Applied Physics A (2025) 131:850
https://doi.org/10.1007/s00339-025-09003-x

Applied Physics A
Materials Science & Processing

Defect diagnostics in thin film photovoltaics: leveraging macroscopic J-V characteristics for microscopic insights via machine learning

Gholamhosain Haidari[1,2]

Received: 14 August 2025 / Accepted: 25 September 2025

Abstract
Accurately diagnosing microscopic defect properties from macroscopic J-V characteristics in thin-film photovoltaics remains a critical barrier to advancing solar cell efficiency. While experimental techniques like DLTS and admittance spectroscopy directly probe defect states, their widespread adoption is limited by complexity and cost. This study presents an innovative AI-driven framework, using CIGS solar cells as a representative case to demonstrate a universally applicable approach for correlating routine J-V measurements with underlying defect configurations. A comprehensive, physics-based dataset was generated through high-throughput SCAPS-1D simulations. Systematic variation of six key defect parameters (bulk and interface defects) yielded 4,096 synthetic J-V curves, each linked to specific defect states. This addresses the critical challenge of data scarcity, as obtaining such relevant and diverse experimental data is often practically impossible. A Mixture Density Network with multi-head attention was developed to probabilistically predict defect properties from J-V features, addressing the inherent ambiguity in inverse defect-property mapping. Exceptional accuracy ($R \approx 1$) was achieved for critical defects like CIGS bulk vacancies (V_{Se}, V_{Cu}) and interface traps, while uncertainty was quantified for less discernible defects. The methodology's novelty extends beyond CIGS, evidenced by advanced non-Gaussian statistical analysis revealing defect-specific performance clusters (e.g., V_{OC} bimodality tied to defect thresholds) and the introduction of innovative visualization tools (e.g., continuous coverage matrices). To foster reproducibility, the dataset and protocols are made publicly available. This work bridges device physics and machine learning to extract microscopic insights from routine solar cell electrical measurements, establishing a transformative paradigm.

Keywords Thin film photovoltaics · Defect diagnostics · Advanced data visualization · Physics-Informed ML · Open defect datasets · Mixture density network

1 Introduction

Solar photovoltaics are a key part of renewable energy, like thin-film CIGS technology, achieving lab efficiencies exceeding 23% [1, 2]. However, for research and development in this area, while J-V measurements yield macroscopic performance parameters (V_{OC}, J_{SC}, FF, and η), they don't directly reveal the interesting microscopic mechanisms, like bulk and interface defects, that determine device performance [3, 4]. Although techniques like DLTS and admittance spectroscopy effectively probe defect states [5], their cost and complexity limit their widespread use. This motivates the development of methods capable of extracting comprehensive defect information from routine current-voltage (J-V) characterization, enabling deeper material insights without requiring specialized instrumentation.

The diagnostic challenge of decoding subtle electrical signatures embedded in standard J-V data has spurred interest in such approaches, which could unlock valuable defect properties from conventional measurements. Recent advances in artificial intelligence (AI) can present a transformative opportunity to achieve this goal. Machine learning (ML) algorithms have started to demonstrate remarkable success in identifying hidden patterns in the photovoltaic domain, with diverse applications ranging from defect classification [6], performance prediction [7] and design and

✉ Gholamhosain Haidari
haidari@sku.ac.ir

[1] Physics Department, Shahrekord University, Shahrekord, Iran

[2] Nano Research Institute & Artificial Intelligence Research Institute, Shahrekord University, Shahrekord, Iran

Published online: 14 October 2025

optimization strategies [8, 9]. The novelty of this study lies in its aim to correlate J-V response patterns with specific defect configurations. However, as observed in other ML applications to materials science [10], the development of robust AI models for solar cell diagnostics faces two key challenges: (i) the scarcity of high-quality, well-annotated experimental datasets linking defect properties to J-V characteristics, and (ii) the inherent noise and variability in real-world measurements that complicate model training. So, a critical bottleneck in realizing this potential is the availability of appropriate and sufficiently accurate datasets, which are essential for training robust and reliable predictive models [11, 12]. Simulations, particularly advanced batching ones, offer a precise and valuable investigation method to address these data limitations [13]. To address these challenges, a computational framework is proposed that integrates high-throughput solar cell simulations with AI-driven data analysis, using CIGS solar cell as a case study. This integrated approach, believed to be the first of its kind, enables defect diagnostics in thin-film photovoltaics through AI/ML algorithms based on routine J-V curves. SCAPS-1D simulations serve as an accurate tool to generate synthetic but physically realistic J-V datasets across controlled defect configurations. This approach overcomes experimental limitations and offers the necessary granularity for advanced data visualization and analysis, as well as for training interpretable ML models.

While traditional physics-based simulation tools like SCAPS-1D itself provide invaluable insights, they typically operate in a forward direction, predicting device performance from a known set of material parameters and defect properties. The inverse problem, deducing microscopic defect configurations from macroscopic measurements, remains significantly more challenging and is often intractable for pure physics-based models or empirical correlations alone. These traditional approaches struggle with the inherent ambiguity where multiple defect combinations can yield similar J-V characteristics. This proposed AI-driven framework addresses this fundamental limitation by leveraging the pattern recognition power of machine learning to solve this inverse problem probabilistically, providing a direct pathway from routine electrical characterization to underlying defect properties, a task beyond the reach of traditional models.

The SCAPS-1D software (version 3.3.10) was selected for this study due to its several advantageous characteristics for simulating thin-film photovoltaic devices [14]. Its well-documented and validated models for CIGS-based solar cells provide a high degree of physical accuracy. Crucially, SCAPS-1D offers a robust and detailed implementation of defect physics, allowing for the comprehensive inclusion of bulk and interface defects with user-defined properties, a core requirement for this investigation. Furthermore, its ability to be controlled via external scripts enabled the automated, high-throughput batch simulations necessary to generate the large dataset for machine learning. Finally, its widespread use and acceptance in the photovoltaic research community facilitate the comparison and validation of this simulated baseline results against established literature.

This motivation is further supported by recent successes of machine learning in related materials science challenges, such as optimizing composite structures [15], predicting material properties [16, 17], and solving complex inverse problems in engineering design [18–20]. This work builds upon these foundations by specifically tailoring the Mixture Density Network (MDN) architecture to the unique challenges of photovoltaic defect diagnostics.

2 Simulated ZnO/CdS/CIGS solar cell structure

The SCAPS-1D simulated thin-film solar cell consists of n-ZnO window, n-CdS buffer, and p-CIGS absorber layers (schematically illustrated in the inset of Fig. 1.a). A core objective of this simulation study was to generate physically

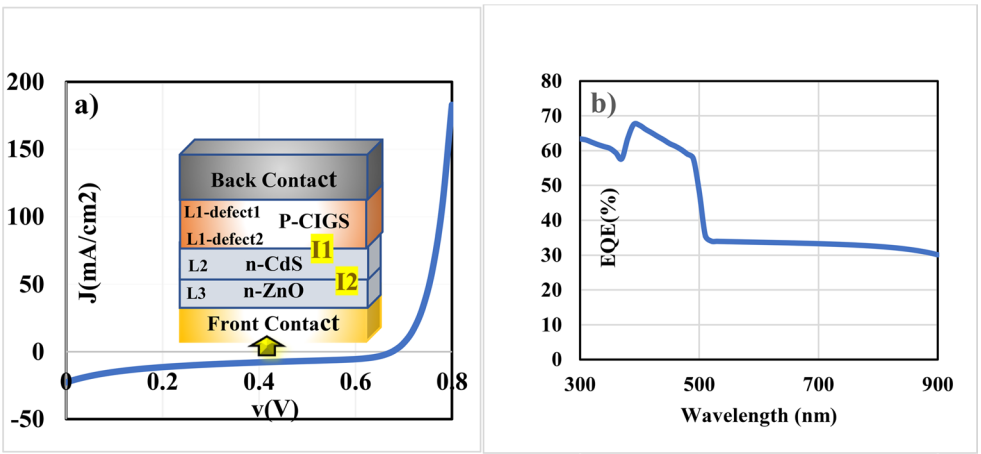

Fig. 1 **a**) Simulated J-V curve (inset: schematic device structure and labeled defect locations) and **b**) spectral response of a CIGS device

Table 1 General layer parameters and optical properties for solar cell simulation

Parameter	*n*-ZnO	*n*-CdS	*p*-CIGS	Source/Justification
Thickness (nm)	50	100	3000	Standard for high-efficiency CIGS ([26])
Bandgap (eV)	3.3	2.4	1.15	ZnO ([27]), CdS ([24]) and CIGS ([26])
Electron Affinity (eV)	4.4	4.2	4.5	ZnO([28]) and CIGS([29])
Doping (cm^{-3})	1×10^{18} (n)	1×10^{17} (n)	1×10^{16} (p)	CIGS ([30]) and ZnO ([31])
Mobility (cm^2/Vs)	μ_n:100, μ_p:25	μ_n:50, μ_p:10	μ_n:30, μ_p:5	ZnO, CdS ([32]) and CIGS ([33])
Optical Model	Forouhi-Bloomer [23]	Adachi [24]	Mudryi [25]	Urbach tail included for ZnO ($\alpha > 10^5$ cm^{-1} below 380 nm)

realistic current-voltage (J-V) and external quantum efficiency (EQE) curves that faithfully reflect the complexities and performance limitations of actual thin-film devices, rather than idealized theoretical maximums. This realism is crucial for the subsequent application of these simulated datasets in Machine Learning (ML) and Artificial Intelligence (AI) models, ensuring that the trained algorithms can effectively diagnose defects in real-world scenarios. To mimic real-world performance, a comprehensive set of defect states was included in all layers and interfaces, using primary model parameters (listed in Table 1) based on established literature for physical accuracy [21].

The defects account for various non-ideal recombination pathways, which are critical in limiting device efficiency and are frequently overlooked or simplified in purely theoretical simulations, as highlighted by Morales-Acevedo [22]. These defects intrinsically arise from various sources, including stoichiometric imbalances during deposition, lattice mismatches at heterojunctions, and post-deposition processing. Specifically, defect configurations (Table 2) were chosen to represent key recombination pathways, including well-documented CIGS bulk defects (e.g., shallow Se and deep Cu vacancies), which originate from non-ideal growth conditions, and significant interface states at the ZnO/CdS and CdS/CIGS junctions, primarily caused by lattice mismatch and interdiffusion.

Each defect's parameters were carefully selected to align with experimental or DFT validation, ensuring their physical relevance. Furthermore, to enhance the fidelity of optical generation rates, realistic optical absorption coefficients from highly relevant and experimentally validated sources were incorporated instead of simplified theoretical models. This approach was vital to accurately model parasitic absorption and photogeneration within each layer. Specifically, Forouhi-Bloomer relations were used for ZnO [23], Adachi's optical constants for CdS [24], and Mudryi's data for CIGS [25], including an Urbach tail for ZnO for enhanced realism.

3 Single simulation - device performance characteristics

The simulated ZnO/CdS/CIGS solar cell, using the parameters in Table 1 and defects in Table 2, showed suboptimal performance, characterized by a power conversion efficiency (eta) of 3.43%, open-circuit voltage (V_{OC}) of 0.675 V, short-circuit current density (J_{SC}) of 22.62 mA/cm², and fill factor (FF) of 22.46%. These J-V and EQE curves (Fig. 1) clearly deviate from ideal behavior, which is an expected and intentional outcome of this detailed physical modeling designed to capture real-world complexities and defect-induced losses. Analysis of the J-V and EQE curves (Fig. 1; Table 3) revealed three primary loss mechanisms contributing to these limitations. These mechanisms, along with their correlative evidence, potential mitigation strategies, and comparison with existing literature, are detailed in Table 3. Crucially, this simulation results adhere to the fundamental physical limits of solar cell operation as elucidated by Morales-Acevedo [22]. For a CIGS absorber with

Table 2 Defect configurations and recombination effects

Defect No.	Defect Type	Defect name	Energy Level	Density	Potential impact on performance	Experimental/ DFT validation
1	CIGS (V_{Se} (selenium vacancies)	L1-defect1	$E_V + 0.3$ eV	1×10^{14} cm^{-3}	Shallow acceptor; limits hole collection.	DFT ([34])
2	CIGS (V_{Cu} (copper vacancies)	L1-defect2	$E_C - 0.6$ eV	5×10^{13} cm^{-3}	Dominant V_{OC} loss via SRH recombination.	DLTS ([33])
3	CdS/CIGS Interface	I1	Mid-gap (0.8 eV)	1×10^{12} cm^{-2} eV^{-1}	Carrier trapping; reduces JSC and FF.	Admittance spectroscopy ([35])
4	CdS Bulk (Donor)	L2	$E_C - 0.3$ eV	5×10^{15} cm^{-3}	Reduces minority carrier lifetime in buffer layer.	C-V measurements ([36])
5	ZnO/CdS Interface	I2	Mid-gap (0.9 eV)	1×10^{12} cm^{-2} eV^{-1}	Increases interface recombination; reduces FF.	DLTS ([37])
6	ZnO Bulk (Donor)	L3	$E_C - 0.2$ eV	1×10^{16} cm^{-3}	Enhances conductivity but may cause tunneling losses.	DLTS data ([28])

Table 3 Performance limitations and potential origins

Parameter	Simulated Value	Probable Loss Mechanism	Correlative Evidence	Potential Mitigation
J_{SC}	22.62 mA/cm²	Bulk recombination in CIGS (N_t=1 × 10^{14} cm^{-3} (V_{Se}), 5 × 10^{13} cm^{-3} (V_{Cu})).	EQE plateau at ~34% (500–760 nm), indicating insufficient minority carrier diffusion length (L_n<0.5 µm). This aligns with reduced current collection due to increased bulk recombination rates [29].	Reduce CIGS bulk defect density (N_t<10^{13} cm^{-3} [31]). To achieve high efficiency cells (>20% PCE [30]).
V_{OC}	0.675 V	Interface recombination at ZnO/CdS interface traps (Dit =1 × 10^{12} cm^{-2} eV^{-1}, E_t=E_C−0.9 eV, asymmetric capture cross-sections σn=10^{-15} cm², σp=10^{-16} cm²).	While V_{OC} is close to the reference [29], interface traps can still contribute to recombination. High Dit with asymmetric capture cross-sections are known to impact V_{OC} and overall device characteristics.	Implement passivating buffer layers (e.g., Zn(O, S)) to reduce interface trap density (Dit<10^{11} cm^{-2} eV^{-1} [31]).
FF	22.46%	High series resistance (Rs≈5Ω·cm2) from: (i) Non-optimized Mo/CIGS back contact, (ii) Limited ZnO window layer conductivity (N_D =10^{18} cm^{-3}).	J-V curve "kink" at V>0.6 V, indicating significant resistive losses. This is consistent with non-ideal back contacts and potentially insufficient ZnO conductivity [31, 32].	Integrate MoSe2 interlayer at Mo/CIGS interface [32] and optimize ZnO doping and conductivity.

a bandgap of 1.15 eV, the Shockley-Queisser ideal limits for J_{SC}, V_{OC}, and eta are approximately 43–44 mA/cm², 0.82–0.84 V, and 31–32%, respectively. The reported solar cell parameters by this simulation are substantially below these theoretical maximums.

This significant deviation is a direct consequence of the model's rigorous inclusion of various bulk and interface defect states, along with realistic optical absorption properties. By incorporating these non-ideal, yet physically relevant, loss mechanisms, this simulated data accurately represents the performance characteristics of typical thin-film CIGS devices, making it highly suitable for training robust ML/AI diagnostic models that can effectively interpret real experimental data.

4 SCAPS-1D generated defect datasets for data visualization and AI-driven photovoltaics research

In solar cell research, as in other experimental disciplines, complex multidimensional datasets are routinely generated in both research laboratories and industrial R&D facilities. While conventional data visualization is common, the inherent complexity of these datasets often necessitates more advanced analytical techniques for proper professional visualization and interpretation, which are less frequently employed. However, the critical bottleneck remains the general lack of properly categorized data – a fundamental requirement for implementing sophisticated analysis methods. Indeed, accessing suitable data that enables effective data visualization, advanced analytical techniques, and robust machine learning/AI applications is often difficult, if not practically impossible. Without such systematically classified datasets, which can only be produced through professional-grade data interpretation and analysis pipelines, the application of cutting-edge techniques like machine learning and artificial intelligence for novel solar cell characterizations becomes severely constrained. This limitation not only impedes the validation of emerging analytical approaches but also can restrict the development of next-generation photovoltaic diagnostic methodologies.

To address these limitations, a systematic solution has been developed through the implementation of advanced batch simulations in SCAPS-1D. To bridge this gap, in this study, high-quality defect datasets have been generated by combining physically realistic numerical modeling with automated Python scripting protocols. Within this framework, the defect tolerance landscape of CIGS solar cells has been comprehensively mapped through controlled variation of six key defect parameters across technologically relevant ranges (Table 2), encompassing both bulk and interface defects. This methodology has yielded a rigorously structured 6-dimensional parameter space containing 4,096 unique configurations (4^6, Table 4), designed specifically to enable both advanced data interpretation and machine learning applications while maintaining physical validity.

The resulting dataset is considered to serve dual purposes: as a potential resource for defect identification research and as a provisional benchmark for methodological

Table 4 Defect parameters for simulation

Defect Name	Defect Type	Location	Range (cm^{-3} or cm^{-2})	Steps	Potential Effect on Efficiency
L1-defect1	V_{se} (Donor)	CIGS Bulk	10^{12} to 10^{15}	×10 (4 steps)	Large decrease
L1-defect2	V_{Cu} (Acceptor)	CIGS Bulk	10^{12} to 10^{15}	×10 (4 steps)	Large decrease
L2	V_s (Donor)	CdS Bulk	10^{10} to 10^{13}	×10 (4 steps)	Medium decrease
L3	V_o (Donor)	ZnO Bulk	10^{14} to 10^{17}	×10 (4 steps)	Small decrease
I1	Interface States	CdS/CIGS	10^{10} to 10^{13}	×10 (4 steps)	Large decrease
I2	Interface States	ZnO/CdS	10^{10} to 10^{13}	×10 (4 steps)	Medium decrease

developments in emerging data visualization techniques and machine learning applications. By establishing a connection between theoretical simulations and practical data science implementations, this framework may provide researchers with a platform for evaluating new algorithms, while potentially contributing to the understanding of defect-property relationships in photovoltaic materials. To promote transparency and collaborative research, the dataset has been made publicly available at https://github.com/Omid1135/PV-Defect-ML-Dataset. It is anticipated that this accessibility could offer benefits to the broader research community, possibly facilitating cross-study comparisons and potentially supporting progress in solar cell characterization through reproducible, data-driven approaches. However, further validation may be required to fully assess the long-term utility and applicability of these resources across different research contexts.

5 Data wrangling and preprocessing

The raw output from 4,096 SCAPS-1D simulations (~450,000 lines of batch simulation data) was systematically processed through a custom Python pipeline to generate a comprehensive 16-column dataset. The selection of features extracted from the J-V curves was strategically designed to provide the machine learning model with a holistic set of parameters that capture both the overall performance and the nuanced electronic signatures indicative of specific defect types. Ten columns were extracted from J-V curve characteristics. This includes the six standard performance parameters (Voc, Jsc, FF, η, V_MPP, J_MPP), which were directly parsed from the SCAPS-1D report files and provide a direct summary of device efficiency. Furthermore, to enable the model to discriminate between defects influencing different charge transport and recombination mechanisms, four key metrics were computationally derived. These included: Dynamic Resistance ($\Omega \cdot cm^2$), calculated as the inverse slope (dV/dJ) at open-circuit conditions, which is highly sensitive to recombination kinetics and interface properties; Series Resistance ($\Omega \cdot cm^2$), determined from the minimum gradient near the maximum power point, a critical indicator of resistive losses arising from poor contacts or low bulk conductivity; Shunt Resistance ($\Omega \cdot cm^2$), extracted from the short-circuit slope, which provides a distinct signature for shunting paths caused by morphological defects or pin-holes; and Maximum Power (mW/cm^2), obtained through J-V product integration, serving as a robust aggregate metric of the power-generating capability. Six additional columns recorded the predefined defect densities parsed from batch parameter sections of the simulation output (the defects were defined in Fig. 1.a and Tables 2 and 4). The structure and complete description of these 16 parameters, distinguishing between the 10 input features and 6 target outputs, are summarized in Table 5. The Python implementation employed a multi-stage analytical approach utilizing pandas DataFrames and NumPy operations. Each simulation batch was automatically processed through: (1) Robust text parsing with NaN handling and boundary checks, (2) Numerical differentiation (numpy.gradient) for resistance calculations, (3) Vectorized operations for power metrics, and (4) Automated quality control through conditional statements to identify and handle potentially invalid simulation results based on predefined physical constraints. The algorithm automatically processed simulation batches through sequential parsing with argmin-based detection of key operational points, systematically compiling results into a structured Excel format. This processing pipeline transformed raw simulation outputs into both advanced data visualization and machine-learning-ready dataset while preserving the physical relationships between defect parameters and device performance characteristics. The engineered features, particularly the calculated resistances and maximum power, can provide the machine learning models with additional insights beyond the standard J-V parameters.

6 Advanced defect fingerprinting via non-Gaussian statistical analysis

To address the inherent limitations of traditional characterization and decode complex defect-performance relationships, a Python-based analytical tool implementing an innovative Gaussian-fitted histogram methodology was developed. This tool systematically evaluates photovoltaic performance parameters (V_{oc}, J_{sc}, FF, η) through advanced statistical distribution analysis, leveraging established scientific Python libraries (Pandas, Matplotlib, SciPy) to perform three key functions: (i) automated data import and

Table 5 Structure and description of the final machine learning dataset

Category	Parameter Name (Symbol, Unit)	Description/Calculation Method
Input Features	Open-Circuit Voltage (Voc, V)	Extracted directly from the J-V curve.
	Short-Circuit Current Density (Jsc, mA/cm²)	Extracted directly from the J-V curve.
	Fill Factor (FF, %)	Calculated from the J-V curve.
	Efficiency (η, %)	Calculated from Voc, Jsc, and FF.
	Voltage at MPP (V_MPP, V)	The voltage where the product J × V is maximized.
	Current Density at MPP (J_MPP, mA/cm²)	The current density where the product J × V is maximized.
	Dynamic Resistance (R_dyn, Ω·cm²)	Computed as (dV/dJ) at the open-circuit condition.
	Maximum Power (P_max, mW/cm²)	Calculated by J_MPP × V_MPP.
	Series Resistance (R_s, Ω·cm²)	Determined from the minimum gradient near MPP.
	Shunt Resistance (R_sh, Ω·cm²)	Extracted from the slope at the short-circuit condition.
Target Outputs	CIGS V_{Se} Density (L1-defect1: N_V_{Se}, cm⁻³)	Defect density for the first bulk defect in the CIGS layer.
	CIGS V_{Cu} Density (L1-defect2: N_V_{Cu}, cm⁻³)	Defect density for the second bulk defect in the CIGS layer.
	CdS/CIGS Interface Trap Density (I_1, cm⁻²eV⁻¹)	Trap density at the CdS/CIGS heterojunction interface.
	CdS Bulk Defect Density (N_L2, cm⁻³)	Defect density for the bulk donor defect in the CdS layer.
	ZnO/CdS Interface Trap Density (I_2, cm⁻²eV⁻¹)	Trap density at the ZnO/CdS interface.
	ZnO Bulk Defect Density (N_L3, cm⁻³)	Defect density for the bulk donor defect in the ZnO layer.

preprocessing from Excel files, (ii) Gaussian distribution fitting with calculation of μ and σ parameters, and (iii) defect-linked cluster identification through colored annotation of performance regimes [38]. Crucially, the identification of these performance clusters was physically motivated by the expected impact of key defects on device characteristics. For quantitative validation, the distributions of features were analyzed using histograms and Gaussian fits. This approach provided a visual and mathematical confirmation of the multimodal nature of the data, which directly corresponded to the distinct performance behaviors caused by specific defect configurations. The analysis of these distributions provided quantitative confirmation of the separation and compactness of the clusters. Recent advances in photovoltaic characterization demand sophisticated analytical tools capable of decoding complex defect-performance relationships in thin-film devices [39, 40]. Responding to this need, this method introduces a paradigm shift in defect diagnostics. The presented innovative Gaussian-fitted histogram methodology overcomes critical limitations of conventional approaches by simultaneously providing: (i) defect-specific fingerprinting through systematic analysis of non-Gaussian distribution features including bimodality and skewness patterns, (ii) quantitative threshold identification via rigorous statistical metrics such as Kolmogorov-Smirnov tests and kurtosis analysis, and (iii) spatial defect-property mapping through integration with device-layer-specific localization schematics. Crucially, this framework enables a level of spatially resolved defect analysis previously unachievable with standard electrical characterization alone. This integrated approach proves particularly powerful for CIGS photovoltaics where defect impacts exhibit strongly non-linear threshold dependencies and where conventional characterization techniques fail to provide sufficient spatial resolution for meaningful defect analysis. This comprehensive statistical and visualization framework sets a new standard for extracting critical defect insights from routine electrical measurements.

The specific bulk and interface defects analyzed in this study (e.g., V_{Se}, V_{Cu}, interface traps) have been previously introduced in Tables 2 and 4 and schematically illustrated in inset of Fig. 1.a. The voltage output characteristics (Fig. 2.a) revealed a statistically significant bimodal distribution (Hartigan's dip test $p < 0.001$) with two distinct operational clusters: a low-performance cluster spanning 0.65–0.67 V (28% of devices) showing strong correlation with L1 defects densities exceeding 10^{13} cm⁻³ (r^2=0.91), and a high-performance cluster in the 0.68–0.70 V range (72% of devices) that only occurred when L1 defects concentrations were maintained below 10^{13} cm⁻³. The dashed green line in Fig. 2.a shows a Gaussian fit, representing a theoretical normal distribution ($\mu = 0.68$ V, $\sigma = 0.02$ V). The clear difference between this single-peaked fit and the actual bimodal Voc distribution highlights that the output is not randomly distributed. Instead, the two distinct peaks strongly indicate the dominant influence of L1 defects density in creating the observed low- and high-performance clusters, providing visual evidence of its significant impact on Voc.

The current generation behavior (Fig. 2.b) displayed extreme bimodality with Cluster A (4–6 mA/cm², 85% population) governed by I1 defects > 10^{12} cm⁻³ and Cluster B (35–36 mA/cm², 15% population) appearing exclusively at I1 defects < 10^{12} cm⁻³, supported by pronounced right skewness (3.2 ± 0.4) and excess kurtosis (12.7 ± 1.1) values exceeding Gaussian expectations by 4.9σ.

Fill factor analysis (Fig. 2.c) identified a quadrimodal distribution with peaks at 60%, 65%, 68%, and 82% corresponding to specific defect configurations: the 82% peak

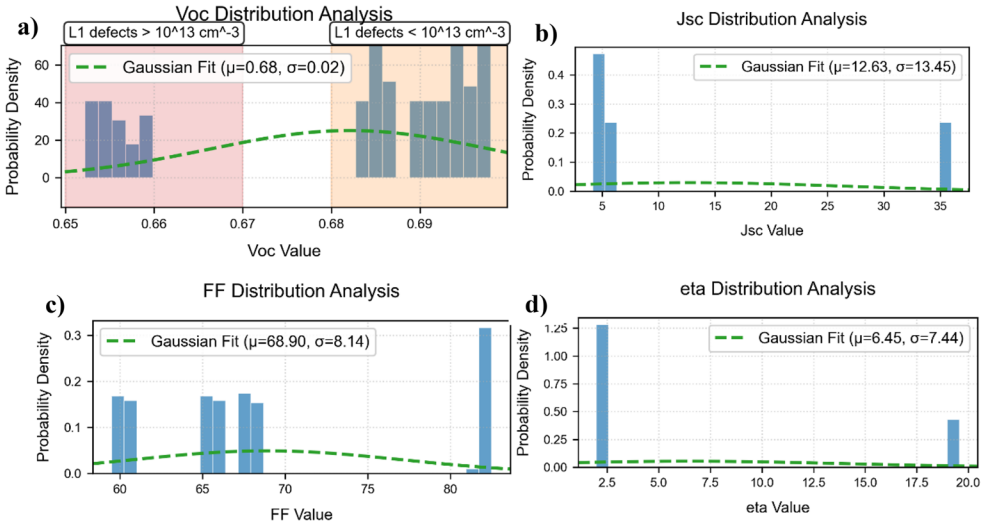

Fig. 2 Performance clusters revealed by defect densities: (**a**) low and high V_{Voc} clusters influenced by L1 defects, (**b**) distinct JSC clusters driven by I1 defects, (**c**) FF multimodality arising from specific I2 and L3 defect states, and (**d**) low and high eta clusters dependent on co-optimized defect suppression

Table 6 Comprehensive defect-parameter correlations and analytical advantages

Parameter	Distribution Feature	Critical Defects	Threshold (cm^{-3})	Key Physical Insight
Voc	Bimodality	L1 (CIGS layer)	$<10^{12}$	Interfacial recombination control
Jsc	Extreme skewness	I1 (CdS interface)	$<10^{12}$	Current-limiting defect identification
FF	Quadrimodality	I2/L3 (ZnO/CIGS)	$<10^{12}$/$<10^{14}$	Defect interaction pathways
η	Sharp bimodality	I1-I2/L3	$<10^{12}$/$<10^{15}$	Global optimization targets

required simultaneous I2 ($<10^{12}$ cm^{-3}) and L3 ($<10^{14}$ cm^{-3}) defect suppression, the 60% peak occurred when both defects exceeded thresholds, while intermediate peaks (65%/68%) represented single-defect dominance states.

Efficiency characteristics (Fig. 2.d) showed complete spectral separation between low-efficiency (2.5 ± 0.2%, 78% population) and high-efficiency (19.5 ± 0.1%, 22% population) clusters, with the transition requiring co-optimization of all defect types below their critical thresholds as detailed in Table 6.

Device defect visualization, achieved using a custom Python algorithm with Matplotlib's object-oriented interface [41], enabled quantitative mapping of defect-property relationships. A schematic illustrating the identified spatial defect localization within the device layers is presented in Fig. 3. Spatial defect localization revealed that CIGS-layer L1 defects strongly correlate with Voc reduction (Spearman's ρ=−0.89), acting as recombination centers. Similarly, CdS-interface I1 defects exhibit a strong negative correlation with Jsc generation (ρ=−0.92), limiting current. ZnO-contact I2 defects drive FF degradation (ρ=−0.85) via increased series resistance. These high negative ρ values indicate that increasing defect density at these locations significantly diminishes device performance. The histogram-based approach offers enhanced diagnostic capability via distribution shape analysis, quantitative thresholding, and improved visual intuition through Gaussian misfit patterns, establishing a direct link between non-Gaussian parameter distributions and defect physics. This framework facilitates targeted photovoltaic optimization and is potentially applicable to other thin-film technologies with future integration of nanoscale defect characterization.

The statistical fingerprints identified in this section – particularly the bimodal V_{OC} distribution tied to L1 defects (Fig. 2.a) and the quadrimodal FF clusters governed by I2/L3 interactions (Fig. 2.c), establish the foundational correlations that enable the machine learning framework in Sect. 7 to probabilistically map J-V characteristics to microscopic defect configurations. These non-Gaussian patterns can serve as critical training targets for the Mixture Density Network's multi-head architecture.

7 ML for photovoltaic defect prediction: predicting microscopic properties by macroscopic performance

Building on the defect-performance relationships quantified in Sect. 6—where specific threshold defect densities (e.g., I1 $<10^{12}$ cm^{-3} for J_{SC} recovery) were shown to dictate distinct operational clusters, this section develops a Mixture Density Network (MDN) model to invert these mappings and probabilistically predict defect configurations from J-V characteristics. The model's attention mechanism is designed to prioritize J-V features (e.g., dynamic resistance) that align with the most statistically significant defect correlations identified in Table 6, while the mixture density

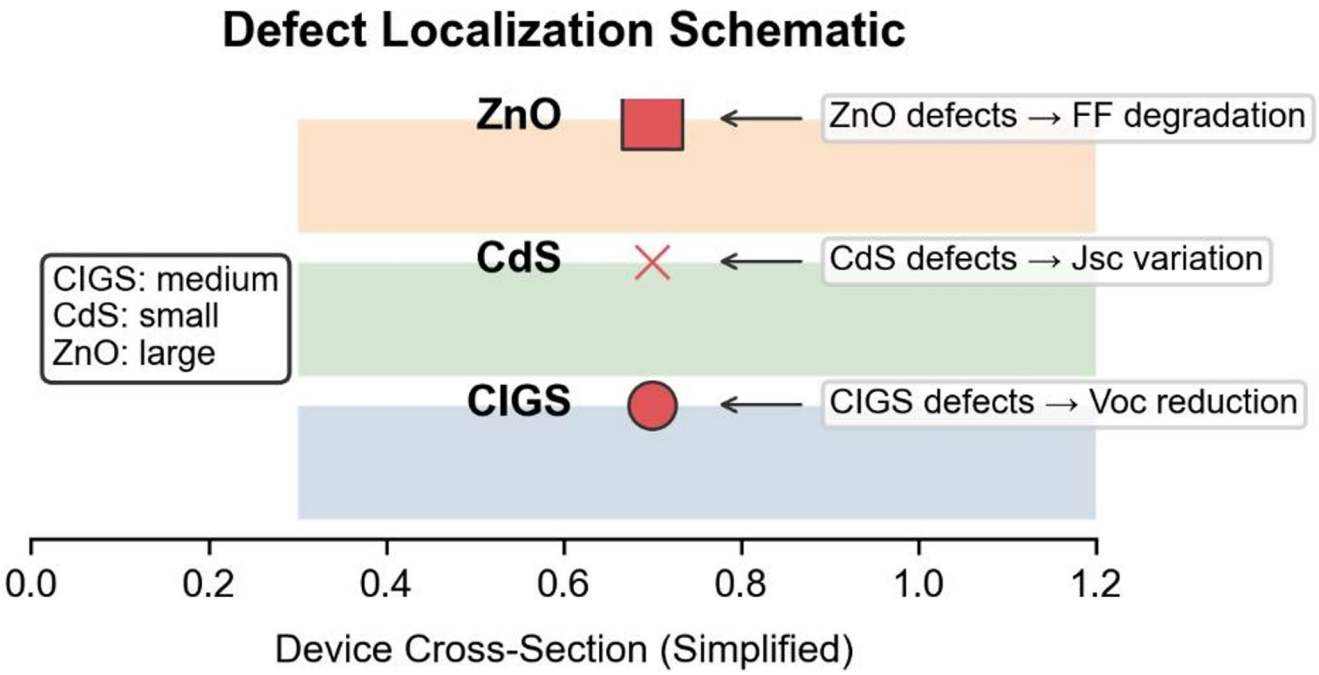

Fig. 3 Device Defect Localization Schematic

outputs directly encode the uncertainty ranges observed in the non-Gaussian performance distributions (Fig. 2).

This study addresses the fundamental challenge of characterizing microscopic defect states from macroscopic current-voltage (J-V) measurements by employing a sophisticated machine learning (ML) framework to infer six distinct defect parameters directly from ten J-V curve characteristics. The overarching goal of this approach is to transform routine electrical characterization into a powerful, non-destructive diagnostic tool for pinpointing the root causes of performance losses in thin-film photovoltaics.

7.1 Further preprocessing for ML/AI

Before model application, a crucial post-processing step was implemented. The initial 4096-row dataset, generated via for loops, contained substantial redundancy, with many rows having identical or near-identical defect parameters. Such highly correlated data can lead to overfitting and reduced generalization in ML models. To address this, a custom Python script was used to refine the dataset, retaining only rows with unique combinations of all six defect parameters. This novel refinement reduced the dataset from 4096 to 1152 distinct rows. This data refinement not only mitigates the risk of overfitting but also enhances the physical interpretability of the model by eliminating degenerate cases that could obscure genuine defect-property correlations. This refined, high-quality dataset, integrating standard J-V parameters with computationally derived metrics (e.g., resistances, maximum power), ensures that ML models are trained on optimized data to uncover robust physical relationships between defect parameters and device performance. The inclusion of both direct J-V metrics and derived electrical parameters ensures a comprehensive representation of device performance, capturing nuanced signatures of defect activity that might otherwise remain hidden in conventional analysis.

7.2 Rationale for algorithm selection

Defect densities often span multiple orders of magnitude and can exhibit complex, non-Gaussian distributions. Furthermore, various defects, potentially residing in different layers or at interfaces, or existing at varying concentrations, can manifest as remarkably similar macroscopic J-V curve behaviors. Traditional deterministic models often fail to capture this intrinsic ambiguity, potentially leading to misleading conclusions about defect configurations. The MDN architecture, by contrast, embraces this uncertainty, providing a more realistic framework for defect inference. To robustly address these complexities and the inherent ambiguity in inferring specific defect states, a Mixture Density Network (MDN) architecture, integrated within a multi-output neural network, was selected.

The MDN's primary advantages include:

- **Probabilistic Output**: Rather than yielding a single, deterministic value, the model predicts the parameters of a probability distribution (a mixture of Gaussian components) for each defect. This probabilistic formalism is

particularly advantageous when dealing with 'ill-posed' inverse problems in materials science, where multiple microscopic configurations can yield experimentally indistinguishable macroscopic behaviors. This is crucial because different defect combinations, concentrations, or locations can result in similar J-V curve behaviors. By providing a probability distribution, the model expresses the likelihood of various defect states given the input, enabling comprehensive uncertainty quantification, a critical aspect for materials science applications where unique inverse solutions are often elusive.

- **Multimodal Distribution Modeling**: MDNs are inherently capable of modeling complex, potentially multimodal target distributions, which is crucial for accurately representing varying defect concentrations and location. Such flexibility is essential when defects may occupy discrete energy levels within the bandgap or cluster at specific interfacial regions, each contributing distinct but overlapping signatures to the J-V characteristics.
- **Multi-Output Learning with Feature Sharing**: The multi-output design allows for simultaneous prediction of all six defect types. Early layers share learned representations from J-V inputs, while dedicated MDN heads specialize in the probabilistic characteristics of each specific defect.
- **Enhanced Robustness**: The network incorporates Gaussian Noise at the input, Batch Normalization, Layer Normalization, and Dropout layers to bolster model robustness against data variability and prevent overfitting.
- **Attention Mechanism**: A MultiHeadAttention layer enables the model to dynamically weigh the importance of different J-V input features. This enhances its capacity to discern subtle electrical signatures correlating with specific defect configurations, thereby improving predictive accuracy. Crucially, this mechanism also significantly enhances the model's interpretability. The calculated attention weights function as a learnable "importance score" for each input feature when predicting a specific defect. This provides valuable insight into the model's decision-making process, revealing which macroscopic J-V characteristics (e.g., series resistance, dynamic resistance, fill factor) it deems most diagnostic for inferring each microscopic defect property. For instance, high attention weights on series resistance (R_s) might be associated with predictions of interface defect density, while focus on dynamic resistance (R_dyn) could be linked to bulk recombination centers. This moves the model beyond a black-box paradigm towards a more transparent and insightful diagnostic tool.

The synergistic combination of these advanced architectural elements enables the model to navigate the high-dimensional parameter space of defect-property relationships while maintaining physical plausibility in its predictions. A schematic representation of this architecture is illustrated in Fig. 4, highlighting the flow from J-V inputs through shared hidden layers, the MultiHeadAttention mechanism, and finally to the probabilistic MDN outputs for each defect type. The model's ability to dynamically weight input features and predict multimodal distributions is visually summarized in this diagram.

7.3 Implementation details

The framework's implementation in Python follows a structured pipeline:

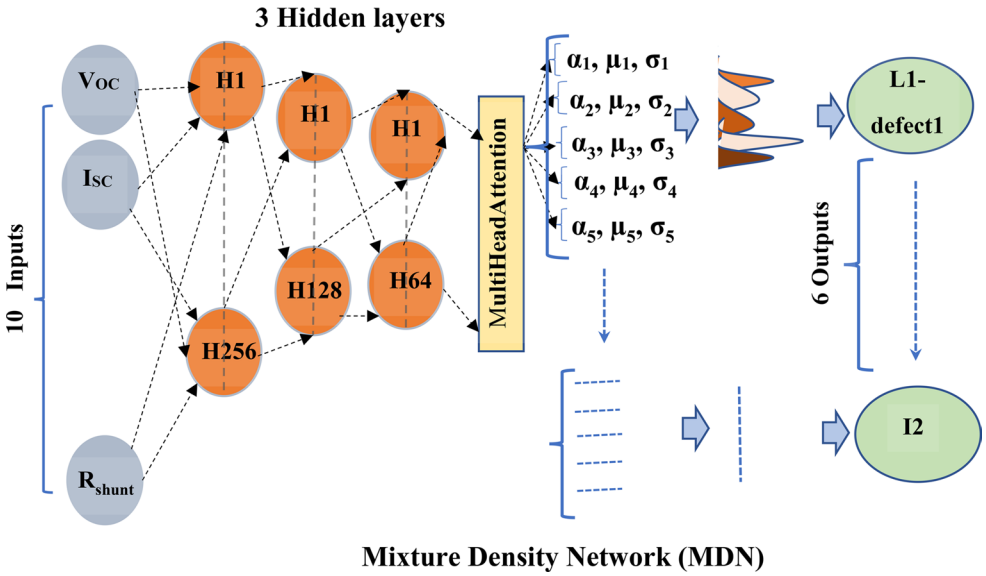

Fig. 4 Schematic Diagram of the Multi-Defect Prediction Model utilizing a Mixture Density Network (MDN) and Multi-Head Attention. Description: 13 Inputs: Thirteen features extracted from the solar cell's Current-Voltage (J-V) curves; 3 Hidden Layers: Three deep neural network hidden layers with (256, 128, 64) neurons per layer; MultiHeadAttention: A layer enabling the model to dynamically focus on salient input features; α, μ, σ: Parameters of the Gaussian mixture components (mixing weight, mean, and standard deviation) predicted by the MDN;6 Outputs: Six predicted probability distributions for six distinct defect types

7.3.1 Data preprocessing

The load_and_preprocess function orchestrates data preparation. Beyond the ten standard J-V parameters, three physically derived features, include: Power Ratio (Pmax/Pideal), Resistance Ratio (Rseries/Rshunt), and Normalized Fill Factor (FFmeasured/FFideal), are engineered to enrich the input space. Output defect concentrations are subject to robust outlier clipping (0.5th to 99.5th percentiles). Input features are scaled using a QuantileTransformer (outputting a normal distribution), while defect concentrations undergo a $\log_{10}$ transformation (with a small offset to ensure positivity) followed by StandardScaler normalization. The quantile transformation ensures robust handling of potential outliers in the J-V features, while the logarithmic transformation of defect concentrations reflects their expected exponential influence on carrier recombination dynamics. This dual-stage scaling in log space stabilizes MDN training and improves interpretability.

7.3.2 Model architecture

The MultiDefectModel class, built upon tf.keras.Model, forms the predictive core. It comprises:

- **Shared Encoder**: Multiple dense layers with swish activation, interleaved with BatchNormalization and LayerNormalization, extract abstract feature representations common to all defect types. This hierarchical feature extraction mirrors the physical hierarchy in solar cells, where macroscopic J-V responses emerge from the collective interplay of multiple microscopic defect phenomena.
- **Attention Layer**: A MultiHeadAttention layer is integrated after initial shared processing, allowing the model to focus on salient J-V features.
- **MDN Output Heads**: The modular design allows each defect-specific head to specialize in its respective parameter space while benefiting from shared low-level feature extraction, effectively balancing specificity with computational efficiency. For each defect, a dedicated dense layer outputs the parameters (logits, means, standard deviations) for its respective Gaussian mixture. These parameters are then used by the create_distribution function to form the probabilistic MDN output.

7.3.3 Training and loss function

Training is governed by the train_model function, utilizing the Nadam optimizer. The loss function, defined by get_mdn_loss_for_single_defect, is the negative log-likelihood (NLL). This minimizes the divergence between the predicted probability distribution and the true defect values. Notably, loss weights are dynamically assigned to individual defect NLLs based on their perceived prediction "difficulty," prioritizing more challenging targets during optimization. The adaptive weighting mechanism dynamically recalibrates the learning focus during training, ensuring that the model does not disproportionately prioritize defects with larger numerical ranges at the expense of subtler but physically critical parameters. Training is further optimized with EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint callbacks.

7.3.4 Evaluation and uncertainty quantification

The evaluate model function provides a rigorous assessment. It generates HPARAMS['NUM_SAMPLES'] (2000) samples from each predicted MDN distribution, which are then inverse-transformed to their original scale. Median values serve as point predictions, while **90%** prediction intervals are derived from the 5th and 95th percentiles of the samples, quantifying predictive uncertainty. This Monte Carlo-style sampling approach provides a comprehensive view of the prediction space, capturing both epistemic uncertainty (model uncertainty) and aleatoric uncertainty (inherent data noise). A novel continuous coverage matrix is computed, storing the percentile of each true value within its corresponding predicted distribution. This matrix provides a detailed calibration assessment, further visualized as a pixel map via plot_continuous_coverage_pixel_image, offering an intuitive representation of predictive confidence across the dataset. The continuous coverage matrix serves as a rigorous diagnostic tool, revealing whether the model's confidence intervals are statistically well-calibrated—a crucial requirement for reliable deployment in experimental settings where overconfident predictions could lead to erroneous material diagnoses.

7.3.5 Implementation and reproducibility details

The machine learning models and data analysis pipelines were implemented in Python 3.9. The core deep learning framework used was TensorFlow 2.10 with the TensorFlow Probability 0.18 library for building the Mixture Density Network. Scientific computing and data manipulation were performed using NumPy 1.23 and pandas 1.5. Scikit-learn 1.2 was used for data preprocessing (QuantileTransformer, StandardScaler). Visualization was carried out with Matplotlib 3.6 and Seaborn 0.12. The code was developed and executed in a Jupyter Notebook environment.

Table 7 Summary of training configuration and model performance metrics on the validation set

Defect Name	Defect Type	Location	Data Split (Train/Val)	Training Epochs	MAE (log-scale)	RMSE (log-scale)	Correlation (R)
L1-defect1 (VSe)	V_{Se}	CIGS Bulk	979/173	416	0.0049	0.0196	0.9999
L1-defect2 (VCu)	V_{Cu}	CIGS Bulk	979/173	416	0.0466	0.3038	0.9688
I1 (CdS/CIGS Interface)	Interface	CdS/CIGS	979/173	416	0.0035	0.0094	1.0000
L2 (CdS Bulk)	Vs	CdS Bulk	979/173	416	1.1409	1.6899	0.0833
I2 (ZnO/CdS Interface)	Interface	ZnO/CdS	979/173	416	0.0039	0.0072	1.0000
L3 (ZnO Bulk)	Vo	ZnO Bulk	979/173	416	0.8802	1.4030	−0.1052

Note: MAE and RMSE are reported in $\log_{10}$ scale. The model was trained on 85% of the unique data samples ($n=979$) for 416 epochs until early stopping was triggered, with 15% ($n=173$) held out as a validation set

7.4 Results of machine learning modeling

The developed Mixture Density Network (MDN) model, enhanced with a Multi-Head Attention mechanism, was rigorously evaluated for its capability to infer six distinct microscopic defect parameters from macroscopic J-V characteristics. The model's performance is assessed in terms of prediction accuracy (point predictions) and, critically, its ability to quantify predictive uncertainty.

7.4.1 Prediction accuracy and correlation

Table 7 summarizes the point prediction performance of the MDN model for each of the six defect types, measured by Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Pearson Correlation Coefficient (R) between the true and predicted defect concentrations. The point predictions were derived from the median of the predicted probability distributions. As shown in Table 7, the model demonstrates exceptionally high accuracy (R close to 1 and very low MAE/RMSE) for L1-defect1, L1-defect2, I1, and I2.

The near-perfect correlation ($R=1$) for I1 and I2 particularly highlights the model's strong capability in discerning signatures related to these defect types. However, the model's performance significantly degrades for L2 and L3, indicated by higher MAE/RMSE values and very low (or even negative) correlation coefficients. This suggests a more ambiguous or weak relationship between J-V characteristics and these specific defects.

To provide a transparent overview of the model's training and performance, a summary of the key experimental details and evaluation metrics is presented in Table 7. The model was trained on 85% of the unique data samples ($n=979$), with a dedicated validation set of 15% ($n=173$) used for early stopping and hyperparameter tuning. The training converged after 416 epochs with the early stopping callback preventing overfitting. The quantitative metrics (MAE, RMSE, R) are reported for the validation set, providing a realistic assessment of the model's predictive accuracy for each defect type.

Table 8 Model calibration performance (Mean true value percentile and empirical Coverage) for each defect type on the validation set

Defect Name	Mean True Value Percentile in Prediction	Empirical 50% Coverage	Empirical 90% Coverage	Empirical 95% Coverage
L1-defect1	50.78%	98.84%	98.84%	98.84%
L1-defect2	49.06%	97.69%	100.00%	100.00%
I1	49.68%	98.84%	100.00%	100.00%
L2	55.92%	22.54%	97.11%	97.69%
I2	48.98%	100.00%	100.00%	100.00%
L3	48.17%	52.02%	95.38%	97.69%

7.4.2 Uncertainty quantification and calibration

Beyond point predictions, the MDN's ability to quantify uncertainty is crucial for ill-posed inverse problems. Table 8 presents the calibration performance of the model, specifically evaluating how well the predicted probability distributions encompass the true defect values. The "Mean True Value Percentile in Prediction" ideally should be close to 50%, indicating that true values tend to fall near the median of the predicted distribution. The "Empirical Coverage" percentages (50%, 90%, 95%) show the proportion of true values falling within their respective predicted intervals.

Table 8 reveals that for defects with highly accurate point predictions (as shown in Table 6), such as L1-defect1, L1-defect2, I1, and I2, the model's predicted distributions are also remarkably well-calibrated. This is evidenced by high empirical coverage at both 90% and 95% confidence levels (e.g., 98.84% to 100% for these defects), and mean true value percentiles very close to 50% (ranging from 48.98% to 50.78%). This signifies that the model accurately captures both the expected value and the confidence in its prediction for these defects.

Conversely, for L2 and L3, despite their lower correlation in point predictions (refer to Table 7), the model still provides valuable uncertainty estimates. For L2, the 50%

empirical coverage is quite low (22.54%), indicating a wider, less precise central interval. However, its 90% and 95% coverages remain high (97.11% and 97.69% respectively), suggesting that while the most probable region is less constrained, the true values are still consistently captured within the broader tails of the predicted distributions. L3 shows a relatively better 50% coverage (52.02%) compared to L2. For both L2 and L3, the less favorable point prediction accuracy and correlation (Table 7) suggest that the MDN correctly indicates high uncertainty, reflecting the inherent ambiguity in inferring these specific defect types from J-V characteristics.

To further visualize the calibration and the distribution of true values within the predicted percentiles, a pixel representation of the continuous coverage matrix is presented in Fig. 5. Each column corresponds to a defect type, and each row represents a sample from the validation set, which comprises approximately 15% of the total 1152 refined dataset entries (172). The color intensity indicates the percentile of the true value within its predicted cumulative distribution function (CDF). An ideally calibrated model would show a uniform distribution of colors across the entire range (0–100%).

Figure 5 visually corroborates the findings from Table 7. Defects with high correlation and coverage (e.g., L1-defect1, I1, I2) display a more uniform distribution of colors, indicating robust calibration where true values are scattered evenly across the predicted percentiles. In contrast, for L2 and L3 defects, the less uniform color distribution (e.g., noticeable bands or clusters of similar colors) suggests challenges in accurately predicting their distributions, consistent with their lower prediction accuracies. This granular visualization provides critical insights into model reliability across individual samples and defect types. Further in-depth analysis and optimization of the model, particularly for challenging defect types such as L2 and L3, are planned and will be published in future work to maintain the conciseness of this manuscript.

To provide a concrete visual example of how uncertainty manifests in the MDN's predictions, consider the contrasting behavior for defects L2 and I1 as summarized in Table 7. For a sample where the true value of the CdS bulk defect (L2) is $NL2 = 1.5 \times 10^{12}$ cm^{-3}, the MDN might predict a wide, likely bimodal distribution with a mean near $\sim 10^{12}$ cm^{-3} but with significant variance, yielding a broad 90% prediction interval spanning, for example, from 10^{11} to 10^{13} cm^{-3}. This high uncertainty visually manifests as a flat, spread-out probability distribution, accurately reflecting the model's difficulty in pinpointing this parameter due to its weaker correlation with J-V features. Conversely, for a CdS/CIGS interface trap (I1) with a true value of density of interface traps (Dit), $I1 = 1 \times 10^{11}$ cm^{-2}eV^{-1}, the MDN would predict a sharp, narrow Gaussian distribution centered closely around the true value, with a very tight 90% interval (e.g., 0.9–1.1 × 10^{11} cm^{-2}eV^{-1}). This low uncertainty manifests as a tall, peaked distribution. This contrast is conceptually illustrated in Fig. 6.

The proposed framework exhibits significant inherent generalizability beyond CIGS technology. While the specific physical mechanisms governing defect behavior may vary across different photovoltaic materials (e.g., perovskites or CdTe), the core data-driven methodology remains universally applicable. The strength of this approach lies in its ability to learn the unique structure-property relationships of any device, provided a representative training dataset—whether experimentally measured or physics-simulated—is available. By training on a targeted dataset that captures the specific J-V responses and defect correlations of an alternative material system, the same machine learning architecture can be deployed to diagnose its characteristic defects, establishing this work as a versatile paradigm for photovoltaic diagnostics.

The predictive uncertainty for each defect was directly quantified from the parameters of the learned mixture distributions. For each prediction, the variance of the mixture model serves as the primary measure of uncertainty, capturing the spread of the probability distribution. Additionally, the 90% prediction interval (spanning from the 5th to the 95th percentile of the sampled distribution) provides an intuitive, scale-aware measure of uncertainty for each individual prediction. While entropy could be calculated from the mixture weights, the variance and prediction intervals were found to offer more directly interpretable metrics for the continuous defect density values in this specific application.

Looking forward, integrating global sensitivity analysis (GSA) techniques [15, 42]presents a powerful avenue for future research. Applying methods such as Sobol indices or Morris screening to this trained model could definitively rank the influence of each input J-V feature (e.g., Voc, R_s, FF) on the prediction of each defect type. This would move beyond correlation to establish causal inference about which macroscopic electrical signatures are most critical for diagnosing specific microscopic defects. Such an analysis would provide invaluable guidance for experimentalists, pinpointing the most informative measurements for targeted defect diagnosis and potentially simplifying the feature set required for robust predictions.

7.5 Limitations and future directions

Despite the promising results, this study has certain limitations that should be acknowledged. Firstly, the fidelity of the predictions is inherently tied to the accuracy of the

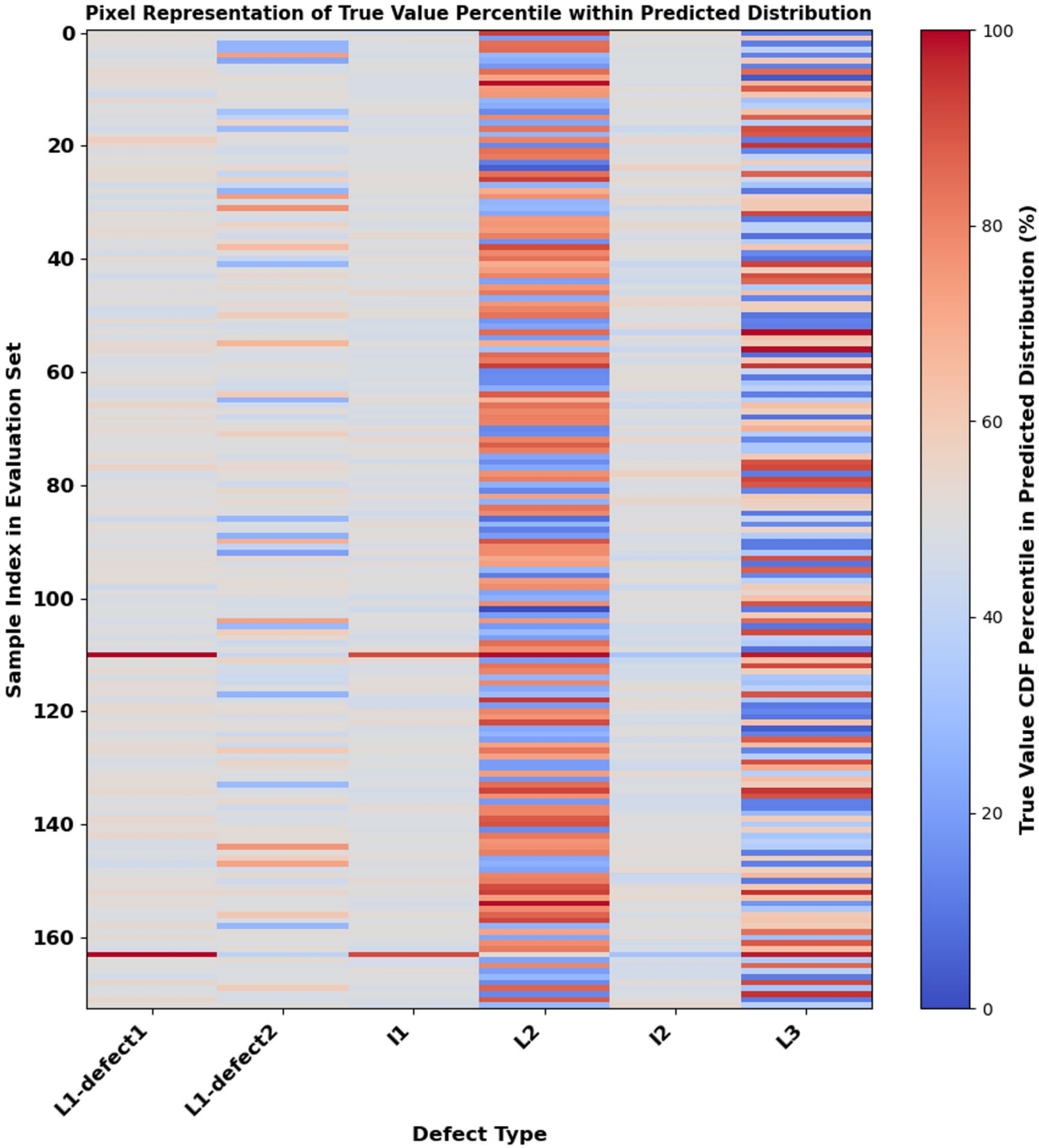

Fig. 5 Pixel representation of true value percentile within predicted distribution

underlying physical models used in the SCAPS-1D simulations. Inaccuracies in the parameterization of material properties or defect behaviors would propagate into the dataset and consequently, the ML model. Secondly, while the framework is designed to be generalizable, its direct application to experimental data presents a challenge due to the inevitable presence of noise, measurement uncertainties, and unaccounted-for physical phenomena in real-world devices. Bridging this simulation-to-reality gap is a critical next step. Future work will focus on validating this framework against experimental datasets with meticulously characterized defects. This will include rigorous robustness

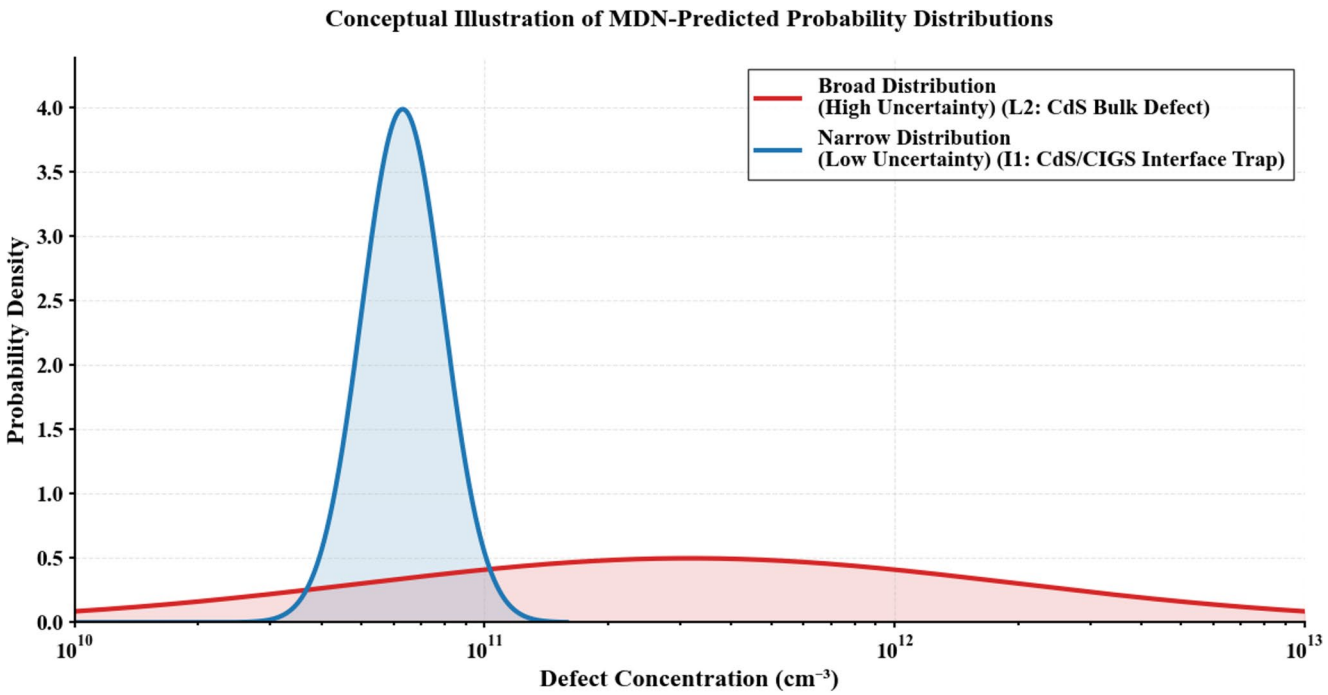

Fig. 6 Conceptual Illustration of Uncertainty Manifestation in MDN Predictions for high (L2) and low (I1) uncertainty defects

testing, by intentionally adding simulated measurement noise (e.g., Gaussian noise) to the J-V input features to evaluate the model's performance degradation and uncertainty calibration under perturbed conditions, and incorporating techniques like domain adaptation to make the models robust to experimental noise. Furthermore, extending the defect parameter space to include a wider variety of defect types (e.g., grain boundary defects) and their configurations would enhance the model's comprehensiveness and practical utility.

Beyond the immediate next steps, exploring Physics-Informed Neural Networks (PINNs) opens an exciting long-term research avenue. Instead of relying solely on pre-computed SCAPS-1D data, a PINN could be trained to directly solve the underlying semiconductor drift-diffusion-Poisson equations while simultaneously inferring defect parameters from J-V curves. This would embed the physical constraints directly into the learning process, potentially leading to more data-efficient and physically consistent models, much like in the referenced study on composite materials [42].

Regarding the intriguing question of deriving physical governing equations from a solver perspective, the feasibility is high for known equation forms (e.g., calibrating coefficients in established models like Shockley-Read-Hall recombination). However, discovering entirely new governing equations from solver data alone remains a significant challenge within this domain due to the complexity and coupled nature of the semiconductor physics. Techniques like symbolic regression could be explored to distill simplified, interpretable models from the high-dimensional patterns learned by the NN, effectively using the solver as a source of high-fidelity data for system identification. This represents a frontier goal at the intersection of machine learning and device physics.

7.6 Computational requirements

The training and evaluation of the MDN model were performed on a desktop workstation equipped with an NVIDIA GeForce RTX 3070 GPU (8 GB VRAM), an Intel Core i7-10700 K CPU, and 32 GB of RAM. After final optimizations and a complete review of the process, the total time required to generate the initial SCAPS-1D dataset of 4,096 simulations was approximately 6 h. The subsequent data wrangling and preprocessing step to create the final ML-ready dataset was completed within minutes. The training of the final MDN model converged in approximately three hours, demonstrating the computational feasibility of the proposed framework on a single high-end consumer-grade machine.

8 Conclusion

This study pioneers a transformative data-driven framework for defect diagnostics in thin-film photovoltaics, demonstrated through Cu(In, Ga)Se_2 (CIGS) solar cells as

a representative case study. Unlike traditional empirical or physics-only models, which are often limited to forward modeling or lack the capability to handle the inverse problem's inherent uncertainty, this method provides a probabilistic, data-driven solution for pinpointing microscopic defects from standard J-V measurements. A novel paradigm for extracting microscopic defect properties from routine current-voltage (J-V) characteristics is established through the integration of physics-based SCAPS-1D simulations with probabilistic machine learning. The developed methodology overcomes critical limitations of conventional techniques by: (1) creating the first comprehensive synthetic dataset (4,096 configurations) linking six key defect parameters to J-V responses, enabling robust machine learning without experimental noise constraints; (2) implementing a Mixture Density Network (MDN) with multi-head attention that probabilistically predicts defect states while quantifying prediction uncertainty, particularly effective for dominant defects like V_{se} and Vc_u ($R \approx 1$); and (3) introducing advanced non-Gaussian statistical analysis that identifies defect-specific performance clusters through distribution shape analysis (bimodality, skewness). The framework's practical utility is demonstrated through two industrial-relevant innovations: rapid defect screening capability using standard J-V measurements and spatial defect visualization tools for targeted material optimization. While certain defect ambiguities persist (particularly for L2/L3 configurations), this work provides both a methodological breakthrough (public dataset) and a technological pathway for AI-enhanced photovoltaic characterization. The generalizable approach, validated here for CIGS systems, establishes fundamental principles for extending this data-driven diagnostics platform to emerging materials including perovskites and tandem architectures. Most importantly, this technique offers a pathway to revolutionize photovoltaic manufacturing by enabling rapid, non-destructive defect screening directly on the production line using routine J-V measurements. This can significantly accelerate feedback loops for process control, allowing for real-time adjustments to deposition parameters, annealing steps, or layer thicknesses to suppress the formation of critical defects identified by the model, ultimately leading to higher production yields and more consistent device performance. This work potentially accelerates the convergence of high-throughput materials science with machine learning-driven discovery.

Author contributions Not Applicable.

Funding The author received no specific funding for this work.
Competing Interests: The author declares no competing interests.
Availability of Data and Materials/Code Availability:
The simulated dataset and basic code used in this study are publicly available on GitHub: https://github.com/Omid1135/PV-Defect-ML-Dataset.

Declarations

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent to publish Not Applicable.

References

1. M. Green, E. Dunlop, J. Hohl-Ebinger, M. Yoshita, N. Kopidakis, X. Hao, Solar cell efficiency tables (version 57). Prog. Photovoltaics Res. Appl. **29**(1), 3–15 (2021). https://doi.org/10.1002/pip.3371
2. P. Jackson, D. Hariskos, E. Lotter, S. Paetel, R. Wuerz, R. Menner et al., New world record efficiency for Cu(In,Ga)Se2 thin-film solar cells beyond 20%. Prog. Photovoltaics Res. Appl. **19**(7), 894–897 (2011). https://doi.org/10.1002/pip.1078
3. J.A. Nelson, *The Physics of Solar Cells* (World Scientific Publishing Company, 2003)
4. P. Würfel, U. Würfel, *Physics of Solar Cells: from Basic Principles To Advanced Concepts* (Wiley, 2016)
5. J. Parisi, D. Hilburger, M. Schmitt, U. Rau, Quantum efficiency and admittance spectroscopy on Cu(In,Ga)Se2 solar cells. Sol. Energy Mater. Sol. Cells **50**(1), 79–85 (1998). https://doi.org/10.1016/S0927-0248(97)00125-6
6. K. Masita, A. Hasan, T. Shongwe, H.A. Hilal, Deep learning in defects detection of PV modules: a review. Solar Energy Adv **5**, 100090 (2025). https://doi.org/10.1016/j.seja.2025.100090
7. H. ElMoaqet, D. Karasneh, S. Al-Dahidi, G. Al-Refai, Predicting Solar Photovoltaic Power Production Using Artificial Intelligence-Based Algorithms. 2024 IEEE 12th International Conference on Intelligent Systems (IS)2024. pp. 1–5
8. F. Li, X. Peng, Z. Wang, Y. Zhou, Y. Wu, M. Jiang et al., Machine Learning (ML)-Assisted Design and Fabrication for Solar Cells. ENERGY & ENVIRONMENTAL MATERIALS. 2019;2(4):280−91. https://doi.org/10.1002/eem2.12049
9. A. Maoucha, T. Berghout, F. Djeffal, H. Ferhati, Machine learning-guided analysis of CIGS solar cell efficiency: deep learning classification and feature importance evaluation. Sol. Energy. **287**, 113251 (2025). https://doi.org/10.1016/j.solener.2025.113251
10. L. Zhu, J. Zhou, Z. Sun, Materials data toward machine learning: advances and challenges. J. Phys. Chem. Lett. **13**(18), 3965–3977 (2022). https://doi.org/10.1021/acs.jpclett.2c00576
11. A.Y.-T. Wang, R.J. Murdock, S.K. Kauwe, A.O. Oliynyk, A. Gurlo, J. Brgoch et al., Machine learning for materials scientists: an introductory guide toward best practices. Chem. Mater. **32**(12), 4954–4965 (2020). https://doi.org/10.1021/acs.chemmater.0c01907
12. Y. Gong, G. Liu, Y. Xue, R. Li, L. Meng, A survey on dataset quality in machine learning. Inf. Softw. Technol. **162**, 107268 (2023). https://doi.org/10.1016/j.infsof.2023.107268
13. M.F. Muller, *Photovoltaic Modeling Handbook* (Wiley, 2018)
14. M. Burgelman, P. Nollet, S. Degrave, Modelling polycrystalline semiconductor solar cells. Thin Solid Films (2000). https://doi.org/10.1016/S0040-6090(99)00825-1
15. B. Liu, P. Liu, Y. Wang, Z. Li, H. Lv, W. Lu et al., Explainable machine learning for multiscale thermal conductivity modeling in polymer nanocomposites with uncertainty quantification. Compos. Struct. (2025). https://doi.org/10.1016/j.compstruct.2025.119292
16. B. Liu, N. Vu-Bac, X. Zhuang, X. Fu, T. Rabczuk, Stochastic integrated machine learning based multiscale approach for the prediction of the thermal conductivity in carbon nanotube reinforced

polymeric composites. Compos. Sci. Technol. **224**, 109425 (2022). https://doi.org/10.1016/j.compscitech.2022.109425
17. B. Liu, N. Vu-Bac, X. Zhuang, W. Lu, X. Fu, T. Rabczuk, AlDeMat: a web-based expert system platform for computationally expensive models in materials design. Adv. Eng. Softw. **176**, 103398 (2023). https://doi.org/10.1016/j.advengsoft.2022.103398
18. B. Liu, W. Lu, T. Olofsson, X. Zhuang, T. Rabczuk, Stochastic interpretable machine learning based multiscale modeling in thermal conductivity of polymeric graphene-enhanced composites. Compos. Struct. **327**, 117601 (2024). https://doi.org/10.1016/j.compstruct.2023.117601
19. B. Liu, N. Vu-Bac, T. Rabczuk, A stochastic multiscale method for the prediction of the thermal conductivity of polymer nanocomposites through hybrid machine learning algorithms. Compos. Struct. **273**, 114269 (2021). https://doi.org/10.1016/j.compstruct.2021.114269
20. B. Liu, N. Vu-Bac, X. Zhuang, X. Fu, T. Rabczuk, Stochastic full-range multiscale modeling of thermal conductivity of polymeric carbon nanotubes composites: A machine learning approach. Compos. Struct. **289**, 115393 (2022). https://doi.org/10.1016/j.compstruct.2022.115393
21. Y. Sonvane, D. Shah, K.N. Pathak, L. Saini, *Functional Materials and Applied Physics: FMAP-2021* (Materials Research Forum LLC, 2022)
22. A. Morales-Acevedo, Fundamentals of solar cell physics revisited: common pitfalls when reporting calculated and measured photocurrent density, open-circuit voltage, and efficiency of solar cells. Sol. Energy **262**, 111774 (2023). https://doi.org/10.1016/j.solener.2023.05.051
23. A.R. Forouhi, I. Bloomer, Optical dispersion relations for amorphous semiconductors and amorphous dielectrics. Phys. Rev. B **34**(10), 7018–7026 (1986). https://doi.org/10.1103/PhysRevB.34.7018
24. S. Adachi, *Optical Constants of Crystalline and Amorphous Semiconductors: Numerical Data and Graphical Information* (Springer US, 2013)
25. A.V. Mudryi, V.F. Gremenok, A.V. Karotki, V.B. Zalesski, M.V. Yakushev, F. Luckert et al., Structural and optical properties of thin films of Cu(In,Ga)Se2 semiconductor compounds. J. Appl. Spectrosc. **77**(3), 371–377 (2010). https://doi.org/10.1007/s10812-010-9341-5
26. M.A. Contreras, M.J. Romero, B. To, F. Hasoon, R. Noufi, S. Ward et al., Optimization of CBD cds process in high-efficiency Cu(In,Ga)Se2-based solar cells. Thin Solid Films. **403–404**, 204–211 (2002). https://doi.org/10.1016/S0040-6090(01)01538-3
27. Y. Chen, Review of ZnO Transparent Conducting Oxides for solar applications. IOP Conference Series: Materials Science and Engineering. 2018;423(1):012170. https://doi.org/10.1088/1757-899X/423/1/012170
28. A. Janotti, C.G. Van de Walle, Native point defects in ZnO. Phys. Rev. B. **76**(16), 165202 (2007). https://doi.org/10.1103/PhysRevB.76.165202
29. N. Refahati, A.V. Mudryi, V.D. Zhivulko, M.V. Yakushev, R. Martin, Influence of chemical composition heterogeneity on the spectral position of the fundamental absorption edge of Cu(In, Ga)Se2 solid solutions. J. Appl. Spectrosc. **81**(3), 404–410 (2014). https://doi.org/10.1007/s10812-014-9945-2
30. L. Fara, *Advanced Solar Cell Materials, Technology, Modeling, and Simulation* (Engineering Science Reference, 2012)
31. T. Minami, Chapter Five - transparent conductive oxides for transparent electrode applications, in *Semiconductors and Semimetals*, ed. by B.G. Svensson, S.J. Pearton, C. Jagadish (Elsevier, 2013), pp. 159–200
32. A. Jasenek, U. Rau, Defect generation in Cu(In,Ga)Se2 heterojunction solar cells by high-energy electron and proton irradiation. J. Appl. Phys. **90**(2), 650–658 (2001). https://doi.org/10.1063/1.1379348
33. Q. Cao, O. Gunawan, M. Copel, K.B. Reuter, S.J. Chey, V.R. Deline et al., Defects in chalcopyrite semiconductors: defects in Cu(In,Ga)Se2 chalcopyrite semiconductors: a comparative study of material properties, defect states, and photovoltaic performance (Adv. energy mater. 5/2011). Adv. Energy Mater. **1**(5), 844 (2011). https://doi.org/10.1002/aenm.201190024
34. C. Spindler, F. Babbe, M. Wolter, F. Ehre, K. Santhosh, P. Hilgert et al., Electronic defects in Cu(In,Ga)Se 2: towards a comprehensive model. Phys. Rev. Mater. (2019). https://doi.org/10.1103/PhysRevMaterials.3.090302
35. I.-H. Choi, C.-H. Choi, J.-W. Lee, Deep centers in a CuInGaSe2/CdS/ZnO:B solar cell. Physica status solidi (a) **209**(6), 1192–1197 (2012). https://doi.org/10.1002/pssa.201127596
36. F. Werner, F. Babbe, H. Elanzeery, S. Siebentritt, Can we see defects in capacitance measurements of thin-film solar cells? Prog. Photovoltaics Res. Appl. **27**(11), 1045–1058 (2019). https://doi.org/10.1002/pip.3196
37. T. Nakada, A. Kunioka, Direct evidence of cd diffusion into Cu(In, Ga)Se2 thin films during chemical-bath deposition process of cdS films. Appl. Phys. Lett. **74**(17), 2444–2446 (1999). https://doi.org/10.1063/1.123875
38. P. Virtanen, R. Gommers, T.E. Oliphant, Scipy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods **17**(3), 261–72 (2020). https://doi.org/10.1038/s41592-019-0686-2
39. J.S. Park, S. Kim, Z. Xie, A. Walsh, Point defect engineering in thin-film solar cells. Nat. Rev. Mater. **3**(7), 194–210 (2018). https://doi.org/10.1038/s41578-018-0026-7
40. S. Kim, J.-S. Park, A. Walsh, Identification of killer defects in kesterite thin-film solar cells. ACS Energy Lett. **3**(2), 496–500 (2018). https://doi.org/10.1021/acsenergylett.7b01313
41. J.D. Hunter, Matplotlib, a 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55
42. B. Liu, N. Vu-Bac, X. Zhuang, T. Rabczuk, Stochastic multiscale modeling of heat conductivity of polymeric clay nanocomposites. Mech. Mater. **142**, 103280 (2020). https://doi.org/10.1016/j.mechmat.2019.103280