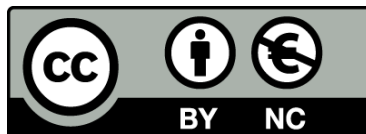




UNIVERSITAT DE
BARCELONA

Combinatorial and Machine Learning Techniques for Complex Thin Film Photovoltaics: Accelerated Research and Process Monitoring Methodologies

Enric Tomás Grau Luque



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial 4.0. Spain License.**

Doctoral thesis

Combinatorial and Machine
Learning Techniques for
Complex Thin Film
Photovoltaics: Accelerated
Research and Process
Monitoring Methodologies

Enric Tomás Grau Luque

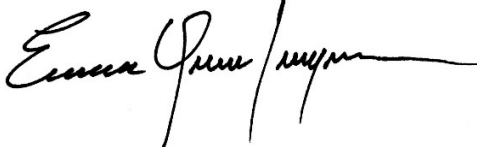


UNIVERSITAT DE
BARCELONA

Combinatorial and Machine Learning Techniques for Complex Thin Film Photovoltaics: Accelerated Research and Process Monitoring Methodologies

Memòria presentada per optar al grau de doctor per la
Universitat de Barcelona

Programa de doctorat en Enginyeria i Ciències
Aplicades

Autor: Enric Tomás Grau Luque 

Directors: Dr. Victor Izquierdo Roca i Dr. Maxim Guc

Tutor: Prof. Dr. Alejandro Pérez Rodríguez



UNIVERSITAT DE
BARCELONA

Aquesta pàgina s'ha deixat en blanc intencionadament

*“The algorithm is very opinionated:
the walls have to match the bedsheets.
It does not do what we want; it has its
own rules.”*

D.B.

ACKNOWLEDGMENTS

First, I'd like to thank to my Tutor Prof. Dr. A.P.R. and my Directors Dr. V.I.R. and Dr. M.G. Their guidance has been a crucial part of this work. For the opportunity they gave me and for their trust on my work, I thank you.

To all my friends and colleagues in IREC, across all groups, I'd like to show my deepest and sincere appreciation and gratefulness. To F.A., A.T., I.B., F.M., P.V., A.L., R.M., D.P., R.F., J.A., A.G., A.N, A.J., and many others who made my day-to-day better and easier. For all the conversation and hangouts, thank you all so much.

To my friends outside the lab that I have met during these 4 years, and those with whom I have shared this experience, I'd like to recognize their contribution and support. Their interest, insights, and unconditional support has helped me go through the worst times and enjoy the best ones. To T.V.W., A.M., F.F., A.Z., M.C., W.E., and everyone else that I might be forgetting, my deepest gratitude goes towards you all.

To my friends back home, who have never stop caring about my path. To I.L., C.C., G.I., M.D., A.S., J.T.A., P.B.H., M.K., B.P., P.S., D.M., J.D., P.P., L.B., J.T.G., and M.F. Thanks to each one of you for keeping up with my life and work, for visiting, and for receiving me.

To B.G., whose instinct and comments over my work has inspired my most creative ideas. For your admirable intelligence and love, I'm in forever in debt.

Finally, and most importantly, to my family. To my parents P.L.H., E.G.A. my sisters C.G.L, M.S.G.L. and M.P.G.L. Being far from home during all these years has not been easy, and without your constant support and interest in my work, I don't think this could have been possible. For your unconditional love, I'm eternally grateful.

TABLE OF CONTENTS

Pg.

ACKNOWLEDGMENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF ABBREVIATIONS AND ACRONYMS	xii
PREFACE AND PUBLICATIONS	xv
PREFACIO Y PUBLICACIONES	xxi
1. INTRODUCTION.....	27
1.1 Motivation.....	27
1.2 State of the art of PV	32
1.2.1 Overview of PV technology.....	32
1.2.2 Thin film photovoltaic technology.....	34
1.2.3 Basic working principle of PV devices.....	38
1.3 Combinatorial analysis for materials and devices	42
1.3.1 Samples for combinatorial analysis	42
1.3.2 Characterization techniques for combinatorial analysis	44
1.3.3 Data analysis for combinatorial approach.....	45
1.4 AI algorithms as support in the materials research.....	46
1.4.1 Introduction and basic principles of AI and ML.....	46
1.4.2 General use of ML in material science	52
1.4.3 Introduction to XAI.....	54
1.4.4 AI in energy, PV devices and materials research.....	56
1.5 Objective of the thesis	58
2 METHODOLOGY	61
2.1 Sample preparation	62
2.2 Characterization techniques.....	64
2.1.1 Raman spectroscopy.....	64
2.1.2 Spectroscopic Normal Reflectance	65
2.1.3 Optoelectronic characterization	65

2.1.4	X-ray fluorescence	65
2.3	Automated measurements.....	65
2.4	Data conditioning, fusion, and traceability.....	69
2.5	Data analysis.....	71
2.6	Methodology for spectrapepper library	74
2.7	Methodology for the pudu library	76
3	Publications	79
3.1	Publication 1	81
3.2	Publication 2	92
3.3	Publication 3	103
3.4	Publication 4	107
4.	FURTHER EXPLORATORY EXPERIMENTS.....	112
4.1	Introduction.....	112
4.2	Methodology.....	114
4.2.1	Sample.....	114
4.2.2	Sample characterization	115
4.2.3	Data processing	116
4.2.4	PC-LDA	117
4.2.5	1D and 2D CNN.....	117
4.2.6	Explainability	118
4.3	Results	118
4.3.1	Exploration of OOD properties.....	118
4.3.2	Sensitivity analysis of activations and model improvement.....	127
4.4	Conclusions of exploratory experiments	130
5.	Conclusions and Outlook	132
	REFERENCES	136
	ANNEXES.....	145
Annex A	146
Annex B	151
Annex C	155

Annex D	160
---------------	-----

LIST OF TABLES

	Pg.
Table 2-1: Example functions from the spectrapepper library categorized by their main purpose according to the experimental step. Full list of the functions and their explanation can be found in detail in the library's repository.....	75
Table 4-1: Experimental setup and parameters of the spectroscopic techniques. Total Aq. Times is the amount in minutes of the total time needed to perform the measurement considering all the 225 cells.	116
Table 4-2: Selected areas and their respective pixel ranges from the 14,640-long vector. In addition, the areas are labeled with the corresponding measurement technique and the axis values in the respective units of that measurement (Shift for Raman and Wavenumber for PL).....	125
Table 4-3: Changes in class classification probability p for each of the incorrectly classified samples (s) after modification of activation values of activations 128, 208 and 522. In bold are the samples that corrected their classification after modification, four in total (samples 12, 56, 58, and 62). Five other samples show statistical benefit but not enough to correct their prediction (samples 26, 140, 45, 122, and 107).	129

LIST OF FIGURES

	Pg.
Figure 1-1: A) Global surface temperature change from years 1 to 2000 and B) Past 170 years of global surface temperature change as observed compared to simulated cases of natural-only and humans & natural causes. Figure extracted from [8].	27
Figure 1-2: Observed impacts of climate change on human systems. The impacts are classified according to the confidence in the attribution to CC (color) and by increasing adverse impact (- sign), positive impact (+ sign), and adverse and positive impacts (\pm sign). Figure extracted from [1].	29
Figure 1-3: Share of primary energy supply from 2010 to 2021 and projection to 2050 in order to accomplish established emission goals. Renewables include solar, wind, hydro, and biomass. Figure extracted from [24].	31
Figure 1-4: Best Research-Cell Efficiencies compiled by the National Renewable Energy Laboratory (NREL). Figure extracted from [43].	33
Figure 1-5: Generic structure of a TF device based on p-n heterojunction. This shows how this technology involves multiscale, multilayer, and multiprocess devices with over 20 critical parameters to control.	35
Figure 1-6: Schematic of the p-n junction.	39
Figure 1-7: Classic circuit model of a photovoltaic cell.	40
Figure 1-8: Current-voltage curve showing its main characteristics.	41
Figure 1-9: A) Diagram of discrete sample set with process temperature and time variations, B) diagram of continuous spread sample with 1 graded layer, C) picture of a discrete sample set and D) picture of a continuous spread graded sample.	43
Figure 1-10: Schematic example to illustrate how the combination of different techniques allows to obtain further insights compared to normal experimentation focused in single techniques. The more characterized a sample is, the more it is possible to visualize the different aspects of its nature.	44
Figure 1-11: Characterization techniques, such as compositional, optical, structural, and optoelectronic, must comply with requirements to be fit for CA. Figure extracted from [78].	45
Figure 1-12: General view of the relationship between AI, ML, and DL.	47
Figure 1-13: Linear regression is performed for V_{oc} and OVC relationship for high performance CIGS solar cells. Colors represent different process temperatures the solar cells were subject to. Figure extracted from [92].	48
Figure 1-14: a) Mean spectra for healthy blood plasma and b) mean spectra for unhealthy (tuberculosis) blood plasma and c) final 2-D dimensionality after PCA. Figure extracted from [95].	50

Figure 1-15: Representation of a three-layer NN with an input, hidden, and output layer. The input is a two-parameter variable, and the hidden layer contains two units. This is arguably one of the simplest forms of a NN.....	52
Figure 1-16: General proper workflow for ML applications in material science. Figure extracted from [100].	54
Figure 1-17: Leveraging AI can enhance human capabilities and expedite discovery within the scientific method. Scientific discovery necessitates the integration of various AI techniques beyond solely data-driven ML. By combining primary AI methods, such as learning, reasoning, and planning, with human-computer interaction, a comprehensive approach emerges. This approach facilitates the integration of multiple knowledge sources, including databases, theory, experiments, and human reasoning, as demonstrated through relevant examples. Figure extracted from [107].	58
Figure 2-1: General flow of the proposed methodology for accelerated research using CA and ML.....	62
Figure 2-2: Samples used for the AlO _x thickness evaluation experiment on A) PET/CIGS, B) Si, and C) PET substrates. The final diagram is an approximation of the measured points. The inner radius of the AlO _x deposition is 1.2 cm, meanwhile the outer radius is 7.6 cm. Extracted from [112].	63
Figure 2-3: Photo of the CZGSe kesterite sample used in the second article. Change of the color is directly related to gradient of [Zn]/[Ge] ratio. Extracted from [112].	64
Figure 2-4: LabVIEW block diagram for measurement synchronization of the v1 system. The system constantly checks the location of the probe and performs measurements when they match the defined points by the user.	67
Figure 2-5: Photo and scheme of A) first version of the system used in the article 1 and B) second version of the system used in article 2.	69
Figure 2-6: Example of a high dimensional spectrum combining Raman and PL spectra for a single measured point. A) shows the raw Raman measurement, B) the PL raw measurement and C) the fused vectors after processing.	71
Figure 2-7: Deconvolution of a Raman spectra from a CZTSe sample using a 325 nm excitation source. Extracted from [116].	72
Figure 2-8: Syntax structure that all functions in the library, that accept spectral data as input, follow. The parameters \mathbf{y} and \mathbf{x} are only for the functions dedicated to spectral processes, which is the main focus of the library. However, there are some functions that can work with any kind of data, but are useful to have in a spectroscopic analysis toolkit (i.e. for the calculation of Spearman and Pearson correlation coefficients).....	75
Figure 4-1: Modified workflow of the methodology to include 1D and 2D CNNs along with respective explainability techniques.	114
Figure 4-2: Compositional ratios of Cu/Sn and Zn/Sn (upper row) and V _{oc} , Efficiency, and J _{sc} optoelectronic for the combinatorial sample (lower row).....	115

Figure 4-3: Schematic of the transformation of the vector of length 14,640 to an image of 120x120 pixels. As $120 \times 120 = 14,400$, 240 pixels have to be deleted in order to be reshaped. In this case, the last 240 pixels were deleted as they offer little information and is the easiest way to accomplish this. Other approaches are possible, such as interpolating or deleting smaller sections across the spectra..... 117

Figure 4-4: Scores in a confusion matrix for training (A) and test sets (B). The new scores of the new model after SA for training (C) and test (D). 119

Figure 4-5: Comparison between the first PC-LDA (left) and the second training (right) for each of the classes (from class 1 to 4 from top to bottom) in terms of importance according to sensitivity analysis. This shows how the vector changes in length after cutting off some of the sections. The importance is the average of the 10 closest spectra to the center of each of the clusters. Removing these sections appears to enhance some of the more important features. ... 120

Figure 4-6: Confusion matrices for Training (A) and Test (B) data sets for the CNN, with averages of 84% and 80%, respectively. Despite the good scores, some overfitting is appreciated, but highly biased by the best performing class, where accuracy is just above 71% with about 29% misclassified as second to first. For the 2D CNN, slightly lower scores are shown, with 0.82 and 0.76 for training and validation. 121

Figure 4-7: Average Grad-CAM visualization of the closest 10 spectra to the center of the cluster from the top performing classification group of $391 < V_{OC}$. From top to bottom, is the first convolutional layer, the second, and third convolutional layer from the CNN. Importance is normalized to 1 for each case..... 122

Figure 4-8: GradCAM results for the 3 convolutional layers (left to right) and the 4 classification groups (top to bottom) 122

Figure 4-9: D1 v D2 plot of the final PC-LDA model (left), MVNLR as $f(D1, D2)$ against V_{OC} (center) and the obtained equation mapped along with the scatter plot of D1 v D2 color graded with the V_{OC} (right)..... 123

Figure 4-10: GradCAM heatmaps for the average of the 10 closest spectra to the center of each of the clusters according to PC-LDA for each of the classification groups (top to bottom) for both convolutional layers (right and left)..... 124

Figure 4-11: Individual R2 for each of the areas when performing regression against V_{OC} (left), Pearson (center left) and Spearman (center right) correlation matrices for all the selected areas, and area 4 (a4) versus area 10 (a10) scatter plot graded with V_{OC} (right)..... 126

Figure 4-12: Regression (top row) and prediction mapping (bottom) for multi linear (left), polynomial quadratic (center) and RBFN (right)..... 127

Figure 4-13: Importance for spectroscopic features according to the change in inner-class probability change and next best-performing probability change. 128

Figure 4-14: Reactivation values for units in the last convolutional layer for correct classifications (Top in green) and incorrect classifications (bottom in brown). Arrows indicate

the units deactivated in the new model and their color indicate the same unit as in red for unit 128, yellow for unit 208, and blue for unit 522. 129

Figure 4-15: Overall confusion matrices showing the scores of A) the original CNN, B) the modified CNN after analysis of activations, C) the incorrect classifications of the original CNN and D) the incorrect classifications of the improved CNN..... 130

LIST OF ABBREVIATIONS AND ACRONYMS

Aluminum Oxide	AlO _x
Artificial Intelligence	AI
Artificial Neural Networks	ANN, NN
Bifacial Photovoltaics	BifPV
Big Data	BD
Building Integrated Photovoltaics	BIPV
Cadmium Sulfide	CdS
Cadmium Telluride	CdTe
Cascaded PCA LDA	PC-LDA
Chemical Vapor Deposition	CVD
Climate Change	CC
Combinatorial Analysis	CA
Convolutional Neural Network	CNN
Copper Indium Gallium Selenide (Cu(In,Ga)Se ₂)	CIGS
Copper Zinc Tin Selenide Kesterites (Cu ₂ ZnSnSe ₄)	CZTSe
Copper Zinc Tin Sulfide kesterites	CZTS
Crystalline Silicon	c-Si
Current-voltage	IV
Deep Learning	DL
Density Functional Theory	DFT
European Union	EU
Explainable Artificial Intelligence	XAI
Fill Factor	FF
Generative Adversarial Network	GAN
Greenhouse Gases	GHGs
Heterojunction with Intrinsic Thin layer	HJT
High-Throughput Experiments	HTE
Institut de Recerca en Energia de Catalunya	IREC

International Energy Agency	IEA
International Panel for Climate Change	IPCC
Leaky Rectified Linear Unit	LeakyReLU
Levelized Cost of Energy	LCOE
Linear Discriminant Analysis	LDA
Local Interpretable Model Agnostic Explanations	LIME
Linear Regression	LR
Machine Learning	ML
Multivariate Curve Resolution	MCR
Multivariate Non-Linear Regression	MVNLR
Neural Networks	NN, ANN
Open Circuit Voltage	VOC
Pasivated Emitter and Rear Cell	PERC
Photoluminescence	PL
Photovoltaics	PV
Physical Vapor Deposition	PVD
Polyethylene Terephthalate	PET
Polycrystalline Silicon	p-Si
Polynomial Regression	PR
Principal Component Analysis	PCA
Quantum Dots	QD
Radial Basis Function	RBF
Random Forest	RF
Rectified Linear Unit	ReLU
Sensitivity Analysis	SA
Short Circuit Current	ISC
Short Circuit Current Density	JSC
Solar Energy Materials and Systems	SEMS
Support Vector Machines	SVMs
Thin films	TF

User Interface	UI
Vehicle Integrated Photovoltaics	VIPV
X-ray diffraction	XRD
X-ray Fluorescence	XRF
Zinc Oxide	ZnO

PREFACE AND PUBLICATIONS

The research presented in this thesis was conducted at the Catalonia Institute for Energy Research (IREC) in the Solar Energy Materials and Systems (SEMS) research group, located in Barcelona, Spain, between the years 2019 and 2023. The work was carried out as part of the research line focused on the development and implementation of innovative, high-throughput research methods for the study of photovoltaic materials and devices, with an emphasis on industrial application of the results. The primary objective of this thesis was to develop and use Artificial Intelligence based on Machine Learning algorithms combined with Combinatorial Analysis to provide new tools for the accelerated research of chalcogenide-based technologies, suitable for thin film photovoltaics and other emerging technologies applications. Specifically, the research is focused on the development and implementation of semi-autonomous optoelectronic and spectroscopic data analysis with Artificial Intelligence methodologies to accelerate the investigation of fundamental physicochemical properties of photovoltaic materials and devices, and making tools available for scientific community and photovoltaic industry to use Artificial Intelligence in their research, workflows and production. The main idea behind the development of such a tools is based on their possibility to provide deeper understanding of the complex behavior of thin film photovoltaic devices, and information about the impact of fabrication parameters on the device performance and efficiency loss/failure mechanisms in a faster way, as well as to reduce the lab-to-market times.

During the course of the doctoral thesis, the performed work and the results obtained have allowed the publication of 4 articles in peer-reviewed journals, 2 of them in high impact factor Q1 journals and 2 articles as open-source and open-access software in a peer-review journal. These 4 articles are a direct result of the work carried out and in alignment with the objectives of this thesis. Additionally, 2 other articles were published with the participation of Enric Tomás Grau Luque (ETGL) as coauthor. The full list of the publications in question, and the role of ETGL, is as follows:

- **Grau-Luque, E.**, Guc, M., Becerril-Romero, I., Izquierdo-Roca, V., Pérez-Rodríguez, A., Bolt, P., Van den Bruele, F., & Ruhle, U. (2022). Thickness evaluation of AlO_x barrier layers for encapsulation of flexible PV modules in industrial environments by normal reflectance and machine learning. *Progress in Photovoltaics: Research and Applications*, 30(3), 229–239. <https://doi.org/10.1002/PIP.3478>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft.

- **Grau-Luque, E.**, Anefnaf, I., Benhaddou, N., Fonoll-Rubio, R., Becerril-Romero, I., Aazou, S., Saucedo, E., Sekkat, Z., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2021). Combinatorial and machine learning approaches for the analysis of $\text{Cu}_2\text{ZnGeSe}_4$: influence of the off-stoichiometry on defect formation and solar cell performance. *Journal of Materials Chemistry A*, 9 (16), 10466–10476. <https://doi.org/10.1039/d1ta01299a>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Data curation, Formal Analysis, Software, Visualization, Writing – original draft.

- **Grau-Luque, E.**, Atlan, F., Becerril-Romero, I., Perez-Rodriguez, A., Guc, M., & Izquierdo-Roca, V. spectrapepper: A Python toolbox for advanced analysis of spectroscopic data for materials and devices. *J. Open Source Software*. 6, 3781 (2021). <https://doi.org/10.21105/joss.03781>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Conceptualization, Data curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft.

- **Grau-Luque, E.**, Becerril-Romero, I., Perez-Rodriguez, A., Guc, M., & Izquierdo-Roca, V. pudu: A Python library for agnostic feature selection and explainability of Machine Learning spectroscopic problems. *Journal of Open Source Software*, 8(92), 5873, <https://doi.org/10.21105/joss.05873>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Conceptualization, Data curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft.

Publications from collaborations:

- Fonoll-Rubio, R., Paetel, S., **Grau-Luque, E.**, Becerril-Romero, I., Mayer, R., Pérez-Rodríguez, A., Guc, M., & Izquierdo-Roca, V. (2022). Insights into the Effects of RbF-Post-Deposition Treatments on the Absorber Surface of High Efficiency $\text{Cu}(\text{In,Ga})\text{Se}_2$ Solar Cells and Development of Analytical and Machine Learning Process Monitoring Methodologies Based on Combinatorial Analysis. *Advanced Energy Materials*, 2103163.

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: supporting data curation, supporting formal analysis, supporting methodology, Software.

- Fonoll-Rubio, R., Becerril-Romero, I., Vidal-Fuentes, P., **Grau-Luque, E.**, Atlan, F., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2022). Combinatorial Analysis Methodologies for Accelerated Research: The Case of Chalcogenide Thin-Film Photovoltaic Technologies. Solar RRL, 2200235.

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: supporting data curation, supporting formal analysis, supporting methodology, Software.

The present doctorate thesis document is structured in 5 chapters including conclusions, and 2 additional sections for references and annexes. These chapters and sections are briefly described below:

Chapter 1 – Introduction: This chapter provides an overview of the current state of the world in terms Climate Change and energy production/consumption, including the role of photovoltaic technologies within this context. An overview of some photovoltaic technologies is presented, including their degree of maturity, advantages, disadvantages, and their potential for further development. The chapter provides an examination of thin film technology, including its unique properties and its alignment with the energy decarbonization roadmap. The chapter is continued with detailed description of the Combinatorial Analysis concepts and its possibility to be applied for thin film photovoltaic technologies. Furthermore, the chapter delves into the utilization of Artificial Intelligence and Machine Learning in photovoltaic materials and devices research, discussing the current state of the field, as well as its future prospects. The aim of this chapter is to provide a comprehensive understanding of the current state of the art of photovoltaic technology with focus on thin film photovoltaics, its potential for future growth, and the role of Combinatorial Analysis, Artificial Intelligence and Machine Learning in advancing the field. Finally, this chapter ends by identifying the gaps and needs of the field and explaining how this align with the objectives of this work.

Chapter 2 – Methodology: This chapter provides a comprehensive overview of the methodology proposed and used in the current work. These include a detailed workflow for the methodology with description of all necessary steps for its implementation, details about the combinatorial samples preparation, and description of the characterization techniques together with the equipment and apparatus used. The chapter contains information about the work implemented for the automation of the measurements and data treatment procedures. All the specific steps performed in the present work related to data conditioning, fusion, and traceability are described together with details about the data analysis approaches and the specific algorithms and programming libraries used. Finally, a description for each of the libraries' structure and working principles is shown to better highlight the methodology behind their development.

Chapter 3 – Summary of results through publications: This chapter begins with a contextualization of the work carried out, providing a general introduction that highlights the significance and relevance of the research conducted. This introduction provides cohesion and continuity to the chapter and serves as a valuable reference point for readers to understand the broader implications of the research. The chapter then presents the four scientific articles published in peer-review journals, which demonstrate the innovative contributions and advancements made in the field. Each article is presented in a clear and concise manner, highlighting the key findings and implications of the research.

The first article, *Thickness Evaluation of AlO_x Barrier Layers for Encapsulation of Flexible PV Modules in Industrial Environments using Normal Reflectance and Machine Learning*, describes and demonstrates a novel characterization methodology based on normal reflectance measurements and Machine Learning algorithms. This methodology enables precise, low-cost, and scalable assessment of the thickness of AlO_x nanometric layers, which are added to flexible photovoltaic devices based on such materials as $\text{Cu}(\text{In,Ga})\text{Se}_2$ (CIGS) and perovskites, to improve solar modules protection through their low water vapor transmission rate. This solution is particularly suitable for roll-to-roll industrial production lines. However, precise control of the thickness of the AlO_x layers is crucial to ensure an effective water barrier performance. Current methods for evaluating such nanometric layers are costly and complex to implement in industrial environments. The proposed methodology is applied to determine the thickness of AlO_x nanolayers deposited on three different substrates relevant for the photovoltaic industry: monocrystalline Si, $\text{Cu}(\text{In,Ga})\text{Se}_2$ flexible modules, and polyethylene terephthalate (PET) flexible encapsulation foil. The methodology demonstrates sensitivity of <10 nm and acquisition times of ≤ 100 ms, making it compatible with industrial monitoring applications. Additionally, a specific design for in-line integration of a normal reflectance system into a roll-to-roll production line for thickness control of nanometric layers is proposed.

The second article, *Combinatorial and machine learning approaches for the analysis of $\text{Cu}_2\text{ZnGeSe}_4$: influence of the off-stoichiometry on defect formation and solar cell performance*, presents a combinatorial approach for the analysis of CZGS ($\text{Cu}_2\text{ZnGeSe}_4$) solar cells. These solar cells are complex systems where changes in one parameter can result in changes in the entire system and, as a consequence, in the overall performance of the devices. In order to overcome the limitations of this promising earth-abundant photovoltaic technology, analyses that take into account this complexity are necessary. The article describes the analysis of a compositional graded sample containing almost 200 solar cells with different Zn/Ge compositions using X-ray fluorescence and Raman spectroscopy. The results are correlated with the optoelectronic parameters of the different cells, providing a deep understanding of the stoichiometric limits and point defects formation in the CZGS compound and the influence of these parameters on the performance of the devices. Intertwined connections between the compositional, vibrational, and

optoelectronic properties of the cells are revealed using a complex analytical approach. The study is further extended by using a Machine Learning algorithm, which confirms the correlation between the properties of the CZGS compound and the optoelectronic parameters, and also allows proposing a methodology for device performance prediction that is compatible with both research and industrial process monitoring environments. This work not only provides valuable insights for understanding and further developing the CZGS photovoltaic technology, but also gives a practical example of the potential of Combinatorial Analysis and Machine Learning for the study of complex systems in materials research.

The third article, *spectrapepper: A Python toolbox for advanced analysis of spectroscopic data for materials and devices*, introduces spectrapepper, a Python library designed to streamline the analysis of complex high-tech materials and devices, such as multi-layered thin film solar cells, using spectroscopy. It integrates several functions for the acquisition, processing, analysis and visualization of spectroscopic data. Spectrapepper enables the design of automated spectroscopy systems and big data analysis, significantly reducing development times for new materials. It has comprehensive documentation and examples are available online, facilitating its access and adoption in the material science community.

The fourth, *pudu: A Python library for agnostic feature selection and explainability of Machine Learning spectroscopic problems*, introduces pudu, a Python library designed to enhance the interpretability of Machine Learning models in spectroscopic data analysis, widely applicable in fields like PV, aiming to increase the transparency and scientific impact of Machine Learning results. It offers four new methods: Importance, Speed, Synergy, and Re-activations, each quantifying the impact of spectral feature changes on model predictions. Suitable for both 1D and 2D classification and regression problems, pudu provides flexibility and localized explanations. It integrates with the main platforms for application of Machine Learning algorithms such as scikit-learn, keras, and pytorch.

Chapter 4 – Further Exploratory Experiments: In this chapter, side experiments that were not part of any publication are presented. In particular, experiments that follow the presented papers are explained and discussed, being a natural follow up and extension of the methodology presented in Chapter 2 and used in Chapter 3. This section serves as the next steps to be followed in order to further advance in the development of thin film photovoltaic devices with the aid of Machine Learning. Multivariate Non-Linear Regressions, Radial Basis Function Networks, Convolutional Neural Networks, and consequent explanation attempts are discussed in this section.

Chapter 5 – Conclusions: In this final chapter of the thesis, a comprehensive summary of the research conducted is provided, focusing on the key findings and conclusions drawn from the study. The chapter begins with an overview of the research objectives and methodology,

highlighting the major contributions made by the research. The main conclusions of the work are then presented, with an evaluation of the extent to which the research objectives were achieved. The significance and implications of the findings are also discussed, placing them in the context of the existing literature and highlighting their potential impact on the field.

References: This section compiles all the references use in this work.

Annexes: This section contains additional information that was not considered to be vital to incorporate in the main body of this work. The annexes mainly present screenshot of the software developed during the doctoral program with the involvement of ETGL and used for different applications.

PREFACIO Y PUBLICACIONES

La investigación presentada en esta tesis se llevó a cabo en el Instituto de Investigación en Energía de Cataluña (IREC) en el grupo de investigación de Materiales y Sistemas de Energía Solar (SEMS), ubicado en Barcelona, España, entre los años 2019 y 2023. El trabajo se realizó como parte de la línea de investigación centrada en el desarrollo e implementación de métodos innovadores de alto rendimiento para el estudio de materiales y dispositivos fotovoltaicos, con énfasis en la aplicación industrial. El objetivo principal de esta tesis fue desarrollar y utilizar Inteligencia Artificial basada en algoritmos de Aprendizaje de Máquinas combinados con Análisis Combinatorio para proporcionar nuevas herramientas para la investigación acelerada de tecnologías basadas en calcógenos, adecuadas para tecnología fotovoltaica de capa fina y otras aplicaciones de tecnologías emergentes. Específicamente, el trabajo se enfoca en el desarrollo e implementación de análisis semi-autónomo de datos optoelectrónicos y espectroscópicos con metodologías de Inteligencia Artificial para acelerar la investigación de propiedades fisicoquímicas fundamentales de materiales y dispositivos fotovoltaicos, y poner a disposición herramientas para la comunidad científica y la industria fotovoltaicos para utilizar la Inteligencia Artificial en sus investigaciones, flujos de trabajo y producción. La idea principal detrás del desarrollo de una herramienta como esta se basa en la posibilidad de proporcionar una comprensión más profunda del comportamiento complejo de los dispositivos fotovoltaicos de capa fina, e información sobre el impacto de los parámetros de fabricación en el rendimiento del dispositivo y los mecanismos de pérdidas de eficiencia de manera más rápida, además de reducir los tiempos de laboratorio al mercado.

Durante el transcurso de la tesis doctoral, el trabajo realizado y los resultados obtenidos permitieron la publicación de 4 artículos en revistas *peer-review*, 2 de ellos en revistas de alto factor de impacto Q1 y 2 artículos como software de código y acceso abierto en revistas *peer-review*. Estos 4 artículos son resultado directo del trabajo realizado y en alineación con los objetivos de esta tesis. Además, se publicaron 2 otros artículos con la participación de Enric Tomás Grau Luque (ETGL) como coautor. La lista completa de las publicaciones en cuestión y el rol de ETGL es la siguiente:

- **Grau-Luque, E.**, Guc, M., Becerril-Romero, I., Izquierdo-Roca, V., Pérez-Rodríguez, A., Bolt, P., Van den Bruele, F., & Ruhle, U. (2022). Thickness evaluation of AlO_x barrier layers for encapsulation of flexible PV modules in industrial environments by normal reflectance and machine learning. *Progress in Photovoltaics: Research and Applications*, 30(3), 229–239. <https://doi.org/10.1002/PIP.3478>

Usando la Taxonomía de Roles de Contribuyentes CRediT, el trabajo de ETGL se puede describir como: Conceptualización, Curación de Datos, Análisis Formal, Investigación, Metodología, Software, Visualización, Escritura - Borrador Original.

- **Grau-Luque, E.**, Anefnaf, I., Benhaddou, N., Fonoll-Rubio, R., Becerril-Romero, I., Aazou, S., Saucedo, E., Sekkat, Z., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2021). Combinatorial and machine learning approaches for the analysis of Cu₂ZnGeSe₄: influence of the off-stoichiometry on defect formation and solar cell performance. *Journal of Materials Chemistry A*, 9 (16), 10466–10476. <https://doi.org/10.1039/d1ta01299a>

Usando la Taxonomía de Roles de Contribuyentes CRediT, el trabajo de ETGL se puede describir como: Curación de Datos, Análisis Formal, Software, Visualización, Escritura - Borrador Original.

- **Grau-Luque, E.**, Atlan, F., Becerril-Romero, I., Perez-Rodriguez, A., Guc, M., & Izquierdo-Roca, V. spectrapepper: A Python toolbox for advanced analysis of spectroscopic data for materials and devices. *J. Open Source Software*. 6, 3781 (2021). <https://doi.org/10.21105/joss.03781>

Usando la Taxonomía de Roles de Contribuyentes CRediT, el trabajo de ETGL se puede describir como: Conceptualización, Curación de Datos, Análisis Formal, Metodología, Software, Visualización, Escritura - Borrador Original.

- **Grau-Luque, E.**, Becerril-Romero, I., Perez-Rodriguez, A., Guc, M., & Izquierdo-Roca, V. pudu: A Python library for agnostic feature selection and explainability of Machine Learning spectroscopic problems. *Journal of Open Source Software*, 8(92), 5873, <https://doi.org/10.21105/joss.05873>

Usando la Taxonomía de Roles de Contribuyentes CRediT, el trabajo de ETGL se puede describir como: Conceptualización, Curación de Datos, Análisis Formal, Metodología, Software, Visualización, Escritura - Borrador Original.

Publicaciones de colaboraciones:

- Fonoll-Rubio, R., Paetel, S., **Grau-Luque, E.**, Becerril-Romero, I., Mayer, R., Pérez-Rodríguez, A., Guc, M., & Izquierdo-Roca, V. (2022). Insights into the Effects of RbF-Post-Deposition Treatments on the Absorber Surface of High Efficiency Cu(In,Ga)Se₂ Solar Cells and Development of Analytical and Machine Learning Process Monitoring Methodologies Based on Combinatorial Analysis. *Advanced Energy Materials*, 2103163.

Usando la Taxonomía de Roles de Contribuyentes CRediT, el trabajo de ETGL se puede describir como: curación de datos de apoyo, análisis formal de apoyo, metodología de apoyo, software de apoyo.

- Fonoll-Rubio, R., Becerril-Romero, I., Vidal-Fuentes, P., **Grau-Luque, E.**, Atlan, F., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2022). Combinatorial Analysis Methodologies for Accelerated Research: The Case of Chalcogenide Thin-Film Photovoltaic Technologies. Solar RRL, 2200235.

Usando la Taxonomía de Roles de Contribuyentes CRediT, el trabajo de ETGL se puede describir como: curación de datos de apoyo, análisis formal de apoyo, metodología de apoyo, software de apoyo.

El presente documento de tesis doctoral está estructurado en 5 capítulos que incluyen conclusiones, y 2 secciones adicionales para referencias y anexos. Estos capítulos y secciones se describen brevemente a continuación:

Capítulo 1 - Introducción: Este capítulo proporciona una visión general del estado actual del mundo en términos de Cambio Climático y producción y consumo de energía, incluyendo el papel de las tecnologías fotovoltaicas dentro de este contexto. Se presenta una visión general de algunas tecnologías PV, incluyendo su grado de madurez, ventajas, desventajas y su potencial para su desarrollo en el futuro. El capítulo proporciona una descripción de la tecnología de capa fina, incluyendo sus propiedades únicas y su alineación con la hoja de ruta de descarbonización energética. El capítulo continúa con una descripción detallada de los conceptos de Análisis Combinatorio y su posibilidad de ser aplicado a tecnologías fotovoltaica de capa fina. Además, el capítulo profundiza en la utilización de Inteligencia Artificial y Aprendizaje de Máquinas en la investigación de materiales y dispositivos PV, discutiendo el estado actual del campo, así como sus perspectivas futuras. El objetivo de este capítulo es proporcionar una comprensión integral del estado actual de la tecnología fotovoltaica con un enfoque en materiales de capa fina, su potencial para un crecimiento futuro y el papel de Análisis Combinatorio, Inteligencia Artificial y Aprendizaje de Máquinas en el avance del campo. Finalmente, este capítulo concluye identificando las brechas y necesidades del campo y explicando cómo esto se alinea con los objetivos de este trabajo.

Capítulo 2 - Metodología: Este capítulo proporciona una visión general de la metodología propuesta y utilizada en el trabajo desarrollado. Esto incluye un flujo de trabajo detallado para la metodología con una descripción de todos los pasos necesarios para su implementación, detalles sobre la preparación de muestras combinatorias y descripción de las técnicas de caracterización junto con los equipos y aparatos utilizados. El capítulo contiene información sobre el trabajo

implementado para la automatización de las técnicas de caracterización y los procedimientos preliminares de tratamiento de datos. Se describen todos los pasos específicos realizados relacionados con el acondicionamiento, fusión y trazabilidad de datos, junto con detalles sobre los enfoques de análisis de datos y los algoritmos y librerías de programación específicos utilizados. Finalmente, se incluye una descripción de la estructura y principios de funcionamiento de cada librería, para así destacar de mejor manera la metodología de desarrollo de cada una.

Capítulo 3 - Resumen de resultados a través de publicaciones: Este capítulo comienza con una contextualización del trabajo realizado, proporcionando una introducción general que destaca la importancia y relevancia de la investigación realizada. Esta introducción proporciona cohesión y continuidad al capítulo y sirve como un punto de referencia valioso para que los lectores comprendan las implicaciones más amplias de la investigación. Luego, el capítulo presenta los cuatro artículos científicos publicados, que demuestran las contribuciones innovadoras y los avances realizados en el campo. Cada artículo se presenta de manera clara y concisa, destacando los hallazgos clave y las implicaciones del trabajo.

El primer artículo, *Thickness evaluation of AlO_x barrier layers for encapsulation of flexible PV modules in industrial environments by normal reflectance and machine learning*, describe y demuestra una novedosa metodología de caracterización basada en medidas de reflectancia normal y algoritmos de Aprendizaje de Máquinas. Esta metodología permite la evaluación precisa, económica y escalable del espesor de capas nanométricas de AlO_x , que se agregan a dispositivos fotovoltaicos flexibles basados en materiales como $Cu(In,Ga)Se_2$ y perovskitas, para mejorar la protección de los módulos solares a través de su baja tasa de transmisión de vapor de agua. Esta solución es especialmente adecuada para líneas de producción industriales de *roll-to-roll*. Sin embargo, el control preciso del espesor de las capas de AlO_x es crucial para garantizar un rendimiento efectivo como barrera contra el agua. Los métodos actuales para evaluar dichas capas nanométricas son costosos y complejos de implementar en entornos industriales. La metodología propuesta se aplica para determinar el espesor de capas nanométricas de AlO_x depositadas en tres sustratos diferentes relevantes para la industria PV: silicio monocristalino, módulos flexibles de $Cu(In,Ga)Se_2$ y lámina de encapsulación flexible de tereftalato de polietileno (PET). La metodología demuestra una sensibilidad de <10 nm y tiempos de adquisición de ≤ 100 ms, lo que la hace compatible con aplicaciones de monitoreo industrial. Además, se propone un diseño específico para la integración en línea de un sistema de reflectancia normal en una línea de producción de *roll-to-roll* para el control del espesor de capas nanométricas.

El segundo artículo, *Combinatorial and machine learning approaches for the analysis of $Cu_2ZnGeSe_4$: influence of the off-stoichiometry on defect formation and solar cell performance*, presenta un enfoque combinatorio para el análisis del material CZGS ($Cu_2ZnGeSe_4$). Las celdas solares, basadas en compuestos quaternarios kesterita como CZGS, son sistemas complejos en los

que cambios en un parámetro pueden resultar en cambios en todo el sistema y, como consecuencia, en el rendimiento general de los dispositivos. Para superar las limitaciones de esta prometedora tecnología fotovoltaica abundante en elementos de tierras raras, son necesarios análisis que tengan en cuenta esta complejidad. El artículo describe el análisis de una muestra que contiene casi 200 celdas solares con diferentes composiciones de Zn/Ge utilizando fluorescencia de rayos X y espectroscopía Raman. Los resultados se correlacionan con los parámetros optoelectrónicos de las diferentes celdas, proporcionando una comprensión profunda de los límites estequiométricos y la formación de defectos puntuales en el compuesto CZGS y la influencia de estos parámetros en el rendimiento de los dispositivos. Se revelan conexiones entrelazadas entre las propiedades composicionales, vibracionales y optoelectrónicas de las celdas mediante un enfoque analítico complejo. El estudio se amplía aún más mediante el uso de un algoritmo de Aprendizaje de Máquinas, que confirma la correlación entre las propiedades del compuesto CZGS y los parámetros optoelectrónicos, y también permite proponer una metodología para la predicción del rendimiento del dispositivo compatible tanto con la investigación como con los entornos de monitoreo de procesos industriales. Este trabajo no solo proporciona información valiosa para comprender y desarrollar aún más la tecnología fotovoltaica CZGS, sino que también da un ejemplo práctico del potencial de Análisis Combinatorio y Aprendizaje de Máquinas para el estudio de sistemas complejos en la investigación de materiales.

El tercer artículo, *spectrapepper: A Python toolbox for advanced analysis of spectroscopic data for materials and devices* presenta spectrapepper, una librería para Python diseñada para agilizar el análisis de materiales y dispositivos, como celdas solares de capa fina, utilizando espectroscopía. Integra varias funciones para la adquisición, procesamiento, análisis y visualización de datos espectroscópicos. Spectrapepper permite el diseño de sistemas de espectroscopía automatizados y el análisis de grandes datos, reduciendo significativamente los tiempos de desarrollo de nuevos materiales. Tiene documentación exhaustiva y ejemplos disponibles en línea, lo que facilita su adopción en la comunidad científica de ciencia de materiales.

El cuarto artículo, *pudu: A Python library for agnostic feature selection and explainability of Machine Learning spectroscopic problems*, presenta una librería para Python diseñada para mejorar la interpretación de modelos de Aprendizaje de Máquinas en el análisis de datos espectroscópicos. Tiene como objetivo aumentar la transparencia e impacto científico de los resultados de Aprendizaje de Máquinas. Ofrece cuatro nuevos métodos: Importancia, Velocidad, Sinergia y Re-activaciones, cada uno cuantificando el impacto de los cambios en las características espectrales en las predicciones del modelo. Adecuado tanto para problemas de clasificación y regresión 1D como 2D, pudu proporciona flexibilidad y explicaciones localizadas. Se integra con las principales plataformas para la aplicación de algoritmos de Aprendizaje de Máquinas como scikit-learn, keras y pytorch.

Capítulo 4 - Experimentos Exploratorios Adicionales: En este capítulo se presentan experimentos secundarios que no formaron parte de ninguna publicación. En particular, se explican y discuten experimentos que siguen a los artículos presentados, siendo una continuación natural y extensión de la metodología presentada en el Capítulo 2 y utilizada en el Capítulo 3. Esta sección muestra los siguientes pasos a seguir para avanzar aún más en el desarrollo de dispositivos fotovoltaicos de capa fina con la ayuda de Inteligencia Artificial. En esta sección se discuten Regresiones no Lineales Multivariadas, Redes de Función de Base Radial, Redes Neuronales Convolucionales y los intentos de explicación consecuentes.

Capítulo 5 - Conclusiones: En este capítulo final, se proporciona un resumen integral de la investigación realizada, centrándose en los hallazgos clave y conclusiones extraídas del trabajo. El capítulo comienza con una visión general de los objetivos de investigación y la metodología, destacando las principales contribuciones realizadas por la investigación. Luego se presentan las principales conclusiones del trabajo, con una evaluación del grado en que se lograron los objetivos de investigación. También se discuten la importancia y las implicaciones de los hallazgos, situándolos en el contexto de la literatura existente y resaltando su impacto potencial en el campo.

Referencias: Esta sección recopila todas las referencias utilizadas en este trabajo.

Anexos: Esta sección contiene información adicional que no se consideró esencial para incorporar en el cuerpo principal de este trabajo. Los anexos principalmente presentan capturas de pantalla del software desarrollado durante el programa de doctorado con la participación de ETGL y utilizado para diferentes aplicaciones.

1. INTRODUCTION

1.1 Motivation

Climate change (CC) has become a critical global issue with far-reaching consequences for the environment, human society, and the economy. It is primarily driven by the increasing concentration of greenhouse gases (GHGs), particularly CO₂, in the atmosphere due to human activities such as fossil fuel combustion, deforestation, and industrial processes [1]. The average global temperature has risen by approximately 1.2°C since the pre-industrial era. This is illustrated in Figure 1-1A where a pronounced spike in temperature is reported between 1850 to 2020. This rapid temperature increase has caused a wide range of changes in the climate system, including more frequent and severe extreme weather events, such as heatwaves, droughts, floods, and storms, as well as sea-level rise, ocean acidification, and alterations in ecosystems and biodiversity [2][3][4]. Climate models project that global temperatures could rise by 1.5°C to 4.8°C by the end of the 21st century, depending on future GHG emissions scenarios. These temperature increases will exacerbate the adverse impacts of CC, including water scarcity, reduced agricultural productivity, and increased risks to human health and well-being [5][6][7].

Changes in global surface temperature relative to 1850–1900

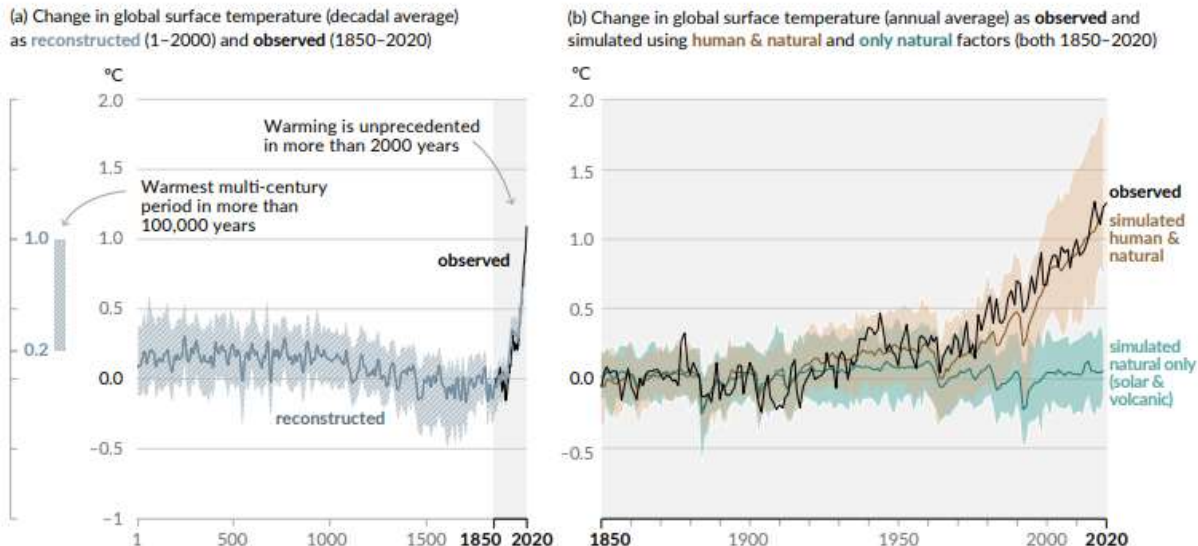


Figure 1-1: A) Global surface temperature change from years 1 to 2020 and B) Past 170 years of global surface temperature change as observed compared to simulated cases of natural-only and humans & natural causes. Figure extracted from [8].

The impacts of CC on the economy are significant and multifaceted, with potential consequences for productivity, infrastructure, and various economic sectors. CC can lead to substantial economic costs, particularly if global temperatures rise by more than 2°C above pre-industrial levels. These

costs can manifest as losses in productivity due to heat stress, reduced agricultural yields, and disruptions to supply chains [9]. Furthermore, the increased frequency and severity of extreme weather events can result in considerable infrastructure damage, necessitating costly repairs and replacements [10]. Economic sectors that are particularly vulnerable to the effects of CC include agriculture, fisheries, and tourism, which often rely heavily on climate-sensitive natural resources [7][11]. For instance, crop yields are expected to decline by 10-25% in some regions, with the most severe reductions occurring in developing countries, where food security is already a pressing concern [5]. Also, CC has been linked to declines in crop yields and increased risks of crop failure, threatening food security and the livelihoods of agricultural workers [5][6].

The effects of CC on ecosystems are profound and diverse, with wide-ranging consequences for species, habitats, and the vital roles they play. As global temperatures continue to rise, ecosystems are experiencing shifts in their distribution, composition, and function, often with cascading effects on biodiversity [4][12][8]. One of the most evident impacts of CC on ecosystems is the alteration of species' geographic ranges, as they move poleward or to higher elevations in search of more suitable habitats [4]. This can lead to the fragmentation and loss of habitat for various species, resulting in declines in their population sizes and increased risks of local or global extinctions [13]. Additionally, CC can exacerbate existing threats to ecosystems, such as habitat loss due to land-use change, pollution, and the spread of invasive species [14]. Changes in temperature and precipitation patterns, as well as the increased frequency of extreme weather events, can also disrupt the timing of key ecological processes, such as flowering, breeding, and migration [3]. These disruptions can cause mismatches between species and their resources, leading to declines in reproductive success and population viability [15]. Furthermore, CC can alter the structure and functioning of ecosystems, affecting processes like nutrient cycling, primary productivity, and decomposition, which in turn can influence ecosystem resilience and the provision of essential services [16]. The degree of these effects varies between industries, as shown in Figure 1-2, which also shows the confidence in human contribution to CC. For instance, crop production in Africa have seen a negative impact due to CC, with high confidence of human contribution.

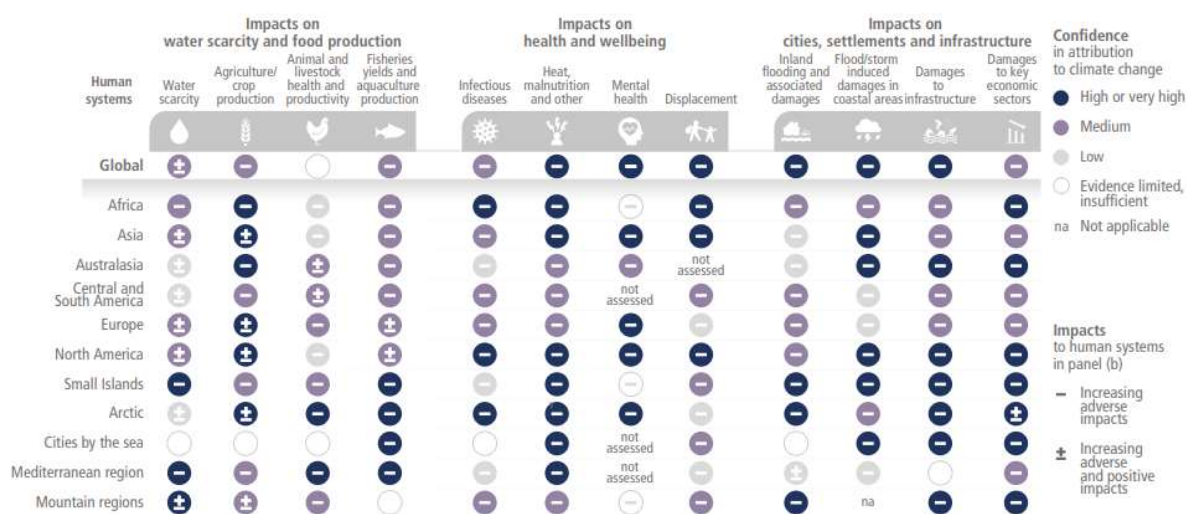


Figure 1-2: Observed impacts of climate change on human systems. The impacts are classified according to the confidence in the attribution to CC (color) and by increasing adverse impact (- sign), positive impact (+ sign), and adverse and positive impacts (\pm sign). Figure extracted from [1].

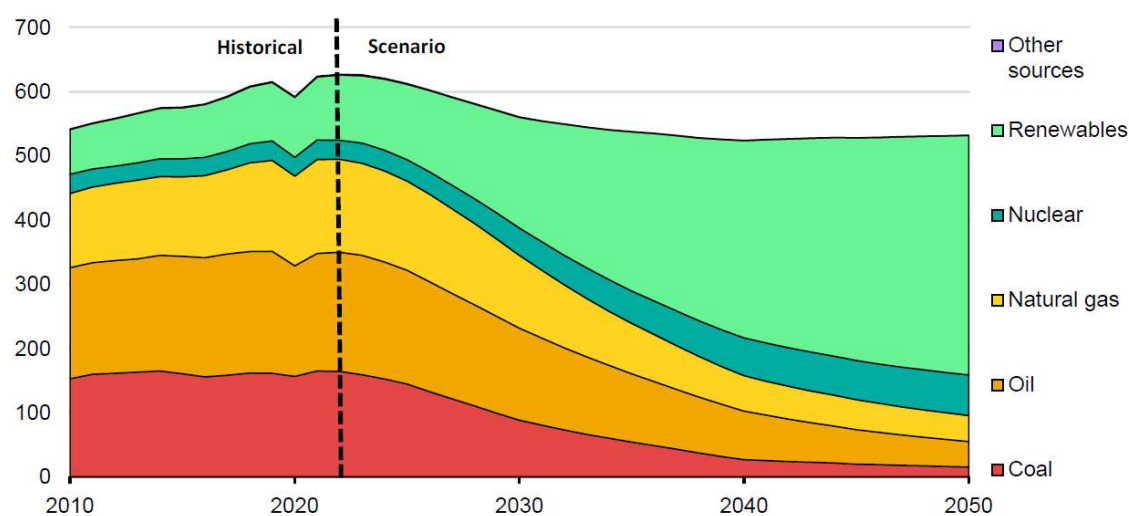
Marine ecosystems are particularly vulnerable to the impacts of CC, as rising ocean temperatures, ocean acidification, and sea level rise pose significant threats to the health and productivity of these systems [11]. Coral reefs, for instance, are at heightened risk of bleaching events and mass die-offs due to warming waters, with severe implications for the rich biodiversity they support and the millions of people who depend on them for food, income, and coastal protection [2]. Given the myriad impacts of CC on ecosystems, it is crucial to implement adaptive management strategies that promote the resilience and conservation of these vital natural resources. Investing in ecosystem-based adaptation measures, such as the restoration and protection of habitats, the establishment of ecological corridors, and the integration of biodiversity conservation into land-use planning and decision-making is paramount to mitigate these effects [1][8].

These kind of disruptions in climate patterns and effects in the environment, ecosystems, and economy, result in job losses, reduced income, and heightened social instability in affected regions [17]. Moreover, CC is projected to lead to the displacement of millions of people due to the increased frequency and intensity of natural disasters, as well as the loss of habitable land [9]. An important factor of the latter is the loss of potable water and increasing sea level that threatens most of habitable land. Glaciers and polar ice caps are melting at an accelerating rate, contributing to an increased risk of flooding in coastal areas [10]. Moreover, sea levels have risen by about 20 centimeters since 1900, and the rate of rise has accelerated in recent decades, posing a threat to coastal communities and ecosystems [3]. Consequently, CC poses significant challenges to achieving sustainable development and poverty alleviation goals, especially in vulnerable regions and communities [9]. To address these challenges, it is essential to implement comprehensive

adaptation and mitigation strategies, such as improving water resource management, promoting climate-resilient agriculture, and enhancing disaster risk reduction efforts [2].

In response to minimize these CC effects, a quick and robust transition to low-carbon and climate-resilient economy, industry, and society is paramount. The development of decarbonization strategies will contribute to society to curb CC, to protect their economies from the negative impacts of CC while also fostering innovation, creating new job opportunities, and improving public health [18]. For this to be possible, it is essential to implement policies and strategies aimed at reducing GHG emissions, fostering resilience, and promoting sustainable development. Investments in renewable energy technologies, such as solar PV, can help drive the transition to a low-carbon economy, while simultaneously creating new employment opportunities and stimulating economic growth [19][20].

The energy sector is a key contributor to CC, as it accounts for approximately 73% of global GHG emissions [21]. As a consequence of the combustion of fossil fuels, such as coal, oil, and natural gas, for electricity generation, transportation, and industrial processes [10]. However, demand for energy has been continuously increasing due to the digitalization of the society, increased demand for the transport of goods, and an increasing globalized economy [1]. Fortunately, social awareness of the problem has driven significant changes in recent years, with a growing emphasis on transitioning to cleaner, more sustainable sources of energy to mitigate the impacts of CC and address energy security concerns [19][22]. Currently, fossil fuels, such as coal, oil, and natural gas, still make up a significant portion of the global energy mix, accounting for approximately 81% of the total primary energy supply in 2020, as shown in Figure 1-3. However, their share is gradually declining, as the growth of renewable energy technologies, such as solar PV, wind, and hydropower, accelerates [23]. Solar PV, in particular, has witnessed remarkable growth in recent years, with global capacity increasing from 40 GW in 2010 to over 714 GW in 2020 [19].



IEA. CC BY 4.0.

Figure 1-3: Share of primary energy supply from 2010 to 2021 and projection to 2050 in order to accomplish established emission goals. Renewables include solar, wind, hydro, and biomass.

Figure extracted from [24].

The electricity generation sector is a critical component of the global energy system, accounting for around 19% of total final energy consumption in 2020 [20]. The share of renewables in global electricity generation reached around 29% in that same year, with a continued upward trend projected to account for almost 50% of global electricity generation by 2030, highlighting the substantial shift taking place within the energy sector [25]. The ongoing transformation of this sector, driven by the increased deployment of renewable energy technologies, is essential for mitigating the impacts of CC, as electricity generation is responsible for approximately 42% of global CO₂ emissions [26][27]. The decarbonization of electricity generation, through the integration of intermittent renewable energy sources like solar and wind, is therefore a crucial aspect of global mitigation efforts. In countries like Chile, the energy sector has also undergone significant changes in recent years, with a focus on diversifying the energy mix and promoting the expansion of renewable energy sources [22]. The share of renewables in Chile's electricity generation increased from 6% in 2010 to 25% in 2020, with solar PV and wind energy being the main drivers of this growth [28]. By 2030, Chile aims to achieve a 70% share of renewable energy in its electricity generation, demonstrating the country's commitment to a low-carbon energy future [17].

The transition to renewable energy sources, particularly solar PV, offers numerous economic, social, and environmental benefits, such as reduced GHG emissions, improved air quality, enhanced energy security, and the creation of new job opportunities [23][17]. However, social and technological challenges remain, including the integration of intermittent renewable energy sources into electricity grids, the need for energy storage solutions, the development of adequate

policy frameworks and financing mechanisms to support the large-scale deployment of renewables, and the improvement of PV module efficiency and integrability [25][29]. Despite these challenges, the ongoing transformation of the global energy sector and the increasing share of renewables in electricity generation signify a positive shift toward a more sustainable and low-carbon future, as they represent a clean, renewable, and abundant source of electricity generation [30][31].

In response to an increasing demand for energy and sustainable sources, PV technologies have experienced rapid advancements in recent years, resulting in significant cost reductions and performance improvements [32][33]. This progress has made PV increasingly competitive with traditional fossil fuel-based energy sources and facilitated its widespread adoption across the globe [30]. The deployment of PV systems not only reduces GHG emissions, but also promotes energy independence, enhances energy security, and creates new job opportunities in the clean energy sector [34][35]. Moreover, the continued development of advanced PV materials and technologies, such as thin film solar cells and devices, holds great potential for further increasing the efficiency, affordability, sustainability, and integrability of solar energy [36][37]. The following subchapter will explore why thin film technology is important in this scenario and can have big deal of impact in tackling this CC issue, and thus by investing in research and development, supporting policy frameworks, and fostering international collaboration, we can accelerate the widespread deployment of PV technologies and their contribution to mitigating CC [34][35].

1.2 State of the art of PV

1.2.1 Overview of PV technology

Photovoltaics (PV) refers to technologies that can transform sunlight directly into electricity by the photovoltaic effect. During the XXI century and specially during the last decade, PV technology has seen remarkable advancements in various aspects, including efficiency improvements, cost reductions, and material innovations. This is illustrated in Figure 1-4 showing how efficiency records have become more common over the past decade with several technologies emerging through the years. This rate of research and innovation has allowed PV technology to transition from an expensive and niche energy source to a mainstream and cost-competitive option for electricity generation. This is also reflected in the global PV market which has grown exponentially, reaching about 750 GW of cumulative installed capacity by the end of 2020 [24]. One of the major drivers for this growth has been the continuous improvement in solar cell efficiency, particularly for crystalline silicon (c-Si) based solar cells, which currently dominate the market, accounting for about 90% of global PV production. The efficiency of commercial c-Si solar cells has reached over 26%, while multicrystalline silicon (or polycrystalline, p-Si) solar cells have achieved efficiencies above 22% [38]. Innovations such as PERC (Passivated Emitter and Rear Cell) and HJT (Heterojunction with Intrinsic Thin Layer) have contributed significantly to

these efficiency improvements. Another development that has increased interest in PV technology are bifacial solar modules, which can capture sunlight on both sides, increasing their overall energy yield [39]. Bifacial technology benefits from albedo, the reflectivity of the ground surface, which can vary depending on factors such as the type of ground cover, season, and location [40]. With these advantages, bifacial modules can generate up to 30% more energy compared to their monofacial counterparts under optimal conditions. The adoption of bifacial solar modules is growing, particularly in utility-scale PV installations, due to their higher energy production and reduced levelized cost of electricity (LCOE) [41].

Apart of the Si-based technologies, emerging thin film photovoltaic (TFPV) technologies, such as perovskite solar cells, have shown great potential, with lab-scale efficiencies reaching over 25%, rivaling those of conventional c-Si solar cells. Perovskites have attracted considerable interest due to their unique optoelectronic properties, low-cost solution-based processing techniques, and the rapid progress in their efficiency [42]. However, challenges in terms of stability, scalability, and potential environmental issues related to lead (Pb) content are still being addressed for these emerging technologies to become commercially viable [34][35]. In recent years, more mature TFPV technologies, such as CdTe and CIGS, also have demonstrated potential for further efficiency improvements and reduced manufacturing costs [34]. The specific interest in thin film (TF) technologies is based on that they offer the advantage of being lightweight, flexible, and suitable for building-integrated photovoltaics (BIPV) applications. The market share of TF technologies, though currently small compared to c-Si, still holds promise due to their unique properties and potential applications.

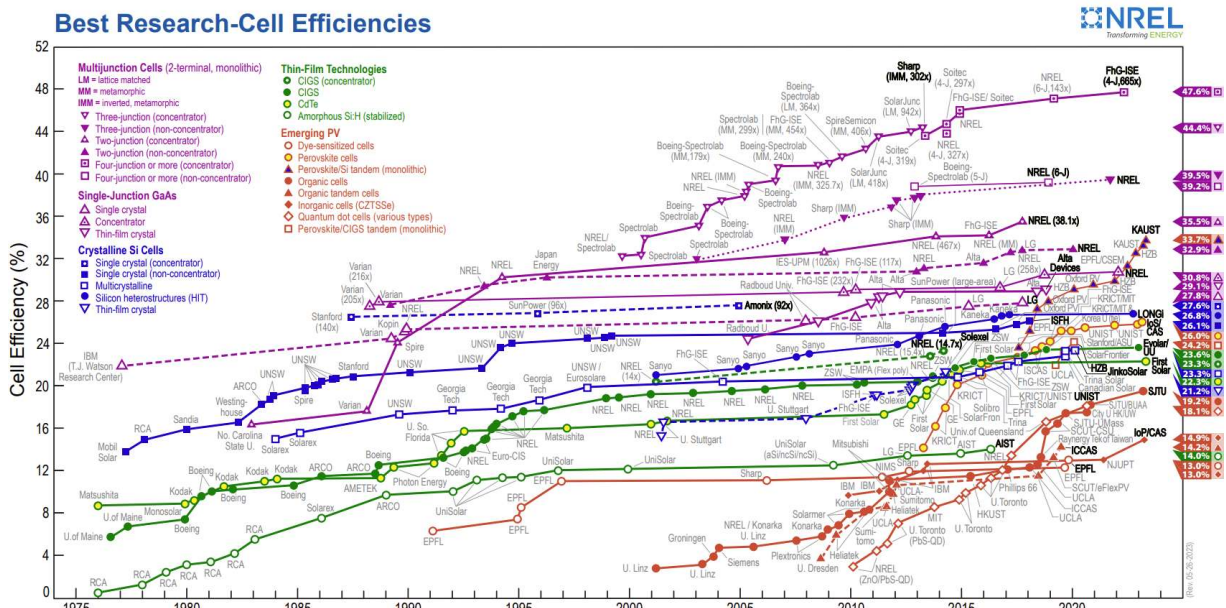


Figure 1-4: Best Research-Cell Efficiencies compiled by the National Renewable Energy Laboratory (NREL). Figure extracted from [43].

1.2.2 Thin film photovoltaic technology

TFPV, which consider functional devices with typical layer structure thicknesses $< 100 \mu\text{m}$ in the context of this thesis, have emerged as a promising alternative complementary to conventional c-Si solar cells, offering several advantages that make them an attractive option for widening the PV application by opening the possibility of its integration into new products. Some notable advantages of TFPV technology that makes them different from conventional c-Si technologies are:

- a) Reduced material usage: These devices are thinner and lighter solar cells compared to their c-Si counterparts [44]. This materials reduction is because TF materials possess a direct bandgap, in contrast with c-Si which is an indirect bandgap semiconductor, requiring the material to be thicker [45].
- b) Compatible with lower production costs: TF devices show a reduction in the use of high value materials as their manufacturing processes often require less material and energy-intensive methods than those employed for c-Si solar cells [34]. This cost reduction can help make solar energy more accessible and affordable, contributing to the global transition towards renewable energy sources [44].
- c) Compatibility with curved surfaces: the possibility to reduce the thickness of the device allows to achieve high flexibility without compromising the mechanical integrability of the devices. This makes this technology interesting for applications where conventional solar cells might be impractical due to their rigidity, bulkiness and weight, such as in building-integrated photovoltaics (BIPV), agrovoltaics (APV), and vehicle integrated photovoltaics (VIPV) applications [46].
- d) Light condition adaptability: Light condition adaptability is significantly enhanced by employing a solid solution, which facilitates band gap tuning. This adaptability enables the device to be fine-tuned for various lighting environments, including indirect sunlight, low irradiance scenarios, or indoor settings illuminated by artificial light sources [47].
- e) Monolithically large-scale areas: The growth of TFPV layers through easily scalable deposition processes, such as Physical Vapor Deposition (PVD) via Sputtering or Chemical Vapor Deposition (CVD) techniques, enables dimensional scalability restricted by the physical capability of the systems and the achievable homogeneity [48]. This contrasts with c-Si technology, which relies on the mechanical integration of various wafer to construct photovoltaic modules [49].

Despite the numerous advantages of TFPV, challenges still remain in terms of efficiency, stability, and scalability. Researchers from the PV community have been working on various TF technologies, such as amorphous silicon (a-Si), CdTe, CIGS, and emerging alternatives like

kesterite and perovskite solar cells or more exotic technologies as quantum dots (QD) or low-dimensionality PV concepts, to overcome these challenges and optimize their performance [37][44]. As progress continues in the development and optimization of TFPV, their potential to become a significant contributor to global renewable energy generation increases, offering a viable and sustainable solution for our growing energy needs.

CIGS solar cells have emerged as a leading TF technology, offering higher efficiencies than both a-Si and CdTe solar cells [50]. The efficiency of CIGS has been steadily increasing, with some laboratory-scale cells achieving over 23% efficiency, surpassing the performance of p-Si solar cells just until the year 2020 [51][38]. The high efficiency of CIGS solar cells can be attributed to factors such as a high absorption coefficient, which allows for the efficient conversion of sunlight into electricity, and the tunable bandgap that enables the optimization of the material's absorption properties [37]. Despite their promising performance, the complex material system and the difficulties in scaling up the production process have limited the widespread adoption [48]. Also, various deposition techniques have been developed for CIGS fabrication, such as co-evaporation, sputtering, and electrodeposition, each with their own set of advantages and challenges [52].

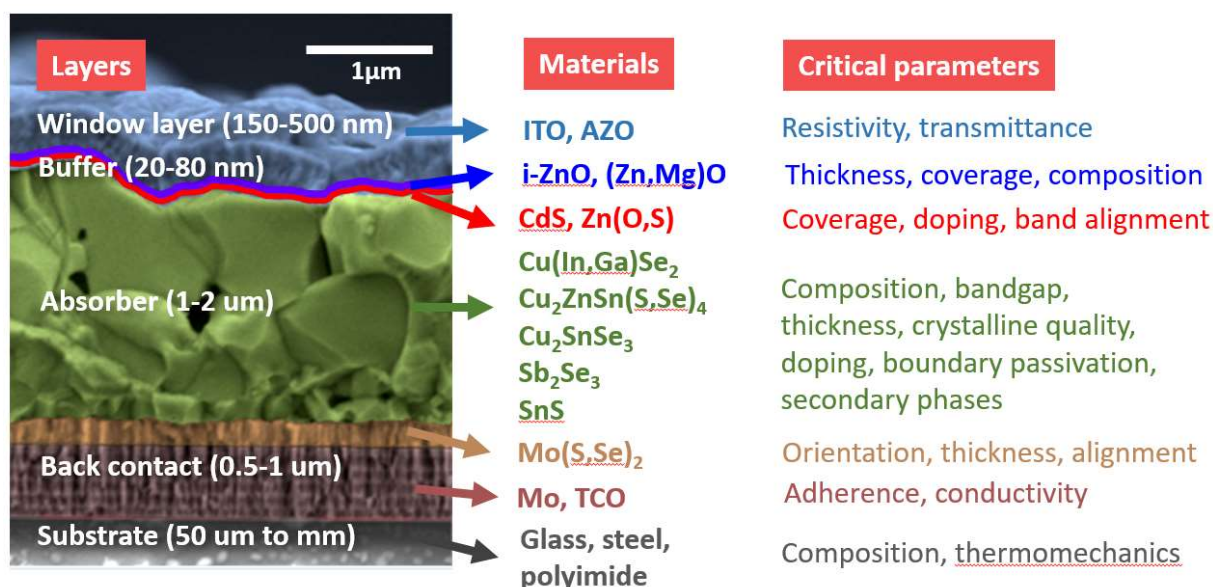


Figure 1-5: Generic structure of a TF device based on p-n heterojunction. This shows how this technology involves multiscale, multilayer, and multiprocess devices with over 20 critical parameters to control.

One of the main challenges of CIGS-based technology, and extensible in general to TFPV technologies, lies in the complexity of the system, which makes it difficult to control the homogeneity in large areas. In the scaling up CIGS production is the need to maintain uniformity in the material's composition and structure across large areas, as any deviation can lead to a significant decrease in the solar cell's efficiency [53]. Interface engineering plays a crucial role in

the performance of CIGS solar cells, as the buffer layer, typically composed of cadmium sulfide (CdS), has a significant impact on the device's overall efficiency and stability [54]. Researchers have been exploring alternative buffer layers, such as zinc oxide (ZnO) and zinc magnesium oxide (ZnMgO), to address concerns related to the use of toxic cadmium (Cd) and to further enhance the performance of CIGS solar cells [52]. Another challenge for CIGS solar cells development is the cell-to-module efficiency gap, which arises due to the differences in performance between small-area laboratory cells and large-area modules produced at a commercial scale. Addressing this gap requires an understanding of the factors affecting module performance, such as interconnect design, current matching, and module encapsulation, and the development of strategies to minimize efficiency losses during the scale-up process [48].

Additionally, another challenge for CIGS-based technologies are supply-chain risks. For instance, materials such as Gallium (Ga) and Indium (In) are considered to be of high economic importance and high supply risk, categorizing them as critical resources [55]. In this context, there are 4 main risk areas where these elements are compromised: supply, demand, concentration, and political risks [56][57]. Depending on the element, supply reduction risk varies drastically. For example, for In, depletion times of reserves are calculated to be about 20 years, meanwhile for Ga is 3000 years. Demand increase risks are 2-fold. The direct increase of demand over rare elements and the decrease in demand of host materials for by-products. For instance, in the case of Cd, Te, In, Ga, and Se, it is considered for their dependence on host materials to be 100% as being by-products of Zn, Cu/Pb, Zn, bauxite (main Aluminum ore), and Cu, in the same order. It is then possible that, shortages of these elements due to demand can be due to these 2 different unrelated phenomena: reduced demand for parent materials and increased demand over the small availability of the materials themselves. Concentration risk is also an important factor, since the production of some of these materials is highly concentrated in some areas and countries. For instance, in 2014, China was the main European Union (EU) supplier of several materials, including Ga with 71% of the supply. The latter is then related to political risks, since conflicts between these regions may lead to shortages of such materials. For these reasons, an increase interest on recycling has been observed in the past decade, however, the estimated effect of recycling rare-earth elements in the supply of such elements is expected to be negligible or at the most complementary [58][59], due to both technical difficulties and high costs of the needed processes. As such, Ga, In, Se, and Te are often to be found of high-risk and difficult supply, particularly damaging for CIGS and CdTe technologies.

In this context, it is important to explore and develop novel technologies based on abundant materials. Kesterite-based solar cell technology is a class of TFPV technology based on the quaternary compound copper zinc tin sulfide (CZTS) or copper zinc tin selenide (CZTSe). This technology has garnered considerable attention due to their attractive properties and potential as an alternative to CIGS solar cells [60]. One of the most appealing aspects of kesterite-based solar cells is the earth-abundant nature of their main constituent elements, which addresses the supply

and environmental concerns associated with other TF materials like CIGS and CdTe [61]. Despite these advantages, kesterite solar cells currently suffer from lower efficiencies compared to CIGS solar cells, with the highest reported efficiency for CZTSSe solar cells being 14.9% as confirmed by the latest version (63) of the “Solar cell efficiency tables” in 2024 [62]. The lower efficiency can be attributed to several factors, such as the presence of defects, high open-circuit voltage (V_{OC}) deficit, and challenges related to the material’s complex stoichiometry and phase stability [63][64]. Recent research efforts have focused on improving the performance of kesterite solar cells through various strategies, including defect engineering, interface engineering, and cation substitution [65][66]. Defect engineering aims to suppress the formation of harmful defects, such as vacancies, antisite defects, and secondary phases, which can negatively impact the solar cell’s performance [65]. Interface engineering, on the other hand, involves optimizing the properties and composition of the buffer layer, typically CdS or Zn(O,S), to improve the overall device performance [63]. Cation substitution, such as replacing some of the copper with silver or indium, can help stabilize the kesterite phase, enhance the material’s optoelectronic properties, and reduce the V_{OC} deficit [64]. Despite the progress made in recent years, further research and development are needed to address the remaining challenges and optimize the performance of kesterite solar cells. As our understanding of the material properties, defect formation mechanisms, and interface interactions in kesterite solar cells continues to improve, these promising earth-abundant and environmentally friendly TFPV technologies could play a significant role in the global transition towards renewable energy [60][66].

In summary, TFPV technologies, including CIGS and kesterite based solar cells, have made significant advancements in recent years, offering a lightweight, flexible, and potentially cost-effective alternatives to widen the application of PV beyond to traditional c-Si solar cells. CIGS solar cells have achieved high efficiencies, surpassing p-Si, but face challenges in scaling up production, maintaining material uniformity, and addressing environmental and supply concerns related to the use of toxic elements and rare-earth materials. Kesterite solar cells, on the other hand, offer a promising earth-abundant and more environmentally friendly alternative, but currently suffer from lower efficiencies due to the presence of defects, high V_{OC} deficit, and complex stoichiometry. Ongoing research efforts in both CIGS and kesterite solar cells are focused on addressing these limitations, exploring strategies such as defect engineering, interface engineering, and alternative buffer layers. As progress continues in the development and optimization of TFPV technologies, their potential to become a significant contributor to global renewable energy generation increases. The future prospects TF technologies are promising, with the potential to revolutionize the solar energy landscape and facilitate the global transition towards a more sustainable and environmentally friendly energy mix. With the above, this thesis focuses in these two promising technologies: CIGS and kesterites, performing experiments with these materials in the first and second publications, respectively. The following subchapter explains how these technologies work, with both being p-n heterojunction-based devices.

1.2.3 Basic working principle of PV devices

To generate the PV effect, semiconductor materials with a bandgap energy that aligns with the energy spectrum of the solar radiation (ranging from 0.5 to 3 eV) are needed. This bandgap energy represents the minimum energy required to excite an electron to a conductive state. Currently, silicon stands as the predominant semiconductor used to produce PV materials and devices, also referred as first-generation technology. It is followed by TF technologies, also known as second-generation technology, which utilize light-absorbing materials such as a-Si, CdTe and, of course, CIGS and kesterites. Other emerging technologies include organic and hybrid solar cells, as well as multi-junction solar cells. These are considered as third-generation technology. All these different technologies can then directly convert incoming light into electricity. Specifically for the case of CIGS and kesterite based devices, as well as for several other inorganic absorber materials, this is possible thanks to the p-n junction formed between two differently doped zones: the p-type and n-type. The p-type is where the majority charge carriers are holes, and the n-type where the majority of the carriers are electrons. When in contact, electrons from the n-region will diffuse towards the p-region, and the holes will diffuse from the p-region to the n-region. The area of this exchange is called the depletion region. In this zone, the n-region is positively charged, and the p-region is negatively charged, creating an electric field oriented from n to p. With the p-n junction formed, incident photons with energy greater than the bandgap of the p-type material will be absorbed and excite electrons, which move from the valence band to the conduction band, leaving holes in the process. The electrochemical potential difference between the materials facilitates the separation of electrons and holes in the depletion region and the movement of the carriers. In quasi-neutral zones, carriers move by diffusion: only those with sufficient diffusion length will be collected. Ultimately, holes are collected at the positive pole of the cell and electrons at the negative pole. This movement is what ultimately generates current. This process is shown in a schematic representation in Figure 1-6.

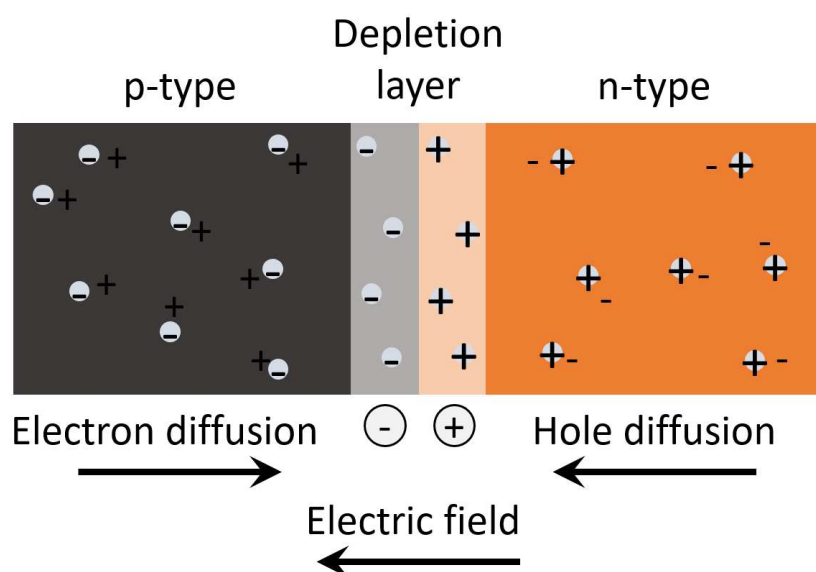


Figure 1-6: Schematic of the p-n junction.

The p-n junction can be either the same material (homojunction) or of two different semiconductors (heterojunction). P-n heterojunctions, first described by Russel Ohl in his 1941 patent “Light-sensitive electric device including silicon” [67], have become an important part for several electronic devices beyond electricity generation, such as rectifiers, photodetectors, diodes, and sensors. In solar energy technology, p-n heterojunctions are used with diverse semiconductor materials to build many different types of devices such as silicone based [68], pervoskites [69], CISE [70], CIGS [71], and kesterites [66].

The efficiency of a real solar cell will never reach 100% respect to the incoming solar radiation, due to several types of losses that occur at different stages. For instance, photons with energy lower than the material’s bandgap are not absorbed, excess energy is lost through thermalization, some photons are reflected off the material’s surface, and electron-holes may succumb to recombination, all leading to a reduced efficiency. Furthermore, the theoretical limit for the efficiency of a solar cell using a single p-n junction is about 30%, what is known as the Shockley-Queisser limit [72]. This limit has been calculated based on the material’s bandgap, assuming an ideal scenario where all recombinations are radiative, charge carriers have infinite mobility, and all photons with energy equal to the material’s bandgap are absorbed. To account for these losses, the solar cell is simplified in a classic model as a circuit with a diode V_d , including a series resistance R_s representing contact and connection resistances, and a shunt resistance R_{sh} representing various leakage currents from the PV source I_{ph} , creating the final device voltage potential V , as represented in Figure 1-7.

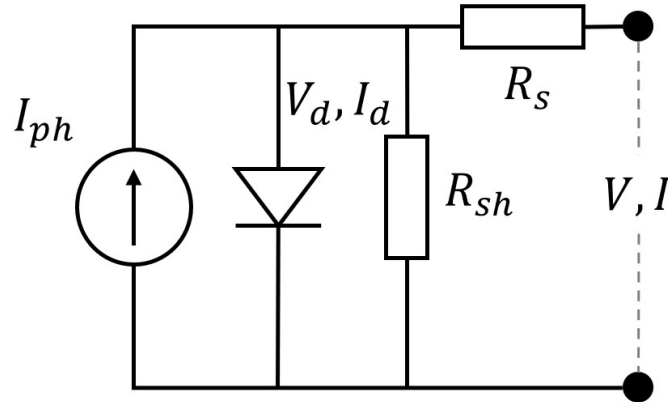


Figure 1-7: Classic circuit model of a photovoltaic cell.

From this classical circuit model of a PV cell, the device can be formulated electrically to ultimately obtain the main electrical properties, namely open circuit voltage V_{OC} , short circuit current I_{SC} , fill factor FF and efficiency η . To understand the calculation these characteristics using the single diode model, we can start with the fundamental circuit equation that encapsulates the behavior of a solar cell through a combination of the photovoltaic current I_{ph} , diode current I_d , series resistance R_s , and shunt resistance R_{sh} . The total current output, I , of the PV cell is described by the following equations:

$$I = I_{ph} - I_d - \frac{V + IR_s}{R_{sh}} \quad \text{Eq. 1-1}$$

$$I_d = I_0 \left(e^{\frac{q(V+IR_s)}{nkT}} - 1 \right) \quad \text{Eq. 1-2}$$

Where I is the final device current, I_0 the reverse-bias saturation current of the diode, q is the electron charge constant $1.602 \cdot 10^{-19} \text{ C}$, n is the ideality factor of the diode, k is the Boltzmann constant $1.638 \cdot 10^{-23} \text{ J/K}$, and T is the temperature in the p-n junction. I_d can also be expanded as a diode current equation as shown in Eq. 1-2. From here, V_{OC} occurs when current $I = 0$, and I_{SC} is then the current flow when $V = 0$. Then, to determine the maximum power point (MPP), we need to find the combination of voltage and current where the product $V \cdot I$ is maximized. This point is crucial because it represents the most efficient operating point of the PV device. With the MPP, the FF is defined as the ratio of the maximum power point $P_{max} = V_{mpp} I_{mpp}$ to the product $V_{OC} I_{SH}$:

$$FF = \frac{V_{mpp} I_{mpp}}{V_{OC} I_{SH}} \quad \text{Eq. 1-3}$$

And the efficiency η of the solar cell is then calculated as the ratio of the maximum power output P_{max} to the input solar power P_{in} :

$$\eta = \frac{P_{max}}{P_{in}} \quad \text{Eq. 1-4}$$

This efficiency is a key metric for evaluating a solar cell as it encapsulates the combined effects of all the parameters of the solar cell. Thus, reflecting both its electrical characteristics and its ability to harness solar energy. Finally, the above characteristics can be graphically illustrated in the Current-Voltage (I-V) curve, where I is plotted as a function of V , as shown in Figure 1-8. In this plot, I_{SC} , V_{OC} , I_{mpp} and V_{mpp} can directly be extracted, and thus FF and efficiency obtained after.

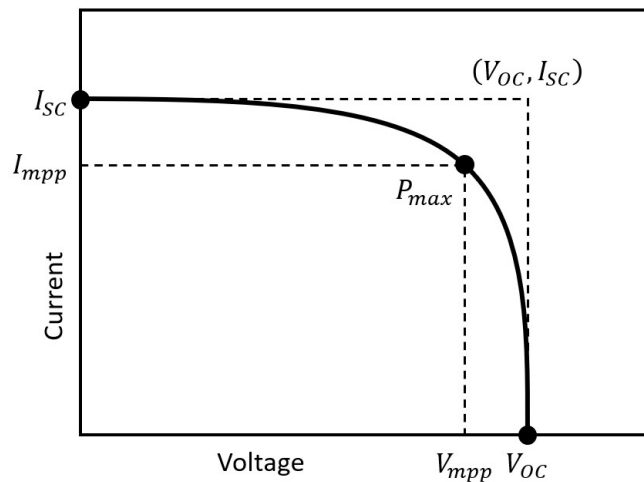


Figure 1-8: Current-voltage curve showing its main characteristics.

Finally, a typical structure of TFPV device based on compounds like chalcopyrite or kesterite type materials is shown in Figure 1-5. Here the p-n heterojunction is formed between absorber (typically of p-type for studied technologies) and buffer layer (typically of n-type for studied technologies), and additional layers for the back and front contacts of the solar cell. The latter are usually more complex due to formation of intermediate layers at the back contacts (e.g. MoS₂ or MoSe₂) and deposition of an extra layer for better isolation at the front contact (e.g. i-ZnO layer). Additional to the importance of each of the layer of the TF solar cell, a significant role is also played by different interphases between the layers, which increases the complexity of the whole structure, making it much more advanced than a simple circuit. This complexity results in the necessity of making an advanced approach in study and characterization, by combining different characterization techniques and by making a special set of samples, which is covered by a combinatorial approach in the analysis of complex systems.

1.3 Combinatorial analysis for materials and devices

Combinatorial Analysis (CA) refers to the process where combinatorial samples or sets of samples are systematically prepared with a specific property intentionally varied in a controlled manner either within a single sample or across multiple samples. This allows for the thorough exploration and understanding of how changes in that property affect outcomes or behaviors. CA has emerged as a powerful technique in materials science and engineering research, enabling the simultaneous study of multiple variables and their interactions in complex systems [73]. This approach allows for a more comprehensive understanding of intricate systems and accelerates the discovery and optimization of novel materials. CA typically involves generating a large number of samples, or a single graded sample, with varying combinations of properties, followed by parallel analysis using a combination of experimental and computational techniques. By investigating the relationship between these variables and the resulting properties, researchers can gain valuable insights into how different factors impact the performance of materials or devices [74]. In materials science, CA has found applications in numerous areas, such as high-throughput screening of catalysts, discovery of new alloys, and optimization of solar cell materials [75]. By utilizing combinatorial methods, researchers can efficiently explore vast parameter spaces and identify optimal combinations that lead to enhanced material properties and performance. The application of CA, in conjunction with advanced data analysis and Machine Learning (ML) techniques, further accelerates the material discovery process [76]. Through these integrated approaches, researchers are not only able to identify correlations, but also uncover underlying mechanisms governing the system, leading to the development of innovative materials with desired properties and functionalities [77]. To effectively conduct CA, several crucial considerations must be accounted for, including sample preparation, characterization techniques, and data analysis approaches.

1.3.1 Samples for combinatorial analysis

Combinatorial samples or sets of samples are those in which a property is deliberately varied in a controlled way in-sample or sample-to-sample, respectively. The analysis of combinatorial samples represents an efficient way of obtaining relevant insights that can be used both for extracting information about fundamental material properties and for technological optimization. In the case of TF, different physical and chemical deposition techniques can be employed for the preparation of combinatorial samples and sets of samples which result in discrete or gradient sample libraries, respectively. A discrete library consists of individual samples in which each of them has a discrete variation of a property, normally related to composition. On the other hand, a gradient library is a single sample with a deliberate inhomogeneity consisting in a continuous variation (gradient) of a property across its surface, e.g. a sample with a graded thickness in one of its layers. Diagrams and real examples of each type of sample library are presented in Figure 1-9. Despite of both approaches being suitable with CA, the two of them present advantages and disadvantages that need to be considered for the choice of one or the other.

When it comes to discrete libraries for materials research, one of its primary benefits is that they are produced under well-defined and controlled conditions. This ensures that each sample within the library is homogeneous, providing a consistent baseline. This homogeneity is crucial for drawing accurate conclusions about the effects of specific variations on material properties and device performance. Moreover, the specific preparation conditions are known for each sample of the library, allowing to directly select the optimal ones for further application and to directly correlate the properties of the samples with the fabrication parameters. However, there are also drawbacks with discrete libraries. The preparation time of several samples can be extensive, especially when ensuring that each sample meets the strict criteria required for homogeneity. Additionally, the resolution for sample-to-sample property variation (i.e., the difference between consecutive samples) is limited by the characteristics and configurations of the fabrication equipment. This can limit the range and resolution of conditions that can be studied.

In contrast, graded libraries offer a different set of advantages and challenges. One of the most significant benefits of graded libraries is their high resolution for property variation. Since graded samples are intentionally produced with inhomogeneities that smoothly varies a certain property across the sample, the effects of these variations on performance can be studied with high detail. Also, a single sample can provide a great deal of information, as it contains a large range of conditions within itself, which is both time-saving and cost-effective. However, the primary drawback of graded libraries is the inherent uncertainty in the fabrication conditions. Since the samples are intentionally inhomogeneous, it can be challenging to pinpoint exact processing conditions and their effects on material properties or device performance, leading to potential ambiguities in the research results.

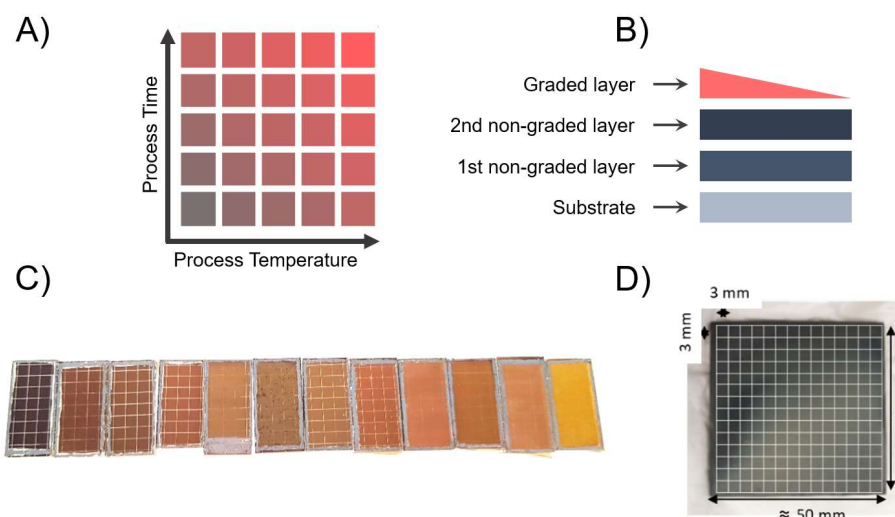


Figure 1-9: A) Diagram of discrete sample set with process temperature and time variations, B) diagram of continuous spread sample with 1 graded layer, C) picture of a discrete sample set and D) picture of a continuous spread graded sample.

1.3.2 Characterization techniques for combinatorial analysis

In terms of characterization, the process of understanding and quantifying the physical, chemical, and structural properties of materials, it is essential to employ multiple techniques to maximize the information gathered about the studied material and increases the chances of uncovering relevant correlations and valuable insights for enhancing future iterations of the devices. Relying solely on a single characterization technique often presents limitations, as no single method can provide a complete picture of a material's multifaceted nature (Figure 1-10), which becomes more critical when speaking about the such a complex systems as TFPV devices. Different characterization techniques focus on different aspects of a material and devices can give a more comprehensive approach on understanding their properties and limitations. For instance, Raman spectroscopy (RS) can analyze chemical composition and molecular structures, Photoluminescence (PL) can deliver optical and electronic information, X-ray fluorescence (XRF) can quantify composition and current-voltage (IV) measurements can measure the optoelectronic parameters of the devices, including their efficiency to convert the light into electricity. All the above information is important for any material being studied but using them all in one study allows for cross-verification of properties and offers a more holistic understanding with more complex, but also more accurate, correlations. In other words, combining different methods can compensate for the limitations of individual techniques and provide a more accurate, comprehensive analysis. By leveraging the combination of the results extracted from these techniques, it is possible to obtain a broad picture of the systems and push knowledge to a deep understanding of the material and its properties. This is crucial for the progress, optimization, and development of novel PV devices technologies and materials.

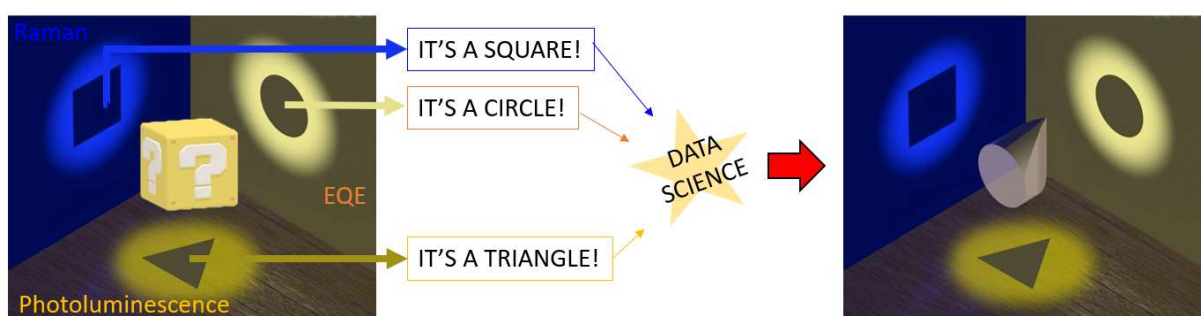


Figure 1-10: Schematic example to illustrate how the combination of different techniques allows to obtain further insights compared to normal experimentation focused on single techniques. The more characterized a sample is, the more it is possible to visualize the different aspects of its nature.

However, to achieve this, the characterization methods must fulfill several requirements, including non-destructive testing, rapid acquisition times, automation capabilities, compatibility with other techniques, and high spatial resolution. These requirements are critical to enable characterization

of samples in bulk that provide relevant statistical data to further understand complex materials and devices. Overall, a study using a combinatorial sample should aim to measure compositional, optical, structural, and optoelectronic properties, all in a non-destructive, fast, and automated way that secures the traceability of the data reliably, as illustrated in Figure 1-11.

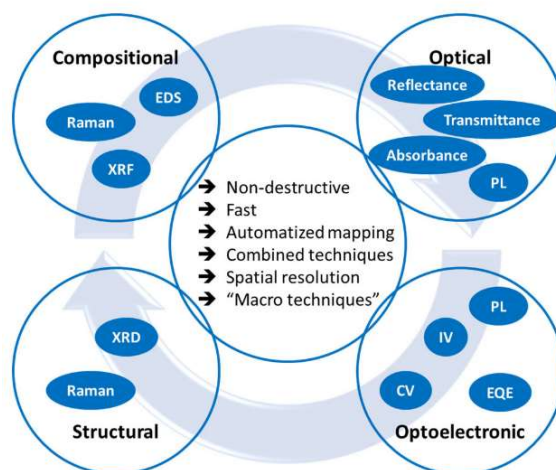


Figure 1-11: Characterization techniques, such as compositional, optical, structural, and optoelectronic, must comply with requirements to be fit for CA. Figure extracted from [78].

1.3.3 Data analysis for combinatorial approach

Data analysis in combinatorial experiments has traditionally depended on standard approaches like correlation and statistical methods, which show good results at identifying relationships between variables in simpler datasets. These conventional methods are valuable for their straightforward analysis and ease of understanding, building great confidence on their use with researchers. However, they often fall short when dealing with the complex, high-dimensional data typical in combinatorial studies. Challenges particularly arise in discerning non-linear relationships and subtle patterns within large datasets where several variables can be considered. Selecting the proper analysis tools for these experiments depends on the complexity and nature of the data. Traditional methodologies are more suitable for experiments with fewer variables and linear relationships, where the simplicity and interpretability of results are paramount. In contrast, modern Artificial Intelligence (AI) and ML techniques are preferred for more complex, multi-dimensional datasets where patterns are not immediately apparent.

Because of this, the field is increasingly moving towards more sophisticated computational techniques. The development of ML and AI has introduced robust tools capable of handling and interpreting the vast and intricate datasets generated in combinatorial experiments. These advanced methodologies excel not only in pattern recognition and predictive modeling but also offer the advantage of processing and analyzing data at a much higher speed than traditional methods. Their

adaptive learning capabilities make them especially suitable for CA, where unexpected relationships and complex interactions may appear. By integrating these technologies, researchers can dig deeper into data, uncovering insights that were previously inaccessible with conventional methods.

This evolution from traditional data processing to a more dynamic, AI and ML-driven analysis represents a significant leap in the field of CA, promising more comprehensive understanding and innovative discoveries in various scientific domains. Unfortunately, challenges in applying ML in CA experiments still exist. These include the need for large and well-annotated datasets for effective training, the complexity of selecting appropriate ML models and features, and the interpretation of ML outputs in a scientifically meaningful way. Furthermore, the holistic analysis of data is also not well developed in the field, leading to a slow incorporation of combinatorial experiments. The development of clear methodologies with accessible tools aiming to simplify the application of AI and ML into CA can greatly improve combinatorial experiments and make CA more appealing for researchers.

1.4 AI algorithms as support in the materials research

1.4.1 Introduction and basic principles of AI and ML

AI and ML are rapidly evolving tools that have been gaining significant attention in recent years, largely due to their capacity to revolutionize a wide range of industries and disciplines [79][77]. They have been used to solve a large number of complex problems and have been applied in various domains, such as natural language processing, computer vision, and robotics [80][81]. With the increasing availability of large datasets and powerful computational resources, it is now possible to develop sophisticated models that can perform tasks that were thought to be too difficult or virtually impossible to accomplish successfully, like natural language interaction, image recognition, and other tasks [80][82]. This success of ML has led to a renewed interest in the field that has also expanded to materials science, where ML techniques have been employed to accelerate the discovery of new materials and optimize existing ones [83][84][85]. The incorporation of AI and ML into materials research has the potential to significantly expedite the development of novel energy materials and further advance renewable energy technologies [86][75]. Additionally to their practical applications, AI and ML have also given rise to a number of ethical and philosophical questions regarding their implications for society and human decision-making [87][88]. As these technologies become more widespread and integrated into our daily lives, it is crucial to ensure that they are transparent, unbiased, and accountable [89][90][91].

Before going further, it is important to define what the AI, ML terms mean. Unfortunately, a standardized definition of these is not yet available, and many variations can be found in the literature. However, in general terms, most definitions would agree that AI is any system, normally

a computer program, that performs a task considered smart or complex in an automated way. With this, it is possible to define ML as an AI that uses data to make predictions over new information, and keeps getting better as more data is available, hence it learns and improves over time. Under this definition, many algorithms and techniques can be mentioned, being the most common and widely used Linear Regression (LR). Other examples include Principal Component Analysis (PCA), Linear Component Analysis (LDA), Random Forest (RF), Support Vector Machine (SVM), and many others. Furthermore, Artificial Neural Networks (ANN) is a subset of ML algorithms that consist of a series of multiple and iterative transformations that decompose a complex problem into a sub-set of simpler problems, like Deep Learning algorithms (DL).

There are several ways to classify ML algorithms, but a common approach is to group them based on their level of supervision, which includes:

- Supervised learning: algorithms that learn from labeled training data and make predictions about unseen data. Examples include LR and SVM.
- Unsupervised learning: algorithms that learn from unlabeled data and find patterns or structure in the data. Examples include k-means clustering and PCA.
- Semi-supervised learning: algorithms that learn from a mix of labeled and unlabeled data. Examples include Label Propagation and Semi-supervised Support Vector Machines.
- Reinforcement learning: algorithms that learn from the consequences of their actions in an environment. Examples include Q-learning and Policy Gradients.

Another way to classify ML algorithms is by their output type:

- Classification: Algorithms that predict categorical output
- Regression: Algorithms that predict continuous output

During the development of this thesis, AI and ML were used, including supervised, unsupervised, and classification methods. Regression methods and ANNs where also tested in the Further Exploratory Experiments.

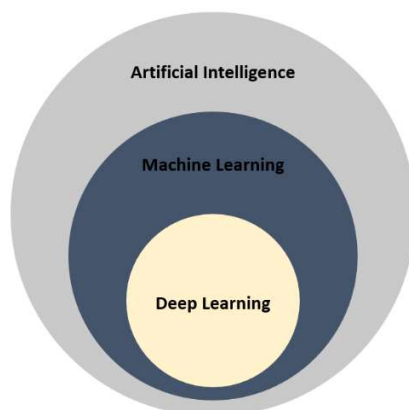


Figure 1-12: General view of the relationship between AI, ML, and DL.

To explain how ML works it is convenient to start with the most basic form, which has been used for over a century for several mathematical and research problems: LR. LR works through least-squares estimation. It takes a basic optimization problem with an objective function of the form:

$$\min \sum_{i=1}^n (\vec{\beta} \cdot \vec{x}_i - y_i)^2 \quad \text{Eq. 1-5}$$

Where \vec{x} is the independent variable, y_i is the dependent variable to predict, and $\vec{\beta}$ is the vector containing the model's parameters. Examples of LR are vast considering its simplicity, low computational needs, and many decades of diverse applications. A good modern example is in [92], where researchers find that there is a linear relationship between V_{OC} and the relative OVC-related Raman peak areas, as illustrated in Figure 1-13.

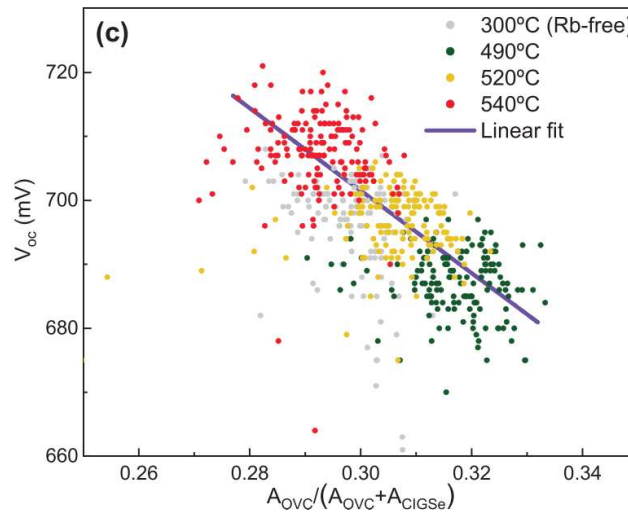


Figure 1-13: Linear regression is performed for V_{OC} and OVC relationship for high performance CIGS solar cells. Colors represent different process temperatures the solar cells were subject to.

Figure extracted from [92].

Other than LR, there are many other algorithms such as the already mentioned PCA, LDA, RF, SVM, QDA, and K-means, among many others. Each of these algorithms has unique characteristics and is best suited for specific types of problems. However, at their core, many of these algorithms share a fundamental principle with LR: the concept of defining and optimizing an objective function, often leveraging large datasets. To show this, we can see how some of these work, in particular PCA and LDA as they have more importance in this thesis. For instance, PCA, is an unsupervised dimension reduction algorithm that looks for the bits of information that better explain the difference between the data [93]. It does this by performing orthogonal linear

transformations of the data to a new coordinate system where the greatest variance between the data is found. This reduction in dimensions means that the information is not selected or deleted in the process but rather translated into different axis. In other words, all the original information is, one way or the other, preserved. This reduction process is performed as many times the user specifies, being capable of making a reduction to a 1-D system. In a similar way to LR, we can translate this into an optimization problem as:

$$w = \max \sum_{i=1}^n (x_i \cdot w_i)^2 \quad \text{Eq. 1-6}$$

Where w is a unit vector of constants w_i such that $\|w\| = 1$, and is equivalent to the transformation that maximizes the variance. The first iteration will find the first principal component, and a second principal component can then be found using that same Equation 1-6. After subtracting the first principal component as

$$\hat{X}_k = X - \sum_{s=1}^{k-1} XW_{(s)}W_{(s)}^T \quad \text{Eq. 1-7}$$

Where X is the matrix containing the observed data, \hat{X}_k is the data transformed into the new dimensional space after subtracting the first principal component. This procedure can be repeated $K - 1$ times being K the initial dimensionality of the problem.

Similarly, LDA is a supervised technique to analyze the difference between classes in a dataset. This makes LDA particularly useful in classification tasks, as it can identify the most important features that discriminate between different classes [94]. In other words, LDA finds the coordinate system that has the largest distance between different groups and the smallest dispersion within each group. For example, if there are 2 categories, it is possible to express the objective function to be optimized when searching for these dimensions as

$$\max \left\{ \frac{\mu_1 - \mu_2}{s_1^2 + s_2^2} \right\} \quad \text{Eq. 1-8}$$

Where s is the variance and μ is the mean position. This equation above can then be generalized for n categories.

By combining the feature reduction capabilities of PCA with the class discrimination capabilities of LDA, an even more powerful tool for data analysis is obtained, known as a cascaded PC-LDA or simply PC-LDA. While PCA focuses on finding the directions of maximum variance in the

data, LDA aims to find the directions that maximize the separation between different classes. With this method, it is possible to identify the most relevant variables for the problem at hand and use them to train a classifier with high performance. PC-LDA is then a combined form of both algorithms, arranged in a sequential way. In other words, PCA is first applied to an intermediate dimension and then transformed by LDA to a final coordinate system. This method is advantageous for the use of both data variability (PCA) and classification group differences (LDA), being just as popular as PCA and LDA used alone since they still may produce better results depending on the specific problem. In particular, these techniques are popular for spectroscopy related problems, since normally spectroscopic measurements are of high dimensionality (vectors of length 1000 or more are common) and several properties that can be extracted that may be difficult to compute are present in spectroscopy data, such as peak position, peak area, peak convolutions, and peak widths, to name a few. An interesting example of this is for autofluorescence spectroscopy over blood plasma for tuberculosis diagnosis [95]. Authors find that PCA is highly effective with 95.2% accuracy to predict tuberculosis through this kind of data, as illustrated in Figure 1-14. Another common use for PC-LDA is for cancer diagnosis and classification. A specific example of this is in [96] where they classify RS measurements using PC-LDA for the diagnosis and distinction of 5 different types of thyroidal tumors, some of them benign and others malignant. The results show that they are able to accurately distinguish between types of tumors in 1 vs 1 comparison (79% to 100% accuracy depending on the comparison pair) and 81% overall accuracy between benign and malignant types. Similar examples are abundant in the literature, particularly with the use of RS [97].

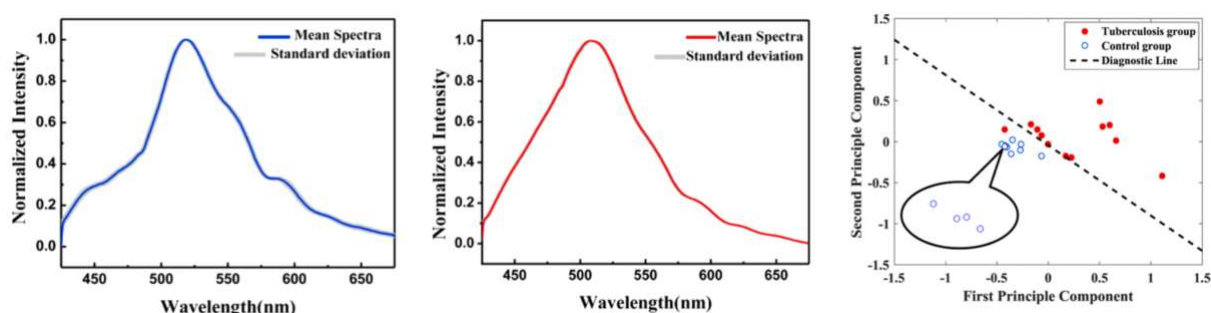


Figure 1-14: a) Mean spectra for healthy blood plasma and b) mean spectra for unhealthy (tuberculosis) blood plasma and c) final 2-D dimensionality after PCA. Figure extracted from [95].

Other more complex type of algorithms widely used nowadays are Artificial Neural Networks (ANN or just NN for Neural Network). ANNs are a type of ML algorithm that are inspired by the structure and function of the human brain [80] and consist of layers of interconnected nodes, called neurons or units, which are used to process and analyze data. ANNs are particularly useful in tasks such as image and speech recognition, natural language processing, and predictive modeling [80]. To represent what a NN does, the following equation is needed:

$$y = f(x; \theta, w) = \varphi(x; \theta)^T w \quad \text{Eq. 1-9}$$

Where x is the input, θ are parameters for φ and w parameters for $\varphi(x)$. With this, what a NN should try to do is to learn θ such as it gets the best approximation to y as possible as in $f \approx f^* = y$, where f^* is the original function. To construct a basic NN, then we can use Eq. 1-9 to stack one after the other as different layers of the NN. As an example, to build a three-layer NN (input layer, 1 hidden layer, and output layer) with an input of two parameters, it is possible to define its components as:

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \quad \text{Eq. 1-10}$$

$$h = \begin{bmatrix} h_1 = g(xW_{,1} + c_1) \\ h_2 = g(xW_{,2} + c_2) \end{bmatrix} \quad \text{Eq. 1-11}$$

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad \text{Eq. 1-12}$$

$$y = hw \quad \text{Eq. 1-13}$$

Where h is the set of functions for each unit (neuron) in the layer that each holds a common function called activation function g , and c a bias value. In general aspects, this is the simplest form of a NN: a feedforward, fully connected network (see an illustration in Figure 1-15), but it can be extended to as many layers with as many units as desired. In that case, if each layer, with its respective units, forms a function f_n , then a NN with N layers will have the following shape:

$$f_n \left(\dots \left(f_3 \left(f_2 \left(f_1(x) \right) \right) \right) \right) \quad \text{Eq. 1-14}$$

The activation function can take many different forms. The most commonly used is Rectified Linear Unit (ReLU), where $g(z) = \max\{0, z\}$. This function is used in many applications and cases; however, it can really be anything that the user desires and as complex as needed. A different approach is to define the activation functions as a radial basis function, in other words, a function that outputs a value in function of the distance to some defined point. A typical function to use is a Gaussian function as

$$\varphi(r) = e^{-(\varepsilon r)^2} \quad \text{Eq. 1-15}$$

Where r is the radius, or distance, to the center. Such NN are defined as Radial Basis Function Networks [98] (RBFN). Radial basis function networks are particularly useful in applications where the input data has a non-linear relationship with the output. They are based on the idea of radial basis functions, which are functions that have a value of 1 at the origin and decrease as the distance from the origin increases. As mentioned, RBFNs use these functions as activation functions in their hidden layers, allowing them to capture non-linear relationships in the data. They are particularly useful in tasks such as function approximation, time-series prediction, and classification problems with non-linearly separable classes. Additionally, RBFNs are known to be robust to noise and outliers in the data and require fewer hidden neurons than other NN architectures. Overall, RBFNs are a powerful tool for solving a variety of ML tasks with non-linear input-output relationship.

With all this, it is patent that artificial NNs can take several shapes and forms, with virtually unlimited possibilities in terms of input and output types, number of layers and units per layer, and how all these elements interact with each other. A (mostly) complete chart of NN structures can be found in the Neural Network Zoo [99].

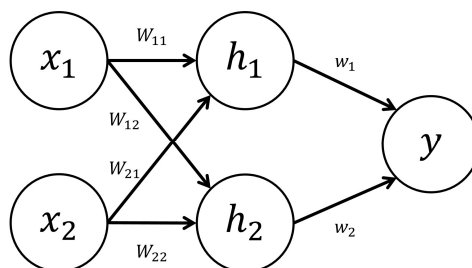


Figure 1-15: Representation of a three-layer NN with an input, hidden, and output layer. The input is a two-parameter variable, and the hidden layer contains two units. This is arguably one of the simplest forms of a NN.

1.4.2 General use of ML in material science

The utilization of ML in materials science typically adheres to a specific workflow: data selection, feature engineering, model building, validation, and result analysis, as illustrated in Figure 1-16. While the exact procedures and details will vary depending on the particular problem at hand, adhering to this structure is crucial for obtaining accurate outcomes and gaining valuable insights into the central research questions.

The initial stage in the ML workflow for materials science involves choosing the appropriate dataset. This step is crucial, as the quality and relevance of the data will significantly impact the accuracy and effectiveness of the ML model. It is important to carefully consider factors such as data source reliability, data completeness, and the presence of any noise or inconsistencies. The

selected data should be representative of the problem being addressed and encompass a wide range of material properties and conditions to ensure that the developed model is robust and applicable to various scenarios. Overall, the data should aim to a well-defined problem.

After selecting the appropriate data, the subsequent critical step involves processing and scaling the data. This phase is essential to prepare the data for effective analysis and model training. Data processing involves a series of steps to clean and organize the data by handling missing values, correcting errors, and removing duplicates. Scaling the data is a key part of this process too, especially for algorithms sensitive to the scale of input features. The goal is to ensure that all features contribute equally to the analysis and model training.

After, the following step is feature engineering, which involves extracting and selecting the most relevant features or attributes from the dataset that will be used as input for the ML model. This process requires domain expertise and a thorough understanding of the materials science problem being addressed. Feature engineering may involve applying transformations, aggregating data, or even creating new features that capture important relationships between variables. The goal is to identify the most informative features that can help the ML model make accurate predictions and uncover hidden patterns in the data, and also remove those that might create bias when processing, like noise, inaccurate data, and measure errors.

The modeling stage involves selecting an appropriate ML algorithm and building the model based on the selected features. Each algorithm has its advantages and drawbacks, and selecting the right one depends on the specific problem, data characteristics, and desired outcomes. Researchers should experiment with different algorithms and parameter settings to identify the best-performing model for their particular problem.

Once the ML model is built, it needs to be tested and validated to ensure that it performs well on unseen data and can generalize to new situations. This is achieved by splitting the dataset into a training set, which is used to build the model, and a testing set, which is used to test the model's performance. Then, the algorithm can be re-trained with a different, but comparable, training set and test if it produces equivalent results. This indicates if the performance is due to the model itself or for a biased training/test set. Common performance metrics include accuracy, precision, recall, F1 score, and mean squared error. Researchers can also use techniques such as cross-validation to get a better understanding of the model's stability and performance across different subsets of data.

The final stage in the ML workflow for materials science is analyzing and interpreting the results obtained from the model. This step involves understanding the relationships and patterns that the model has uncovered, assessing the model's strengths and limitations, and determining how the findings can be applied to the problem at hand. The insights gained from the ML model can be

used to guide further research, inform decision-making, or even be integrated into other computational tools and techniques to accelerate the discovery and development of new materials.

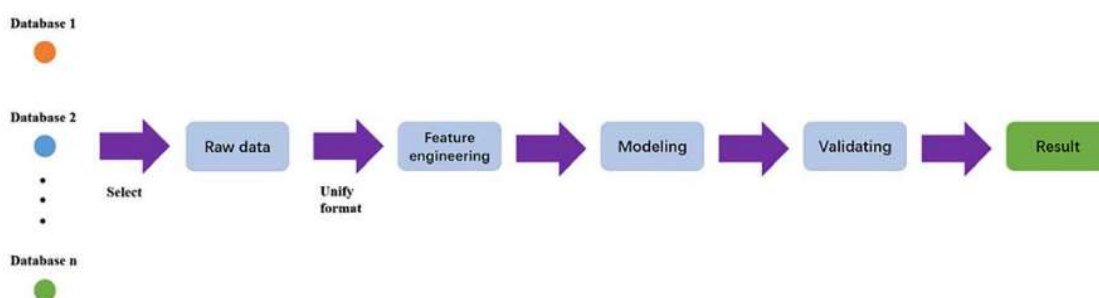


Figure 1-16: General proper workflow for ML applications in material science. Figure extracted from [100].

1.4.3 Introduction to XAI

From the last sections, it is clear how complex ML algorithms can get, particularly when using a combination of algorithms, dividing the problem into a subset of problems with different algorithms, or when more layers and units are added to a NN model. This complexity makes it difficult to really understand what the AI is actually doing and to exactly know how and why is able to make predictions. Explainable Artificial Intelligence (XAI) is an emerging field that aims to make AI systems more transparent, understandable, and accountable to humans. With the rapid advancements in AI, there has been a growing concern about the lack of interpretability and transparency of AI models, especially in critical decision-making scenarios [101][102]. This line of research, however, is subject to ongoing discussion on the right questions to make, and the ethics involving the explanation of algorithmic decisions. This is paramount as cases of bias can be life changing [103], and these biases are difficult to approach [104][88]. Furthermore, this topic has gain so much importance, that European regulators have established the explanation of life affecting decisions by a computer program a right [87]. To tackle this issue, not only methodologies are required, but also proper definitions, correct questions, and high ethical standards [88][91]. Particularly in natural sciences, as mentioned, ML has become increasingly popular for its ability to quickly and efficiently analyze large amounts of data. Its versatility and accessibility through various libraries and products have also contributed to its widespread use. However, without the ability to explain the results obtained from ML algorithms, their scientific value may be diminished, and the consistency of future research may be affected [90]. XAI is therefore crucial for ensuring the validity and significance of ML-based results in the field of natural sciences, and every other field.

One way to perform the latter is to find a function that approximates the algorithm to be analyzed to a simpler form that can be interpreted by humans. An example of such is LIME [89], or Local Interpretable Model-Agnostic Explanations, a popular technique that aims to provide interpretable

and transparent explanations for the predictions made by any black-box model, including NNs. The technique works by perturbing the input data in a local region around the instance being explained, and then training a simple linear model on the perturbed data. This allows LIME to explain the predictions made by the black-box model by approximating the decision boundary in the vicinity of the instance being explained, and attributing importance to the input features based on their contribution to the prediction. The output of LIME is a human-understandable explanation of the model's decision-making process, which can be used to improve the trustworthiness, accountability, and interpretability of ML models. The latter can be expressed as

$$\mathcal{L}(f, g, \Pi_x) = \sum_{z, z' \in Z} \Pi_x(z) (f(z) - g(z'))^2 \quad \text{Eq. 1-16}$$

Where $f(x)$ is the model to be explained, Π_x is a proximity measure between an instance z to x , and $g \in G$ is a model where G is the class of linear models, such that $g(z') = w_g \cdot z'$. Then it is possible to minimize this function with a complexity measure $\Omega(g)$ that is low enough that is interpretable by humans as the loss function expressed like

$$\xi(x) = \min_{g \in G} \mathcal{L}(f, g, \Pi_x) + \Omega(g) \quad \text{Eq. 1-17}$$

With this is possible to find an approximation $\mathcal{L}(f, g, \Pi_x)$ that humans can understand.

A different approach is to improve the interpretability of ML algorithms is through the use of sensitivity analysis, which involves systematically varying the input features and measuring the resulting change in the prediction. By comparing the predictions obtained with the original values of the features to those obtained with the modified values, it is possible to understand how the prediction changes as each feature is varied. An example of this is RELIEF [105], a feature selection method that detects statistical significant features according the change in the target. This can be modelled as:

$$\text{Variable type} \begin{cases} \text{if nominal} \rightarrow \Delta(y_k, \hat{y}_k) = \begin{cases} 0 & \text{if } y_k = \hat{y}_k \\ 1 & \text{if } y_k \neq \hat{y}_k \end{cases} \\ \text{if numerical} \rightarrow \Delta(y_k, \hat{y}_k) = \frac{(y_k - \hat{y}_k)}{\mu_k} \end{cases} \quad \text{Eq. 1-18}$$

Where $\Delta(y_k, \hat{y}_k)$ denotes the difference between the original instance y_k and the perturbed instance \hat{y}_k and μ_k is a normalization parameter to transfer the domain to a relative scale. This method can successfully identify what features are more important when making a particular prediction and, in the aggregate, see what features are more relevant on the possible outputs. However, this method is originally limited to two classification groups and its random perturbation nature yields strictly

stochastic results. Despite this, the method can be easily generalized to a desired number of classification groups and the random perturbations can also be replaced with intended and deterministic changes by considering:

$$I_j = P_M(x) - P_M(R_j(x)) \quad \text{Eq. 1-19}$$

Where P_M is the probability function of the ML model M , $x \in X$ is an array of dimensions $h \times w$, and R is a function of local perturbation of feature $j \in J$. With this, the value of I will determine how important a feature is compared to others measured by the probability change in classification.

Finally, XAI can appear in different shapes and forms with vastly different approaches and levels of complexity. Regardless, the exploration of XAI underscores its critical role in enhancing transparency, fostering trust, and ensuring ethical AI deployment, thereby bridging the gap between advanced AI technologies and their practical, understandable, and responsible application in various domains.

1.4.4 AI in energy, PV devices and materials research

As mentioned, AI and ML have emerged as transformative tools in various fields, including energy and PV devices and materials. In fact, it is foreseen that the widespread use of these tools, in conjunction with CA, can shorten development times for novel materials by a factor of 10, from 10 to 20 years to just a few years [75][86][73][106]. Moreover, there is evidence for 1,000 times acceleration in the rate of the discovery of novel amorphous alloys with the power of combining high-throughput experiments (HTE) with ML models [87]. In contrast with traditional methods for discovering new materials, such as the empirical trial and error method and density functional theory (DFT), that typically require a long research and development cycle, are of high cost with low efficiency, and have difficulty keeping pace with the development of materials science today [100], AI and ML have shown great potential for the discovery, optimization, and characterization of advanced materials for PV devices. ML algorithms can analyze vast amounts of data, identify patterns, correlations, and optimal material configurations, enabling researchers to focus their efforts on the most promising candidates [77][81][84]. As such, these tools have the potential to revolutionize the way we discover, design, optimize, and manufacture devices, enabling faster innovation and implementation of sustainable technologies.

As so, successful research has been achieved using ML and AI in the field of energy and PV materials. For example, Mahmood et al. (2021) reviews several examples where ML has been applied successfully for improvement and discovery of organic solar cells. Ren et al. (2018) combined ML and HTE iteratively to accelerate the discovery of new metallic glasses for energy storage applications. In addition to material discovery, ML can be utilized for CA and HTE. For instance, Fonoll et al. (2022) discussed the importance of sample preparation, characterization

techniques, and analysis approaches in CA. They highlighted the use of ML in studying the relationships between material properties, enabling researchers to gain insights into the factors affecting device performance. Furthermore, ML models can analyze large amounts of data generated from experiments and simulations, identifying patterns and correlations that can help optimize device structures and configurations [79][76]. Also, AI can assist in automating the analysis and interpretation of data from various characterization techniques, enhancing the efficiency and accuracy of material characterization. Moreover, AI can be employed to study the relationship between material structure and properties, ultimately guiding the design of novel materials with tailored functionalities. For example, Vasudevan et al. (2019) explored the use of AI in materials science, focusing on high-throughput library generation and ML techniques to discover new materials and understand the underlying physics. Lastly, AI can help improve the reliability and reproducibility of experimental results by automating data processing and analysis. This can lead to more robust conclusions and a better understanding of the underlying phenomena in material systems, ultimately accelerating the development of advanced materials for PV applications [79].

A key limiting factor of ML models is that, generally, the predictive space is within the input realm. In other words, predictions and conclusions from these kinds of experiments, in most cases, can only be used for data that is within the parameters of the experiment. However, in scientific discoveries, it is generally preferred to predict far outside the training distributions. For instance, much of materials research aims to identify ways to produce top-performing materials that are, by definition, beyond the confines of the available data. To overcome this, the common single-hypothesis experimentation must become obsolete, transferring over to experiments designs that are combinatorial in nature [107]. The idea that the search for new materials with outstanding properties and new mechanisms require a broader search through composition- processing-structure-property space than could be afforded by conventional one-sample-at-a-time techniques, has been patent for over a century [108]. Far from being a novel idea, there is still paths to pave in order to reach the full potential of computation in material science and PV technology. Once this point is reached, where knowledge extraction catches up the HTE synthesis and characterization, the limit to rate of new materials discovery becomes the decision making, i.e., what materials to pursue next given the knowledge of materials discovered so far and processing conditions needed to make them [76].

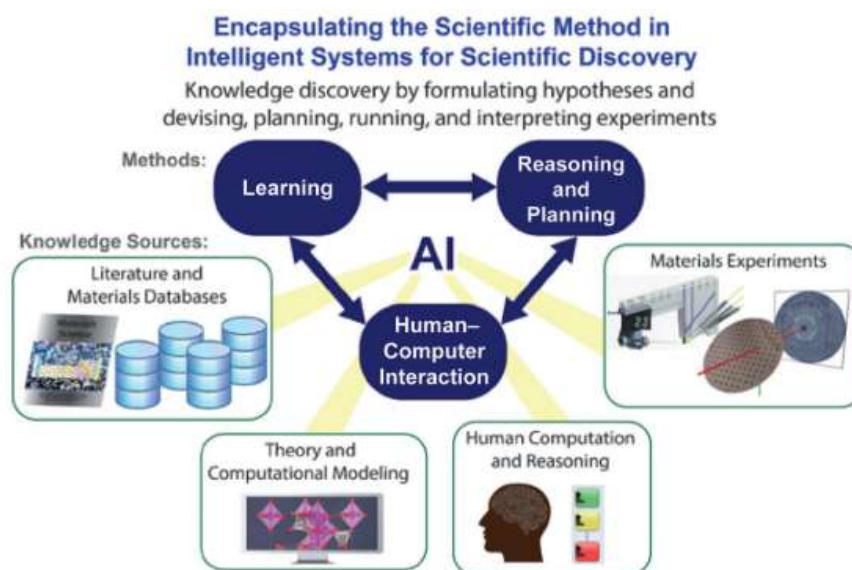


Figure 1-17: Leveraging AI can enhance human capabilities and expedite discovery within the scientific method. Scientific discovery necessitates the integration of various AI techniques beyond solely data-driven ML. By combining primary AI methods, such as learning, reasoning, and planning, with human-computer interaction, a comprehensive approach emerges. This approach facilitates the integration of multiple knowledge sources, including databases, theory, experiments, and human reasoning, as demonstrated through relevant examples. Figure extracted from [107].

1.5 Objective of the thesis

TFPV devices hold immense potential to disrupt different industries by bringing cheap and sustainable energy. This means that, either directly or indirectly, this technology can replace fossil fuel generated power, with a clean and affordable alternative and, furthermore, it can allow solar energy to be used in places and applications that other technologies can't. However, even though much progress has been made in the past years, a long way ahead of improvement can be foreseen, with room for improvement in terms of efficiency, material usage, and production scaling. The evidence suggests that, to achieve better results for TFPV, traditional experiment designs must step aside to give way for CA experiments driven by AI and ML, that are able to considerably reduce research and development times. These technologies have been under research with more interest just over the past few years in the field of PV materials and devices with promising results. However, despite this notable applications and results obtained so far in the field with AI and ML, the implementation of these tools, including also CA, even though more common as time passes, has been rather slow for this research field [81]. This is mainly due to several barriers between researchers and these tools: the availability of large amounts of data, proper pre-processing of data sets, lack of multi-disciplinary groups with experts of enough computational knowledge, and more [85][109]. Additionally, the difficult interpretation of results that decrease trust in ML models, as

discussed in the previous subsection, also contributes to this issue, even more considering that XAI is difficult to implement even for experts in the field [110][111]. In other words, the application of AI tools requires substantial theoretical, statistical, analytical, and programming skills from a research team. To overcome these barriers and accelerate the implementation of these technologies and decrease research and development times, it is paramount to achieve the following points:

- Facilitate high-throughput data acquisition: Ensuring the availability of tools that can handle large-scale data collection efficiently is crucial (from the point of view of time and human resources consumption). This will streamline the gathering of large datasets necessary for properly perform CA experiments and training and testing AI models.
- Automate data processing: It's essential to simplify the data processing workflow by making it generalized and automated. Reliance on specialist knowledge should be reduced to make the process more accessible and efficient.
- Establish clear ML and CA protocols: The development of straightforward ML and CA methodologies with predictable outcomes is vital. This will allow for automated research processes, making AI applications more reliable and easier to replicate.
- Democratize ML result interpretation: Making the interpretation of ML results more accessible is key. This approach will create greater trust in AI technologies and enhance the understanding and insights derived from these tools, broadening their application across various fields.

Accomplishing the above points will significantly streamline the integration of AI in research settings, paving the way for more efficient, accessible, and reliable technological advancements in the field of TFPV. As so, the identification of these needs and problems have inspired this thesis, which proposes as its main goal “the development of innovative CA techniques based on AI and ML algorithms for the accelerated research and development of relevant chalcogenide-based TFPV materials, including CIGS, CZTSSe, and other emerging technologies, to reduce their lab-to-market times and improvement cycles.” To accomplish this main goal, the following three objectives are also defined:

- Objective 1: “Design and implement autonomous systems to obtain high amounts of data in large area / large number of samples using different spectroscopic (Raman, PL, reflectance, transmittance) and optoelectronic (IV, EQE, IQE, CV) techniques, that will enable innovative big data-based research based on the correlation of physicochemical properties of the materials with the device performance.”

- Objective 2: “Develop new methodologies based in AI and ML algorithms for big data processing. This will include fast-rate data conditioning and processing using CA results and AI-based strategies.”
- Objective 3: “Make tools available for non-expert scientists for easy implementation of AI data processing and interpretation of results in an accessible way.”

Through the realization of this work, the experiments, developed tools, and results have aligned with this objectives and main goal. This is further detailed in the following section and also reflected in the included articles in this compendium.

2 METHODOLOGY

The methodology developed and used during this thesis is schematically shown in Figure 2-1. Overall, the methodology is divided in five steps. It starts with the synthesis and characterization of a combinatorial sample or of a combinatorial set of samples. These must have a compositional or process condition variation to allow for the study of its impact in the final performance. Also, the samples must be comprehensively characterized, using techniques that encompass compositional, structural, optical, and electrical properties so that there is a holistic view in the study. The second step is for the characterization data obtained to be divided into features and targets. This means that the desired property to be studied (i.e. open circuit voltage, efficiency, etc.) is defined as the target and the feature is all the rest of the data that is used to make a prediction or classification in terms of the target. In other words, the data must be divided as data to make a prediction and data to be predicted. In the next step the data is subjected to conditioning and fusion respecting traceability and preparing it for their input into the ML algorithm. Conditioning refers to the necessary steps to process the data and remove undesired information that might bias or interfere with the results. Fusion in this case refers to the process of merging (fusioning) the data selected as features into one single vector. This is a crucial step for this methodology as it simplifies in great deal how much processing must be done to the data, as there is no further information that needs to be directly extracted from spectroscopic data, specifically. Then, traceability is the process of correctly assigning the measurements to a measured spot by keeping track of which measured point uses what measurement, as different spots may use the same measurement if the measurements are of a too large area compared to others. This can be challenging when different measurement techniques measure over different areas, which is often the case. The fourth stage in the methodology is to analyze the data by applying ML. The present work proposes the use of, but not limited to, PCA, LDA, or cascaded PC-LDA classification algorithms as a powerful tool to process spectroscopic data. The results of such models lead to the classification of the data which allows making decisions about the most relevant and optimum production parameters and generates knowledge about the critical properties of the materials and devices. These results, however, must be evaluated in terms of overfitting and efficacy by studying the training, test, and validation results. Finally, in the last step, the methodology allows to select critical samples, techniques, and spectral ranges to generate more solid feedback for further technology improvement.

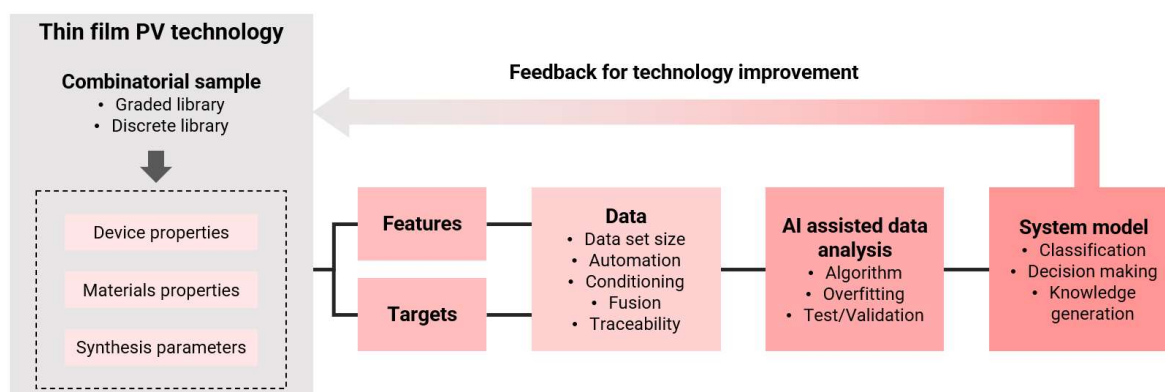


Figure 2-1: General flow of the proposed methodology for accelerated research using CA and ML.

It should be noted that while the combinatorial sample preparation and the characterization processes are tailored specifically for TFPV technologies, their direct applicability to other technological areas may be limited. However, further steps of the methodology are more universal since they are related to the data manipulation and application of ML and barely depend on the analyzed materials. This generalization makes it possible to adapt and extend the proposed methodology to other complex multilayer and multicomponent systems relatively ease.

For the included experimental articles in this compendium (articles 1 and 2), this methodology was applied, with different needs and complexity levels in each case. For instance, the first article makes use only one spectroscopic technique, thus data fusion and traceability was a rather simple process, in contrast to the second article where three spectral vectors were used or each of the measured points. Regardless of these differences, both cases successfully obtain insightful results, which are discussed in each article accordingly. The third and fourth articles, in contrast, are open-access and open-source tools to help implement this methodology. The following subsections explain in more detail each of these steps, highlighting key considerations and details essential for its implementation, and the following chapter introduces the articles in question where their specific details are presented.

2.1 Sample preparation

During this thesis there have been two different sets of samples prepared in collaboration with colleagues from the SEMS group at IREC, and from the Dutch Organization for Applied Scientific Research (TNO). Both of these sample sets were developed with graded variations of one of their components. The combinatorial sample set used in the first study in cooperation with TNO is subdivided in three sets with different substrate materials used: Si, PET/CIGS and PET. On top of these substrates a nanometric layer of AlO_x was deposited using a laboratory-scale rotary spatial atomic layer deposition (ALD) reactor. The nominal thickness of the AlO_x layers was changed

from 15 nm up to 75 nm. Moreover, due to the used technological process, a graded layer of AlO_x was deposited with the gradient in a radial shape, thicker towards the center and thinner towards the edges (Figure 2-2). This allows to develop a methodology based on normal reflectance (NF) spectroscopy for monitoring AlO_x nanolayer thickness. More details and information about these samples can be found in the respective published article also presented in Section 3.

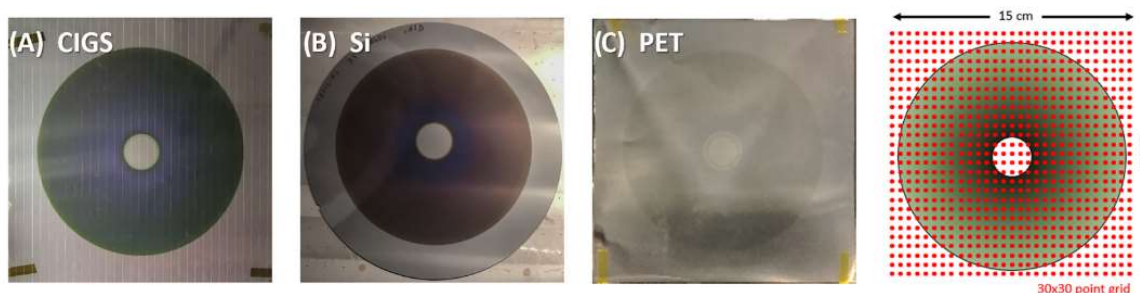


Figure 2-2: Samples used for the AlO_x thickness evaluation experiment on A) PET/CIGS, B) Si, and C) PET substrates. The final diagram is an approximation of the measured points. The inner radius of the AlO_x deposition is 1.2 cm, meanwhile the outer radius is 7.6 cm. Extracted from [112].

For the sample used in the second study, a CZGSe combinatorial sample was synthesized with a compositional gradient of $[\text{Zn}]/[\text{Ge}]$ ratio (Figure 2-3). The sample was made by sputtering of the metallic precursor and its subsequent selenization. The sputtering was performed on a soda lime glass substrate covered by a metallic Mo layer. The solar cell devices were then finished with the standard procedure for SEMS devices, by depositing the CdS layer (using chemical bath deposition), i-ZnO layer (using sputtering of Zn metal in O atmosphere), and $\text{In}_2\text{O}_3\text{-SnO}_2$ layer (using sputtering of In-Sn alloy in O atmosphere). With the use of Raman spectra measured under different excitation conditions, the effects of the graded $[\text{Zn}]/[\text{Ge}]$ ratio on structural and compositional properties of the compound and on the performance of the solar cell devices was explored in detail. More details and information about this sample can be found in the respective published article also presented in Section 3.

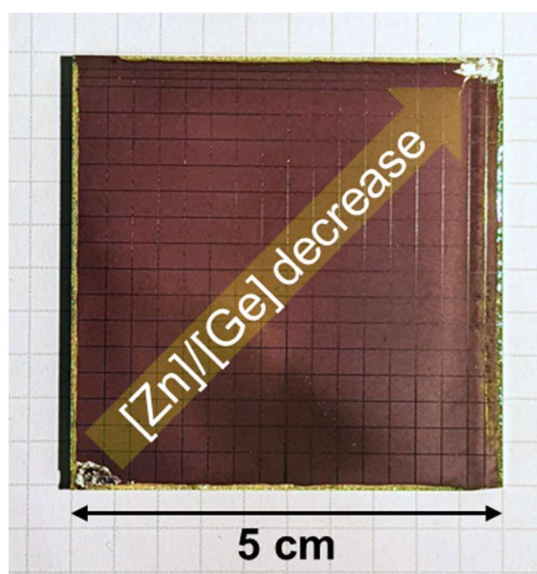


Figure 2-3: Photo of the CZGSe kesterite sample used in the second article. Change of the color is directly related to gradient of $[Zn]/[Ge]$ ratio. Extracted from [112].

2.2 Characterization techniques

Given the intrinsic complexity of the compounds and layer structures in TFPV devices, it is crucial that the characterization of the sample library is thorough and exhaustive. This entails applying multiple characterization techniques to the same defined small area of the sample, known as the analysis area or pixel cell. Such a comprehensive approach is necessary to maximize the data gathered and to uncover potential correlations. The chosen methods for characterization should be non-destructive, possess a spatial resolution equal to or finer than the pixel cell size, and ideally offer fast data acquisition times and automation capabilities. The techniques must be non-destructive, so it is possible to conduct numerous measurements on the same sample without altering its properties. Additionally, a high spatial resolution is essential to detect and analyze the variations in properties present in graded combinatorial samples accurately. The techniques utilized in this study, which are detailed below, meet these criteria.

2.1.1 Raman spectroscopy

Raman measurements have been performed using IREC developed Raman setups optimized for the UV–Visible spectral region (based on Horiba Jobin Yvon FHR640 monochromator) and NIR–IR region (based on Horiba Jobin Yvon iHR320 monochromator). The first system is coupled with an open electrode CCD detector cooled down to $-132\text{ }^{\circ}\text{C}$ and the second with NIR enhanced CCD detector cooled down to $-75\text{ }^{\circ}\text{C}$. Solid state lasers ($\lambda_{\text{ex}} = 532, 633, 785\text{ nm}$), and gas He-Cd lasers ($\lambda_{\text{ex}} = 442\text{ nm}$) were used as excitation sources. Different gratings for the light dispersion were employed to optimize the spectral resolution. The measurements were always performed with laser

power density in the range 25-150 W cm⁻² by using a macrosport with a diameter in the range of 50–70 μm depending on the excitation wavelength. Finally, the use of unpolarized laser beam allowed to minimize the impact of the crystalline orientation in the Raman spectra.

2.1.2 Spectroscopic Normal Reflectance

A probe with a broad emission (400 – 1000 nm, approximately) halogen lamp as illumination source was used for the reflectance measurements. The IREC designed probe was coupled to an XY-crane to enable mapping measurements and the acquired signal was processed through a compact CCD spectrometer (Thorlabs CCS200). A vacuum chuck table was employed to ensure the flatness of flexible samples during measurements. A spot size of ~100 μm and acquisition times in the 10-100 ms range (depending on the type of sample analyzed) were employed for the measurements.

2.1.3 Optoelectronic characterization

I-V measurements were performed under illumination and in dark conditions have been performed to evaluate the final device performance. I-V characteristics were acquired on complete devices using a Sun 3000 AAA solar simulator from Abet Technology (uniform illumination area of 15 x 15 cm²) calibrated with a Si reference solar cell under AM1.5 illumination. Sample temperature around 25 °C was kept during the measurements.

2.1.4 X-ray fluorescence

Compositional measurements and thickness estimation of the different layers were determined with an X-Ray fluorescence (XRF) equipment (Fischerscope XVD) calibrated by inductively coupled plasma (ICP). The measurements were done using a 50 kV accelerating voltage, a Ni10 filter to reduce background signal, and an integration time per measuring point of 45 seconds. The equipment in question comes equipped with a measurement analyzer software that permits the estimation of compositions and thickness of the layer stacks by calculating attenuations in the subsequent layers, this required sample calibration that was analyzed by ICP technique.

2.3 Automated measurements

In accordance with the developed methodology described above, the use of fast, automated mapping measurement procedures is important due to the high number of measurements to be performed on the sample libraries that, otherwise, could result in extremely high acquisition times for obtaining the high-statistics that are desirable for AI application. There are significant advances in automation of the measurement systems, that are self-controlled, do not require significant sample preparation time or the permanent control and supervision of an operator. In the case that these systems are not available, standard spectroscopic systems can be automated through the coupling of measuring probe-heads to programmable motorized XYZ gantry systems or translation stages combined with the use of optical fibers or with the use of detectors that can be integrated

within the probe-heads. This approach leads to a significant reduction of the labor and time needed for acquiring a statistically relevant amount of data and to increase the size of the data sets. Automation is then a critical component to significantly contribute to the evolution, enhancement, and development of any research area, including TFPV materials and devices. It enables an enhanced efficiency and suitability into the research process with fast and consistent data acquisition and reduced long-term costs that yields better products.

In this thesis work an automated spectroscopic platform was developed and implemented to streamline the measurements and preliminary analysis of various spectroscopic techniques, including multiwavelength RS, PL, and NF. This system was designed with a modular approach, comprising of several components such as a large area and high precision X-Y gantry crane, multiple modal probe, light excitation source, spectroscopic detector, and a system controller (consisting of both hardware and software). The centralized control of these modules was achieved through the implementation of custom software, utilizing LabVIEW for the control of the equipment and Python for processing the obtained data. LabVIEW, with its robust ability to interface seamlessly with hardware, is particularly well suited for managing the complex coordination required among the various components of the system. Also, beyond its intuitive graphical programming interface, LabVIEW is also commonly supported natively on several commercially available pieces of hardware, making it easy to implement and merge with other systems. On the other hand, Python brings to the table its advanced data analysis capabilities, courtesy of several extensive libraries available. Additionally, the customization and scalability offered by both LabVIEW and Python are important for both research and process monitoring settings where specific needs and modifications are required. Moreover, the strong community support and comprehensive documentation available for both languages greatly aid in troubleshooting. Together, the integration of LabVIEW's hardware-oriented precision with Python's powerful data handling and analysis proficiency forms a powerful combination, making them ideal for creating an automated, efficient, and adaptable spectroscopic platform.

More specifically, LabVIEW was used for the automation and coordination of the gantry systems and spectrometers. In other words, LabVIEW is accountable for the measurement, acquisition, and movement. This is shown in Figure 2-4 where the block diagram (fundamental way of programming in LabVIEW) is shown. The figure shows only the section of the program that checks for the position of the probe to decide whether to perform a measurement or not. Additionally, some basic pre-processing is done with LabVIEW. For instance, normally more than one measurement is acquired per point to both smoothen the spectra by performing an average and also to remove artifacts such as cosmic rays, in the case of RS for example [113][114]. These types of operations are simple and require little processing resources, so LabVIEW can handle these in operations right after a measurement and keep reduced acquisition times. Even though it is possible to easily integrate Python to perform these operations within LabVIEW, it is decided for these to be performed as mentioned for a leaner code structure.

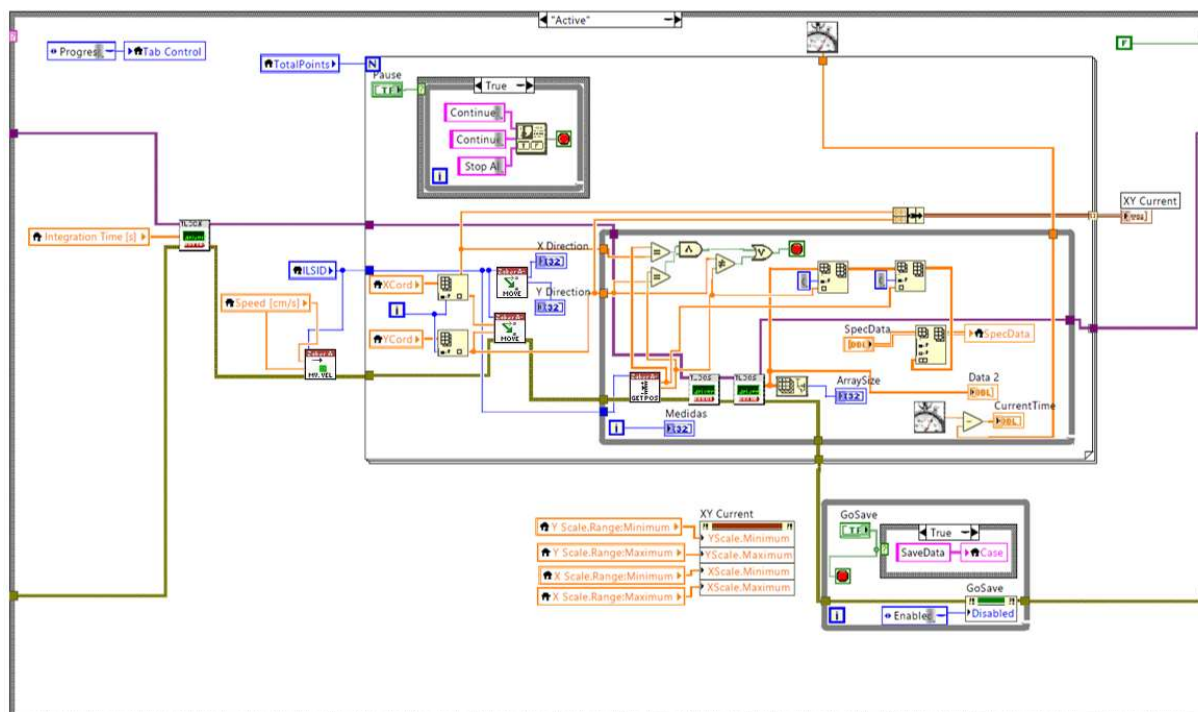


Figure 2-4: LabVIEW block diagram for measurement synchronization of the v1 system. The system constantly checks the location of the probe and performs measurements when they match the defined points by the user.

Then Python is reserved to handle more complex and resource intensive processing after the measurements and before the data is used for analysis and for procedures that need the measurement, or a sub-set of the measurement, to be completed. For example, for RS, pre-processing steps will include axis calibration and baseline removal. The first one requires for a reference measurement to be performed either at the beginning or end of a sample measurement. Both calibration measurements are recommended when the total time is considered too long, since conditions may change significantly during that time. These references contain well-known characteristic peaks, such as Si which shows a peak at 520 cm^{-1} , and the peaks are fitted with a distribution, normally Lorentz distribution in this case. With this, it is possible to more accurately check how the axis is shifted. When two calibration measurements are performed, the average shift can be used or the spectras can individually be corrected with the closest calibration measurement in terms of time. This step normally would require a manual identification and fitting of the peak, but it is automated in just one single line of code using Python and the spectrapepper library (spectrapepper is the third article in this compendium). After, the baseline may be removed. For this, a b-spline is fitted, under the peaks through points in the spectra to remove the baseline below. For this step, no generalized and fully automated way has been found yet in the literature since it shows very specific needs and parameters for each problem. Because of this, the points for the

curve fit must be found manually, but this only needs to be done once, since preserving the code with the parameters will secure repeatability through any other dataset from the same material. This is still true for normalizing to specific peak areas or peak ratios. This procedure can be visualized in Code 2-1 as a functional piece of code. As mentioned, this same code secures consistency and repeatability, as the same “formula” can be used for other measurements of the same sample or other samples of the same material just by changing the data source file. Another example is shown in Annex D, where 196 spectras are processed in just 0.1 seconds.

```
import spectrapepper as spep
# Load the data to be processed and calibrated.
x, y = spep.load('raman_measurement.txt')
# Load the calibration data from Si sample.
x_si, y_si = spep.load('si_calibration.txt')
# This function checks automatically for the peak, fits
# a curve, and extracts the shift of the axis.
x_shift = spep.shiftrf(y_si, x_si, ref_peak=520)
# The shift is added to the measurement axis.
x = x + x_shift
# Remove baseline.
y = spep.bspbaseline(y, x, points=[160, 315, 450, 530])
# Normalize the spectra to the maximum value.
y = spep.normtoratio(y, x, r1=[190, 220], r2=[165, 190])
```

Code 2-1: Example code for processing RS data from data acquired in the LabVIEW system.
Total effective lines of code are seven.

In a first instance, a first version (v1) of the software was used in the first article. This system consisted in a compact CCD spectrometer (Thorlabs CCS200) coupled to a broad emission (400-1000 nm, approximately) halogen light source and a NR probe. The probe was attached to a XY gantry system for the automated movement. Both the spectrometer and the gantry system are connected to the control unit (PC) and programmed with LabVIEW. This had to be done in such a way that the movement of the crane and the spectrometer measurement were fully coordinated, ensuring consistency and repeatability. In this case, the measurements were performed while the crane was in movement, emulating the conditions in a roll-to-roll environment where the CIGS sample are produced. A picture and diagram of this system is presented in Figure 2-5A, and the user interface (UI) is attached in Anex A. An evolved second version (v2) was used for the second article, mainly improved in the ability of performing multi-technique measurements in a quasi-simultaneous way and in the same spots. The diagram and photo of this system is shown in Figure 2-5B, and the UI included in Anex B. The system was also further developed in cooperation with the SEMS members to use a more advanced version in real process-monitoring conditions. The UI of this system is included in Anex C.

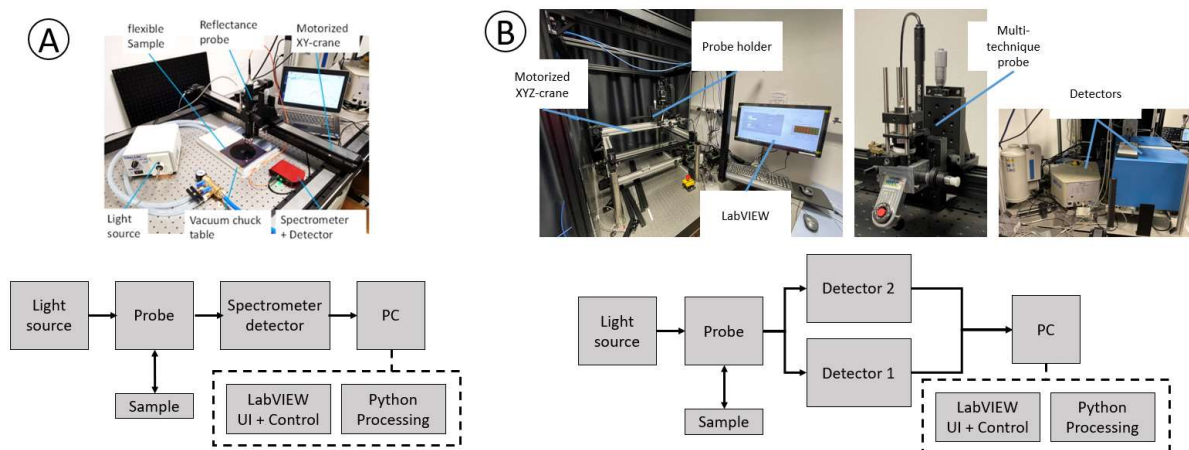


Figure 2-5: Photo and scheme of A) first version of the system used in the article 1 and B) second version of the system used in article 2.

2.4 Data conditioning, fusion, and traceability

Once the data has been acquired, the next step in the methodology is the conditioning and fusion for the utilization in ML algorithms considering the necessity of the data to be traceable. In this process, a critical point is data separation into targets and features. Targets are the properties to predict or classify by the ML algorithm, while features, also known as descriptors, are the variables used to make that prediction. Choosing the correct targets and features to be used in AI assisted methodologies is one of the most critical steps in data analysis, as the input of irrelevant features or inadequate targets will lead to no or confusing results, but a good selection of these will increase the possibilities for a successful experiment with insightful and interpretable results [81][109][115]. In the case of TFPV devices, various targets can be defined such as fabrication parameters, chemical composition of a specific layer, or, more commonly, optoelectronic data of the final solar cell (efficiency, open circuit voltage (V_{OC}), short circuit current (J_{SC}), and fill factor (FF)). In most cases, the target data will be scalars, each associated with a specific sample of a discrete library or a specific area of a graded sample. On the other hand, features can be the results provided by the characterization techniques such as Raman, PL or XRF data as well as the fabrication parameters as described above. This means that features may have heterogeneous data of one- (scalars) or high- (vectors, images) dimensionality, which adds additional complexity on the data treatment related to heterogeneous data fusion. An important remark is that the data selected for features should not be in the targets in the same workflow.

Data pre-processing can also be required for specific data types or measurement techniques. The main objective of data pre-processing is avoiding the introduction of non-relevant features that are not directly related to the sample itself, but rather to the equipment (e.g. instabilities, characteristics of certain components, design limitations, artifacts...) or to the measuring environment

(temperature effects, background illumination and shadows...). This is especially critical when using spectroscopic data in which noise, artifacts, spikes or background signals may add non-relevant information in the spectra. The data arising from each different characterization technique have different pre-processing requirements. For example, in the case of RS it is commonly necessary to calibrate the spectral range and correct peak positions with some reference sample, remove spikes and subtract the baseline. Figure 2-6 shows an illustrative example of spectroscopic data (Raman and PL) before and after pre-processing (and fusing, which is explained below).

The data conditioning process should be completed with a data standardization process, which consists in scaling up the data so that they are numerically comparable among them. For instance, if a scalar data has a maximum value of 5 and is to be fused with a vector that has a maximum value of 5,000, normalization may be necessary for the ML algorithm to accurately consider the scalar feature inside the fused data vector. This could be done by normalizing each technique to its global maximum, or through other methods such as standardizing or normalizing each step of the merged vector from 0 to 1, also known as Min-Max scaling. For instance, standardization transforms the data in the way that it has a mean of zero and a standard deviation of one. It subtracts the mean value of the data and divides by the standard deviation, effectively re-scaling or standardizing the range of the data. This approach assumes that the data follow a Gaussian or normal distribution and scales them accordingly. It maintains the shape of the original distribution and the outliers remain as outliers. On the other hand, Min-Max scaling scales and translates the data within a specified range, typically between 0 and 1. This method subtracts the minimum value from each step of the data series and divides by the range of the data set (i.e., maximum value minus the minimum value). This technique bounds the data but doesn't change their distribution. While it is a simple and common scaling method, Min-Max scaling can be significantly influenced by outliers in the data, causing a majority of the normalized data to be squeezed in a smaller interval. In this regard, data normalization is not a straightforward procedure, and the best option needs to be evaluated case by case to ensure that the original information is not altered or that artifacts do not appear in the process, which could greatly affect the data analysis results.

Once the process of data conditioning is performed, the data of each measured point needs to be fused into a single vector that can be fed into the ML algorithm. In the case of homogeneous data, i.e. when all the data are of the same type, either scalar features or vector features can be joined together in a single vector with higher dimension in a straightforward way. Such a vector then becomes a part of the input file for the ML algorithm, and specific indicators must be created and saved for each of these vectors to keep the traceability of the data. Figure 2-6 shows an example of such a high dimensional spectrum (vector) which is obtained by concatenating Raman and PL spectrum (after their conditioning).

For the publications in this compendium, the above procedure was used, however for a different number of characterization techniques. For the first article, only NF was used for the

characterization technique proposed. These spectra are obtained from the spectrometer on a 0 to 1 scale, and no further processing was necessary other than the subtraction of the bare base material spectra to the deposited AlO_x measurement (as explained in the published article). From there, the dataset was scaled using standardization (as explained above) before being fed to the ML algorithm as a full vector, without extracting any features beforehand. On the other hand, for the second experiment, Raman spectra measured under 442 nm, 532 nm, and 785 nm excitation wavelengths are used for the characterization of the graded sample. Before being introduced as features into the PC-LDA, these were processed separately by wavelength and then merged into a single vector. In this case, the spectras were normalized to the ratios of peaks 172 and 205 cm^{-1} for all wavelengths and then normalized to the global maximum of each wavelength, effectively taking the scale to a global maximum of 1 to each wavelength, to finally merge them together by cell.

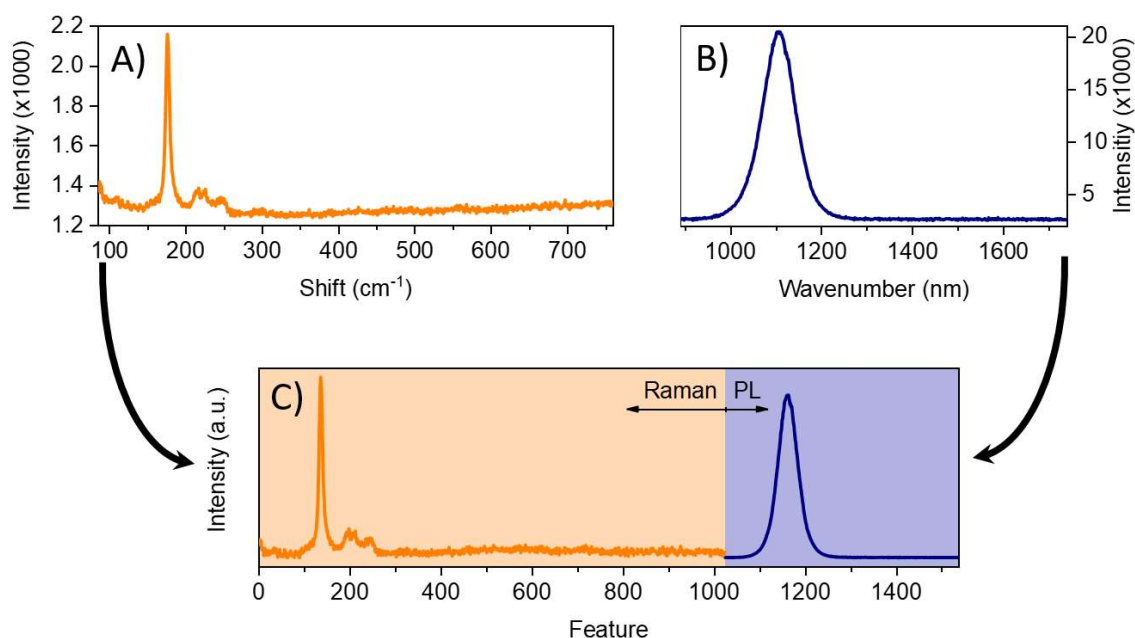


Figure 2-6: Example of a high dimensional spectrum combining Raman and PL spectra for a single measured point. A) shows the raw Raman measurement, B) the PL raw measurement and C) the fused vectors after processing.

2.5 Data analysis

Data analysis is the process of examining and interpreting the information extracted from characterization to gain insights and conclusions about the studied samples. In spectroscopic analysis, this typically involves using specific indicators taken directly from the spectra, like comparing area ratios in different spectral regions, identifying peak positions or inflection points, or measuring peak widths, often through spectral or peak fitting. For instance, RS can reveal a variety of aspects such as crystalline quality, structure type, presence of defects, secondary phases,

or variations in layer thickness. These aspects are subtly included in the spectra through changes in the features of the peaks: position, full width at half maximum (FWHM), absolute or relative intensity, symmetry, etc. However, calculating these properties automatically and without supervision in a generalized way can be challenging due to the data's complexity. For instance, to accurately calculate the areas of peaks in a Raman measurement (for example), a detailed deconvolution of the spectra needs to be performed. This can be illustrated in Figure 2-7, where the full deconvolution of the measurement is performed to extract more accurate area values of the peaks at 176 cm^{-1} and 250 cm^{-1} [116]. Even though this approach yields more accurate results, it implies the deep knowledge of the spectra and the possible peak constitution of the data based on the present materials and structures in the sample. Additionally, this process requires long processing times, as normally a combination of manual and automated processes is needed, along with considerable computational resources for larger data sets. A different approach to perform such task is to use Multivariate Curve Resolution (MCR), which covers several algorithms for mixture analysis in spectroscopic data in an automated way [117]. This is performed by making several peak fittings in known peak locations under an optimization function to minimize the difference between the sum of the fittings and the analyzed spectra. However, these techniques do require to be adapted in detail to each problem, which also requires expert knowledge and familiarity of the possible peaks due to materials and structures present in the studied samples. Even though the recognition of appearing peaks can be approximated (for example using PCA), aiming for a more automated process, it still requires deep knowledge of limitations and restrictions of each problem, which implies larger computational resources needed for each deconvolution, thus making it difficult to implement in industrial processes and high-throughput and big data (BD) experiments.

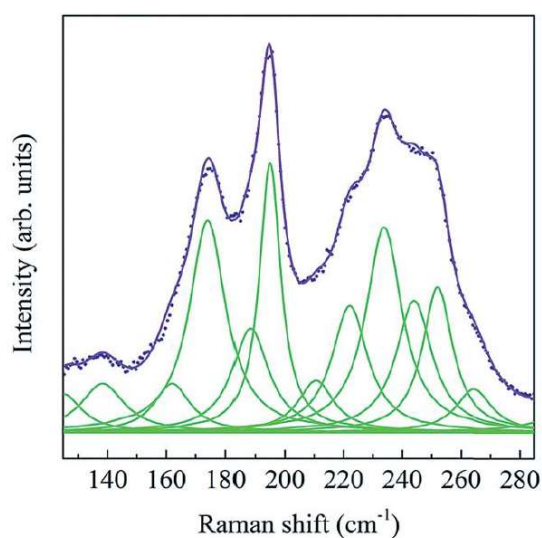


Figure 2-7: Deconvolution of a Raman spectra from a CZTSe sample using a 325 nm excitation source. Extracted from [116].

With this in mind, traditional methods of spectral analysis have several limitations including:

- They are often slow and may require a large human and computational resources, making them inefficient for large-scale studies.
- Expert knowledge and experience are necessary for accurate analysis, limiting accessibility.
- They are not easily adaptable as a universal method, as each experiment needs specific parameter adjustments.
- Analyzing large datasets is challenging because of the significant computational resources required.
- The accuracy of the results heavily relies on precise and careful data processing and conditioning.

To address these issues, the used methodology incorporates ML algorithms, which can handle such data effectively with minimal need for human oversight, while still providing valuable insights. This approach uses dimension reduction algorithms that consider all the available information and simplify it, making it easier to handle and understand. This method leads to more efficient and effective data analysis for TF materials, improving both the speed and the quality of the analysis. Specifically, for the first article PC-LDA is used, meanwhile LDA is used for the second article. These are closely related algorithms, the first one being a cascaded combination of PCA and LDA. PC-LDA, and its separate parts as PCA and LDA, stand out in data analysis and classification for their useful characteristics, making them a reliable choice for high-dimensionality data due to their simplicity, interpretability, and efficiency in handling smaller datasets. These methods, also notable for their computational speed, are often preferred for analysis when resources, such as data and computational power, are limited. Additionally, their susceptibility to overfitting is less in smaller datasets, a common challenge in more complex models, and the PCA aspect of PC-LDA excels in feature extraction and dimensionality reduction, crucial for high-dimensional data handling. In contrast, while algorithms like RF, SVM, QDA, and ANNs can also perform well with high-dimensionality problems, they each have limitations. RF, for instance, is not a dimension reduction algorithm, discarding information in the process that might be important far ahead, and lacks the capability for lower-dimensional visual representations, limiting its use in further validation and physical model development. SVM, although effective, does not maintain the relationship between classes due to its nature of randomization in its initialization. QDA, although closer to PCA and LDA in working principle, tends to exhibit poor test and validation performance and a tendency towards heavy overfitting due to its quadratic nature. ANNs are more difficult to interpret due to their complexity, are less efficient with smaller datasets, are more computational demanding, require to be carefully designed by experienced programmers, and do not offer lower-dimension visualizations. In this regard, it is important to test between these algorithms, specially LDA and PCA as standalones as they share a lot of the benefits of PC-LDA and may offer better

results in specific cases, reason why the first article of this thesis uses PC-LDA and the second uses LDA.

Finally, in order to facilitate the use of the ML by non-experts with a specific focus on spectroscopic data, the third article describes the spectrapepper Python package that includes specific tools to conditioning the spectroscopic data and also visualize ML results from dimension reduction algorithms. Additionally, the fourth article describes the pudu library for gaining insights into the ML results. With respect to the ML algorithms, other open-source libraries have done good work simplifying their application. In particular, the library scikit-learn [118] has been used for the programming of PCA, LDA, and PC-LDA.

2.6 Methodology for spectrapepper library

The Spectrapepper library is built on two key ideas: simplicity and flexibility. Simplicity means it's easy to use, even if the user is not an expert with coding. It uses a straightforward approach, offering functions that stand on their own, without requiring knowledge of classes, methods or other libraries. This makes it more accessible, especially for beginners. The user only deals with functions that take in parameters and return results, all while working with standard Python lists. Flexibility means the library can handle different tasks easily. You can use the functions with one or many spectras without changing the function name or its parameters. It doesn't matter the size or type of the spectras; the functions are designed to work with any spectral data, as long as it's formatted correctly. This correct format means having the data in rows, where each row represents one spectra. This design choice makes Spectrapepper versatile and user-friendly, helping the user to focus on the experimental and analytical challenges. The syntax structure is also standardized across the library. This means that functions have straightforward and intuitive names and consistent parameters. In particular, the first parameter will always be the spectral data, symbolized with a y . The x-axis is always the second parameter but not mandatory, as sometimes it is irrelevant for the operation to be performed (i.e. when normalizing to the max value of y). This syntax is illustrated in Figure 2-8.

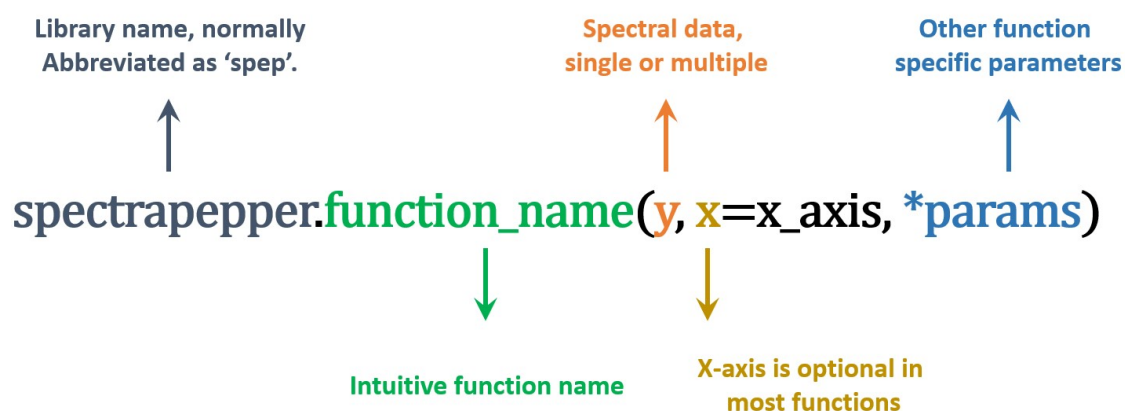


Figure 2-8: Syntax structure that all functions in the library, that accept spectral data as input, follow. The parameters y and x are only for the functions dedicated to spectral processes, which is the main focus of the library. However, there are some functions that can work with any kind of data, but are useful to have in a spectroscopic analysis toolkit (i.e. for the calculation of Spearman and Pearson correlation coefficients)

One of the goals of the library is to tackle all steps in research and industrial processes, offering at least a number of functions for data acquisition, processing, analysis, and visualization, as shown in Table 2-1. For example, for data acquisition, cosmic ray functions are included, which are essential to clean the raw outputs in Raman spectroscopy [114][113]. For data processing, common procedures are included, like baseline removal techniques, noise removal, smoothing, and normalization techniques. For analysis, spectrapepper contains functions for calculating common spectral characteristics, like full width half maximum, areas, averages and standard deviations, and asymmetry. Finally, some functions help to visualize results and the data, for example with stack plots, covariance matrices, and confusion matrices. Some functions may be used in multiple steps, like the “pearson” and “spearman” functions that both calculate the respective correlation coefficients matrices and also plots them.

Table 2-1: Example functions from the spectrapepper library categorized by their main purpose according to the experimental step. Full list of the functions and their explanation can be found in detail in the library’s repository.

Acquisition	Processing	Analysis	Visualization
cosmicdd	bpsbaseline	fwhm	stackplot
cosmicmed	alsbaseline	avg	confusionmatrix
cosmicmp	normtoratio	median	spearman
	normtglobalmax	sdev	pearson
	normtymax	asymmetry	
	lowpass	crosscorrelation	
	moveavg		

2.7 Methodology for the pudu library

Spectroscopy is all about understanding how light interacts with materials, focusing on changes observed in spectral data. This data usually presents itself as peaks that differ in shape, symmetry, intensity, position, and complexity. Even the smallest alterations can indicate significant differences in the material being studied, although sometimes large changes might not be as impactful [119]. Therefore, properly analyzing these variations is crucial. The same principle applies when using ML with spectroscopic data, where detecting changes in spectral features is key. The pudu library, a direct result of this thesis, introduces four methods designed to assess such ML models based on the concept of change, namely importance, speed, synergy, and re-activations. These techniques aim to help scientists delve deeper into their spectroscopic data analysis, extending beyond just the initial ML findings.

Importance: Importance quantifies the relevance of the features according to the changes in the prediction according to defined sequential perturbations on the features. Thus, Importance is measured in probability or target value difference for classification or regression problems, respectively. In a formal way, let $x \in X$ be a 2-D array of dimensions $h \times w$. Let P_M be the probability function of the model M . Then, $P_M(x)$ is the probability of x to belong to a classification class according to the problem solved by M . Considering $j \in J$ the feature in position (h_j, w_j) of x , then the local importance (LI) for said feature j is defined as:

$$LI_j = P_M(x) - P_M(R_j(x)) \quad \text{Eq. 2-1}$$

Where R is a function of local perturbation of feature j . Then, the relative importance (RI) can be denoted as:

$$RI_j = \frac{LI_j - \min(LI)}{\max(LI) - \min(LI)} \quad \text{Eq. 2-2}$$

Where LI contains all the LI_j of sample x . Then, importance is the difference in a model's classification probability according to change in the features.

Speed: Speed quantifies how fast a prediction changes according to perturbations in the features. For this, the Importance is calculated at different perturbation levels, and a line is fitted to the obtained values and the slope is extracted as the Speed value. This is better defined considering states of R with different set parameters R_1, R_2, \dots . As for Importance for x , LI of feature j using the different perturbations would be LI_1, LI_2, \dots . Then, Speed is the slope calculated according to the linear fit of the LI points as $(1, LI_{j,1}), (2, LI_{j,2}), \dots$. Then, the speed is how fast the Importance changes according to change in the feature, or how sensitive it is. These can have positive or negative values, depending on the slope. A positive value means that a bigger change will produce

a bigger change in the prediction. A negative value means that bigger changes produce smaller changes in the prediction.

Synergy: Peaks in spectral data can change at the same time as other peaks. However, their relationship can be difficult to pinpoint and understand, especially in more complex mixtures and materials. Synergy helps to explore these relationships of change by perturbing simultaneously pairs of areas of interest. For this, consider a feature $j^* \in J$ and a distinct feature $j \in J$ from x_i . Both are perturbed under R obtaining $x_{j^*,j}$. Then, the local importance obtained is $LI_{j^*,j}$. Then,

$$LI_{j^*} = (LI_{j^*,1}, LI_{j^*,2}, \dots \forall j \neq j^* \in J) \quad \text{Eq. 2-3}$$

The synergy then indicates how features complement each other in terms of change and the effect on the prediction.

Activations and re-activation: Convolutional Neural Networks (CNN) can result in highly complex structures. As such, understanding how the final form of a CNN relates to the input data can be certainly challenging, but if done correctly can yield great benefits, as shown in [120]. Re-activation attempts to evaluate this structure in terms of change, thus better understanding how spectral characteristics affect the final shape of such networks. To do so, consider the following definitions:

Units: In a convolutional layer $l \in L$, where L is the group of all convolutional layers in the model M , the number of units in K is defined by the size of the input (h, w) , kernel size (k_h, k_w) , strides (s_h, s) and the filters f . Specifically, the number of units can be calculated as:

$$H_0 = (h - k_h)/s_h + 1 \quad \text{Eq. 2-4}$$

$$W_0 = (w - k_w)/s_w + 1 \quad \text{Eq. 2-5}$$

$$units = f * H_0 * W_0 \quad \text{Eq. 2-6}$$

Where (H_0, W_0) are the dimensions of the output of layer l .

Activation map: As defined in [121], for x , take the activation map $A_k(x)$ for each of the units k . Then a_k is the activation distribution for each individual units for $X_s \in X$, where X_s is a subset of all samples X . Then, all the activations belonging to the p quantile as $P(a_k > T_k)$ were T_k is the value above which the quantile exists.

Re-activation map: The above can be evaluated based in feature perturbations considering x , the original data, and x_j , the perturbed input in feature j , and evaluate the difference as $B_k(x_j) - B_k(x) = \Delta B_{k,j} \forall j \in J$, where B is the pre-activation map of unit k . From here we can extract the distribution Δb_k and then pass the data through the activation function to obtain a_k . Finally, select the p quantile as $P(\Delta a_k > T_k) = p$. In this case, X_s is the set of perturbed samples derived from.

The latter accounts then for difference in unit activations after perturbation that would account for a re-activation. For example, if unit k has an activation value of u , and after perturbation the same unit k obtains a value of $u^* = u \rightarrow \Delta u = 0$, then it is not re-activated considering an activation function of ReLU or LeakyReLU. In other words, this looks for significant changes in the activation map according to change, meaning significant a value that would be considered an activation in $A_k(x)$.

With this, it is possible to obtain the following information:

- How many units are re-activated, in units of change
- What feature produces more unit re-activations, per unit of change
- What unit is re-activated the most, per unit of change
- Which feature re-activates what unit the most times

3 PUBLICATIONS

This thesis is structured in the shape of article compendium, where the scientific articles in which this manuscript is founded are collected in this section and will constitute the following subchapters and are presented chronologically as they were submitted. The first two of these articles have been published in high impact factor journals under the titles:

- Thickness evaluation of AlO_x barrier layers for encapsulation of flexible PV modules in industrial environments by normal reflectance and machine learning. *Progress in Photovoltaics: Research and Applications*, 30 (3), 229–239. <https://doi.org/10.1002/PIP.3478>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft.

- Combinatorial and machine learning approaches for the analysis of Cu₂ZnGeSe₄: influence of the off-stoichiometry on defect formation and solar cell performance. *Journal of Materials Chemistry A*, 9 (16), 10466–10476. <https://doi.org/10.1039/d1ta01299a>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Data curation, Formal Analysis, Software, Visualization, Writing – original draft.

The third and fourth articles are open-source and open-access software that have been developed over the past years involved in the program in response to the difficulties found in the literature and the performed research itself. The detailed guide and explanation of these softwares are published as the following peer-reviewed articles:

- spectrapepper: A Python toolbox for advanced analysis of spectroscopic data for materials and devices. *Journal of Open Source Software*, 6 (67), 3781. <https://doi.org/10.21105/joss.03781>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Conceptualization, Data curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft.

- pudu: A Python library for agnostic feature selection and explainability of Machine Learning spectroscopic problems. *Journal of Open Source Software*, 8 (92), 5873. <https://doi.org/10.21105/joss.05873>

Using the Contributor Role Taxonomy CRediT, ETGL work can be described as: Conceptualization, Data curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft.

The first study acts as a proof of concept of the methodology proposed in Section 2 of the thesis, demonstrating the efficacy of CA in conjunction with ML for spectroscopy in TF research and as a valuable tool for monitoring production processes in industrial environment and using an early version of the automated system. The first functions included in the spectrapepper library were coded and used during this experiment. The second publication also follows the same methodology, with ML and CA, focusing on research objectives aimed at enhancing understanding and gaining deeper insights into material properties. This experiment, however, uses more spectroscopic measurements than the first, further showing the capabilities and versatility of the methodology. The third work is the open-access and open-source library spectrapepper, containing all the functions and procedures used for the spectroscopic processing and analysis performed for the first two articles and for the subsequent work, allowing for a seamless and simple integration of AI methodologies for HTE in IREC, from data acquisition to data analysis. Finally, the fourth article is the open-access and open-source library pudu, which deals with explainability and interpretability for the results from ML models, allowing for the extraction of deeper insights of such results from the performed experiments. This library was published and developed after the publication of the first three articles, but it naturally appeared as a necessity and logical consequence of the methodology, since further and better interpretation of the ML results were needed to fully take advantage of CA and ML analysis. However, this library is used in the posterior exploratory experiments (Section 4), showcasing its potential and usefulness.

3.1 Publication 1



Received: 17 June 2021 | Accepted: 12 September 2021

DOI: 10.1002/pip.3478

RESEARCH ARTICLE



Thickness evaluation of AlO_x barrier layers for encapsulation of flexible PV modules in industrial environments by normal reflectance and machine learning

Enric Grau-Luque¹ | Maxim Guc¹ | Ignacio Becerril-Romero¹ |
Víctor Izquierdo-Roca¹ | Alejandro Pérez-Rodríguez^{1,2} | Pieter Bolt³ |
Fieke Van den Bruele⁴ | Ulfert Ruhle⁵

¹Catalonia Institute for Energy Research (IREC), Barcelona, Spain

²IN²UB, Departament Enginyeria Electrònica i Biomèdica, Universitat de Barcelona, Barcelona, Spain

³TNO, Department of Solar Technology and Applications, High Tech Campus, Eindhoven, The Netherlands

⁴TNO, Holst Centre, High Tech Campus, Eindhoven, The Netherlands

⁵Filison AG, Niederhasli, Switzerland

Correspondence

Ignacio Becerril-Romero and Víctor Izquierdo-Roca, Catalonia Institute for Energy Research (IREC), Jardins de les Dones de Negre 1, 08930 Sant Adrià del Besòs, Barcelona, Spain.
Email: ibecerril@irec.cat and vizquierdo@irec.cat

Funding information

ACCIÓ; European Commission; European Regional Development Fund; Ministerio de Ciencia e Innovación; Generalitat de Catalunya, Grant/Award Number: TECSPR18-1-0048; University of Barcelona; European Union, Grant/Award Number: H2020-LCE-2017-RES-IA

Abstract

Flexible photovoltaic (PV) devices, such as those based on Cu (In,Ga)Se₂ (CIGS) and perovskites, use polymeric front sheets for encapsulation that do not provide sufficient protection against the environment. The addition of nanometric Al_xO layers by spatial atomic layer deposition (S-ALD) to these polymeric materials can highly improve environmental protection due to their low water vapor transmission rate and is a suitable solution to be applied in roll-to-roll industrial production lines. A precise control of the thickness of the AlO_x layers is crucial to ensure an effective water barrier performance. However, current thickness evaluation methods of such nanometric layers are costly and complex to incorporate in industrial environments. In this context, the present work describes and demonstrates a novel characterization methodology based on normal reflectance measurements and either on control parameter-based calibration curves or machine learning algorithms that enable a precise, low-cost, and scalable assessment of the thickness of AlO_x nanometric layers. In particular, the proposed methodology is applied for precisely determining the thickness AlO_x nanolayers deposited on three different substrates relevant for the PV industry: monocrystalline Si, Cu (In,Ga)Se₂ multistack flexible modules, and polyethylene terephthalate (PET) flexible encapsulation foil. The proposed methodology demonstrates a sensitivity <10 nm and acquisition times ≤100 ms which makes it compatible with industrial monitoring applications. Additionally, a specific design for in-line integration of a normal reflectance system into a roll-to-roll production line for thickness control of nanometric layers is defined and proposed.

KEYWORDS

AlO_x, CIGS, encapsulation, flexible PV, machine learning, normal reflectance, process monitoring, thickness assessment

1 | INTRODUCTION

Light weight and flexible photovoltaic (PV) modules fully exploit the technological capabilities of thin film PVs since; besides their inherent

advantages such as reduced fragility and adaptation to curved surfaces that open the way to numerous applications, they can be fabricated through high throughput roll-to-roll (RtR) processes. This type of intensive production reduces both the economic and energy

costs of PV, leading to devices with an increased energy return on energy invested (EROI) ratio, key for the expansion of solar energy.¹ A critical step in the fabrication of flexible PV modules is the implementation of a suitable encapsulation architecture. Contrarily to the standard rigid modules that typically employ glass sheets for this purpose, the encapsulation of flexible PV modules, like those based on, for example, Cu (In,Ga)Se₂ (CIGS) or perovskite absorber materials, relies on the use of flexible transparent polymeric front sheets.² These commonly present reduced water vapor barrier properties and require additional protective layers to ensure a proper environmental protection and long-term preservation of the modules.^{2,3} In this regard, AlO_x-based nanolayers deposited on the polymeric front sheets by atomic layer deposition (ALD) have been demonstrated to possess a high conformality and compactness which confer them a very low water vapor transmission rate.^{4–9} However, conventional ALD is an extremely slow deposition technique (200 nm/h in RTR configuration to achieve 25- to 30-nm thick layer)⁴ incompatible with industrial high throughput processing. On the other hand, spatial ALD (S-ALD) is a technological alternative to standard ALD in which, in simple terms, the samples move between spatially separated half-reaction zones where the deposition takes place. As a consequence, the deposition rate is mainly limited by the amount of deposition areas and the time required to move the samples between them achieving deposition rates ~1 nm/s which are fully compatible with RTR web speeds.^{10,11} In the case of CIGS, on which this work is focused, it has also been demonstrated that, in addition to the employment of AlO_x-deposited polymeric front sheets, AlO_x nanolayers can also be directly applied on the top electrode of the devices and still provide high end barrier properties.⁵ In this context, S-ALD-deposited AlO_x barriers break ground to low-cost and RTR-compatible front sheet solutions that represent a step forward for the production flexible PV devices. Furthermore, the use of AlO_x nanolayers in PV is not limited to encapsulation and AlO_x layers have also been successfully employed for interface engineering and passivation in thin films solar cells and other devices^{12–16}.

Like with most applications in which nanocoatings are employed, a precise thickness control of the applied AlO_x nanolayers is crucial for them to function properly and provide the desired effect. In the case of flexible PV module encapsulation, the water vapor barrier properties of AlO_x nanolayers have been reported to drop dramatically when their thickness falls below 10 nm,⁸ while their brittleness increases with layer thickness making it more prone to fracture under bending stresses on flexible substrates.¹⁷ As such, a precise thickness control is required to ensure an adequate functionality of the AlO_x layers for their use as water vapor barriers in flexible PV devices. Likewise, precise thickness control is also critical for other uses of AlO_x layers like interface passivation¹⁸. In this context, the development of methodologies and tools that can be implemented at RTR lines for in-line process monitoring represents a strategic technological advance for improving and optimizing PV module production at mass scale.

The reduced thickness of nanocoatings commonly requires the use of very specific characterization techniques like those based on X-ray photoelectron spectroscopy,^{19,20} atomic force and

electron microscopy,^{21,22} Rutherford backscattering spectroscopy,²³ ellipsometry,^{4,13,15,16,18,24–26} or transmittance-reflectance spectroscopy,^{8,27,28} among others. These techniques require long acquisition times, sample destruction, and/or high-energy (deep UV or X-ray) excitation wavelengths. Moreover, some of the mentioned techniques suffer from limitations related to the characteristics of the nanocoated substrate (high roughness and/or multistack configuration, sensitivity to high excitation energies, etc.) and/or cannot be applied to large area analyses. All these issues make the implementation of the existing methods in high throughput production lines very technically challenging and economically costly.

Another critical issue regarding the scaling up of the production of flexible PV devices is layer homogeneity. In fact, homogeneity is usually considered one of the most important barriers for the transfer of CIGS PV devices from the laboratory to the industry level.²⁹ In this regard, the implementation of techniques that allow performing fast large area mappings to monitor layer homogeneity is also of great relevance for the industry.

In this framework, this work describes and demonstrates a novel nondestructive, fast, precise, low-cost, and scalable characterization method for determining the thickness of AlO_x nanometric layers (from a few to around 100 nm) that is compatible with both research and industrial process monitoring environments. This approach is based on normal reflectance measurements and takes advantage of the impact of quantum confinement (QC) effects on the optical properties of AlO_x nanolayers as a consequence of their nanometric thickness: optical bandgap (E_g) and optical constants (n and k).³⁰ The viability and effectiveness of the technique are demonstrated for precisely determining the thickness of AlO_x nanometric layers deposited by S-ALD on different relevant materials for the PV industry with different characteristics: monocrystalline Si wafers (smooth), a complete CIGS module (multistack substrate), and polyethylene terephthalate (PET) polymer encapsulating sheets (rough). We demonstrate that the slight QC-induced changes in the optical properties of the AlO_x layers can be detected from normal reflectance measurements with a simple setup and be used for layer thickness determination with resolutions better than 10 nm through their combination with machine learning algorithms. The compatibility of the proposed approach with fast micro (tenths of microns) and macro (up to several m²) mapping analyses, depending on the application, as well as the possibility of modifying the system employed, for better compatibility with in-line industrial process monitoring, is analyzed and discussed.

2 | EXPERIMENTAL

2.1 | Data acquisition and sample description

Figure 1 shows the normal reflectance probe that was implemented in this work and used for the evaluation of the thickness of AlO_x nanolayers with a broad emission (400–1000 nm, approximately) halogen lamp as illumination source. The probe was coupled to an XY crane to enable mapping measurements, and the acquired signal was

FIGURE 1 (A) Schematic of the normal reflectance probe based in a broad emission halogen lamp and (B) picture of the implemented system

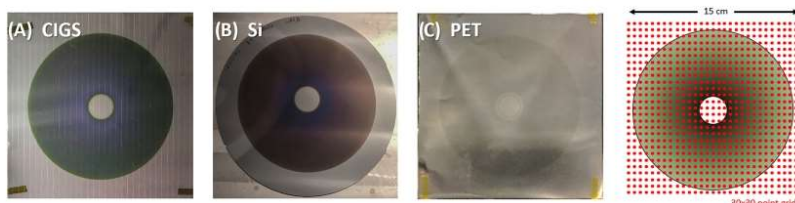
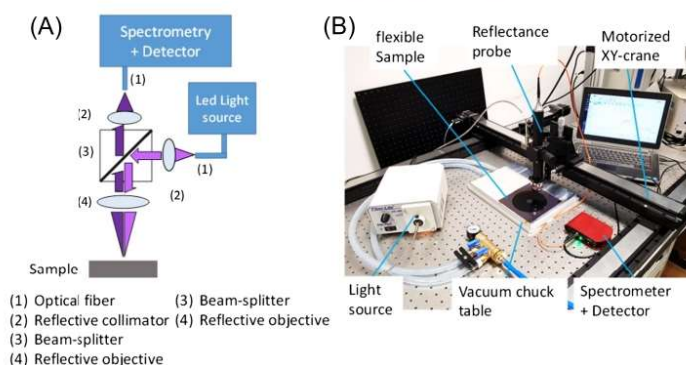


FIGURE 2 Picture of (A) the Cu(In,Ga)Se₂ (CIGS)-device, (B) Si, and (C) polyethylene terephthalate (PET) samples coated with a 75-nm AlO_x layer. The last is the approximate schematic of the samples and measuring points. Inner and outer radii of the deposition area are 1.2 and 7.6 cm, respectively

processed through a compact CCD spectrometer (Thorlabs CCS200). A vacuum chuck table was employed to ensure the flatness of flexible samples during measuring. A spot size of $\sim 100\ \mu\text{m}$ and acquisition times in the 10- to 100-ms range (depending on the type of sample analyzed) were employed for the measurements.

The system described above was employed on different sets of samples for the evaluation of the thickness of nanometric AlO_x layers deposited by means of a laboratory-scale rotary spatial ALD reactor³¹ on relevant substrates for the PV industry: (i) three monocrystalline Si substrates coated with 25-, 50-, and 75-nm AlO_x layers; (ii) six $15 \times 15\text{-cm}^2$ complete (nonencapsulated) flexible Cu(In,Ga)Se₂ thin film PV devices (on polyimide foil substrate with Mo back electrode, CdS buffer layer and Al-doped zinc oxide front electrode) with 15-, 25-, 30-, 50-, 60-, and 75-nm AlO_x layers (referred to as "CIGS" in the text and figures); and (iii) six $15 \times 15\text{-cm}^2$ PET foil samples with 15-, 25-, 30-, 50-, 60-, and 75-nm AlO_x layers. It should be noted that these thicknesses should be taken only as nominal deposition values. The AlO_x layers were deposited using trimethylaluminum (TMA) and water as precursors for aluminum and oxygen, respectively. Layers were deposited at 100°C, at 30 rpm, with 50 sccm TMA/950 sccm dilution and 750sccm H₂O at 50°C/750 sccm dilution. The coated area had a donut shape (see Figure 2A–C) with slight thickness variations along the radial direction (decreasing from the center to the edges) which allowed to test the sensitivity of the proposed methodology. For the Si and CIGS samples, the normal reflectance

measurements were carried out in a mapping configuration (30×30 measuring points grid, approximately) covering the whole area of the samples (see Figure 2D). In the case of the PET samples, a special low-reflectance (<5% in the 300- to 700-nm range) holder had to be employed due to the high transparency of this material to the excitation wavelengths used for the analysis. In addition, the use of such holder prevented the use of the XY crane for the acquisition of large area mappings and the measurements were performed point-by-point, manually (15 points per sample).

Additional measurements were carried out on the Si-based samples along the radial direction (9 points per sample) by means of ellipsometry (Horiba Jobin Yvon, Uvisel) to corroborate that the signal obtained from the surface of the substrate material varied as a consequence of the change of the optical properties of AlO_x with layer thickness. The measurements were made with a 70° angle of incidence.

2.2 | Methodology

In order to quantify the differences in the normal reflectance spectra and translate them into thickness data, the following control parameter (q_t) was defined:

$$q_t = \sum_x |A_s(x) - A_{ref}(x)| \quad (1)$$

where $A_s(x)$ is the integrated intensity of the normal reflectance spectra of the AlO_x layer in a selected spectral range x (400–900 nm in the case of the halogen lamp and 600–700 nm for the 660 nm LED) for a specific measuring point and $A_{ref}(x)$ is the average integrated intensity in the same range of all the measuring points corresponding to the uncoated base (substrate). The use of q_t reduces the impact of potential sample-to-sample fluctuations in the normal reflectance spectra as a consequence of changes in the reflectivity of the base material (substrate) not related to the AlO_x layers. In this way, the use of this parameter allows improving the accuracy of the AlO_x thickness evaluation.

In order to provide a useful methodology for the thickness quantification, the q_t parameter was calculated for each measuring point and, then, all the q_t values obtained were averaged for each sample and plotted versus the nominal deposition thickness for each type of substrate/ AlO_x sample. Through the fitting of such data, calibration curves were subsequently calculated to show the potential of the proposed methodology for predicting the thickness of an AlO_x layer from normal reflectance measurements.

Alternatively, a machine learning-driven methodology based on the combination of principal component analysis (PCA) and linear discriminant analysis (LDA) algorithms was employed to quantify the normal reflectance spectra and translate them into thickness data. PCA and LDA are both dimension-reduction algorithms, and their combination was selected due to its wide-spread use for spectral data analysis in different methods and fields of application^{32–37}. The goal of this algorithm is to learn to classify data into distinct groups defined by the user and make predictions on new input data. In order to test and implement the machine learning-based PCA-LDA algorithm, the Python programming environment³⁸ with the Scikit-Learn library³⁹ was used. All the experimental data were randomly divided in 70% for training and 30% for testing, and the input features were the same as those used to calculate the q_t parameter using Equation 1. In the case of the in-sample analysis, the data points were divided in five groups corresponding to rings in the sample with rings 1 and 5 representing the outer and inner extremes, respectively (see Figure S1). To evaluate the performance of the algorithm, training and test scores are used. These values are calculated as the number of correctly classified points divided by the total amount; thus, these values range from 0 to 1, being 0 no correct classifications at all and 1 when all points are correctly classified. The training scores were above 0.8 for the halogen lamp and above 0.6 for the LED light source (see Figures 7 and 8).

3 | RESULTS

3.1 | Initial validation: Si/ AlO_x samples

Before employing normal reflectance, preliminary ellipsometry measurements were carried out on Si/ AlO_x samples with three different nominal layer thicknesses (25, 50, and 75 nm) and slight in-sample radial thickness gradient (see Section 2 for further details) in order to

corroborate that the optical properties of the samples change with the thickness of the AlO_x layer. The results are shown in Figure 3B. It can be observed that the Ψ and Δ angles of the complex reflectance ratio present slight in-sample changes along the radial direction within the different points measured (see Figure 3A) and abrupt sample-to-sample differences as a consequence of the varying thicknesses. These changes consist mainly of a blue-shift of the spectra as the AlO_x layer thickness is reduced. It should be noted that these variations represent the change of the complex reflectance of the Si-air structure as a consequence of the presence of an intermediate AlO_x layer with varying thicknesses. Since the optical properties of a layered structure are strongly intertwined with and can be derived from its complex reflectance ratio (using an adequate model), it can be concluded that the variation of the latter indicates a change in the optical properties of AlO_x . Further analysis of the ellipsometry measurements to determine which specific properties are modified when the thickness of the AlO_x layers is varied is beyond the scope of this work. However, taking into account that the deposition conditions were identical for the three Si/ AlO_x samples, the observed thickness-induced blue-shift can be attributed to QC effects.^{40–42} However, other effects different to QC that may be contributing to the changes observed in the ellipsometry measurements cannot be discarded. The results that will be presented throughout this work are based on the correlation of these slight changes of the optical properties of AlO_x with normal reflectance measurements.

In this way, the Si/ AlO_x samples were subsequently measured in similar positions along the radial direction with the normal reflectance system described in Section 2 (see Figure 1) in order to make an initial validation of the technique to detect the small variations observed by ellipsometry. Figure 3C shows the raw normal reflectance data acquired on the Si/ AlO_x samples. A similar shape can be observed for the spectra measured at different points along the radial direction of each sample with a broad band having the maximum position in the 620- to 720-nm region, approximately. However, the intensity of this band is observed to vary from the center (lower) to the edge (higher) of the sample in consonance with the thickness gradient of the deposited layers. The maximum intensity is reached for the bare uncoated Si substrate. On the other hand, clear sample-to-sample differences can be spotted not only as abrupt differences in the intensity of the reflectance spectra (with higher thickness leading the lower intensity) but also as a different shape of the reflectance band (with a faster decrease of the short wavelength part of the band as layer thickness increases). In addition, a higher in-sample variability is found as the nominal thickness of the deposited AlO_x layer increases. All these variations correlate well with those observed in the ellipsometry measurements (Figure 3B) and indicate that normal reflectance is sensitive both to AlO_x thickness variations in the order of tens of nm (sample-to-sample) and also to the slight nanometric variations (expected to be well below 10 nm) found within the layers.

Once that the sensitivity of the normal reflectance technique was proved to be high enough to detect small thickness variations, the proposed methodology was applied to the same Si samples coated with 25-, 50-, and 75-nm AlO_x layers but carrying out a mapping

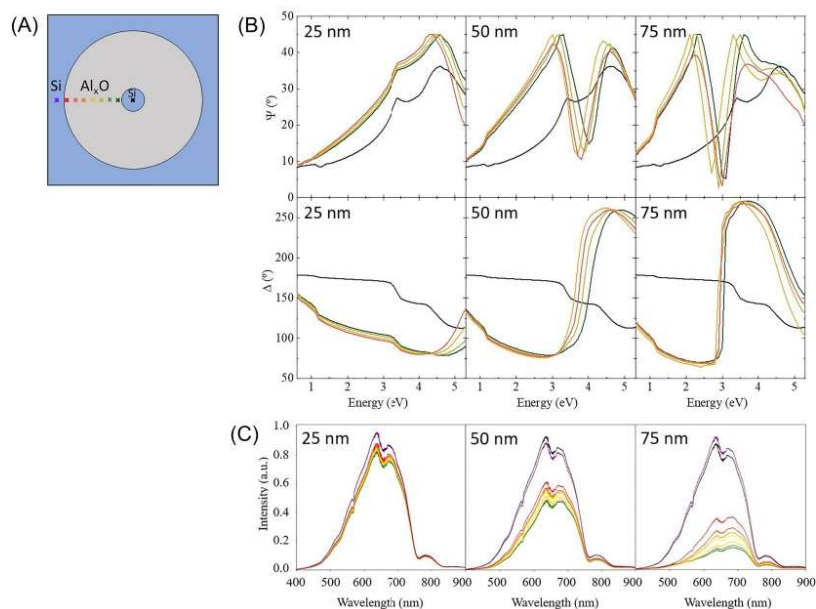


FIGURE 3 (A) Color code legend of the different measuring points. (B) Complex reflectance ratio angles of the Si/ AlO_x samples measured by ellipsometry in different positions along the radial direction. (C) Raw normal reflectance spectra of the Si/ AlO_x samples measured in similar positions

analysis covering the full surface of the sample (Figure 4A). The raw data (Figure 4A, left) show three different groups of spectra that have been highlighted with different colors in the figure. The groups with the highest (purple) and lowest (gray) intensities correspond to the Si substrate and to the measuring table, respectively. The last group (green-yellow gradient) encompasses all the AlO_x -coated points. As expected, the AlO_x spectra have the same characteristics as those shown in Figure 3, that is, similar shape, gradual in-sample variations, and more abrupt sample-to-sample differences with changes in the reflectance band intensity and shape, with higher variability in the thicker samples. In order to quantify the differences in the normal reflectance spectra and translate them into thickness data with high sensitivity, the q_t parameter was calculated for each point measured in every sample using Equation 1. The obtained values are represented in the form of mappings in Figure 4A (right) where the bare Si, measuring table, and donut-shaped AlO_x areas can be clearly distinguished. Furthermore, differences in q_t can be observed in the radial direction confirming the sensitivity of the technique to slight thickness variations below 10 nm. On the other hand, the mappings further confirm that in-sample thickness variability increases with the nominal thickness of the deposited AlO_x layer, with distribution limits for the q_t parameter roughly ranging from 50–100, 250–350, and 400–550 for the 25-, 50-, and 75-nm samples, respectively. This indicates that the developed methodology can also be employed for high-resolution large area homogeneity control of deposited AlO_x layers.

The data presented in Figure 4A were employed to calculate the average q_t in the donut-shaped AlO_x -covered areas. By plotting the average q_t versus the nominal thickness of the AlO_x layers, a calibration curve is obtained for correlating the reflectance data with the thickness of the AlO_x layers (Figure 5A). It can be observed that there appears to exist a quadratic relationship between the nominal thickness of the AlO_x layers and the q_t parameter. As such, these data demonstrate that normal reflectance offers a feasible and precise method for thickness assessment of AlO_x layers deposited on monocrystalline Si substrate.

3.2 | Application of normal reflectance to AlO_x thickness estimation on CIGS and PET

Once that the proposed methodology was demonstrated for Si-based samples, it was applied to AlO_x layers (15, 25, 30, 50, 60, and 75 nm) deposited on relevant substrates regarding the encapsulation of thin film flexible PV devices: CIGS and PET.

In the case of the CIGS modules, a similar mapping to that described for the Si samples was performed. Three representative cases (25, 50, and 75 nm) are shown in Figure 4B, while the complete results for all the different samples can be consulted in Figure S2. The raw spectra shown in the figures reveal a high degree of similarity to those obtained on Si samples with a broad band having the maximum

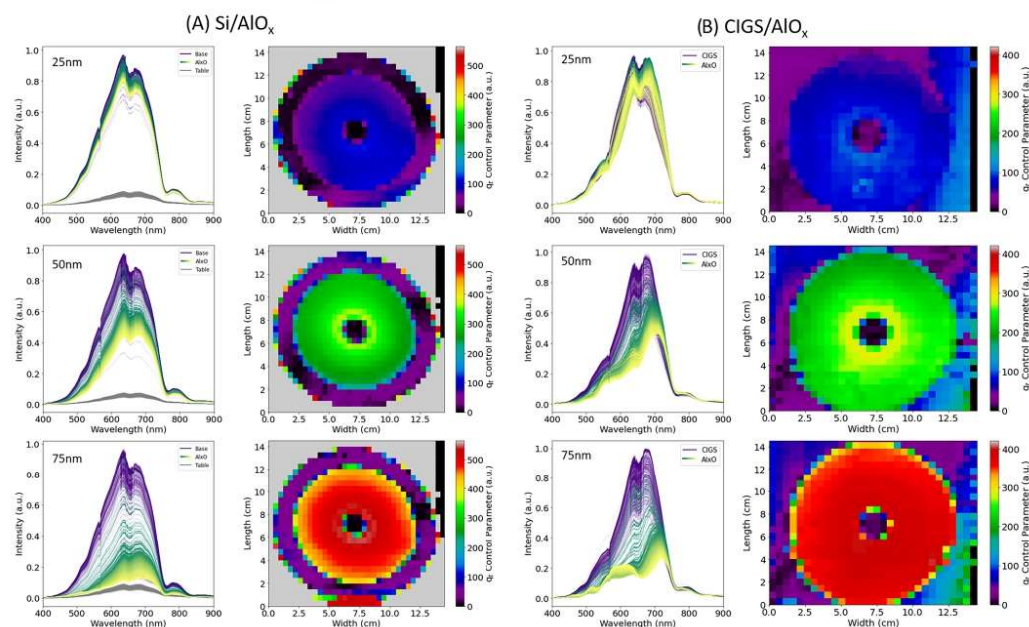


FIGURE 4 Raw reflectance spectra (left) and mappings of the calculated q_t parameter (right) for Si/AIO_x (A) and Cu(In,Ga)Se₂ (CIGS)/AIO_x (B) samples

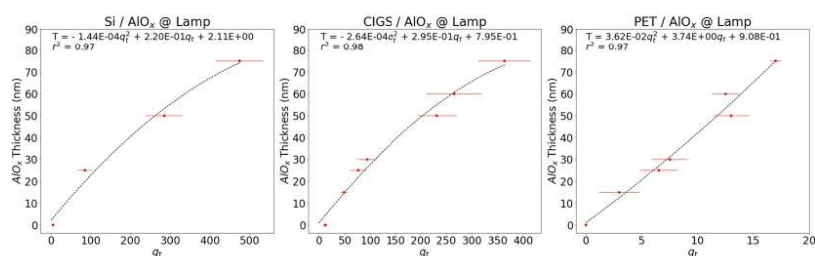


FIGURE 5 Average q_t (red dots) and fitted calibration curves (dashed curves) obtained for AIO_x deposited on Si (A), Cu(In,Ga)Se₂ (CIGS) (B), and polyethylene terephthalate (PET) (C). The error bars correspond to the standard deviation (σ) of the different measured points

position in the 620- to 720-nm region and with similar in-sample and sample-to-sample variability characteristics. However, a higher variability of the reflectance signal arising from the base (CIGS substrate) is detected for thicknesses ≥ 50 nm, as well as a lower deviation for the thin AIO_x coatings ≤ 30 nm. This is detected despite the fact that for low AIO_x thickness values (≤ 30 nm), the intensity and shape of the spectra from the base and the AIO_x-covered areas present similar values. However, the q_t mappings show that the use of this parameter allows overcoming this issue and distinguishing the bare substrates from the donut-shaped AIO_x-covered areas, showing to be sensitive

to even the small in-sample AIO_x thickness changes. The strong variations detected in the signal of the base material are related to local inhomogeneities of the CIGS absorber and/or CdS buffer layers of the devices, which are the main layers that reflect the excitation light employed (400–900 nm). The possibilities to overcome this obstacle will be discussed later on.

In the case of PET, the need of using a special low-reflectance holder prevented carrying out large area mappings. The spectra acquired are shown in Figure S3. Again, the spectra present similar characteristics as those of Si and CIGS samples. However, due to the

low reflectivity of the base material (PET), almost no shape changes of the reflection band are detected which indicates that the sensitivity of the proposed methodology is lower for this material. Despite the lower sensitivity, the changes of the band intensity are marked enough to allow detecting differences between the different samples.

The average q_i and calibration curves obtained for AlO_x deposited on CIGS and PET are shown in Figure 5B,C, respectively. These data demonstrate that normal reflectance can also be employed for determining the thickness of AlO_x encapsulation barrier layers deposited on multistack CIGS and rough PET substrates with high accuracy.

3.3 | Implementation of machine learning algorithms: CIGS test case

Although the use of the q_i control parameter and calibration curves is a perfectly suitable methodology for obtaining precise thickness measurements as demonstrated above, the amount of data generated in the large area mapping measurements carried out in this work provides an ideal test environment for the implementation of more advanced analysis techniques based on machine learning algorithms that could be suitable for analysis in industrial applications. As a proof of concept, a machine learning algorithm based on PCA-LDA (see more details in Section 2) was applied to the experimental data obtained in the present study for the CIGS samples. These samples were selected as the most relevant case for the present study since large area mappings were performed onto them, and they present a rough surface that makes thickness estimation challenging by other techniques. Both the capacity to detect sample-to-sample and in-sample variations were tested. The results are shown in Figure 6.

Regarding the sample-to-sample analysis, it can be observed that the PCA-LDA algorithm enables a clear classification of the samples with different nominal thicknesses yielding a test score very close to 1 (Figure 6A). This is somehow remarkable taking into account that, as already mentioned, all the samples present a slight thickness radial gradient which inevitably produces a broadening in the classification groups complicating classification. For example, it is interesting to note that although some overlapping is clearly observed for the 50- and 60-nm samples, the algorithm is still capable of correctly classifying the points according to AlO_x nominal layer thickness.

As for the in-sample variability, the 75-nm AlO_x sample was analyzed by dividing the data points in groups corresponding to five

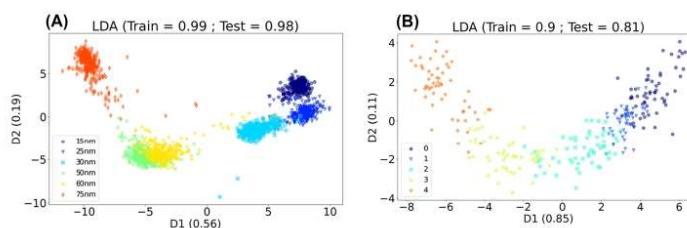
rings in the sample with rings 1 and 5 representing the outer and inner extremes, respectively (see Figure S1). As shown in Figure 6B, the algorithm enables to effectively classify the points by the different thickness ring to which they belong proving, again, the high sensitivity of the proposed normal reflectance methodology to thickness variations <10 nm. Although some overlapping can be observed between the different groups, this is mostly due to the thickness grading existent within the rings.

These results prove the feasibility of employing a machine learning-driven analysis coupled to the proposed normal reflectance approach as a powerful alternative to that based on calibration curves for process monitoring in an industrial environment enabling measuring nanocoating thicknesses with sensitivities <10 nm. Furthermore, it should be taken into account that the amount of data employed for the analysis (~ 2640 spectra) is far from being considered "big data" and, as such, that this methodology can be further improved for a more precise classification using a higher number of training inputs.

4 | DISCUSSION: IMPLEMENTING NORMAL REFLECTANCE FOR INDUSTRIAL IN-LINE PROCESS MONITORING

Although normal reflectance has previously been employed for thin film thickness evaluation at research^{43–45} and process monitoring²⁷ levels, this was done through the use of complex and expensive systems, long acquisition times, and models based on previous knowledge on the optical properties of the material and/or for thicknesses larger than those used in the applications described here. In this way, the results presented in this work represent a laboratory proof of concept of a completely different and innovative approach that demonstrates the feasibility of directly employing normal reflectance data for a precise and fast determination of the thickness of nanometric AlO_x layers deposited on Si, CIGS, and PET employing a simple and inexpensive system. In this regard, it should be noted that the main novelty of this work lies in the fact that the methodology proposed does not require to understand the physical meaning of the differences observed in the normal reflectance spectra but only to be able to detect these differences and correlate them to the thickness of the AlO_x nanometric layers. However, the main focus of this methodology is its implementation in industrial environments, especially for process monitoring of AlO_x barrier layers deposited by

FIGURE 6 PCA-LDA sample-to-sample (A) and in-sample (B) thickness classification results for Cu(In,Ga)Se₂ (CIGS) samples. The in-sample variation analysis was performed on the 75-nm AlO_x sample with the data grouped by rings in the radial direction (see Figure S1)



S-ALD in RTR configuration. In this context, several aspects must be considered in order to make the normal reflectance methodology more industrially-friendly.

The first aspect that should be considered is the substitution of the excitation light source based on a standard halogen lamp by a more stable, maintenance-free, versatile, and low-cost one. In this regard, LED-based light sources are more appropriate for the industry. The feasibility of employing a monochromatic LED light source (660 nm) was tested for the Si and CIGS samples presented above. Taking a look at the calibration curves obtained under LED illumination (Figure 7), it can be seen that a very high dispersion is obtained for the q_t parameter in the case of the Si base. As for CIGS, except for the 25- and 30-nm samples which seem undistinguishable due to their similar q_t values, the results show that the LED excitation source works in a fairly similar fashion as the halogen lamp.

Additionally, the same machine learning-based analysis performed for the halogen lamp measurements in the CIGS samples was applied to the spectral data acquired with the LED light source (Figure 8). Regarding sample-to-sample analysis (Figure 8A), the test score for point classification is significantly lower than in the case of the halogen lamp. It can be observed that this is due to the fact that the 15- to 30-nm and 50- to 60-nm samples present a high overlapping leading to misclassification of the groups. This is in accordance with the results obtained with the calibration curve presented in Figure 7 for these samples. Similarly, the in-sample classification (Figure 8B) also presents a lower score and higher overlapping than in the case of the halogen lamp.

All these dispersion/overlapping issues, though, can be resolved by tailoring the LED excitation employed to the characteristics of the

sample to maximize the reflectance signal. Furthermore, the use of several multiplexed LED sources with different wavelengths would open the way to further optimization of the signal acquisition. Either way, the results shown in Figures 7 and 8 indicate that the methodologies presented in this work are versatile in terms of the possibility of employing different illumination sources tailored to the characteristics of the material to be analyzed, enabling the optimization of the system in a simple way.

Another critical aspect that should be taken into account for industrial implementation of a normal reflection-based monitoring tool is the methodology employed for calculating q_t . In this work, $A_{ref}(x)$ was defined in Equation 1 as the average integrated intensity of all the spectra corresponding to the base material (see Section 2 for further details). It was defined in such manner because the samples were analyzed only after AlO_x deposition. Although the use of this parameter has been shown to be critical for obtaining measurements with high precision, the high variability of the reflectance signal throughout the different points of the base material is one of the reasons why the measurements present a high dispersion hindering the differentiation of AlO_x layers with low thicknesses from the bare substrate. Nevertheless, in an industrial process monitoring environment, two optical probes located before and after the AlO_x deposition process and synchronized to measure in the exact same position would allow calculating $A_{ref}(x)$ for each measuring point (instead of using an average value) minimizing the signal fluctuations related to the inhomogeneities of the base material and improving the accuracy and reliability of the methodology. Figure 9 schematically depicts the design of such a system. On the other hand, it should be taken into account that besides the dispersion introduced by the inhomogeneity

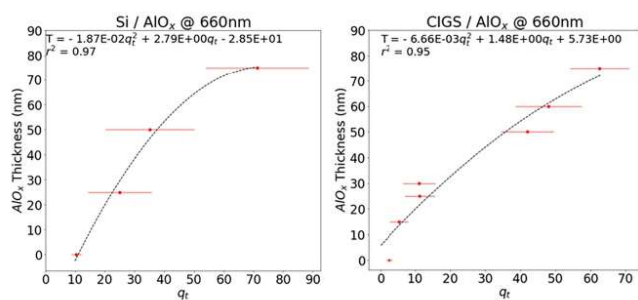


FIGURE 7 Example of reflectance calibration curves obtained with a 660-nm LED source for Si/ AlO_x (A) and Cu (In,Ga) Se_2 (CIGS)/ AlO_x samples (B)

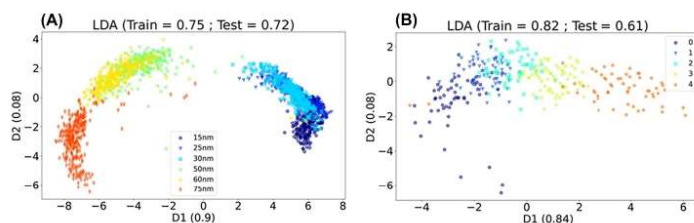


FIGURE 8 PCA-LDA sample-to-sample (A) and in-sample (B) thickness classification results for Cu (In,Ga) Se_2 (CIGS) samples measured with an LED light source. The in-sample variation analysis was performed on the 75 nm of AlO_x sample with the data grouped by rings in the radial direction (see Figure S1)

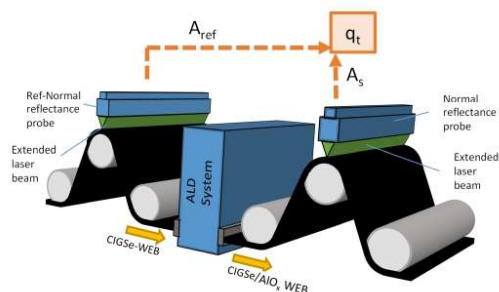


FIGURE 9 Integration of a normal reflectance-based process monitoring tool in a RTR S-ALD AlO_x deposition process

of the base materials, the samples analyzed in this work had an intentional AlO_x thickness gradient along the radial direction which has introduced additional dispersion. This has also importantly contributed to the difficulties observed to distinguish the AlO_x deposited areas from the bare substrate for low layer thicknesses. However, the results presented throughout the work have shown that the methodology employed is sensitive to these slight thickness changes of a few nm. As such, if $A_{\text{ref}}(x)$ and q_t are estimated individually for each measuring point, the resolution of the measurements would clearly be well below 10 nm allowing to precisely estimate the thickness of the AlO_x layers and carrying out high-resolution homogeneity control of deposited AlO_x layers in large areas. Moreover, changing the measuring spot from μm to cm size would allow performing both micro and macro homogeneity evaluations of the thickness of the layers.

Finally, as demonstrated in this work for low amount of spectra, in an industrial environment where an extremely large amount of data are expected to be obtained continuously, the implementation of machine learning algorithms for data analysis would also represent an advantageous strategy for improving the precision of the measurements thanks to its higher resilience to both sample and instrumental related fluctuations, fast training and classification, continuous self-improvement, and versatility in comparison to the use of calibration curves.

5 | CONCLUSIONS

In this work, a novel solution has been proposed and demonstrated for determining the thickness of AlO_x nanometric coatings using normal reflectance measurements: a nondestructive, fast, precise, low-cost, and scalable characterization method that can be implemented both in research and industrial process monitoring environments. The approach is based on detecting variations in the normal reflectance signal of a base/ AlO_x /air sample originated as a consequence of the varying nanolayer thickness. The viability of the proposed solution for the analysis of AlO_x layer thickness in PV

devices has been demonstrated employing a self-designed normal reflectance system and analyzing AlO_x nanolayers deposited on Si, CIGS, and PET substrates. Large area mappings covering the full surface of the samples have been performed, and methodologies based both on control parameter-based calibration curves and machine learning algorithms have been developed to relate the reflectance signal to the thickness of the AlO_x layers for each type of sample. These methodologies have been proven to be sensitive to thickness variations below 10 nm and have been demonstrated to be reliable for monitoring the AlO_x layers thickness in large area industrial environments with high resolution. Additionally, the limitations of the technique as well as the most critical aspects that should be regarded to implement a normal reflectance-based tool for industrial process monitoring have been discussed. As such, this work paves the way for developing a novel characterization technology that has direct application for monitoring industrial AlO_x -based encapsulation processes for flexible thin film PV modules but that can also be extended to many other industrial applications that require a precise and simple way of evaluating the thickness of nanocoatings.

ACKNOWLEDGEMENTS

This work has been partially funded by the European Union H2020 Framework Program (H2020-LCE-2017-RES-IA) under Grant Agreement no. 792245 (SuperPV). The authors from IREC and the University of Barcelona are grateful to Solar-ERA.NET DURACIS project (Spanish subproject funded by MICIIN nr. PCIN-2017-041), are supported by the European Regional Development Funds (ERDF, FEDER Programa Competitivitat de Catalunya 2007–2013), and belong to the SEMS (Solar Energy Materials and Systems) Consolidated Research Group of the “Generalitat de Catalunya” (Ref. 2017SGR 862). M.G. acknowledges the financial support from ACCIÓ-Generalitat de Catalunya within the TECNIOSpring Plus fellowship (TECSPR18-1-0048).

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Enric Grau-Luque <https://orcid.org/0000-0002-8357-5824>

Maxim Guc <https://orcid.org/0000-0002-2072-9566>

Ignacio Beceril-Romero <https://orcid.org/0000-0002-7087-6097>

Victor Izquierdo-Roca <https://orcid.org/0000-0002-5502-3133>

Alejandro Pérez-Rodríguez <https://orcid.org/0000-0002-3634-1355>

Pieter Bolt <https://orcid.org/0000-0002-9336-6321>

Fieke Van den Bruele <https://orcid.org/0000-0002-8188-296X>

REFERENCES

- Beceril-Romero I. *Alternative Substrates for Sustainable and Earth-Abundant Thin Film Photovoltaics*. PhD Thesis. Barcelona: Universitat de Barcelona; 2019. <http://diposit.ub.edu/dspace/handle/2445/147001>
- Dhere NG. Flexible packaging for PV modules, Proc. SPIE 7048, Reliability of Photovoltaic Cells, Modules, Components, and Systems. 2008;70480R. <https://doi.org/10.1117/12.795718>

3. Leterrier Y. Durability of nanosized oxygen-barrier coatings on polymers. *Prog Mater Sci.* 2003;48(1):1-55. [https://doi.org/10.1016/S0079-6425\(02\)00002-6](https://doi.org/10.1016/S0079-6425(02)00002-6)
4. Hirvikorpi T, Laine R, Vähä-Nissi M, et al. Barrier properties of plastic films coated with an Al₂O₃ layer by roll-to-roll atomic layer deposition. *Thin Solid Films.* 2014;550:164-169. <https://doi.org/10.1016/j.tsf.2013.10.148>
5. Garcia PF, McLean RS, Hegedus S. Encapsulation of Cu(In,Ga)Se₂ solar cell with Al₂O₃ thin-film moisture barrier grown by atomic layer deposition. *Sol Energy Mater Sol Cells.* 2010;94(12):2375-2378. <https://doi.org/10.1016/j.solmat.2010.08.021>
6. Zhang Y, Bertrand JA, Yang R, George SM, Lee YC. Electroplating to visualize defects in Al₂O₃ thin films grown using atomic layer deposition. *Thin Solid Films.* 2009;517(11):3269-3272. <https://doi.org/10.1016/j.tsf.2008.12.052>
7. Cooper R, Upadhyaya HP, Minton TK, Berman MR, Du X, George SM. Protection of polymer from atomic-oxygen erosion using Al₂O₃ atomic layer deposition coatings. *Thin Solid Films.* 2008;516(12):4036-4039. <https://doi.org/10.1016/j.tsf.2007.07.150>
8. Groner MD, George SM, McLean RS, Garcia PF. Gas diffusion barriers on polymers using Al₂O₃ atomic layer deposition. *Appl Phys Lett.* 2006;88(5):051907. <https://doi.org/10.1063/1.2168489>
9. Garcia PF, McLean RS, Reilly MH, Groner MD, George SM. Ca test of Al₂O₃ gas diffusion barriers grown by atomic layer deposition on polymers. *Appl Phys Lett.* 2006;89(3):031915. <https://doi.org/10.1063/1.2221912>
10. Poedt P, Cameron DC, Dickey E. Spatial atomic layer deposition: A route towards further industrialization of atomic layer deposition. *J Vac Sci Technol A: Vac Surf Films.* 2012;30(1):010802. <https://doi.org/10.1116/1.3670745>
11. Poedt P, Knaepen R, Illiberi A, Roozeboom F, van Asten A. Low temperature and roll-to-roll spatial atomic layer deposition for flexible electronics. *J Vac Sci Technol A: Vac Surf Films.* 2012;30(1):01A142. <https://doi.org/10.1116/1.3667113>
12. Li J, Wang Y, Wan F, et al. Passivation via atomic layer deposition Al₂O₃ for the performance enhancement of quantum dot photovoltaics. *Sol Energy Mater Sol Cells.* 2020;209:110479. <https://doi.org/10.1016/j.solmat.2020.110479>
13. Fan P, Sun Z, Wilkes GC, Gupta MC. Low-temperature laser generated ultrathin aluminum oxide layers for effective c-Si surface passivation. *Appl Surf Sci.* 2019;480:35-42. <https://doi.org/10.1016/j.apsusc.2019.02.023>
14. Ojeda-Durán E, Monfil-Leyva K, Andrade-Arvizu J, et al. CZTS solar cells and the possibility of increasing VOC using evaporated Al₂O₃ at the CZTS/CdS interface. *Sol Energy.* 2020;198:696-703. <https://doi.org/10.1016/j.solener.2020.02.009>
15. Choi S, Kamikawa Y, Nishinaga J, Yamada A, Shibata H, Niki S. Lithographic fabrication of point contact with Al₂O₃ rear-surface-passivated and ultra-thin Cu(In,Ga)Se₂ solar cells. *Thin Solid Films.* 2018;665:91-95. <https://doi.org/10.1016/j.tsf.2018.08.044>
16. Lim JWM, Chan CS, Xu L. High quality hydrogenated amorphous silicon thin films with enhanced growth rates for surface passivation in an Al₂O₃ based ICP reactor. *Procedia Eng.* 2016;139:56-63. <https://doi.org/10.1016/j.proeng.2015.09.216>
17. Hwang B-U, Kim D-I, Cho S-W. Role of ultrathin Al₂O₃ layer in organic/inorganic hybrid gate dielectrics for flexibility improvement of InGaZnO thin film transistors. *Org Electron.* 2014;15(7):1458-1464. <https://doi.org/10.1016/j.orgel.2014.04.003>
18. Vermang B, Fjallstrom V, Gao X, Edoff M. Improved rear surface passivation of Cu(In,Ga)Se₂ solar cells: A combination of an Al₂O₃ rear surface passivation layer and nanosized local rear point contacts. *IEEE J Photovolt.* 2014;4(1):486-492. <https://doi.org/10.1109/jphotov.2013.2287769>
19. Gunter PLJ, Niemantsverdriet JW. Thickness determination of uniform overlayers on rough substrates: A comparison of calculations for Al₂O₃/Al to x-ray photoelectron spectroscopy and atomic force microscopy experiments on technical aluminum foils. *J Vac Sci Technol A: Vac Surf Films.* 1995;13(3):1290-1292. <https://doi.org/10.1116/1.579552>
20. Wang X, Xiang J, Wang W, Zhao C, Zhang J. Dependence of electrostatic potential distribution of Al₂O₃/Ge structure on Al₂O₃ thickness. *Surf Sci.* 2016;651:94-99. <https://doi.org/10.1016/j.susc.2016.04.001>
21. Suárez-Campos G, Cabrera-German D, Castelo-González AO, et al. Characterization of aluminum oxide thin films obtained by chemical solution deposition and annealing for metal-insulator-metal dielectric capacitor applications. *Appl Surf Sci.* 2020;513:145879. <https://doi.org/10.1016/j.apsusc.2020.145879>
22. Hyde GK, McCullen SD, Jeon S. Atomic layer deposition and incompatibility of titanium nitride nano-coatings on cellulose fiber substrates. *Biomed Mater.* 2009;4(2):025001. <https://doi.org/10.1088/1748-6041/4/2/025001>
23. Martínez G, Shutthanandan V, Thevuthasan S, Chessa JF, Ramana CV. Effect of thickness on the structure, composition and properties of titanium nitride nano-coatings. *Ceram Int.* 2014;40(4):5757-5764. <https://doi.org/10.1016/j.ceramint.2013.11.014>
24. Erlat AG, Henry BM, Grovenor CRM, Briggs AGD, Chater RJ, Tsukahara Y. Mechanism of Water Vapor Transport through PET/AlOxNy Gas Barrier Films. *J Phys Chem B.* 2004;108(3):883-890. <https://doi.org/10.1021/jp036244y>
25. Kobayashi NP, Donley CL, Wang S-Y, Williams RS. Atomic layer deposition of aluminum oxide on hydrophobic and hydrophilic surfaces. *J Cryst Growth.* 2007;299(1):218-222. <https://doi.org/10.1016/j.jcrysgro.2006.11.224>
26. Srinivasan K, Kottantharayil A. Aluminium oxide thin film deposited by spray coating for p-type silicon surface passivation. *Sol Energy Mater Sol Cells.* 2019;197:93-98. <https://doi.org/10.1016/j.solmat.2019.03.048>
27. Şakalak H, Yılmaz K, Gürsoy M, Karaman M. Roll-to-roll initiated chemical vapor deposition of super hydrophobic thin films on large-scale flexible substrates. *Chem Eng Sci.* 2020;215:115466. <https://doi.org/10.1016/j.ces.2019.115466>
28. González-Ramírez JE, Fuentes J, Hernández LC, Hernández L. Evaluation of the thickness in nanolayers using the transfer matrix method for modeling the spectral reflectivity. *Res Lett Phys.* 2009;2009:1-4. <https://doi.org/10.1155/2009/594175>
29. Bermudez V, Perez-Rodríguez A. Understanding the cell-to-module efficiency gap in Cu(In,Ga)(S,Se)₂ photovoltaics scale-up. *Nat Energy.* 2018;3(6):466-475. <https://doi.org/10.1038/s41560-018-0177-1>
30. Lin T, Wan N, Xu J, Xu L, Chen K-J. Size-dependent optical properties of SnO₂ nanoparticles prepared by soft chemical technique. *J Nanosci Nanotechnol.* 2010;10(7):4357-4362. <https://doi.org/10.1166/jnn.2010.2203>
31. Illiberi A, Frijters C, Ruth M. Atmospheric spatial atomic layer deposition of ZnO buffer layers for flexible Cu(In,Ga)Se₂ solar cells. *J Vac Sci Technol A.* 2018;36(5):051511. <https://doi.org/10.1116/1.5040457>
32. Ralbovsky NM, Lednev IK. Raman spectroscopy and chemometrics: A potential universal method for diagnosing cancer. *Spectrochim Acta A Mol Biomol Spectrosc.* 2019;219:463-487. <https://doi.org/10.1016/j.saa.2019.04.067>
33. Chauhan R, Kumar R, Kumar V, Sharma K, Sharma V. On the discrimination of soil samples by derivative diffuse reflectance UV-vis-NIR spectroscopy and chemometric methods. *Forensic Sci Int.* 2021;319:110655. <https://doi.org/10.1016/j.forsciint.2020.110655>
34. Guleken Z, Ünübol B, Bilici R, et al. Investigation of the discrimination and characterization of blood serum structure in patients with opioid use disorder using IR spectroscopy and PCA-LDA analysis. *J Pharm Biomed Anal.* 2020;190:113553. <https://doi.org/10.1016/j.jpba.2020.113553>

35. Chopri R, Sharma S, Singh R. Forensic analysis of red lipsticks using ATR-FTIR spectroscopy and chemometrics. *Forensic Chem.* 2020;17:100209. <https://doi.org/10.1016/j.forc.2019.100209>
36. Visneschi-Necrasov T, Barreira JCM, Cunha SC, Pereira G, Nunes E, Oliveira MPP. Phylogenetic insights on the isoflavone profile variations in Fabaceae spp.: Assessment through PCA and LDA. *Food Res Int.* 2015;76:51-57. <https://doi.org/10.1016/j.foodres.2014.11.032>
37. Wang Y, Zhu J, Chen X. Autofluorescence spectroscopy of blood plasma with multivariate analysis methods for the diagnosis of pulmonary tuberculosis. *Optik.* 2020;224:165446. <https://doi.org/10.1016/j.jjleo.2020.165446>
38. Python Software Foundation. Python language reference, version 3.9.0. Available at <http://www.python.org>
39. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-Learn: Machine learning in python. *J Mach Learn Res.* 2011;12(85):2825-2830.
40. Alonso MI, Marcus IC, Garriga M, Goñi AR, Jedrzejewski J, Balberg I. Evidence of quantum confinement effects on interband optical transitions in Si nanocrystals. *Phys Rev B.* 2010;82(4):045302. <https://doi.org/10.1103/physrevb.82.045302>
41. Fairbrother A, Izquierdo-Roca V, Fontané X. ZnS grain size effects on near-resonant Raman scattering: optical non-destructive grain size estimation. *CrystEngComm.* 2014;16(20):4120. <https://doi.org/10.1039/c3ce42578a>
42. Becerril-Romero I, Sylla D, Placidi M. Transition-metal oxides for kesterite solar cells developed on transparent substrates. *ACS Appl Mater Interfaces.* 2020;12(30):33656-33669. <https://doi.org/10.1021/acsami.0c06992>
43. Harbecke B, Heinz B, Grosse P. Optical properties of thin films and the Berreman effect. *Appl Phys A.* 1985;38(4):263-267. <https://doi.org/10.1007/bf00616061>
44. Ventura SD, Birgin EG, Martínez JM, Chambouleyron I. Optimization techniques for the estimation of the thickness and the optical parameters of thin films using reflectance data. *J Appl Phys.* 2005;97(4):043512. <https://doi.org/10.1063/1.1849431>
45. Ohlídal M, Ohlídal I, Franta D, Králík T, Ják M, Eliáš M. Optical characterization of thin films non-uniform in thickness by a multiple-wavelength reflectance method. *Surf Interface Anal.* 2002;34(1):660-663. <https://doi.org/10.1002/sia.1382>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Grau-Luque E, Guc M, Becerril-Romero I, et al. Thickness evaluation of AlO_x barrier grain size for encapsulation of flexible PV modules in industrial environments by normal reflectance and machine learning. *Prog Photovolt Res Appl.* 2022;30(3):229-239. doi: 10.1002/ppp.3478

3.2 Publication 2

Cite this: *J. Mater. Chem. A*, 2021, 9,
10466

Combinatorial and machine learning approaches for the analysis of $\text{Cu}_2\text{ZnGeSe}_4$: influence of the off-stoichiometry on defect formation and solar cell performance†

Enric Grau-Luque,^a Ikram Anefnaf,^{bc} Nada Benhaddou,^{bc} Robert Fonoll-Rubio,^a Ignacio Becerril-Romero,^a Safae Aazou,^{bc} Edgardo Saucedo,^{ib ad} Zouheir Sekkat,^{bce} Alejandro Perez-Rodriguez,^{af} Victor Izquierdo-Roca^{ib *a} and Maxim Guc^{ib *a}

Solar cells based on quaternary kesterite compounds like $\text{Cu}_2\text{ZnGeSe}_4$ are complex systems where the variation of one parameter can result in changes in the whole system, and, as consequence, in the global performance of the devices. In this way, analyses that take into account this complexity are necessary in order to overcome the existing limitations of this promising Earth-abundant photovoltaic technology. This study presents a combinatorial approach for the analysis of $\text{Cu}_2\text{ZnGeSe}_4$ based solar cells. A compositional graded sample containing almost 200 solar cells with different [Zn]/[Ge] compositions is analyzed by means of X-ray fluorescence and Raman spectroscopy, and the results are correlated with the optoelectronic parameters of the different cells. The analysis results in a deep understanding of the stoichiometric limits and point defects formation in the $\text{Cu}_2\text{ZnGeSe}_4$ compound, and shows the influence of these parameters on the performance of the devices. Then, intertwined connections between the compositional, vibrational and optoelectronic properties of the cells are revealed using a complex analytical approach. This is further extended using a machine learning algorithm. The latter confirms the correlation between the properties of the $\text{Cu}_2\text{ZnGeSe}_4$ compound and the optoelectronic parameters, and also allows proposing a methodology for device performance prediction that is compatible with both research and industrial process monitoring environments. As such, this work not only provides valuable insights for understanding and further developing the $\text{Cu}_2\text{ZnGeSe}_4$ photovoltaic technology, but also gives a practical example of the potential of combinatorial analysis and machine learning for the study of complex systems in materials research.

Received 12th February 2021
Accepted 7th April 2021

DOI: 10.1039/d1ta01299a

rsc.li/materials-a

Introduction

$\text{Cu}_2\text{ZnSn}(\text{S},\text{Se})_4$ (CZTSSe) based compounds, more widely known as kesterites, are considered as the natural earth-abundant and low toxicity successors of the more mature inorganic thin film photovoltaic (PV) technologies $\text{Cu}(\text{In},\text{Ga})\text{Se}_2$

(CIGS) and CdTe which are based on scarce and/or toxic materials.¹ Although a considerable amount of progress in the technological development and fundamental understanding of kesterites has been achieved in the last years, the record power conversion efficiency at laboratory scale has barely evolved since 2014 and is stagnated at around 13%.² Sn is often regarded as the main culprit of this stagnation due, mainly, to the volatility of $\text{Sn}(\text{S},\text{Se})_x$ species³ that leads to morphological and compositional problems,⁴ and the instability of the Sn oxidation state that may lead to the formation of deep defects^{5,6} ultimately causing kesterite PV devices to exhibit a high V_{oc} deficit. In this regard, the substitution of Sn by Ge is currently regarded as a promising strategy to improve the kesterite technology. Ge doping (CZTSSe:Ge) and alloying (CZTGSSe) has been demonstrated to enhance the performance of kesterite devices significantly by improving the V_{oc} of the devices which is commonly attributed to the formation of liquid phases and better intermixing during high temperature synthesis, to improvements in carrier lifetime and to a reduction of band tailing.^{5,7–9} In

^aCatalonia Institute for Energy Research-IREC, Sant Adrià de Besòs, Barcelona, Spain. E-mail: vizquierdo@irec.cat; mguc@irec.cat

^bDepartment of Chemistry, Faculty of Sciences, Mohammed V University in Rabat, Morocco

^cOptics & Photonics Center, Moroccan Foundation for Advanced Science, Innovation and Research – MASIR, Rabat, Morocco

^dPhotovoltaic Group, Electronic Engineering Department, Universitat Politècnica de Catalunya, C. J. Girona 31, Barcelona, 08034, Spain

^eDepartment of Applied Physics, Osaka University, 2-1 Yamadaoka, Suita, Osaka, Japan

^fDepartament d'Enginyeria Electrònica i Biomèdica, IN2UB, Universitat de Barcelona, C/ Martí i Franquès 1, 08028 Barcelona, Spain

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1ta01299a



addition, the partial substitution of Sn by Ge increases the bandgap of the kesterite semiconductor material enabling the creation of a graded bandgap through the development of in-depth compositional engineering strategies.^{10,11}

On the other hand, the total substitution of Sn by Ge (CZGSSe) appears as an even more promising approach since, in addition to completely avoiding Sn-related issues, the wider bandgap of the material (~ 1.5 eV for CZGSe and ~ 2.2 eV for CZGS)^{12,13} opens the door to semi-transparent, tandem and photocatalytic water splitting applications. Significant advances have been made in the last years in pure Ge kesterite solar cells leading to CZGSe devices with efficiencies of up to 8.5%.¹⁴ Although promising, this value is very far from those of the best CZTSe devices. The highest efficiency levels reported for CZGSe are commonly achieved imitating the standard CZTSe and employing off-stoichiometry Cu-poor Zn-rich absorber compositions.^{14–17} However, this compositional ratio might not be the optimal one for CZGSe and might be one of the reasons holding back the development of this technology towards higher efficiencies as compared with the standard CZTSe. As such, fundamental studies that investigate the formation of CZGSe, secondary phases and point defects off-stoichiometry represent a very valuable asset for understanding this material and for paving the way to the development of strategies that may lead to a further development of the technology. In this context, Gunder *et al.* carried out a detailed investigation of defect formation in off-stoichiometric CZGSe powder.¹⁸ However, mainly Zn-rich and Cu-rich powder samples were synthesized, with few Cu-poor samples.

Finally, solar cells based on quaternary kesterite compounds and multilayer stacks like CZGSe are complex systems where the variation of one parameter can result in changes in the whole system, and, as consequence, in the global performance of the devices. In this way, analyses that take into account this complexity are necessary in order to overcome the existing limitations of this promising Earth-abundant photovoltaic technology.

In this work, we present a systematic study of a combinatorial CZGSe sample comprised by almost 200 individual solar cells with different $[\text{Zn}]/[\text{Ge}]$ compositions in the Cu-poor regime. Structural, compositional and optoelectronic characterization are applied in a combinatorial way. Firstly, a complex analytical approach allows defining the off-stoichiometric limits of formation of the CZGSe kesterite phase and the optimum compositional range to obtain the highest efficiencies (up to 6.3%) in terms of the $[\text{Zn}]/[\text{Ge}]$ ratio. This includes the study of solar cell performance dependence on the concentration of point defects. Secondly, we demonstrate the potential of applying machine learning (ML) for the analysis of the combinatorial sample. The ML methodology proves to enable effective prediction of cell efficiency based only on Raman spectra and compositional data, while the resulting discriminants show a linear correlation with point defect concentration and device efficiency pointing at a strong fundamental interconnection between point defects, Raman spectra, composition and cell performance. This work serves both as a fundamental study that provides valuable results for the development of the CZGSe

technology and as a powerful example of how combinatorial analysis and machine learning can be used to unravel the critical parameters that govern the performance of complex optoelectronic devices.

Experimental details and methods

Sample preparation

A CZGSe combinatorial sample was prepared through the selenization of a compositionally graded Cu/Zn/Ge metallic stack precursor deposited by DC magnetron sputtering (Alliance AC450) on a 5×5 cm² soda-lime glass/Mo substrate. In order to generate a compositional gradient, the Cu and Zn precursor layers were homogeneously deposited over the substrate while the Ge layer was deposited without substrate rotation generating a thickness gradient and, in turn, a $[\text{Zn}]/[\text{Ge}]$ compositional gradient. A 3-zone tubular furnace (Hobersal) was employed to synthesize the CZGSe absorber. The 3 zones were kept at the same temperature during the whole process to ensure spatial homogeneity throughout the entire length of the furnace. Samples were placed inside a graphite box (69 cm³) together with crucibles containing 100 mg of Se (Alfa-Aesar powder, 200 mesh, 99.999%) and 5 mg of GeSe₂ (American Elements, power, 99.999%) to perform a 2-step reactive thermal annealing in a Se + Ge atmosphere. It consisted in a first stage in which the furnace was kept at 330 °C and 1.5 mbar Ar pressure for 30 minutes and a second step at 480 °C and 1 bar Ar pressure for 15 minutes. The heating rate was set to 20 °C min⁻¹ in both steps. The samples were let to cool down naturally.

The as-synthesized absorber was submitted to a chemical etching in diluted KCN (2% w/v, room temperature, 2 min). Immediately after, a CdS layer was deposited by chemical bath deposition (the process is detailed in ref. 19). The solar cell structure was then completed with i-ZnO (50 nm) and ITO (200 nm, 60 Ω sq⁻¹ sheet resistance) layers deposited by DC-magnetron sputtering (Alliance Concept CT100). The sample was then scribed into 196 individual 3×3 mm² solar cells (see Fig. S1†) using a manual microdiamond scriber (MR200 OEG). Neither anti-reflective coating nor metallic grids were used in the devices presented in this work.

Characterization techniques

The elemental composition of the different cells of the combinatorial sample was determined by X-ray fluorescence (XRF) using a Fischerscope XDV system with a 1 mm spot diameter, a 50 kV acceleration voltage, a Ni10 filter and a 45 s acquisition time. Raman analysis with blue (442 nm) and green (532 nm) excitation wavelengths were performed on the bare absorber, while measurements with NIR (785 nm) were performed in complete devices using Horiba Jobin Yvon FHR640 and iHR320 monochromators coupled with CCD detectors. The first monochromator is optimized for the UV and visible spectral ranges and was used with 442 nm (He–Cd gas laser) and 532 nm (solid state laser) excitation wavelengths. The second monochromator is optimized for the NIR range and was used with a 785 nm (solid state laser) excitation wavelength. The power



density of the lasers was kept below 150 W cm^{-2} and the spot size was $\sim 70 \mu\text{m}$. The measurements were performed in a backscattering configuration through a specific probe designed at IREC.

The J - V characteristics of the devices were obtained under simulated AM1.5 illumination (1000 W m^{-2} intensity at room temperature) using a pre-calibrated Class AAA solar simulator (Abet Technologies Sun 3000).

Machine learning methodology

A machine learning (ML) driven methodology based on a linear discriminant analysis (LDA) algorithm was employed to deepen into the complex dependence of solar cell optoelectronic parameters and composition on the different parameters found in the analysis of the Raman spectra. LDA is a dimension-reduction algorithm, capable of reducing high-dimensionality problems into a bi-dimensional one, discerning and employing the most relevant dimensions of the dataset. In order to test and implement the machine learning based LDA algorithm, the Python programming environment²⁰ with the Scikit-Learn library²¹ was used. All the Raman spectra measured under different excitation conditions for each cell were used as input features (588 spectra, in total), and the data were randomly divided in 70% for training and 30% for testing. The algorithm was trained for 3 different classification targets, namely $[\text{Zn}]/[\text{Ge}]$ ratio, V_{oc} and efficiency. For each trained algorithm, the data was divided in 4 classification groups of approximately an equal amount of data. The amount of experimental data employed for the analysis (196 cells) is far from being considered "big data" and the results presented below are susceptible to further improvement for a more precise classification through the use of a higher number of training inputs. Nevertheless, this approach illustrates the applicability and potential of this methodology for material analysis by spectroscopic techniques.

Results and discussions

Off-stoichiometry limits and secondary phases

The chemical composition of every individual solar cell of the combinatorial samples was obtained by XRF. The mappings of the cationic ratios are presented in the Fig. S1† and, in Fig. 1, the obtained values are combined with the different off-stoichiometry kesterite types (see ref. 22 and 23 for more details about the off-stoichiometry lines of kesterite type compounds). It can be observed that the compositions of the different cells of the combinatorial samples cross the A- (almost perpendicularly), J- and L-type lines. On the one hand, the $[\text{Zn}]/[\text{Ge}]$ ratio covers a wide range from almost 0.7 to 1.4, allowing to explore not only the typical Zn-rich compositions commonly used for kesterite type compounds, but also the Zn-poor region. This allows investigating the origin of the positive effect of Zn-rich compositions in kesterite PV devices. On the other hand, the $[\text{Cu}]/([\text{Zn}]+[\text{Ge}])$ ratio was maintained well below 1, ensuring the Cu-poor condition for all cells. It is worth mentioning that the non-stoichiometric compositions were quite different from

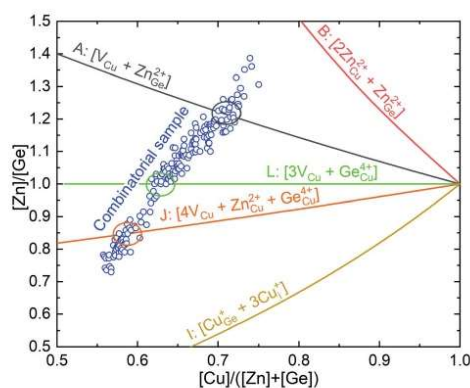


Fig. 1 Cationic ratios of the CZGSe combinatorial sample (blue open circles) represented together with off-stoichiometric types of kesterite compounds (solid lines).^{22,23}

previous studies of the same compound,¹⁸ where mainly Zn-rich and Cu-rich powder samples were synthesized and the few Cu-poor samples contained secondary phases.

Raman spectroscopy was used as the main tool to analyse phase formation in the CZGSe absorber layer with varying $[\text{Zn}]/[\text{Ge}]$ ratio. A multiwavelength analysis allowed to detect possible secondary phases and variations in the main kesterite phase (see Fig. 2). The blue excitation wavelength is well-known to be highly sensitive for detecting the ZnSe secondary phase.^{22,24} As shown in the figure, this secondary phase was found in cells with Zn-rich composition (see the spectrum of the $[\text{Zn}]/[\text{Ge}] = 1.25$ cell in Fig. 2, left). What is more, the cells on which a strong ZnSe peak was detected, also presented a shift to lower wavenumbers and an increase of the full width at half maximum (FWHM) of the main and the second most intense peaks of the kesterite phase under different excitation wavelengths not sensitive to ZnSe (see the spectrum of a cell with $[\text{Zn}]/[\text{Ge}] = 1.25$ under green excitation wavelength in Fig. 2, middle). This, according to the previous interpretations of the kesterite type compounds, can be related with an increased Cu/Zn disorder²⁵ or with a phonon confinement effect due to a low grain size in the absorber.²⁶ Since all the cells of the combinatorial sample were processed at the exact same temperature, the appearance of strong variations in the Cu/Zn disordering is unlikely (*e.g.* see ref. 27 and 28). In addition, it is hard to envision how the formation of the ZnSe phase could influence Cu/Zn disordering in the CZGSe compound. On the other hand, even taking into account the optimal temperature treatment for the formation of good crystalline quality CZGSe phase,^{16,29} the presence of ZnSe grains can greatly influence the formation and size of the kesterite grains,³⁰ leading to a worsening of its crystalline quality and grain size, and causing the appearance of phonon confinement in agreement to the observed red shift and broadening of the kesterite Raman peaks.



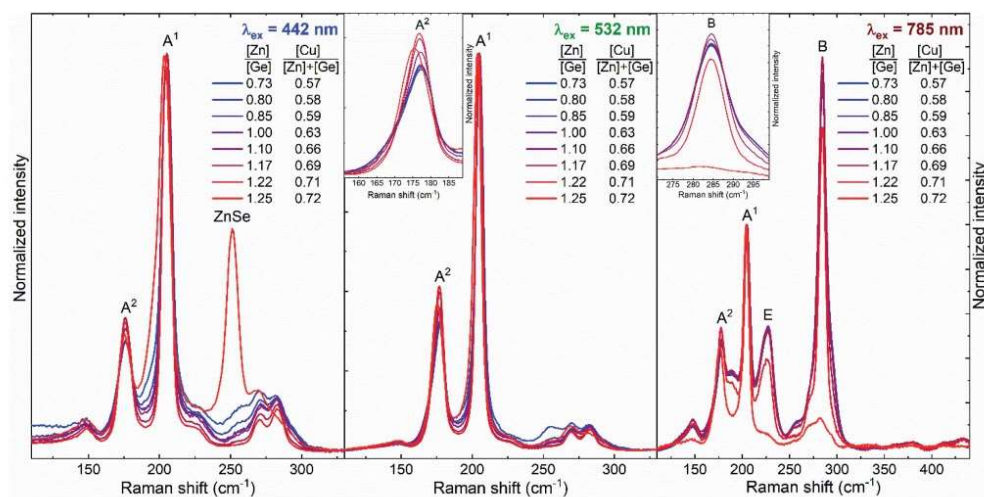


Fig. 2 Examples of Raman scattering spectra measured in cells with different compositions under different excitation wavelength.

In addition to the ZnSe phase, the blue excitation wavelength is also sensitive to the GeSe₂ phase. This is a 2D compound that has a direct band gap of ~2.7 eV (close to the energy of the blue laser line -2.8 eV) and has its main Raman peak at around 210 cm⁻¹.³¹ The latter is strongly overlapped with the main peak of CZGSe (A¹ symmetry peak at 205 cm⁻¹), which compromises the detection of this secondary phase. However, a strong resonance with the blue excitation wavelength should result in the appearance of, at least, a shoulder at the high wavenumber side of the main kesterite peak. However, the spectra in Fig. 2 (left) show no evidence of the presence of GeSe₂, even for the lowest [Zn]/[Ge] ratios (highest Ge-content). Another Ge-based binary compound is GeSe. The orthorhombic crystalline polymorph of this binary compound has a direct band gap of about 1.53 eV and an intense A₁ symmetry mode at 188 cm⁻¹.³² This band gap value is very close to the resonant condition of this secondary phase under a 785 nm excitation wavelength. Nevertheless, no clear Raman peak of GeSe phase can be observed in the spectra acquired (Fig. 2, right). On the other hand, although the properties of the amorphous Ge_xSe_{1-x} phase strongly depend on the *x* value, an intense Raman band close to 200 cm⁻¹ and assigned to the stretching mode of the GeSe_{4/2} corner-sharing tetrahedra can be defined as representative for most of the compositional polymorphs of these amorphous phases.³³ Moreover, a GeSe₉ liquid phase, recently found as one of the intermediate phases during the formation of CZGSe,¹⁶ might remain at the surface of the absorber layer. In the spectra measured under the green excitation wavelength, a slight broadening of the main peak of CZGSe with the decreasing [Zn]/[Ge] ratio can be observed. However, it cannot be unequivocally ascribed to the presence of amorphous Ge_xSe_{1-x} or liquid GeSe₉ secondary phases since it could also be explained by intrinsic changes in the kesterite phase, that will be further discussed. Finally, elemental Ge and

Ge-containing ternary phases (like Cu₂GeSe₃ or Cu₂GeSe₄) can be formed in very Zn-poor conditions. Although a good sensitivity of Raman spectroscopy to these phases is expected, their narrow band gap (below 1 eV, see ref. 34–37 for band gap and fingerprint Raman spectra) does not allow working in resonant conditions making the detection of small amounts of elemental Ge and Ge-containing ternary phases very challenging. In this way, it can be concluded that no Ge-related secondary phases are forming even for very Zn-poor compositions, or that the amount of these secondary phases is negligibly small, as no strong/sharp changes of the spectra of the cells of the CZGSe combinatorial sample with different compositions can be observed (in the [Zn]/[Ge] = 0.7–1.2 range). This differentiates the pure Ge-based from the pure Sn-based kesterites, where the formation of Sn-containing secondary phases has been observed even at Zn-rich compositions.^{38,39} Finally, it should be noted that no Cu-containing binary secondary phases are expected due to Cu-poor composition of all the cells of the combinatorial sample (see Fig. 1).

According to the phase formation analysis performed by means of Raman spectroscopy presented above, only a ZnSe secondary phase was clearly detected in the combinatorial sample under Zn-rich compositions. ZnSe becomes the dominant phase for [Zn]/[Ge] > 1.2 in some of the cells. This squeezes the upper limit in which the pure CZGSe kesterite phase can be formed to [Zn]/[Ge] ratios close to 1.2. On the contrary, the lower limit for the formation of the pure CZGSe kesterite phase can be considered close to 0.7, comparable with CZTSSe compounds.^{23,38} However, the presence of small amount of Ge-based secondary phases is hard to exclude from the data presented. Nevertheless, taking into account previous studies^{18,22} and the results presented here, it can be concluded that replacing of Sn with Ge does not lead to a significant shortening



of the off-stoichiometry range for the formation of the CZGSe kesterite type structure. Likewise, the stoichiometry flexibility of CZGSe compounds might be one of the culprits for the existing limitations of the PV devices in the same way as for other kesterite-based compounds.⁴

Point defect formation

As mentioned above (see Fig. 1), the chemical composition of most of the cells is in the range from J- to A-type kesterite off-stoichiometric lines. According to this, the main point defect expected in the combinatorial sample are copper vacancies (V_{Cu}) and zinc or germanium in copper position (Zn_{Cu} and Ge_{Cu}).^{22,38,40} In order to define the influence of these point defects on the Raman spectra, specific cells with chemical compositions close to the three off-stoichiometric lines crossed by the combinatorial sample (A-, J- and L-type lines) were analysed. Measurements under different excitation wavelengths resulted in the detection of the characteristic features that differentiate the Raman spectra of the different off-stoichiometry types of the CZGSe kesterite compound (see Fig. 3). Here, the spectra of the cells around the J- and L-type lines present a great similarity with just very subtle changes in the intensity of the band at 176 cm^{-1} and E/B symmetry peaks in the high wavenumber range ($220\text{--}300\text{ cm}^{-1}$), which are mainly observed under 785 nm excitation (Fig. 3, right). In contrast, strong differences are observed for the spectra of the cells close to the A-type line with a decrease of the relative intensity and width of the peaks, except for the band at 176 cm^{-1} . In previous works, the change in the intensity of the second most intense Raman band in the CZTSe compound was correlated, mainly, with a change of the concentration of V_{Cu} point defects, and was shown to have a crucial impact on the

properties of the CZTSe absorber and on device performance.^{41–43} Taking this into account, it can be inferred that in the combinatorial sample analysed in this work, there is a higher concentration of V_{Cu} for the cells close to the J- and L-type lines, and it is reduced for the cells around the A-type line. This is in line with the observations made above, where an increase of the intensity of the Raman band at 176 cm^{-1} is observed with the increasing $[Zn]/[Ge]$ ratio (see Fig. 2). However, the concomitant increase of the $[Cu]/([Zn] + [Ge])$ ratio (see Fig. 1), even if much smaller than the increase of the $[Zn]/[Ge]$ ratio, is also expected to have a critical influence on the concentration of V_{Cu} .

Taking a look at Fig. 3, a great similarity between the spectra corresponding to cells around the J- and L-type lines can be seen regardless the presence of the Zn_{Cu} point defect in some cells and its absence in others. This allows concluding that the Zn_{Cu} defect has a low influence on the Raman scattering spectra of the CZGSe compound. On the other hand, the Ge_{Cu} substitutional defect presents a more significant influence on the Raman spectra, but mainly in the high wavenumber range ($220\text{--}300\text{ cm}^{-1}$), where the relative intensity of the peaks increases with the higher Ge content (or lower $[Zn]/[Ge]$ ratio). Nevertheless, it is hard to strictly distinguish the influence of the two substitutional point defects on the intensity of the peaks at the high frequency range and both of them will be considered in the analysis of the influence of the point defects on device performance presented in the next section.

Finally, it should be borne in mind that, for the analysis carried out above, only cells without ZnSe were used since the presence of this phase results in significant changes in the spectra measured under any excitation condition (e.g. see Fig. S2,† where spectra of the cells with similar compositions close to A-type off-stoichiometric line, but with and without

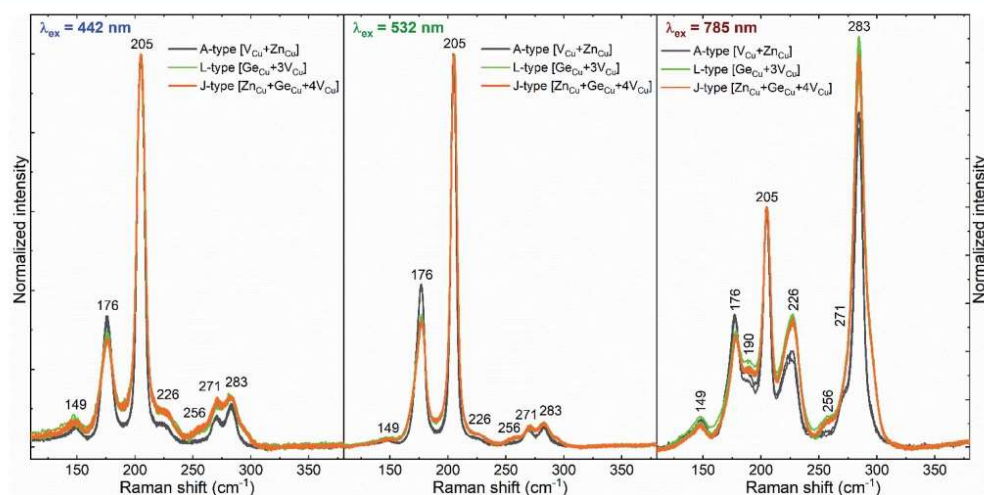


Fig. 3 Examples of Raman spectra in the vicinity of different off-stoichiometric type lines measured under different excitation wavelength.



Paper

ZnSe phase are presented). This indicates that Raman spectroscopy is a critical investigation tool to control the quality and phase purity of the kesterite type compounds.²²

Influence of point defects on device performance

This section studies the influence of the point defects detected by Raman spectroscopy on solar cell performance. First, the analysis is focused on the dependence of the optoelectronic properties of the devices on the relative integrated intensity of the band at 176 cm^{-1} (calculated as $A_{176}/(A_{176} + A_{205})$ with A_{176} calculated in the $168\text{--}183\text{ cm}^{-1}$ range and A_{205} calculated in the $198\text{--}211\text{ cm}^{-1}$ range from the spectra measured under 532 nm excitation wavelength), which is inversely related to V_{Cu} concentration as previously reported.^{41,43} Fig. 4a, shows a clear dependence of the efficiency of the solar cells with the concentration of V_{Cu} , with the highest efficiency achieved for a certain optimum concentration range of this defect, which corresponds to a relative integrated intensity of the Raman band at 176 cm^{-1} lying in the $0.325\text{--}0.350$ range. Outside this optimum range, a deficit (higher band intensity) or excess (lower band intensity) of V_{Cu} defect concentration is expected, both having a negative influence on device performance. As such, three different regions (deficit, excess and optimum V_{Cu}) can be distinguished. A more detailed analysis of the evolution of the optoelectronic properties with defect concentration (Fig. 4 b-d) reveals that the main driving force behind the evolution of solar cell efficiency are the changes in the fill factor (FF) and short circuit current density (J_{sc}). Both parameters exhibit a similar tendency with defect concentration. Copper vacancies are a well-known beneficial point defect in kesterite and chalcopyrite based solar cells, which leads to the formation of a shallow acceptor level and has a strong influence on the electrical conductivity of the absorber layer.^{44,45} In this way, a deficit of this beneficial defect leads to a decrease of the charge carrier concentration. However, an excess results in the formation of a high amount of scattering centres which significantly decreases the mobility of the charge carriers. Both effects, have a direct influence on the electrical conductivity of the absorber layer and, in turn, on the FF and J_{sc} of the final devices as observed in Fig. 4. On the other hand, the open circuit voltage (V_{oc}) shows only a slight dependence on V_{Cu} concentration, with only a sharp increase in the optimum defect range. These results differ from a previously published analysis of CZTSe samples, where the V_{oc} was found to depend on the change of the relative intensity of the second most intense kesterite band (I_2 around 170 cm^{-1}).⁴² However, it should be noted that only cells around the A-type off-stoichiometric line were selected in the mentioned reference, which strongly reduces the number of possible parameters that can influence solar cell performance, at least from the point of view of point defects.

Then, on a second stage, the influence of Zn_{Cu} and Ge_{Cu} substitutional point defects on the optoelectronic properties of the devices was analysed. According to previous results, an increase of the relative intensity of the peaks in the high wavenumber range correlates with a higher concentration of

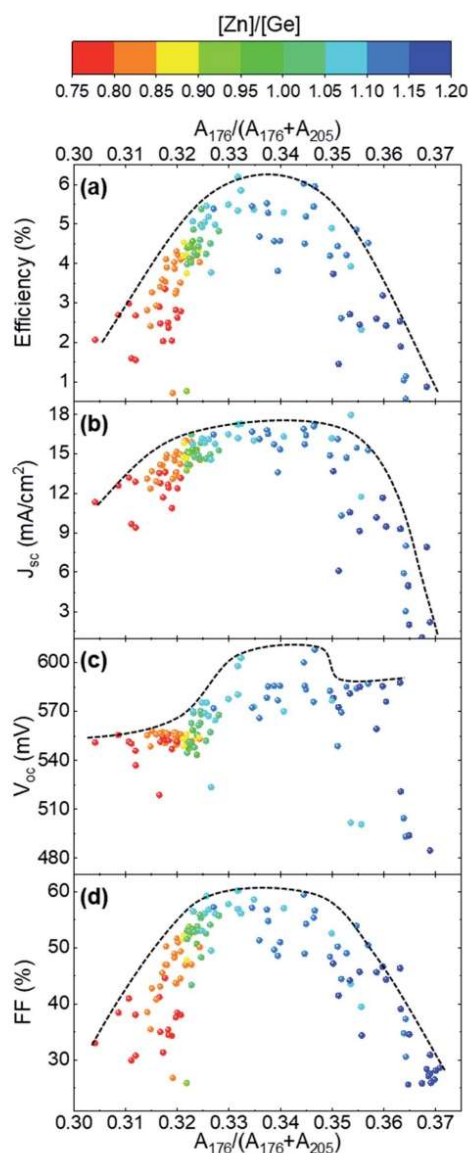


Fig. 4 Dependence of optoelectronic properties (efficiency (a), short circuit current (b), open circuit voltage (c), fill factor (d)) of the cells from the relative integrated intensity of the Raman band at 176 cm^{-1} . The colour scale corresponds to the $[\text{Zn}]/[\text{Ge}]$ ratio.

substitutional defects.^{41–43} Fig. 5a shows a parabolic dependence of solar cell efficiency with the variation of the relative integrated intensity of the Raman peaks in the range $235\text{--}300\text{ cm}^{-1}$. As in the case of V_{Cu} (see Fig. 4a), this shape is mainly



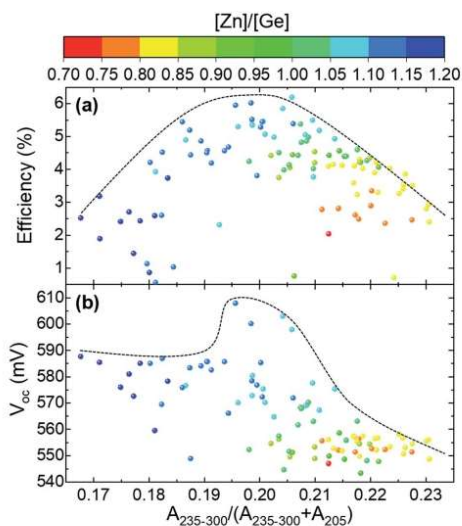


Fig. 5 Dependence of (a) efficiency and (b) open circuit voltage of the cells on the relative integrated intensity of the Raman peaks in the range 235–300 cm^{-1} . The colour scale corresponds to the [Zn]/[Ge] ratio.

governed by the changes in FF and J_{sc} (Fig. S3†) although in a less pronounced manner. This can be related to a slightly lower influence of the substitutional defects on these optoelectronic parameters. On the other hand, the analysis of the dependence of the V_{oc} on the relative integrated intensity of the peaks in the 235–300 cm^{-1} range (Fig. 5b) shows a clearer influence of the substitutional defects on this parameter. Similarly to the analysis above, three regions can be distinguished: (1) low amount of defects ($A_{235-300} < 0.190$); (2) optimum amount of defects ($0.190 < A_{235-300} < 0.215$); (3) high amount of defect ($A_{235-300} > 0.215$). A relatively constant V_{oc} value can be observed in region 1, followed by a sharp decrease in the second one, and a gentle decrease in region 3. Taking into account the [Zn]/[Ge] ratio, it can be seen that region 3 corresponds to Zn-poor conditions, for which Ge_{Cu} substitutional defects are expected to prevail over the Zn_{Cu} defects. The former defect can form a deep donor defect (based on *first principle calculations* of the familiar Sn-containing kesterite compounds⁴⁶) that increases the amount of non-radiative recombination, which finally decreases the V_{oc} .^{47,48}

A further analysis of the Raman spectra measured under different excitation wavelengths allows to establish additional failure mechanisms that lead to the decrease of the performance of the solar cells outside the optimum compositional range. A deep analysis of the spectra measured under blue and NIR excitations reveals that the lower efficiency in these ranges can be explained by two factors: the appearance of the ZnSe secondary phase and the change of the band gap of the absorber. The former effect can be seen in Fig. 6a where the

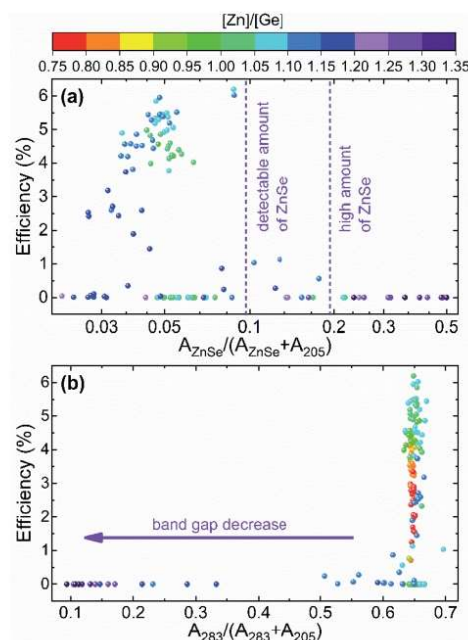


Fig. 6 Dependence of solar cell efficiency with (a) relative integrated intensity of the ZnSe peak (spectra measured under 442 nm excitation), (b) relative integrated intensity of the peak at 283 cm^{-1} (spectra measured under 785 nm excitation). The colour scale corresponds to the relative integrated intensity of the band at 176 cm^{-1} calculated from the spectra measured under 532 nm excitation.

relative intensity of the ZnSe peak is presented. The figure shows a clear decrease of solar cell performance with the increasing content of the ZnSe phase, proving the strong detrimental effect of this secondary phase. On the other hand, the band gap of CZGSe (around 1.5 eV) exhibits a resonant behaviour under NIR excitation conditions that leads to an increase of the intensity of the LO components of the E and B symmetry modes in kesterite type compounds.^{22,49,50} In the present study, this is observed by a strong enhancement of B symmetry peak at 283 cm^{-1} (see Fig. 2 and 3). However, the latter is not similar in all the cells, with some of them showing a rather low intensity of this peak (Fig. 6b). This can be related to the distancing of the CZGSe band gap from the excitation laser line. Previously, several works mentioned the effect of defects in the cations sublattice on the band gap of kesterite materials, but mainly the disorder of the Cu/Zn cations was discussed,⁵¹⁻⁵³ while changes in the concentration of point defects was just briefly tackled.^{54,55} As mentioned above, in the combinatorial sample analysed in this work, it is not expected to have a significant difference in Cu/Zn disorder, while a clear change in the concentration of V_{Cu} , Zn_{Cu} and Ge_{Cu} is observed. This implies that the concentration of point defects can lead to



significant enough changes in the band gap of the CZGSe material that result in the observed decrease of the resonant effect in the Raman spectra.

Finally, the complex analysis presented above reveals that the optimum compositional cationic range that allows achieving high efficiency devices is $[\text{Zn}]/[\text{Ge}] = 1.05\text{--}1.15$. This is in agreement with previously published values for high efficiency solar cells based on CZGSe.^{16,17,29} Furthermore, the efficiency of the devices seems to be less dependent on the $[\text{Cu}]/([\text{Zn}] + [\text{Sn}])$ ratio, as long as the Cu-poor condition is respected (e.g. similar efficiencies were obtained for $[\text{Cu}]/([\text{Zn}] + [\text{Sn}]) = 0.67$ in ref. 29 and 0.78 in ref. 17). Deviations from this optimum cationic ratio range towards stoichiometric and Zn-poor compositions result in an increase of the amount of both V_{Cu} and substitutional defects. This can be assumed to be the reason for the slight increase of the FWHM of the Raman peaks in the spectra measured in the cells with $[\text{Zn}]/[\text{Ge}] \leq 1.0$, as discussed above (see Section 3.1 and Fig. 2). On the contrary, an increase of the Zn concentration in the system with $[\text{Zn}]/[\text{Ge}] > 1.15$, results in the decrease of both V_{Cu} and substitutional defects, in bandgap narrowing, and in an increase of the probability of the detrimental ZnSe secondary phase being formed. Note that the ZnSe phase was found to form also in cells within the optimum cationic range strongly influencing device performance and, as such, the formation of this phase should be controlled during the production process.

Machine learning approach for device performance analysis

The results presented in the previous section clearly indicate that there is a complex dependence between solar cell performance and the different parameters obtained from Raman spectroscopy, such as $V_{\text{Cu}}/\text{Zn}_{\text{Cu}}/\text{Ge}_{\text{Cu}}$ defect concentration or the presence of secondary phases. In this regard, the application of machine learning to the analysis of the Raman spectroscopic data can not only significantly reduce the analysis time, but also yield methodologies for solar cell efficiency prediction. For the present study, a LDA dimension-reduction algorithm was applied. This type of algorithm is widely used for spectral data analysis in different methods and fields of application.^{56–61} This is because of the high dimensionality nature of the spectroscopic data analysed in this work and the ability of LDA to reduce these dimensions to just a few while preserving most of the information, which allows for feature extraction rather than feature selection. Using the Raman spectra of each cell obtained with different wavelengths as input, the LDA algorithm was used to classify the different cells of the combinatorial sample according to the following classification targets: efficiency, V_{oc} and $[\text{Zn}]/[\text{Ge}]$ ratio. The 2-dimensional outputs for each classification targets are presented in Fig. 7 as a function of two discriminants (D1 and D2, with labels a, b, c for the three mentioned targets, respectively), along with the classification groups and training/test scores. As already mentioned above, the amount of data employed for the analysis (196 cells) is far from being considered “big data” and, as such, this methodology and the results presented below should be taken as a first approach and are susceptible to further

improvement for a more precise classification through the use of a higher number of training inputs. In the case of the efficiency and V_{oc} targets, a defined data classification clustering is observed with comparable scores. In Fig. 7a, it can be observed that low efficiency (<3.4%) cells are relatively well classified and separated from the groups with medium (3.4–4.4%) and high (>4.4%) efficiency. On the contrary, for the V_{oc} target, while the cells with the highest voltages are well differentiated, the rest of groups show high overlapping (Fig. 7b). However, despite the relatively low classification score, it is worth noticing that there is no clear overfitting occurring by comparing the individual and overall scores of the algorithm, indicating that a larger data set could greatly improve the classification.

In the case of the $[\text{Zn}]/[\text{Ge}]$ ratio target shown in Fig. 7c, the classification resulted in a clear sequential correlation between the discriminants, from lower to higher ratios. The first 2 groups, cells with 0–0.85 and 0.85–1.05 $[\text{Zn}]/[\text{Ge}]$ ratios, show significant overlapping leading to misclassifications in both training and test data which is reflected in the LDA algorithm scores. On the other hand, a good clustering is shown for the $1.05 < [\text{Zn}]/[\text{Ge}] < 1.15$ and $[\text{Zn}]/[\text{Ge}] > 1.15$ groups. Even though the discriminants in LDA algorithms are of unclear nature and do not necessarily follow an underlying physical concept (at least in a straightforward way), the resulting curve in Fig. 7c shows great resemblance with that obtained from the analytical analysis of the influence of point defects on device performance (Fig. 4a). Bearing this in mind, the relative integrated intensity of the Raman peak at 176 cm^{-1} and efficiency of the solar cells in the combinatorial sample were plotted against the classification discriminants D1c and D2c (Fig. 8). The latter were mainly selected as they show the clearest differentiation between the different classification groups, and, thus, were expected to have a more pronounced dependence from the physical parameters of the solar cell devices. Despite the fact that none of the analytical parameters (relative integrated intensity of the Raman peak and solar cell efficiency) was directly used for the LDA algorithm, a pronounced linear-like correlation between parameters and the LDA discriminants can be observed. Moreover, a closer look to the obtained correlation, allows to define that the points that deviate from the correlation forming a cloud in Fig. 8a (highlighted with a dotted oval) correlate with zero efficiency solar cells. In this way, the behaviour of these cells is probably not related to the concentration of V_{Cu} defects, but to other critical parameters of the devices (*i.e.* presence of ZnSe in absorber, or bad absorber/buffer interface, *etc.*). Similarly, in case of the correlation of cell efficiency with the D2c parameter (Fig. 8b), the points that lie far from the proposed dashed line probably present some additional issues, not directly related to the absorber layer itself that is strictly analysed in the present study. These correlations, however, firstly, allow to get a glimpse of the possible physical meaning behind the D1c and D2c discriminants, and, secondly, prove that there exists a strong correlation of the V_{Cu} defects with the Raman spectra $[\text{Zn}]/[\text{Ge}]$ ratio and solar cell efficiency. In this way, the variations of the V_{Cu} parameter are directly reflected on the other parameters. These multi-variable correlations are intrinsic to the CZGSe material itself and of



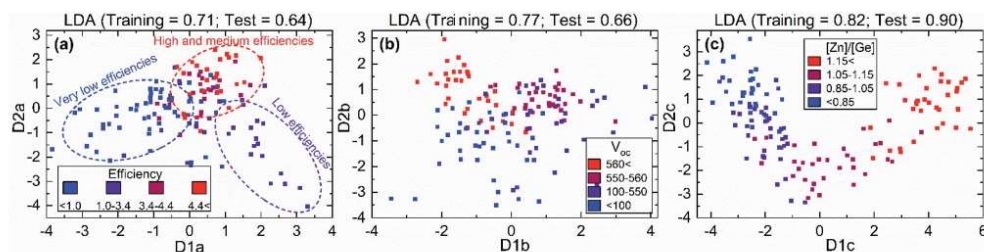


Fig. 7 Results of the machine learning analysis of the Raman spectroscopic data for different classification targets: (a) efficiency, (b) V_{Cu} and (c) $[\text{Zn}]/[\text{Ge}]$ ratio. The training and test scores obtained for each run are presented on the top of the panels.

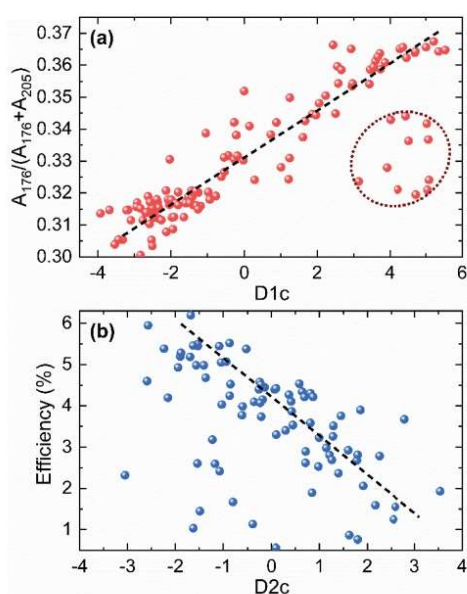


Fig. 8 (a) Relative integrated intensity of the Raman peak at 176 cm^{-1} plotted against discriminant $D1c$, and (b) efficiency of the solar cells plotted against discriminant $D2c$ of the $[\text{Zn}]/[\text{Ge}]$ ratio LDA classification target. The explanation of the dotted oval line can be found in the main text.

importance to derive its properties. Moreover, this finding makes possible to predict solar cell efficiency using only Raman spectra and compositional data. This has an enormous potential both for research and industrial process monitoring applications. In addition, the results presented here illustrate a powerful example of why combinatorial analysis should be established as a standard procedure for the study of complex systems such as thin film solar cells based on chalcogenide compounds in order to deepen into the critical parameters that govern their efficiency.

Conclusions

This work has presented a complex analysis of the formation of secondary phases and point defects under off-stoichiometry conditions, as well as of their effect on PV performance, in a combinatorial CZGSe sample comprised by almost 200 individual solar cells covering a wide range of $[\text{Zn}]/[\text{Ge}]$ compositions in the Cu-poor regime. Firstly, an analytical approach based on Raman spectroscopy and XRF has allowed defining the off-stoichiometric limits of formation of the CZGSe kesterite phase: $0.7 < [\text{Zn}]/[\text{Ge}] < 1.2$. It has been observed that, close to the top limit, the probability of forming ZnSe increases and, above it, it may become the dominant phase, while the formation of other secondary phases is almost negligible in the whole range studied. As for defect formation, the footprint of V_{Cu} and substitutional Zn_{Cu} and Ge_{Cu} defects on the vibrational properties of the CZGSe material has been analysed. Strong V_{Cu} -induced variations in the Raman spectra have been found, especially close to the J- and L-type off-stoichiometry lines, whereas a softer effect (mainly at high wavenumbers) has been observed for the substitutional defects. Finally, the complex analytical correlation of compositional, spectroscopic and optoelectronic data for each of the 200 solar cells, has allowed revealing that V_{Cu} controls the J_{sc} and FF of the devices while the substitutional defects have their main influence on the V_{oc} leading to an optimum cationic compositional range of $[\text{Zn}]/[\text{Ge}] = 1.05\text{--}1.15$ for achieving the high efficiency ($\sim 6\%$) solar cell devices. This has been further explored through the application of an LDA machine learning algorithm. Using just Raman spectra as input, the algorithm employed has been shown to be able to classify the different cells in terms of composition and optoelectronic parameters. These results confirm the deep intrinsic intertwining of the V_{Cu} defect concentration with the Raman spectra, $[\text{Zn}]/[\text{Ge}]$ ratio and solar cell efficiency, and represent a powerful solar cell performance prediction methodology both for research and industrial process monitoring environments. As such, this work not only provides valuable insight for understanding and further developing the CZGSe photovoltaic technology and but also gives a practical example of the potential of combinatorial analysis and machine learning for the study of complex systems in materials research.



Authors contributions

Enric Grau-Luque: formal analysis, software, visualization, writing – original draft; Ikram Anefnaf: formal analysis, resources, writing – review & editing; Nada Benhaddou: formal analysis, resources; Robert Fonoll-Rubio: investigation; Ignacio Becerril-Romero: validation, writing – review & editing; Safae Aazou: funding acquisition, project administration, writing – review & editing; Edgardo Saucedo: conceptualization, funding acquisition, project administration; Zouheir Sekkat: funding acquisition, supervision; Alejandro Perez-Rodriguez: funding acquisition, supervision; Victor Izquierdo-Roca: conceptualization, methodology, validation, writing – review & editing; Maxim Guc: data curation, formal analysis, investigation, methodology, validation, visualization, writing – original draft.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements no 777968 (INFINITE-CELL project) and 952982 (Custom-Art project), and was partially supported by the Spanish Ministry of Science, Innovation and Universities under the WINCOST (ENE2016-80788-C5-1-R) project. Authors from IREC belong to the SEMS (Solar Energy Materials and Systems) Consolidated Research Group of the "Generalitat de Catalunya" (ref. 2017 SGR 862) and are grateful to European Regional Development Funds (ERDF, FEDER Programa Competitivitat de Catalunya 2007–2013). MG acknowledges the financial support from Spanish Ministry of Science, Innovation and Universities within the Juan de la Cierva fellowship (IJC2018-038199-1).

References

- C. Candelise, M. Winkler and R. Gross, *Prog. Photovoltaics Res. Appl.*, 2012, **20**, 816–831.
- W. Wang, M. T. Winkler, O. Gunawan, T. Gokmen, T. K. Todorov, Y. Zhu and D. B. Mitzi, *Adv. Energy Mater.*, 2014, **4**, 1301465.
- J. J. Scragg, T. Ericson, T. Kubart, M. Edoff and C. Platzer-Björkman, *Chem. Mater.*, 2011, **23**, 4625–4633.
- R. Fonoll-Rubio, J. Andrade-Arvizu, J. Blanco-Portals, I. Becerril-Romero, M. Guc, E. Saucedo, F. Peiró, L. Calvo-Barrio, M. Ritzer, C. S. Schnohr, M. Placidi, S. Estradé, V. Izquierdo-Roca and A. Pérez-Rodríguez, *Energy Environ. Sci.*, 2021, **14**, 507–523.
- C. J. Hages, S. Levchenko, C. K. Miskin, J. H. Alsmeyer, D. Abou-Ras, R. G. Wilks, M. Bär, T. Unold and R. Agrawal, *Prog. Photovoltaics Res. Appl.*, 2015, **23**, 376–384.
- S. Giraldo, M. Neuschitzer, T. Thersleff, S. López-Marino, Y. Sánchez, H. Xie, M. Colina, M. Placidi, P. Pistor, V. Izquierdo-Roca, K. Leifer, A. Pérez-Rodríguez and E. Saucedo, *Adv. Energy Mater.*, 2015, **5**, 1501070.
- S. Giraldo, E. Saucedo, M. Neuschitzer, F. Oliva, M. Placidi, X. Alcobé, V. Izquierdo-Roca, S. Kim, H. Tampo, H. Shibata, A. Pérez-Rodríguez and P. Pistor, *Energy Environ. Sci.*, 2018, **11**, 582–593.
- S. Kim, K. M. Kim, H. Tampo, H. Shibata, K. Matsubara and S. Niki, *Sol. Energy Mater. Sol. Cells*, 2016, **144**, 488–492.
- A. D. Collord and H. W. Hillhouse, *Chem. Mater.*, 2016, **28**, 2067–2073.
- J. Andrade-Arvizu, V. Izquierdo-Roca, I. Becerril-Romero, P. Vidal-Fuentes, R. Fonoll-Rubio, Y. Sánchez, M. Placidi, L. Calvo-Barrio, O. Vigil-Galán and E. Saucedo, *ACS Appl. Mater. Interfaces*, 2019, **11**, 32945–32956.
- J. Márquez, H. Stange, C. J. Hages, N. Schaefer, S. Levchenko, S. Giraldo, E. Saucedo, K. Schwarzburg, D. Abou-Ras, A. Redinger, M. Klaus, C. Genzel, T. Unold and R. Mainz, *Chem. Mater.*, 2017, **29**, 9399–9406.
- S. Levchenko, R. Caballero, L. Dermenji, E. V. Telesh, I. A. Victorov, J. M. Merino, E. Arushanov, M. Leon and I. V. Bodnar, *Opt. Mater.*, 2015, **40**, 76–80.
- M. León, S. Levchenko, R. Serna, G. Gurieva, A. Nateprov, J. M. Merino, E. J. Friedrich, U. Fillat, S. Schorr and E. Arushanov, *J. Appl. Phys.*, 2010, **108**, 093502.
- L. Choubrac, M. Bär, X. Kozina, R. Félix, R. G. Wilks, G. Brammertz, S. Levchenko, L. Arzel, N. Barreau, S. Harel, M. Meuris and B. Vermang, *ACS Appl. Energy Mater.*, 2020, **3**, 5830–5839.
- L. Choubrac, G. Brammertz, N. Barreau, L. Arzel, S. Harel, M. Meuris and B. Vermang, *Phys. Status Solidi*, 2018, **215**, 1800043.
- N. Benhaddou, S. Aazou, R. Fonoll-Rubio, Y. Sánchez, S. Giraldo, M. Guc, L. Calvo-Barrio, V. Izquierdo-Roca, M. Abd-Lefdil, Z. Sekkat and E. Saucedo, *J. Mater. Chem. C*, 2020, **8**, 4003–4011.
- S. Sahayaraj, G. Brammertz, B. Vermang, T. Schnabel, E. Ahlswede, Z. Huang, S. Ranjbar, M. Meuris, J. Vleugels and J. Poortmans, *Sol. Energy Mater. Sol. Cells*, 2017, **171**, 136–141.
- R. Gunder, J. A. Márquez-Prieto, G. Gurieva, T. Unold and S. Schorr, *CrystEngComm*, 2018, **20**, 1491–1498.
- M. Neuschitzer, Y. Sanchez, S. López-Marino, H. Xie, A. Fairbrother, M. Placidi, S. Haass, V. Izquierdo-Roca, A. Perez-Rodriguez and E. Saucedo, *Prog. Photovoltaics Res. Appl.*, 2015, **23**, 1660–1667.
- G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA, 2009.
- G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa and A. Mueller, *GetMobile Mob. Comput. Commun.*, 2015, **19**, 29–33.
- S. Schorr, G. Gurieva, M. Guc, M. Dimitrievska, A. Pérez-Rodríguez, V. Izquierdo-Roca, C. S. Schnohr, J. Kim, W. Jo and J. M. Merino, *J. Phys. Energy*, 2020, **2**, 012002.
- A. Lafond, L. Choubrac, C. Guillot-Deudon, P. Deniard and S. Jobic, *Z. Anorg. Allg. Chem.*, 2012, **638**, 2571–2577.
- M. Dimitrievska, H. Xie, A. J. Jackson, X. Fontané, M. Espindola-Rodríguez, E. Saucedo, A. Pérez-Rodríguez, A. Walsh and V. Izquierdo-Roca, *Phys. Chem. Chem. Phys.*, 2016, **18**, 7632–7640.



- 25 M. Y. Valakh, O. F. Kolomys, S. S. Ponomaryov, V. O. Yukhymchuk, I. S. Babichuk, V. Izquierdo-Roca, E. Saucedo, A. Pérez-Rodríguez, J. R. Morante, S. Schorr and I. V. Bodnar, *Phys. Status Solidi RRL*, 2013, **7**, 258–261.
- 26 M. Dimitrievska, A. Fairbrother, A. Pérez-Rodríguez, E. Saucedo and V. Izquierdo-Roca, *Acta Mater.*, 2014, **70**, 272–280.
- 27 A. Ritscher, M. Hoelzel and M. Lerch, *J. Solid State Chem.*, 2016, **238**, 68–73.
- 28 D. M. Többens, G. Gurieva, S. Levchenko, T. Unold and S. Schorr, *Phys. Status Solidi*, 2016, **253**, 1890–1897.
- 29 N. Benhaddou, S. Aazou, Y. Sánchez, J. Andrade-Arvizu, I. Beceril-Romero, M. Guc, S. Giraldo, V. Izquierdo-Roca, E. Saucedo and Z. Sekkat, *Sol. Energy Mater. Sol. Cells*, 2020, **216**, 110701.
- 30 A. Fairbrother, X. Fontané, V. Izquierdo-Roca, M. Placidi, D. Sylla, M. Espindola-Rodríguez, S. López-Mariño, F. A. Pulgarín, O. Vigil-Galán, A. Pérez-Rodríguez and E. Saucedo, *Prog. Photovoltaics Res. Appl.*, 2014, **22**, 479–487.
- 31 Y. Yang, S. C. Liu, W. Yang, Z. Li, Y. Wang, X. Wang, S. Zhang, Y. Zhang, M. Long, G. Zhang, D. J. Xue, J. S. Hu and L. J. Wan, *J. Am. Chem. Soc.*, 2018, **140**, 4150–4156.
- 32 H. R. Chandrasekhar and U. Zwick, *Solid State Commun.*, 1976, **18**, 1509–1513.
- 33 E. Sleetcx, L. Tichý, P. Nagels and R. Callaerts, *J. Non. Cryst. Solids*, 1996, **198–200**, 723–727.
- 34 M. Fujii, S. Hayashi and K. Yamamoto, *Jpn. J. Appl. Phys.*, 1991, **30**, 687–694.
- 35 B. K. Sarkar, A. S. Verma and P. S. Deviprasad, *Phys. B*, 2011, **406**, 2847–2850.
- 36 G. Marcano, C. Rincón, G. Marin, G. E. Delgado, A. J. Mora, J. L. Herrera-Pérez, J. G. Mendoza-Alvarez and P. Rodríguez, *Solid State Commun.*, 2008, **146**, 65–68.
- 37 S. G. Choi, A. L. Donohue, G. Marcano, C. Rincón, L. M. Gedvilas, J. Li and G. E. Delgado, *J. Appl. Phys.*, 2013, **114**, 033531.
- 38 L. E. Valle Rios, K. Neldner, G. Gurieva and S. Schorr, *J. Alloys Compd.*, 2016, **657**, 408–413.
- 39 I. Beceril-Romero, L. Acebo, F. Oliva, V. Izquierdo-Roca, S. López-Mariño, M. Espindola-Rodríguez, M. Neuschitzer, Y. Sánchez, M. Placidi, A. Pérez-Rodríguez, E. Saucedo and P. Pistor, *Prog. Photovoltaics Res. Appl.*, 2018, **26**, 55–68.
- 40 G. Gurieva, L. E. Valle Rios, A. Franz, P. Whitfield and S. Schorr, *J. Appl. Phys.*, 2018, **123**, 161519.
- 41 M. Dimitrievska, A. Fairbrother, E. Saucedo, A. Pérez-Rodríguez and V. Izquierdo-Roca, *Appl. Phys. Lett.*, 2015, **106**, 073903.
- 42 M. Dimitrievska, A. Fairbrother, E. Saucedo, A. Pérez-Rodríguez and V. Izquierdo-Roca, *Sol. Energy Mater. Sol. Cells*, 2016, **149**, 304–309.
- 43 M. Dimitrievska, F. Oliva, M. Guc, S. Giraldo, E. Saucedo, A. Pérez-Rodríguez and V. Izquierdo-Roca, *J. Mater. Chem. A*, 2019, **7**, 13293–13304.
- 44 I. Repins, N. Vora, C. Beall, S. H. Wei, F. Yan, M. Romero, G. Teeter, H. Du, B. To, M. Young and R. Noufi, in *Materials Research Society Symposium Proceedings*, Cambridge University Press, 2012, vol. 1324, pp. 97–108.
- 45 M. Grossberg, J. Krustok, C. J. Hages, D. M. Bishop, O. Gunawan, R. Scheer, S. M. Lyam, H. Hempel, S. Levchenko and T. Unold, *J. Phys. Energy*, 2019, **1**, 044002.
- 46 S. Chen, A. Walsh, X.-G. Gong and S.-H. Wei, *Adv. Mater.*, 2013, **25**, 1522–1539.
- 47 M. M. Islam, M. A. Halim, T. Sakurai, N. Sakai, T. Kato, H. Sugimoto, H. Tampo, H. Shibata, S. Niki and K. Akimoto, *Appl. Phys. Lett.*, 2015, **106**, 243905.
- 48 S. Levchenko, J. Just, A. Redinger, G. Larramona, S. Bourdais, G. Dennler, A. Jacob and T. Unold, *Phys. Rev. Appl.*, 2016, **5**, 024004.
- 49 M. Guc, S. Levchenko, I. V. Bodnar, V. Izquierdo-Roca, X. Fontane, L. V. Volkova, E. Arushanov and A. Pérez-Rodríguez, *Sci. Rep.*, 2016, **6**, 19414.
- 50 M. Guc, A. P. Litvinchuk, S. Levchenko, M. Y. Valakh, I. V. Bodnar, V. M. Dzhagan, V. Izquierdo-Roca, E. Arushanov and A. Pérez-Rodríguez, *RSC Adv.*, 2016, **6**, 13278–13285.
- 51 G. Rey, A. Redinger, J. Sendler, T. P. Weiss, M. Thevenin, M. Guennou, B. El Adib and S. Siebentritt, *Appl. Phys. Lett.*, 2014, **105**, 112106.
- 52 J. J. S. Scragg, J. K. Larsen, M. Kumar, C. Persson, J. Sendler, S. Siebentritt and C. Platzer Björkman, *Phys. Status Solidi*, 2016, **253**, 247–254.
- 53 M. Valentini, C. Malerba, F. Menchini, D. Tedeschi, A. Polimeni, M. Capizzi and A. Mittiga, *Appl. Phys. Lett.*, 2016, **108**, 211909.
- 54 D. P. Halliday, R. Claridge, M. C. J. Goodman, B. G. Mendis, K. Durose and J. D. Major, *J. Appl. Phys.*, 2013, **113**, 223503.
- 55 T. Gershon, B. Shin, T. Gokmen, S. Lu, N. Bojarczuk and S. Guha, *Appl. Phys. Lett.*, 2013, **103**, 193903.
- 56 N. M. Ralbovsky and I. K. Lednev, *Spectrochim. Acta, Part A*, 2019, **219**, 463–487.
- 57 R. Chauhan, R. Kumar, V. Kumar, K. Sharma and V. Sharma, *Forensic Sci. Int.*, 2021, **319**, 110655.
- 58 Z. Guleken, B. Ünübol, R. Bilici, D. Sarbal, S. Toraman, O. Gündüz and S. Erdem Kuruca, *J. Pharm. Biomed. Anal.*, 2020, **190**, 113553.
- 59 R. Choppi, S. Sharma and R. Singh, *Forensic Chem.*, 2020, **17**, 100209.
- 60 T. Visnevschi-Necrasov, J. C. M. Barreira, S. C. Cunha, G. Pereira, E. Nunes and M. B. P. P. Oliveira, *Food Res. Int.*, 2015, **76**, 51–57.
- 61 Y. Wang, J. Zhu and X. Chen, *Optik*, 2020, **224**, 165446.



3.3 Publication 3



spectrapepper: A Python toolbox for advanced analysis of spectroscopic data for materials and devices.

Enric Grau-Luque¹, Fabien Atlan¹, Ignacio Becerril-Romero¹, Alejandro Perez-Rodriguez^{1, 2}, Maxim Guc¹, and Victor Izquierdo-Roca¹

¹ Catalonia Institute for Energy Research (IREC), Jardins de les Dones de Negre 1, 08930 Sant Adrià de Besòs, Spain ² Departament d'Enginyeria Electrònica i Biomèdica, IN2UB, Universitat de Barcelona, C/ Martí i Franqués 1, 08028 Barcelona, Spain

DOI: [10.21105/joss.03781](https://doi.org/10.21105/joss.03781)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Rachel Kurchin](#) 

Reviewers:

- [@stuartcampbell](#)
- [@ksunden](#)

Submitted: 23 September 2021

Published: 19 November 2021

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Statement of need

In recent years, the complexity of novel high-tech materials and devices has increased considerably. This complexity is primarily in the form of increasing numbers of components and broader ranges of applications. An example of the latter is the last generation of thin-film solar cells, which comprise several functional micro- and nano- layers including back contact, absorber, buffer, and transparent front contact. Most of these layers are complex multicomponent compounds (Cu(In,Ga)Se₂, Sb₂Se₃, CdTe, CdS, Zn(O,S), ZnO:Al, etc.) that require fine-tuning of their physicochemical properties to ensure functionality and high performance (Chopra et al., 2004; Powalla et al., 2018). This embedded complexity means that further development of such devices requires advanced characterization and methodologies that allow correlating the physicochemical data of the different layers (chemical composition, structural properties, defect concentration, etc.) with the performance of the final devices in a fast, precise, and reliable way. In this regard, non-destructive methodologies based on spectroscopic characterization techniques (Raman, photoluminescence, X-ray fluorescence, reflectance, transmittance, etc.) have already been demonstrated to possess a high versatility and potential for this type of analyses (Dimitrievska et al., 2019; Guc et al., 2017; Oliva et al., 2016). These spectroscopy-based methodologies can provide deep information that encompasses the complexity of novel materials and devices in a non-destructive way, providing a profound understanding of their properties, failure mechanisms, and possible improvements (Grau-Luque et al., 2021). The latest advances in the application of spectroscopic methodologies for complex materials and devices include the implementation of combinatorial analysis (CA), artificial intelligence (AI) and machine learning (ML), that have been already used in few studies and are slowly becoming more common (Chen et al., 2020). Furthermore, the widespread use of this kind of tools in both laboratory environments and on-line/in-line monitoring of production lines is predicted to shorten development times by a factor of 10, from 10 to 20 years to just a few years (Aspuru-Guzik & Persson, 2018; Correa-Baena et al., 2018; Maine & Garnsey, 2006; Mueller et al., 2016). Unfortunately, several barriers for researchers to implement CA, AI, and ML remain (Gu et al., 2019; Mahmood & Wang, 2021). One of them is the proper pre-processing of spectroscopic data that allows not only to emphasize the relevant changes in the spectra, but also to combine data obtained from different techniques and instruments. Also, the use of ML requires substantial amounts of high-quality data for a precise analysis of the physicochemical parameters of new materials and devices, which necessitates the use of automated systems for massive characterization measurements. In other words, the implementation of automated high-throughput experiments and the capability to perform big-data pre-processing to enhance features of spectroscopic data for ML, and subsequent CA, requires deep theoretical, statistical, analytical, and programming knowledge. Therefore, simple and practical platforms that help researchers to apply such tools are paramount to accelerate their



universal adoption and ultimately shorten the development times of new materials and devices (Butler et al., 2018).

Overview

`spectrapepper` is a Python package that aims to ease and accelerate the use of advanced tools such as machine learning and combinatorial analysis, through simple, straightforward, and intuitive code and functions. This library includes a wide range of tools for spectroscopic data analysis in every step, including data acquisition, processing, analysis, and visualization. Ultimately, `spectrapepper` enables the design of automated measurement systems for spectroscopy and the combinatorial analysis of big data through statistics, artificial intelligence, and machine learning. `spectrapepper` is built in Python 3 (Van Rossum & Drake, 2009), and also uses third-party packages including `numpy` (Harris et al., 2020), `pandas` (Reback et al., 2021), `scipy` (Virtanen et al., 2020), and `matplotlib` (Hunter, 2007), and encourages the user to use `scikit-learn` (Pedregosa et al., 2011) for machine learning applications. `spectrapepper` comes with full documentation, including quick start, examples, and contribution guidelines. Source code and documentation can be downloaded from <https://github.com/spectrapepper/spectrapepper>.

Features

A brief and non-exhaustive list of features includes:

- Baseline removal functions.
- Normalization methods.
- Noise filters, trimming tools, and despiking methods (Barton & Hennelly, 2019; Whitaker & Hayes, 2018).
- Chemometrics algorithms to find peaks, fit curves, and deconvolve spectra.
- Combinatorial analysis tools, such as Spearman, Pearson, and n-dimensional correlation coefficients.
- Tools for ML applications, such as data merging, randomization, and decision boundaries.
- Sample data and examples.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 952982 (Custom-Art project) and Fast Track to Innovation Programme under grant agreement no. 870004 (Solar-Win project). Authors from IREC belong to the SEMS (Solar Energy Materials and Systems) Consolidated Research Group of the "Generalitat de Catalunya" (ref. 2017 SGR 862) and are grateful to European Regional Development Funds (ERDF, FEDER Programa Competitivitat de Catalunya 2007–2013). MG acknowledges the financial support from Spanish Ministry of Science, Innovation and Universities within the Juan de la Cierva fellowship (IJC2018-038199-I).

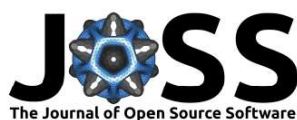
References

Aspuru-Guzik, A., & Persson, K. (2018). Materials Acceleration Platform - Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods with Ar-

- tificial Intelligence. *Report of the Clean Energy Materials Innovation Challenge Expert Workshop, January*, 1–108. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974>
- Barton, S. J., & Hennelly, B. M. (2019). An Algorithm for the Removal of Cosmic Ray Artifacts in Spectral Data Sets. *Applied Spectroscopy*, 73(8), 893–901. <https://doi.org/10.1177/0003702819839098>
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547–555. <https://doi.org/10.1038/s41586-018-0337-2>
- Chen, C., Zuo, Y., Ye, W., Li, X., Deng, Z., & Ong, S. P. (2020). A Critical Review of Machine Learning of Energy Materials. *Advanced Energy Materials*, 10(8), 1903242. <https://doi.org/10.1002/aenm.201903242>
- Chopra, K. L., Paulson, P. D., & Dutta, V. (2004). Thin-film solar cells: an overview. *Progress in Photovoltaics: Research and Applications*, 12(2-3), 69–92. <https://doi.org/10.1002/PIP.541>
- Correa-Baena, J.-P., Hippalgaonkar, K., Van Duren, J., Jaffer, S., Chandrasekhar, V. R., Stevanovic, V., Wadia, C., Guha, S., & Buonassisi, T. (2018). Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule*. <https://doi.org/10.1016/j.joule.2018.05.009>
- Dimitrievska, M., Oliva, F., Guc, M., Giraldo, S., Saucedo, E., Pérez-Rodríguez, A., & Izquierdo-Roca, V. (2019). Defect characterisation in Cu₂ZnSnSe₄ kesterites: Via resonance Raman spectroscopy and the impact on optoelectronic solar cell properties. *Journal of Materials Chemistry A*, 7(21), 13293–13304. <https://doi.org/10.1039/c9ta03625c>
- Grau-Luque, E., Anefnaf, I., Benhaddou, N., Fonoll-Rubio, R., Becerril-Romero, I., Aazou, S., Saucedo, E., Sekkat, Z., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2021). Combinatorial and machine learning approaches for the analysis of Cu₂ZnGeSe₄: Influence of the off-stoichiometry on defect formation and solar cell performance. *Journal of Materials Chemistry A*, 9(16), 10466–10476. <https://doi.org/10.1039/d1ta01299a>
- Gu, G. H., Noh, J., Kim, I., & Jung, Y. (2019). Machine learning for renewable energy materials. *Journal of Materials Chemistry A*, 7(29), 17096–17117. <https://doi.org/10.1039/c9ta02356a>
- Guc, M., Hariskos, D., Calvo-Barrio, L., Jackson, P., Oliva, F., Pistor, P., Perez-Rodriguez, A., & Izquierdo-Roca, V. (2017). Resonant Raman scattering based approaches for the quantitative assessment of nanometric ZnMgO layers in high efficiency chalcogenide solar cells. *Scientific Reports 2017 7:1*, 7(1), 1–11. <https://doi.org/10.1038/s41598-017-01381-4>
- Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). *Array programming with NumPy* (No. 7825; Vol. 585, pp. 357–362). Nature Research. <https://doi.org/10.1038/s41586-020-2649-2>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Mahmood, A., & Wang, J.-L. (2021). Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy & Environmental Science*, 14(1), 90–105. <https://doi.org/10.1039/d0ee02838j>
- Maine, E., & Garnsey, E. (2006). Commercializing generic technology: The case of advanced materials ventures. *Research Policy*, 35, 375–393. <https://doi.org/10.1016/j.respol.2005.12.006>

- tificial Intelligence. *Report of the Clean Energy Materials Innovation Challenge Expert Workshop, January*, 1–108. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974>
- Barton, S. J., & Hennelly, B. M. (2019). An Algorithm for the Removal of Cosmic Ray Artifacts in Spectral Data Sets. *Applied Spectroscopy*, 73(8), 893–901. <https://doi.org/10.1177/0003702819839098>
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547–555. <https://doi.org/10.1038/s41586-018-0337-2>
- Chen, C., Zuo, Y., Ye, W., Li, X., Deng, Z., & Ong, S. P. (2020). A Critical Review of Machine Learning of Energy Materials. *Advanced Energy Materials*, 10(8), 1903242. <https://doi.org/10.1002/aenm.201903242>
- Chopra, K. L., Paulson, P. D., & Dutta, V. (2004). Thin-film solar cells: an overview. *Progress in Photovoltaics: Research and Applications*, 12(2-3), 69–92. <https://doi.org/10.1002/PIP.541>
- Correa-Baena, J.-P., Hippalgaonkar, K., Van Duren, J., Jaffer, S., Chandrasekhar, V. R., Stevanovic, V., Wadia, C., Guha, S., & Buonassisi, T. (2018). Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule*. <https://doi.org/10.1016/j.joule.2018.05.009>
- Dimitrievska, M., Oliva, F., Guc, M., Giraldo, S., Saucedo, E., Pérez-Rodríguez, A., & Izquierdo-Roca, V. (2019). Defect characterisation in Cu₂ZnSnSe₄ kesterites: Via resonance Raman spectroscopy and the impact on optoelectronic solar cell properties. *Journal of Materials Chemistry A*, 7(21), 13293–13304. <https://doi.org/10.1039/c9ta03625c>
- Grau-Luque, E., Anefnaf, I., Benhaddou, N., Fonoll-Rubio, R., Becerril-Romero, I., Aazou, S., Saucedo, E., Sekkat, Z., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2021). Combinatorial and machine learning approaches for the analysis of Cu₂ZnGeSe₄: Influence of the off-stoichiometry on defect formation and solar cell performance. *Journal of Materials Chemistry A*, 9(16), 10466–10476. <https://doi.org/10.1039/d1ta01299a>
- Gu, G. H., Noh, J., Kim, I., & Jung, Y. (2019). Machine learning for renewable energy materials. *Journal of Materials Chemistry A*, 7(29), 17096–17117. <https://doi.org/10.1039/c9ta02356a>
- Guc, M., Hariskos, D., Calvo-Barrio, L., Jackson, P., Oliva, F., Pistor, P., Perez-Rodriguez, A., & Izquierdo-Roca, V. (2017). Resonant Raman scattering based approaches for the quantitative assessment of nanometric ZnMgO layers in high efficiency chalcogenide solar cells. *Scientific Reports 2017 7:1*, 7(1), 1–11. <https://doi.org/10.1038/s41598-017-01381-4>
- Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). *Array programming with NumPy* (No. 7825; Vol. 585, pp. 357–362). Nature Research. <https://doi.org/10.1038/s41586-020-2649-2>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Mahmood, A., & Wang, J.-L. (2021). Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy & Environmental Science*, 14(1), 90–105. <https://doi.org/10.1039/d0ee02838j>
- Maine, E., & Garnsey, E. (2006). Commercializing generic technology: The case of advanced materials ventures. *Research Policy*, 35, 375–393. <https://doi.org/10.1016/j.respol.2005.12.006>

3.4 Publication 4



puDu: A Python library for agnostic feature selection and explainability of Machine Learning spectroscopic problems

Enric Grau-Luque¹, Ignacio Becerril-Romero¹, Alejandro Perez-Rodriguez^{1,2}, Maxim Guc¹, and Victor Izquierdo-Roca¹

¹ Catalonia Institute for Energy Research (IREC), Jardins de les Dones de Negre 1, 08930 Sant Adrià de Besòs, Spain. ² Departament d'Enginyeria Electrònica i Biomèdica, IN2UB, Universitat de Barcelona, C/ Martí i Franqués 1, 08028 Barcelona, Spain.

DOI: [10.21105/joss.05873](https://doi.org/10.21105/joss.05873)

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [Arfon Smith](#)

Reviewers:

- [@hbaniecki](#)
- [@aksholokhov](#)

Submitted: 10 July 2023

Published: 12 December 2023

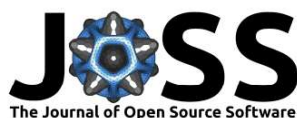
License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Statement of need

Spectroscopic techniques (e.g. Raman, photoluminescence, reflectance, transmittance, X-ray fluorescence) are an important and widely used resource in different fields of science, such as photovoltaics (Fonoll-Rubio et al., 2022) (Grau-Luque et al., 2021), cancer (Bellisola & Sorio, 2012), superconductors (Fischer et al., 2007), polymers (Easton et al., 2020), corrosion (Haruna et al., 2023), forensics (P. V. Bhatt & Rawtani, 2023), and environmental sciences (Estefany et al., 2023), to name just a few. This is due to the versatile, non-destructive and fast acquisition nature that provides a wide range of material properties, such as composition, morphology, molecular structure, optical and electronic properties. As such, machine learning (ML) has been used to analyze spectral data for several years, elucidating their vast complexity, and uncovering further potential on the information contained within them (Goodacre, 2003) (Luo et al., 2022). Unfortunately, most of these ML analyses lack further interpretation of the derived results due to the complex nature of such algorithms. In this regard, interpreting the results of ML algorithms has become an increasingly important topic, as concerns about the lack of interpretability of these models have grown (Burkart & Huber, 2021). In natural sciences (like materials, physical, chemistry, etc.), as ML becomes more common, this concern has gained especial interest, since results obtained from ML analyses may lack scientific value if they cannot be properly interpreted, which can affect scientific consistency and strongly diminish the significance and confidence in the results, particularly when tackling scientific problems (Roscher et al., 2020).

Even though there are methods and libraries available for explaining different types of algorithms such as SHAP (Lundberg et al., 2017), LIME (Ribeiro et al., 2016), or GradCAM (Selvaraju et al., 2017), they can be difficult to interpret or understand even for data scientists, leading to problems such as miss-interpretation, miss-use and over-trust (Kaur et al., n.d.). Adding this to other human-related issues (Krishnâ1 et al., 2022), researchers with expertise in natural sciences with little or no data science background may face further issues when using such methodologies (Zhong et al., 2022). Furthermore, these types of libraries normally aim for problems composed of image, text, or tabular data, which cannot be associated in a straightforward way with spectroscopic data. On the other hand, time series (TS) data shares similarities with spectroscopy, and while still having specific needs and differences, TS dedicated tools can be a better approach. Unfortunately, despite the existence of several libraries that aim to explain models for TS with the potential to be applied to spectroscopic data, they are mostly designed for a specialized audience, and many are model-specific (Rojat et al., 2021). Moreover, spectral data normally manifests as an array of peaks that are typically overlapped and can be distinguished by their shape, intensity, and position. Minor shifts in



these patterns can indicate significant alterations in the fundamental properties of the subject material. Conversely, pronounced variations in the other case might only indicate negligible differences. Therefore, comprehending such alterations and their implications is paramount. This is still true with ML spectroscopic analysis where the spectral variations are still of primary concern. In this context, a tool with an easy and understandable approach that offers spectroscopy-aimed functionalities that allow to aim for specific patterns, areas, and variations, and that is beginner and non-specialist friendly is of high interest. This can help the different stakeholders to better understand the ML models that they employ and considerably increase the transparency, comprehensibility, and scientific impact of ML results (U. Bhatt et al., 2020) (Belle & Papantonis, 2021).

Overview

pudu is a Python library that quantifies the effect of changes in spectral features over the predictions of ML models and their effect to the target instances. In other words, it perturbs the features in a predictable and deliberate way and evaluates the features based on how the final prediction changes. For this, four main methods are included and defined. *Importance* quantifies the relevance of the features according to the changes in the prediction. Thus, this is measured in probability or target value difference for classification or regression problems, respectively. *Speed* quantifies how fast a prediction changes according to perturbations in the features. For this, the *importance* is calculated at different perturbation levels, and a line is fitted to the obtained values and the slope, or the rate of change of *importance*, is extracted as the *speed*. *Synergy* indicates how features complement each other in terms of prediction change after perturbations. Finally, *re-activations* account for the number of unit activations in a Convolutional Neural Network (CNN) that after perturbation, the value goes above the original activation criteria. The latter is only applicable for CNNs, but the rest can be applied to any other ML problem, including CNNs. To read in more detail how these techniques work, please refer to the [definitions](#) in the documentation.

pudu is versatile as it can analyze classification and regression algorithms for both 1- and 2-dimensional problems, offering plenty of flexibility with parameters, and the ability to provide localized explanations by selecting specific areas of interest. To illustrate this, [Figure 1](#) shows two analysis instances using the same *importance* method but with different parameters. Additionally, its other functionalities are shown in examples using scikit-learn (Pedregosa et al., 2011), keras (Chollet et al., 2018), and localreg (Marholm, 2022) are found in the documentation, along with XAI methods including LIME and GradCAM.

pudu is built in Python 3 (Van Rossum & Drake, 2009) and uses third-party packages including numpy (Harris et al., 2020), matplotlib (Caswell et al., 2021), and keras. It is available in both PyPI and conda, and comes with complete documentation, including quick start, examples, and contribution guidelines. Source code and documentation are available in <https://github.com/pudu-py/pudu>.

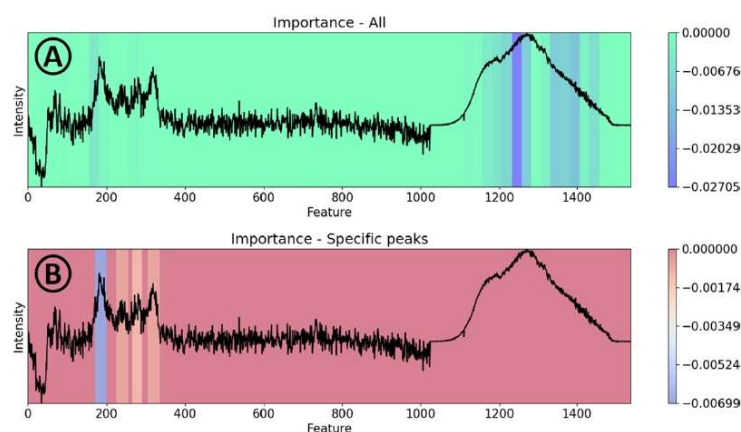


Figure 1: Two ways of using the same method *importance* by A) using a sequential change pattern over all the spectral features and B) selecting peaks of interest. These spectras are measured from thin-film photovoltaic samples and are correlated to their performance using ML, as explained in (Fonoll-Rubio et al., 2022). In A), the vector was divided in window sizes of 25 pixels were perturbed individually. The impact of the peak in the range of 1200-1400 opaques the impact of the rest. In contrast, in B) specific ranges are defined, so only the first four main peaks are selected to be analyzed and better visualize their impact in the prediction.

Acknowledgements

Co-funded by the European Union (GA N° 101058459 Platform-ZERO). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union (EU) or European Health and Digital Executive Agency (HADEA). Neither the EU nor the granting authority can be held responsible for them. This project has received funding from the EU's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie GA N° 801342 (Tecniospring INDUSTRY) and the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ). This work has received funding from the EU's Horizon 2020 Research and Innovation Programme under GA N° 958243 (SUNRISE project). Authors from IREC belong to the MNT-Solar Consolidated Research Group of the "Generalitat de Catalunya" (ref. 2021 SGR 01286) and are grateful to European Regional Development Funds (ERDF, FEDER Programa Competitivitat de Catalunya 2007–2013).

Authors contribution with CRediT

- Enric Grau-Luque: Conceptualization, Data curation, Software, Writing – original draft
- Ignacio Becerril-Romero: Investigation, Methodology, Writing – review & edition
- Alejandro Perez-Rodriguez: Funding acquisition, Project administration, Resources, Supervision
- Maxim Guc: Formal analysis, Validation, Methodology, Writing – review & edition
- Victor Izquierdo-Roca: Funding acquisition, Project administration, Supervision

References

- Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4, 39. <https://doi.org/10.3389/FDATA.2021.688969>
- Bellisola, G., & Sorio, C. (2012). Infrared spectroscopy and microscopy in cancer research and diagnosis. *American Journal of Cancer Research*, 2(1), 1. [/pmc/articles/PMC3236568/](https://pubmed.ncbi.nlm.nih.gov/23236568/) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3236568/>
- Bhatt, P. V., & Rawtani, D. (2023). Spectroscopic Analysis Techniques in Forensic Science. *Modern Forensic Tools and Devices: Trends in Criminal Investigation*, 149–197. <https://doi.org/10.1002/9781119763406.CH8>
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). *Explainable Machine Learning in Deployment*. <https://doi.org/10.1145/3351095.3375624>
- Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70, 245–317. <https://doi.org/10.1613/JAIR.1.12228>
- Caswell, T. A., Droettboom, M., Lee, A., Andrade, E. S. de, Hunter, J., Hoffmann, T., Firing, E., Klymak, J., Stansby, D., Varoquaux, N., Nielsen, J. H., Root, B., May, R., Elson, P., Seppänen, J. K., Dale, D., Lee, J.-J., McDougall, D., Straw, A., ... Ivanov, P. (2021). *matplotlib/matplotlib: REL: v3.4.2*. <https://doi.org/10.5281/ZENODO.4743323>
- Chollet, F., Others, Chollet, F., & Others. (2018). Keras: The Python Deep Learning library. *Astrophysics Source Code Library*, ascl:1806.022. <https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C/abstract>
- Easton, C. D., Kinnear, C., McArthur, S. L., & Gengenbach, T. R. (2020). Practical guides for x-ray photoelectron spectroscopy: Analysis of polymers. *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, 38(2). <https://doi.org/10.1116/1.5140587>
- Estefany, C., Sun, Z., Hong, Z., & Du, J. (2023). Raman spectroscopy for profiling physical and chemical properties of atmospheric aerosol particles: A review. *Ecotoxicology and Environmental Safety*, 249, 114405. <https://doi.org/10.1016/J.ECOENV.2022.114405>
- Fischer, Ø., Kugler, M., Maggio-Aprile, I., Berthod, C., & Renner, C. (2007). Scanning tunneling spectroscopy of high-temperature superconductors. *Reviews of Modern Physics*, 79(1), 353–419. <https://doi.org/10.1103/REVMODPHYS.79.353>
- Fonoll-Rubio, R., Paetel, S., Grau-Luque, E., Becerril-Romero, I., Mayer, R., Pérez-Rodríguez, A., Guc, M., & Izquierdo-Roca, V. (2022). Insights into the Effects of RbF-Post-Deposition Treatments on the Absorber Surface of High Efficiency Cu(In,Ga)Se₂ Solar Cells and Development of Analytical and Machine Learning Process Monitoring Methodologies Based on Combinatorial Analysis. *Advanced Energy Materials*, 2103163. <https://doi.org/10.1002/AENM.202103163>
- Goodacre, R. (2003). Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vibrational Spectroscopy*, 32(1), 33–45. [https://doi.org/10.1016/S0924-2031\(03\)00045-6](https://doi.org/10.1016/S0924-2031(03)00045-6)
- Grau-Luque, E., Anefnaf, I., Benhaddou, N., Fonoll-Rubio, R., Becerril-Romero, I., Aazou, S., Saucedo, E., Sekkat, Z., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2021). Combinatorial and machine learning approaches for the analysis of Cu₂ZnGeSe₄: influence of the off-stoichiometry on defect formation and solar cell performance. *Journal of Materials Chemistry A*, 9(16), 10466–10476. <https://doi.org/10.1039/d1ta01299a>

- Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). *Array programming with NumPy* (No. 7825; Vol. 585, pp. 357–362). *Nature Research*. <https://doi.org/10.1038/s41586-020-2649-2>
- Haruna, K., Obot, I. B., & Saleh, T. A. (2023). Infrared Spectroscopy in Corrosion Research. *Corrosion Science*, 261–289. <https://doi.org/10.1201/9781003328513-9>
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (n.d.). *Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning*. <https://doi.org/10.1145/3313831.3376219>
- Krishná1, S., Han¹, T. H., Gu, A., Pombra, J., Jabbari, S., Wu, Z. S., & Lakkaraju, H. (2022). *The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective*. <https://arxiv.org/abs/2202.01602v3>
- Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30. <https://github.com/slundberg/shap>
- Luo, R., Popp, J., & Bocklitz, T. (2022). Deep Learning for Raman Spectroscopy: A Review. *Analytica*, 3(3), 287–301. <https://doi.org/10.3390/analytica3030020>
- Marholm, S. (2022). *sigvaldm/localreg: Multivariate RBF output*. <https://doi.org/10.5281/ZENODO.6344451>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python* (Vol. 12, pp. 2825–2830). <http://scikit-learn.sourceforge.net>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. <https://doi.org/10.48550/arxiv.1602.04938>
- Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., & Díaz-Rodríguez, N. (2021). *Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey*. <https://arxiv.org/abs/2104.00950v1>
- Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8, 42200–42216. <https://doi.org/10.1109/ACCESS.2020.2976199>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). <https://doi.org/10.1109/iccv.2017.74>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*; CreateSpace. *Scotts Valley, CA*, 242. ISBN: 978-1-4414-1269-0
- Zhong, X., Gallagher, B., Liu, S., Kaikhura, B., Hiszpanski, A., & Han, T. Y. J. (2022). Explainable machine learning in materials science. *Npj Computational Materials* 2022 8:1, 8(1), 1–19. <https://doi.org/10.1038/s41524-022-00884-7>

4. FURTHER EXPLORATORY EXPERIMENTS

4.1 Introduction

After exploring the dimension reduction approach with PCA, LDA and PC-LDA, in the last period of this thesis two new approaches were explored for AI driven analysis for TFPV materials and devices as an extension and logical step forward for improvement of the methodology used, as schematically shown in Figure 4-1. These two techniques are of different nature and tackle two different problems and questions that derive from ML results and can be used together to gain further insights from an experiment. The first question is about out-of-distribution (OOD) properties of TFPV materials. In other words, this first extension of the methodology explores an AI driven way to detect hypothetical compositional characteristics of TFPV devices for improved performance. This is done with a combination of the PC-LDA and CNN using explainability techniques and Multivariate Non-Linear Regressions (MVNLR) for prediction of these OOD properties. Essentially, a PC-LDA and CNN models are trained to perform the same classification task and are analyzed to see what features are more important, and this information is cross validated between the models to see what features are consistently affecting the results, regardless of the model used. This combination is motivated by the objective of enhancing our comprehension of the combinatorial spectroscopic data at our disposal and facilitating data-informed decisions. As this kind of research in this particular field is still fresh and largely unpublished, it promises significant potential for impact. To maximize the utilization of CNNs, two unique networks are designed: a 1D network that directly interfaces with the existing data, and a 2D network requiring data transformation into a 2D matrix (an image), a novel process for this kind of data. The incorporation of a 2D CNN is motivated by the fact that these models more widely used and studied and also attract more interest from the scientific community. Additionally, 2D CNNs have the benefit of exploiting other data properties that may be overlooked by a 1D version, such as spatial correlations among different characterization techniques (8 spectroscopic measurements are used in this experiment) and the detection of intricate patterns. Upon completion of the CNNs' training and testing, sensitivity analysis and GradCAM techniques are applied to specific instances. This allows to gain insights into the algorithms' decision-making processes, yielding useful explanations that will further facilitate a more thorough dataset analysis.

The second approach explores in more depth the use of CNNs for the classification of TFPV using a modern technique called dissection. This method studies the structure and behavior of the CNN with the premise that deeper understanding of such models may allow the user to modify the model and obtain enhanced and customized results and to obtain deeper insights into the data. This is motivated by the fact that ML techniques, particularly CNNs, have revolutionized various scientific disciplines and offer unprecedented capabilities for data analysis, prediction, and even the discovery of new materials with desirable properties. However, the application of CNNs in the field of energy materials faces unique challenges and opportunities that warrant focused investigation. In particular, a critical issue, as discussed in the introduction of this work, is the

interpretability of these type of models. While CNNs are powerful tools for pattern recognition and prediction, their 'black-box' nature makes them difficult to interpret, becoming an increasingly important topic in research and applications [101]. This is particularly problematic in material science, where understanding the underlying mechanisms is crucial for the development and optimization of new materials. The lack of interpretability can also affect the trust that researchers place in the model's results, which is essential for their broader acceptance and application in the scientific community. The above signifies that even though results from ML models can seem impactful due to good classifications and logistic regressions, they may lack real scientific value due to the lack of explanation, diminishing the significance of the results [90]. Therefore, there is a pressing need to develop and apply methods that not only improve the performance of CNNs but also make their decision-making processes more transparent. Enhancing interpretability can increase researchers' trust in these models, ensuring that they are not just statistically accurate but also scientifically meaningful. To tackle this problem, several tools have raised in the past years, including techniques such as SHAP [122], LIME [89], and GradCAM [123]. However, it is well documented that different explainability methods show disagreement in metrics such as feature importance rank and sign agreement (whether or not a feature has a positive or negative impact in the output), highlighting this important problem with post-hoc explanations [110]. A way to improve this is to use methods that not only aim to explain decisions of models but also try to reason their inner workings. Furthermore, understanding the intricacies of a CNN model can facilitate its manipulation and improvement, thereby increasing its reliability and efficacy. This can be performed with techniques such as dissection [121], which attempts to align individual units of a CNN with local features in the data. This allows the model to become more interpretable by assigning specific roles to individual units. For instance, researchers have been able to observe that there are units that are activated by specific concepts in images, such as objects, parts, materials, and colors [124]. This allows to deliberately and intentionally manipulate models to achieve desired results, such as removing or including specific objects in images, aiming to increase the interpretability, completeness, reliability and efficacy. This technique, in this case, is adapted to work with spectroscopic data and, therefore, spectroscopic features.

This section shows these two expansions of the methodology with a combinatorial CZTSe-based sample divided in 225 cells, each measured 8 times under different conditions, 4 with PL and 4 with Raman using 785 nm, 532 nm, 442 nm and 325 nm excitation wavelengths each, along with XRF and IV measurements. Thus, the same data derived from the same sample and processing are used for both approaches, which is explained below. Both approaches are natural next steps from the methodology described in this thesis (Section 2), and this section introduces the preliminary results obtained in the explorative work using them. A more developed technique of this kind may be used for synthetic experiments to skip and avoid physical ones, reducing research and development times further beyond what automatization in experiments can achieve.

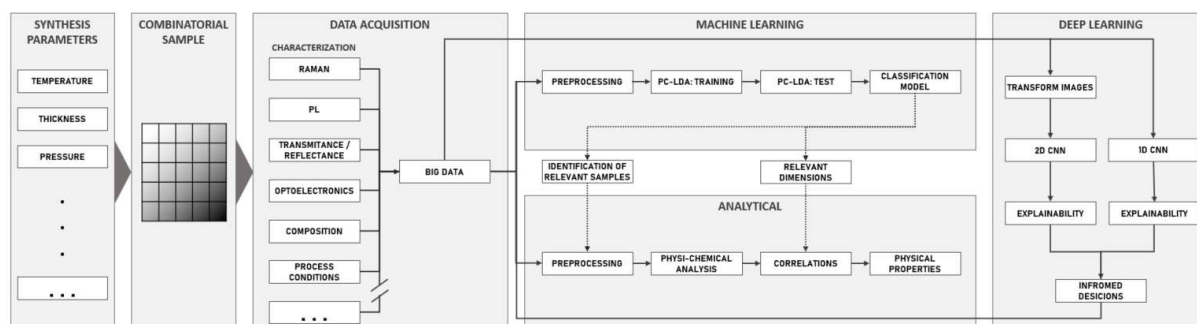


Figure 4-1: Modified workflow of the methodology to include 1D and 2D CNNs along with respective explainability techniques.

4.2 Methodology

4.2.1 Sample

The samples chosen for this experiment consist of 225 solar cells, sourced from a 15x15 combinatorial CZTSe-based sample. The absorbers' composition (Figure 4-2) presents noticeable variability in terms of Copper, Tin, and Zinc, resulting in significant in-sample diversity across all optoelectronics. The highest V_{OC} values typically emerge in proximity to the central and upper regions of the sample, exhibiting middle $[Zn]/[Sn]$ ratio values and leaning towards a mid-low to low $[Cu]/[Zn]$ ratio. On the contrary, the peak J_{SC} values are usually found on the lower side of the sample, coinciding with medium and mid-low $[Zn]/[Sn]$ ratios and medium $[Cu]/[Sn]$ ratios. Highest efficiencies are detected closer to the sample's center, with diminishing values in a radial pattern. An important observation is that several cells appear devoid of any optoelectronic properties, mainly in the right edge and bottom and left-bottom edges. These samples, despite their seeming lack of utility, are preserved in the analysis for potential spectral property exploration if deemed necessary.

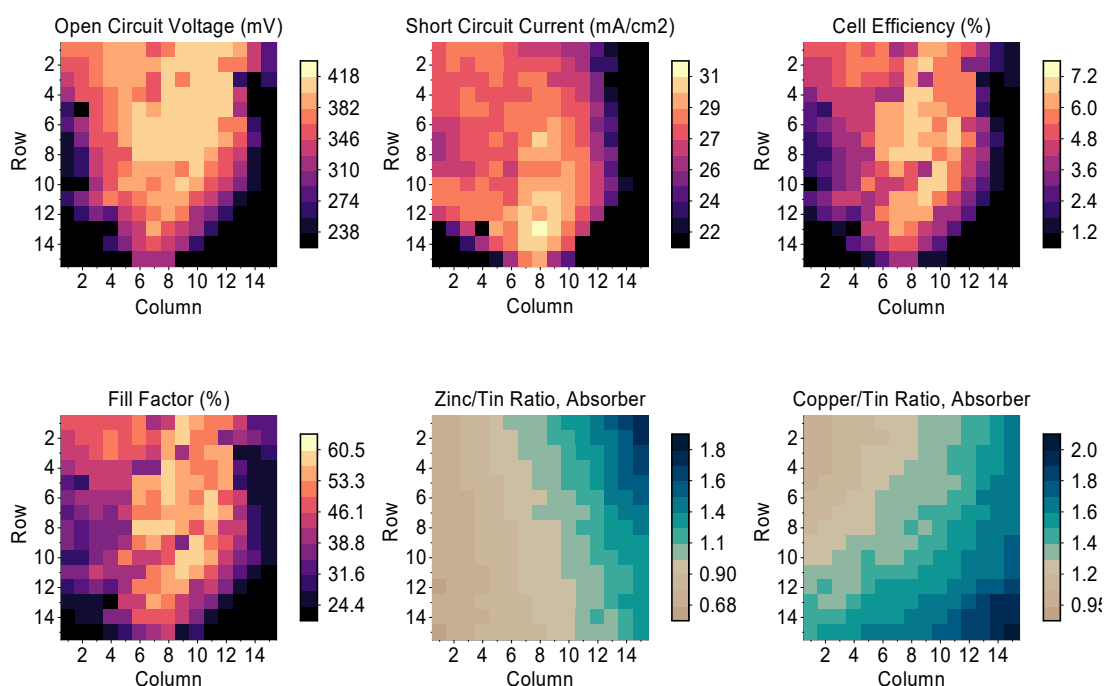


Figure 4-2: V_{OC} , J_{SC} , and Efficiency optoelectronic for the combinatorial sample (upper row) and Fill Factor and Compositional ratios of Zn/Sn and Cu/Sn (lower row).

4.2.2 Sample characterization

Each cell is measured using XRF, IV, Raman, and PL techniques, with the latter two employed under four distinct wavelengths: 325 nm, 442 nm, 532 nm, and 785 nm. Consequently, when all spectra are fused, a 1-D vector of length 15,344 is produced. Information regarding laser power, acquisition times, and averages for each of the techniques can be found in Table 4-1. Additionally displayed in the same table, the total acquisition time represents the collective minutes of acquisition throughout the entire sample, that is, taking into consideration all of the 225 cells, culminating in a grand total of 40 hours and 20 minutes. The time recorded does not factor in elements such as sample preparation, equipment calibration, processing, or movement times. It is worth noticing, that in the current study the main focus was on the research objectives and on testing of the methodologies, thus the used timings were selected to obtained relatively high quality of the spectra (with low signal to noise ratio), and these values can be further decreased for the more specific applications.

Table 4-1: Experimental setup and parameters of the spectroscopic techniques. Total Aq. Times is the amount in minutes of the total time needed to perform the measurement considering all the 225 cells.

	Raman				PL			
Wavelength (nm)	785	532	442	325	785	532	442	325
Power (mW)	3.9	3.5	5.5	3.6	3.75	5.2	1.2	3.2
Acquisition time (s)	40	20	20	120	1	0.2	0.1	0.3
Acquisitions (n°)	3	5	3	3	3	5	3	3
Total Aq. Time (m)	450	375	225	1350	11.3	3.75	1.13	3.34
Spectra length	2000	1024	1024	984	1210	512	5076	3514

4.2.3 Data processing

Following the methodology, the eight spectra were fused into a single vector for each cell, as has been performed in past studies [125][92]. This approach allows our algorithm to discern between techniques and concurrently incorporate their respective benefits for the classification process. For this merge to be effective, the techniques were scaled to a comparable scale. The Raman spectras were normalized to their main peak ratios for each of the wavelengths, meanwhile for the PL measurements were normalized to the global maximum of each of the wavelengths. With this, the maximum value is restrained to 1 for all of the techniques, and thus remain comparable when fusion is performed. For all measured points, the spectra were merged in descending source wavelength with Raman first followed by PL. In other words, first Raman 785, 532, 442, and 325 nm followed by PL 785, 532, 442, and 325 nm. The resulting vector of 15,344 is then reduced to a final it is of length 14,640 after deleting small ranges at the beginning and end of each spectra that normally contain artifacts left over from data processing, which focuses on the main ranges of this kind of data, normally on more central regions (Figure 4-3).

For the 2D CNN a final step is performed to transform the is represented as images of 120x120 pixels, which contains a total of 14,400 values. To match the obtained vector of length 14,640 after data processing, 240 additional pixels are removed from the end of the vector, without affecting any of the relevant areas.

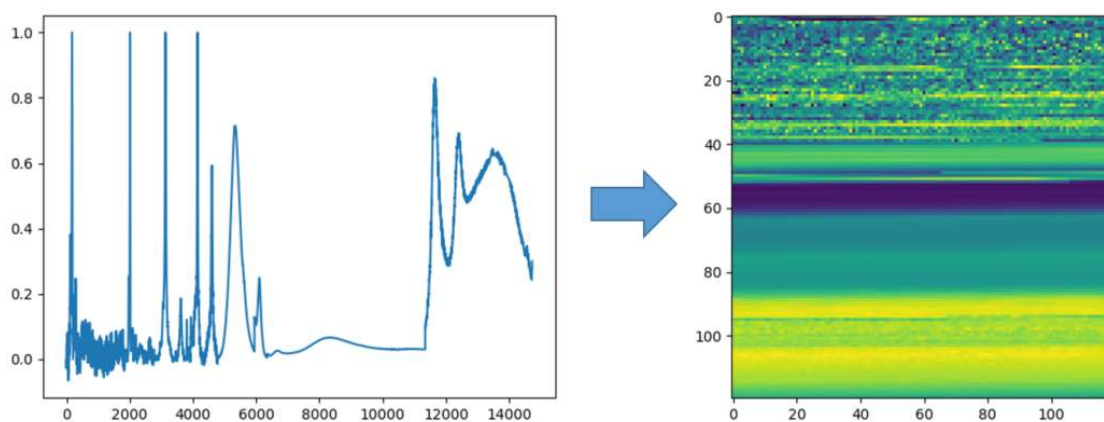


Figure 4-3: Schematic of the transformation of the vector of length 14,640 to an image of 120x120 pixels. As $120 \times 120 = 14,400$, 240 pixels have to be deleted in order to be reshaped. In this case, the last 240 pixels were deleted as they offer little information and is the easiest way to accomplish this. Other approaches are possible, such as interpolating or deleting smaller sections across the spectras.

4.2.4 PC-LDA

In the PC-LDA model, the chosen target is V_{OC} , which is intended for the classification objective. The feature selection is carried out using the 1D vector of length 15,344. This approach is expected to enhance observation of the methodologies that optimally inform the analyses of V_{OC} . With V_{OC} designated as the target, the PC-LDA model undergoes training for four classification groups: $259 < V_{OC}$, $259 \leq V_{OC} < 353$, $353 \leq V_{OC} < 391$, and $391 \leq V_{OC}$. Upon completion of the initial iteration, a sensitivity analysis is initiated to identify the sections of the spectra that are most relevant to the algorithm. This permits the elimination of spectra sections that hinder processing, ending with a feature length of 14,640. Consequently, this enables a further iteration to secure more faithful results and ultimately the selection of crucial sections for performing logistic regression over V_{OC} .

4.2.5 1D and 2D CNN

For the OOD procedure, the architecture of the 1D CNN is composed of three distinctive 1D convolutional layers, each followed by batch normalization, max pooling, and a dropout layer. Initially, a 1D convolutional layer utilizes 4 filters with a 2-unit kernel and 1-unit stride in the 14640×1 input vector. This layer's output undergoes a sequence of normalization, max pooling with a 16-unit pool size, and dropout at a rate of 0.20. A second 1-D convolutional layer with 8 filters and a 4-unit kernel followed by batch normalization, max pooling with an 8-unit pool size, and dropout with rate of 0.20. The convolutional layers used ReLU as activation. Following these layers, a 1-D Global Average Pooling layer links to the final dense layer with four neurons and a softmax activation function, which provides a probability distribution over four classes. In contrast

with the 1D CNN, the 2D CNN shows 2 convolutional layers (1 input, 1 hidden) and one output dense layer. The first convolutional layer utilizes 2 filters with a 2x2 kernel and 1x1 stride, followed by a dropout layer with 0.3 rate. The second layer utilizes 4 filters with a 4x4 kernel and 2x2 stride, followed by a dropout layer with the same 0.3 rate. Both layers use LeakyReLU with a -0.01 coefficient. These are followed by a Flatten layer and then the final dense layer of 4 units and a softmax activation. Just as the PC-LDA, the 1D and 2D CNNs are trained to classify V_{OC} in four classification groups: $259 < V_{OC}$, $259 \leq V_{OC} < 353$, $353 \leq V_{OC} < 391$, and $391 \leq V_{OC}$.

For performing dissection, however, a slightly different CNN architecture was used. This is due to the fact that this CNN was design in a later time with the goal of improving results observed in previous attempts. Specifically, the constructed CNN is designed to classify the same data 4 classes according to the V_{OC} value of the cell as measured in the IV curve: $V_{OC} < 303$, $303 \leq V_{OC} < 363$, $363 \leq V_{OC} < 396$, and $396 \leq V_{OC}$. The resulting CNN is composed of three distinctive 1-D convolutional layers, each followed by batch normalization, max pooling, and a dropout layer. Initially, a 1-D convolutional layer utilizes 4 filters with an 8-unit kernel and 2-unit stride in the 14640x1 input vector. This layer's output undergoes a sequence of normalization, max pooling with a 16-unit pool size, and dropout at a rate of 0.25. A second 1-D convolutional layer with 8 filters and a 4-unit kernel followed by batch normalization, max pooling with an 8-unit pool size, and dropout with rate of 0.25. Lastly, a third 1-D convolutional layer with 16 filters of 16-unit kernel follow too by batch normalization, 8-unit max pooling, and dropout with 0.25 rate. All convolutional layers used LeakyReLU as activation function with a threshold of 0.05. Following these layers, a 1-D Global Average Pooling layer links to the final dense layer with four neurons and a softmax activation function, which provides a probability distribution over four classes.

4.2.6 Explainability

For the PC-LDA, sensitivity analysis is performed using the pudu library. For both of the 1D CNN and 2D CNN, sensitivity analysis along with GradCAM are used. This last one is applied for each of the convolutional layers of the networks. This allows comparing their results for more insightful information about the classification processes.

For the last CNN analysis, the pudu library was also used. In this case, the sensitivity analysis was performed to quantify not only the probability impact of feature changes but also the activation values of the units of the last convolutional layer.

4.3 Results

4.3.1 Exploration of OOD properties

Scores for the PC-LDA are exhibited in Figure 4-4 in the confusion matrices of A) and B). High training and test scores are evident, with a floor value of 0.82 for the 259 – 353 mV range in the

training set. The test set indicates a lower score of 0.73 for the cells performing at the lowest level; nonetheless, overfitting is discernible in the highest-performing data points, which achieve 100% accuracy. Sensitivity analysis (SA) reveals a significant influence of artifacts within the spectra. In other words, peaks and valleys in negligible zones of the spectra, especially at the start and tail sections, yield a relatively large impact on the classification of specific instances, as demonstrated in Figure 4-5. This information allows for the removal of these spectra zones and the training of an alternate PC-LDA model. As shown in the figure, the resultant vector is more concise, now measuring 14,640 in length. Figure 4-4C and D also displays how scores change following this process, with marked enhancements in both the training and test set. The mitigation of overfitting is clear, indicated by closer scores within groups and between training and test scores. This suggests a more equitable evaluation and comparison of the data. The established clusters are depicted in **Error! No se encuentra el origen de la referencia.** (left), exhibiting clear separation and continuity.

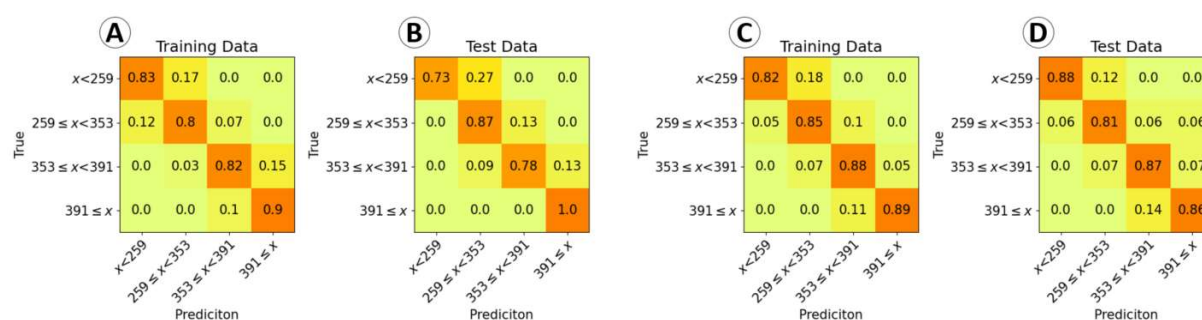


Figure 4-4: Scores in a confusion matrix for training (A) and test sets (B). The new scores of the new model after SA for training (C) and test (D).

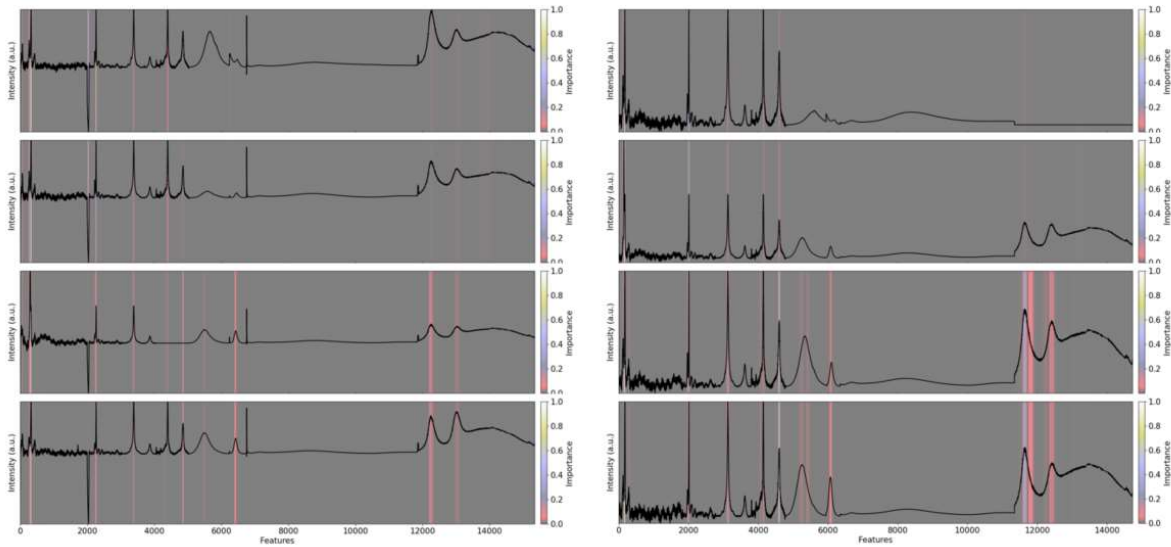


Figure 4-5: Comparison between the first PC-LDA (left) and the second training (right) for each of the classes (from class 1 to 4 from top to bottom) in terms of importance according to sensitivity analysis. This shows how the vector changes in length after cutting off some of the sections. The importance is the average of the 10 closest spectra to the center of each of the clusters. Removing these sections appears to enhance some of the more important features.

With the newly derived dimensions D1 and D2, a MVNLR can be performed in the shape of $f(D1, D2) = y$, with y representing the V_{OC} . A 2nd degree polynomial is selected due to the enhanced results compared to degree 1 and superior interpretability and visual representation relative to degree 3 and beyond, where results see only marginal improvements. Nonetheless, D1 and D2 represent an already uninterpretable combination of thousands of higher dimensions, making it challenging, if not impossible, to derive insights beyond the fact that V_{OC} can be predicted from this data. The regression, represented by Eq. 4-1, exhibits good correlation with R2 scores of 0.85 as shown in Figure 4-9; **Error! No se encuentra el origen de la referencia.** (center). The visual mapping of the resulting equation in an expanded solution space, seen in Figure 4-9 (right), highlights continuity and also mathematically suggests a higher-performing cluster with elevated D1 and D2 values of the ranges 6-8 and 2-5, respectively.

$$y = 328 + 29.3 \cdot D_1 - 19.4 \cdot D_2 - 2.89 \cdot D_1^2 + 6.95 \cdot A_5 \cdot A_{14} - 2.45 \cdot A_{14}^2 \quad \text{Eq. 4-1}$$

$$y = 489 - 22.5 \cdot A_5 + 1.17 \cdot A_{14} \quad \text{Eq. 4-2}$$

$$y = 520 + 20.0 \cdot A_5 - 2.77 \cdot A_{14} - 3.44 \cdot A_5^2 - 0.40 \cdot A_5 \cdot A_{14} - 0.007 \cdot A_{14}^2 \quad \text{Eq. 4-3}$$

The utilization of CNNs takes a comparable methodology applied to discern crucial elements within the data. The performance of this model is favorable, displaying scores that are comparable with the PC-LDA technique, as shown in Figure 4-6. With the employment of CNN, however, we gain access to additional explainability tools such as Grad-CAM, along with more in-depth

potential analysis of the algorithm via the intersection of activation patterns in units and layers. In light of this, the average activation map of the ten nearest spectra, as previously deduced with the PC-LDA algorithm, for the top-performing cells ($391 \leq V_{OC}$) for each convolutional layer is portrayed in Figure 4-7. Figure 4-8 compiles the results for the four classification groups and the three convolutional layers. A distinctive common characteristic, consistent across all classes and layers, is the total lack of focus directed towards the final segment of the vector, belonging to the 325 nm PL measurement. Additionally, for the second layer, dispersed attention is primarily allocated to Raman 532, 442, 325 and PL 785, 532, and 442, while the final convolution layer assigns markedly more importance to the Raman 532 nm spectra, and notably to PL 442 for the highest performance cells, as well as for the second-best classification. However, for the bottom groups, attention pivots towards PL 442 in the worst performance cells and towards the 785, 532, and 442 Raman spectra for the second-worst. This indicates, in line with PC-LDA, that the different classification groups stimulate the algorithm in quite dissimilar ways. Nonetheless, in general terms, it appears that local patterns are scarce, while macro patterns exert a greater influence on the final decision of the CNN.

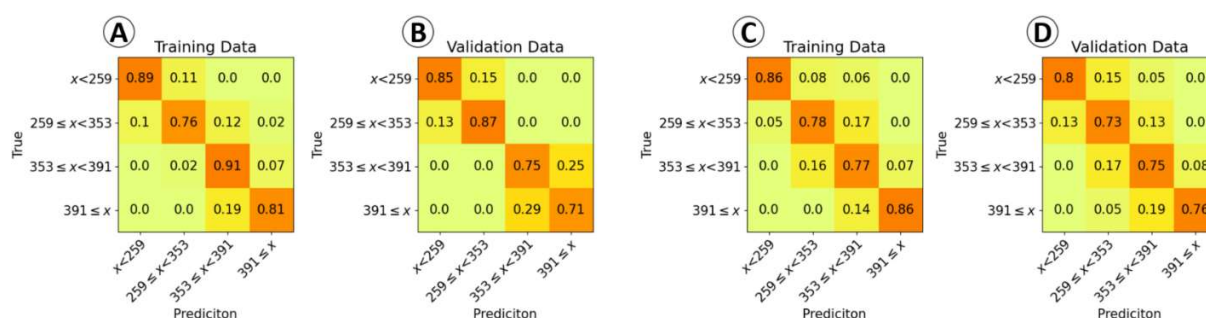


Figure 4-6: Confusion matrices for Training (A) and Test (B) data sets for the CNN, with averages of 84% and 80%, respectively. Despite the good scores, some overfitting is appreciated, but highly biased by the best performing class, where accuracy is just above 71% with about 29% misclassified as second to first. For the 2D CNN, slightly lower scores are shown, with 0.82 and 0.76 for training and validation.

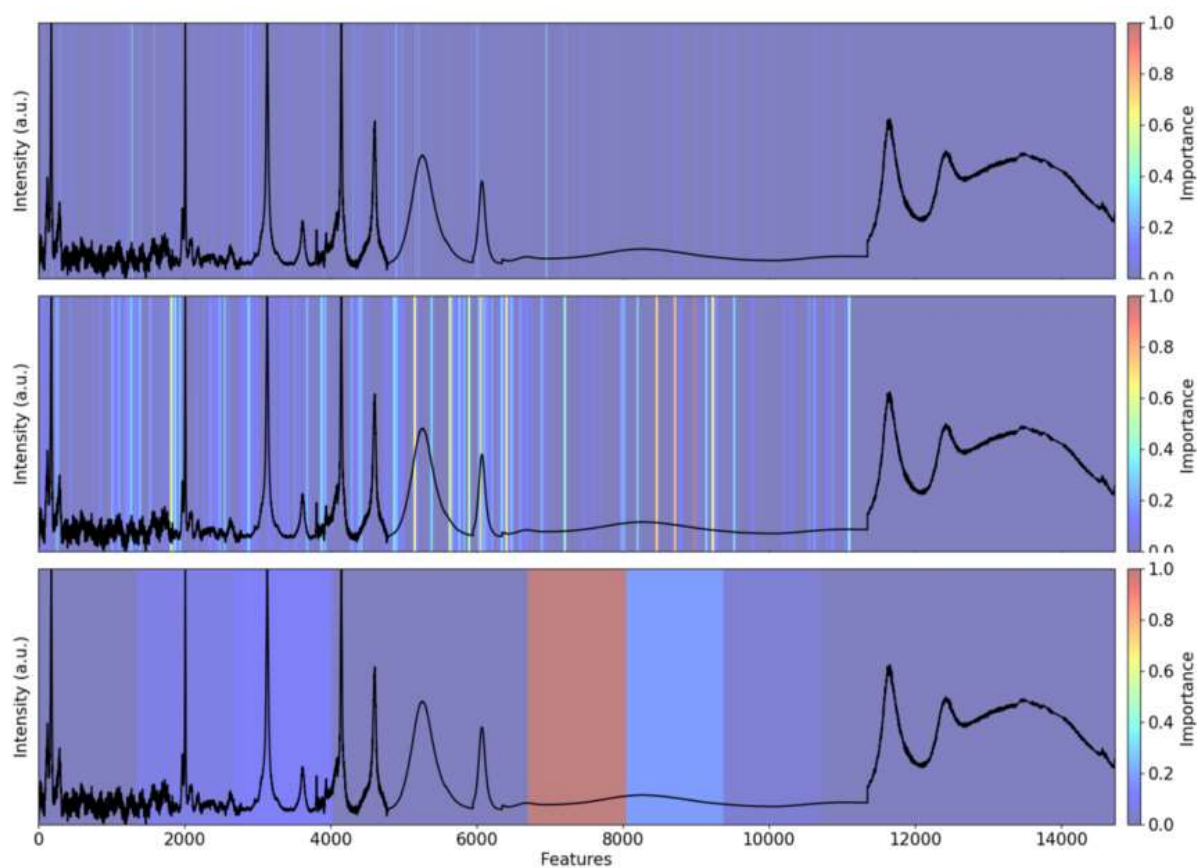


Figure 4-7: Average Grad-CAM visualization of the closest 10 spectra to the center of the cluster from the top performing classification group of $391 < V_{OC}$. From top to bottom, is the first convolutional layer, the second, and third convolutional layer from the CNN. Importance is normalized to 1 for each case.

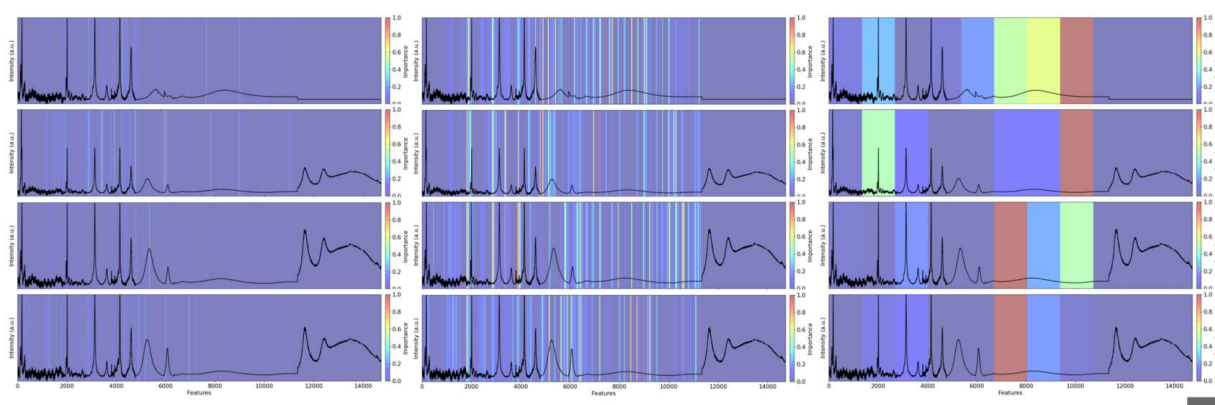


Figure 4-8: GradCAM results for the 3 convolutional layers (left to right) and the 4 classification groups (top to bottom)

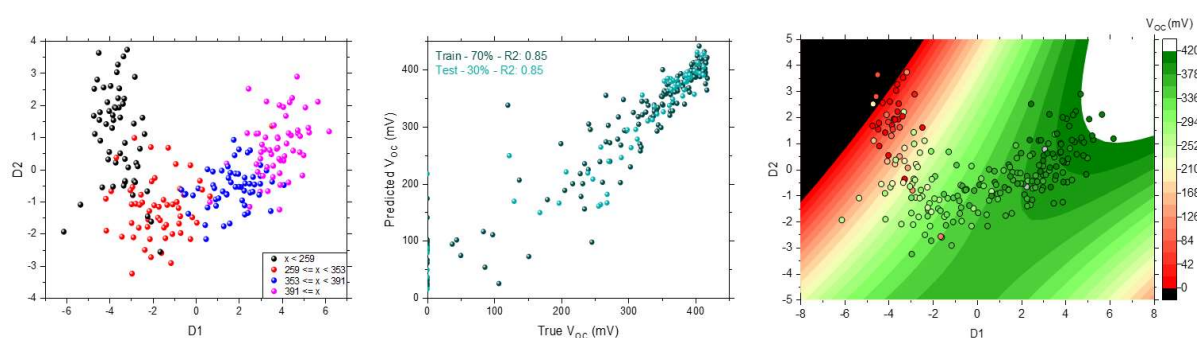


Figure 4-9: D1 v D2 plot of the final PC-LDA model (left), MVNLR as $f(D1, D2)$ against V_{OC} (center) and the obtained equation mapped along with the scatter plot of D1 v D2 color graded with the V_{OC} (right).

For the 2D CNN, the GradCAM mappings are assembled in Figure 4-10. These mappings, originally presented in the form of images, have been flattened into 1D vector format for ease of visualization. In this instance, a slightly different behavior is observed compared to the first case, as anticipated due to the inherent differences between the algorithms. Specifically, attention appears to be directed to the PL 325 vector, especially for the second group ($259 \leq V_{OC} < 353$). Aside from this peculiarity, the remaining activations bear resemblance to those from the 1-D case, with the exception of the fourth group in the first layer, which yields no activations at all for GradCAM. This can be explained in a two-fold manner: firstly, the classification of this group primarily relies on more “general” patterns detected by the final convolutional layer. Secondly, this layer appears to solely account for negative impacts in the classification, given its use of LeakyReLU activations instead of ReLU, in contrast to the 1D CNN.

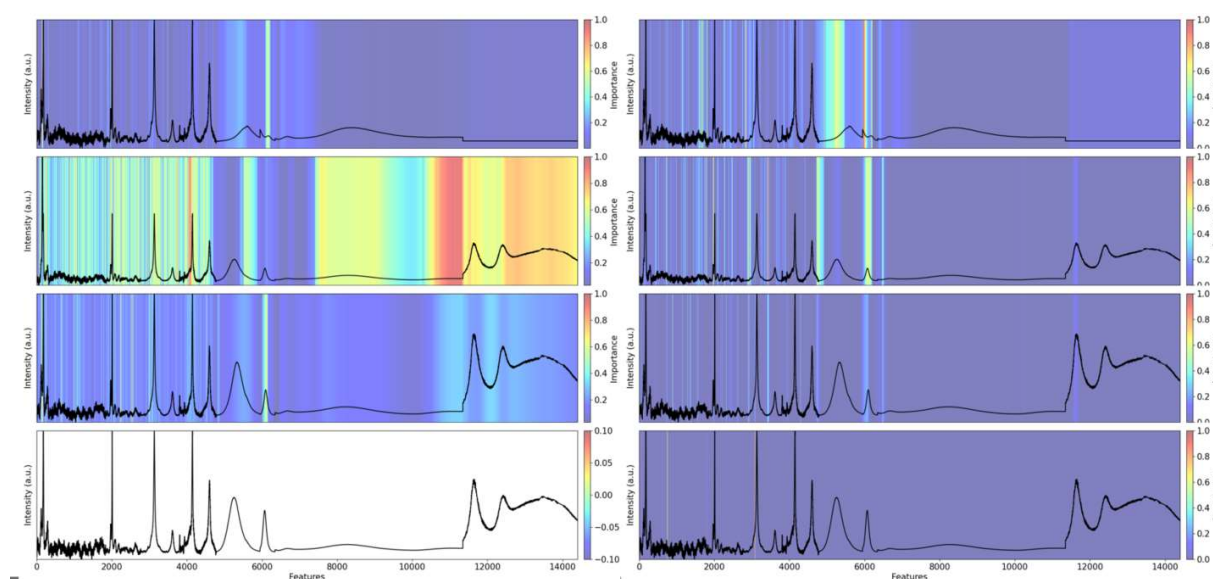


Figure 4-10: GradCAM heatmaps for the average of the 10 closest spectras to the center of each of the clusters according to PC-LDA for each of the classification groups (top to bottom) for both convolutional layers (right and left).

Following the analysis conducted, we were able to more accurately pinpoint potentially significant areas within the spectra. Due to this process, several considerations were taken into account: 1) the full area of the peaks did not necessarily need to be utilized, 2) varying impacts were observed in different sections of the peaks, potentially suggesting the presence of smaller peaks, 3) areas of utmost significance were commonly found on or near the peaks, and 4) despite their theoretical lack of interest, some sections still displayed activations and significance. However, caution should be exercised due to the fact that the models, despite their high accuracy, did not reach full generalization, potentially leading to the erroneous use of these sections. Moreover, the preceding analysis should be interpreted with caution, and prioritizing good criteria for area selection is paramount. Hence, the sections deemed unimportant but are highlighted by the methodologies, which are more likely to complicate subsequent analysis and divert attention from impactful areas, were omitted. Following these guidelines, 21 areas were selected, as shown in Table 4-2, and the correlation between them is examined with both Pearson and Spearman coefficients. Separately, the correlation with V_{OC} was examined using a quadratic regression independently for each area in the form of $f(A_n) = C_0 + C_1 \cdot A_n + C_2 \cdot A_n^2 = y$, with the results represented as R2 scores in Figure 4-11. Areas A14, A16, and A3 stood out as the most predictive, while areas A8, A2, and A1 were the least. Given the high correlation between these areas, they are presumed to contain the same, or similar, information, limiting their combined predictive power due to data redundancy. Therefore, the next highest scoring, yet unrelated, area to A14 was determined to be A5 (Figure 4-11). With these two areas, interpretable correlations can be further investigated. As they can be visualized in 2D due to their limited variable count, the results are more accessible for human interpretation. Quadratic polynomial regression (PR) and radial basis function network

(RBFN) were applied, with the outcomes depicted alongside Multivariate Linear Regression in Figure 4-12. Good linearity was observed for voltages of 250 mV and above, although correlation was poor for lower voltages, mirroring the findings in [125]. However, this issue was resolved with non-linear regression, yielding improved agreement with lower voltages and an overall R2 score of 0.76. This model, shown in Eq. 4-3, presents a digestible equation comprising only six terms. The equation allows for the existence of high-performing cells not accomplished in the investigated sample, both with elevated and decreased A5 and A14 areas, although the latter possibility lies beyond the established parameter range but is theoretically plausible due to symmetry. The feasibility of achieving these values is subject to further investigation and presents an intriguing research question.

Table 4-2: Selected areas and their respective pixel ranges from the 14,640-long vector. In addition, the areas are labeled with the corresponding measurement technique and the axis values in the respective units of that measurement (Shift for Raman and Wavenumber for PL).

Area	Measurement	Pixel range	Shift range (cm ⁻¹)
A1	Raman 785	90 – 145	159 – 180
A2	Raman 785	145 – 200	180 – 207
A3	Raman 785	250 – 320	231 – 262
A4	Raman 532	1950-1985	262 – 343
A5	Raman 532	1985 – 2030	529 – 653
A6	Raman 442	3020 – 3070	272 – 332
A7	Raman 442	3070 – 3200	417 – 516
A8	Raman 442	3500 – 3700	516 – 665
A9	Raman 325	3915 – 3960	630 – 913
A10	Raman 325	4025 – 4100	346 – 430
A11	Raman 325	4100 – 4215	464 – 537
A12	Raman 325	4400 – 4510	901 – 1039
A13	Raman 325	4510 – 4680	1039 – 1250
Area	Measurement	Pixel range	Wavelength (nm)
A14	PL 785	5060 – 5560	1126 – 1365
A15	PL 785	5560 – 5775	1365 – 1468
A16	PL 532	6000 – 6150	1077 – 1326
A17	PL 442	6440 – 7000	462 – 546
A18	PL 442	7000 – 10000	546 – 994
A19	PL 325	11470 – 12090	358 – 451
A20	PL 325	12090 – 12660	451 – 537
A21	PL 325	12660 – 14510	537 – 816

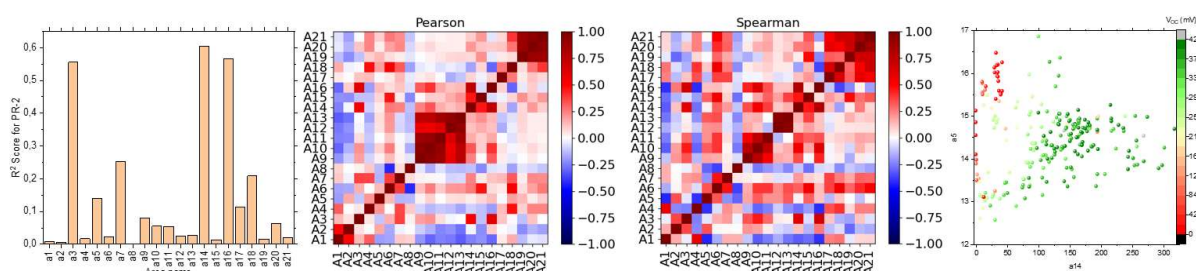


Figure 4-11: Individual R² for each of the areas when performing regression against V_{OC} (left), Person (center left) and Spearman (center right) correlation matrices for all the selected areas, and area 4 (a4) versus area 10 (a10) scatter plot graded with V_{OC} (right).

RBFN, unlike the initial two, is incapable of deriving a specific equation due to its nature as an NN. As such, the fundamental processes through which it predicts are inherently opaque to human understanding, even when expandability techniques are employed, as discussed in section 11.4.3. This opaqueness is to be expected, as a trade-off typically exists between interpretability and accuracy. Nevertheless, the best scores are achieved by RBFN, at 0.83. As a result, a notably more complex surface is revealed, with two peaks evident for the optimal voltages (around 200 and 350 of A14). Despite the differences in these mappings, certain commonalities are identifiable. Firstly, consensus is reached that the least effective cells exhibit equally large A5 and small A14 areas. Secondly, it appears that the majority of the importance is held by A14, meaning that the performance is more dependent on this area for the measured space. However, all agree that it might be mathematically possible for A5 to exert more influence, though this does not hold true for the first case. Thirdly, they also seem to concur that increasing performance becomes more challenging as the quality of the sample improves. This can be observed in the enlarging step size, which grows with V_{OC} . In spectral terms, this implies that it is easier to enhance a poor performing cell than it is to improve a cell that already performs well.

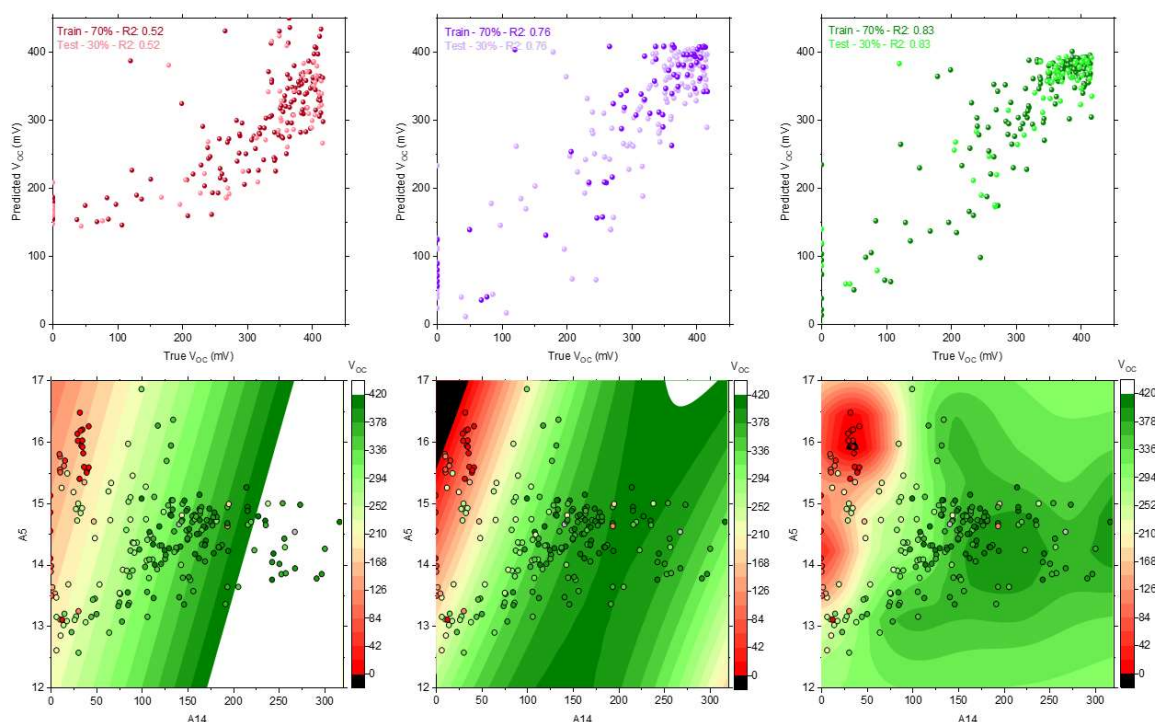


Figure 4-12: Regression (top row) and prediction mapping (bottom) for multi linear (left), polynomial quadratic (center) and RBFN (right).

4.3.2 Sensitivity analysis of activations and model improvement

It is observed that, overall, the highest absolute values of importance are focused on the crucial peaks of the spectroscopic measurements, however with much more attention to PL data than to Raman. For instance, the main PL curve for 442 nm appears to have the most absolute importance across all classes (Figure 4-13A and C). When analyzing the evolution of classification to the next best-performing class (Figure 4-13B and D), importance is attributed also to sections of PL 442 nm with more protagonism of PL 785 nm and 532 nm compared to the inner-class case. All Raman spectra though show some activity, but it is small compared to these PL numbers. The latter is consistent with the nature of this methodologies, since PL is normally associated with the band gap, meanwhile Raman with more related to structural and defect properties, which is indirectly related to the band gap. Overall, most importance is attributed to features in the ranges of 5000-6000, 6000-6500, 8000-11000, and 11000-13000.

For the re-activations, there are specific units related to specific features, as expected to such 1D CNN. In particular, more defined relationships exist between activation-feature pairs 464-76, 288-35, 480-77, 464-77, and 510-76. This means that PL with 442 nm, belonging to feature 76 and 77, is closely related with activations in units 464, 480, and 510, meanwhile PL 785 nm, belonging to feature 35, is tethered with unit 288. Furthermore, when diving by classification, this last

relationship is dominated by the first classification group of $V_{OC} < 303$ mV, and the other pairs are more related with the top 2 classes. Overall, for correctly classified vectors, most activations are related to features in window range between 34-43 and 75-79.

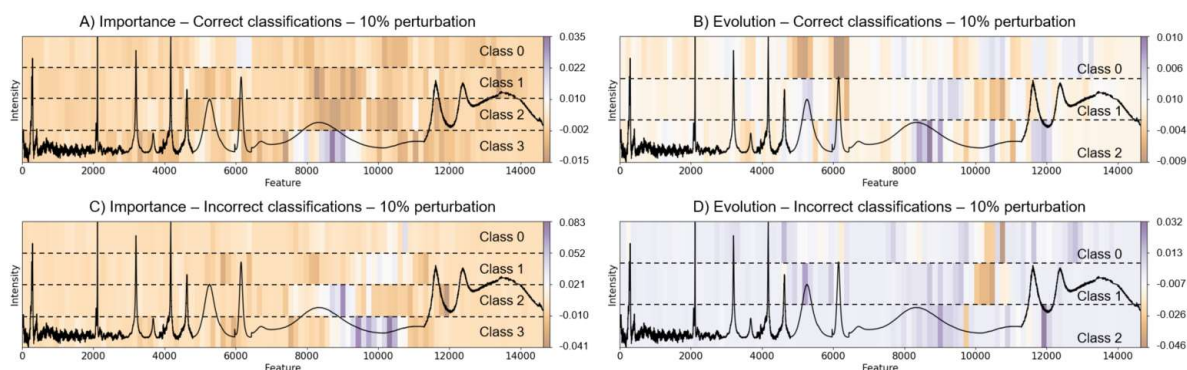


Figure 4-13: Importance for spectroscopic features according to the change in inner-class probability change and next best-performing probability change.

By identifying the critical units' re-activations that perform good and bad classifications, it is possible to individually check how these activations affect the overall performance of the model. These re-activations are shown in Figure 4-14 by class and overall. In general, it is clear to see that sets of units focus on different classes, presumably according to where each class contains its most characteristic features. For instance, overall correct prediction seems to be driven mostly by units 464, 288, 480, 44, and 510. In more detail, for correct predictions of Class 1 shows more activations in units 288 and 304, meanwhile Class 2 activate more units 510 and 542. In contrast Classes 3 and 4 show similar units being activated, namely 463 and 480, with difference in the bottom part of their respective lists. When analyzing incorrect classification, several units appear as having influence in this miss-classifications. Overall, units 558, 128, and 46 seem to have the most influence in these errors. With this information, and considering the class-specific re-activations, we can try deactivating by setting to 0, each of these activations and see how they affect the final prediction. After try and error, it is detected that by deactivating units 128, 208, and 522, an improvement in the overall scores, going from 0.87 to 0.88 is achieved, as shown in Figure 4-15 . Even though the difference of only 1 percentage point, it does represent the correct classification of around 17% of the incorrect classifications, or 4 out of 23, as shown in Table 4-3. Furthermore, five other samples show statistical improvement out of the probability function, though not enough change to correct their classification, which translates to that 39%, or 9 out of 23, perceived benefits on the changes to the activations. As consequence, however, there was 1 sample that switched from correctly classified to incorrectly classified.

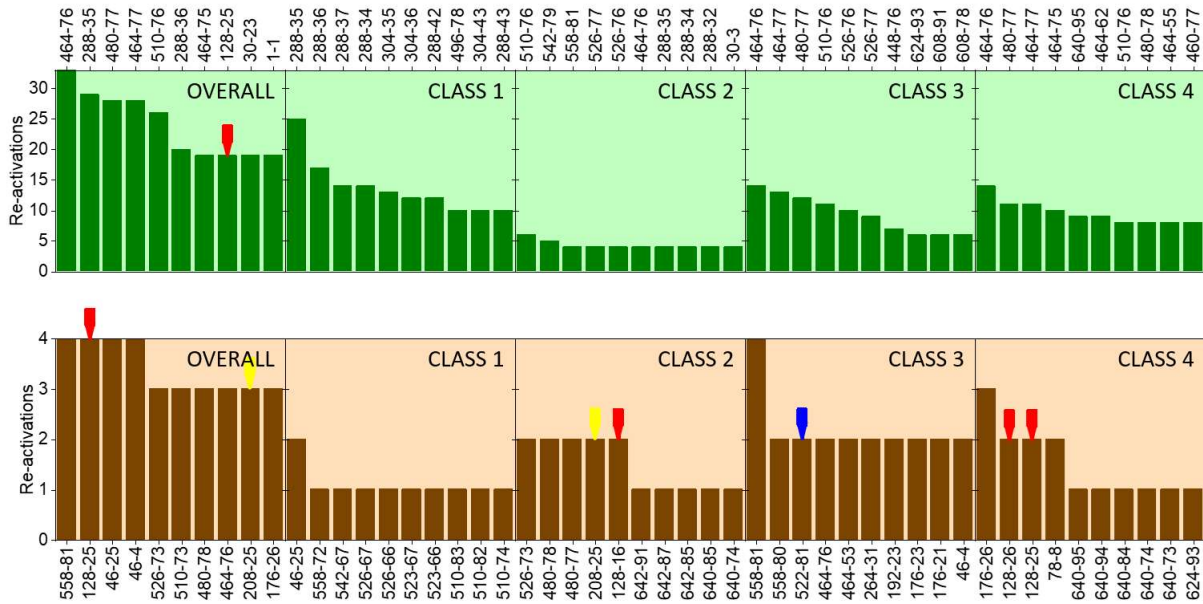


Figure 4-14: Reactivation values for units in the last convolutional layer for correct classifications (Top in green) and incorrect classifications (bottom in brown). Arrows indicate the units deactivated in the new model and their color indicate the same unit as in red for unit 128, yellow for unit 208, and blue for unit 522.

Table 4-3: Changes in class classification probability p for each of the incorrectly classified samples (s) after modification of activation values of activations 128, 208 and 522. In bold are the samples that corrected their classification after modification, four in total (samples 12, 56, 58, and 62). Five other samples show statistical benefit but not enough to correct their prediction (samples 26, 140, 45, 122, and 107).

	$V_{oc} < 303$				$303 \leq V_{ov} < 363$						$363 \leq V_{ov} < 396$						$396 \leq V_{ov}$						
s	78	144	38	105	12	26	140	89	127	146	173	56	45	122	107	109	113	134	58	63	99	124	138
Δp_0	-0.02	0	0	0	-0.12	-0.01	0	0	0	0	0	0	0	0	0	0	0	0	0.01	-0.01	0	0	0
Δp_1	0.02	0	0	0	0.16	0.02	0.01	0	0	0	0	-0.08	-0.05	-0.03	-0.02	0	0	0	-0.09	0.03	0	0	0
Δp_2	0	0	0	0	-0.06	-0.02	-0.01	0	0	0	0	0.08	0.05	0.03	0.01	0	0	0	-0.06	-0.06	0	0	0
Δp_3	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0.14	0.04	0	0	0

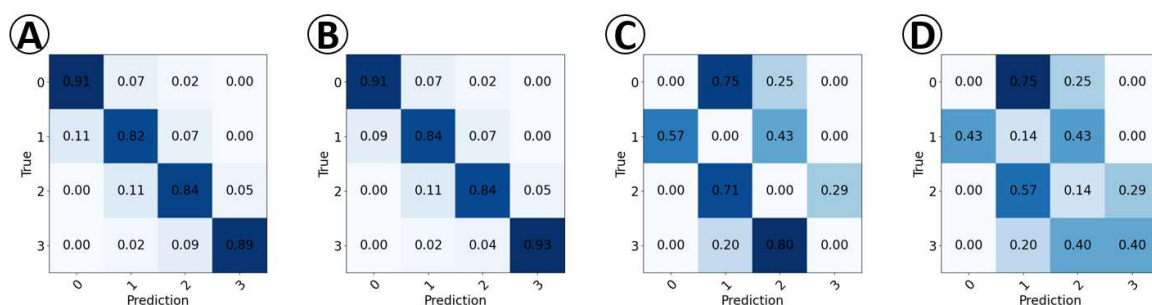


Figure 4-15: Overall confusion matrices showing the scores of A) the original CNN, B) the modified CNN after analysis of activations, C) the incorrect classifications of the original CNN and D) the incorrect classifications of the improved CNN.

4.4 Conclusions of exploratory experiments

In this study, various ML and XAI techniques are leveraged to conduct a comprehensive analysis of spectroscopic measurements obtained from a combinatorial sample and to better understand the mechanisms of CNNs using this data. The sample is synthesized with varying ratios of Copper, Tin, and Zinc elements, and segmented into a 15x15 grid, for a total of 225 cells. These cells undergo thorough characterization via optoelectronic and compositional techniques, supplemented by Raman and PL spectroscopy measured under wavelengths of 785, 532, 442, and 325 nm. In a first stage, the study commences with the application of PC-LDA, which facilitates the refinement of the vector feature by omitting irrelevant sections of the spectra. The algorithm demonstrates good classification capabilities, yielding an average accuracy of 0.86 for both training and validation sets. Subsequently, a 1D CNN is configured for the same classification task, producing comparable results with average accuracy scores of 0.84 for both the training and testing sets. A 2D CNN is then designed for the same classification problem but with a reconfigured version of the features transformed into 120x120 images as opposed to a 14,640 vector, resulting in accuracy scores of 0.81 and 0.75 for the training and test sets, respectively. Application of explainability methodologies elucidates how specific sections, along with local and macro patterns, influence the decisions of the algorithms. With this knowledge, specific areas are chosen to perform Pearson and Spearman correlation matrices. At the same time, quadratic regression is done with each of the selected areas to predict V_{OC} . The obtained R^2 scores are cross-referenced with the correlation matrices to identify the optimal pair of areas (A5 and A14) that exhibit high individual predictability but low intercorrelation. Different regression techniques are then implemented to predict V_{OC} using these areas. Radial Basis Function Networks (RBFN) yield the best result with an R^2 coefficient of 0.86, whereas MVNLR obtains 0.76. Despite the superior performance of RBFN, MVNLR is preferred due to the readability of the equation, facilitating a more thorough comprehension of the outcomes. These outcomes suggest the potential for higher V_{OC} with increasing A5 and A14 areas under the MVNLR model. However, the feasibility of this prediction remains questionable. In contrast, the more accurate RBFN model presents a clear V_{OC} limitation based on the provided solution space.

On a second stage of the exploratory experiments, a shift of focus is turned into the detailed analysis and explanation of a third CNN for better understanding the failure of classification of these TFPV cells according to their performance. This is achieved by measuring the change in unit activations (re-activations) in the last convolutional layer of the achieved model and associating them with specific features and correct or incorrect classifications. It is shown that it is possible to better explore the structure of CNN in the context of advanced characterization of PV materials and devices, achieving better understanding on their reasoning, innerworkings, and results. This is achieved by being able to associate activations and specific features, protecting clues on how exactly the model is making decisions in each instance. Furthermore, by better understanding the latter, it is possible to modify the CNN to improve its performance, achieving a successful improvement by correctly classifying 4 additional cells after the modification of the CNN.

Overall, these findings highlight the potential impact of future research, emphasizing the significance of improving our understanding of ML and AI models in the field of materials research. The advancement of this kind of research may mean not only the gain of deeper insights into TFPV materials but also to the better comprehension of ML models. The latter is paramount to further leverage AI in research and may lead to create computational experiments that may replace physical in-lab experiment, further shortening material discovery and improvement beyond the current capabilities of AI.

5. CONCLUSIONS AND OUTLOOK

This thesis aimed to leverage cutting-edge analysis techniques based on CA and AI to investigate the physicochemical and optoelectronic attributes of chalcogenide-based TFPV materials and other emerging technologies. The primary objective was to develop innovative CA approaches based on AI and ML to accelerate research and development of TFPV materials, including but not limited to chalcopyrite and kesterite compounds, and reduce their lab-to-market times. This main goal was subject to three objectives, namely the design and implementation of automated, high-throughput, multi-technique characterization systems, the creation of ML methodologies for CA data processing, and create approachable and accessible tools to implement all the above in a seamless and straightforward way.

For this, the work began with the development and implementation of an automated spectroscopic platform, facilitating automated measurements and preliminary analysis of multitechnique spectroscopy, including RS, PL, and NF. A subsequent study demonstrated the effectiveness of PC-LDA algorithm in assessing the thickness of AlO_x barrier layers in flexible PV combinatorial samples deposited on top of industrially relevant substrates. The model provided accurate and reliable thickness measurements using NF spectroscopy data, proving to be non-destructive, fast, and cost-effective. This study revealed the CA and ML combination potential in quality control and process optimization in the PV industry.

In a second stage, an evaluation of the AI strategy led to a proposed methodology workflow based on dimension reduction algorithms, mainly PCA, LDA and PC-LDA, that simplifies the preprocessing of spectroscopic data and offers deeper insights into relevant material and PV device processes. To streamline this methodology, the Python library "spectrapeper" was developed, covering procedures from data acquisition to analysis. A subsequent study examined the influence of off-stoichiometry on defect formation and solar cell performance in CZGSe TFs using the methodology based on CA and LDA. The analysis revealed that variations in Ge content significantly impacted defect formation and device performance, and allowed to define the optimal composition ranges of the CZGSe compound to produce the high efficiency solar cells.

At a final stage of the doctorate program, the need to further explore the derived results from the methodology, in particular the ML results from the dimension reduction, raised as a natural consequence of the performed work. This was also sustained by an abundance of evidence in the literature and also by the lack of tools for explainability in spectroscopic data. Thus, to facilitate deep data-driven analysis of such outcomes the "pudu" library was created for sensitivity analysis, helping identify crucial spectra parts for algorithmic classification.

With the above it is possible to affirm that the objectives established for this thesis are fulfilled as follows:

- Objective 1: The first objective, focusing on the development of autonomous systems for high-throughput data collection using various spectroscopic and optoelectronic techniques, is effectively met by the capabilities of the developed measuring system and the spectrapepper library. The automated system was developed in IREC for HTE of spectroscopic characterization. The first article uses an early version for automated single measurement, meanwhile the second article uses a second version where multiple techniques can be performed quasi-simultaneously in the same spot. On the other hand, spectrapepper facilitates the efficient collection and processing of large-scale data, and it is partially used in the automated system. The ability to rapidly acquire and process a vast amount of data from different instruments ensures a comprehensive dataset, essential for in-depth analysis and understanding of the physicochemical and optoelectronic properties of TFPV materials.
- Objective 2: The second objective involves the development of AI algorithms for efficient spectroscopic data processing, a task crucial for handling the extensive data generated in future TFPV research. This objective is tackled two-fold. First, the use of dimension reduction algorithms in the proposed methodology, namely PCA, LDA, and PC-LDA, allows for a simplification of spectral processing since no specific features need to be extracted. Second, the spectrapepper library addresses this need by offering automated and generalized tools for the processing of large datasets of spectroscopic data. This automation not only streamlines the workflow but also reduces the requirement for specialized expertise, making the process more accessible and efficient.
- Objective 3: Fulfilling the third objective entails the creation of tools that are easy to use and implement in scientific and industrial settings. This is achieved with the two libraries spectrapepper and pudu, currently accessible as open-access and open-source. These libraries contain a broad spectrum of tools that researchers can easily implement in their data processing and analysis. Most of these tools are aimed to be single-line commands with clear names, purposes, and parameters so they can fit any kind of demand or needs and can be implemented in custom systems due to their open-source nature.

In summary, the integration of these solutions directly supports the main goal. The comprehensive data collection and processing capabilities of the automated system coupled with the spectrapepper functions, the clear and user-friendly ML framework, and the availability of deeper insights for spectroscopic problems with pudu allow for a fast experimental analysis cycle and may help to better understand TFPV materials and devices. This is pivotal for advancing the field of TFPV technologies, enabling more efficient, reliable, and optimized TFPV materials and devices.

Finally, 2 follow-up experiments that employ the ML framework, both libraries, and other XAI techniques to analyze spectroscopic measurements from a combinatorial CZTSe sample are

presented. These are two approaches that naturally follow the proposed methodology used in this thesis, extending it towards better interpretability and deeper insights into the data. The first, performs an analysis using PC-LDA with promising classification results. Further analysis using 1D and 2D CNN also achieves high accuracy scores. The application of XAI reveals how specific sections and patterns influence algorithmic decisions. Specific areas are chosen to perform correlation matrices and regression, predicting hypothetical compositions for enhanced V_{OC} using areas with high individual predictability and low intercorrelation. Predicting V_{OC} using these areas with different regression techniques highlights RBFN as superior. However, MVNR is favored due to its interpretability. The study emphasizes the importance of understanding ML and AI models in materials research, promising to positively impact future research. The second uses a more advanced technique, namely re-activation and dissection analysis, to better explore how incorrect predictions are formed and how to use this knowledge to improve the performance of CNN models. The approach successfully analyzes how activations relate to specific features in the data, which allows to modify the CNN to improve its classification performance. Both of these approaches are then presented as a natural extension of the methodology used in the published articles and contains great potential for further development and research.

In short, the culmination of this thesis includes the development of AI procedures for the analysis of advanced characterization data of PV materials and devices, and the development of open-source software designed for researchers with little coding experience. These results have positioned the IREC-SEMS research group at the forefront of next-generation PV technologies by providing a robust and versatile CA and AI methodology and tools that has benefited various projects, including Solar-Win (H2020, GN 870004), In4CIS (Proyectos de I+D+I Programación Conjunta Internacional 2019, GN PCI2019-111837-2), SUNRISE (H2020, GN 958243), and Platform-ZERO (H2020, GN 101058459). Moreover, the results of the thesis allowed to strongly consolidate two research lines of the SEMS group: “Advanced characterization of the PV materials and devices” and “Development, methodologies and prototyping of sensors for photovoltaics and process monitoring”. This was possible by providing new tools and possibilities for the advanced analysis of spectroscopic data for TFPV.

After the development of this work, and after all these years, it is clear to me, and hopefully to the reader as well, that we stand at the forefront of a new era in scientific discovery and technological advancement. AI continues to reshape the landscape of research and the content of our daily lives, transcending every boundary and seeping into virtually every field and application. These technologies, just as have done with me, will inspire and empower researchers to dig deeper into the unknown, find solutions to some of the most pressing challenges faced by humanity and, hopefully, reveal the mysteries of our universe. As our knowledge grows, so too will our ability to harness AI to improve the quality of life for people around the world. However, it is essential to maintain a sense of humility and responsibility as we unleash this power. We must not forget the human element at the heart of our endeavors, as our progress will be judged not only by the

sophistication of our algorithms but also by our commitment to collaboration, empathy, and ethical considerations. Let us embrace the possibilities that AI offers, fostering a future where technology and human ingenuity work hand in hand to create a more sustainable, equitable, and compassionate world for all. I firmly hold the conviction that, by preserving our curiosity and strengthening our human values, there will be no limit to what we can achieve, no questions we cannot answer, and no algorithm we cannot explain.

Gracias totales.

REFERENCES

- [1] IPCC, “Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change,” Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2022. doi: doi:10.1017/9781009325844.
- [2] L. Berrang-Ford, J. D. Ford, and J. Paterson, “Are we adapting to climate change?,” *Glob. Environ. Chang.*, vol. 21, no. 1, pp. 25–33, Feb. 2011, doi: 10.1016/J.GLOENVCHA.2010.09.012.
- [3] R. G. Taylor *et al.*, “Ground water and climate change,” *Nat. Clim. Chang. 2012 34*, vol. 3, no. 4, pp. 322–329, Nov. 2012, doi: 10.1038/nclimate1744.
- [4] R. Garrido *et al.*, “Potential impact of climate change on the geographical distribution of two wild vectors of Chagas disease in Chile: *Mepraia spinolai* and *Mepraia gajardoi*,” *Parasites and Vectors*, vol. 12, no. 1, pp. 1–16, Oct. 2019, doi: 10.1186/S13071-019-3744-9/TABLES/2.
- [5] G. S. Malhi, M. Kaur, and P. Kaushik, “Impact of Climate Change on Agriculture and Its Mitigation Strategies: A Review,” *Sustain. 2021, Vol. 13, Page 1318*, vol. 13, no. 3, p. 1318, Jan. 2021, doi: 10.3390/SU13031318.
- [6] V. Karimi, E. Karami, and M. Keshavarz, “Climate change and agriculture: Impacts and adaptive responses in Iran,” *J. Integr. Agric.*, vol. 17, no. 1, pp. 1–15, Jan. 2018, doi: 10.1016/S2095-3119(17)61794-5.
- [7] N. K. Arora, “Impact of climate change on agriculture production and its sustainable solutions,” *Environ. Sustain. 2019 22*, vol. 2, no. 2, pp. 95–96, Jun. 2019, doi: 10.1007/S42398-019-00078-W.
- [8] IPCC, “Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change,” Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896.
- [9] R. S. J. Tol, “The Economic Impacts of Climate Change,” <https://doi.org/10.1093/reep/rex027>, vol. 12, no. 1, pp. 4–25, Feb. 2018, doi: 10.1093/REEP/REX027.
- [10] S. C. Pryor and R. J. Barthelmie, “Climate change impacts on wind energy: A review,” *Renew. Sustain. Energy Rev.*, vol. 14, no. 1, pp. 430–437, Jan. 2010, doi: 10.1016/J.RSER.2009.07.028.
- [11] D. Soto, J. León-Muñoz, J. Dresdner, C. Luengo, F. J. Tapia, and R. Garreaud, “Salmon farming vulnerability to climate change in southern Chile: understanding the biophysical, socioeconomic and governance links,” *Rev. Aquac.*, vol. 11, no. 2, pp. 354–374, May 2019, doi: 10.1111/RAQ.12336.
- [12] IPCC, “Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Field, C.B., V.R. Barros, D.J. Dokken, K.J.,” Cambridge, United Kingdom and New York, NY, USA, 2014.
- [13] A. Urquiza and M. Billi, “Water markets and social–ecological resilience to water stress in the context of climate change: an analysis of the Limarí Basin, Chile,” *Environ. Dev. Sustain.*, vol. 22, no. 3, pp. 1929–1951, Mar. 2020, doi: 10.1007/S10668-018-0271-3/METRICS.

- [14] M. Billi, G. Blanco, and A. Urquiza, “What is the ‘Social’ in Climate Change Research? A Case Study on Scientific Representations from Chile,” *Minerva*, vol. 57, no. 3, pp. 293–315, Sep. 2019, doi: 10.1007/S11024-019-09369-2/TABLES/1.
- [15] D. Araya-Osses, A. Casanueva, C. Román-Figueroa, J. M. Uribe, and M. Paneque, “Climate change projections of temperature and precipitation in Chile based on statistical downscaling,” *Clim. Dyn.*, vol. 54, no. 9–10, pp. 4309–4330, May 2020, doi: 10.1007/S00382-020-05231-4/FIGURES/13.
- [16] F. J. Fernández, M. Blanco, R. D. Ponce, F. Vásquez-Lavín, and L. Roco, “Implications of climate change for semi-arid dualistic agriculture: a case study in Central Chile,” *Reg. Environ. Chang.*, vol. 19, no. 1, pp. 89–100, Jan. 2019, doi: 10.1007/S10113-018-1380-0/METRICS.
- [17] S. Nasirov, A. Girard, C. Peña, F. Salazar, and F. Simon, “Expansion of renewable energy in Chile: Analysis of the effects on employment,” *Energy*, vol. 226, p. 120410, Jul. 2021, doi: 10.1016/J.ENERGY.2021.120410.
- [18] W. Nordhaus, “Climate Change: The Ultimate Challenge for Economics,” *Am. Econ. Rev.*, vol. 109, no. 6, pp. 1991–2014, 2019, doi: 10.1257/AER.109.6.1991.
- [19] International Energy Agency, “Energy Technology Perspectives 2023,” Paris, 2023. doi: 10.1787/9789264109834-en.
- [20] E. De Cian and I. Sue Wing, “Global Energy Consumption in a Warming Climate,” *Environ. Resour. Econ.*, vol. 72, no. 2, pp. 365–410, Feb. 2019, doi: 10.1007/S10640-017-0198-4/TABLES/14.
- [21] R. Schaeffer *et al.*, “Energy sector vulnerability to climate change: A review,” *Energy*, vol. 38, no. 1, pp. 1–12, Feb. 2012, doi: 10.1016/J.ENERGY.2011.11.056.
- [22] Y. Simsek, Á. Lorca, T. Urmee, P. A. Bahri, and R. Escobar, “Review and assessment of energy policy developments in Chile,” *Energy Policy*, vol. 127, pp. 87–101, Apr. 2019, doi: 10.1016/J.ENPOL.2018.11.058.
- [23] M. B. Hayat, D. Ali, K. C. Monyake, L. Alagha, and N. Ahmed, “Solar energy—A look into power generation, challenges, and a solar-powered future,” *Int. J. Energy Res.*, vol. 43, no. 3, pp. 1049–1067, Mar. 2019, doi: 10.1002/ER.4252.
- [24] I. Energy Agency, “Energy Technology Perspectives 2023,” 2023, Accessed: Apr. 11, 2023. [Online]. Available: www.iea.org
- [25] M. Child, C. Kemfert, D. Bogdanov, and C. Breyer, “Flexible electricity generation, grid exchange and storage for the transition to a 100% renewable energy system in Europe,” *Renew. Energy*, vol. 139, pp. 80–101, Aug. 2019, doi: 10.1016/J.RENENE.2019.02.077.
- [26] S. Ambec and C. Crampes, “Decarbonizing Electricity Generation with Intermittent Sources of Energy,” <https://doi.org/10.1086/705536>, vol. 6, no. 6, pp. 1105–1134, Nov. 2019, doi: 10.1086/705536.
- [27] A. Rode *et al.*, “Estimating a social cost of carbon for global energy consumption,” *Nat.* 2021 5987880, vol. 598, no. 7880, pp. 308–314, Oct. 2021, doi: 10.1038/s41586-021-03883-8.
- [28] A. Zurita *et al.*, “State of the art and future prospects for solar PV development in Chile,” *Renew. Sustain. Energy Rev.*, vol. 92, pp. 701–727, Sep. 2018, doi: 10.1016/J.RSER.2018.04.096.
- [29] L. Ramirez Camargo, J. Valdes, Y. Masip Macia, and W. Dorner, “Assessment of on-site steady electricity generation from hybrid renewable energy systems in Chile,” *Appl. Energy*, vol. 250, pp. 1548–1558, Sep. 2019, doi: 10.1016/J.APENERGY.2019.05.005.

- [30] A. Zurita *et al.*, “State of the art and future prospects for solar PV development in Chile,” *Renew. Sustain. Energy Rev.*, vol. 92, no. November 2017, pp. 701–727, 2018, doi: 10.1016/j.rser.2018.04.096.
- [31] R. Prävǎlie, C. Patriche, and G. Bandoc, “Spatial assessment of solar energy potential at global scale. A geographical approach,” *J. Clean. Prod.*, vol. 209, pp. 692–721, Feb. 2019, doi: 10.1016/J.JCLEPRO.2018.10.239.
- [32] P. K. Nayak, S. Mahesh, H. J. Snaith, and D. Cahen, “Photovoltaic solar cell technologies: analysing the state of the art,” *Nat. Rev. Mater.* 2019 44, vol. 4, no. 4, pp. 269–285, Mar. 2019, doi: 10.1038/s41578-019-0097-0.
- [33] T. M. Razykov, C. S. Ferekides, D. Morel, E. Stefanakos, H. S. Ullal, and H. M. Upadhyaya, “Solar photovoltaic electricity: Current status and future prospects,” *Sol. Energy*, vol. 85, no. 8, pp. 1580–1608, Aug. 2011, doi: 10.1016/J.SOLENER.2010.12.002.
- [34] S. Nabi Mughal, R. K. Jarial, S. Mughal, and Y. R. Sood, “A Review on Solar Photovoltaic Technology and Future Trends,” *NCRACIT Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* © 2018 IJSRCSEIT, vol. 1, no. 4, pp. 227–235, 2018, Accessed: Apr. 04, 2023. [Online]. Available: <https://www.researchgate.net/publication/324922616>
- [35] M. Tawalbeh, A. Al-Othman, F. Kafiah, E. Abdelsalam, F. Almomani, and M. Alkasrawi, “Environmental impacts of solar photovoltaic systems: A critical review of recent progress and future outlook,” *Sci. Total Environ.*, vol. 759, p. 143528, Mar. 2021, doi: 10.1016/J.SCITOTENV.2020.143528.
- [36] T. D. Lee and A. U. Ebong, “A review of thin film solar cell technologies and challenges,” *Renew. Sustain. Energy Rev.*, vol. 70, pp. 1286–1297, Apr. 2017, doi: 10.1016/J.RSER.2016.12.028.
- [37] T. D. Lee and A. Ebong, “Thin film solar technologies: A review,” *2015 12th Int. Conf. High-Capacity Opt. Networks Enabling/Emerging Technol. HONET-ICT 2015*, Jan. 2016, doi: 10.1109/HONET.2015.7395441.
- [38] M. A. Green, E. D. Dunlop, J. Hohl-Ebinger, M. Yoshita, N. Kopidakis, and X. Hao, “SOLAR CELL EFFICIENCY TABLES (VERSION 58),” 1099, doi: 10.1002/pip.3444.
- [39] W. Gu, T. Ma, S. Ahmed, Y. Zhang, and J. Peng, “A comprehensive review and outlook of bifacial photovoltaic (bPV) technology,” *Energy Convers. Manag.*, vol. 223, p. 113283, Nov. 2020, doi: 10.1016/J.ENCONMAN.2020.113283.
- [40] E. G. Luque, F. Antonanzas-Torres, and R. Escobar, “Effect of soiling in bifacial PV modules and cleaning schedule optimization,” *Energy Convers. Manag.*, vol. 174, pp. 615–625, 2018, doi: 10.1016/j.enconman.2018.08.065.
- [41] C. D. Rodríguez-Gallegos *et al.*, “Global Techno-Economic Performance of Bifacial and Tracking Photovoltaic Systems,” *Joule*, vol. 4, no. 7, pp. 1514–1541, Jul. 2020, doi: 10.1016/J.JOULE.2020.05.005.
- [42] Q. Tao, P. Xu, M. Li, and W. Lu, “Machine learning for perovskite materials design and discovery,” *npj Comput. Mater.* 2021 71, vol. 7, no. 1, pp. 1–18, Jan. 2021, doi: 10.1038/s41524-021-00495-8.
- [43] National Renewable Energy Laboratory, “Best Research-Cell Efficiencies,” p. 1, 2023, [Online]. Available: <https://www.nrel.gov/pv/assets/pdfs/best-research-cell-efficiencies.pdf>
- [44] E. T. Efazl *et al.*, “A review of primary technologies of thin-film solar cells,” *Eng. Res. Express*, vol. 3, no. 3, p. 032001, Sep. 2021, doi: 10.1088/2631-8695/AC2353.

- [45] N. L. Muttumthala and A. Yadav, "A concise overview of thin film photovoltaics," *Mater. Today Proc.*, vol. 64, pp. 1475–1478, Jan. 2022, doi: 10.1016/J.MATPR.2022.04.862.
- [46] J. A. Luceño-Sánchez, A. M. Díez-Pascual, and R. P. Capilla, "Materials for Photovoltaics: State of Art and Recent Developments," *Int. J. Mol. Sci.* 2019, Vol. 20, Page 976, vol. 20, no. 4, p. 976, Feb. 2019, doi: 10.3390/IJMS20040976.
- [47] M. Li *et al.*, "Indoor Thin-Film Photovoltaics: Progress and Challenges," *Adv. Energy Mater.*, vol. 10, no. 28, p. 2000641, Jul. 2020, doi: 10.1002/AENM.202000641.
- [48] V. Bermudez and A. Perez-Rodriguez, "Understanding the cell-to-module efficiency gap in Cu(In,Ga)(S,Se)₂ photovoltaics scale-up," *Nat. Energy*, vol. 3, no. 6, pp. 466–475, 2018, doi: 10.1038/s41560-018-0177-1.
- [49] W. Hoffmann and T. Pellkofer, "Thin films in photovoltaics: Technologies and perspectives," *Thin Solid Films*, vol. 520, no. 12, pp. 4094–4100, Apr. 2012, doi: 10.1016/J.TSF.2011.04.146.
- [50] T. Feurer *et al.*, "Progress in thin film CIGS photovoltaics – Research and development, manufacturing, and applications," *Prog. Photovoltaics Res. Appl.*, vol. 25, no. 7, pp. 645–667, Jul. 2017, doi: 10.1002/PIP.2811.
- [51] B. J. Stanbery, D. Abou-Ras, A. Yamada, and L. Mansfield, "CIGS photovoltaics: reviewing an evolving paradigm," *J. Phys. D. Appl. Phys.*, vol. 55, no. 17, p. 173001, Dec. 2021, doi: 10.1088/1361-6463/AC4363.
- [52] S. Shi *et al.*, "Recent progress in the high-temperature-resistant PI substrate with low CTE for CIGS thin-film solar cells," *Mater. Today Energy*, vol. 20, p. 100640, Jun. 2021, doi: 10.1016/J.MTENER.2021.100640.
- [53] M. K. van der Hulst *et al.*, "A systematic approach to assess the environmental impact of emerging technologies: A case study for the GHG footprint of CIGS solar photovoltaic laminate," *J. Ind. Ecol.*, vol. 24, no. 6, pp. 1234–1249, Dec. 2020, doi: 10.1111/JIEC.13027.
- [54] M. Neuschitzer *et al.*, "Optimization of CdS buffer layer for high-performance Cu₂ZnSnSe₄ solar cells and the effects of light soaking: elimination of crossover and red kink," *Prog. Photovoltaics Res. Appl.*, vol. 23, no. 11, pp. 1660–1667, Nov. 2015, doi: 10.1002/PIP.2589.
- [55] E. and Sme. Directorate-General for Internal Market, Industry, "Study on the Critical Raw Materials for the EU," 2023.
- [56] E. Gervais, S. Shammugam, L. Friedrich, and T. Schlegl, "Raw material needs for the large-scale deployment of photovoltaics – Effects of innovation-driven roadmaps on material constraints until 2050," *Renew. Sustain. Energy Rev.*, vol. 137, p. 110589, Mar. 2021, doi: 10.1016/J.RSER.2020.110589.
- [57] C. Helbig, A. M. Bradshaw, C. Kolotzek, A. Thorenz, and A. Tuma, "Supply risks associated with CdTe and CIGS thin-film photovoltaics," *Appl. Energy*, vol. 178, pp. 422–433, Sep. 2016, doi: 10.1016/J.APENERGY.2016.06.102.
- [58] T. E. Graedel *et al.*, "Recycling Rates of Metals: A Status Report," 2011.
- [59] M. Redlinger, R. Eggert, and M. Woodhouse, "Evaluating the availability of gallium, indium, and tellurium from recycled photovoltaic modules," *Sol. Energy Mater. Sol. Cells*, vol. 138, pp. 58–71, Jul. 2015, doi: 10.1016/J.SOLMAT.2015.02.027.
- [60] S. Giraldo, Z. Jehl, M. Placidi, V. Izquierdo-Roca, A. Pérez-Rodríguez, and E. Saucedo, "Progress and Perspectives of Thin Film Kesterite Photovoltaic Technology: A Critical Review," *Adv. Mater.*, vol. 31, no. 16, p. 1806692, Apr. 2019, doi:

- 10.1002/ADMA.201806692.
- [61] S. Delbos, “Kesterite thin films for photovoltaics : a review,” *EPJ Photovoltaics*, vol. 3, p. 35004, 2012, doi: 10.1051/EPJPV/2012008.
- [62] M. A. Green *et al.*, “Solar cell efficiency tables (Version 63),” *Prog. Photovoltaics Res. Appl.*, vol. 32, no. 1, pp. 3–13, Jan. 2024, doi: 10.1002/PIP.3750.
- [63] Nisika, K. Kaur, and M. Kumar, “Progress and prospects of CZTSSe/CdS interface engineering to combat high open-circuit voltage deficit of kesterite photovoltaics: a critical review,” *J. Mater. Chem. A*, vol. 8, no. 41, pp. 21547–21584, Oct. 2020, doi: 10.1039/D0TA06450E.
- [64] J. Li, D. Wang, X. Li, Y. Zeng, and Y. Zhang, “Cation Substitution in Earth-Abundant Kesterite Photovoltaic Materials,” *Adv. Sci.*, vol. 5, no. 4, p. 1700744, Apr. 2018, doi: 10.1002/ADVS.201700744.
- [65] Q. Tian and S. Liu, “Defect suppression in multinary chalcogenide photovoltaic materials derived from kesterite: progress and outlook,” *J. Mater. Chem. A*, vol. 8, no. 47, pp. 24920–24942, Dec. 2020, doi: 10.1039/D0TA08202C.
- [66] M. He, K. Sun, M. P. Suryawanshi, J. Li, and X. Hao, “Interface engineering of p-n heterojunction for kesterite photovoltaics: A progress review,” *J. Energy Chem.*, vol. 60, pp. 1–8, Sep. 2021, doi: 10.1016/J.JEACHEM.2020.12.019.
- [67] R. S. Ohl, “Light-sensitive electric device including silicon,” US2443542A, Sep. 17, 1941
- [68] S. DeWolf, A. Descoeurdes, Z. C. Holman, and C. Ballif, “High-efficiency silicon heterojunction solar cells: A review,” *Green*, vol. 2, no. 1, pp. 7–24, Mar. 2012, doi: 10.1515/GREEN-2011-0018/MACHINEREADEABLECITATION/RIS.
- [69] L. Huang, J. Wang, and Y. Zhu, “Efficient p-n Heterojunction Perovskite Solar Cell without a Redundant Electron Transport Layer and Interface Engineering,” *J. Phys. Chem. Lett.*, vol. 12, no. 9, pp. 2266–2272, Mar. 2021, doi: 10.1021/ACS.JPCLETT.1C00417/SUPPL_FILE/JZ1C00417_SI_001.PDF.
- [70] S. Wagner, J. L. Shay, P. Migliorato, and H. M. Kasper, “CuInSe₂/CdS heterojunction photovoltaic detectors,” *Appl. Phys. Lett.*, vol. 25, no. 8, pp. 434–435, Oct. 1974, doi: 10.1063/1.1655537.
- [71] W. J. Lee *et al.*, “Behavior of Photocarriers in the Light-Induced Metastable State in the p-n Heterojunction of a Cu(In,Ga)Se₂ Solar Cell with CBD-ZnS Buffer Layer,” *ACS Appl. Mater. Interfaces*, vol. 8, no. 34, pp. 22151–22158, Aug. 2016, doi: 10.1021/ACSAMI.6B05005/SUPPL_FILE/AM6B05005_SI_001.PDF.
- [72] W. Shockley and H. J. Queisser, “Detailed Balance Limit of Efficiency of p-n Junction Solar Cells,” *J. Appl. Phys.*, vol. 32, no. 3, pp. 510–519, Mar. 1961, doi: 10.1063/1.1736034.
- [73] E. Maine and E. Garnsey, “Commercializing generic technology: The case of advanced materials ventures,” *Res. Policy*, vol. 35, pp. 375–393, 2006, doi: 10.1016/j.respol.2005.12.006.
- [74] P. J. McGinn, “Thin-Film Processing Routes for Combinatorial Materials Investigations-A Review,” *ACS Comb. Sci.*, vol. 21, no. 7, pp. 501–515, Jul. 2019, doi: 10.1021/ACSCOMBSCI.9B00032.
- [75] A. Aspuru-Guzik and K. Persson, “Materials Acceleration Platform - Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods with Artificial Intelligence,” *Rep. Clean Energy Mater. Innov. Chall. Expert Work.*, no. January, pp. 1–108, 2018, Accessed: Jun. 07, 2021. [Online]. Available:

- <http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974>
- [76] R. K. Vasudevan *et al.*, “Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics,” *MRS Commun.*, vol. 9, no. 3, pp. 821–838, Sep. 2019, doi: 10.1557/MRC.2019.95.
 - [77] 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, Jul. 2018, doi: 10.1038/s41586-018-0337-2.
 - [78] R. Fonoll-Rubio *et al.*, “Combinatorial Analysis Methodologies for Accelerated Research: The Case of Chalcogenide Thin-Film Photovoltaic Technologies,” *Sol. RRL*, p. 2200235, Jul. 2022, doi: 10.1002/SOLR.202200235.
 - [79] G. Carleo *et al.*, “Machine learning and the physical sciences,” *Rev. Mod. Phys.*, vol. 91, no. 4, p. 045002, Dec. 2019, doi: 10.1103/RevModPhys.91.045002.
 - [80] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 91, no. 5. MIT Press, 2016. doi: 10.1017/CBO9781107415324.004.
 - [81] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, and S. P. Ong, “A Critical Review of Machine Learning of Energy Materials,” *Adv. Energy Mater.*, vol. 10, no. 8, p. 1903242, Feb. 2020, doi: 10.1002/AENM.201903242.
 - [82] Y. Goldberg, “A Primer on Neural Network Models for Natural Language Processing,” *J. Artif. Intell. Res.*, vol. 57, pp. 345–420, Oct. 2015, doi: 10.1613/jair.4992.
 - [83] Y. Liu, O. C. Esan, Z. Pan, and L. An, “Machine learning for advanced energy materials,” *Energy AI*, vol. 3, p. 100049, Mar. 2021, doi: 10.1016/j.egyai.2021.100049.
 - [84] Z. Lu, “Computational discovery of energy materials in the era of big data and machine learning: A critical review,” *Mater. Reports Energy*, vol. 1, no. 3, p. 100047, Aug. 2021, doi: 10.1016/J.MATRE.2021.100047.
 - [85] G. H. Gu, J. Noh, I. Kim, and Y. Jung, “Machine learning for renewable energy materials,” *J. Mater. Chem. A*, vol. 7, no. 29, pp. 17096–17117, Jul. 2019, doi: 10.1039/c9ta02356a.
 - [86] J.-P. Correa-Baena *et al.*, “Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing,” *Joule*, 2018, doi: 10.1016/j.joule.2018.05.009.
 - [87] Goodman Bryce and Flaxman Seth, “European Union Regulations on Algorithmic Decision Making and a ‘Right to Explanation,’” *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.
 - [88] D. Boyd and K. Crawford, “Critical Questions for Big Data,” *Information, Commun. Soc.*, vol. 15, no. 5, pp. 662–679, Jun. 2012, doi: 10.1080/1369118X.2012.678878.
 - [89] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” *NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess.*, pp. 97–101, Feb. 2016, doi: 10.48550/arxiv.1602.04938.
 - [90] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, “Explainable Machine Learning for Scientific Insights and Discoveries,” *IEEE Access*, vol. 8, pp. 42200–42216, 2020, doi: 10.1109/ACCESS.2020.2976199.
 - [91] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” *Proc. - 2018 IEEE 5th Int. Conf. Data Sci. Adv. Anal. DSAA 2018*, pp. 80–89, Jan. 2019, doi: 10.1109/DSAA.2018.00018.

- [92] R. Fonoll-Rubio *et al.*, “Insights into the Effects of RbF-Post-Deposition Treatments on the Absorber Surface of High Efficiency Cu(In,Ga)Se₂ Solar Cells and Development of Analytical and Machine Learning Process Monitoring Methodologies Based on Combinatorial Analysis,” *Adv. Energy Mater.*, p. 2103163, Jan. 2022, doi: 10.1002/AENM.202103163.
- [93] X. Kong, C. Hu, and Z. Duan, *Principal component analysis networks and algorithms*, 1st ed. Singapore: Springer Singapore, 2017. doi: 10.1007/978-981-10-2915-8.
- [94] A. J. Izenman, “Linear Discriminant Analysis,” in *Modern Multivariate Statistical Techniques*, 1st ed. Springer, New York, NY, 2008, pp. 237–280. doi: 10.1007/978-0-387-78189-1_8.
- [95] Y. Wang, J. Zhu, and X. Chen, “Autofluorescence spectroscopy of blood plasma with multivariate analysis methods for the diagnosis of pulmonary tuberculosis,” *Optik (Stuttg.)*, vol. 224, p. 165446, Dec. 2020, doi: 10.1016/j.ijleo.2020.165446.
- [96] D. O’Dea *et al.*, “Raman spectroscopy for the preoperative diagnosis of thyroid cancer and its subtypes: An in vitro proof-of-concept study,” *Cytopathology*, vol. 30, no. 1, pp. 51–60, Jan. 2019, doi: 10.1111/CYT.12636.
- [97] N. M. Ralbovsky and I. K. Lednev, “Raman spectroscopy and chemometrics: A potential universal method for diagnosing cancer,” *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 219, pp. 463–487, 2019, doi: 10.1016/j.saa.2019.04.067.
- [98] S. A. Billings and G. L. Zheng, “Radial basis function network configuration using genetic algorithms,” *Neural Networks*, vol. 8, no. 6, pp. 877–890, Jan. 1995, doi: 10.1016/0893-6080(95)00029-Y.
- [99] F. Van Veen and S. Leijnen, “The Neural Network Zoo,” 2019. <https://www.asimovinstitute.org/neural-network-zoo>
- [100] J. Wei *et al.*, “Machine learning in materials science,” *InfoMat*, vol. 1, no. 3, pp. 338–358, Sep. 2019, doi: 10.1002/INF2.12028.
- [101] N. Burkart and M. F. Huber, “A Survey on the Explainability of Supervised Machine Learning,” *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021, doi: 10.1613/JAIR.1.12228.
- [102] U. Bhatt *et al.*, “Explainable Machine Learning in Deployment,” 2020, doi: 10.1145/3351095.3375624.
- [103] S. M. and L. K. Julia Angwin, Jeff Larson, “Machine Bias,” 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [104] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science (80-.)*, vol. 356, no. 6334, pp. 183–186, Apr. 2017, doi: 10.1126/SCIENCE.AAL4230.
- [105] K. Kira and L. A. Rendell, “A Practical Approach to Feature Selection,” *Mach. Learn. Proc. 1992*, pp. 249–256, Jan. 1992, doi: 10.1016/B978-1-55860-247-2.50037-1.
- [106] T. Mueller, A. G. Kusne, and R. Ramprasad, “Machine Learning in Materials Science: Recent Progress and Emerging Applications,” in *Reviews in Computational Chemistry*, Wiley, 2016, pp. 186–273. doi: 10.1002/9781119148739.ch4.
- [107] C. P. Gomes, B. Selman, and J. M. Gregoire, “Artificial intelligence for materials discovery,” *MRS Bull.*, vol. 44, no. 7, pp. 538–544, Jul. 2019, doi: 10.1557/MRS.2019.158.
- [108] D. Sheppard, “Robert Le Rossignol, 1884-1976: Engineer of the ‘Haber’ process,” *Notes Rec. R. Soc. Lond.*, vol. 71, no. 3, pp. 263–296, 2017, doi: 10.1098/RSNR.2016.0019.

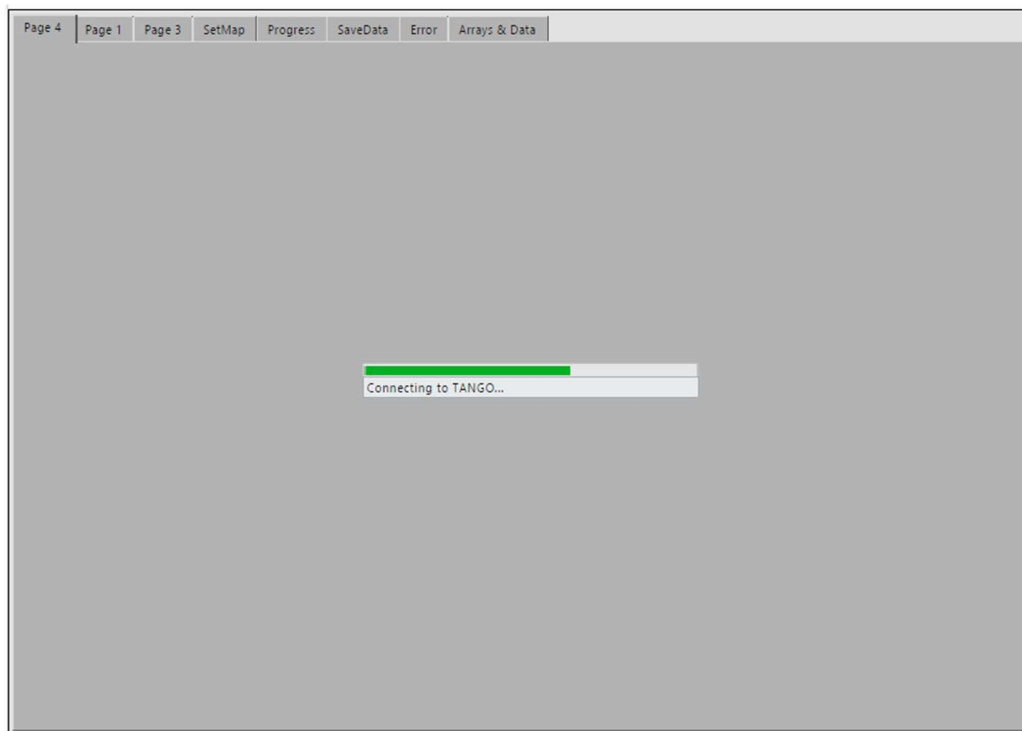
- [109] A. Mahmood and J.-L. Wang, "Machine learning for high performance organic solar cells: current scenario and future prospects," *Energy Environ. Sci.*, vol. 14, no. 1, pp. 90–105, Jan. 2021, doi: 10.1039/d0ee02838j.
- [110] S. Krishn 1 *et al.*, "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective," Feb. 2022, Accessed: Sep. 06, 2023. [Online]. Available: <https://arxiv.org/abs/2202.01602v3>
- [111] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning", doi: 10.1145/3313831.3376219.
- [112] E. Grau-Luque *et al.*, "Thickness evaluation of AlOx barrier layers for encapsulation of flexible PV modules in industrial environments by normal reflectance and machine learning," *Prog. Photovoltaics Res. Appl.*, vol. 30, no. 3, pp. 229–239, Mar. 2022, doi: 10.1002/pip.3478.
- [113] D. A. Whitaker and K. Hayes, "A simple algorithm for despiking Raman spectra," *Chemom. Intell. Lab. Syst.*, vol. 179, pp. 82–84, Aug. 2018, doi: 10.1016/j.chemolab.2018.06.009.
- [114] S. J. Barton and B. M. Hennelly, "An Algorithm for the Removal of Cosmic Ray Artifacts in Spectral Data Sets," *Appl. Spectrosc.*, vol. 73, no. 8, pp. 893–901, Aug. 2019, doi: 10.1177/0003702819839098.
- [115] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big Data of Materials Science: Critical Role of the Descriptor," *Phys. Rev. Lett.*, vol. 114, no. 10, p. 105503, Mar. 2015, doi: 10.1103/PhysRevLett.114.105503.
- [116] M. Dimitrievska *et al.*, "Defect characterisation in Cu₂ZnSnSe₄ kesterites via resonance Raman spectroscopy and the impact on optoelectronic solar cell properties," *J. Mater. Chem. A*, vol. 7, no. 21, pp. 13293–13304, May 2019, doi: 10.1039/C9TA03625C.
- [117] A. de Juan and R. Tauler, "Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review," *Anal. Chim. Acta*, vol. 1145, pp. 59–78, Feb. 2021, doi: 10.1016/J.ACA.2020.10.051.
- [118] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," 2011. Accessed: Jun. 07, 2021. [Online]. Available: <http://scikit-learn.sourceforge.net>.
- [119] J. C. Vickerman and I. S. Gilmore, *Surface Analysis-The Principal Techniques*, 2nd Editio. Wiley, 2009.
- [120] D. Bau *et al.*, "GaN dissection: Visualizing and understanding generative adversarial networks," *7th Int. Conf. Learn. Represent. ICLR 2019*, 2019.
- [121] D. Bau, B. Zhou, A. Khosla, O. Aude, and A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representation," 2017.
- [122] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Jan. 18, 2023. [Online]. Available: <https://github.com/slundberg/shap>
- [123] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization." pp. 618–626, 2017. doi: 10.1109/iccv.2017.74.
- [124] D. Bau, J. Y. Zhu, H. Strobel, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 48, pp. 30071–30078, Dec. 2020, doi: 10.1073/PNAS.1907375117/-/DCSUPPLEMENTAL.

- [125] E. Grau-Luque *et al.*, “Combinatorial and machine learning approaches for the analysis of $\text{Cu}_2\text{ZnGeSe}_4$: influence of the off-stoichiometry on defect formation and solar cell performance,” *J. Mater. Chem. A*, vol. 9, no. 16, pp. 10466–10476, Apr. 2021, doi: 10.1039/d1ta01299a.

ANNEXES

Annex A

User interface (UI) of the software developed using LabVIEW and used in the first article. This first version (v1) makes NF measurements and allows flexibility in several variables and parameters as shown in the UI. This version was developed solely by E.T.G.L.



Page 4 Page 1 Page 3 SetMap Progress SaveData Error Arrays & Data

User
NoUser

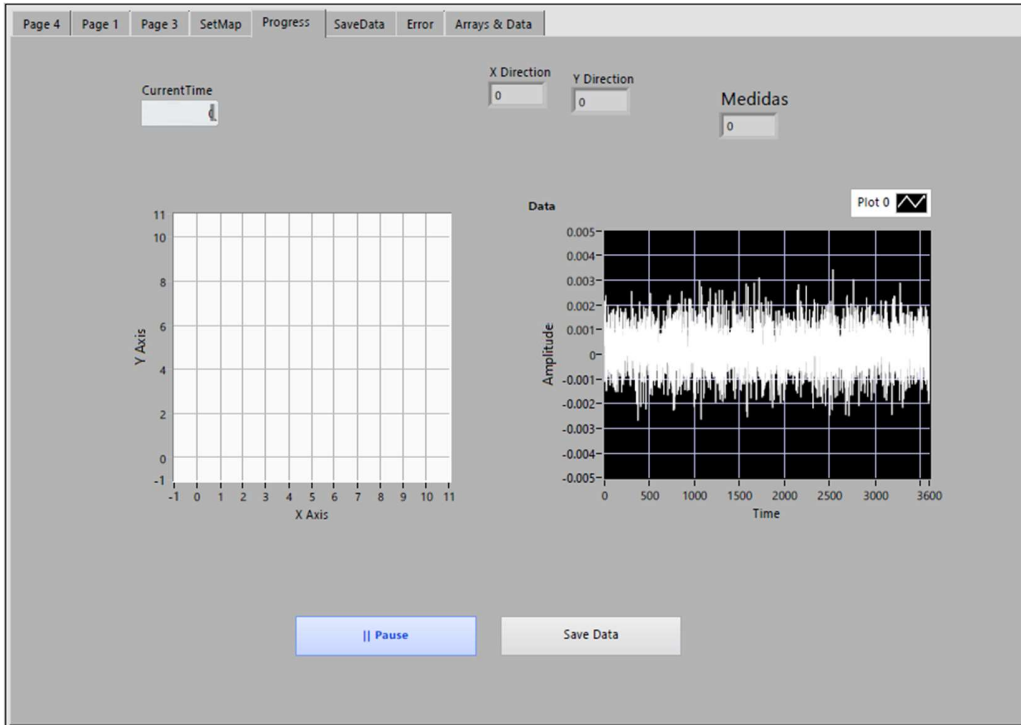
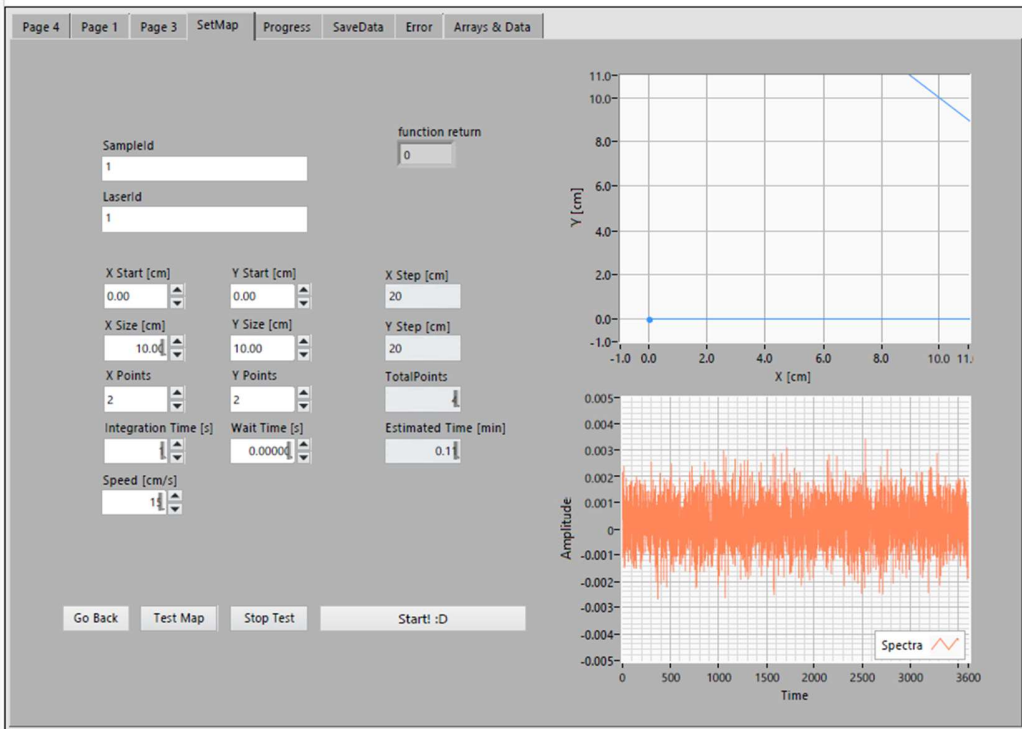
Select

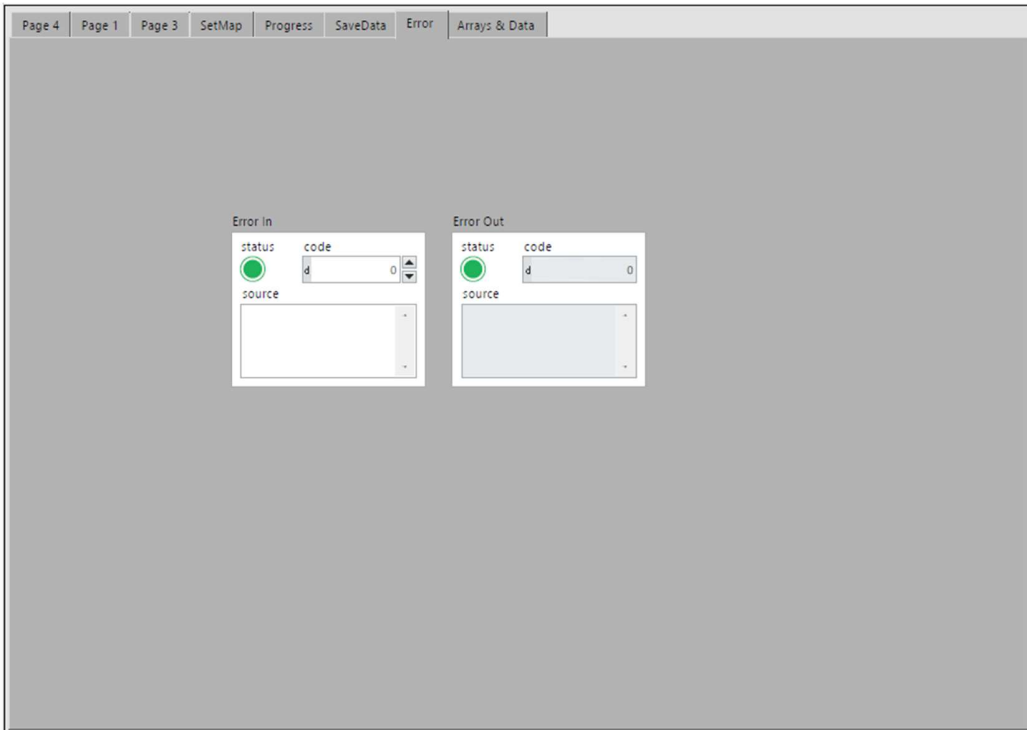
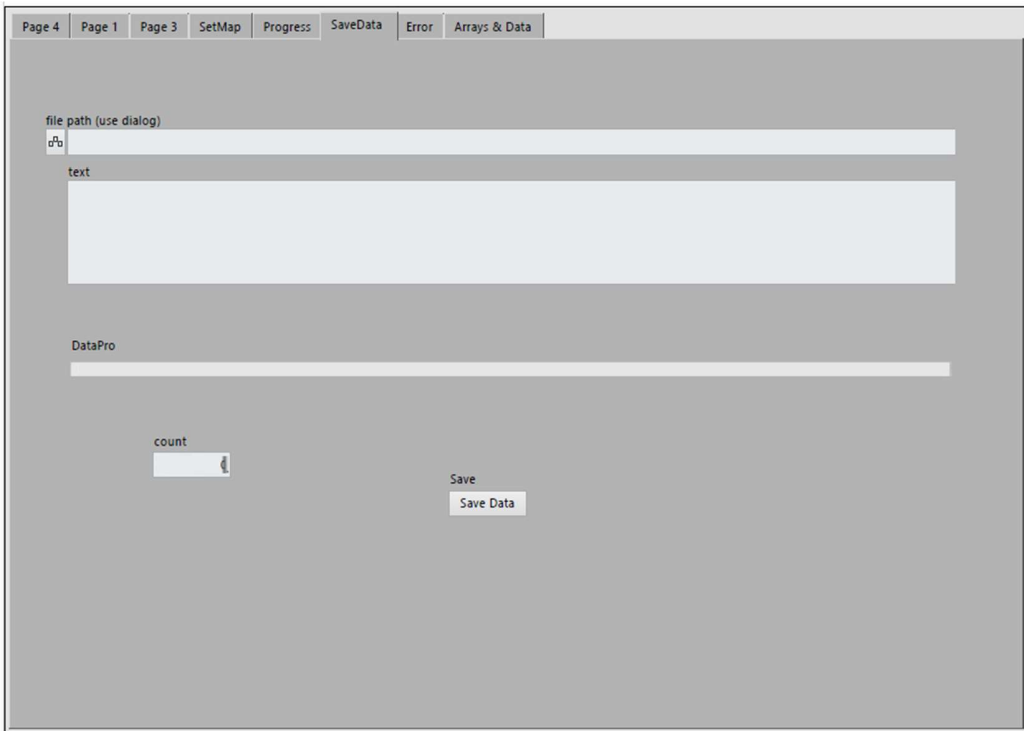
Page 4 Page 1 Page 3 SetMap Progress SaveData Error Arrays & Data

SetX SetY SetZero
Set Zero

LimitX LimitY
2 2

SetXY
Set XY





Page 4 | Page 1 | Page 3 | SetMap | Progress | SaveData | Error | Arrays & Data

XCoord: 0 | YCoord: 0 | IABaudRate: 57600

SpectroName: | SpectroSN: |

SpecData:

0	0	0
0	0	0

Distance: 0

CCS Handle: |

VISA resource name: |

ArraySize: 0

loadWait: 10

SpectroConnected: | TangoConnected:

X Scale.Range:Minimum: 0

Y Scale.Range:Minimum: 0

X Scale.Range:Maximum: 0

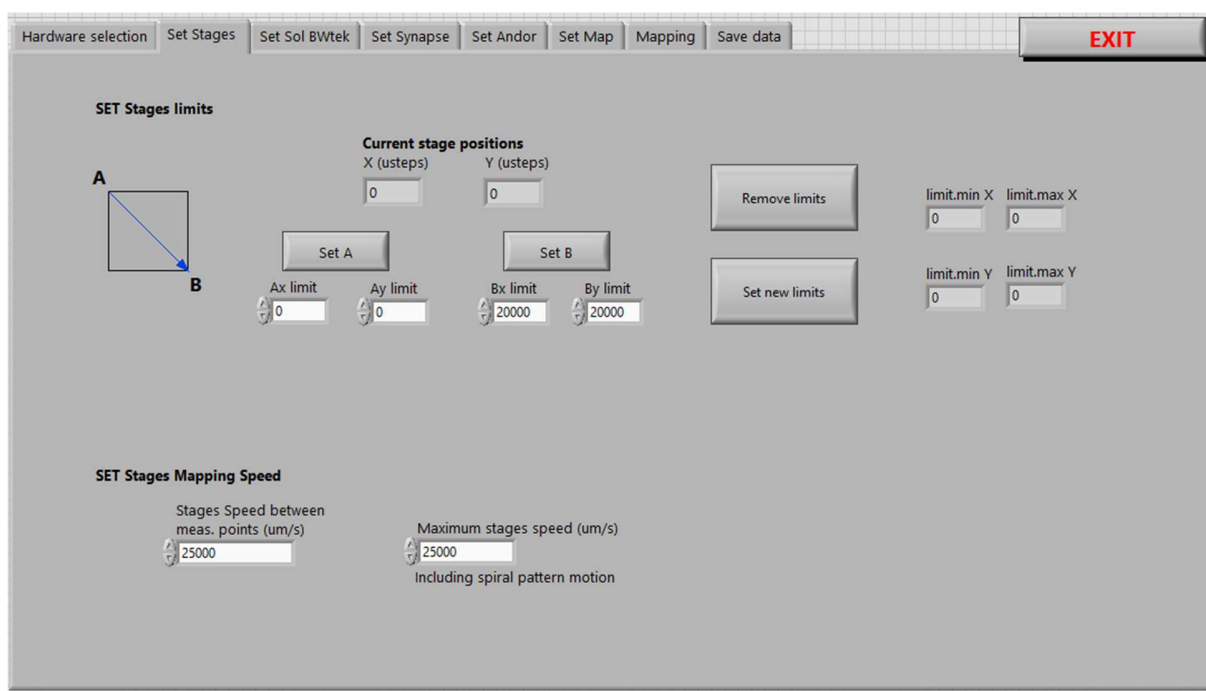
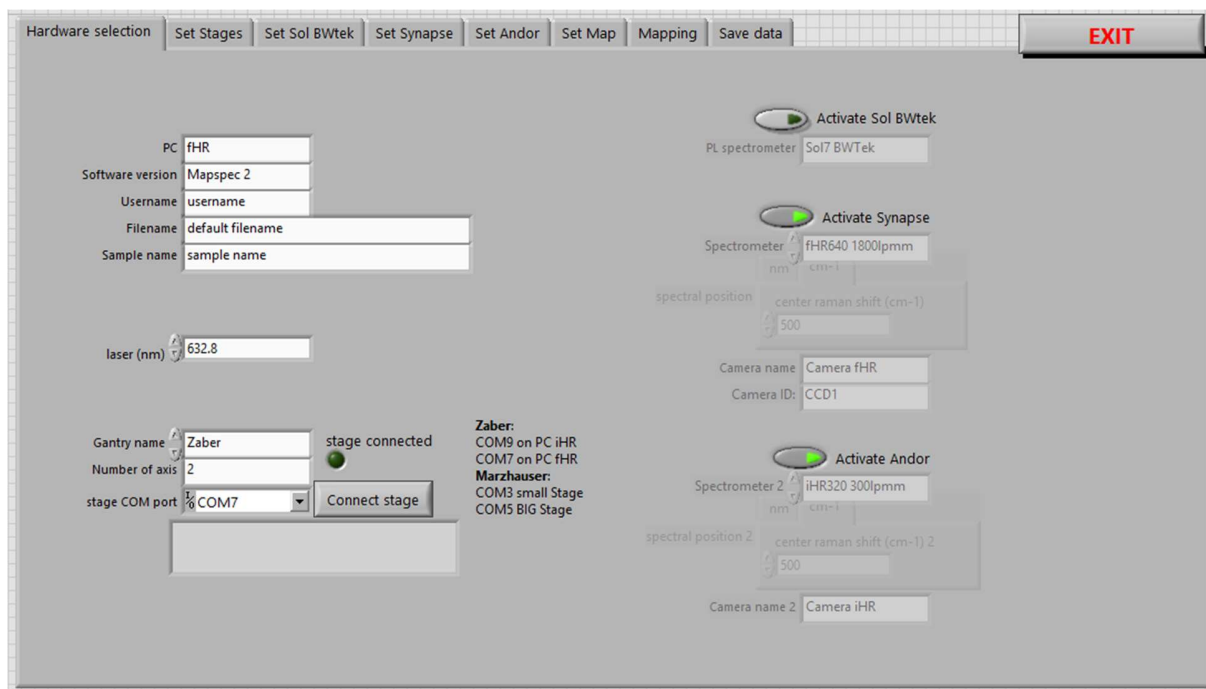
Y Scale.Range:Maximum: 0

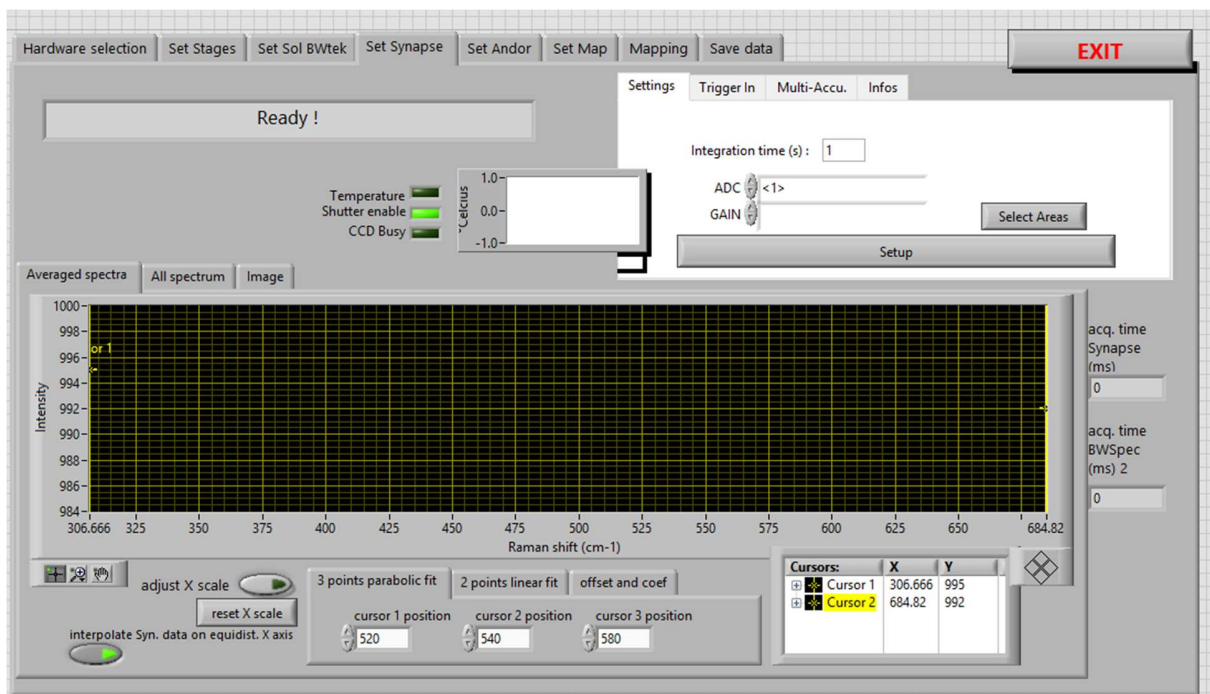
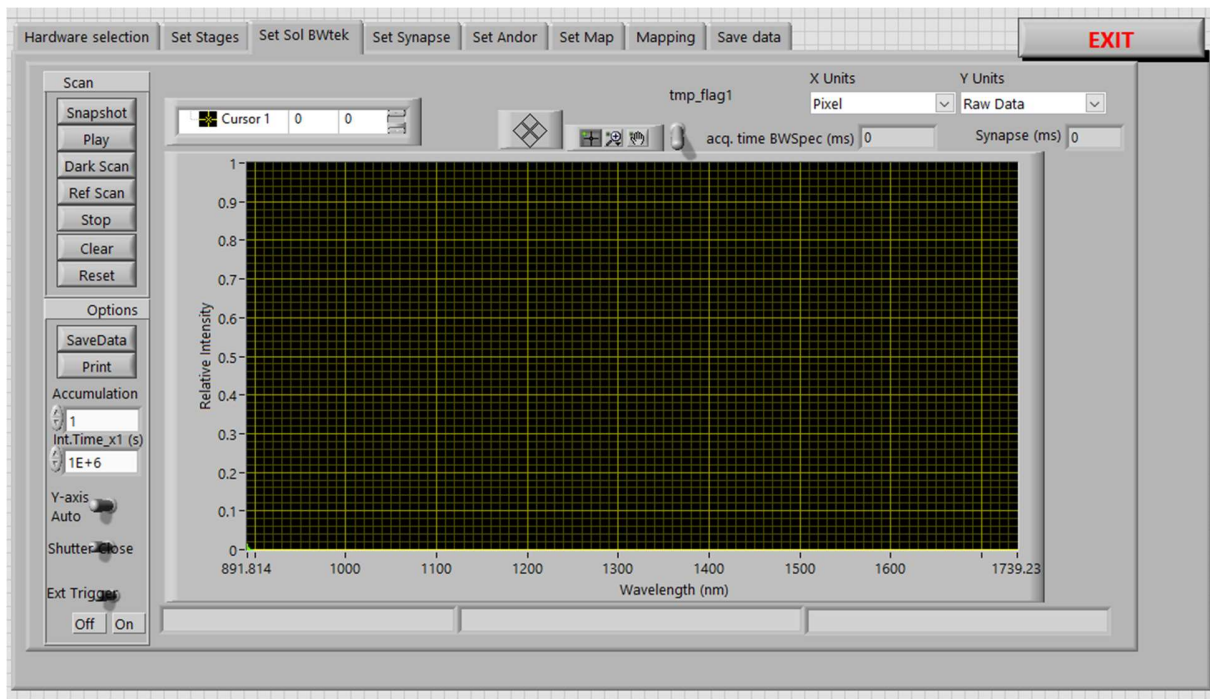
ILSID: 3

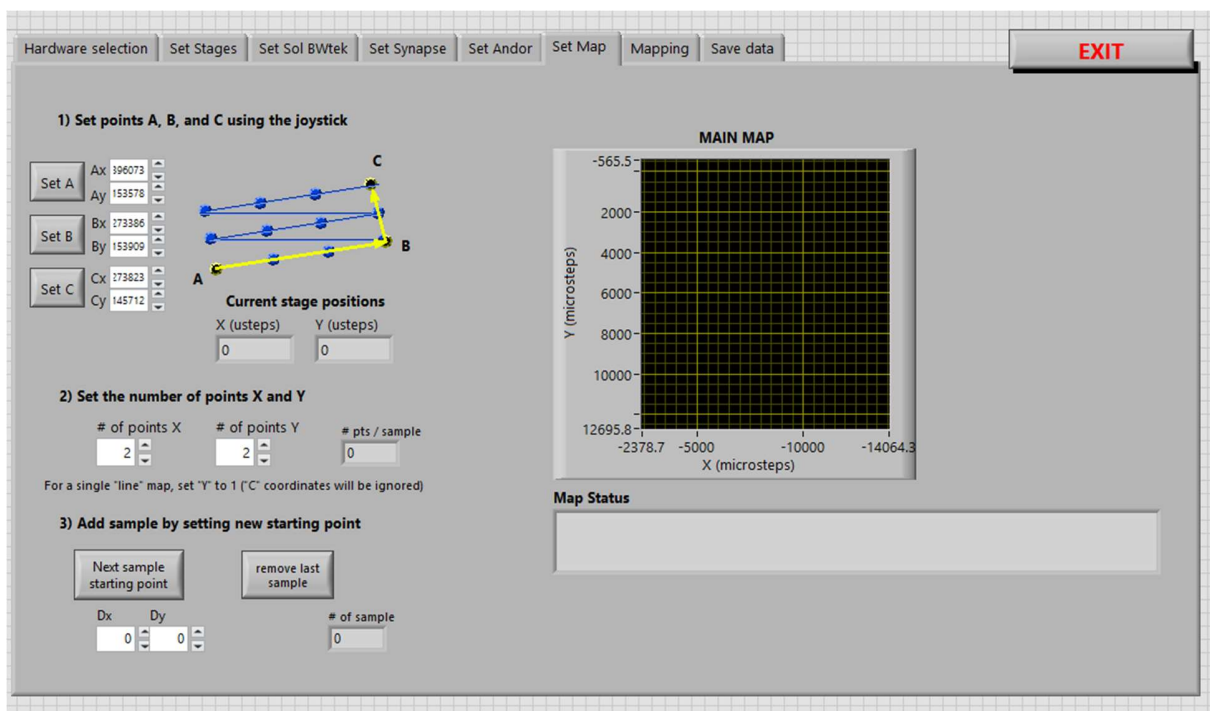
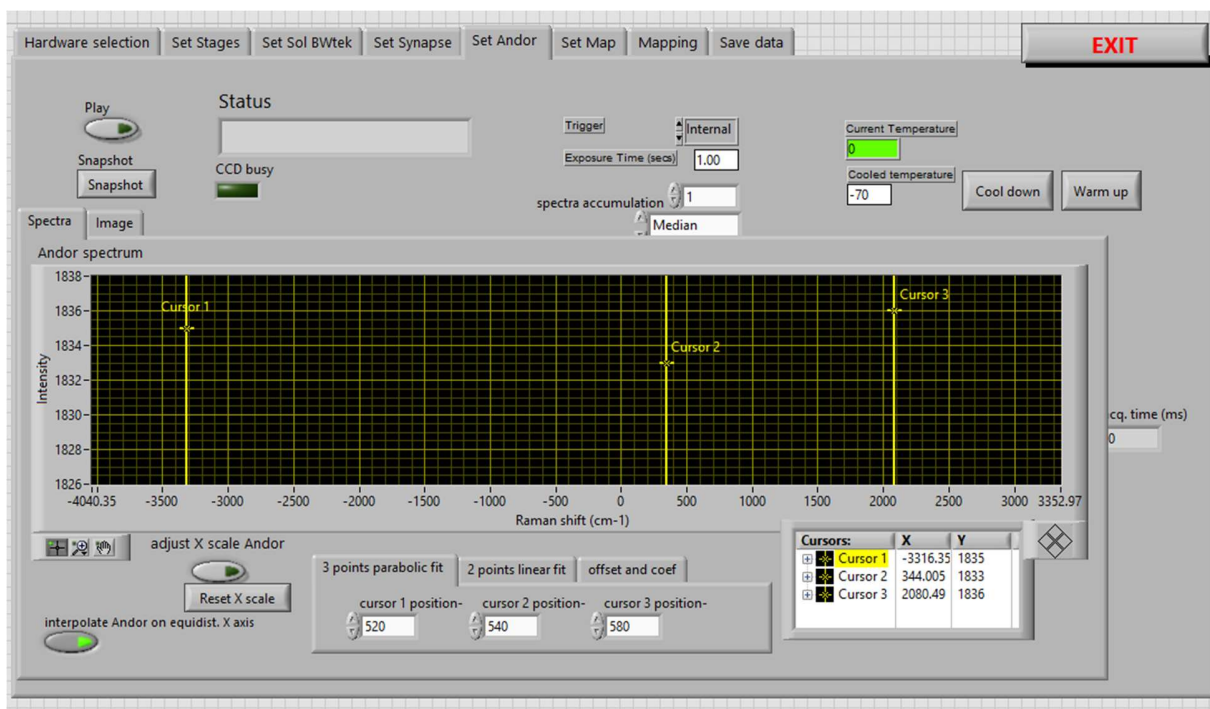
Case: Page 4

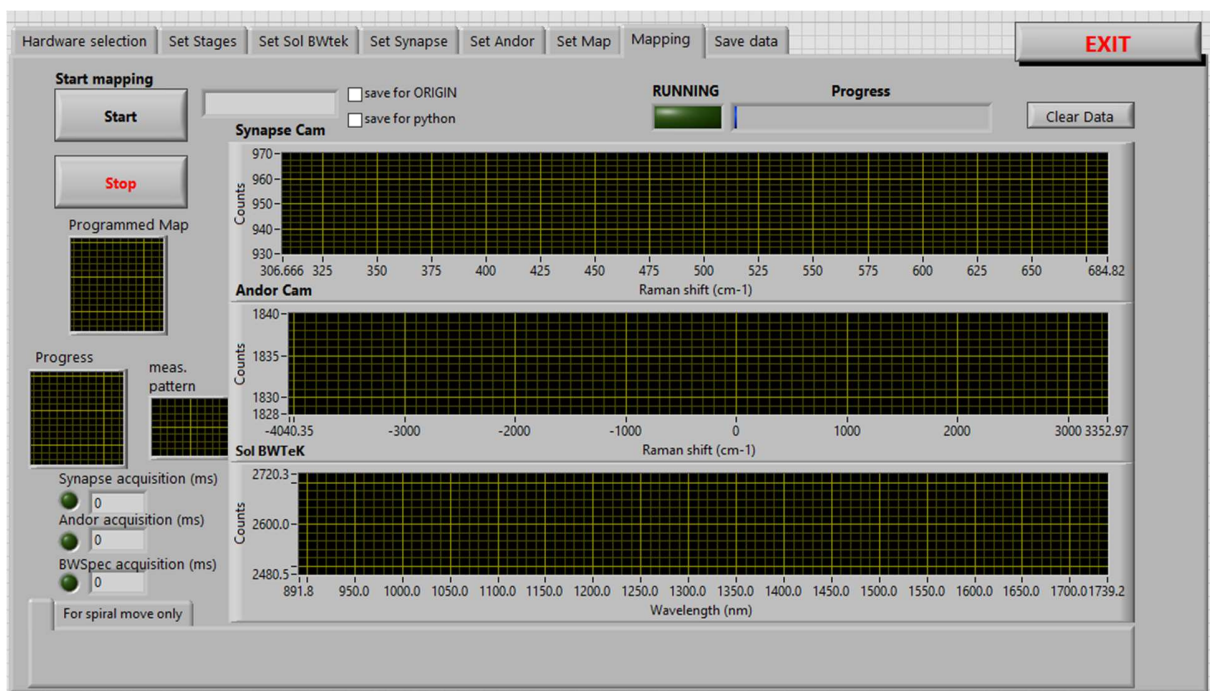
Annex B

User interface (UI) of the software developed using LabVIEW and used in the second article and exploratory experiments. This second version (v2) can perform multiple techniques quasi-simultaneously, including Raman and Photoluminescence. This version used the v1 as a starting point and is the result of cooperation with members of the SEMS teams in IREC.







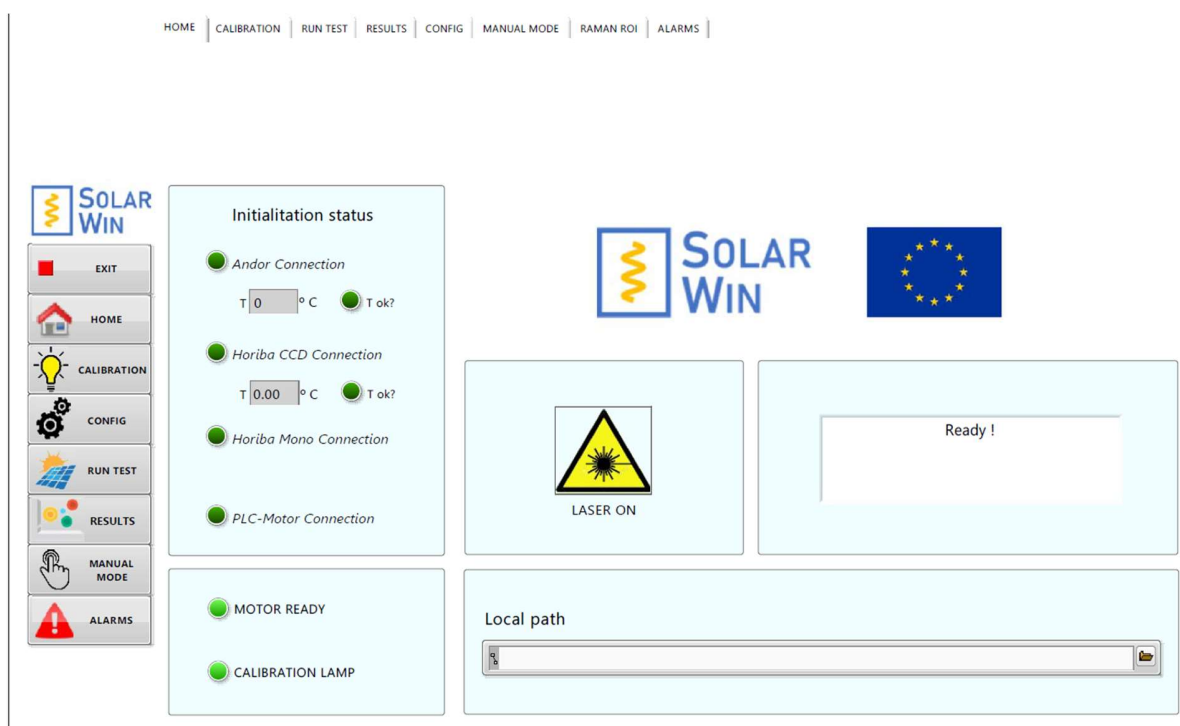


The screenshot shows the configuration page of the 'Mapping' software. At the top, there is a menu bar with options: Hardware selection, Set Stages, Set Sol BWTek, Set Synapse, Set Andor, Set Map, Mapping, Save data, and an EXIT button. The main area is divided into three columns for configuring different cameras:


- Sol BWTek:**
 - Data Format 2: .txt
 - data name prefix 2: prefix_Sol-
 - data name suffix 2: -suffix_Sol
 - data name previsualization: (empty field)
 - Save PL data (one file per sample)
- Synapse:**
 - Data Format: .txt
 - data name prefix: prefix_Synapse-
 - data name suffix: -suffix_Synapse
 - data name previsualization: (empty field)
 - Save Synapse data (one file per sample)
- Andor:**
 - Data Format 3: .txt
 - data name prefix 3: prefix_Andor-
 - data name suffix 3: -suffix_Andor
 - data name previsualization: (empty field)
 - Save Andor data (one file per sample)

Annex C

User interface (UI) of the software in production version for the Solar-Win project, based on the v1 and v2 softwares and greatly improved and optimized by the SEMS team in IREC. This version is further designed to be used in real process monitoring on industrial environments.

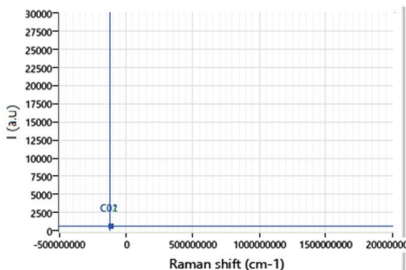


HOME | CALIBRATION | RUN TEST | RESULTS | CONFIG | MANUAL MODE | RAMAN ROI | ALARMS



CALIBRATION

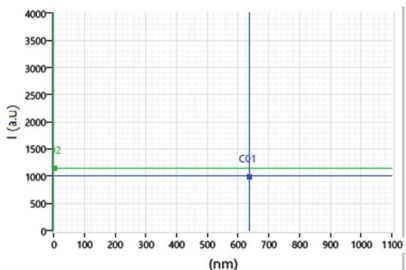
RAMAN



Data Raman

X1	-116954	Y1	612
X2	-116954	Y2	612

PL



Data PL

X1	634	Y1	995
X2	0	Y2	1151

RUN CALIBRATION

CALIBRATION LAMP
 LAMP READY
 Time left

CALIBRATION POSITIONS

Calibration name

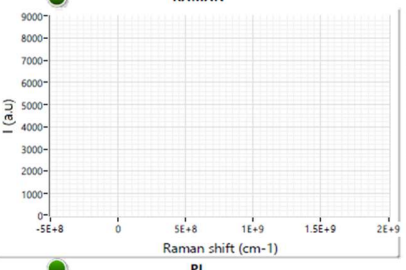
Comment

ROI RAMAN

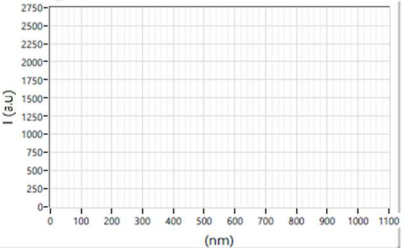
HOME | CALIBRATION | RUN TEST | RESULTS | CONFIG | MANUAL MODE | RAMAN ROI | ALARMS



RAMAN



PL



TEST SETUP

RUN TEST

ABORT TEST

Comment

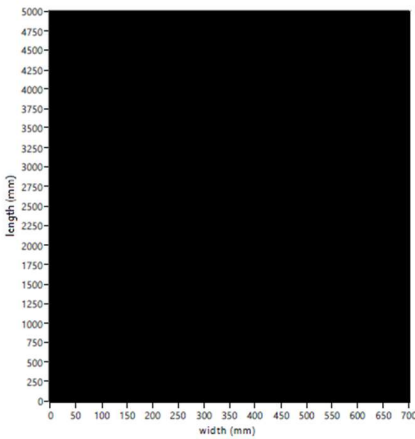
Measurement

X Motor Pos (mm)

Y Initial Pos (mm)

Y End Pos (mm)

PLATE ROLL SCAN Parameter RO



length (mm) vs width (mm)

HOME | CALIBRATION | RUN TEST | RESULTS | CONFIG | MANUAL MODE | RAMAN ROI | ALARMS

- EXIT
- HOME
- CALIBRATION
- CONFIG
- RUN TEST
- RESULTS
- MANUAL MODE
- ALARMS

RESULTS 1

RESULTS 2

MANUAL CONFIGURATION

R0	X low 0	X high 0
0	0	0
R1	X low 1	X high 1
0	0	0
R2	X low 2	X high 2
0	0	0
R3	X low 3	X high 3
0	0	0

PL

Area: 0, Area Mean: 0, Area std: 0

Max

Max: 0, Max Mean: 0, Max std: 0

FWHM

FWHM: 0, FWHM Mean: 0, FWHM std: 0

CONFIGURATION 1

RAMAN

R0	R0
0	R0 Mean
0	R0 std
R1	R1
0	R1 Mean
0	R1 std
R2	R2
0	R2 Mean
0	R2 std
R3	R3
0	R3 Mean
0	R3 std

HOME | CALIBRATION | RUN TEST | RESULTS | CONFIG | MANUAL MODE | RAMAN ROI | ALARMS

- EXIT
- HOME
- CALIBRATION
- CONFIG
- RUN TEST
- RESULTS
- MANUAL MODE
- ALARMS

Andor InGaAs Configuration

UPDATE

Integration time (ms): 10

Set Temp (°C): 0

Wavelength: 0

Acquisition Mode: Single Scan

Accumulations: 1

Accumulation Cycle Time (s): 0.00

Kynetic Cycle Time (s): 0.00

Acquisition time (ms): 0.00

Horiba CCD Configuration

UPDATE

Integration time (ms): 10

Multiaquisition

Shutter Mode: Shutter Before First

Accumulations: 1

N of cleans: 1

Horiba Monochr. Configuration

UPDATE

Grating

Target Grating: 1800

Current Grating: 1200

Wavelength: 0

Current Blaze: 0

Units: nm

Description: 1

Front in: Slit

Motor Configuration

UPDATE

Speed: 100

Position: 0

Acc: 500

Dec: 500

Operation Mode: Absolute

Position Force: 100

InPos: 5

End?: 0

Raman Area limits Configuration

UPDATE

X low 0	X high 0	X low 0	X high 0
0	0	0	0
X low 1	X high 1	X low 1	X high 1
0	0	0	0
X low 2	X high 2	X low 2	X high 2
0	0	0	0
X low 3	X high 3	X low 3	X high 3
0	0	0	0

HOME | CALIBRATION | RUN TEST | RESULTS | CONFIG | MANUAL MODE | RAMAN ROI | ALARMS

- EXIT
- HOME
- CALIBRATION
- CONFIG
- RUN TEST
- RESULTS
- MANUAL MODE
- ALARMS

Andor InGaAs Configuration

UPDATE

Integration time (ms)

Set Temp (°C)

Wavelength

Acquisition Mode: Single Scan

Accumulations

Accumulation Cycle Time (s)

Kynetic Cycle Time (s)

Acquisition time (ms)

Horiba CCD Configuration

UPDATE

Integration time (ms)

Multiaquisition

Shutter Mode: Shutter Before First

Accumulations

N of cleans

Horiba Monochr. Configuration

UPDATE

Grating

Target Grating: 1800

Current Grating: 1200

Wavelength: 0

Current Blaze:

Units: nm

Description: 1

Front in: Slit

Motor Configuration

UPDATE

Speed

Position

Acc

Dec

Operation Mode: Absolute

Position Force

InPos

EndP

Raman Area limits Configuration

X low 0	X high 0	X low 0	X high 0
0	0	0	0
X low 1	X high 1	X low 1	X high 1
0	0	0	0
X low 2	X high 2	X low 2	X high 2
0	0	0	0
X low 3	X high 3	X low 3	X high 3
0	0	0	0

Configuration 1 **UPDATE**

Configuration 2 **UPDATE**

HOME | CALIBRATION | RUN TEST | RESULTS | CONFIG | MANUAL MODE | RAMAN ROI | ALARMS

Cont. Spectra

- EXIT
- HOME
- CALIBRATION
- CONFIG
- RUN TEST
- RESULTS
- MANUAL MODE
- ALARMS

Y Pixels: 20 to 240

X Pixels: 1 to 1024

Amplitude: -443 to -680

Image YMax Image XMax

Image YMin Image XMin

UPDATE

Manual Exit

Annex D

Example python code for processing Raman data as received from the LabVIEW software. For 196 measurements, it takes 0.1 seconds to perform all the basic processing, including area calculations of the main peaks. The output plots are also shown below.

```
import matplotlib.pyplot as plt
import spectrapepper as spep
import numpy
import time

start = time.time()

# Load data set.
x, y = spep.load_spectras()

# Plot the raw data
for i in y:
    plt.plot(x, i)
plt.xlim(100, 600)
plt.title('Raw data')
plt.show()

# Remove baseline.
y = spep.bspline(y, x, points=[160, 315, 450, 530])

# Normalize the spectra to the maximum value.
y = spep.normtoratio(y, x, r1=[190, 220], r2=[165, 190])

# Calculate areas.
areas = spep.areacalculator(y, x, limits=[[165, 190], [190, 220], [230, 260], [370, 420],
[420, 470], [470, 530]])

# Plot processed data
for i in y:
    plt.plot(x, i)
plt.xlim(100, 600)
plt.title('Processed data')
plt.show()

print('6 areas calculated for each spectras: ', numpy.shape(areas))

print('Total time (no plots): ', (time.time()-start), ' s.')

>> 6 areas calculated for each spectras: (196, 6)
>> Total time (no plots): 0.09651398658752441 s.
```

