

Process Insights into Perovskite Thin-Film Photovoltaics from Machine Learning with In Situ Luminescence Data

Felix Laufer, Sebastian Ziegler, Fabian Schackmar, Edwin A. Moreno Viteri, Markus Götz, Charlotte Debus, Fabian Isensee, and Ulrich W. Paetzold*

Large-area processing remains a key challenge for perovskite solar cells (PSCs). Advanced understanding and improved reproducibility of scalable fabrication processes are required to unlock the technology's economic potential. In this regard, machine learning (ML) methods have emerged as a promising tool to accelerate research and unlock the control needed to produce large-area solution-processed perovskite thin films. However, a suitable dataset allowing the analysis of a scalable fabrication process is currently missing. Herein, a unique labeled in situ photoluminescence (PL) dataset for blade-coated PSCs is introduced and explored with unsupervised k-means clustering, demonstrating the feasibility to derive meaningful insights from such data. Correlations between the obtained clusters and the measured performance of PSC reveal that the in situ PL signal encodes information about the perovskite thin-film quality. Detrimental mechanisms during thin-film formation are detected by identifying spatial differences in PL patterns and, consequently, of device performance. In addition, k-nearest neighbors is applied to predict the performance of PSCs, motivating further investigations into ML-based in-line process monitoring of scalable PSC fabrication to detect, understand, and ultimately minimize process variations across iterations.

technology demonstrates outstanding power conversion efficiencies (PCEs), exceeding 25%.^[3] Despite numerous favorable optoelectronic properties of perovskite semiconductors, four key challenges remain and delay the successful commercialization of perovskite solar cells (PSCs): 1) the long-term stability, 2) the toxicity of the contained lead, 3) upscaling to large-areas, and 4) unlocking cost-effective, reliable large-scale production (high throughput and high yield).^[2,4] Traditional efforts in material science and device engineering in the field are based on countless trial-and-error experiments. However, these approaches for material discovery, process development, characterization, full device evaluation, and stability testing are often complicated, expensive, laborious, and time-consuming given the large experimental parameter space.^[5] These drawbacks motivate the implementation of autonomous experimentation methods and data-driven techniques like machine learning (ML).^[6,7]


1. Introduction

Photovoltaics (PV) is one of the most auspicious technologies for the transition of the global energy mix toward a sustainable future. Within PV, hybrid metal-halide perovskite semiconductors have emerged since 2009 as a promising absorber material class for the next-generation thin-film solar cells.^[1,2] Already today, the

In an increasing number of research fields, ML methods are employed to identify yet undiscovered correlations and to provide insights into fundamental working principles. Besides pattern extraction, ML can be utilized to make classifications or predictions and to uncover new insights into the studied data. For this reason, ML algorithms are successfully adopted to an increasing number of applications in materials science,^[8–11] encompassing,

F. Laufer, F. Schackmar, E. A. Moreno Viteri, U. W. Paetzold
Light Technology Institute (LTI)
Karlsruhe Institute of Technology (KIT)
Engesserstrasse 13, 76131 Karlsruhe, Germany
E-mail: ulrich.paetzold@kit.edu

S. Ziegler, F. Isensee
Division of Medical Image Computing
German Cancer Research Center (DKFZ)
Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/solr.202201114>.

© 2023 The Authors. Solar RRL published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

DOI: 10.1002/solr.202201114

S. Ziegler, F. Isensee
Applied Computer Vision Lab
Helmholtz Imaging
Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

F. Schackmar, U. W. Paetzold
Institute of Microstructure Technology (IMT)
Karlsruhe Institute of Technology (KIT)
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

M. Götz, C. Debus
Helmholtz AI
Steinbuch Centre for Computing (SCC)
Karlsruhe Institute of Technology (KIT)
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

most recently, also the field of perovskite PV.^[12–15] ML is employed for the discovery of new perovskite semiconductors^[16–18] and to showcase the accelerated development of lead-free perovskites.^[19–21] Moreover, the long-term stability of PSCs has been investigated by combining ML with high-throughput experimentation^[22,23] or by applying ML to datasets generated through data mining of previous publications on the topic of perovskite stability.^[24] Time series models have been developed for accelerated material stability evaluation and performance forecasting in humid environments.^[25–27] Furthermore, data-driven approaches can be used to generate predictive models for optoelectronic characteristics such as the bandgap^[28–31] based on theoretical physical material features like ionization energy, atomic/molecular sizes, or lattice constant. These material parameters can be extracted from open-access material properties databases.^[32–36] Yet, ML applications in PV are not limited to accelerated material discovery, the prediction of theoretical material properties, or the extraction of new knowledge from previous publications through data mining. As demonstrated for silicon PV, ex situ deployment of these methods in combination with imaging techniques enables monitoring and quantification of the current operational status of fully processed solar cells during operation in the field.^[37–45]

The key ingredient for unlocking large-scale and thus economically viable production of PSCs is the upscaling of the thin-film formation process to large areas using scalable deposition techniques. To this end, the entangled phases of film formation, i.e., drying, nucleation, and crystal growth, must be carefully monitored and controlled to obtain high-quality optoelectronic thin films.^[4] The ideal perovskite thin film exhibits morphology that features large grain sizes, high film density, low surface roughness, and uniform crystal structure without pinholes and imperfections on large areas. Considering the complexity of the entangled film formation phases, scaling the technology is intricate and requires an enhanced understanding of the highly complex thin-film formation.^[4] To this end, in situ characterization methods, are needed to non-invasively monitor the quality of the thin film during its formation on large areas with a temporal resolution of less than one second.^[46] Fulfilling these requirements, in situ photoluminescence (PL) imaging is used to gain insight into the complex fabrication process of the perovskite thin film. In situ PL imaging has great potential to monitor, control and research the perovskite thin-film formation.^[46–50] In this work, we introduce and analyze a unique, labeled dataset containing in situ PL data from 1129 blade-coated perovskite thin-film solar cells processed under the very exact same conditions, layer stacks, and precursor materials, so that performance variations are solely caused by fluctuations in the fabrication process itself. In contrast to the commonly used spin-coating fabrication process, blade-coating is closer to an industrial manufacturing scenario since it is scalable to larger areas. Data acquisition was performed with an in-house-built imaging setup^[50] during the vacuum-assisted quenching of blade-coated perovskite thin films, which initiates the drying and crystallization of high-quality perovskite absorber layers.^[51–53] This dataset offers the potential to enhance the understanding of the scalable fabrication of perovskite layers, a critical step toward the commercialization of perovskite PV (see **Figure 1A**). The application of ML techniques to analyze the perovskite layer formation allows

identifying the circumstances and causes of unintended deviations across iterations from a previously optimized fabrication process. In the conventional fabrication process, unintentional deviations from the optimized experimental process cannot be detected and evaluated until the layer stack is completed into a full device, which requires several additional experimental process steps after the perovskite layer is deposited. Elucidating the link between the imaged PL intensity acquired in situ during the perovskite deposition step and the performance of the finished solar cells will ultimately enable performance predictions prior to the solar cells actually going through the several remaining processing steps to finalize the full devices. Additionally, implementing early detection of fabrication flaws in an industrial production line can mitigate costs through the application of preventive maintenance strategies, resulting in savings in both time and materials.

To exploit the potential of in situ PL imaging for monitoring perovskite formation, this work explores supervised and unsupervised ML algorithms on the generated in situ PL dataset to detect unintended process variations introduced by upscaling of the fabrication process. First, the advantage of acquiring in situ data during the perovskite formation over ex situ PL data is revealed by highlighting the correlation between expert-chosen in situ PL features and the corresponding PCE. Then, unsupervised *k*-means clustering is used for initial data exploration of human-encoded features extracted from the luminescence data. Subsequently, the dependence on data encoding by human experts is fully removed by clustering the entire PL transients. It is demonstrated that *k*-means clustering creates transient PL clusters that correlate with the performance of the final PSC. In addition, *k*-means clustering is used to identify adverse process mechanisms in the formation of perovskite thin films. Spatial correlations of the generated transient PL clusters and, consequently, of solar cell performance are detected. Areas with increased detrimental perovskite thin-film properties are identified through differences in the spatial distribution of PL data patterns assigned to the different clusters. Finally, after the initial exploration of the dataset using unsupervised *k*-means clustering, the supervised ML algorithm *k*-nearest neighbors (kNN) is applied to predict PV parameters and perovskite layer thickness, which has a major influence on solar cell performance. The promising results motivate further investigations using more advanced supervised ML techniques like neural networks in conjunction with the entire image time series data. In summary, this work demonstrates that ML-based analysis of our introduced unique in situ PL dataset offers high potential for in-line process monitoring and can accelerate the commercialization of perovskite thin-film PV by identifying process variations at an early stage during device fabrication.

2. Results and Discussion

2.1. Experimental In Situ Luminescence Dataset for Perovskite PV

Applying ML methods to data acquired in situ during the fabrication process of PSCs is a promising strategy for improving process understanding and reproducibility. However, to train ML

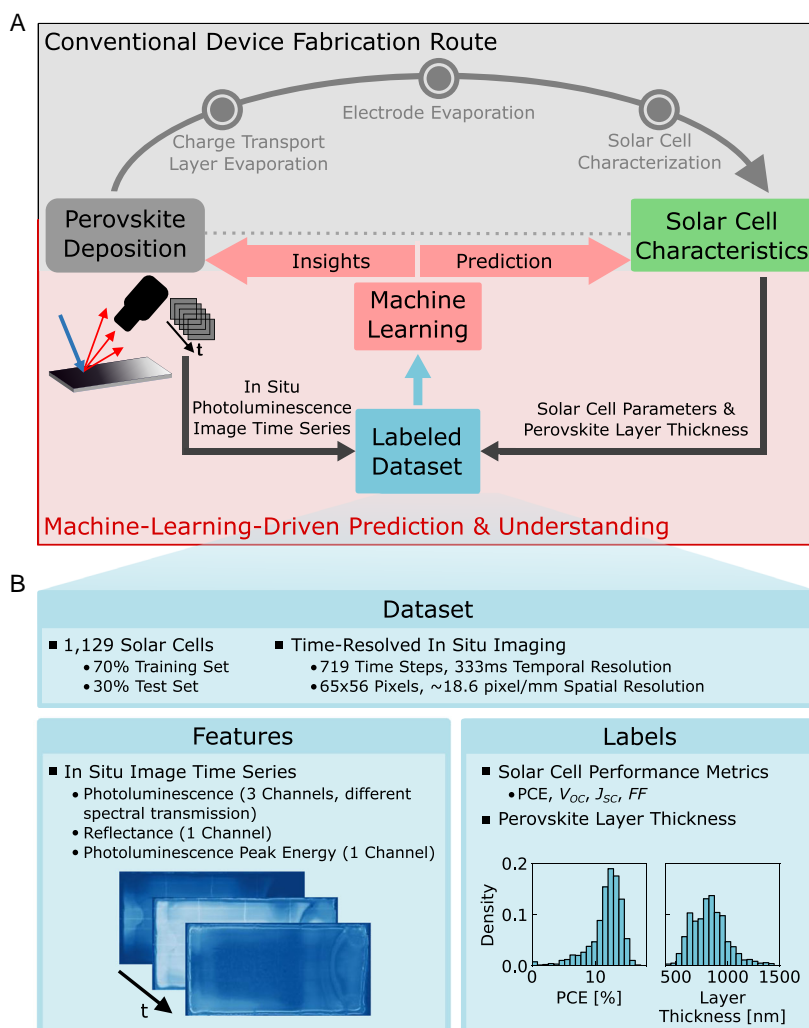


Figure 1. Process understanding and prediction through machine learning. A) Schematic illustration of a feedback loop driven by machine learning (ML) methods applied to experimental in situ data. Training an ML model on in situ data enables directly predicting the solar cell characteristics without having to complete the fabrication process. Furthermore, ML helps to understand and to optimize the perovskite deposition by revealing underlying patterns the model uses for its prediction. B) Description of the generated dataset containing 1129 solar cells. For each solar cell, time-resolved in situ imaging was performed during perovskite formation yielding multiple channels of image time series (videos) which can be used as features when implementing ML algorithms. Furthermore, performance metrics as well as the perovskite layer thickness were acquired and can be used as labels.

models, large in situ datasets are necessary. To this end, 1129 solar cells were fabricated using the blade coating deposition technique (see Figure S1, Supporting Information). The vacuum quenching process of the perovskite layer was monitored using a PL imaging setup. The camera captures images of the entire $32 \times 64 \text{ mm}^2$ substrate with the blade-coated perovskite layer on top (see Figure 2A and S2, Supporting Information).

Accordingly, the generated dataset contains time-resolved in situ images acquired during the formation of the perovskite layer (see Figure 1B). The imaging setup outputs four channels containing image time series ($2D + t$) measured through different spectral filters, capturing reflectance (one channel) and different parts of the PL spectrum (three channels). Furthermore, the three PL channels with different spectral transmissions were used to compute an image time series of spatially resolved PL peak energy (one channel) using the method introduced by

Chen et al.^[54] In a preprocessing step, the images were cropped into 32 smaller patches (65×56 pixels each), only depicting the active area of a singular solar cell. For all experimental iterations, the transient data was aligned at the beginning of the evacuation of the vacuum chamber. 170 s after the start of data acquisition, i.e., after ≈ 505 time steps, the chamber was flooded with ambient air before data acquisition ended after a total of 240 s, resulting in a time series of 719 time steps at 3 frames per second. Consequently, the image time series in the dataset encompasses the drying and crystallization of the blade-coated perovskite thin films.

All solar cells were then built to completion and their performance was measured to determine current-density-voltage curves. For each solar cell in the dataset, the PV parameters (PCE, open-circuit voltage (V_{OC}), short-circuit current density (J_{SC}), fill factor (FF)), measured backward and forward, were

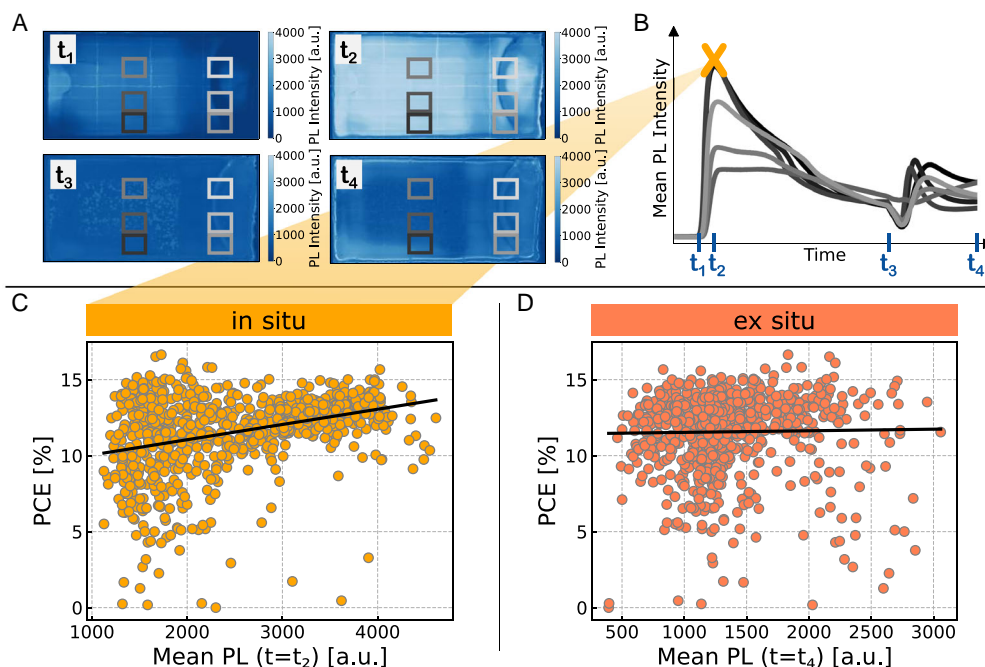


Figure 2. Rationale for using in situ data. A) Exemplary photoluminescence (PL) images of a blade-coated perovskite thin-film layer acquired at four different points in time during the vacuum quenching process. The rectangles mark six out of a total of 32 areas on the substrate, which coincide with the active area of a solar cell. B) Graph displaying the transient behavior of the spatial mean PL intensity of the active areas of six solar cells marked in (A). The orange cross at $t = t_2$ marks the transients' maxima as an in situ feature. Starting at $t = t_3$, the vacuum chamber is again vented with ambient air. C,D) Scatter plots showing each solar cell as a data point scattered over its measured power conversion efficiency (PCE) and the detected mean PL intensity of the in situ feature at $t = t_2$ (C) or of the ex situ feature at $t = t_4$ (D). The black lines represent the fits of the linear regression corresponding to the calculated correlation coefficients.

added as parameters to the dataset. In this work, the mean between backward and forward measurement is used as a label for all performance metrics. Furthermore, the mean perovskite layer thickness of each solar cell's active area was determined with a profilometer along three spatially offset scan lines. The thickness information was added to the dataset as an additional parameter. Auxiliary information such as the substrate ID and the position of each solar cell within its substrate is provided as well.

The complete dataset contains data from a total of 1129 solar cells. For the evaluation of the predictive capabilities of ML models, the dataset was split into training and test datasets, containing 780 and 349 solar cell samples, respectively. The training set is used for model development, i.e., the selection of hyperparameters, and for training the models, while the test set allows for an unbiased evaluation of the models using previously unseen data. The split was performed on a substrate level, moving all (up to 32) solar cells originating from the same substrate in unison to either the training or test split. This way, the evaluation of the model using the test set can be viewed as actual predictions on newly generated experimental data. Furthermore, the training dataset was divided into five subsets for fivefold cross-validation, again assigning solar cells originating from the same substrate to the same data subset (see Experimental Section for more information).

All solar cells were fabricated using the same materials, the same experimental methods, and the same experimental

parameters. Thus, this dataset can be utilized for the application of ML techniques to improve process understanding, as well as for in-line performance prediction, for process monitoring, and for in-line detection of inferior perovskite thin-film quality due to differences between blade coating iterations. These potential applications highlight that this experimental dataset is a crucial stepping stone enabling data-driven in-line process monitoring to accelerate the industrialization of perovskite PV. The dataset is made publicly available to the community (<https://doi.org/10.5281/zenodo.7503391>).^[55]

2.2. Benefits of Acquiring In Situ Photoluminescence Data

The initial investigation of the training dataset in this work demonstrates the value of the generated dataset by showing that the in situ PL data correlates with solar cell performance and can be used to predict it. To decrease the complexity of the problem, the data's dimensionality is reduced by conducting feature reduction using domain knowledge, i.e., expert knowledge about the underlying research area and the generated data. For each time step, the spatially resolved patches in the PL images coinciding with the solar cells' active area are encoded by the mean value of the PL intensity of the patch (see Figure 2A,B). This emphasizes the temporal sequence of the experiment and allows the investigation of small differences in the material formation process between iterations, which are revealed by variations in the temporal evolution of PL intensity. Among the resulting transients

for each channel, only the data acquired through the 725 nm long pass filter is selected, as it captures the entire PL emission spectrum of the used perovskite material (optical band gap of approximately 1.59 eV). Hence, the data used in this work focuses on the transient evolution of the PL intensity while keeping the analysis of spatial (in-)homogeneity in the image patches for future work.

The benefits of recording in situ data during the perovskite formation process are highlighted by comparing the correlation of the solar cell performance with in situ PL features versus ex situ PL features. First, the in situ PL transients are examined with a simple, intuitive approach, e.g., by correlating the most prominent feature of the transients, the maximum in PL, to the solar cell performance (see Figure 2C). It is apparent that the PCE of the solar cell tends to increase with the maximum of the transient PL. For high maximum mean PL values, the corresponding PCEs are mostly higher than 11% (see Figure 2C). For low maximum mean PL, the mean PCE decreases and the distribution widens. This implies that even for low mean maximum PL some solar cells show high performance, but the ratio of low performing solar cells (PCE < 10%) increases. The correlation between the in situ feature and the PCE is quantified by a Pearson correlation coefficient of $r = 0.3386$ with a fit quality measure of $r^2 = 0.1147$ (see Table 1) and by the increasing slope of the fitted line obtained by linear regression (see Figure 2C). While the expert's domain knowledge is crucial to encode the in situ PL transient, the choice of the data point used as the input feature is somewhat sensitive and highly influences the result. However, already the simple approach provided here reveals the above-mentioned correlation. In the second step, the last data points of the transients ($t = t_4$) are selected as representative of ex situ data to demonstrate the benefit of acquiring in situ data during the vacuum quenching process over ex situ data. When compared to the corresponding in situ plot, the ex situ data points are more dispersed (see Figure 2D) indicating that the ex situ data encodes less information content regarding the subsequently measured PCE of the solar cell when compared to the previously presented in situ data. The small correlation between the ex situ feature and the PCE is demonstrated by a correlation coefficient of $r = 0.0182$ with a fit quality measure of $r^2 = 0.0003$ (see Table 1) as well as by the nearly flat linear regression line (see Figure 2D).

Finally, the correlation coefficients confirm a higher correlation between the in situ feature and the labels compared to the ex situ features (see Table 1). In addition to the difference in correlation for PCE, the correlation coefficient for the in situ feature is also particularly high for FF compared to the ex situ value. The highest correlation for both the in situ as well as ex situ feature is found for perovskite layer thickness. This is due to the fact that the performance of the solar cell is influenced

by the quality of the entire device, while the perovskite layer thickness is completely determined during the process step represented in the dataset. The other steps do not add any uncertainty to the layer thickness, whereas the performance parameters may be subject to variation introduced by the other processing steps, leading to a lower correlation. In summary, in situ PL data is considerably more insightful compared to ex situ PL data and, consequently, highly advantageous for in-the-loop ML-based process monitoring.

2.3. Initial Exploration of the Dataset Using Unsupervised Machine Learning

To explore and analyze the dataset further, the unsupervised ML technique of k -means clustering is applied to the training set. The main use case of unsupervised ML is a data exploration and pattern extraction. The k -means clustering algorithm divides data into k clusters, each containing data with similar characteristics (see Experimental Section for more information).^[56–58] To ensure comparability, the number of clusters k is fixed. In the Figure S3, Supporting Information, the choice of the number of clusters $k = 4$ is most suitable according to the established elbow method^[59] in combination with cross-validation for all examples presented in the following.

By selecting features, the entire transient is reduced into a small number of values, making it easy to interpret the data. In this initial exploration, prominent features of the PL transients are selected as input for the clustering algorithm. In the process, gradually more and more input features are introduced, leading to an increase in information, but also to higher complexity (see Figure 3A–C).

First, the maximum mean PL introduced earlier is used as the input feature for the k -means clustering algorithm (see Figure 3D). The aforementioned trend that the PCE of the solar cell increases with the maximum of the transient PL is confirmed by the differences in the PCE distributions of the four clusters (see Figure 3G and S4, Supporting Information). Having clustered the maximum mean PL intensity as a single in situ feature, the time of the transient's maximum is introduced as an additional second input feature (see Figure 3B). Examining the two features, the existence of two distinct groups within the dataset is evident (see Figure 3E). For a number of solar cells, the detection of the maximum mean PL is delayed when compared to most others. The solar cells with delayed maximum mean PL are assigned to cluster 0 and display the worst general performance (see Figure 3H and S4, Supporting Information). The group without delayed time of maximum mean PL is grouped into the three remaining clusters depending on the maximum mean PL intensity (see Figure 3H). Cluster 1 to cluster 3 show increasing median and mean PCE. Again, the best-performing

Table 1. Correlation coefficients confirm a higher correlation between the in situ feature and the labels.

a)	PCE	V_{oc}	J_{sc}	FF	Layer thickness
In situ feature	0.3386 ($r^2 = 0.1147$) ^{b)}	0.0496 ($r^2 = 0.0025$)	0.1707 ($r^2 = 0.0291$)	0.4225 ($r^2 = 0.1785$)	−0.6080 ($r^2 = 0.3696$)
Ex situ feature	0.0182 ($r^2 = 0.0003$)	0.1249 ($r^2 = 0.0156$)	−0.0116 ($r^2 = 0.0001$)	−0.0313 ($r^2 = 0.0010$)	0.6091 ($r^2 = 0.3710$)

^{a)} r^2 : coefficient of determination, PCE: power conversion efficiency, V_{oc} : open-circuit voltage, J_{sc} : short-circuit current density, FF: fill factor; ^{b)} Bold values indicate higher correlation for given label.

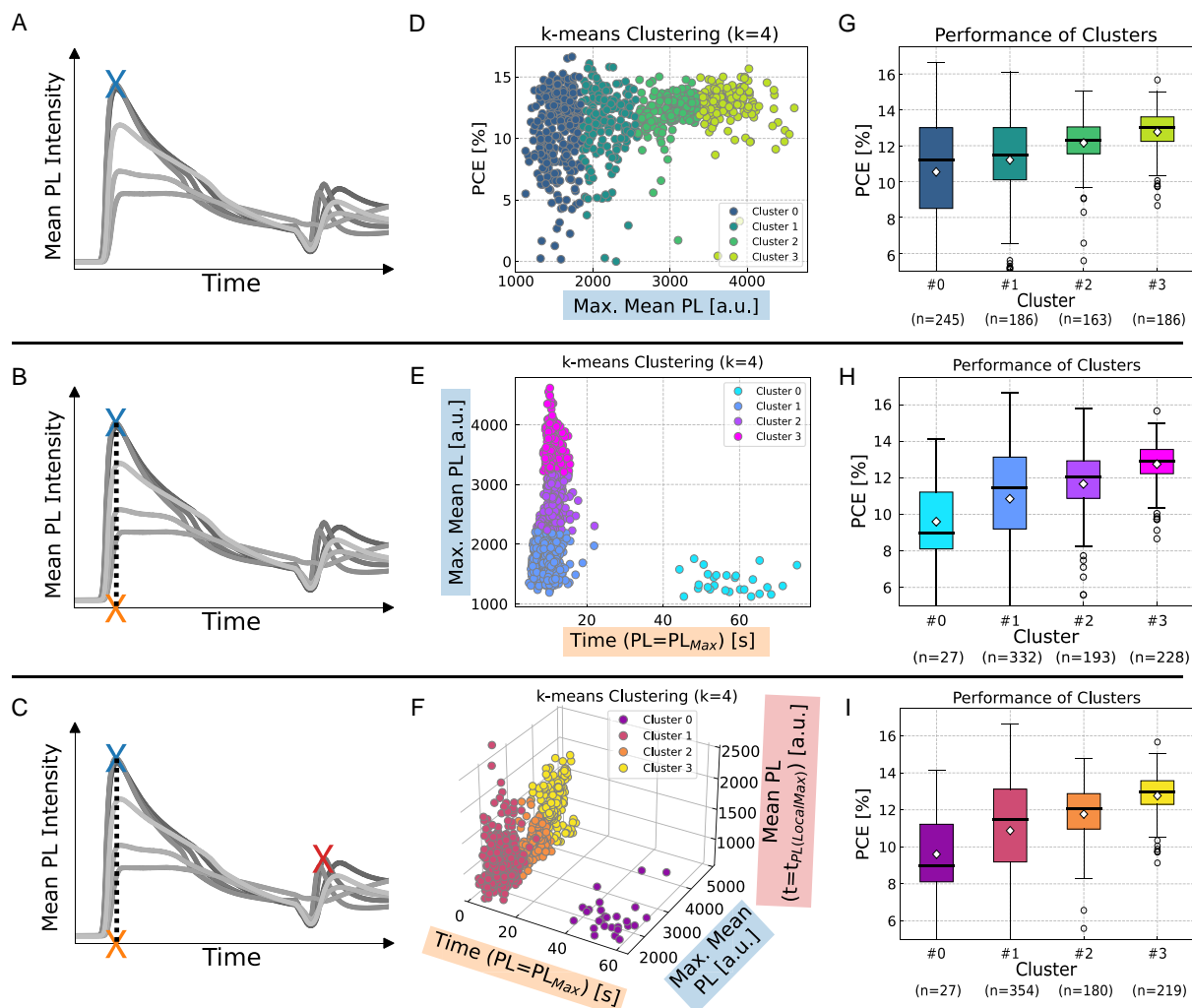


Figure 3. Clustering of human-readable multi-feature data. A–C) Exemplary mean photoluminescence transients. First, one single data point is extracted from the transient and used as feature when clustering the data (A). Afterwards, a second (B) and third data point (C) is added to the model input. D–F) Scatter plots showing each solar cell as a data point. The single feature values are scattered over the PCE (D). The data points are scattered in two-dimensional space when using two features as model input (E) and in 3D space when using three features (F). G–I) The PCE distributions of the resulting clusters are displayed as boxplots for the clustering of one feature (G), two features (H), and three-feature data (I). Mean and median values are indicated by white diamonds and black lines, respectively.

cluster 3 is of low variance even though 228 solar cells are assigned to it.

The addition of a third feature further increases the information contained in the input and allows for potentially better clustering of the data as more information is available. As the number of features increases, the input provides a more holistic representation of the entire data. However, it also leads to a more complex problem, which is harder for humans to interpret, thus requiring the use of machine learning. To cluster three-feature samples, the local maximum of the oscillation of the mean PL transient (see Figure 3C) is added to the existing features. The oscillation is attributed to the start of the venting process, i.e., the opening of the vacuum valve, subsequent to the evacuation time interval and the consequent increase of the pressure in the experimental chamber. While the two distinct groups within the fitted clusters are evident, variance is added through the introduction of the third

feature, especially to the group without delayed PL maxima (see Figure 3F). The borders between different clusters are placed similarly to the two-feature clustering case which leads to a similar assignment of solar cells to the clusters (see Figure 3I).

It is demonstrated that the corresponding general solar cell performances of the generated multi-feature clusters differ substantially. This allows for distinguishing favorable from less favorable properties of the acquired in situ PL transient. Other feature sets of two or three parameters extracted from the mean PL transients, the resulting clusters, and corresponding performance can be found in Figure S5 and S6, Supporting Information.

Plotting and qualitatively interpreting features with more than three parameters is increasingly difficult, which necessitates the use of automated methods that can deal with such a high-dimensional feature space.

2.4. Clustering In Situ Photoluminescence without Human-Encoded Features

By selecting features, the entire curve is boiled down into a small number of values, making it easier for the human expert to interpret the data. However, potentially valuable information is removed by making it human-readable. Therefore, in the following experiment, the entire mean PL transients are used as input for the ML algorithm. Without the need to identify the transients' features with the highest possible information content, no prior decisions are made by a human expert for encoding the data.

Applying k-means clustering to the entire mean PL transient data of the training dataset shows that unsupervised ML can be used to identify in situ transient data patterns which correspond to varying performance of the final solar cells. It is observed that the clusters differ with respect to various characteristics of the transients (see **Figure 4A**). Comparing the transients assigned to the different clusters, a temporal offset of the PL signal onset can be recognized. The transients assigned to cluster 3 show an early onset of the PL signal while the transients' onsets of cluster 0 are considerably delayed. The variation in onset time can be explained by spatial differences in the wet-film thickness (see **Figure 4D** for boxplots of layer thickness information). For areas with thicker wet films and hence more material, the film takes longer to dry, and therefore the PL signal onset is delayed when compared to areas with thinner wet films.^[60] In addition to the PL onset time, the transients assigned to the clusters also differ regarding the height of the initial maximum (see **Figure 4B**).

The transients of cluster 0 have low initial maximal PL peaks. The average peak height increases with cluster number and cluster 3 only contains transients with a high initial peak. The differences in PL peak height can also be explained by different layer thicknesses. However, for the absolute PL intensity also the quenching process plays a major role. The underlying self-assembled monolayer (2PACz) also differs spatially in layer thickness since it was blade-coated as well. This leads to spatial differences in charge carrier extraction and therefore to locally different quenching of the PL.^[60] In addition to differences in absolute PL intensity, the four clusters also differ regarding the relative behavior over time (see **Figure 4C**). The transients of cluster 0 with low absolute PL intensity show merely a small relative decrease over time while the relative decrease after the initial maximum gets steeper with increasing cluster number. Having investigated the differences in the transients' characteristics, the clusters are examined regarding the performance of the corresponding solar cells. It is confirmed that clustering the entire mean PL transient curves can be used for extracting data patterns that correlate with the performance of the full devices. PCE as well as FF increase from cluster 0 to cluster 3 coinciding with higher PL signal peaks and earlier PL signal onsets (see **Figure 4D** and S7, Supporting Information). This suggests a correlation between the PL signal onset time and the transients' peak height with the photovoltaic parameters PCE and FF. Both, mean and median PCE decline from cluster 3 to cluster 0 which manifests a general trend. Similar to the clustering examples shown previously, the PCE variance of the best-performing

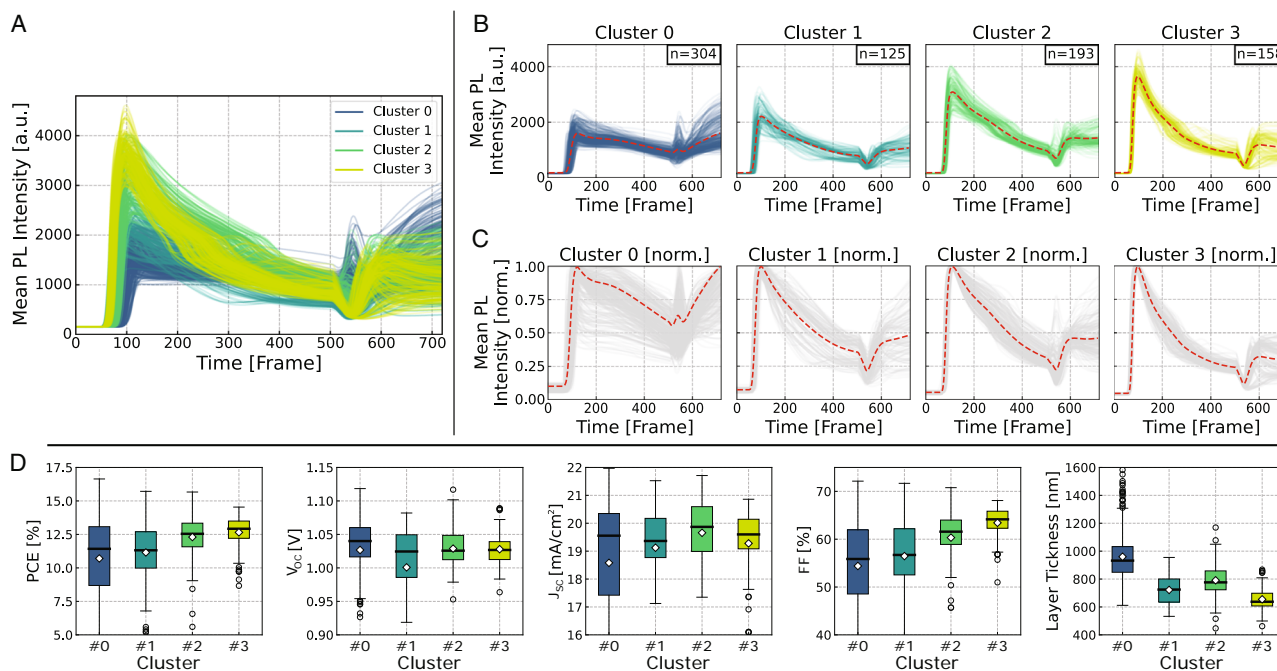


Figure 4. Clustering photoluminescence transients without human input. A) Depiction of the mean PL transients of all the training samples. The transients are color-coded with regard to the assigned cluster. B) Four different diagrams displaying the mean PL transients which were assigned to the four clusters. The mean transient curve of each cluster is illustrated as the dashed red line. C) The four diagrams display the normalized transients assigned to the different clusters. The mean transient curve of each cluster is illustrated as the dashed red line. D) The distributions of the performance parameters of the resulting clusters are displayed. Next to the PCE, also the boxplots displaying the distributions of open-circuit voltage (V_{OC}), short-circuit current (J_{SC}), fill factor (FF), and perovskite layer thickness are shown. Mean and median values are indicated by white diamonds and black lines, respectively.

cluster is small, but it increases for decreasing mean PCE of the clusters. This allows the identification of data patterns in cluster 3 which lead to merely well-performing solar cells and nearly no solar cells of low performance. 87.3% of the solar cells assigned to cluster 3 perform better than the mean PCE of the complete (training) dataset (11.57%) highlighting the above-average performance of the solar cells in cluster 3 (comparable results for previously shown clustering applications are displayed in Figure S8, Supporting Information). The same trend can be observed for the FF in a more pronounced manner. Mean and median FF decline from cluster 3 to cluster 0 as well. Also, 93.7% of the solar cells assigned to cluster 3 show a higher FF than the mean FF of the complete dataset (58.03%). For V_{OC} and J_{SC} a trend is not apparent. However, the median V_{OC} decreases from cluster 0 to cluster 3 (cluster 1 being an outlier from the general trend) which suggests an inverse correlation between V_{OC} and transients' peak height and PL signal onset time. In addition to clustering the mean PL transients, *k*-means can also be applied to the transients of the other channels in the dataset, e.g., reflectance and PL peak energy (see Figure S9 and S10, Supporting Information).

2.5. Spatial Correlation of Clusters Reveals Detrimental Process Mechanism

Upon investigation of the four PL transient clusters, it is evident that the variance of the performance metrics of cluster 0 is much larger compared to the other clusters. This raises the question of whether the PL patterns associated with cluster 0 also correlate with solar cell performance, or whether this subset of the data does not correlate with the PL transients.

To answer this question, cluster 0 is further subdivided by performing a second round of clustering on the data samples (cells) assigned to it. It is observed that assigning the samples to three subclusters allows further differentiation between PL transient patterns with different solar cell performances. The number of subclusters $k_{sub} = 3$ is determined by the implementation of the elbow method (see Figure S11, Supporting Information). To emphasize the importance of the transients' peak height and the PL signal onset time (which were shown to strongly affect the clustering, see Figure 4) the 304 data samples used for the sub-clustering step are truncated by discarding the subsequent chamber venting. It is observed that the PL signal onset time has a profound impact on the result (see Figure 5A). When compared to the PCE distribution of the original cluster 0, the solar cells assigned to cluster 0_a on average perform poorer while the solar cells attributed to cluster 0_c on average showcase better PCEs (see Figure 5B). Both mean and median PCE differ for the three subclusters indicating that the extracted data patterns lead to differences in performance. A comparison of the centroid transients fitted in two rounds of clustering with the centroids of a corresponding single round of clustering with $k = 6$ illustrates the differences in the data patterns found (see Figure S11, Supporting Information). In a second round of clustering, the subset of data is divided into several low-intensity PL subclusters. However, when clustering with $k = 6$ initially, the PL transients are not divided into clusters in the same way, with low-intensity PL transients, in particular, being less sharply resolved. In summary, sub-clustering

confirms the correlation with solar cell performance also for data assigned to cluster 0 (see Figure 5C).

Subsequently, it is shown that the *k*-means clustering detects data patterns indicative of poor solar cell performance, thereby identifying detrimental process mechanisms during perovskite thin-film formation. Having previously examined the fitted clusters in terms of solar cell performance metrics, the position of each solar cell on the substrate during fabrication is used as an additional feature to evaluate the clusters.

Taking into account the additional feature, spatial correlations of the generated clusters of the transient PL patterns and, consequently, of solar cell performance are identified. It is found that the spatial histograms of the different clusters differ considerably. First, an accumulation of solar cells assigned to the cluster with the poorest performance, e.g., cluster 0_a, is found on the right substrate edge (see Figure 5E). Nearly no solar cell which is grouped into cluster 0_a was positioned in the left half of the substrate during the fabrication process. However, over 80% of the solar cells of cluster 0_a were positioned in the rightmost quarter of the large substrate. Investigating the morphology of a blade-coated perovskite thin film (see Figure 5D) reveals an arch-shaped inhomogeneity as a reason for the spatial accumulation of low-performing solar cells. The large inhomogeneity is caused by solution backflow after the coating process (the blade coating applicator is moved from left to right). The resulting spatial differences in the amount of material led to layer thicknesses that deviate from the target layer thickness defined by suitable experimental parameters.

Second, solar cells assigned to the best-performing cluster were also predominantly located in a certain sub-area of the substrate. However, in contrast to cluster 0_a, almost no solar cell assigned to cluster 3 was positioned in the substrate area affected by the material backflow (see Figure 5F). Thickness deviations caused by decreasing the amount of solution available for distribution during coating as well as the material flow towards the upper and lower substrate edges result in the majority of solar cells being located in the center of the leftmost quarter of the substrate. The original clusters reveal the same insight when investigating their spatial histograms in Figure S12, Supporting Information. Investigating the spatial distribution of solar cells' PCE (see Figure S13, Supporting Information) confirms the general worse performance of solar cells positioned in the substrate area affected by the material backflow.

In summary, the potential of unsupervised ML to reveal new insights into the fabrication process is demonstrated by extracting transient PL patterns indicative of poor solar cell performance. Next to correlations with performance metrics and perovskite layer thickness, the transients also encode information about spatial differences in the experiment process and can be used to identify detrimental process mechanisms during the experimental procedure.

2.6. Predictive Capability of Supervised Machine Learning on Previously Unseen Data

Demonstrating the ML model's ability to predict the final solar cell performance parameters based on the in situ PL dataset shows that it can potentially be used for in-the-loop process

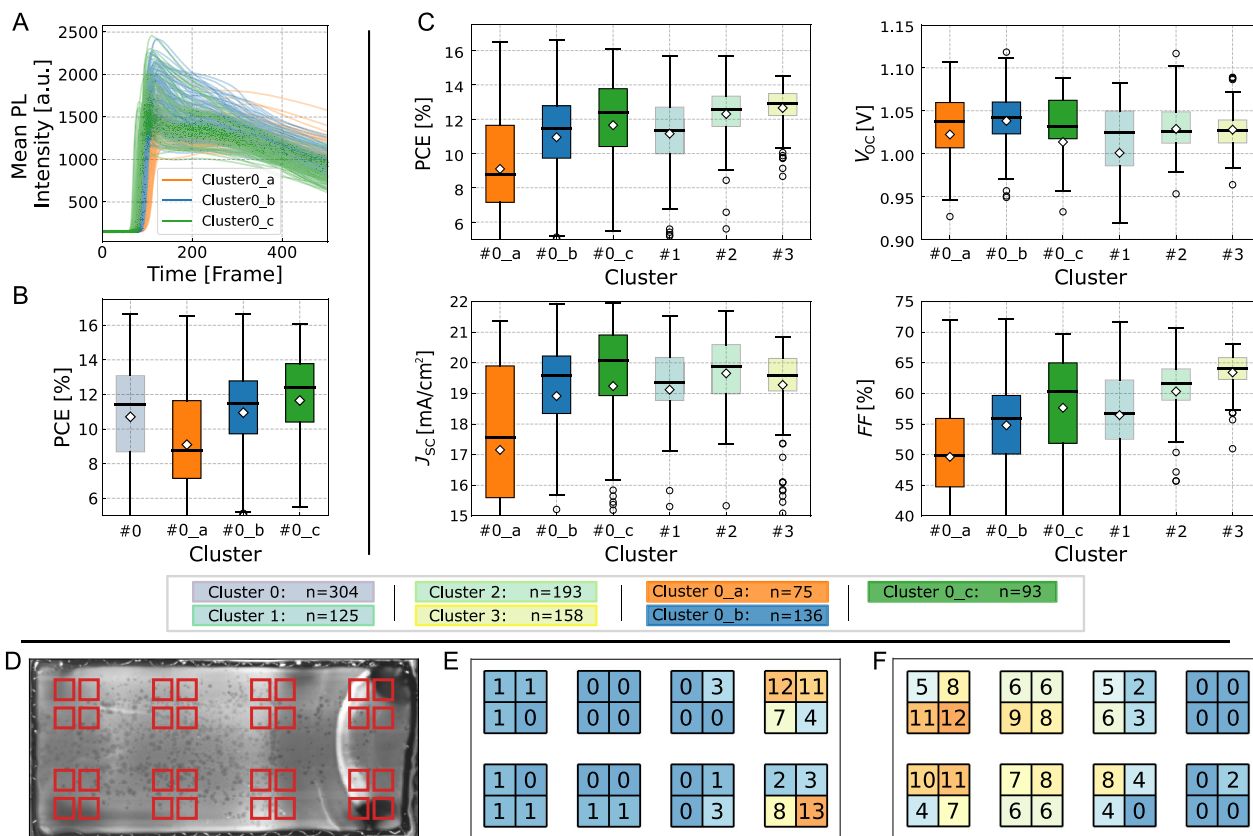


Figure 5. Investigation of the clusters showcases the spatial correlation of clusters. A) The mean PL transients previously assigned to cluster 0 are truncated and sub-clustered. The colors of the curves indicate which subcluster they are assigned to upon second clustering. B) Distribution of the original cluster 0 (box with higher transparency) alongside the distributions of the newly generated subclusters. Mean and median values are indicated by white diamonds and black lines, respectively. C) The distributions of the performance parameters of the three subclusters are displayed as boxplots next to the distributions of the remaining original clusters 1 to 3. D) Image of an exemplary blade-coated perovskite thin-film layer. The blade is moved from the left to the right-hand side. The red rectangles mark the positions of the 32 solar cells located on the substrate. E, F) Spatial histograms represented by heat maps showing the number of solar cells in each of the 32 possible positions on the substrate assigned to a single cluster. The heat maps show the spatial histogram of the solar cells assigned to the cluster performing the poorest, cluster 0_a, (E) and to the best-performing cluster, cluster 3 (F).

monitoring of perovskite thin-film quality. The detection of deviations from the optimal experimental process will be a critical step towards an ML-guided active feedback loop that will allow real-time adjustment of process parameters.

After prior exploration of the dataset using unsupervised ML, the predictive capability of the supervised ML algorithm *k*-nearest neighbors (kNN)^[61] is investigated on previously unseen data in the test set. To predict the target value of each new sample in the test set, kNN determines the *k* samples in the training set which are most similar to the unseen test sample based on a similarity metric. The mean of these *k* most similar training samples is then used as a prediction for the test sample (see Experimental Section for more information).

First, the optimal number of neighbors *k* must be determined for the prediction of each target label. Using fivefold cross-validation on the training set, the quality of the model prediction is assessed for each fold by calculating the mean absolute error (MAE) between the predicted values and the actual measured values. The model's general performance on the validation set is then obtained by averaging

the five MAE values. That way, the optimal value of *k* can be determined without using the data from the test set.

Comparing the cross-validation results highlights substantial reductions in MAE when using the kNN regressor instead of a dummy mean regressor, which always predicts the mean of the current training set (see Table 2). Compared to the dummy mean regressor, the prediction of PCE with the kNN model is more accurate, showing a reduction in MAE from 1.8967% to 1.5408% (absolute), corresponding to a relative decrease in MAE in PCE of 18.76%. While the model does not perform extensively better than the dummy mean regressor in terms of V_{OC} showing an 8.44% improvement from 0.0308 to 0.0282 V, the MAE in J_{SC} is reduced from 1.4262 to 1.2456 mA cm⁻², a 12.66% decrease, and the MAE in FF is reduced by 22.30% (relative) from 6.5632% to 5.0997% (absolute). Reducing the MAE from 137.9262 to 79.2875 nm, the kNN model performs particularly well in predicting the perovskite layer thickness, showing a 42.51% decrease in MAE.

Table 2. Reduction of prediction error using k -nearest neighbors regressor.

a)	Cross-validation on training set			Test set			k
	MAE using dummy mean regressor	MAE using kNN	Reduction of prediction error	MAE using dummy mean regressor	MAE using kNN	Reduction of prediction error	
PCE [%]	1.8967	1.5408	-18.76%	1.9354	1.5156	-21.69% ^{b)}	14
V_{OC} [V]	0.0308	0.0282	-8.44%	0.0357	0.0296	-17.09%	24
J_{SC} [mA cm^{-2}]	1.4262	1.2456	-12.66%	1.3742	1.1400	-17.04%	15
FF [%]	6.5632	5.0997	-22.30%	6.553	5.0983	-22.20%	14
Layer thickness [nm]	137.9262	79.2875	-42.51%	139.7959	67.6807	-51.59%	18

^{a)}kNN: k -nearest neighbors, MAE: mean absolute error, PCE: power conversion efficiency, V_{OC} : open-circuit voltage, J_{SC} : short-circuit current density, FF: fill factor;
^{b)}Bold values indicate prediction improvement on previously unseen data.

Second, the optimized models are evaluated on previously unseen data in the test dataset. After determining k using only the training set, the kNN models are applied to the test set. Compared to the dummy mean regressor, i.e., always predicting the mean of the entire training dataset, the model's prediction accuracy improved substantially, showing smaller MAEs (see Table 2). It is highlighted that the prediction accuracy of PCE and FF are improved by 21.69% (relative) from 1.9354%

to 1.5156% (absolute) and by 22.20% (relative) from 6.553% to 5.0983% (absolute), respectively. In addition, the MAE for predicting V_{OC} decreased by 17.09% from 0.0357 to 0.0296 V and for J_{SC} by 17.04% from 1.3742 to 1.1400 mA cm^{-2} . Again, the improvement in accuracy is highest in the prediction of layer thickness, showing a 51.59% reduction in MAE for the test data with kNN compared to the dummy mean regressor, corresponding to a reduction from 139.7959 to 67.6807 nm.

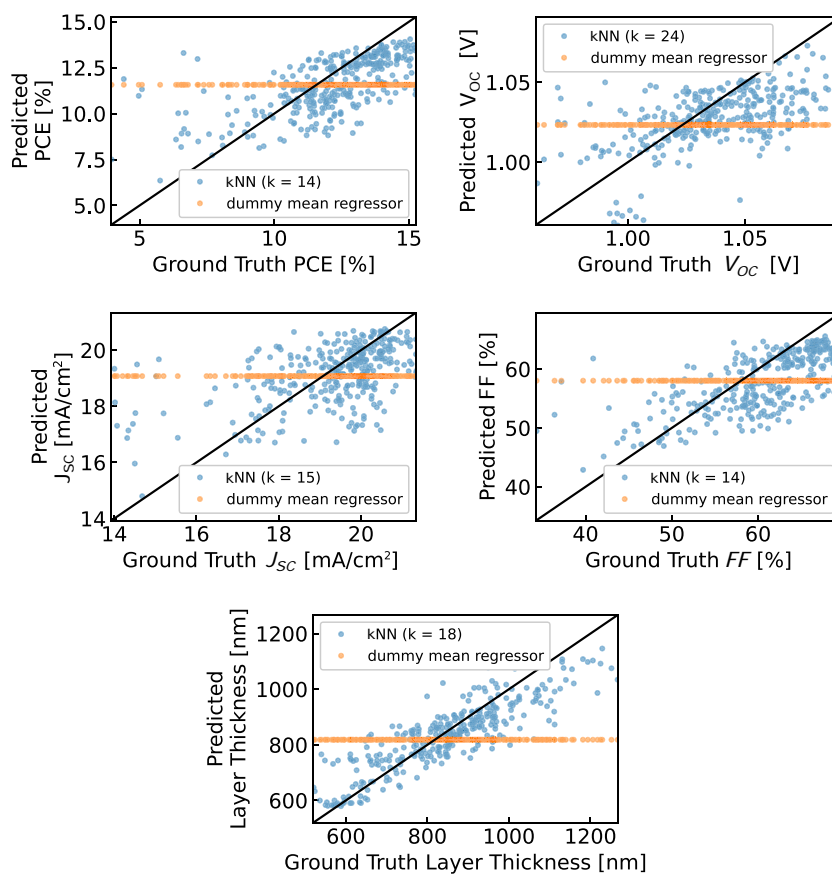


Figure 6. Showcase of the predictive capability of the kNN algorithm. Diagrams comparing the predicted values of the test set samples for each label with the ground truth. The diagonal black lines indicate where the predictions would perfectly match the ground truth. The dummy mean regressor simply predicts the training set mean of the label for all the test set samples. For each label, the optimal number of neighbors k is determined on the training set using cross-validation. For illustration purposes, the limits of the axes were chosen in a way that the highest and lowest two percent of the data are excluded (see all data in Figure S14, Supporting Information).

The visualization of the prediction results on the test set underlines that the kNN models achieve a considerably higher prediction accuracy than the dummy mean regressor (see Figure 6). For illustration purposes, the limits of the axes were chosen such that the highest and lowest two percent of the data were removed (see the visualization of all test data in Figure S14, Supporting Information). In addition to making predictions based on the mean PL transients, kNN can also be applied to the transients of the other channels in the dataset, e.g., reflectance and PL peak energy (see Figure S15, Supporting Information). There is a tendency for high target values to be underestimated, while low target values are overestimated. This motivates further investigations into the dataset employing more elaborate ML algorithms without prior feature extraction. Here, the implementation of convolutional neural networks using the entire acquired 2D + t data offers a promising way to improve prediction accuracy.

However, the prediction results achieved using the supervised kNN demonstrate that the dataset containing in situ PL data can be used to predict perovskite layer thickness and performance parameters of solar cells. This represents a critical step toward in-line process monitoring enabled by ML-based solar cell performance prediction, which will help accelerate the upscaling of perovskite PV technology.

3. Conclusion

This work reports on ML for process understanding of the scalable perovskite thin-film formation by generating a unique in situ PL dataset and analyzing it with ML. The analysis of the perovskite layer formation using ML enables the early identification of unintended variations across iterations during device fabrication. To this end, we introduce a dataset containing multi-channel PL image time series of 1129 PSCs acquired in situ during the vacuum quenching of the perovskite layer and the corresponding solar cell performance metrics as well as the perovskite layer thickness as labels. After generating transients from the image time series, first, the correlation between an expert-chosen in situ PL feature and the corresponding solar cell performance highlights the advantage of acquiring in situ data during the perovskite formation compared to ex situ PL data. For an initial exploration of the data, prominent features of the PL transients are selected and used as input for the unsupervised ML algorithm *k*-means clustering. This manual selection of features by human experts introduces bias into the analysis. Hence, in the next step, entire PL transients are clustered to reduce the dependency on human expert input. It is shown that *k*-means clustering generates clusters containing different PL transient patterns which correlate with the performance of the final PSC. Moreover, spatial correlations of the generated clusters of PL transients and, consequently, of solar cell performance are identified. Substrate areas with unfavorable perovskite thin-film properties are detected displaying the model's ability to detect detrimental process mechanisms during the experimental procedure. Finally, the supervised ML technique kNN is applied to unseen test data for tentative predictions of solar cell performance and perovskite layer thickness. The promising prediction results motivate further investigations into the realization of

ML-based in-line performance prediction for PSCs. Here, the application of more sophisticated ML algorithms, such as convolutional neural networks, using the entire raw data including spatial information provides an opportunity to further improve prediction accuracy. In summary, this work demonstrates that ML-based analysis of in situ PL data has a high potential for in-line processing monitoring and can accelerate the successful commercialization of perovskite thin-film PV.

4. Experimental Section

PSC Fabrication: The used materials and the fabrication process are based on the ones described in detail in the authors' group's previous work described in Ref. [60].

Perovskite Ink Fabrication: For the fabrication of the solar cells, the double cation perovskite (DCP) composition $\text{Cs}_{0.17}\text{FA}_{0.83}\text{Pb}(\text{I}_{0.91}\text{Br}_{0.09})_3$ was used. An ink was prepared by dissolving PbI_2 (0.875 M, TCI Chemicals), and PbBr_2 (0.125 M, TCI Chemicals) in a mix of *N,N*-dimethylformamide (DMF, anhydrous, Sigma-Aldrich), dimethyl sulfoxide (DMSO, anhydrous, Sigma-Aldrich) in a ratio 4:1 (vol%). Afterward, the PbX_2 solution was added to $\text{CH}(\text{NH}_2)_2\text{I}$ (FAI, 0.825 M, GreatCell Solar) and CsI (0.175 M, abcr) and then diluted 2:1 (vol%) with γ -butyrolactone (GBL, Sigma-Aldrich). Before deposition, 2.4 vol% *L*- α -phosphatidylcholine (Sigma Aldrich) solution (0.5 mg mL⁻¹ in DMSO) was added.

Solar Cell Fabrication: The glass substrates with pre-patterned ITO (Luminescence Technology) were cleaned in acetone and isopropanol in an ultrasonic bath for 15 and 5 min, respectively, and then cleaned using an oxygen plasma for 3 min. Afterward, NiO_x was sputtered as a 10 nm thick hole transport layer (NiO_x target by Kurt J. Lesker Company, 99.995% metallic purity, see Ref.[62] for more details). Then, a 1 min low power oxygen plasma was applied before blade coating a 2PACz solution (>98%, TCI Chemicals, 1.5 mg mL⁻¹ in ethanol) on top. For blade coating, a Zehntner ZAA 2300.H automatic film applicator and a ZUA 2000 universal applicator were used with a blading gap of 100 μm . 16 μL 2PACz solution were blade-coated onto the 32 \times 64 mm² substrate. The substrate was coated twice in the forward direction at a blade speed of 16 mm s⁻¹ and afterward annealed for 10 min at 100 °C. For the blade coating of the perovskite layer, the parameters were changed to 25 μL ink volume and 25 mm s⁻¹ blading speed. After the deposition of the perovskite layer, the samples were placed in a self-built vacuum chamber (see Ref.[60] for more details) which was then evacuated for 3 min. After completion of the vacuum-quenching process, the chamber was vented with ambient air and the samples were annealed for 30 min at 150 °C. All blade coating and successive annealing steps were performed in ambient conditions of ≈ 21 °C and 45% relative humidity. The large samples were then cut into eight 16 \times 16 mm² samples. To finalize the devices, a 25 nm C60 fullerene (Sigma Aldrich, 98%) electron transport layer, a 5 nm BCP (Luminescence Technology) interfacial layer, and 100 nm silver back-contact were deposited by thermal evaporation. Through the usage of a shadow mask during deposition of the back contact, each sample yields four cells with an active area of 10.5 mm² per solar cell.

Photoluminescence Imaging: The used in situ PL imaging system is based on the setup introduced in our previous work.^[50] A monochrome sCMOS camera (CS2100M-USB Quantalux, 1,920 \times 1,080 pixels, Thorlabs) was equipped with a lens (MVL25M23, Thorlabs) and a wheel loaded with four different filters was placed between the camera and the samples. A microcontroller was used to synchronize the camera's trigger (10 ms exposure time) and the filter wheel's rotation (180 rpm). The filter wheel was loaded with: 1) a 725 nm long pass (Edmund Optics, stacked below a 620 nm long pass, RG620), 2) a 780 nm long pass (RG780, Thorlabs, stacked on top of a 715 nm long pass, RG715, Thorlabs), 3) a 775 nm short pass combined with a 665 nm long pass (Edmund Optics and RG665, Thorlabs), and 4) a neutral density filter with adjustable transmittance (two stacked linear polarizers LPVISE200-A, Thorlabs). For excitation, two blue LED bars (LDL2, 146X30BL2-WD, CCS Inc.) with a center wavelength of 467 nm were mounted in parallel and tilted towards

each other, enabling illumination of the samples (≈ 0.08 suns) without visible reflections of the LED bars in the images.

Solar Cell Characterization: Current-density–voltage characteristics (J – V) of the solar cells were measured using a class AAA 21-channel LED solar simulator (Wavelabs Solar Metrology Systems Sinus-70) under AM1.5G spectrum (100 mW cm^{-2}) in a nitrogen atmosphere. The intensity was calibrated using a silicon reference solar cell filtered with a KG5 band pass (Newport). The J – V scans of the cells were measured in backward and forward directions with a shadow mask (aperture size 7.84 mm^2) using a scanning rate of circa 0.6 V s^{-1} (Keithley 2400 source measurement unit). Using a Peltier element controlled by a microcontroller, the temperature of the solar cells was kept constant at $25 \text{ }^\circ\text{C}$.

Machine Learning Methods: ML methods can be categorized into: i) supervised and ii) unsupervised learning algorithms, which differ in the type of data they receive as input for the training of the model.

In unsupervised learning, ML models are trained without any labels of the input data. This approach is of high interest for data exploration since it can identify patterns in the data without the often cumbersome process of prior data annotation and therefore help with finding the underlying structures of the dataset. Clustering is a commonly used unsupervised learning approach that separates samples in a dataset into groups of similar properties. The identified clusters can then be interpreted and analyzed further to identify relevant information in the dataset.^[13,14,58] The most widely adapted clustering technique is the k -means^[56] algorithm. It divides the unlabeled (training) data into k clusters, defined by a cluster centroid, such that each sample is assigned to the cluster with the nearest centroid. The cluster centroids are learned in an iterative process that minimizes the squared sum of distances between each sample and its corresponding centroid.^[57] The elbow method,^[59] a heuristic which plots the sum of squared error as a function of the number of clusters, can be used to choose a suitable number of clusters k by selecting the elbow of the plot. In this work, the unsupervised ML method k -means clustering is employed for the exploration and analyses of the (training) dataset.

Supervised ML models are trained on labeled data. Using a set of data samples (training set), the model then learns the mapping between the input feature and the given label. When provided with a new input sample (test set), the trained model can predict the target for the new input data. The kNN^[61] algorithm is a simple supervised ML technique, which is also distance-based like the unsupervised k -means clustering. kNN assumes that similar samples exist at close distances to each other. Therefore, it predicts a value for each sample in the test set based on the k examples in the training set which are closest (most similar) to the test sample. To compute the distance, the default distance metric (Euclidean distance) implemented in scikit-learn's^[63] KNeighborsRegressor is used in this work. For regression problems, kNN then returns the mean of the k closest training examples as the prediction for each test sample. The optimal k can be determined using cross-validation on the training set. In this work, the kNN is employed for predictions of the solar cell performance metrics and the perovskite layer thickness.

To enable an unbiased evaluation of the trained model on the test set, the concept of cross-validation is applied to train and optimize the model only on the training set. To implement fivefold cross-validation, the training dataset is divided into five subsets, assigning groups of solar cells of the same substrate to the same data subset (per-substrate stratification). This allows training and optimizing the ML models on four out of the five subsets while the fifth subset is used for validation. This process is repeated five times using each subset for validation once while the four remaining subsets are used for training. Finally, the previously unseen data in the test set is used only to evaluate the model which has been trained and optimized on the training set using cross-validation.

Computational Methods: All ML models presented in this study were built using the scikit-learn (1.0.2)^[63] library in Python (3.8.6).^[64] The data was preprocessed using scaling algorithms provided by scikit-learn. The code was written with the additional Python packages NumPy (1.22.3),^[65] pandas (1.4.1),^[66] SciPy (1.8.0),^[67] h5py (3.6.0),^[68] matplotlib (3.5.1),^[69] PyTorch (1.11.0),^[70] and TorchMetrics (0.7.0).^[71] Furthermore, the packages tiffle (2021.3.4),^[72] OpenCV (4.5.1.48),^[73] and Pillow (9.0.1)^[74] were used for preprocessing of the PL images.

The computational experiments were run on the bwUniCluster 2.0+GFB-HPC cluster system located at the Steinbuch Centre for Computing at Karlsruhe Institute of Technology.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The financial support by Helmholtz Association through AI-InSu-Pero (ZT-I-PF-5-106), the Bundesministerium für Wirtschaft und Klimaschutz through the project CAPITANO (3EE1038B), the Initiating and Networking Funding of the Helmholtz Association (HYIG of U.W.P. (VH-NG-1148)) as well as the Karlsruhe School of Optics & Photonics (KSOP) is gratefully acknowledged. This work is supported by the Helmholtz Association Initiative and Networking Fund under the Helmholtz AI platform grant. The authors acknowledge support from the state of Baden-Württemberg through bwHPC. Part of this work was funded by Helmholtz Imaging, a platform of the Helmholtz Incubator on Information and Data Science. The authors express their gratitude to Simon Ternes for jointly developing the initial version of the imaging setup. The authors would like to thank the members of the Perovskite Taskforce at KIT.

Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

U.W.P. and F.L.: Conceptualization; F.S.: Methodology; F.L., F.S., and E.A.M.V.: Investigation; F.L. and S.Z.: Data Curation; F.L. and S.Z.: Software; F.L., and S.Z.: Formal Analysis; S.Z., M.G., C.D., and F.I.: Validation; F.L.: Writing – Original Draft; S.Z., F.S., M.G., C.D., F.I., and U.W.P.: Writing – Review & Editing; F.L.: Visualization; M.G., C.D., F.I., and U.W.P.: Project Administration; M.G., C.D., F.I., and U.W.P.: Funding Acquisition; M.G., F.I. and U.W.P.: Resources; C.D., F.I., and U.W.P.: Supervision.

Data Availability Statement

The generated dataset (<https://doi.org/10.5281/zenodo.7503391>)^[55] as well as the python code (<https://github.com/AI-InSu-Pero/ML-PerovskitePV-InSituLuminescence>) generated for this work are made publicly available to the community.

Keywords

clustering, datasets, in situ characterization, machine learning, performance prediction, perovskite solar cells, process monitoring

Received: January 10, 2023
Published online: February 16, 2023

- [1] H. J. Snaith, *J. Phys. Chem. Lett.* **2013**, *4*, 3623.
[2] J. P. Correa-Baena, M. Saliba, T. Buonassisi, M. Grätzel, A. Abate, W. Tress, A. Hagfeldt, *Science* **2017**, *358*, 739.

- [3] National Renewable Energy Laboratory (NREL), Best Research-Cell Efficiency Chart. <https://www.nrel.gov/pv/cell-efficiency.html> (accessed: January 2023).
- [4] I. A. Howard, T. Abzieher, I. M. Hossain, H. Eggers, F. Schackmar, S. Ternes, B. S. Richards, U. Lemmer, U. W. Paetzold, *Adv. Mater.* **2019**, *31*, 1806702.
- [5] K. Rajan, *Mater. Today* **2005**, *8*, 38.
- [6] J. P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, T. Buonassisi, *Joule* **2018**, *2*, 1410.
- [7] A. Agrawal, A. Choudhary, *APL Mater.* **2016**, *4*, 053208.
- [8] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
- [9] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.
- [10] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakithodi, C. Kim, *npj Comput. Mater.* **2017**, *3*, 54.
- [11] J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, *npj Comput. Mater.* **2019**, *5*, 83.
- [12] Q. Tao, P. Xu, M. Li, W. Lu, *npj Comput. Mater.* **2021**, *7*, 23.
- [13] L. Zhang, M. He, S. Shao, *Nano Energy* **2020**, *78*, 105380.
- [14] M. Srivastava, J. M. Howard, T. Gong, M. Rebello Sousa Dias, M. S. Leite, *J. Phys. Chem. Lett.* **2021**, *12*, 7866.
- [15] R. E. Kumar, A. Tiisonen, S. Sun, D. P. Fenning, Z. Liu, T. Buonassisi, *Matter* **2021**, *5*, 1353.
- [16] K. Takahashi, L. Takahashi, I. Miyazato, Y. Tanaka, *ACS Photonics* **2018**, *5*, 771.
- [17] G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, *Front. Mater.* **2016**, *3*, 19.
- [18] J. Kirman, A. Johnston, D. A. Kuntz, M. Askerka, Y. Gao, P. Todorović, D. Ma, G. G. Privé, E. H. Sargent, *Matter* **2020**, *2*, 938.
- [19] J. Im, S. Lee, T. W. Ko, H. W. Kim, Y. K. Hyon, H. Chang, *npj Comput. Mater.* **2019**, *5*, 37.
- [20] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, *Nat. Commun.* **2018**, *9*, 3405.
- [21] X. Cai, Y. Zhang, Z. Shi, Y. Chen, Y. Xia, A. Yu, Y. Xu, F. Xie, H. Shao, H. Zhu, D. Fu, Y. Zhan, H. Zhang, *Adv. Sci.* **2022**, *9*, 2103648.
- [22] Y. Zhao, J. Zhang, Z. Xu, S. Sun, S. Langner, N. T. P. Hartono, T. Heumueller, Y. Hou, J. Elia, N. Li, G. J. Matt, X. Du, W. Meng, A. Osvet, K. Zhang, T. Stubhan, Y. Feng, J. Hauch, E. H. Sargent, T. Buonassisi, C. J. Brabec, *Nat. Commun.* **2021**, *12*, 2191.
- [23] S. Sun, A. Tiisonen, F. Oviedo, Z. Liu, J. Thapa, Y. Zhao, N. T. P. Hartono, A. Goyal, T. Heumueller, C. Batali, A. Encinas, J. J. Yoo, R. Li, Z. Ren, I. Marius Peters, C. J. Brabec, M. G. Bawendi, V. Stevanovic, J. Fisher, T. Buonassisi, *Matter* **2021**, *4*, 1305.
- [24] Ç. Odabaşı, R. Yıldırım, *Sol. Energy Mater. Sol. Cells* **2020**, *205*, 110284.
- [25] R. J. Stoddard, W. A. Dunlap-Shohl, H. Qiao, Y. Meng, W. F. Kau, H. W. Hillhouse, *ACS Energy Lett.* **2020**, *5*, 946.
- [26] J. M. Howard, Q. Wang, E. Lee, R. Lahoti, T. Gong, M. Srivastava, A. Abate, M. S. Leite, *J. Phys. Chem. Lett.* **2020**, *13*, 2254.
- [27] J. M. Howard, Q. Wang, M. Srivastava, T. Gong, E. Lee, A. Abate, M. S. Leite, *J. Phys. Chem. Lett.* **2022**, *2022*, 2254.
- [28] W. A. Saidi, W. Shadid, I. E. Castelli, *npj Comput. Mater.* **2020**, *6*, 36.
- [29] G. Pilania, A. Mannodi-Kanakithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2016**, *6*, 19375.
- [30] R. Li, Q. Deng, D. Tian, D. Zhu, B. Lin, *Crystal* **2021**, *11*, 818.
- [31] P. Omprakash, B. Manikandan, A. Sandeep, R. Shrivastava, P. Viswesh, D. B. Panemangalore, *Comput. Mater. Sci.* **2021**, *196*, 110530.
- [32] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **2012**, *58*, 227.
- [33] D. D. Landis, J. S. Hummelshøj, S. Nestorov, J. Greeley, M. Dultzak, T. Bligaard, J. K. Nørskov, K. W. Jacobsen, *Comput. Sci. Eng.* **2012**, *14*, 51.
- [34] Y. Cai, W. Xie, Y. T. Teng, P. C. Harikesh, B. Ghosh, P. Huck, K. A. Persson, N. Mathews, S. G. Mhaisalkar, M. Sherburne, M. Asta, *Chem. Mater.* **2019**, *31*, 5392.
- [35] E. I. Marchenko, S. A. Fateev, A. A. Petrov, V. V. Korolev, A. Mitrofanov, A. V. Petrov, E. A. Goodilin, A. B. Tarasov, *Chem. Mater.* **2020**, *32*, 7383.
- [36] T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli, A. Vijayan, A. Kulkarni, A. Y. Anderson, B. Primera Darwich, B. Yang, B. L. Coles, C. A. R. Perini, C. Rehermann, D. Ramirez, D. Fairen-Jimenez, D. Di Girolamo, D. Jia, E. Avila, E. J. Juarez-Perez, F. Baumann, F. Mathies, G. S. Anaya González, G. Boschloo, G. Nasti, G. Paramasivam, G. Martínez-Denegri, H. Näsström, H. Michaels, H. Köbler, H. Wu, I. Benesperi, M. Ibrahim Dar, I. B. Pehlivan, I. E. Gould, et al., *Nat. Energy* **2021**, *7*, 107.
- [37] L. Bommers, M. Hoffmann, C. Buerhop-Lutz, T. Pickel, J. Hauch, C. Brabec, A. Maier, I. M. Peters, *Prog. Photovoltaics: Res. Appl.* **2021**, *30*, 597.
- [38] L. Bommers, T. Pickel, C. Buerhop-Lutz, J. Hauch, C. Brabec, I. M. Peters, *Prog. Photovoltaics Res. Appl.* **2021**, *29*, 1236.
- [39] Z. Abdullah-Vetter, Y. Buratti, P. Dwivedi, A. Sowmya, T. Trupke, Z. Hameiri, in *Conf. Record of the IEEE Photovoltaic Specialists Conf.*, Institute of Electrical and Electronics Engineers Inc., Fort Lauderdale, FL, USA **2021**, pp. 745–749.
- [40] Y. Buratti, A. Sowmya, R. Evans, T. Trupke, Z. Hameiri, *Prog. Photovoltaics Res. Appl.* **2022**, *30*, 276.
- [41] Y. Buratti, Z. Abdullah-Vetter, A. Sowmya, T. Trupke, Z. Hameiri, in *Conf. Rec. IEEE Photovoltaics Specialist Conf.*, Fort Lauderdale, FL, USA **2021**, pp. 97–100.
- [42] U. Otamendi, I. Martinez, M. Quartulli, I. G. Olaizola, E. Viles, W. Cambarau, *Sol. Energy* **2021**, *220*, 914.
- [43] V. Kumar, P. Maheshwari, *Prog. Photovoltaics Res. Appl.* **2021**, *30*, 880.
- [44] A. M. Karimi, J. S. Fada, M. A. Hossain, S. Yang, T. J. Peshek, J. L. Braid, R. H. French, *IEEE J. Photovoltaics* **2019**, *9*, 1324.
- [45] M. R. Ahan, A. Nambi, T. Ganu, D. Nahata, S. Kalyanaraman, in *SenSys 2021 – Proc. of the 2021 19th ACM Conf. on Embedded Networked Sensor Systems*, Coimbra, Portugal **2021**, pp. 485.
- [46] K. Schötz, F. Panzer, *J. Phys. Chem. A* **2021**, *125*, 2209.
- [47] K. Suchan, J. Just, P. Becker, E. L. Unger, T. Unold, *J. Mater. Chem. A* **2020**, *8*, 10439.
- [48] J. J. van Franeker, K. H. Hendriks, B. J. Bruijners, M. W., G. M. Verhoeven, M. M. Wienk, R. A., J. Janssen, J. J. van Franeker, K. H. Hendriks, B. J. Bruijners, M. M. Wienk, R. A. J. Janssen, *Adv. Energy Mater.* **2017**, *7*, 1601822.
- [49] F. Babbe, C. M. Sutter-Fella, *Adv. Energy Mater.* **2020**, *10*, 1903587.
- [50] S. Ternes, F. Laufer, P. Scharfer, W. Schabel, B. S. Richards, I. A. Howard, U. W. Paetzold, *Sol. RRL* **2021**, *6*, 2100353.
- [51] B. Abdollahi Nejand, D. B. Ritzer, H. Hu, F. Schackmar, S. Moghadamzadeh, T. Feeny, R. Singh, F. Laufer, R. Schmager, R. Azmi, M. Kaiser, T. Abzieher, S. Gharibzadeh, E. Ahlswede, U. Lemmer, B. S. Richards, U. W. Paetzold, *Nat. Energy* **2022**, *7*, 620.
- [52] F. Schackmar, H. Eggers, M. Frericks, B. S. Richards, U. Lemmer, G. Hernandez-Sosa, U. W. Paetzold, *Adv. Mater. Technol.* **2021**, *6*, 2000271.

- [53] H. Eggers, F. Schackmar, T. Abzieher, Q. Sun, U. Lemmer, Y. Vaynzof, B. S. Richards, G. Hernandez-Sosa, U. W. Paetzold, *Adv. Energy Mater.* **2020**, *10*, 1903184.
- [54] B. Chen, J. Peng, H. Shen, T. Duong, D. Walter, S. Johnston, M. M. Al-Jassim, K. J. Weber, T. P. White, K. R. Catchpole, D. Macdonald, H. T. Nguyen, *Adv. Energy Mater.* **2019**, *9*, 1802790.
- [55] F. Laufer, S. Ziegler, F. Schackmar, E. A. Moreno Viteri, M. Götz, C. Debus, F. Isensee, U. W. Paetzold, In Situ Photoluminescence Imaging Dataset of Blade-Coated Perovskite Photovoltaics (v1.0) [Data set], Zenodo, **2023**, <https://doi.org/10.5281/zenodo.7503391>.
- [56] J. MacQueen, *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1967**, p. 281.
- [57] Y. Liu, O. C. Esan, Z. Pan, L. An, *Energy AI* **2021**, *3*, 100049.
- [58] F. Häse, L. M. Roch, P. Friederich, A. Aspuru-Guzik, *Nat. Commun.* **2020**, *11*, 4587.
- [59] R. L. Thorndike, *Psychometrika* **1953**, *184*, 267.
- [60] F. Schackmar, F. Laufer, R. Singh, A. Farag, H. Eggers, S. Gharibzadeh, B. Abdollahi Nejad, U. Lemmer, G. Hernandez-Sosa, U. W. Paetzold, *Adv. Mater. Technol.* **2022**, 2201331.
- [61] N. S. Altman, *Am. Stat.* **1992**, *46*, 175.
- [62] A. Farag, R. Schmager, P. Fassl, P. Noack, B. Wattenberg, T. Dippell, U. W. Paetzold, *ACS Appl. Energy Mater* **2022**, *5*, 6700.
- [63] Scikit-learn: machine learning in Python <https://scikit-learn.org/stable/#> (accessed: January 2023).
- [64] Python <https://www.python.org/> (accessed: January 2023).
- [65] NumPy: The fundamental package for scientific computing with Python <https://numpy.org/> (accessed: January 2023).
- [66] pandas: Python Data Analysis Library <https://pandas.pydata.org/> (accessed: January 2023).
- [67] SciPy: Fundamental algorithms for scientific computing in Python <https://scipy.org/> (accessed: January 2023).
- [68] The HDF5 Library & File Format: High-performance data management and storage suite <https://www.hdfgroup.org/solutions/hdf5/> (accessed: January 2023).
- [69] Matplotlib: Visualization with Python <https://matplotlib.org/> (accessed: January 2023).
- [70] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, S. Chintala, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, Curran Associates, Inc, Vancouver, BC, Canada, 8–14 December **2019**.
- [71] Torchmetrics: <https://torchmetrics.readthedocs.io/en/stable/> (accessed: January 2023).
- [72] TiffFile: Read and write TIFF files <https://pypi.org/project/tiffFile/> (accessed: January 2023).
- [73] OpenCV: Computer Vision library <https://opencv.org/> (accessed: January 2023).
- [74] Pillow: Python Imaging Library <https://pypi.org/project/Pillow/> (accessed: January 2023).