidate transport materials. However, material identity is categorical, so the trained model can interpolate among observed categories only in a statistical sense; it cannot predict the behavior of an unseen ETL or HTL material without additional descriptors such as electron affinity, ionization energy, mobility, and defect energetics.

As shown in Fig. 2, the Spearman heatmap condenses the association among engineered features and photovoltaic outputs.

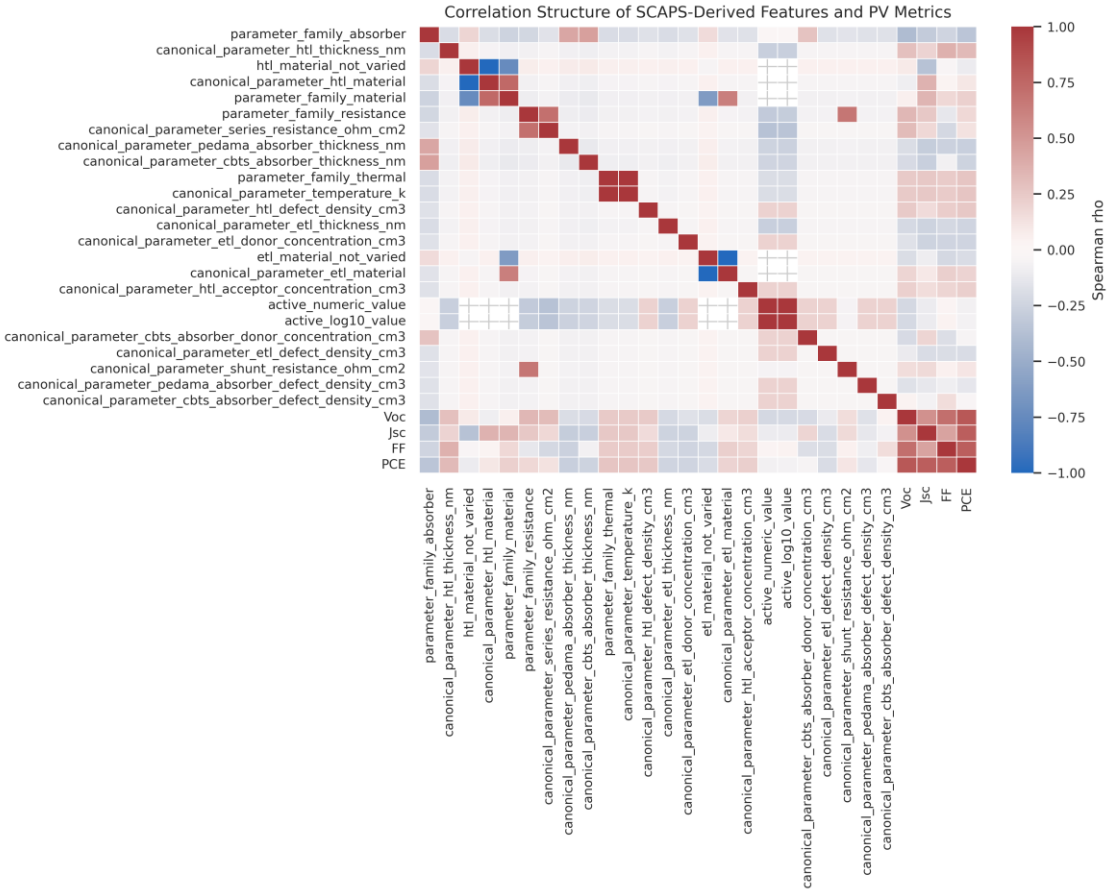

Fig. 2. Correlation heatmap of the engineered SCAPS features and target photovoltaic metrics. The map was generated from the cleaned dataset and highlights the monotonic associations used for exploratory data analysis.

Fig. 2 indicates that family-level absorber terms, thermal variables, transport-layer indicators, and selected layer-thickness features carry meaningful monotonic information about target performance. The heatmap should not be read as causal proof because the dataset is assembled from one-factor sweeps, but it is valuable as a screening tool for identifying variables that deserve greater attention in factorial simulations.

Table IV is cited here to compare all trained regressors on the held-out test split.

TABLE IV: Test-set and cross-validation performance of the trained regression models.

Target	Model	R^2	MAE	MSE	RMSE	CV R^2
V_{oc}	RandomForest	0.6677	0.02813	0.00622	0.07887	0.5895
V_{oc}	Ridge	-0.5407	0.1049	0.02884	0.1698	-81.08
V_{oc}	Lasso	-0.008451	0.1225	0.01887	0.1374	-0.04768
V_{oc}	SVR	0.671	0.02985	0.006158	0.07848	0.636
J_{sc}	RandomForest	0.8423	0.553	2.406	1.551	0.7988
J_{sc}	Ridge	-0.1686	2.348	17.82	4.222	-288.8

J_{sc}	Lasso	0.01878	2.554	14.97	3.869	-104.7
J_{sc}	SVR	0.3301	0.9797	10.22	3.196	0.4287
FF	RandomForest	0.6564	0.4412	0.886	0.9413	0.5689
FF	Ridge	-1.743	1.557	7.072	2.659	-164.7
FF	Lasso	-0.1023	1.178	2.842	1.686	-0.0341
FF	SVR	0.8398	0.2871	0.4131	0.6427	0.4996
PCE	RandomForest	0.8417	1.007	5.165	2.273	0.7789
PCE	Ridge	-0.7808	4	58.12	7.623	-224
PCE	Lasso	0.09931	4.136	29.39	5.422	-103.9
PCE	SVR	0.4454	1.565	18.1	4.254	0.4489

The Random Forest models achieved R^2 values of 0.6677 for V_{oc}, 0.8423 for J_{sc}, 0.6564 for FF, and 0.8417 for PCE. The corresponding RMSE values were 0.0789 V, 1.5512 mA cm^{-2}, 0.9413 percentage points, and 2.2726 percentage points. These values indicate that the nonlinear ensemble captured the major structure of J_{sc} and PCE particularly well. V_{oc} and FF were more difficult because several low-voltage and low-fill-factor examples originate from narrow leakage or contact regimes that are underrepresented relative to the baseline-like records. SVR slightly outperformed Random Forest for V_{oc} and FF on the particular held-out split, with R^2 values of 0.6710 and 0.8398, respectively, but Random Forest remained superior for J_{sc} and PCE and provided more direct feature-importance interpretation. Ridge and Lasso baselines performed poorly for this dataset, confirming that linear shrinkage alone is insufficient for the nonlinear and thresholded SCAPS response surface. As shown in Fig. 3, the Random Forest importance analysis identifies the dominant SCAPS control variables.

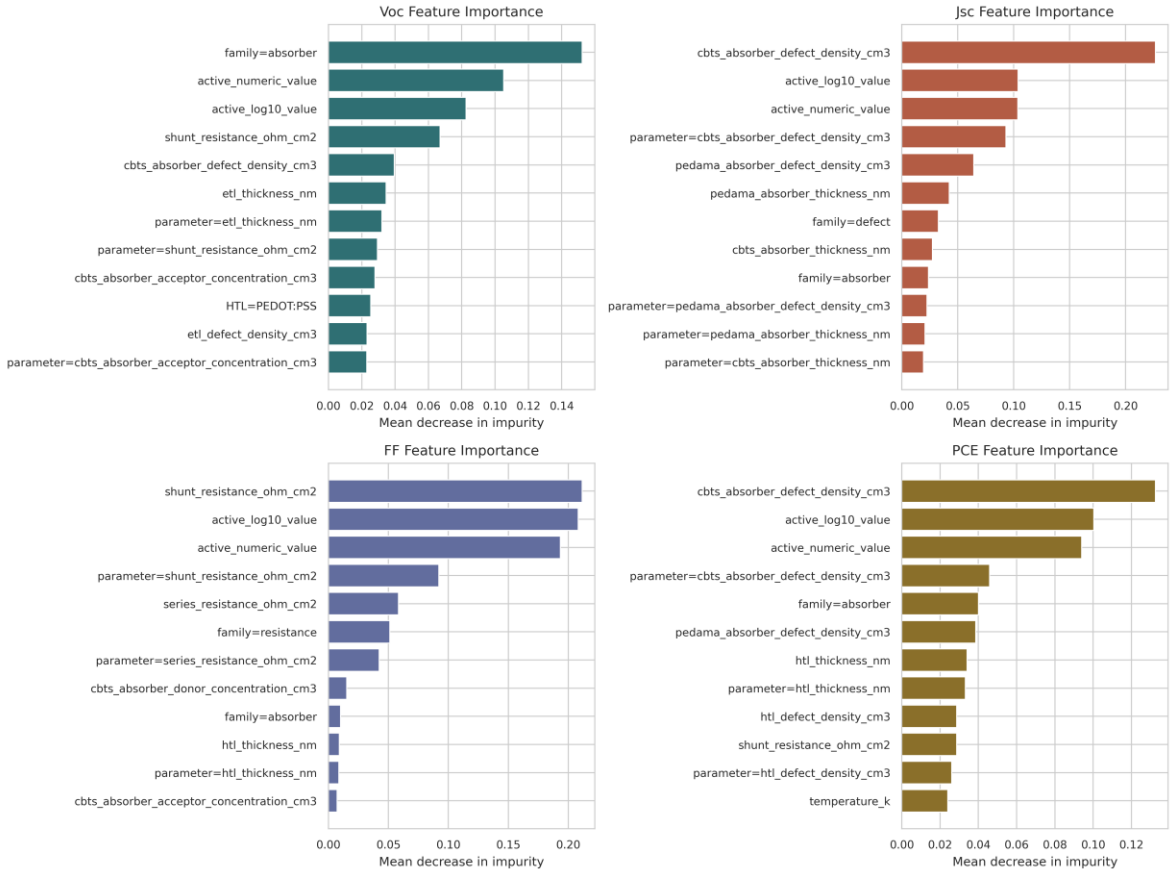

Fig. 3. Random Forest feature-importance ranking for V_{oc}, J_{sc}, FF, and PCE. Importance values are based on mean decrease in impurity from the tuned Random Forest estimators.

Fig. 3 shows that the dominant predictors are physically plausible. V_{oc} is most influenced by absorber-family terms, active parameter magnitude, the logarithmic active value, shunt resistance, and Cu_2BaSnS_3 absorber defect density. J_{sc} is dominated by Cu_2BaSnS_3 defect density, active-value features, and $PeDAMA_4Pb_5I_{16}$ defect or thickness terms. FF is governed primarily by shunt resistance, active-value terms, and series resistance. PCE combines these mechanisms, with Cu_2BaSnS_3 defect density, active-value descriptors, absorber family, $PeDAMA_4Pb_5I_{16}$ defect density, and HTL thickness emerging as key variables. This pattern is consistent with semiconductor-device expectations: deep absorber defects enhance Shockley-Read-Hall

recombination, shunt pathways reduce the slope and voltage retention of the J-V curve, and series resistance reduces the maximum-power operating point.
Table V is cited here to provide the numerical feature-importance summary used in the preceding physical interpretation.

TABLE V: Top Random Forest predictors for each photovoltaic target.

Target	Top predictors
V_{oc}	family=absorber (0.152); active_numeric_value (0.105); active_log10_value (0.083); shunt_resistance_ohm_cm^2 (0.067); cbts_absorber_defect_density_cm^3 (0.040)
J_{sc}	cbts_absorber_defect_density_cm^3 (0.227); active_log10_value (0.104); active_numeric_value (0.104); parameter=cbts_absorber_defect_density_cm^3 (0.093); pedama_absorber_defect_density_cm^3 (0.064)
FF	shunt_resistance_ohm_cm^2 (0.211); active_log10_value (0.208); active_numeric_value (0.193); parameter=shunt_resistance_ohm_cm^2 (0.092); series_resistance_ohm_cm^2 (0.058)
PCE	cbts_absorber_defect_density_cm^3 (0.133); active_log10_value (0.100); active_numeric_value (0.094); parameter=cbts_absorber_defect_density_cm^3 (0.046); family=absorber (0.040)

As shown in Fig. 4, the actual-versus-predicted plots show held-out agreement between SCAPS-derived and machine-learning values.

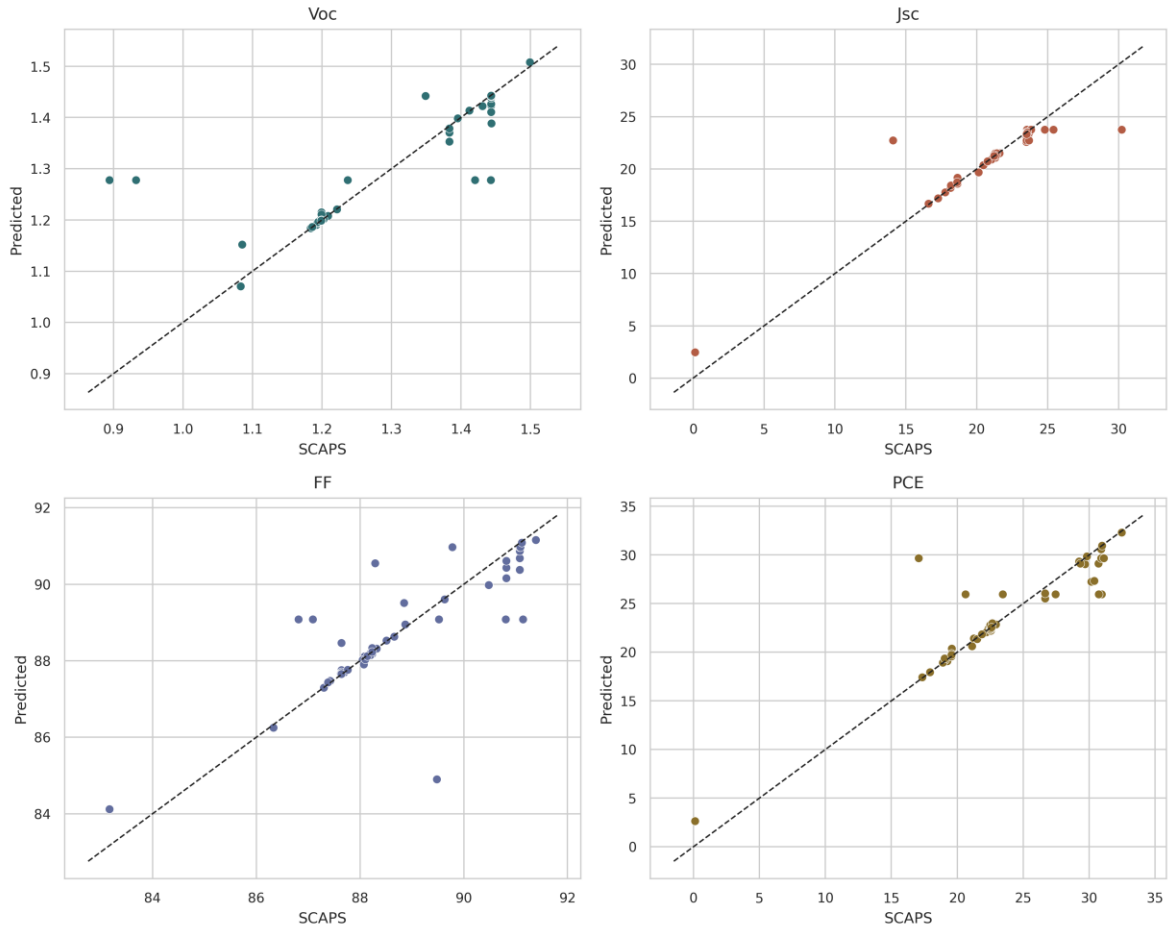

Fig. 4. Actual versus predicted held-out values for the four target variables. The dashed line represents the ideal y = x relation.

The scatter plots in Fig. 4 show tight agreement for the central J_{sc} and PCE regions and moderate dispersion for FF and V_{oc} at the domain edges. The largest J_{sc} and PCE residuals occur near low-current or high-current extremes where fewer samples are available. This pattern is expected for a Random Forest trained on one-factor sweep data because tree ensembles interpolate robustly within populated feature regions but cannot infer unsampled coupled physics beyond the observed sweep structure.
As shown in Fig. 5, the residual diagnostics reveal whether prediction errors are centered around zero across the target ranges.

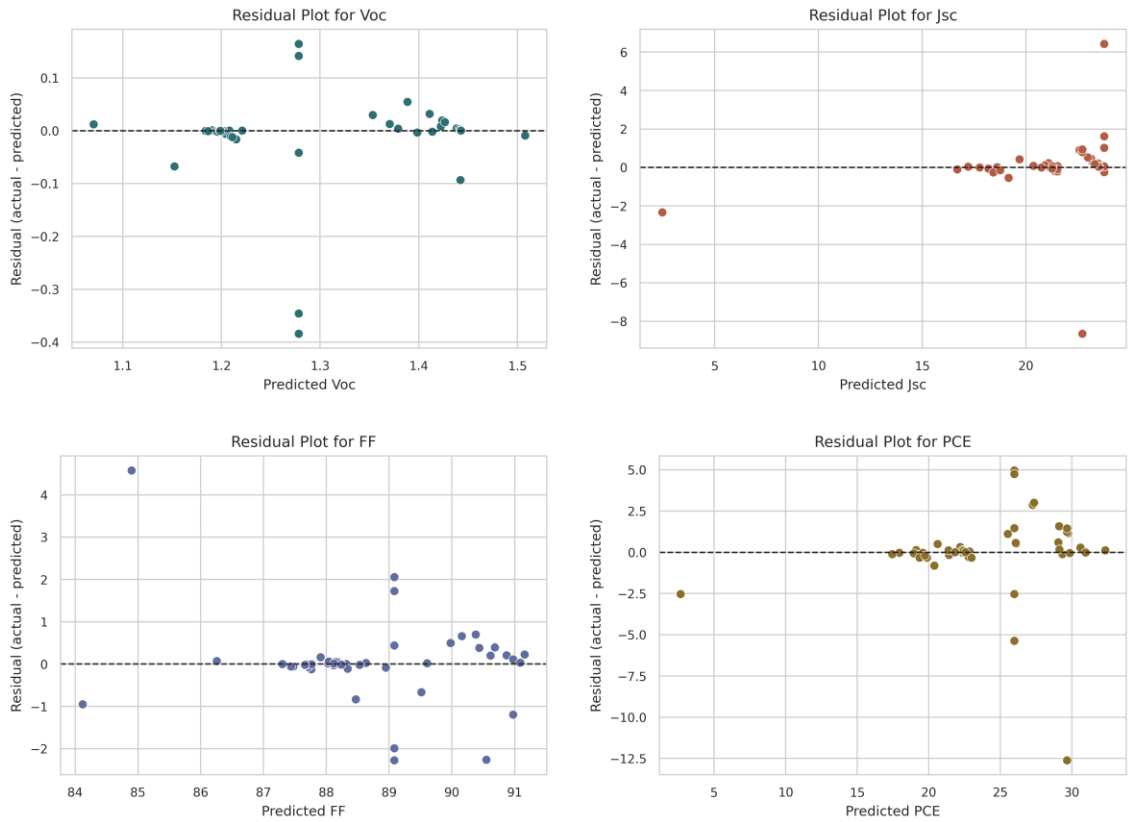

Fig. 5. Residual plots for Random Forest predictions of V_{oc}, J_{sc}, FF, and PCE. Residuals are defined as actual minus predicted values.

Table VI is cited here to summarize the residual statistics that support the diagnostic interpretation in Fig. 5.

TABLE VI: Random Forest held-out residual statistics.

Target	n	Mean	SD	Min.	Max.
FF	54	0.04122	0.9492	-2.271	4.581
J_{sc}	54	0.04322	1.565	-8.638	6.438
PCE	54	0.03462	2.294	-12.6	4.976
V_{oc}	54	-0.008963	0.07909	-0.3839	0.1648

Table VII is cited here for the one-sample residual-bias test calculated from the saved Random Forest prediction file.

TABLE VII: One-sample t-test for held-out residual bias relative to zero.

Target	n	Mean residual	SD	z statistic	approx. p value
FF	54	0.04122	0.9492	0.3191	0.7497
J_{sc}	54	0.04322	1.565	0.2029	0.8392
PCE	54	0.03462	2.294	0.1109	0.9117
V_{oc}	54	-0.008963	0.07909	-0.8328	0.405

The mean residuals are small relative to the target ranges: −0.0090 V for V_{oc}, 0.0432 mA cm^{-2} for J_{sc}, 0.0412 percentage points for FF, and 0.0346 percentage points for PCE. These values indicate little systematic bias in the average held-out predictions. The standard deviations reveal the expected target-dependent spread, with PCE residuals having the largest dispersion because PCE compounds voltage, current, and fill-factor errors. As shown in Fig. 6, the residual histograms provide a second view of the model-error distribution.

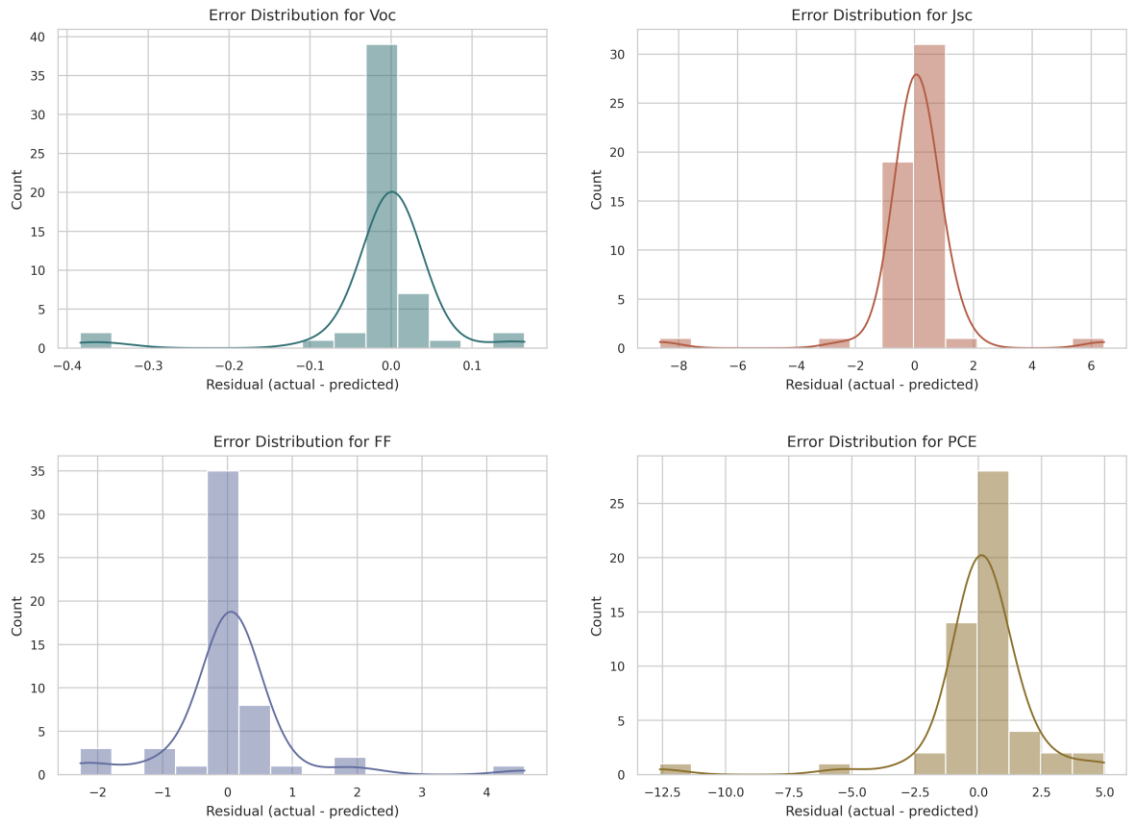

Fig. 6. Error-distribution histograms for Random Forest residuals across V_{oc}, J_{sc}, FF, and PCE.

Fig. 6 confirms that most errors cluster near zero while a few tail cases control the RMSE. The asymmetry in the PCE error distribution is especially important for engineering use: a surrogate model may identify promising regions, but final device claims should still be verified by SCAPS reruns or experimental characterization. This is a practical distinction between screening accuracy and certification-level prediction.

As shown in Fig. 7, the cross-validation comparison shows how model families generalize during grid search.

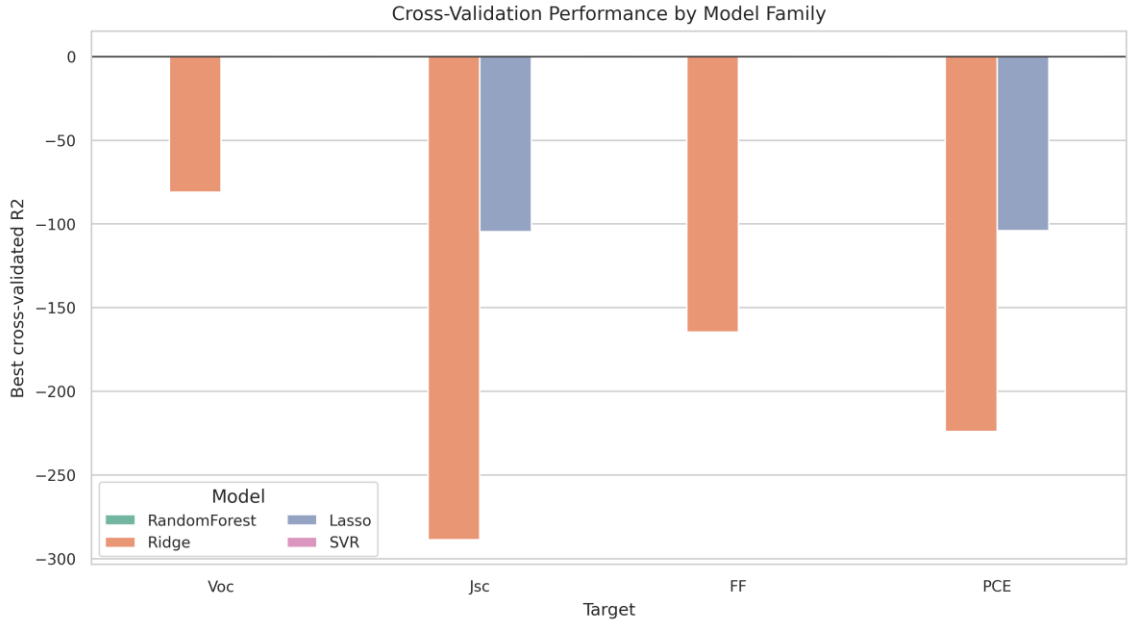

Fig. 7. Cross-validation performance by target and model family. The plotted values are the best cross-validated R^2 scores selected during GridSearchCV.

Fig. 7 reinforces the conclusion that nonlinear models are necessary for this dataset. Ridge and Lasso are strongly limited by the one-factor nonlinear response, especially for J_{sc} and PCE. SVR performs competitively on V_{oc} and FF but does not provide the same straightforward feature-importance ranking. Random Forest therefore offers the best balance between predictive performance, robustness to mixed feature types, and interpretability for the present SCAPS-derived data.
Table VIII is cited here to compare the best SCAPS-derived device with the corresponding Random Forest predictions.

TABLE VIII: SCAPS-derived and Random Forest-predicted values for the best simulated device.

Parameter	SCAPS Derived Value	ML Predicted Value	Absolute Error	Relative Error (%)
V_{oc}	1.551	1.507	0.0436	2.811
J_{sc}	23.71	23.36	0.3459	1.459
FF	91.28	90.81	0.4715	0.5165
PCE	33.57	32.33	1.238	3.687

The best simulated device was obtained in the Cu_2BaSnS_3 absorber-defect-density sweep at a varied parameter value of 1.0×10^{10} cm^{-3}. The SCAPS-derived values were V_{oc} = 1.5510 V, J_{sc} = 23.7098 mA cm^{-2}, FF = 91.28%, and PCE = 33.57%. The Random Forest predicted 1.5074 V, 23.3639 mA cm^{-2}, 90.8086%, and 32.3323%, respectively. The relative errors were 2.81% for V_{oc}, 1.46% for J_{sc}, 0.52% for FF, and 3.69% for PCE. For a simulator-trained surrogate, these errors are sufficiently small to support rapid screening and ranking, although the highest-efficiency point should not be used as a final claim without direct SCAPS confirmation.

The physical trends inferred from the model align with expected photovoltaic loss pathways. When absorber defect density increases, recombination centers reduce minority-carrier lifetime and lower the collection probability. This suppresses J_{sc} and can also reduce V_{oc} by shrinking the quasi-Fermi level splitting. When shunt resistance is low, leakage current dominates the J-V curve near short circuit and open circuit, which explains the dramatic minimum V_{oc} of 0.2351 V, FF of 25.00%, and PCE of 1.38% observed in the shunt-resistance sweep. When series resistance increases, the current at the maximum-power point is penalized, producing the expected FF and PCE reduction while V_{oc} and J_{sc} remain comparatively stable.

Thickness-dependent behavior is more subtle. Absorber thickening can initially increase optical absorption and current, but excessive thickness can increase recombination and transport losses, especially when mobility or lifetime is limited. The $PeDAMA_4Pb_5I_{16}$ absorber-thickness sweep shows J_{sc} decreasing from a maximum near 21.30 mA cm^{-2} to lower values at larger thickness, while the Cu_2BaSnS_3 thickness sweep exhibits a broader current response with J_{sc} increasing from 11.7463 to a maximum near 21.5871 mA cm^{-2} before declining. Such nonmonotonic behavior is exactly the type of relationship for which tree ensembles are useful.

Material screening of ETL and HTL layers also appears in the dataset. ETL variation produced PCE values from 17.06% to 31.11%, while HTL variation produced PCE values from 19.75% to 31.11%. These ranges confirm that carrier-selective contacts and transport-layer band alignment can significantly reshape device performance even when absorber properties are unchanged. In practical engineering terms, this suggests that absorber optimization should be coupled with transport-layer screening rather than treated as an isolated problem.

Temperature effects were moderate but systematic over the simulated interval. The temperature sweep covered 250 to 400 K for complete records, with V_{oc} decreasing from 1.4654 V to 1.3892 V and PCE decreasing from 31.71% to 29.02%. This behavior is consistent with the increase in saturation current and enhanced thermally activated recombination at higher temperature. The small increase in J_{sc} across the same sweep reflects the fact that current can be less temperature-sensitive than voltage in many simulated photovoltaic stacks.

The study also clarifies an important limitation of the present data design. Because the SCAPS outputs are one-factor sweeps, the model has limited evidence for true parameter interactions. A Random Forest can represent interactions if they are present in the training data, but it cannot infer unobserved combinations such as simultaneous variation of absorber defect density, interface defect density, mobility, electron affinity, and contact work function. Therefore, the surrogate should be understood as a fast interpolator and trend interpreter over the existing simulation campaign, not as a global optimizer for the full device design space.

The comparison with HMM methodology is also instructive. Hidden Markov Models are designed for sequential latent-state problems, such as degradation pathways, time-resolved bias stress, thermal cycling, or manufacturing process histories [15]. The available SCAPS data are independent steady-state parameter sweeps; assigning an artificial sequence to them would risk a misleading statistical model. The correct future use of

HMMs would require sequential photovoltaic data, for example repeated SCAPS or experimental measurements under aging, illumination, or temperature-cycling protocols.

The residual-bias test supports this interpretation from a statistical perspective. The held-out mean residuals are close to zero for all targets, so the model does not show a large systematic offset over the random test split. The practical concern is therefore not average bias but local error at sparse or extreme sweep regions. In photovoltaic optimization, this distinction matters: an unbiased model can still overestimate a small number of high-value candidates, and those candidates are precisely the ones most likely to be selected for follow-up. A conservative workflow should therefore use the model to rank and prioritize candidates, then re-evaluate shortlisted devices in SCAPS.

The negative cross-validated R^2 values observed for some Ridge and Lasso configurations should not be interpreted as failure of regularization in general. They reflect the mismatch between linear models and the specific nonlinear sweep data used here. Regularized linear models are valuable when response surfaces are approximately additive and monotonic, or when interpretability of signed coefficients is the primary goal. In this dataset, abrupt shunt-resistance degradation, defect-density thresholds, and nonmonotonic thickness response violate those assumptions. Their weak performance is therefore itself a useful diagnostic result.

The strongest engineering implication is that defect management should be prioritized before fine optimization of less sensitive thickness or transport variables. The high-efficiency point in the Cu_2BaSnS_3 defect-density sweep occurs at the lowest simulated defect density, and the Random Forest importance ranking repeatedly elevates absorber-defect terms for J_{sc} and PCE. This suggests that process routes capable of reducing bulk traps, suppressing deep-level recombination, and improving crystallinity would likely produce larger performance gains than marginal tuning of already benign layer thicknesses.

At the same time, resistive design remains essential for fill-factor preservation. The FF model assigns dominant importance to shunt resistance and active-value features, while series resistance also appears among the leading terms. In fabricated devices, these variables correspond to contact quality, pinhole density, film continuity, electrode sheet resistance, interfacial reactions, and leakage pathways. The simulation-derived surrogate therefore points to a balanced optimization strategy: reduce absorber defects to improve voltage and current while simultaneously maintaining high shunt resistance and low series resistance to protect FF.

Another practical implication concerns dataset expansion. The current feature table contains several parameters requested by the broader project description only when they appear in the workbook. Future simulation campaigns should explicitly include absorber bandgap, electron affinity, interface defect density, electron mobility, hole mobility, buffer-layer thickness, back-contact work function, and transport-layer electronic parameters in a single design-of-experiments table. Doing so would convert the present one-factor surrogate into a more complete architecture optimizer and would enable stronger claims about interactions among band alignment, recombination, and carrier extraction.

From an engineering standpoint, the surrogate workflow offers several benefits. It reduces the time needed to identify high-impact parameters, provides a ranked list of variables for follow-up simulation, and helps distinguish robust improvements from narrow sweep artifacts. It also creates a reproducible bridge between SCAPS exports and publication-ready statistical evidence. Future work should extend this pipeline to factorial or Sobol sampling, include interface defect density and mobility explicitly, and benchmark gradient-boosted models such as XGBoost, LightGBM, and CatBoost. Explainable techniques such as SHAP values could supplement impurity-based importance rankings and improve local interpretation of individual high-efficiency candidates [20].

Reproducibility is a central practical outcome of this work. The workflow separates raw simulator input, cleaned learning table, fitted estimators, figures, and manuscript interpretation. This is important because SCAPS studies can otherwise become difficult to audit: a reported value may depend on a particular worksheet, a particular row in a parameter sweep, or a manual transfer of J-V metrics from the simulator into a summary table. Here, every reported model metric is available in model_performance_metrics.csv, every Random Forest prediction used for residual analysis is available in test_set_predictions_random_forest.csv, and every feature-importance statement is traceable to feature_importance_random_forest.csv. This file-level traceability is especially valuable for journal review because it allows a reader to distinguish between simulator-derived evidence, machine-learning evidence, and author interpretation.

Uncertainty should nevertheless be interpreted conservatively. The reported test split contains 54 held-out records per target, which is adequate for an internal surrogate assessment but not equivalent to external validation on independently designed SCAPS experiments. The cross-validation scores are also lower than some test-set scores, particularly for V_{oc} and FF, indicating that the apparent predictability depends on how the one-factor sweep records are partitioned. A stronger future validation strategy would reserve entire parameter-sweep families as external groups. Such grouped validation would answer a more demanding question: whether a model trained on thickness, resistance, and material sweeps can predict a new defect-density sweep or an unseen transport-layer family.

The present feature-importance values should also be read as model explanations rather than universal physical constants. Mean decrease in impurity can distribute importance across correlated variables, and the current feature table deliberately contains both active_numeric_value and active_log10_value. Therefore, importance assigned to these generic magnitude descriptors must be interpreted together with the canonical-parameter and family indicators. The physically meaningful conclusion is not that a generic numeric column is itself a device parameter, but that parameter magnitude, especially for concentration-like sweeps, carries strong predictive information when combined with the identity of the varied parameter.

For experimental translation, the most useful output of the model is not a single predicted efficiency but a ranked strategy for additional simulation and fabrication. The analysis suggests first reducing absorber defect density, then preserving high shunt resistance and low series resistance, and then co-optimizing transport-layer thickness and material selection. Such a sequence is consistent with device engineering practice: recombination control sets the achievable voltage and current envelope, while contact and resistance optimization determine how much of that envelope is retained at the maximum-power point. The SCAPS-ML workflow therefore acts as a decision-support layer rather than a replacement for semiconductor-device reasoning.

A final methodological implication is that publication-quality machine-learning studies in photovoltaics should report both performance and data geometry. A high R^2 value is less informative if the training set is dominated by repeated near-baseline records, and a moderate R^2 value can still be scientifically valuable if it correctly identifies the dominant physical loss channel. In this manuscript, the PCE Random Forest achieved $R^2 = 0.8417$ and RMSE = 2.2726 percentage points, while also identifying defect density and resistance-related mechanisms that are consistent with SCAPS physics. That combination of prediction and interpretation is the main reason the surrogate is useful despite the limitations of the available one-factor sweep design.

IV. Conclusion

This paper reconstructed a complete scientific workflow from SCAPS-1D simulation exports, Python analysis code, generated result tables, figures, and trained model artifacts. The irregular workbook was converted into a 270-record machine-learning dataset with complete V_{oc}, J_{sc}, FF, and PCE labels. The final data covered absorber, defect, transport-layer, material, resistance, and temperature sweeps, thereby capturing a meaningful portion of the simulated design space for the studied lead-free photovoltaic architecture.

Random Forest regression provided a strong and interpretable surrogate for the SCAPS outputs. The held-out R^2 values were 0.6677 for V_{oc}, 0.8423 for J_{sc}, 0.6564 for FF, and 0.8417 for PCE. The best SCAPS-derived PCE was 33.57%, and the Random Forest estimate for that device was 32.33%, giving a relative PCE error of 3.69%. Feature importance and correlation analysis identified absorber-family terms, defect density, shunt resistance, series resistance, layer thickness, and active parameter magnitude as the dominant drivers.

The main limitation is the one-factor sweep structure of the source simulations. The resulting model is useful for interpolation, screening, and mechanistic interpretation within the simulated domain, but it should not be used for unconstrained extrapolation or global optimization. Future SCAPS campaigns should vary multiple parameters simultaneously, include a dedicated external validation set, and test boosted-tree, Bayesian, Gaussian-process, and deep-learning surrogates. Sequential models such as HMMs should be reserved for degradation or process-history datasets rather than independent static sweeps.

Ethics approval

Not applicable. This study did not involve human participants, animals, or clinical data requiring ethical approval.

Consent to participate

Not applicable. This study did not involve human subjects requiring informed consent.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data, images, or personal information requiring consent for publication.

Competing interests

The authors declare no competing interests.

Author contributions: All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by KCD, AS, RKS, SS and SS. The first draft of the manuscript was written by RKS. AS and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability: All simulation data are available in the SCAPS-1D output files (BaCoCl_.xlsx), available from the corresponding author upon reasonable request.

Acknowledgment: We would like to express our gratitude to Dr. Marc Burgelman and their team from Ghent University, Belgium, for providing access to the SCAPS simulator, which has been instrumental in conducting the simulations for this research. This work is supported by the research project (Project Id 3847) funded by the Council of Science and Technology, Uttar Pradesh, India as per letter no. CST/D-826.

References

[1] W. Shockley and H. J. Queisser, "Detailed balance limit of efficiency of p-n junction solar cells," Journal of Applied Physics, vol. 32, no. 3, pp. 510-519, 1961.
[2] A. Kojima, K. Teshima, Y. Shirai, and T. Miyasaka, "Organometal halide perovskites as visible-light sensitizers for photovoltaic cells," Journal of the American Chemical Society, vol. 131, no. 17, pp. 6050-6051, 2009.
[3] N.-G. Park, "Perovskite solar cells: an emerging photovoltaic technology," Materials Today, vol. 18, no. 2, pp. 65-72, 2015.
[4] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature, vol. 559, no. 7715, pp. 547-555, 2018.
[5] M. Burgelman, P. Nollet, and S. Degrave, "Modelling polycrystalline semiconductor solar cells," Thin Solid Films, vol. 361-362, pp. 527-532, 2000.
[6] M. Burgelman, J. Verschraegen, S. Degrave, and P. Nollet, "Modeling thin-film PV devices," Progress in Photovoltaics: Research and Applications, vol. 12, no. 2-3, pp. 143-153, 2004.
[7] A. Saidarsan, S. Guruprasad, A. Malik, P. Basumatary, and D. S. Ghosh, "A critical review of unrealistic results in SCAPS-1D simulations: Causes, practical solutions and roadmap ahead," Solar Energy Materials and Solar Cells, vol. 279, Art. no. 113230, 2025.
[8] S. Banik, A. Das, B. K. Das, and N. Islam, "Numerical simulation and performance optimization of a lead-free inorganic perovskite solar cell using SCAPS-1D," Heliyon, vol. 10, no. 1, Art. no. e23985, 2024.
[9] S. Karthick, S. Velumani, and J. Boucle, "Chalcogenide BaZrS3 perovskite solar cells: A numerical simulation and analysis using SCAPS-1D," Optical Materials, vol. 126, Art. no. 112250, 2022.
[10] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[11] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[12] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
[13] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55-67, 1970.
[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B, vol. 58, no. 1, pp. 267-288, 1996.
[15] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.
[16] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2021.
[17] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
[18] C. Odabasi and R. Yildirim, "Performance analysis of perovskite solar cells in 2013-2018 using machine-learning tools," Nano Energy, vol. 56, pp. 770-791, 2019.
[19] C. Chen, A. Maqsood, and T. J. Jacobsson, "The role of machine learning in perovskite solar cell research," Journal of Alloys and Compounds, vol. 960, Art. no. 170824, 2023.
[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017.