Discovering Process Dynamics for Scalable Perovskite Solar Cell Manufacturing with Explainable AI

Lukas Klein[†] *Sebastian Ziegler*[†] *Felix Laufer* *Charlotte Debus* *Markus Götz* *Klaus Maier-Hein* *Ulrich W. Paetzold*[‡]* *Fabian Isensee*[‡] *Paul F. Jäger*[‡]

† Contributed equally, ‡ Contributed equally

L. Klein, Dr. P. F. Jäger
Interactive Machine Learning Group, German Cancer Research Center, Heidelberg, Germany

L. Klein
Institute for Machine Learning, ETH Zürich, Zürich, Switzerland

L. Klein, S. Ziegler, Prof. K. Maier-Hein, Dr. F. Isensee, Dr. P. F. Jäger
Helmholtz Imaging, German Cancer Research Center, Heidelberg, Germany

S. Ziegler, Prof. K. Maier-Hein, Dr. F. Isensee
Division of Medical Image Computing, German Cancer Research Center, Heidelberg, Germany

F. Laufer, T.T. Prof. U. W. Paetzold
Light Technology Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany
ulrich.paetzold@kit.edu

T.T. Prof. U. W. Paetzold
Institute of Microstructure Technology, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany

Dr. C. Debus, Dr. M. Götz
Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany

Dr. C. Debus, Dr. M. Götz
Helmholtz AI, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany

Keywords: *Perovskite solar cells, Deep learning, Explainable artificial intelligence (XAI), Energy materials science, Knowledge discovery*

Large-area processing of perovskite semiconductor thin-films is complex and evokes unexplained variance in quality, posing a major hurdle for the commercialization of perovskite photovoltaics. Advances in scalable fabrication processes are currently limited to gradual and arbitrary trial-and-error procedures. While the in-situ acquisition of photoluminescence videos has the potential to reveal important variations in the thin-film formation process, the high dimensionality of the data quickly surpasses the limits of human analysis. In response, this study leverages deep learning and explainable artificial intelligence (XAI) to discover relationships between sensor information acquired during the perovskite thin-film formation process and the resulting solar cell performance indicators, while rendering these relationships humanly understandable. We further show how gained insights can be distilled into actionable recommendations for perovskite thin-film processing, advancing towards industrial-scale solar cell manufacturing. Our study demonstrates that XAI methods will play a critical role in accelerating energy materials science.

1 Introduction

Perovskite solar cells (PSCs) have been established as one of the most promising candidates for next-generation photovoltaics. Since the emergence of hybrid organic-inorganic metal halide perovskite semiconductors, power conversion efficiencies (PCEs) of PSCs have improved vastly, exceeding 30% PCE in perovskite/silicon tandem photovoltaics [1]. Despite numerous advantages, [2, 3, 4], the technology has not reached the market yet due to insufficient device stability and the lack of cost-effective and reliable large-scale production [5, 6]. Large-area perovskite thin-films can be deposited using thermal sublimation in vacuum [7, 8, 9] or scalable solution-based deposition techniques like blade coating [10, 11], slot-die coating [12, 13], and ink-jet printing [14, 15]. State-of-the-art solution-processed multi-cation perovskite thin-films require fast solvent extraction to rapidly reach the level of supersaturation and initiate prompt crystallization [16, 17]. The crystallization process heavily affects the perovskite thin-film

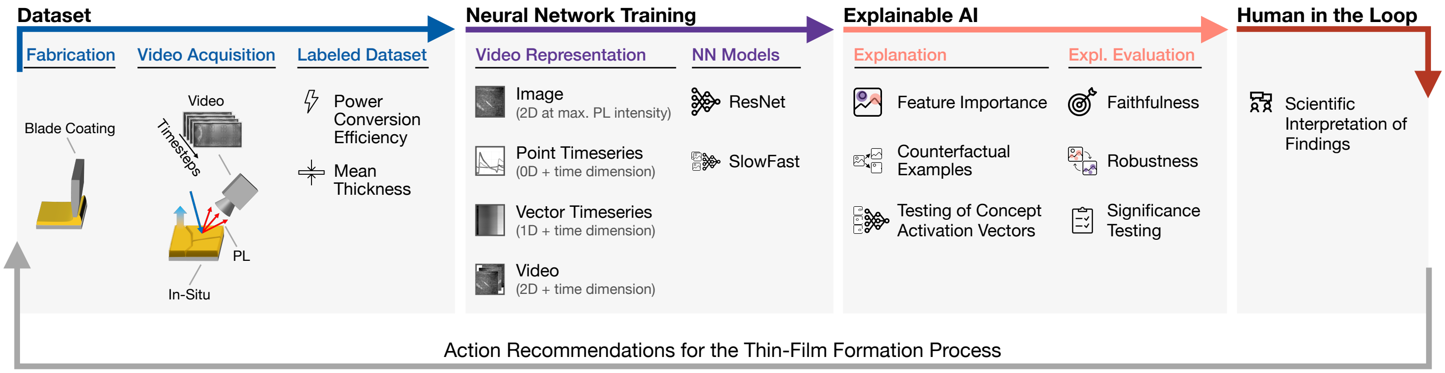


Figure 1: **Overview of the Experiment Setup** Process diagram of the experimental setup. Leads from dataset acquisition to the interpretation of the findings by scientists. *Dataset* abstractly visualizes the acquisition of the videos and labels. *Neural network training* showcases the different representations and neural network architectures used to predict the labels. *Explainable AI* lists XAI methods and XAI evaluation approaches. Finally, quantitative findings are interpreted by scientists to connect them to actionable recommendations, with the possibility of leveraging them for constructing a new enhanced dataset and closing the circle. Abbreviations: NN: neural network, PL: photoluminescence.

formation process and is the key step in producing high-quality perovskite thin-films. In practice, this crystallization process is very difficult to control, as it is heavily dependent not only on the layer stack, deposition, and materials but also on external process parameters such as temperature, as well as lab-specific equipment. Optimal parameters cannot be easily transferred between setups and have to be re-determined for each manufacturing site following a trial-and-error procedure [18, 19, 20]. However, even when nominally identical process parameters are applied, the PSC quality varies due to deviating real-world process parameters resulting from small human or technical inconsistencies infeasible to measure. Consequently, the entire thin-film formation process is hard to optimize for specific setups, leading to poor reproducibility. Hence, a standardized and quantitative way of determining optimal process parameters is lacking to reduce the significant volatility in PSC quality.

Machine learning (ML) has recently been applied to specific optimization problems in various fields, including materials sciences, as it outperforms humans in finding correlations and clues in highly complex data [21, 22, 23]. Specifically, in perovskite research, ML has been used to optimize specific parameters on tabular data, e.g. material choice [24], bandgaps [25], compositional ionic radii [26] or optimizing specific characteristics like the morphology or crystal structure utilizing scanning electron microscope (SEM) [27] and grazing incidence x-ray diffraction (GIXD) images [28]. However, the current application of ML in perovskite research is only working with low-dimensional ex-situ data, looking exclusively at the final thin-film, but not the perovskite formation process itself. We argue, that only by understanding the full process in a data-driven manner we can discover new insights about the underlying mechanisms that lead to volatility in PSC quality.

We address this challenge by introducing a data-driven concept for knowledge discovery. This concept combines deep learning (DL) with multiple explainable artificial intelligence (XAI) methods. While DL is able to find patterns in complex data that would be infeasible to find through traditional analyses, we use XAI to render the found patterns to human-understandable concepts, which can be translated by material scientists into actionable conclusions. To our knowledge, it is the first time that XAI is used to such an extent on high-dimensional data for knowledge discovery as well as PSC fabrication. Based on this setup, we are not only able to find evidence in favor or against existing hypotheses but also uncover unprecedented insights leading to the establishment of new hypotheses regarding reliable large-scale PSC manufacturing. These insights are generated on the basis of a unique high-dimensional dataset, containing in-situ photoluminescence (PL) intensity videos of the perovskite thin-film formation (see our previous study by [29]). While process parameters in this dataset are nominally identical, the video data captures the real-world process parameters by showing the thin-film formation process they produce. By doing so, we do not limit ourselves to prior assumptions regarding high-impact parameters but enable an unbiased inclusion of all possible real-world process parameters. This methodology offers two main advantages: First, we do not limit our prior set of information and allow the identification of unanticipated findings. Second, we can also discover multiple distinct process parameters that cause the same finding, as changes in the thin-film formation process can be achieved by several different actions, which vary in suitability depending on the specific setup.

In the following sections, we first describe the experimental setup leveraging ML and XAI to gain a deeper understanding of thin-film formation processes leading to high-quality PSCs. Based on this setup, we present four key findings derived from the XAI analysis. These findings ultimately allow us to formulate action recommendations on the fabrication process. Concluding, we discuss the limitations and future potential of our approach.

2 A data-driven approach for knowledge discovery facilitates the experimentation process

Dataset This study builds on our publicly available dataset published by [29] that contains in-situ PL video data of 1,129 PSCs (Figure 1). The videos were acquired using PL imaging, which is a non-invasive, easily accessible, versatile measurement technique capable of monitoring the perovskite crystallization in situ on large areas with spatial resolution as well as sub-second temporal resolution. The PL videos were recorded during the vacuum-based quenching of blade-coated perovskite thin-films distributed over 38 substrates using nominally the exact same process conditions. Consequently, the video data depicts the drying and crystallization of the blade-coated perovskite thin-films. Four filters were used to capture the characteristic PL of the underlying processes: a neutral density filter (R_{ND}), measuring the reflectance, two longpass filters, capturing the PL with wavelengths longer than $725nm$ (PL_{LP725}) and $780nm$ (PL_{LP780}), respectively, and a $775nm$ shortpass capturing short-wave PL (PL_{SP775}) combined with a longpass to remove the excitation light [30]. All solar cells were fabricated incorporating a double cation perovskite absorber layer with the composition $\mathrm{Cs_{0.17}\ FA_{0.83}Pb(I_{0.91}Br_{0.09})^3}$. Subsequent to the processing of the perovskite thin-film, the full device stack of the PSC was completed. The PCE of the PSCs as well as the mean thickness (mTh) of the perovskite thin-film serve as labels for our neural network (NN) training, allowing them to learn a relationship between the videos and the quality of a PSC. This ultimately allows the prediction of PSC quality, i.e. PCE, before completing the perovskite thin-film into a functional solar cell device. A detailed description of the data acquisition process can be found in Supplementary subsection A.2.

Figure 2 depicts a characteristic PL signal in a point timeseries (Point TS) data representation, where each line represents the average PL per filter over the whole dataset. Characteristic features of the PL signal can be attributed to different phases during the perovskite thin-film formation, which we extend from [6]: In *Phase I*, the evacuation of the vacuum chamber leads to an accelerated drying of the wet-film due to increased solvent evaporation rates. No PL signal is detected yet as the precursor materials are still dissolved in the ink and no perovskite semiconductor material is formed. With the nucleation onset of perovskite crystallites in *Phase II*, perovskite nuclei and small grains start to emit a strong PL signal. During crystallization (*Phase III*) larger grains are formed by coalescing and ripening of smaller ones. Non-radiative recombination at grain boundaries and a reduced outcoupling of luminescence photons emitted from the solid perovskite thin-film - due to total internal reflection - reduce the overall emitted PL signal. *Phase IV* starts with the venting of the vacuum chamber creating the final film surface morphology. The evolution of the PL signal during that phase remains not fully understood but is hypothesized to correlate with the thin-film's final morphology, i.e., surface roughness [31].

Neural Network Training and **Explainable AI Methods** The DL models employed in this work are trained on different representations of the high-dimensional data as shown in Figure 1: the original video, image, point timeseries (Point TS), and vector timeseries (Vector TS). Detailed descriptions of each representation and their respective data preprocessing are available in section 8. The DL models, specialized for each representation, are trained to predict the labels, i.e. PCE or mTh. Model architectures for each representation are described in section 8 (Neural Network Architectures) and chosen hy-

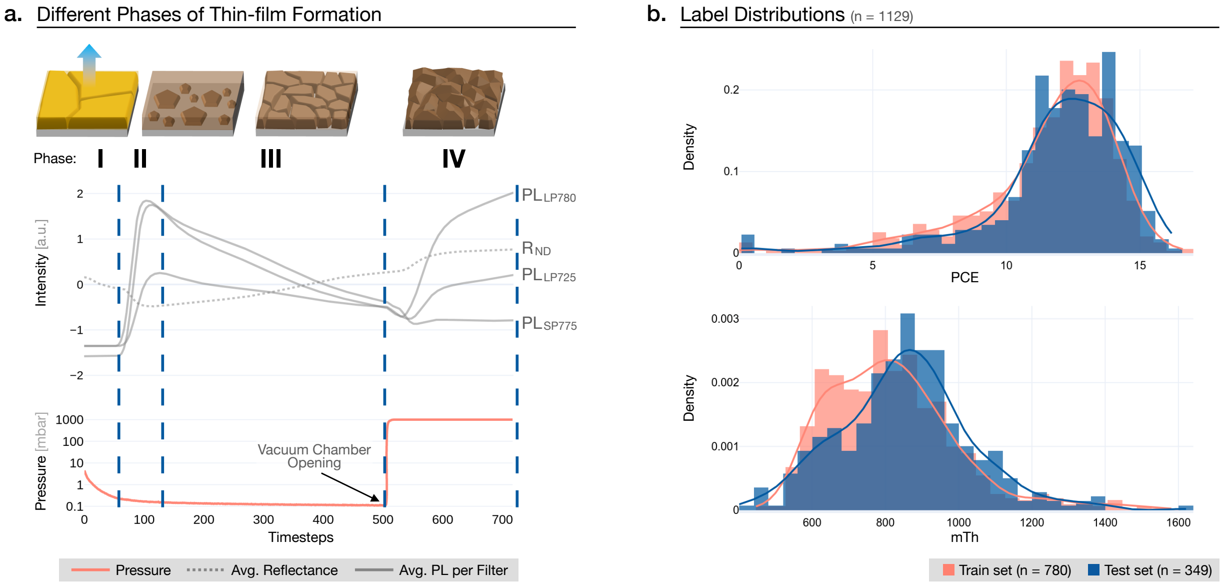


Figure 2: **Description of utilized data** **a.** The figure shows the four different thin-film formation phases based on the average Point TS PL and reflectance intensity for each of the four filters. Below, the simultaneous change of the air pressure in mbar is depicted. **b.** Distribution of both labels, separated into train and testset. Abbreviations: Point TS: point timeseries, PL: photoluminescence.

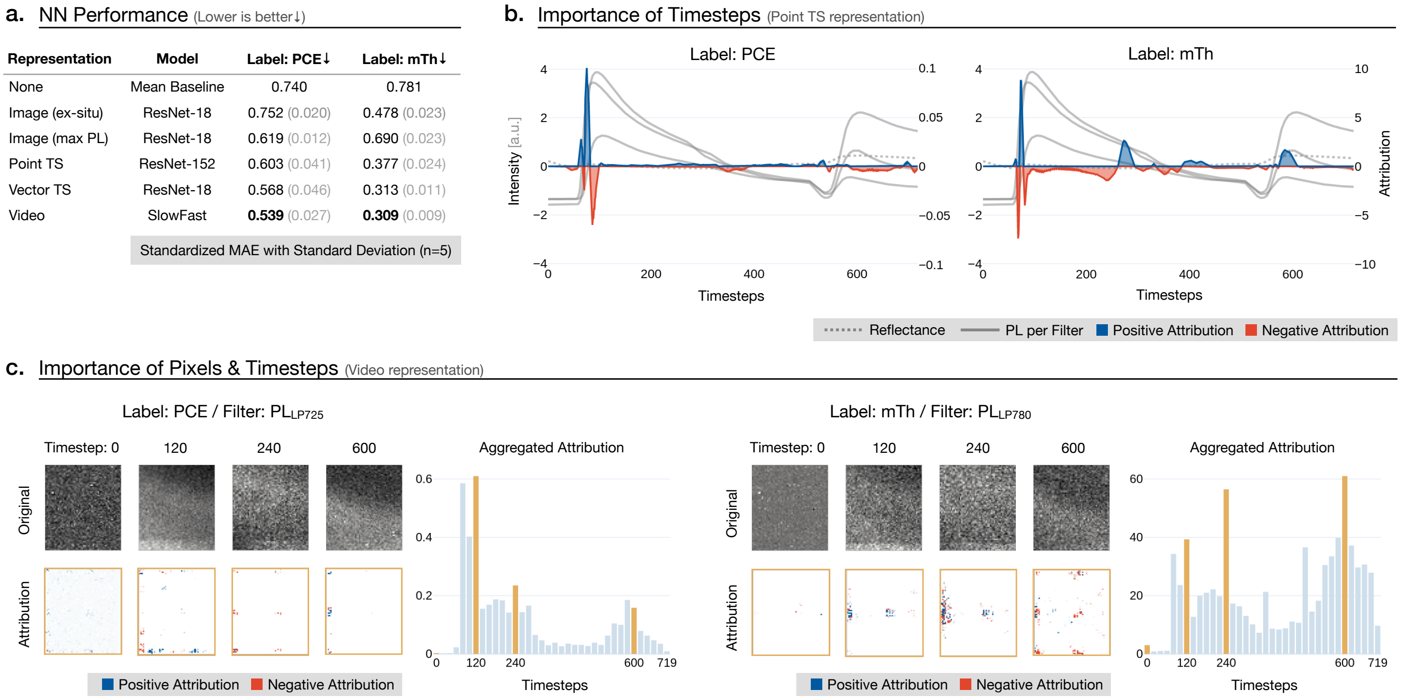
perparameters and augmentation techniques can be found in Supplementary E. Building on these models, multiple XAI methods are used to render the ML mapping between PL input data and predicted PCE or mTh humanly interpretable. To understand which input features and phases are most important to our models, we apply attribution methods [32, 33] to compute either local explanations, i.e. explaining a model's behavior on a single observation (i.e. single sample), or global explanations, i.e. explaining patterns that are present in general. However, the computed attribution maps only indicate the importance of individual features without revealing the underlying reasons and causal effects leading to the importance. To answer this question we deploy counterfactual explanations (CEs) [34, 35, 36] and the Testing of Concept Activation Vectors (TCAV) [37]. CEs alter the input observation to receive a specific counterfactual outcome and simulate "what if" -scenarios. For TCAV, on the other hand, building on the CE analysis we define concepts that occur in the data, e.g. a high PL peak, and test the importance of each concept to specific layers of the DL model (detailed description is provided in methods section 8).

In the following sections, we present our key findings derived from the multi-dimensional in-situ PL dataset by XAI. Since all the different data representations, labels, and XAI methods yield various combinations, we show only the most relevant results. A comprehensive overview of the DL and XAI results obtained during this study can be found in Supplementary B.

3 Temporal progression of in-situ photoluminescence contains more information compared to the spatial dimension

While two-dimensional data (e.g. images) and correlations therein can be captured and processed reasonably well by human experts, correlations in three-dimensional data (e.g. videos) are hardly accessible. In fact, our XAI analysis shows that the temporal progression of the PL data contained in in-situ videos recorded during the perovskite thin-film formation, i.e. the vacuum-based quenching step, contains much more information about the device performance and perovskite thin-film thickness compared to single ex-situ PL images.

DL models trained on representations containing time information outperform DL models trained on spatial information alone (Figure 3 (a.)). When limiting the data to only one frame (image representa-



Representation	Model	Label: PCE↓	Label: mTh↓
None	Mean Baseline	0.740	0.781
Image (ex-situ)	ResNet-18	0.752 (0.020)	0.478 (0.023)
Image (max PL)	ResNet-18	0.619 (0.012)	0.690 (0.023)
Point TS	ResNet-152	0.603 (0.041)	0.377 (0.024)
Vector TS	ResNet-18	0.568 (0.046)	0.313 (0.011)
Video	SlowFast	**0.539** (0.027)	**0.309** (0.009)

Figure 3: **NN model performance and XAI explanations about the temporal progression** **a.** NN performance is measured in standardized mean absolute error (sMAE) to compare between labels. The mean baseline is computed by calculating the label's mean on the training set and using it as prediction for every case in the testset. **b.** Attribution-map showing which timesteps (over all four filters) of the Point TS attribute either positive (blue) or negative (red) to the prediction of each label. The scale of the attribution differs between both labels, as it depends on the scale of the labels. **c.** Attribution-map for the video data of label PCE and filter PL_{LP725} (left), and label mTh and filter PL_{LP780} (right). Both graphics show four frames and their attribution-maps, selected based on the aggregated absolute attribution per timestep to their right. Abbreviations: NN: neural network, TS: timeseries, PL: photoluminescence, PCE: power conversion efficiency, mTh: mean thickness, MAE: mean absolute error.

tion), thereby neglecting the temporal dimension, choosing the timestep influences the prediction performance differently for each label. While for PCE using the frame at maximum PL intensity (in-situ) yields better performance than the final frame (ex-situ) of the thin-film formation, it is vice versa for mTh. This suggests that mTh is more dependent on process phases after the maximum PL intensity frame, while this timestep contains substantial information for PCE prediction. The significant variation in PCE and mTh prediction performance results from the fact that the PCE label, other than the mTh label, encompasses effects and correlations of the subsequent layer processing on top of the perovskite thin-film, given that the PCE is determined for the full PSC. Parity plots, showing the correlation between predicted and ground truth labels, can be found in Supplementary Figure S14. Attribution-maps showing how important each timestep for the DL model is (Figure 3 (b.)) reveal that the models neither identify single dominant timesteps nor consider each timestep as equally relevant, but rather highlights distinct time periods of high relevance for the models trained on either the PCE label or the mTh label. Importantly, these periods are also reflected in the video representation when aggregating attribution per frame, while the spatial attribution within frames does not show recognizable patterns (Figure 3 (c.)).

Our analysis highlights that the model focuses on time periods that coincide with the defined phases. The model leverages the information in these phases to successfully differentiate between sequences resulting in high- or low-performing PSCs. The predictive performance of the models extensively increases when including temporal information, and for in-situ data is always substantially better than the mean baseline. This emphasizes the successful learning of non-trivial relationships, which represents a requirement for the subsequent XAI analysis. The findings suggest that the temporal dimension provides crucial characteristics for understanding the perovskite thin-film formation, such as the timing of the different phases during thin-film formation which is not present in individual images. Thus, not only the acquisition of in-situ data compared to ex-situ but also the inclusion of a temporal dimension is vitally important for PSC process monitoring and optimization.

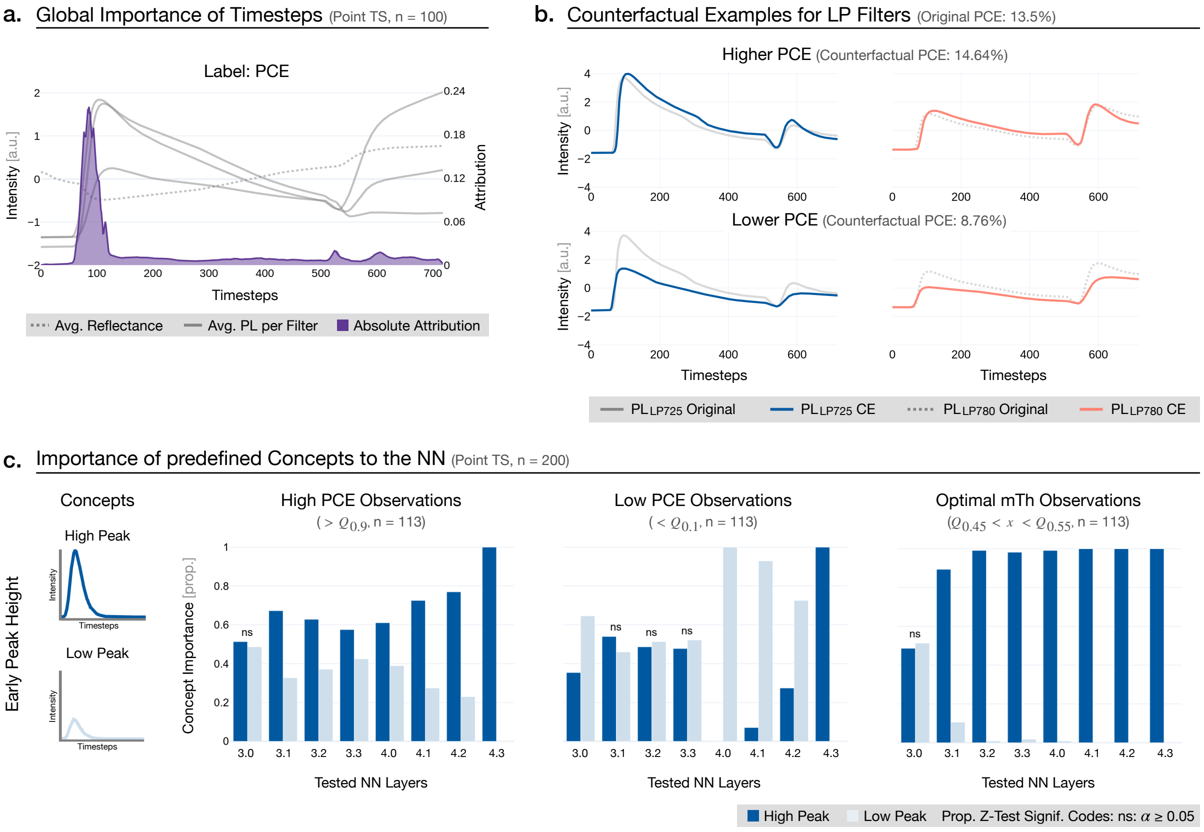


Figure 4: **Diverse XAI results highlight the importance of *Phase II* (nucleation onset)** **a.** Absolute attribution-map for PCE averaged over 100 observations. **b.** Generated CEs of the LP filters yielding either high or low PCE prediction. See Supplementary Figure S8 for the other two filters and Supplementary Figure S9 and S10 for the CEs of the same observation's other representations. The original predicted PCE for this observation is shown in the header and the PCE predicted based on the artificially computed CEs are presented behind the label. **c.** For TCAV, we test the last eight layers of the NN, as they capture more semantic information than earlier layers. For each of the eight layers, we observe whether the layer is more sensitive to the concepts "High Peak" or "Low Peak" of the concept class "Early Peak Height", or whether there is no significant difference (ns, based on proportion z-test with a significance level (α) of 0.05). This is done based on selecting contrasting data subsets, to see the difference in importance of the concepts to e.g. high and low PCE observations (see Supplementary B for high mTh observations). Abbreviations: NN: neural network, TS: timeseries, PL: photoluminescence, PCE: power conversion efficiency, mTh: mean thickness, TCAV: Testing of Concept Activation Vectors.

4 High photoluminescence peak intensity at nucleation onset induces higher quality perovskite films

We show and quantify that the quality of blade-coated perovskite thin-films strongly correlates with the PL intensity close to the onset of the nucleation and crystallization phase. This onset is apparent at the start of *Phase II*. When visualizing the global importance over 100 observations, i.e. PL data recorded from 100 PSCs, we observe that models predicting PCE (Figure 4 (a.)) and mTh (Figure 6 (a.)) both show the highest absolute attribution to *Phase II*. Figure 5 indicates the importance of each of the four filters by their mean absolute attribution $\overline{|A_F|}$, as they contribute in different extends to the final prediction.

To substantiate our analysis, we artificially generate CEs (detailed description is provided in methods section 8) of the PL intensity curves (Figure 4 (b.), see Supplementary B for the other two filters), such that the model predicts substantially higher or lower PCE compared to the original observation. These CEs reveal that when moderately increasing the nucleation onset peak the model predicts higher PCE values and vice versa (Figure 4 (b.)), thereby confirming our initial observation. These results are further reaffirmed for mTh, as a decreased PL intensity of the nucleation onset results in higher mTh prediction (Figure 6 (b.)). To predict a lower mTh, however, no substantial change in the PL intensity is required, suggesting that lower measured mTh values in our dataset still fall into an optimal range, and only for higher values the PL intensity course is substantially different.

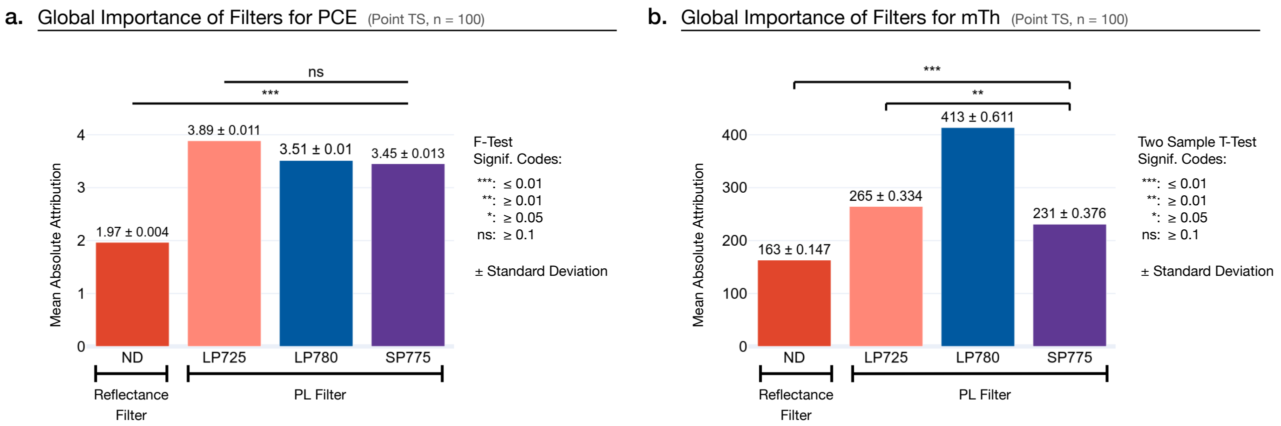


Figure 5: **The importance of each filter differs between labels** **a.** Both figures show the mean absolute attribution of each filter $(|\overline{A_F}|)$ over $n = 100$ observations with standard deviation, in order to assess the importance of each filter. Higher is always more important. F-Test to determine the difference between all filters and only between the PL filters. The difference between all three PL filters is not significant $(\alpha \geq 0.1)$ **b.** Two sample T-Test to determine the difference between ND and SP775 filter and LP725 and SP775 filter. Both are significantly different from each other.

To further reinforce the findings of the CE analysis, we deploy TCAV to determine the extent to which the concepts of high and low nucleation onset peaks affect the model prediction. We split the whole dataset via quantiles (Q_x) into two subsets for both labels, to not only observe the general importance of the concepts to the model, but specifically when predicting observation subsets with properties we are interested in: high PCE $(> Q_{0.9})$ and low PCE $(< Q_{0.1})$ observations, and optimal $(Q_{0.45} < x < Q_{0.55})$ and high $(> Q_{0.9})$ mTh observations $(\forall Q_x : n = 113)$. We do not use low mTh observations, as the data shows the highest, thus optimal, PCE around $800nm$ (Supplementary A), and the CE analysis revealed that lower mTh values do not necessarily result from substantially different PL intensity curves. Figure 4 (c.) shows that when predicting high PCE observations the concept of “High Peak” is more important to the model whereas when predicting low PCE observations the concept of “Low Peak” is more important. Equivalently, in the case of mTh, the concept of “High Peaks” is more important than “Low Peaks” for the optimal and high mTh subset (only optimal is shown in Figure 4 (c.), see Supplementary B for high mTh observations). Both TCAV findings reconfirm the CE-based conclusions.

In summary, our data-driven approach shows that a higher peak in *Phase II* leads to improved PSC quality. Consequently, practical guidelines for future experimental work are derived from XAI analysis: Process parameters and ink formulations shall be optimized toward maximizing the PL peak height during *Phase II*. This data-driven finding complements experimental trial-and-error analysis in literature, where it was shown that changes in the rate of evacuating the vacuum chamber impact not only the PL onset time and the PL peak height but also the perovskite thin-film quality [38]. The actionable recommendation for future processes is to increase the evacuation rate to achieve higher PL peaks, which is indicative of higher solar cell performance.

5 High photoluminescence peak intensity at the start of the chamber venting induces thick and rough perovskite thin-films

To fabricate high-quality perovskite thin-films, a homogenous layer morphology is critical. Our study shows that increased thin-film thickness and roughness (the latter is highly correlated with the thickness measurement, see Supplementary A) can be inferred from the XAI analysis of in-situ PL data. The vacuum quenching of the perovskite material and the subsequent venting (starting at around $t = 505$) strongly affect the crystallization and the morphology of the perovskite layer [38]. Indeed, our DL models predicting mTh show besides the high absolute attribution to *Phase II* also attribution to *Phase IV* (Figure 6 (a.)). Specifically, there is first a small attribution peak at around $t = 510$, before the dip in PL intensity, and then a large attribution concentration after the dip. For PCE observations, only a

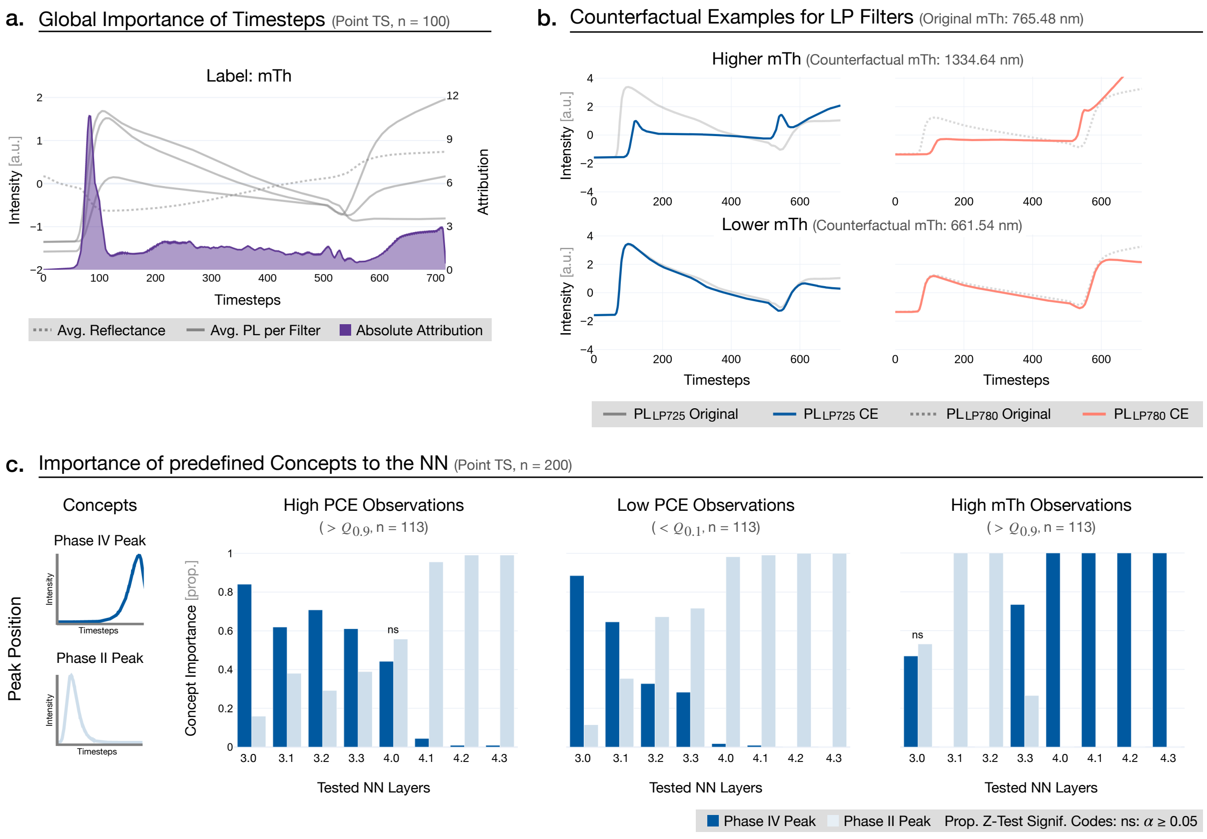


Figure 6: **Diverse XAI results highlight the importance of *Phase IV* (surface morphology)** **a.** Absolute attribution-map for mTh averaged over 100 observations. **b.** Generated CEs of the LP filters yielding either high or low mTh prediction. See Supplementary Figure S8 for the other two filters and Supplementary Figure S9 and S10 for the CEs of the same observation's other representations. The original predicted mTh for this observation is shown in the header and the mTh predicted based on the artificially computed CEs is presented behind the label. **c.** For TCAV, we test the last eight layers of the NN, as they capture more semantic information than earlier layers. For each of the eight layers, we observe whether the layer is more sensitive to the concepts "Phase IV Peak" or "Phase II Peak" of the concept class "Peak Position", or whether there is no significant difference (ns, based on proportion z-test with a significance level (α) of 0.05). This is done based on selecting contrasting data subsets, to see the difference in importance of the concepts to e.g. high and low PCE observations (see Supplementary B for optimal mTh observations). Abbreviations: NN: neural network, TS: timeseries, PL: photoluminescence, PCE: power conversion efficiency, mTh: mean thickness, TCAV: Testing of Concept Activation Vectors.

smaller attribution spike at $t = 510$ can be observed (Figure 4 (a.)). The CEs for mTh observations in Figure 6 (b.) (see Supplementary B for the other two filters) reveal that along with a low PL peak in *Phase II*, a high PL intensity during *Phase IV* leads to higher mTh. Alterations leading to lower mTh are only very minor.

After determining the importance of *Phase IV*, we compare the two concepts "Phase II Peak" and "Phase IV Peak" to further distinguish between the two most important time periods to the NN. While both concepts are equally important for high PCE observations, "Early Peak" is more important for low PCE observations (Figure 6 (c.)). The results refine the conclusion that especially for low PCE values, *Phase II* is more important than *Phase IV*. Also for mTh observations, both concepts are generally important (Supplementary B), with "Late Peak" being moderately more important than "Early Peak", confirming the importance of *Phase IV* previously observed in the CE experiments.

In summary, our analysis reveals that the perovskite thin-film roughness correlates to the timing of the venting step. We conclude further that residual solvent contained in the thin-film leads to increased surface roughness, resulting in increased PL outcoupling, i.e. high PL signal during venting. In contrast, perfectly dry perovskite thin-films exhibit no change in morphology, i.e. no significant change in PL, during venting. The actionable recommendation derived from our XAI analysis is to optimize the processing such that the PL does not increase after the venting, i.e. to prevent the formation of rough and therefore thick layers, which are more likely to result in bad-performing solar cells. This can be achieved by extending the evacuation times which dries the thin-film and eliminates the PL increase during venting.

6 Superior crystal growth is reflected in a steeper photoluminescence intensity decay during the crystallization phase

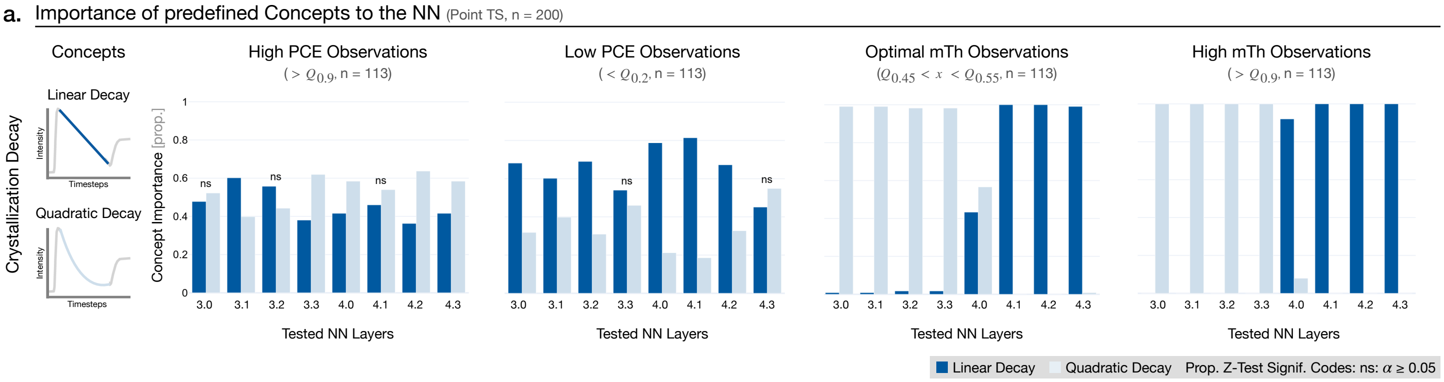


Figure 7: **TCAV concept importance to analyze the decay during *Phase III* (crystallization)** For each of the eight layers, we observe whether the layer is more sensitive to the concepts "Linear Decay" or "Quadratic Decay" of the concept class "Crystallization Decay", or whether there is no significant difference (ns, based on proportion z-test with a significance level (α) of 0.05). We select contrasting data subsets to see the difference in importance of the concepts to e.g. optimal and high mTh observations. Abbreviations: NN: neural network, PCE: power conversion efficiency, mTh: mean thickness, TCAV: Testing of Concept Activation Vectors.

Next to nucleation, the phase of crystallization and crystal growth (*Phase III*) is of critical importance for the morphology of the perovskite thin-films. By means of our XAI analysis, we find that high-performing PSCs correlated with a steeper decrease in PL intensity during *Phase III* when compared to low-performing PSCs. When revisiting the CE analysis in Figure 4 (b.) and Figure 6 (b.) we observe that the PL intensity slope apparent in *Phase III* is steeper when predicting a higher PCE or a lower mTh. This is also reaffirmed by the cluster analysis of [29] showing that clusters having a higher mean PCE also exhibit a steeper slope in *Phase III*. To understand the underlying effect behind this difference in decay slope, we use TCAV to test the importance of the two concepts "Linear Decay" and "Quadratic Decay" to the model (Figure 7). We find that the concept "Quadratic Decay" is more important for observations with high PCE, while the concept "Linear Decay" is more important for observations with low PCE. It is possible that this correlation may be spurious, as a high nucleation onset peak in *Phase II* in PCE could result in a more quadratic crystallization decay in *Phase III*, while a lower peak results in linear decay. Therefore, future work needs to verify the causal effect behind the change in decay. In the case of mTh, the model is sensitive to both concepts and no unique characteristic for optimal or high observations can be defined.

In summary, the data-driven analysis reveals that a fast superlinear decay of the PL signal correlates with higher performance. In addition, the correlation suggests that the crystallites grow and coalesce into larger crystals which reduces the number of grain boundaries and promotes the extraction of charge carriers. Grain boundaries exhibit a high defect density which reduces radiative recombination and therefore lead to a decrease in emitted PL. Thus, a high importance of the concept "Quadratic Decay" for observations with high PCE could possibly be caused by the higher charge extraction of high-performing PSCs leading to a stronger reduction of radiative recombination during the crystal growth phase illustrated in the steeper decrease in PL. For low-performing solar cells, charge carrier extraction is lower, resulting in a high rate of radiative recombination and therefore in a flatter decrease of emitted PL signal over time.

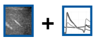
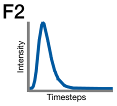
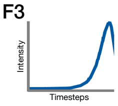
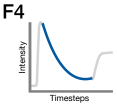
	Finding revealed by XAI	Scientific Interpretation	Actionable Recommendation
F1 	**Temporal information** increases predictive performance of the models extensively	Temporal dimension provides crucial characteristics for understanding the perovskite thin-film formation	Acquisition of temporal in-situ data is vitally important for process monitoring and optimization
F2 	A **higher peak in Phase II** leads to improved PSC quality	Changes in evacuation rate improve perovskite thin-film quality and impact PL onset time and PL peak height	Increase the evacuation rate to achieve higher PL peaks, which is indicative of higher PCEs
F3 	Perovskite thin-film roughness correlates to the **timing of the venting step**	Increased PL signal during venting caused by increased surface roughness due to residual solvent contained in the thin-film	Extending the vacuum quenching time eliminates the PL increase during venting, by preventing the formation of rough and therefore thick layers
F4 	Fast **superlinear decay** of the PL signal correlates with higher performance	Higher charge extraction of high-performing PSCs leads to a stronger reduction of radiative recombination during crystal growth phase	Improve charge extraction resulting in a faster decrease of emitted PL signal

Table 1: Overview of all findings and the recommendations derived from them.

7 Conclusions

Our analysis shows that fluctuation in the quality of PSCs processed with nominally identical conditions can be understood by investigating the thin-film formation process with DL and XAI, bringing us one step closer to the industrial usage of PSCs. We are able to infer actionable recommendations just by analyzing the video dataset and without having to carry out extensive and costly trial-and-error experiments. This is possible due to our unique approach of leveraging diverse XAI methods, going beyond mere feature importance, combined with deep learning-based modeling of video data, to generate new insights that would not have been identified by human experts.

While this data-driven approach can vastly accelerate and facilitate experimentation in materials science, some possible limitations need to be taken into account. In DL, overfitting is a common problem, where no meaningful relationships are learned. Therefore, general conclusions can only be drawn if sufficient prediction performance on unseen test data is achieved. Our quantitative testset evaluation (Figure 3) and parity plots (Figure S14) show that our models are well capable of predicting unseen data to find general patterns. Differences between the two labels are due to the fact that PCE can only be measured after the completion of the solar cell while mTh only depends on the perovskite layer. Since the subsequent production steps can introduce irregularities that adversely affect PCE but are impossible to predict from only the videos it is expected that mTh is predicted overall more accurately. This is also reflected in the parity plots (Figure S14) showing PCE predictions for low-PCE solar cells being consistently overestimated, due to the error leading to a low PCE only occurring in not imaged subsequent steps. Apart from these cases however, a high-quality prediction performance for subsequent XAI steps is achieved. Further, human interaction with XAI is prone to confirmation biases and overinterpretation. To mitigate these potential pitfalls, we not only apply diverse XAI methods that confirm observations from different perspectives but also perform a large-scale quantitative evaluation (Supplementary C). Furthermore, a data-driven approach is naturally limited by the dataset used for analysis. A higher spatial resolution which makes prominent defects and crystal structures better visible (e.g. SEM) would widen the range of potential insights but is infeasible to obtain for in-situ videos. Naturally, there is a possibility of unobserved parameters, not captured in our dataset, but still affecting the labels. However, we deem the possibility of important unobserved parameters and confounders as rather low, since the information-rich video data captures almost all important processes of the thin-film formation, with the exception of the succeeding production steps to finalize the PSC when measuring PCE. Further, the inference from XAI results to underlying causal variables is performed by human experts, so as to control against potential confounders. Lastly, we want to stress that the techniques applied in this manuscript should also be regarded as a general concept for experimental material researchers to assess and enhance their experimental setups. The XAI methods are not limited to the dataset we have used as an example. Consequently, a similar analysis can be conducted to interpret and improve other fabrication processes in materials science.

Our encouraging action recommendations exemplify the usage of XAI methods in materials science and PSC research and showcase data-driven approaches as key tools for the development of upcoming photovoltaic technologies. Moreover, this work highlights the importance of investigating PSCs in a scalable experimental setup to tackle current reliability issues in large-scale PSC production. To this end, it is important to understand the perovskite thin-film formation process which is investigated in this work.

8 Methods

Data Preprocessing We create several data representations in– or excluding time and/or spatial dimensions. The original dataset consists of width (w), height (h), and time (t). The **video** representation contains all available information $\{w, h, t\}$. They contain 719 timesteps (i.e. frames) acquired at a rate of 3 frames per second. Each frame is an image sized 65 × 56 pixels with a spatial resolution of 18.6 $pixels/mm$ and covers the active area of a PSC. The point timeseries (**Point TS**) contains only the temporal information by aggregating each frame via its mean $\{\overline{(w, h)}|t\}$. The **image** representation on the other hand only covers spatial information. It consists of a single frame at a given time point t, for example, the one with the maximum PL signal $\{(w, h)|t_{maxPL}\}$ or the last frame of each video $\{(w, h)|t_{|T|}\}$, simulating an ex-situ approach. The vector timeseries (**Vector TS**) is similar to the Point TS, aggregating only the w dimension $\{\overline{(w|h)}, t\}$ resulting in a 2D representation that includes temporal as well as some spatial information. Additionally, we tried spectral analysis by converting the Point TS to a spectrogram. However, there were not many prominent frequencies in the data, resulting in lower prediction performance. Each of the representations contains the four filters. They are concatenated along the channel dimension. Further, the data was standardized using the z-transformation with the mean and standard deviation of the training set. Additionally, each model is trained with different data augmentations such as flips and blurs. A detailed list of all augmentations for each model can be found in Supplementary E. For the testset, the data was only standardized using mean and standard deviation again from the training set and no augmentations were applied.

Neural Network Training and Testing We use the same train and test split as [29], excluding 30% of the 1,129 PSCs as a held-out test set stratified on a substrate level. For model development, we apply 5-fold cross-validation to the training set. The average score of the 5 different models is then used to determine the best configurations. This ensures a reliable model evaluation mitigating potential overfitting on only one validation set. The final model evaluation as seen in Figure 3 (a.) is done on the test set. sMAE is used as a metric: It standardizes the common MAE by dividing it by the standard deviation of the respective ground truth label to render scores comparable across different labels and value ranges. The complete results table for the test set including also unnormalized scores is available in the Supplementary D.

Neural Network Architectures Since we use different data representations with varying dimensionality we need to adapt the neural network architecture to the representation. Different architectures such as VGGs [39], ResNets [40], PreActResNets [41], EfficientNets [42], and Wide ResNets [43] were compared against each other during development and the final architecture for each data representation and label and final models were selected based on the highest cross-validated performance as measured by sMAE. Overall, the ResNet architecture [40] had the highest performance for the 1D and 2D representations. For the Point TS which only has one dimension (time), we apply a ResNet-152 using 1D convolutions. For the image and Vector TS representations, we use a ResNet-18 with 2D convolutions. The four filters from our dataset are presented to the networks as input channels, analogous to how red, green, and blue are used when processing natural images.

Learning from entire video sequences is substantially more challenging. Using a standard ResNet analogous to the 1D and 2D representations would result in an architecture that needs to accumulate information across 4 × 65 × 56 × 719 input tensors. Computation time aside, the curse of dimensionality coupled with a limited number of training cases would impede optimization, particularly generalization performance. Thus, a purpose-built solution is required. SlowFast [44] is highly optimized for video data and yields state-of-the-art performance on common video data benchmarks. It consists of two convolutional branches that require different inputs. Its central concept is that a video contains static (slow) information, e.g. objects that are present in the video at all times or only change very slowly, and dynamic (fast) information, e.g. movements or other actions. Accordingly, the two branches of SlowFast focus on the two distinct types of information. The slow pathway processes fewer frames by using a large temporal stride, allowing the network to detect static information efficiently. On the other hand, the fast pathway inputs more video frames but uses a more lightweight sub-architecture by using fewer convolutional kernels. In the end, both branches are concatenated and followed by a classification or regression head.

Neural Network Hyperparameters We use the mean absolute error (MAE) as a loss function to train our regression neural networks. All models are trained for 1,000 epochs, either using the AdamW [45] or the Madgrad [46] optimizer and a cosine annealing learning rate scheduler [47]. A table with detailed information on all hyperparameters for each model is available in Supplementary E.

Depending on the representation we make extensive use of data augmentation. This allows to slightly change the input data every time the model sees it during training, ensuring more variability and thus, better generalization. While we only z-standardize the data for the Point TS representation, we additionally use flips, blurs, and spatial transformations for the other representations. A detailed list of augmentations used for each model is available in Supplementary E.

Attribution Methods Due to the risk of confirmation bias and unfaithful explanations [33], we compute each attribution-map for all representations and labels with four different attribution methods. These include Guided Backpropagation (GBP) [48], Guided Gradient-weighted Class Activation Mapping (GGC) [49], Integrated Gradients (IG) [32], and Expected Gradients (EG) [50]. All results shown in the main paper are based on EG due to the explanation evaluation results (see Supplementary C), as it shows a good balance between robustness and faithfulness without the prior selection of a baseline value (possibly biasing the explanations). The explanations generated by all other methods are in Supplementary C. Local explanations are computed on test set observations. As there are no significant differences between train and test set explanations, global explanations are computed on the full dataset to leverage the substantially larger size compared to the test set.

The most apparent solution to measure the sensitivity of a model's output to its input is the respective gradient. However, vanilla gradients are prone to gradient shattering [51] and ignoring global effects in the input space. Thus, they can e.g. be combined with deconvolutional networks [52] which aim to invert the data flow of a NN, to reconstruct the discriminative input space of an activation or output node. While both approaches are almost equivalent [53], they differ in their backwards pass because, for non-linear functions such as the Rectified Linear Unit (ReLU), deconvolutions compute “switches” during the forward pass to invert the function. In the case of ReLU for example, this results in a sign indicator function computed on the higher-layer's reconstruction instead of the layer input, which would be the case in backpropagation (for more detailed information see Section 3.4 in [48]). GBP combines both backwards pass approaches by masking out the values for which at least one of the approaches is negative, guiding the gradient by an additional signal from the higher layers on top of the usual backpropagation.

We combine GBP with GradCAM, a method leveraging the idea that convolutional neural networks transform spatial to semantic information by attributing to the semantic information, which is then back-projected into the input space. The resulting GGC takes the element-wise product between GBP and the non-negative GradCAM attributions, leveraging both the semantic information from GradCAM and the more fine-grained spatial information in the input space from GBP. We back-project from the last block in the ResNet and the multipathway fusion block in the SlowFast architecture.

IG on the other hand computes a path integral between a baseline value x_0 and the true value x_j of each of the j input features (i.e. pixels or timesteps).

$$\mathrm{IG}_j(x, x_0) = (x_j - x_{0j}) \int_{\alpha=0}^{1} \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x_j} d\alpha \tag{1}$$

However, the prior selection of a baseline value in IG is not always clear, and performing multiple path integrals over several baseline values can be inefficient. Thus, EG avoids the selection of a baseline value, by leveraging a probabilistic baseline D computed over a sample of observations.

$$\mathrm{EG}_j(x) = \underset{x_0 \sim D,\ \alpha \sim U(0,1)}{\mathbb{E}} \left[\frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x_j}\ d\alpha \right] \tag{2}$$

In application, this expectation is approximated via a mini-batch sampling approach for x_0 and α.

Counterfactual Examples To generate CEs, we use the Genetic Counterfactuals (GeCo) algorithm [54] together with the respective models in Figure 3 (a.). GeCo computes plausible (assuring that they could be real) and feasible (assuring they can actually be computed) CEs in a short time. It relies on a genetic algorithm, which is customized to favor searching CEs with the smallest number of changes. To achieve the short computation time, it utilizes novel optimizations such as the Δ-representation of candidate counterfactuals and only partial evaluation of the classifier. This speed in computation is especially important for our task, as it would not be feasible to compute CEs for high-dimensional data such as videos or the Vector TS representation with other methods such as Diverse Counterfactuals (DiCE) [55] or Diffusion Visual Counterfactual Explanations (DVCEs) [56], even with very high computing resources.

As our labels are continuous, we leverage the CEs to visualize how an observation has to be changed to receive either a substantially higher or lower PCE ($> 13.93\%$ and $< 9.22\%$) or mTh ($> 1300nm$ and $< 700nm$) prediction compared to the ground truth value.

TCAV We leverage TCAV [37] to identify concepts that are most important to the model's predictions. The technique uses a Concept Activation Vector (CAV), v, to quantify the importance of a particular concept to the model's predictions. A CAV is a high-dimensional vector that is learned by training a linear model on the activations of a hidden layer l and two datasets of examples, $C = [c_1, c_2]$, that are representative of the concepts. The CAV is then the unit length normal vector to the linear decision boundary of the model, pointing in the direction of c_1, while c_2 lies in the opposite direction. We then calculate the sensitivity $S_{C,l}$ of the output into the direction of the CAV by taking the directional derivative:

$$S_{C,l}(c_1) = \nabla h_l(f_l(c_1)) \cdot v_C^l \tag{3}$$

With $f()$ being the part of the model up to the hidden layer l and $h()$ the part of the model from the hidden layer to the output. We use a sign-test to test if the output for a specific observation is more sensitive to concepts one or two. If the directional derivative in the direction of the CAV is positive it is more sensitive to c_1 and if negative more sensitive to c_2. We compute the concept importance score by averaging the sign-test result for the respective high/low PCE or mTh subsets X_q.

$$\mathrm{CoIm_{C,l,q}} = \frac{|\mathrm{x} \in \mathrm{X_q} : \mathrm{S_{C,l}(x)} > 0|}{|\mathrm{X_k}|} \tag{4}$$

We sample the datasets C for all of our six concepts (*Phase IV* and *Phase II* Peak, High and Low Peak, Linear and Quadratic Decay) separately for each filter based on extracted summary statistics of each Point TS and specific random permutations to not create out-of-distribution (OOD) examples (Supplementary F for a more detailed description). Each concept is sampled 100 times. To ensure robust and trustworthy CAVs, we also evaluate the linear classifier which is trained to separate both concepts in the layer-activation output space. For PCE observations (high and low PCE, as the CAV is independent of the observations, and only dependents on the DL model and concepts), we train a Lasso-regression [57] which reaches a test set (33% split, $n_{test} = 66$) accuracy of 99% for the *Phase IV* and *Phase II* Peak concept separation, 95% for the High and Low Peak and 61% for the Linear and Quadratic Decay. In the case of mTh, the test set accuracies are 100%, 89%, and 62% respectively. See Supplementary F for the hyperparameters of the linear classifier.

Faithfulness Metrics We apply Sensitivity-N [58], Insertion and Deletion [59] to evaluate the faithfulness of the explanations to the models. The distribution of each evaluation score is approximated from 500 observations for the Point TS, 250 for the image and Vector TS, and 50 for the video representation. For all metrics, the score for a total unfaithful (random) attribution-map is zero.

Sensitivity-N is a metric that is satisfied when the sum of the attributions A for any subset of n features is equal to the variation of the output $f()$ caused by removing the features in the subset. x_s is the set of all subsets of features from cardinality 1 to n.

$$\sum_{j=1}^{n} A_j(x) = f(x) - f(x_{[x_s=0]}) \quad \forall x_s = [x_1, \cdots, x_n] \subseteq x \tag{5}$$

We measure how much the sum of the attributions (left-hand side) and the variation in the output (right side) correlate when calculating each side for all subsets in x_s. We compute the Pearson correlation between both sides for n is equal to the values of 1, 3, 12, 41, 144, and 501. However, it is not efficient for the larger values of n to compute the correlation for all possible x_s. To approximate this, we draw each subset 100 times via Monte Carlo sampling. Each point in Supplementary Figure S12 is then the mean of the sampled score distribution for each value of n. If an XAI method assigns positive and negative attribution exactly the opposite way, negative correlation values are also possible.

Deletion deletes input features one at a time by replacing them with a baseline value based on their attribution score. For the Vector TS, image, and video representations we use zero as the baseline value, for the Point TS we use the implementation presented by [60], as in this case, a zero value does not correspond with an informationless state. Insertion gradually inserts features into a baseline input. The baseline input is an extremely blurred or smoothed version of the input observation ($\sigma = 5$), to simulate an informationless state without a distribution shift in the testing data, creating an OOD example, a problem discussed by [61].

We are inserting or deleting the features with the highest to the lowest attribution for both evaluation metrics and compute the neural network output at each step for every observation. In the original implementation of the metrics the area under the curve (AUC) value of the output for all steps is computed. However, this only works in the case of a classification task. Thus we implemented the area between the curve (ABC) computation, an adaptation of Insertion/Deletion to the regression task, suggested by [62].

Robustness Metrics To evaluate the robustness of our explanation we implemented the Sensitivity-Max and Infidelity metrics [63]. Both are based on the idea that a small perturbation of the input x should optimally also result only in a small change in the explanation. Infidelity calculates the expected mean-squared error (MSE) between the attribution-map A multiplied by a random variable input perturbation I and the differences between the neural network output f() at its input and perturbed input.

$$\text{Infd}(A, f, x) = \mathop{\mathbb{E}}_{I \sim D}[(I^T A(f,x) - (f(x) - f(x-I)))^2] \quad I = x - \epsilon \quad \epsilon \sim N(0, \sigma^2) \tag{6}$$

We implemented the difference to a noisy baseline as the input perturbation which subtracts a Gaussian random vector with a standard deviation $\sigma = 0.01$ from the input observation to receive the input perturbation I, following the distribution D.

Sensitivity-Max, however, measures the maximum change in the explanation with a small perturbation of the input x. Specifically, it measures the maximum sensitivity of an attribution-map A by sampling multiple observations s (in our case $n_s = 10$) from a subspace of an L-infinity ball with a defined input neighborhood radius ($r = 0.02$) and approximating the equation via Monte Carlo sampling.

$$\mathrm{Sens_{Max}(A, f, x, r) = \max_{||s-x||\leq r} ||A(f, s) - A(f, x)||} \tag{7}$$

Sensitivity-Max is upper-bounded for attribution-maps which are locally Lipshitz continuous.

Acknowledgements
This work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. We thank Lars Krämer, Karol Gotkowski and Carsten Lüth for insightful discussions and feedback.

References

[1] National Renewable Energy Laboratory (NREL), Best Research-Cell Efficiency Chart, URL `https://www.nrel.gov/pv/cell-efficiency.html`.

[2] A. Al-Ashouri, E. Köhnen, B. Li, A. Magomedov, H. Hempel, P. Caprioglio, J. A. Márquez, A. B. M. Vilches, E. Kasparavicius, J. A. Smith, N. Phung, D. Menzel, M. Grischek, L. Kegelmann, D. Skroblin, C. Gollwitzer, T. Malinauskas, M. Jošt, G. Matič, B. Rech, R. Schlatmann, M. Topič, L. Korte, A. Abate, B. Stannowski, D. Neher, M. Stolterfoht, T. Unold, V. Getautis, S. Albrecht, *Science* **2020**, *370*, 6522 1300.

[3] Y. Hou, E. Aydin, M. D. Bastiani, C. Xiao, F. H. Isikgor, D.-J. Xue, B. Chen, H. Chen, B. Bahrami, A. H. Chowdhury, A. Johnston, S.-W. Baek, Z. Huang, M. Wei, Y. Dong, J. Troughton, R. Jalmood, A. J. Mirabelli, T. G. Allen, E. V. Kerschaver, M. I. Saidaminov, D. Baran, Q. Qiao, K. Zhu, S. D. Wolf, E. H. Sargent, *Science* **2020**, *367*, 6482 1135.

[4] M. A. Ruiz-Preciado, F. Gota, P. Fassl, I. M. Hossain, R. Singh, F. Laufer, F. Schackmar, T. Feeney, A. Farag, I. Allegro, H. Hu, S. Gharibzadeh, B. A. Nejand, V. S. Gevaerts, M. Simor, P. J. Bolt, U. W. Paetzold, *ACS Energy Letters* **2022**, *7*, 7 2273.

[5] J.-P. Correa-Baena, M. Saliba, T. Buonassisi, M. Grätzel, A. Abate, W. Tress, A. Hagfeldt, *Science* **2017**, *358*, 6364 739.

[6] I. A. Howard, T. Abzieher, I. M. Hossain, H. Eggers, F. Schackmar, S. Ternes, B. S. Richards, U. Lemmer, U. W. Paetzold, *Advanced Materials* **2019**, *31*, 26 1806702.

[7] T. Abzieher, T. Feeney, F. Schackmar, Y. J. Donie, I. M. Hossain, J. A. Schwenzer, T. Hellmann, T. Mayer, M. Powalla, U. W. Paetzold, *Advanced Functional Materials* **2021**, *31*, 42 2104482.

[8] J. Li, H. Wang, X. Y. Chin, H. A. Dewi, K. Vergeer, T. W. Goh, J. W. M. Lim, J. H. Lew, K. P. Loh, C. Soci, T. C. Sum, H. J. Bolink, N. Mathews, S. Mhaisalkar, A. Bruno, *Joule* **2020**, *4*, 5 1035.

[9] K. B. Lohmann, J. B. Patel, M. U. Rothmann, C. Q. Xia, R. D. J. Oliver, L. M. Herz, H. J. Snaith, M. B. Johnston, *ACS Energy Letters* **2020**, *5*, 3 710, pMID: 32296733.

[10] B. Abdollahi Nejand, D. B. Ritzer, H. Hu, F. Schackmar, S. Moghadamzadeh, T. Feeney, R. Singh, F. Laufer, R. Schmager, R. Azmi, M. Kaiser, T. Abzieher, S. Gharibzadeh, E. Ahlswede, U. Lemmer, B. S. Richards, U. W. Paetzold, *Nature Energy* **2022**, *7*, 7 620.

[11] B. Chen, Z. J. Yu, S. Manzoor, S. Wang, W. Weigand, Z. Yu, G. Yang, Z. Ni, X. Dai, Z. C. Holman, J. Huang, *Joule* **2020**, *4*, 4 850.

[12] J. Li, J. Dagar, O. Shargaieva, O. Maus, M. Remec, Q. Emery, M. Khenkin, C. Ulbrich, F. Akhundova, J. A. Márquez, T. Unold, M. Fenske, C. Schultz, B. Stegemann, A. Al-Ashouri, S. Albrecht, A. T. Esteves, L. Korte, H. Köbler, A. Abate, D. M. Többens, I. Zizak, E. J. W. List-Kratochvil, R. Schlatmann, E. Unger, *Advanced Energy Materials* **2023**, *n/a*, n/a 2203898.

[13] A. S. Subbiah, F. H. Isikgor, C. T. Howells, M. De Bastiani, J. Liu, E. Aydin, F. Furlan, T. G. Allen, F. Xu, S. Zhumagali, S. Hoogland, E. H. Sargent, I. McCulloch, S. De Wolf, *ACS Energy Letters* **2020**, *5*, 9 3034.

[14] F. Mathies, E. J. W. List-Kratochvil, E. L. Unger, *Energy Technology* **2020**, *8*, 4 1900991.

[15] F. Schackmar, H. Eggers, M. Frericks, B. S. Richards, U. Lemmer, G. Hernandez-Sosa, U. W. Paetzold, *Advanced Materials Technologies* **2021**, *6*, 2 2000271.

[16] N.-G. Park, K. Zhu, *Nature Reviews Materials* **2020**, *5*, 5 333.

[17] S. Ternes, J. Mohacsi, N. Lüdtke, H. M. Pham, M. Arslan, P. Scharfer, W. Schabel, B. S. Richards, U. W. Paetzold, *ACS Applied Materials & Interfaces* **2022**, *14*, 9 11300, pMID: 35195981.

[18] L. Gu, F. Fei, Y. Xu, S. Wang, N. Yuan, J. Ding, *ACS Applied Materials & Interfaces* **2022**, *14*, 2 2949, pMID: 34985243.

[19] B. Abdollahi Nejand, I. M. Hossain, M. Jakoby, S. Moghadamzadeh, T. Abzieher, S. Gharibzadeh, J. A. Schwenzer, P. Nazari, F. Schackmar, D. Hauschild, L. Weinhardt, U. Lemmer, B. S. Richards, I. A. Howard, U. W. Paetzold, *Advanced Energy Materials* **2020**, *10*, 5 1902583.

[20] F. Mathies, H. Eggers, B. S. Richards, G. Hernandez-Sosa, U. Lemmer, U. W. Paetzold, *ACS Applied Energy Materials* **2018**, *1*, 5 1834.

[21] Y. C. Goh, X. Q. Cai, W. Theseira, G. Ko, K. A. Khor, *Scientometrics* **2020**, *125*, 2 1197.

[22] J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, *npj Computational Materials* **2019**, *5*, 1 83.

[23] X. Tang, *BJR Open* **2019**, *2*, 1 20190031.

[24] C. Odabaşı, R. Yıldırım, *Energy Technology* **2020**, *8*, 12 1901449.

[25] V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, K. S. Kim, *The Journal of Physical Chemistry C* **2020**, *124*, 16 8905.

[26] C. Li, H. Hao, B. Xu, Z. Shen, E. Zhou, D. Jiang, H. Liu, *Computational Materials Science* **2021**, *198* 110714.

[27] A. Ali, H. Park, R. Mall, B. Aïssa, S. Sanvito, H. Bensmail, A. Belaidi, F. El-Mellouhi, *Chemistry of Materials* **2020**, *32*, 7 2998.

[28] V. Starostin, V. Munteanu, A. Greco, E. Kneschaurek, A. Pleli, F. Bertram, A. Gerlach, A. Hinderhofer, F. Schreiber, *npj Computational Materials* **2022**, *8*, 1 101.

[29] F. Laufer, S. Ziegler, F. Schackmar, E. A. Moreno Viteri, M. Götz, C. Debus, F. Isensee, U. W. Paetzold, *Solar RRL* **2023**, *7*, 7 2201114.

[30] S. Ternes, F. Laufer, P. Scharfer, W. Schabel, B. S. Richards, I. A. Howard, U. W. Paetzold, *Solar RRL* **2022**, *6*, 3 2100353.

[31] F. Mathies, E. R. Nandayapa, G. Paramasivam, M. F. Al Rayes, V. R. F. Schröder, C. Rehermann, E. J. W. List-Kratochvil, E. L. Unger, *Mater. Adv.* **2021**, *2* 5365.

[32] M. Sundararajan, A. Taly, Q. Yan, In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17. JMLR.org, **2017** 3319–3328.

[33] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, **2017**, URL `http://arxiv.org/abs/1702.08608`.

[34] S. Wachter, B. D. Mittelstadt, C. Russell, *CoRR* **2017**, *abs/1711.00399*.

[35] S. Dandl, C. Molnar, M. Binder, B. Bischl, In T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, H. Trautmann, editors, *Parallel Problem Solving from Nature – PPSN XVI*. Springer International Publishing, Cham, ISBN 978-3-030-58112-1, **2020** 448–469.

[36] I. Stepin, J. M. Alonso-Moral, A. Catala, M. Pereira-Fariña, *Information Sciences* **2022**, *618* 379.

[37] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. sayres, In J. Dy, A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, **2018** 2668–2677, URL `https://proceedings.mlr.press/v80/kim18d.html`.

[38] F. Schackmar, F. Laufer, R. Singh, A. Farag, H. Eggers, S. Gharibzadeh, B. Abdollahi Nejand, U. Lemmer, G. Hernandez-Sosa, U. W. Paetzold, *Advanced Materials Technologies* **2023**, *8*, 5 2201331.

[39] K. Simonyan, A. Zisserman, In *International Conference on Learning Representations*. **2015** .

[40] K. He, X. Zhang, S. Ren, J. Sun, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. **2016** .

[41] K. He, X. Zhang, S. Ren, J. Sun, In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, **2016** 630–645.

[42] M. Tan, Q. Le, In *International conference on machine learning*. PMLR, **2019** 6105–6114.

[43] S. Zagoruyko, N. Komodakis, In *BMVC*. **2016** .

[44] C. Feichtenhofer, H. Fan, J. Malik, K. He, In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. **2019** .

[45] I. Loshchilov, F. Hutter, In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, **2019** URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

[46] A. Defazio, S. Jelassi, *J. Mach. Learn. Res.* **2023**, *23*, 1.

[47] I. Loshchilov, F. Hutter, In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, **2017** URL `https://openreview.net/forum?id=Skq89Scxx`.

[48] J. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, In *ICLR (workshop track)*. **2015** URL `http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a`.

[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, In *2017 IEEE International Conference on Computer Vision (ICCV)*. **2017** 618–626.

[50] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, S.-I. Lee, *Nature Machine Intelligence* **2021**, *3*, 7 620.

[51] D. Balduzzi, M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, B. McWilliams, In D. Precup, Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, **2017** 342–350, URL `https://proceedings.mlr.press/v70/balduzzi17b.html`.

[52] M. D. Zeiler, R. Fergus, In D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, editors, *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, ISBN 978-3-319-10590-1, **2014** 818–833.

[53] K. Simonyan, A. Zisserman, In Y. Bengio, Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. **2015** URL `http://arxiv.org/abs/1409.1556`.

[54] M. Schleich, Z. Geng, Y. Zhang, D. Suciu, *Proc. VLDB Endow.* **2021**, *14*, 9 1681–1693.

[55] R. K. Mothilal, A. Sharma, C. Tan, In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20. Association for Computing Machinery, New York, NY, USA, ISBN 9781450369367, **2020** 607–617, URL https://doi.org/10.1145/3351095.3372850.

[56] M. Augustin, V. Boreiko, F. Croce, M. Hein, In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., **2022** 364–377, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/025f7165a452e7d0b57f1397fed3b0fd-Paper-Conference.pdf.

[57] R. Tibshirani, *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58*, 1 267.

[58] M. Ancona, E. Ceolini, C. Öztireli, M. H. Gross, In *International Conference on Learning Representations*. **2017** .

[59] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, **2018**, URL https://arxiv.org/abs/1806.07421.

[60] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. **2019** 4197–4201.

[61] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., **2019** URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

[62] N. Hama, M. Mase, A. B. Owen, Deletion and insertion tests in regression models, **2022**, URL https://arxiv.org/abs/2205.12423.

[63] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., **2019** URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a7471fdc77b3435276507cc8f2dc2569-Paper.pdf.

[64] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, *Information* **2020**, *11*, 2 125.

[65] F. Isensee, P. Jäger, J. Wasserthal, D. Zimmerer, J. Petersen, S. Kohl, J. Schock, A. Klein, T. Roß, S. Wirkert, P. Neher, S. Dinkelacker, G. Köhler, K. Maier-Hein, batchgenerators - a python framework for data augmentation, **2020**, URL https://doi.org/10.5281/zenodo.3632567.