Understanding Scalable Perovskite Solar Cell Manufacturing with Explainable AI

Lukas Klein[*]
Interactive Machine Learning Group, DKFZ
Institute for Machine Learning, ETH Zürich
Helmholtz Imaging, DKFZ
Heidelberg, Germany
lukas.klein@dkfz.de

Sebastian Ziegler[*]
Division of Medical Image Computing, DKFZ
Helmholtz Imaging, DKFZ
Heidelberg, Germany
sebastian.ziegler@dkfz.de

Felix Laufer
Light Technology Institute, KIT

Charlotte Debus
Steinbuch Centre for Computing, KIT
Helmholtz AI, KIT

Markus Götz
Steinbuch Centre for Computing, KIT
Helmholtz AI, KIT

Klaus Maier-Hein
Division of Medical Image Computing, DKFZ
Helmholtz Imaging, DKFZ

Ulrich W. Paetzold[†]
Light Technology Institute, KIT
Institute of Microstructure Technology, KIT
Karlsruhe, Germany

Fabian Isensee[†]
Division of Medical Image Computing, DKFZ
Helmholtz Imaging, DKFZ
Heidelberg, Germany

Paul F. Jäger[†]
Interactive Machine Learning Group, DKFZ
Helmholtz Imaging, DKFZ
Heidelberg, Germany

Abstract

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. Through a diverse set of XAI methods, we explain not only *what* characteristics are important but also *why*, allowing material scientists to translate findings into actionable conclusions. Our study demonstrates that XAI methods will play a critical role in accelerating energy materials science.

[*] Contributed equally

[†] Contributed equally

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

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 perovskite semiconductors, power conversion efficiencies (PCEs) of PSCs have improved vastly, exceeding 30% PCE in perovskite/silicon tandem photovoltaics [NREL, 2023]. Despite numerous advantages, [Al-Ashouri et al., 2020; Hou et al., 2020; Ruiz-Preciado et al., 2022], the technology has not reached the market yet due to insufficient device stability (degrading PCE over time) and the lack of cost-effective and reliable large-scale production [Correa-Baena et al., 2017; Howard et al., 2019]. The crystallization process during manufacturing heavily affects the perovskite thin-film 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 [Gu et al., 2022; Abdollahi Nejand et al., 2020; Mathies et al., 2018]. 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 production parameters is lacking to reduce the significant volatility in PSC quality.

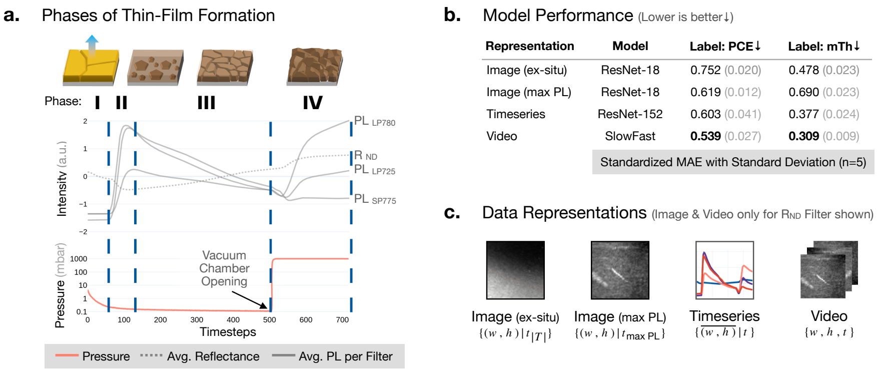
Machine learning (ML) has recently been applied to specific optimization problems in various fields, including materials sciences, as it outperforms humans in finding correlations and clues in highly complex data [Goh et al., 2020; Schmidt et al., 2019; Tang, 2019]. Specifically, in perovskite research, ML has been used to optimize specific parameters on tabular data, e.g. material choice [Odabaşı and Yıldırım, 2020], bandgaps [Gladkikh et al., 2020], compositional ionic radii [Li et al., 2021] or optimizing specific characteristics like the morphology or crystal structure utilizing scanning electron microscope (SEM) [Ali et al., 2020] and grazing incidence x-ray diffraction (GIXD) images [Starostin et al., 2022]. However, the current application of ML in perovskite research is only working with low-dimensional data, looking exclusively at the final thin-film (ex-situ), but not the perovskite formation process itself (in-situ). 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 can find patterns in complex data that would be infeasible to find through traditional analyses, we use XAI methods from the areas of feature importance, counterfactual examples, and concept testing to render these patterns humanly understandable, which then 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.[1]

2 Predicting the quality of perovskite solar cells

Dataset This study builds on the publicly available dataset published by Laufer et al. [2023] that contains in-situ photo-luminescence (PL) video data of 1,129 PSCs (Figure 1). 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. However, since small variations in the process parameters resulting from small human or technical inconsistencies are always present, the data contains a wide range of quality within the PSCs that cannot be explained by looking at the defined process parameters since they are nominally the same. 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 [Ternes et al., 2022]. Subsequent to the processing of the perovskite thin-film, the full device

[1] Code available at: `https://github.com/MIC-DKFZ/perovskite-xai`.


a. Phases of Thin-Film Formation

b. Model Performance (Lower is better↓)

Representation	Model	Label: PCE↓	Label: mTh↓
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)
Timeseries	ResNet-152	0.603 (0.041)	0.377 (0.024)
Video	SlowFast	**0.539** (0.027)	**0.309** (0.009)

Standardized MAE with Standard Deviation (n=5)

c. Data Representations (Image & Video only for R_{ND} Filter shown)

Figure 1: **a.** The figure shows the four different thin-film formation phases based on the average PL and reflectance intensity timeseries for each of the four filters. Below, the simultaneous change of the air pressure in mbar is depicted. **b.** Model performance is measured in standardized mean absolute error (sMAE) to compare between labels. **c.** Examples of the four data representations used.

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 model training, allowing to learn the relationship between the videos and the quality of a PSC. While a higher PCE always indicates a better solar cell quality the mTh should not be too thick as it is highly correlated with the thin-film roughness, and a homogenous layer morphology is critical for a high-quality PSC (see Appendix A for the relationship between PCE and mTh).

We transform the data into four different representations that are depicted in Figure 1 (c.). Each representation focuses on different aspects of the videos, allowing to compare several XAI methods across varying data dimensionalities. While the video representation contains all available information the two image representations only cover spatial aspects by selecting one frame from the video. We chose the frame with the highest PL (in-situ) and the last frame (ex-situ). The timeseries represents each frame as a mean, resulting in one line per PL filter and only containing temporal information. Figure 1 (a.) depicts a characteristic PL signal in the timeseries data representation. Characteristics of the PL signal can be attributed to different phases during the perovskite thin-film formation, which we extend from Howard et al. [2019]: 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 thin-film surface morphology, i.e. surface roughness [Mathies et al., 2021].

Model Training For each data representation and label we test several architectures and undergo an extensive hyperparameter tuning on the training set to obtain best possible predictions (see Appendix B for all hyperparameters). While for the timeseries and image-based representations ResNet [He et al., 2016] architectures worked best, we used a Slowfast model [Feichtenhofer et al., 2019] for the videos. Figure 1 (b.) shows the standardized MAE on the test set averaged over 5 training runs. For both labels the video representation combining temporal and spatial information yielded the best performance. Models trained on representations containing time information outperform models trained on spatial information alone. When limiting the data to only one frame (image representation), thereby neglecting the temporal dimension, choosing the timestep influences the prediction performance differently for each label. In general, mTh prediction is more accurate than PCE prediction because PCE can only measured on the complete PSC requiring additional processing steps not captured in the videos while mTh only depends on the captured processing step. A parity plot is available in Figure 7. Overall, the trained models

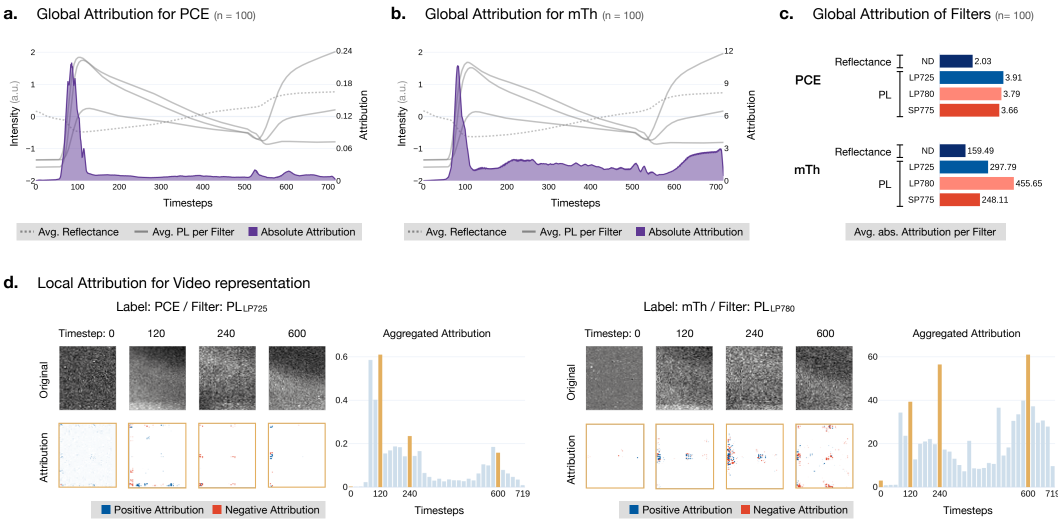


Figure 2: **a.** & **b.** Absolute attribution-map for PCE and mTh averaged over 100 observations. **c.** The average absolute attribution of each filter over the whole dataset, to assess the importance of each filter when predicting PCE or mTh. **d.** Attribution-map for the video data of label PCE 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.

show a good performance in predicting the general trends and build a reliable foundation for the following XAI analysis.

3 Feature Importance

To understand which input features and phases are most important to our models, we apply several attribution methods [Sundararajan et al., 2017; Doshi-Velez and Kim, 2017; Erion et al., 2021] to compute either local explanations, i.e. explaining a model's behavior on a single observation, or global explanations, i.e. explaining patterns that are present in general (see Appendix D for a detailed methodology and list of the attribution methods). The diversity in attribution methods enlarges the trustworthiness of the results. Figure 2 (a.&b.) shows the global attribution computed via Expected Gradients (also called Gradient SHAP) for PCE and mTh averaged over 100 timeseries observations (see Appendix subsection D.1 for all representations and attribution methods). Our analysis highlights that the model focuses on time periods that coincide with the defined phases. We observe that models predicting PCE and mTh both show the highest absolute attribution to *Phase II*, the onset of the nucleation and crystallization phase. In addition, models predicting mTh also show attribution to *Phase IV*. Specifically, there is a small attribution peak at around $t = 510$ before the dip in PL intensity, and a large attribution concentration after the dip. The attributed periods of the vacuum quenching starting at around $t = 505$ and the subsequent venting strongly affect the crystallization and the morphology of the perovskite layer [Schackmar et al., 2023]. For PCE observations, only a smaller attribution spike at $t = 510$ can be observed. 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 2 (d.))

To examine the importance of each of the four filters, we show their mean absolute attribution in Figure 2, discovering that they contribute to different extends to the final prediction. Further, the importance of each filter differs between PCE and mTh prediction. In case of PCE, the filtered PL intensities are substantially more important than the reflectance. However, for mTh, PL_{LP780} appears much more important than the other two PL intensities.

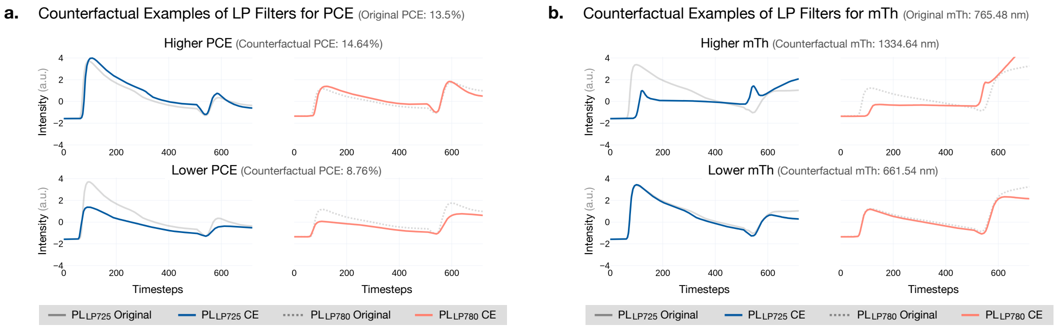


Figure 3: **a.** & **b.** Generated CEs of the LP filters yielding either high or low PCE and mTh prediction. The original predicted PCE or mTh for this observation is shown in the header and the PCE or mTh predicted based on the artificially computed CEs is presented behind the label.

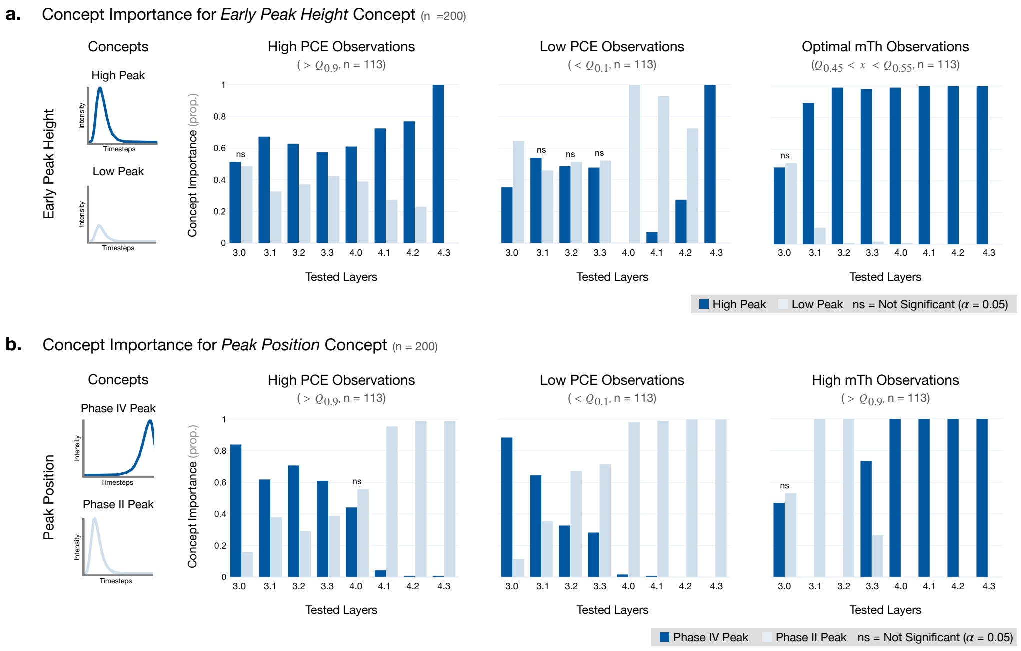
4 Counterfactual Examples

The derived attribution maps highlight the significance of individual features but do not explain the underlying behavior responsible for such importance. To answer this question we deploy counterfactual explanations (CEs) [Wachter et al., 2017; Dandl et al., 2020; Stepin et al., 2022] and the Testing of Concept Activation Vectors (TCAV) [Kim et al., 2018]. CEs alter the input observation to receive a specific counterfactual outcome and simulate "what if" -scenarios. To generate CEs, we use the Genetic Counterfactuals (GeCo) algorithm [Schleich et al., 2021], which computes plausible (assuring that they could be real) and feasible (assuring they can actually be computed) CEs in a short time. 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.

The CEs for the two most important filters to the models reveal that when moderately increasing the nucleation onset peak during *Phase II* the model predicts higher PCE values and vice versa (Figure 3 (a.), see Appendix E for all filters and representations). Subsequently for mTh, a decreased PL intensity of the nucleation onset results in higher mTh prediction (Figure 3 (b.)), and a high PL intensity during *Phase IV* leads to higher mTh. To predict a lower mTh, however, no substantial change in the PL intensity is required, suggesting that lower measured mTh values in the dataset still fall into an optimal range, and only for higher values the PL intensity course is substantially different.

5 Testing of Concept Activation Vectors (TCAV)

Based on the CE analysis we define the two concept classes of "Early Peak Height" and "Peak Position" to test the importance of each including concept to specific layers of the model. For each concept class C, we sample two datasets of examples, $C = [c_1, c_2]$, that are representative of each of the two concepts we want to test against each other. 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 (see Appendix A), and the CE analysis revealed that lower mTh values do not necessarily result from substantially different PL intensity curves. We sample two datasets of examples, $C = [c_1, c_2]$, that are representative of each of the two concepts we want to test. Each of the four datasets is sampled separately for each filter using summary statistics from each timeseries and specific permutations to avoid out-of-distribution (OOD) examples (see Appendix F for more details).


a. Concept Importance for *Early Peak Height* Concept (n =200)

b. Concept Importance for *Peak Position* Concept (n = 200)

Figure 4: **a.** & **b.** For TCAV, we test the last eight layers of the model, 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 of the concept classes “Early Peak Height” and “Peak Position”, or whether there is no significant difference (ns, based on proportion z-test with a significance level (α) of 0.05).

For the concept classes of “Early Peak Height”, Figure 4 (a.) 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 (a.), see Appendix subsection F.1 for high mTh observations). As feature importance and CE analysis determined also the importance of *Phase IV*, we compare in “Peak Position” the two concepts “Phase II Peak" and “Phase IV Peak" to further distinguish between the two most important time periods to the model. While both concepts are equally important for high PCE observations, “Early Peak" is more important for low PCE observations (Figure 4 (b.)). 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, with “Late Peak" being moderately more important than “Early Peak”, confirming the importance of *Phase IV* previously observed in the CE experiments (see Appendix subsection F.1 for low mTh). Both TCAV findings reconfirm the CE-based conclusions.

In summary, our data-driven approach shows two findings: that a higher peak in *Phase II* leads to improved PSC quality and that the perovskite thin-film roughness correlates to the timing of the venting step. The first 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 [Schackmar et al., 2023]. Subsequently, we would recommend for future processes to increase the evacuation rate to achieve higher PL peaks, which is indicative of higher solar cell performance. Based on the second finding, we conclude 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. Thus we would recommend optimizing 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 can be achieved by extending the evacuation times which dries the thin-film and eliminates the PL increase during venting.

6 Conclusion & Discussion

In our work, we applied a diverse set of XAI methods in collaboration with material scientists to answer scientific questions, which would not be possible by traditional human analysis. One critical aspect is the application of different methods to first detect important features by attribution methods and second understand why these features are important through CE and TCAV methods. Although all methods adequately addressed the specified task, we observed certain limitations. For CEs there are more effective methods available such as Diverse Counterfactuals (DiCE) [Mothilal et al., 2020] or Diffusion Visual Counterfactual Explanations (DVCEs) [Augustin et al., 2022], however, their computational complexity is significantly larger. The high speed in the computation of GeCo is especially important for our task, as it would be otherwise computationally infeasible to compute CEs for high-dimensional data such as videos. Further, TCAV is only suitable to discover "new knowledge" to a limited extend, as tested concepts have to be defined in advance, thereby inheriting a potential bias from prior assumptions.

As XAI methods only explain correlations, leaving final causal inferences to the judgement of the material scientist, the explanations are limited by the dataset. Naturally, there is a possibility of unobserved parameters, not captured in our dataset, but still affecting the labels. However, the information-rich video data captures the result of the interactions among all parameters influencing the thin-film formation by recording the actual formation itself. Therefore the possibility of important unobserved parameters and confounders is minimized. Additionally, since the interpretation of XAI results in relation to underlying causal variables is conducted by human experts, they also account for potential confounding factors.

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. We are able to infer insights just by analyzing the video dataset and without having to carry out extensive and costly trial-and-error experiments. Our encouraging insights 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.

Acknowledgments and Disclosure of Funding

This work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. The authors acknowledge the the financial support of the Helmholtz Association (Program-Oriented Funding Phase IV, MTET Topic 1, 38.01.05 / the project AI-INSU-PERO, ZT-I-PF-5-106 / the project AutoPeroSol, ZT-PF-3-020), the Karlsruhe School of Optics and Photonics, and the German Federal Ministry of Education and Research (Solar Tap innovation platform). We thank Lars Krämer, Karol Gotkowski and Carsten Lüth for insightful discussions and feedback.

7 Supplementary Material

A Relationships of PCE and mTh

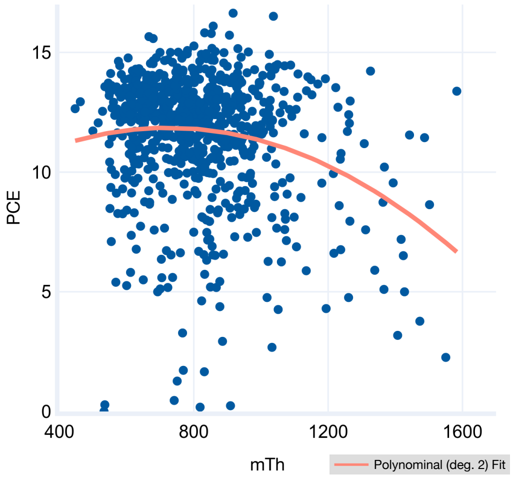


Figure 5: Relationship between mTh and PCE, modeled with a second-degree polynomial regression.

When displaying mTh against PCE in Figure 5 and fitting a second-degree polynomial regression, we observe that on average the highest PCE is around 800nm mTh. While lower mTh only leads to a minor decrease in PCE, an increase in mTh leads to lower PCE. From approximately 1,000nm on we observe a negative correlation between PCE and mTh.

The thickness measurements were performed using a profilometer where a stylus is moved over the thin-film's surface (see Figure 6). By removing all material (up to the transparent conductive oxide) at multiple positions and taking them as reference points, the profilometer surface scans can be used to determine the thickness of the perovskite layer after subtracting the thickness of all other (evaporated) layers with well-defined thicknesses. Given the same solution and material volume as perovskite thin-films with a smooth surface, the averaging of the acquired thickness over the scan length of thin-films with rough surfaces results in an increased thin-film layer thickness value. Therefore, high layer thicknesses measured using a profilometer with subsequent averaging over the entire scan correlate with increased surface roughness.

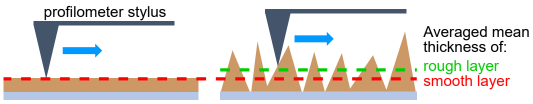


Figure 6: Schematic representation of the relationship between thickness measurement and thin-film roughness.

B Neural Network Hyperparameters & Augmentations

Label	Representation	Model	Optimizer	LR	Augmentation
PCE	Image$_{ex-situ}$	ResNet-18	Madgrad	0.0001	Flip ($p = 0.5$) One of: Motion Blur ($p = 0.2$) Median Blur ($p = 0.1$) Blur ($p = 0.1$) Piecewise Affine ($p = 0.5$) z-transformation
	Image$_{maxPL}$	ResNet-18	Madgrad	0.0001	Random Flip ($p = 0.5$) Gaussian Blur ($p = 0.5$) z-transformation
	Timeseries	ResNet-152	Madgrad	0.0001	z-transformation
	Video	SlowFast	AdamW	0.001	z-transformation Gamma Transform Gaussian Blur ($p = 0.15$) Random Flip ($p = 0.3$) Blank Rectangles
mTh	Image$_{ex-situ}$	ResNet-18	Madgrad	0.0001	Random Flip ($p = 0.5$) Gaussian Blur ($p = 0.5$) z-transformation
	Image$_{maxPL}$	ResNet-18	Madgrad	0.0001	Random Flip ($p = 0.5$) Gaussian Blur ($p = 0.5$) z-transformation
	Timeseries	ResNet-152	Madgrad	0.0001	z-transformation
	Video	SlowFast	AdamW	0.001	z-transformation Gamma Transform Gaussian Blur ($p = 0.15$) Random Flip ($p = 0.3$) Blank Rectangles

Table 1: Used hyperparameters for training the NNs. Augmentations for 2D representations were all implemented using Albumentations [Buslaev et al., 2020] while 3D augmentations were implemented with Batchgenerators [Isensee et al., 2020]. All runs were trained for 1,000 epochs with a batch size of 256 using a cosine annealing learning rate (LR) scheduler.

C Neural Network Performance Evaluation

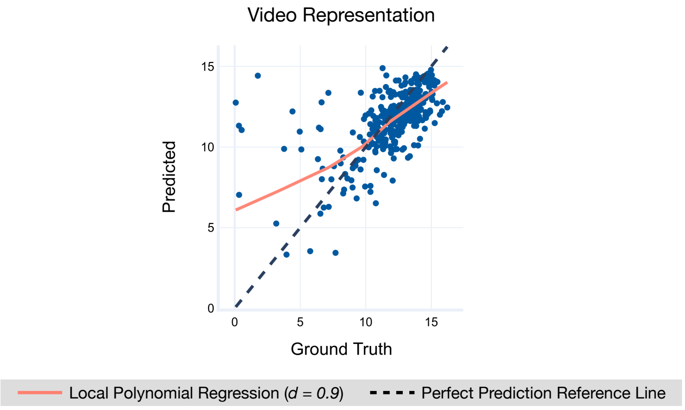


Figure 7: Parity plot for the video representation. Local Polynomial Regression to compare error trend with perfect prediction reference line.

Figure Figure 7 shows a parity plot for the Slowfast model trained on the video data representation and using PCE as target. We observe that low-PCE cases are consistently overestimated by the model. This is due to the fact that PCE can only be measured on the complete PSC while the videos only capture the perovskite layer manufacturing and not its subsequent steps. If errors occur in these subsequent processing steps the final PCE is lower while the perovskite layer may be actually of good quality. Therefore it is expected to overestimate these cases. Besides that, the model can reliably distinguish between high and low-quality PSCs.

D Attribution Methods

Due to the risk of confirmation bias and unfaithful explanations [Doshi-Velez and Kim, 2017], we compute each attribution-map for all representations and labels with four different attribution methods. These include Guided Backpropagation (GBP) [Springenberg et al., 2015], Guided Gradient-weighted Class Activation Mapping (GGC) [Selvaraju et al., 2017], Integrated Gradients (IG) [Sundararajan et al., 2017], and Expected Gradients (EG) [Erion et al., 2021]. 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 [Balduzzi et al., 2017] and ignoring global effects in the input space. Thus, they can e.g. be combined with deconvolutional networks [Zeiler and Fergus, 2014] 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 [Simonyan and Zisserman, 2015], 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 Springenberg et al. [2015]). 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.

D.1 Feature Importance for all Representations, Labels and Attribution Methods

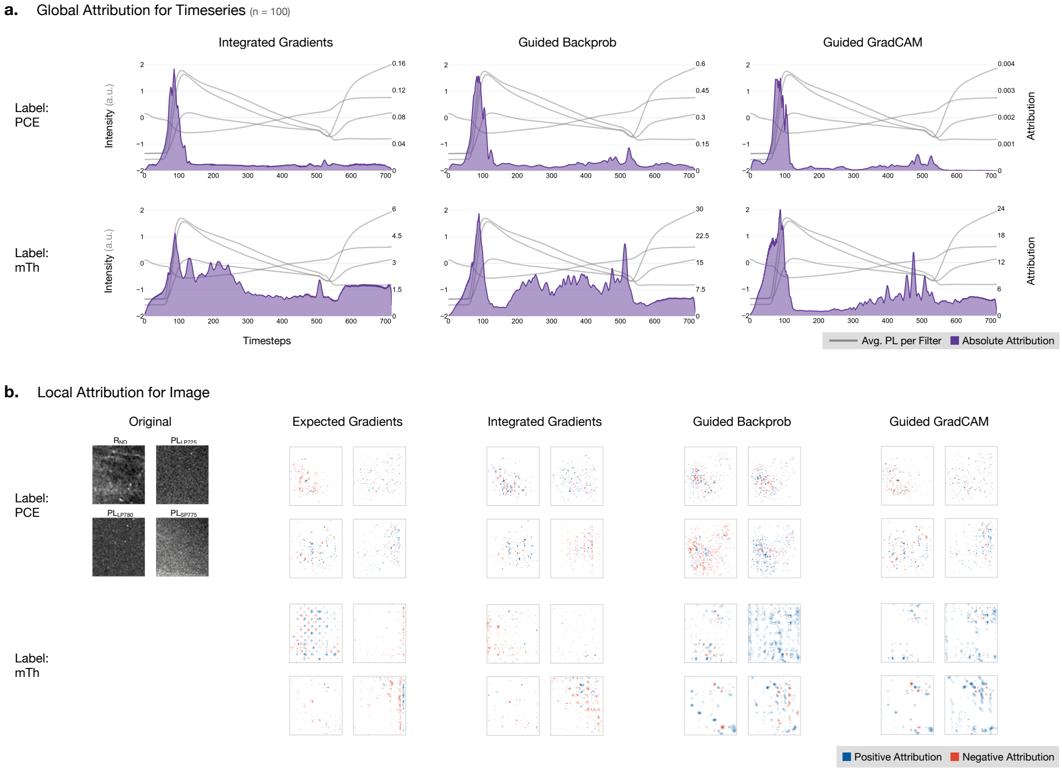


Figure 8: **a.** Absolute attribution-maps for PCE and mTh for IG, GBP, and GGC averaged over 100 observations. The maps show very similar characteristics across all XAI methods. Only IG on mTh attributes more to the phase after the nucleation onset compared to GBP GGC and EG (Figure 5). **b.** As the features in the image representation are location invariant, we can not produce global explanations by averaging the absolute attribution-maps. Thus the local explanation in positive and negative attribution per filter is shown for a single observation for both labels and all four XAI methods.

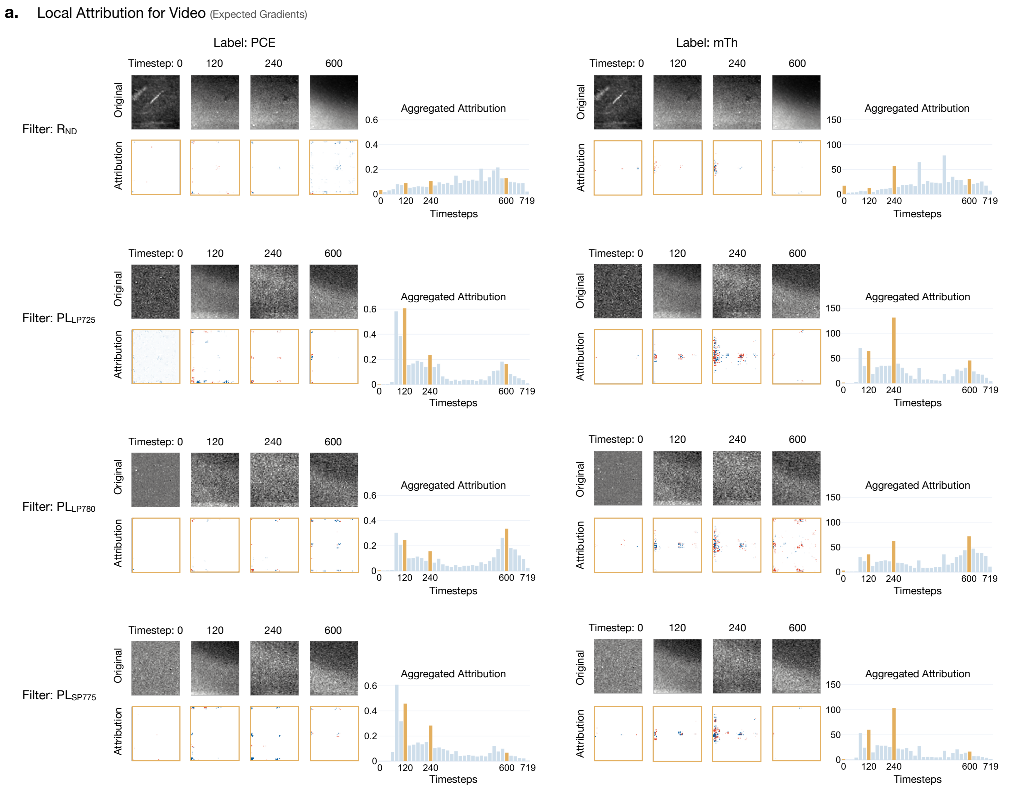


Figure 9: EG-based attribution-maps of four selected frames at the timesteps 0, 120, 240, and 600 for both labels and all filters. The aggregated absolute attribution shows the importance of each frame of the filter. Also, it shows the same attribution pattern as for the timeseries representation in 8.

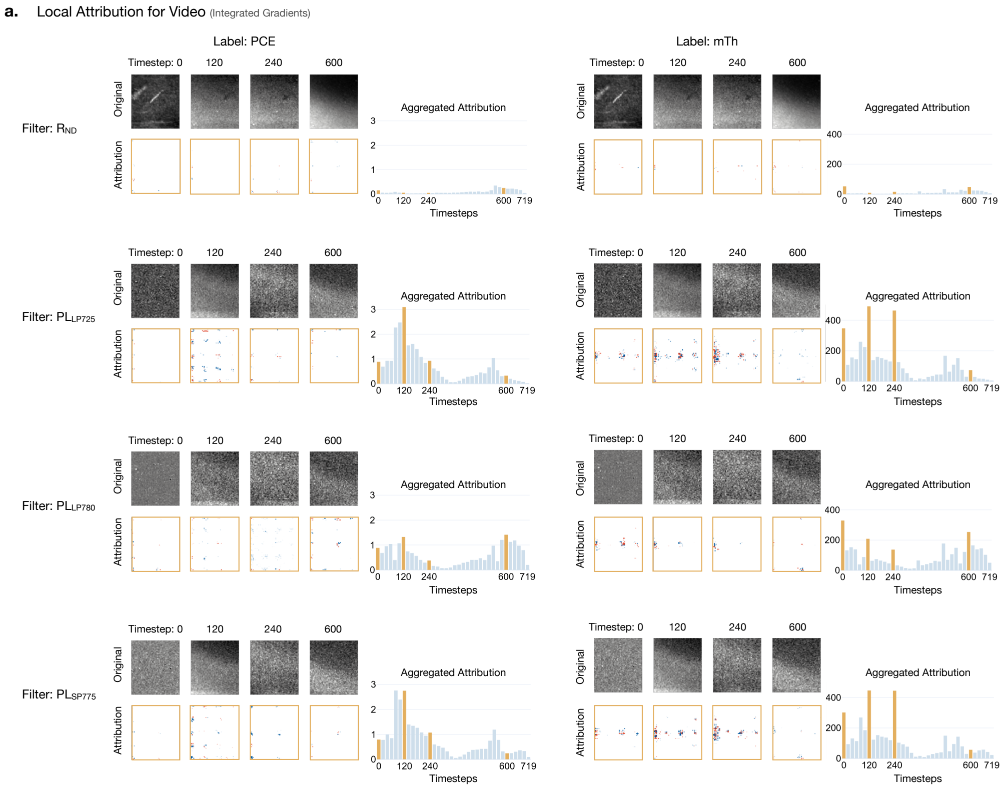


Figure 10: IG-based attribution-maps of four selected frames at the timesteps 0, 120, 240, and 600 for both labels and all filters. The aggregated absolute attribution shows the importance of each frame of the filter.

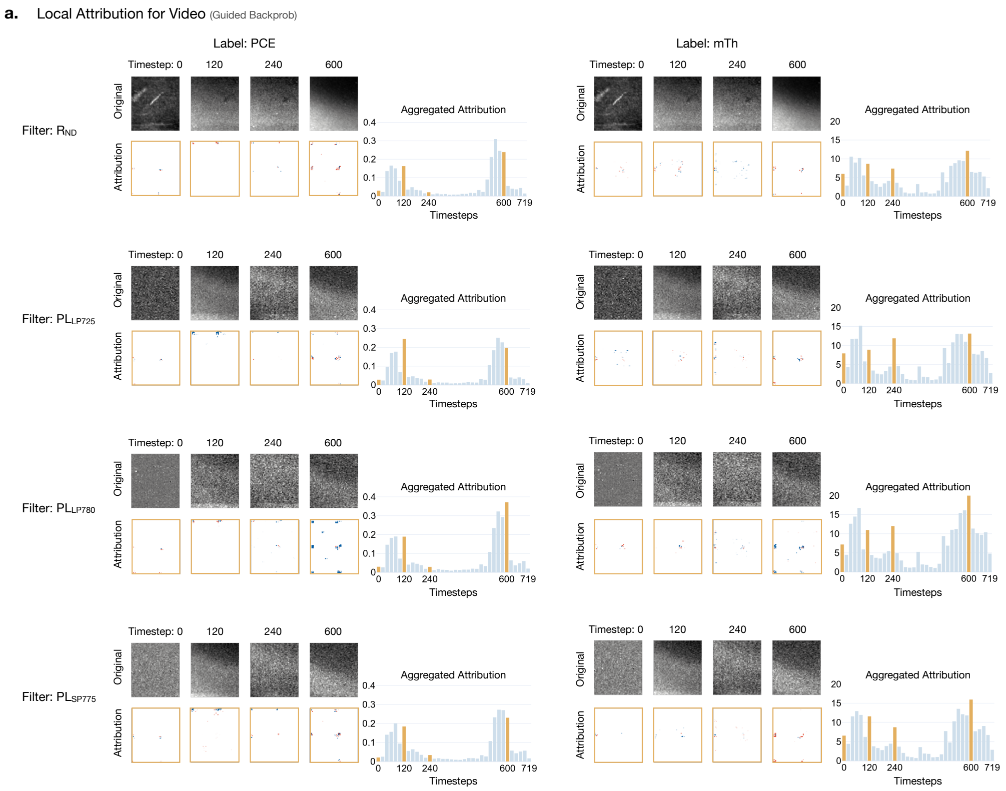


Figure 11: GBP-based attribution-maps of four selected frames at the timesteps 0, 120, 240, and 600 for both labels and all filters. The aggregated absolute attribution shows the importance of each frame of the filter.

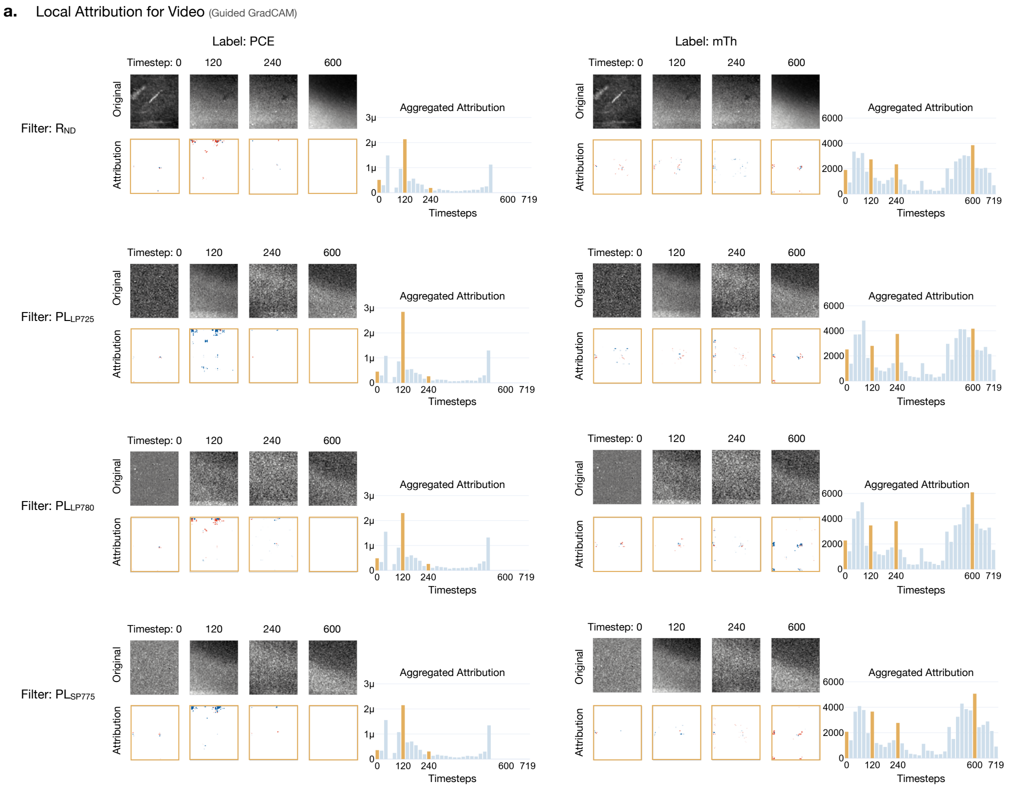


Figure 12: GGC-based attribution-maps of four selected frames at the timesteps 0, 120, 240, and 600 for both labels and all filters. The aggregated absolute attribution shows the importance of each frame of the filter.

E Counterfactual Examples

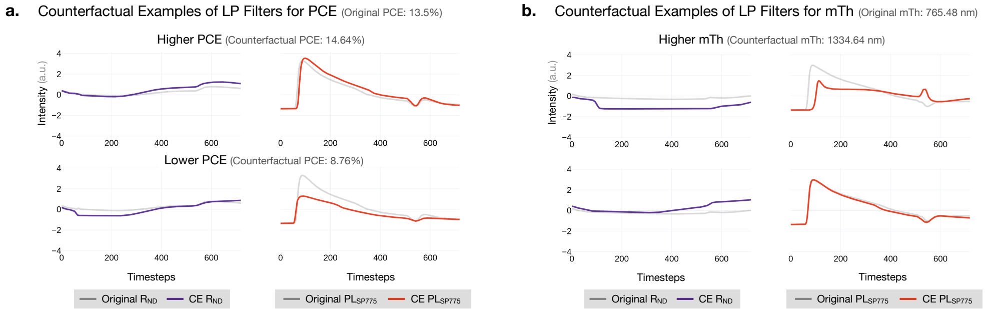


Figure 13: **a. & b.** Artificial generated CEs of a timeseries for substantially higher or lower PCE and mTh for filters R_{ND} and PL_{SP775}.

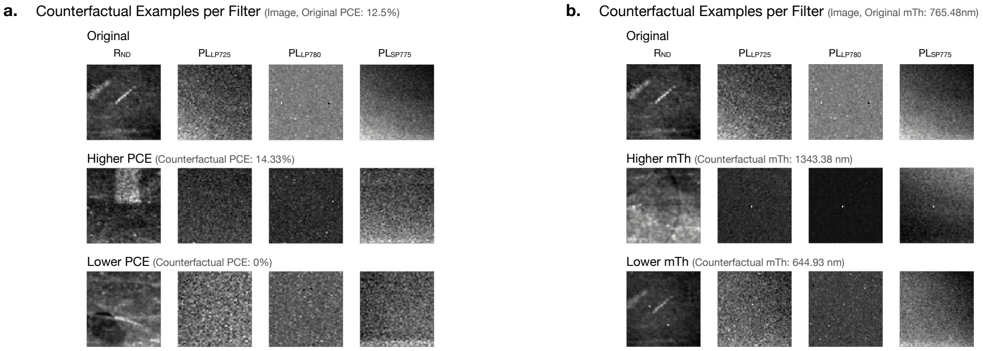

a. Counterfactual Examples per Filter (Image, Original PCE: 12.5%)
Original
RND
PLLP725
PLLP780
PLSP775
Higher PCE (Counterfactual PCE: 14.33%)
Lower PCE (Counterfactual PCE: 0%)
b. Counterfactual Examples per Filter (Image, Original mTh: 765.48nm)
Higher mTh (Counterfactual mTh: 1343.38 nm)
Lower mTh (Counterfactual mTh: 644.93 nm)

Figure 14: **a. & b.** Artificial generated CEs of the image representation per filter for PCE and mTh. We observe specific artifacts generated in the images by the genetic algorithm in the R_{ND} filter. For all other filters, mainly the brightness is controlled. Interestingly, in the case of lower PCE, the algorithm generated a CE that achieves a 0% PCE prediction.

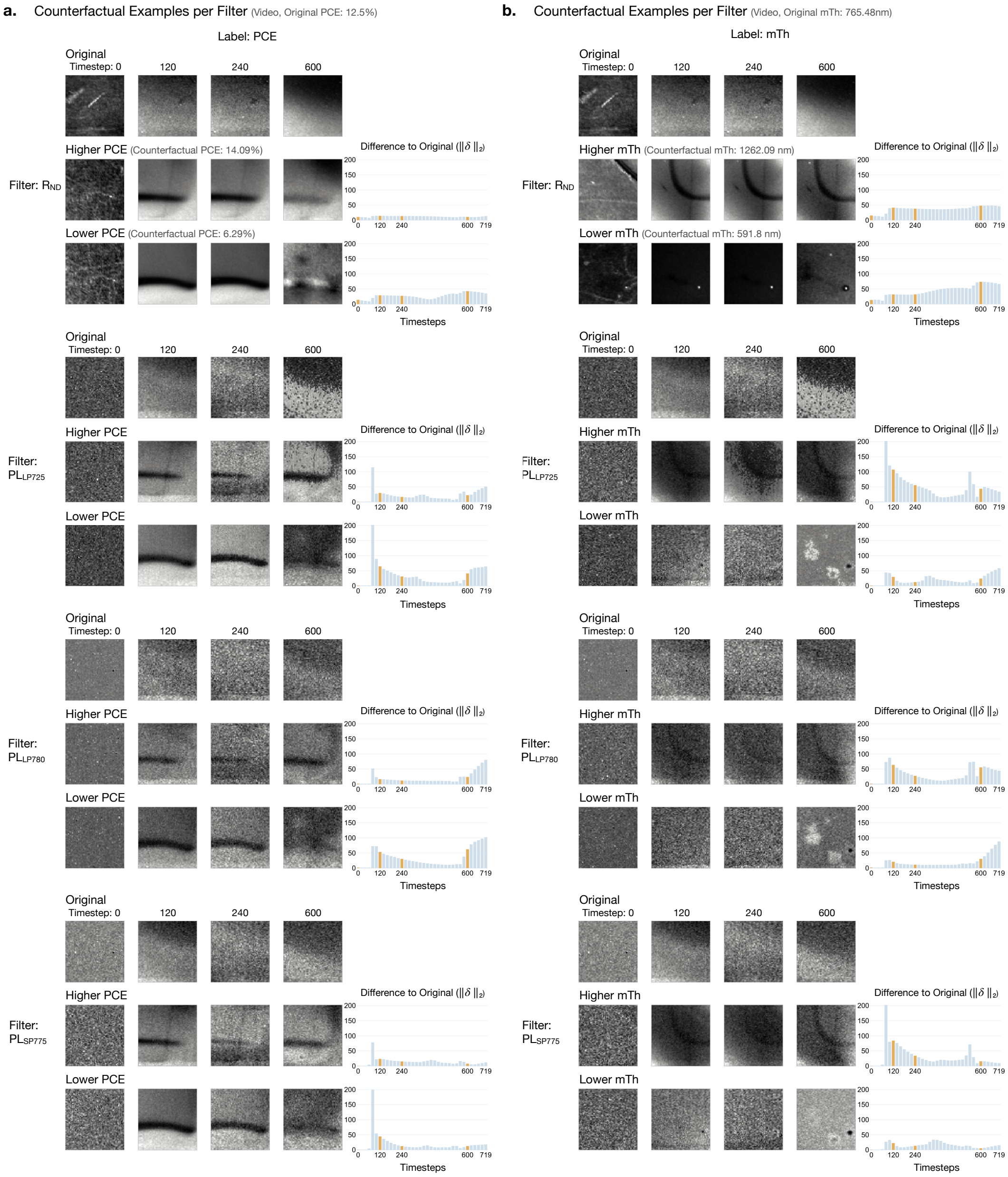


Figure 15: **a. & b.** Artificially generated CEs of the video representation for both labels. Each filter shows four original frames together with the counterfactual frames of the substantially higher or lower prediction. To get a more high-level perspective of what was altered in the video CEs, we present for each filter the Euclidean distance ($||\delta||_2$) between each counterfactual frame and the original frame to see which frames were altered the most. As for the timeseries representation, we observe that most alternations were done during the nucleation onset and surface morphology formation phases.

F TCAV

We leverage TCAV [Kim et al., 2018] 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.

Label	alpha (α)	max iterations	tolerance (δ)
PCE	0.02	50,000	1e-7
mTh	0.02	100,000	1e-8

Table 2: Hyperparameters of the Lasso-regression [Tibshirani, 1996] used to linearly divide the concepts c_1 and c_2. For each label the hyperparameters are the same for all three concept sets. The random state was constant during all experiments.

The sampling of each concept class:

$$\begin{aligned} &\text{EarlyPeakHeight(F)}_{\text{Low,High}} \sim \text{N}(\text{Q}_{0.15|0.85}(\hat{\text{X}}), 0.5 * \sigma(\hat{\text{X}})) \\ &\hat{X} = \{max(\{x_{t=0}, \dots, x_{360}\})|x \in X^n\}_F^n \end{aligned} \tag{4}$$

Early Peak Height Choose $n = 100$ random locations during *Phase II*. For each filter F, take the 0.85 (high peak) and respectively 0.15 (low peak) quantile value ($Q_{0.15|0.85}$) of the maximum values from $t = 0$ to $t = 360$ ($\hat{X}$). For each random location sample a peak from a normal distribution with the mean of the quantile value and variance of $0.5 * \sigma(\hat{X})$ of all *Phase II* maximum values. Resample and interpolate to a fitting curve. Repeat for each filter and combine filters for each observation.

$$\begin{aligned} &\text{PeakPosition(F)}_{\text{Early,Late}} \sim \text{N}(\hat{\mu}, 0.25 * |\hat{\mu}|) \\ &\hat{\mu} = \bar{\hat{X}} - C \quad \hat{X} = \{max(\{x_{t=0|505}, \dots, x_{360|719}\})|x \in X^n\}_F^n \end{aligned} \tag{5}$$

Peak Position Choose $n = 100$ random locations during *Phase II* and *Phase IV*. For each of the 200 locations and the four filters F, sample a peak from a normal distribution with the mean equal to the mean over the maximum values of either of both phases ($\hat{X}$) and subtract a regularization constant not to produce OOD observations. The variance is then equal to $0.25 * |\hat{\mu}|$. Resample and interpolate to a fitting curve. Repeat for each filter and combine filters for each observation.

F.1 TCAV results for all concepts

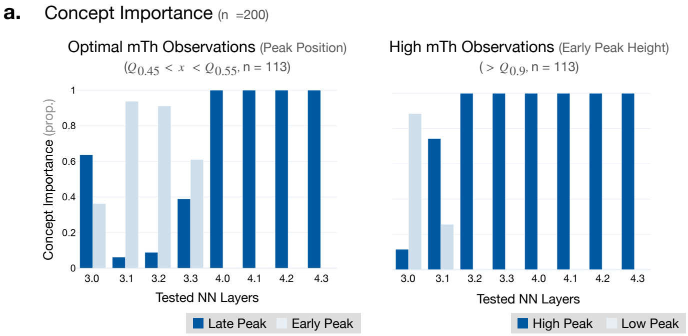


Figure 16: For the peak position concept, the late peak is moderately more important than the early peak for both mTh subsets (see Figure 4 (b.) for higher mTh). In the case of the early peak height concept, the model is always sensitive to high peaks (see Figure 4 (a.) for optimal mTh observations).

References

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, and U. W. Paetzold, "Vacuum-assisted growth of low-bandgap thin films ($fa_{0.8}ma_{0.2}sn_{0.5}pb_{0.5}i_3$) for all-perovskite tandem solar cells," *Advanced Energy Materials*, vol. 10, no. 5, p. 1902583, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/aenm.201902583

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, and S. Albrecht, "Monolithic perovskite/silicon tandem solar cell with $>$29% efficiency by enhanced hole extraction," *Science*, vol. 370, no. 6522, pp. 1300–1309, 2020. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.abd4016

A. Ali, H. Park, R. Mall, B. Aïssa, S. Sanvito, H. Bensmail, A. Belaidi, and F. El-Mellouhi, "Machine learning accelerated recovery of the cubic structure in mixed-cation perovskite thin films," *Chemistry of Materials*, vol. 32, no. 7, pp. 2998–3006, 2020. [Online]. Available: https://doi.org/10.1021/acs.chemmater.9b05342

M. Augustin, V. Boreiko, F. Croce, and M. Hein, "Diffusion visual counterfactual explanations," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 364–377. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/025f7165a452e7d0b57f1397fed3b0fd-Paper-Conference.pdf

D. Balduzzi, M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, and B. McWilliams, "The shattered gradients problem: If resnets are the answer, then what is the question?" in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 342–350. [Online]. Available: https://proceedings.mlr.press/v70/balduzzi17b.html

A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, feb 2020. [Online]. Available: https://doi.org/10.3390%2Finfo11020125

J.-P. Correa-Baena, M. Saliba, T. Buonassisi, M. Grätzel, A. Abate, W. Tress, and A. Hagfeldt, "Promises and challenges of perovskite solar cells," *Science*, vol. 358, no. 6364, pp. 739–744, 2017. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aam6323

S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *Parallel Problem Solving from Nature – PPSN XVI*, T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, and H. Trautmann, Eds. Cham: Springer International Publishing, 2020, pp. 448–469.

F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017. [Online]. Available: http://arxiv.org/abs/1702.08608

G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 620–631, Jul 2021. [Online]. Available: https://doi.org/10.1038/s42256-021-00343-w

C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, and K. S. Kim, "Machine learning for predicting the band gaps of abx3 perovskites from elemental properties," *The Journal of Physical Chemistry C*, vol. 124, no. 16, pp. 8905–8918, 2020. [Online]. Available: https://doi.org/10.1021/acs.jpcc.9b11768

Y. C. Goh, X. Q. Cai, W. Theseira, G. Ko, and K. A. Khor, "Evaluating human versus machine learning performance in classifying research abstracts," *Scientometrics*, vol. 125, no. 2, pp. 1197–1212, Nov 2020. [Online]. Available: https://doi.org/10.1007/s11192-020-03614-2

L. Gu, F. Fei, Y. Xu, S. Wang, N. Yuan, and J. Ding, "Vacuum quenching for large-area perovskite film deposition," *ACS Applied Materials & Interfaces*, vol. 14, no. 2, pp. 2949–2957, 2022, pMID: 34985243. [Online]. Available: https://doi.org/10.1021/acsami.1c22128

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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, and E. H. Sargent, "Efficient tandem solar cells with solution-processed perovskite on textured crystalline silicon," *Science*, vol. 367, no. 6482, pp. 1135–1140, 2020. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aaz3691

I. A. Howard, T. Abzieher, I. M. Hossain, H. Eggers, F. Schackmar, S. Ternes, B. S. Richards, U. Lemmer, and U. W. Paetzold, "Coated and printed perovskites for photovoltaic applications," *Advanced Materials*, vol. 31, no. 26, p. 1806702, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.201806702

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, and K. Maier-Hein, "batchgenerators - a python framework for data augmentation," Jan. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3632567

B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2668–2677. [Online]. Available: https://proceedings.mlr.press/v80/kim18d.html

F. Laufer, S. Ziegler, F. Schackmar, E. A. Moreno Viteri, M. Götz, C. Debus, F. Isensee, and U. W. Paetzold, "Process insights into perovskite thin-film photovoltaics from machine learning with in situ luminescence data," *Solar RRL*, vol. 7, no. 7, p. 2201114, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/solr.202201114

C. Li, H. Hao, B. Xu, Z. Shen, E. Zhou, D. Jiang, and H. Liu, "Improved physics-based structural descriptors of perovskite materials enable higher accuracy of machine learning," *Computational Materials Science*, vol. 198, p. 110714, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025621004419

F. Mathies, H. Eggers, B. S. Richards, G. Hernandez-Sosa, U. Lemmer, and U. W. Paetzold, "Inkjet-printed triple cation perovskite solar cells," *ACS Applied Energy Materials*, vol. 1, no. 5, pp. 1834–1839, 2018. [Online]. Available: https://doi.org/10.1021/acsaem.8b00222

F. Mathies, E. R. Nandayapa, G. Paramasivam, M. F. Al Rayes, V. R. F. Schröder, C. Rehermann, E. J. W. List-Kratochvil, and E. L. Unger, "Gas flow-assisted vacuum drying: identification of a novel process for attaining high-quality perovskite films," *Mater. Adv.*, vol. 2, pp. 5365–5370, 2021. [Online]. Available: http://dx.doi.org/10.1039/D1MA00494H

R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 607–617. [Online]. Available: https://doi.org/10.1145/3351095.3372850

NREL, "Best Research-Cell Efficiency Chart," 2023. [Online]. Available: https://www.nrel.gov/pv/cell-efficiency.html

C. Odabaşı and R. Yıldırım, "Assessment of reproducibility, hysteresis, and stability relations in perovskite solar cells using machine learning," *Energy Technology*, vol. 8, no. 12, p. 1901449, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ente.201901449

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, and U. W. Paetzold, "Monolithic two-terminal perovskite/cis tandem solar cells with efficiency approaching 25%," *ACS Energy Letters*, vol. 7, no. 7, pp. 2273–2281, 2022. [Online]. Available: https://doi.org/10.1021/acsenergylett.2c00707

F. Schackmar, F. Laufer, R. Singh, A. Farag, H. Eggers, S. Gharibzadeh, B. Abdollahi Nejand, U. Lemmer, G. Hernandez-Sosa, and U. W. Paetzold, "In situ process monitoring and multichannel imaging for vacuum-assisted growth control of inkjet-printed and blade-coated perovskite thin-films," *Advanced Materials Technologies*, vol. 8, no. 5, p. 2201331, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/admt.202201331

M. Schleich, Z. Geng, Y. Zhang, and D. Suciu, "Geco: Quality counterfactual explanations in real time," *Proc. VLDB Endow.*, vol. 14, no. 9, p. 1681–1693, may 2021. [Online]. Available: https://doi.org/10.14778/3461535.3461555

J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, p. 83, Aug 2019. [Online]. Available: https://doi.org/10.1038/s41524-019-0221-0

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," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a

V. Starostin, V. Munteanu, A. Greco, E. Kneschaurek, A. Pleli, F. Bertram, A. Gerlach, A. Hinderhofer, and F. Schreiber, "Tracking perovskite crystallization via deep learning-based feature detection on 2d x-ray scattering data," *npj Computational Materials*, vol. 8, no. 1, p. 101, May 2022. [Online]. Available: https://doi.org/10.1038/s41524-022-00778-8

I. Stepin, J. M. Alonso-Moral, A. Catala, and M. Pereira-Fariña, "An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information," *Information Sciences*, vol. 618, pp. 379–399, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002002552201218X

M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3319–3328.

X. Tang, "The role of artificial intelligence in medical imaging research," *BJR Open*, vol. 2, no. 1, p. 20190031, Nov. 2019.

S. Ternes, F. Laufer, P. Scharfer, W. Schabel, B. S. Richards, I. A. Howard, and U. W. Paetzold, "Correlative in situ multichannel imaging for large-area monitoring of morphology formation in solution-processed perovskite layers," *Solar RRL*, vol. 6, no. 3, p. 2100353, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/solr.202100353

R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x

S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *CoRR*, vol. abs/1711.00399, 2017. [Online]. Available: http://arxiv.org/abs/1711.00399

M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.