DeepDeg: Forecasting and explaining degradation in novel photovoltaics

Felipe Oviedo[1,2], David S. Hayden[1], Thomas Heumeuller[5,6], Jonas Wortmann[5,6], Jose Dario Perea[4], Richa Naik[1], Hansong Xue[3], Juan Lavista[2], John Fisher III[1], Christoph J Brabec[5,6] and Tonio Buonassisi[1]

[1]Massachusetts Institute of Technology, Cambridge, 02139, MA, USA.
[2]AI for Good, Microsoft, Redmond, 98052, WA, USA.
[3]Solar Energy Research Institute of Singapore (SERIS), National University of Singapore, 7 Engineering Drive 1, Singapore, 117574, Singapore.
[4]Departamento Ingeneria Bioquimica, Universidad ICESI, Cali, Colombia.
[5]Helmholtz-Institute Erlangen-Nürnberg (HI ERN), High-Throughput Photovoltaics (HTPV), Erlangen, Germany.
[6]Department of Materials Science and Engineering, Institute of Materials for Electronics and Energy Technology (i-MEET), Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

Abstract

Degradation is a technical and market hurdle in the development of novel photovoltaics and other energy devices. Understanding and addressing degradation requires complex, time-consuming measurements on multiple samples. To address this challenge, we present *DeepDeg*, a machine learning model that combines deep learning, explainable machine learning, and physical modeling to: 1) forecast hundreds of hours of degradation, and 2) explain degradation in novel photovoltaics. Using a large and diverse dataset of over 785 stability tests of organic solar cells, totaling 230,000 measurement hours, DeepDeg is able to accurately predict degradation dynamics and explain the physiochemical factors driving them using few initial hours of degradation. We use cross-validation and a held-out dataset of over 9,000 hours of degradation of PCE10:OIDTBR to evaluate our model. We demonstrate that by using DeepDeg, degradation characterization and screening can be accelerated by 5-20x.

Keywords: solar cells, degradation, stability, energy devices, machine learning, forecasting, explainability, photovoltaics

One major bottleneck for novel energy devices, such as photovoltaics (PV) or batteries, is degradation. Energy devices must operate under stringent environmental conditions during many years to be competitive technologies. For example, commercial PV modules have very low degradation rates around 0.5% to 2.0% per year [1, 2], and novel PV requires similarly low degradation rates. Nevertheless, the degradation rates of novel PV such as perovskites and organic photovoltaics (OPV) are significantly higher [3, 4]. The photoactive materials in these devices are sensitive to environmental conditions, and their interaction with other parts of the device introduces additional degradation pathways [3, 5, 6]. Traditionally, controlled experiments have been used to identify and mitigate degradation. This approach is limited in novel PV by a slow learning rate and reduced generalizable insights. Novel batteries and fuel cell technologies share similar challenges [7–9]: all have stringent long-term performance requirements, require complex and time-consuming characterization, and present hard-to-identify degradation dynamics.

Recent advances in machine learning (ML) have allowed accelerated measurement and mitigation of degradation [10, 11]. In PV, ML has been used to forecast the degradation of figures of merit of solar cells, such as power conversion efficiency, mainly in crystalline silicon solar cells and modules [12, 13], along with some discrete instances of novel PV technologies [14–17]. These approaches rely either on modelling a single figure of merit and having adequate understanding of degradation pathways [6, 17–19], supplementary characterization [19–22], or a large quantity of experimental data [18, 23, 24]. In the case of batteries, substantial work has occurred in forecasting degradation of commercial technologies with similar practical constraints as PV [25–27]. Previous works in batteries have used a variety of statistical and ML methodologies: from analytic or physically-motivated models [17, 25, 28] to deep learning methods [29]. There is an inherent trade-off between the performance of these models and their human explainability [17, 30, 31], *i.e.* models that predict degradation as the parametric exponential decay will be more interpretable (and useful for experimentalists) than deep learning models, but might be less accurate.

Contributions: In this work we propose *DeepDeg*, a flexible ML framework that satisfies these requirements by decoupling the forecasting and explainability problems. The DeepDeg framework consists of: a) a deep learning model to accurately forecast degradation in novel PV using the initial hours of degradation of multivariate device characteristics (in this work, the current-voltage (JV) [32]), and b) an ML explainability framework to attribute and predict the impact of various physical factors during degradation on a given figure of merit. For this, we curate a large dataset of JV degradation in OPV. We successfully demonstrate accurate forecasts of the degradation trends and explanation of driving factors of degradation with as little of 5% to 15% of the total degradation time, speeding up degradation characterization by an approximate factor of 5 to 20 times.

1 Main text

Model overview: We propose *DeepDeg*, an ML model for forecasting and explaining device degradation. Our approach combines high-throughput device stability tests, deep learning and explainable ML based on physical or chemical models. The model is developed to (1) **forecast** the degraded JV characteristics of solar cells based simply on the initial hours of measured degradation, and (2) **explain** and predict the impact of physical or chemical factors, referred in this work as *"driving factors"* of degradation, on the change of any arbitrary figure of merit.

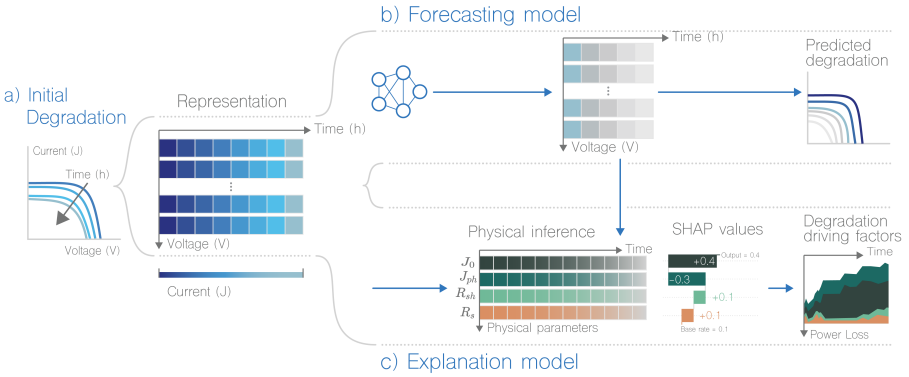

Fig. 1 Overview of DeepDeg: a) The initial hours of degradation of a PV device are represented as a multivariate time series of current density-voltage characteristics $J(V,t)$. b) A forecasting model consisting of a neural network, trained on degradation data from other solar cells, is used to predict future degradation dynamics in a single shot. c) Time-regularized physical inference is used to fit the degradation dynamics to the dynamics of various physical parameters in time according to a device model. A surrogate explanation model and SHAP values are used to quantify the effect of each physical parameter on a figure of merit, such as power loss at the maximum power point ΔP_{MPP}.

DeepDeg is summarized in Fig. 1. For a given solar cell sample, the model uses the initial hours of JV degradation under controlled illumination and humidity conditions (Fig. 1a). Based on this initial degradation data and the complete degradation dynamics of other samples, a deep learning model (the "Forecasting model", Fig. 1b) is trained to predict the complex evolution of the multivariate JV characteristics. The multivariate JV data is represented as a multivariate time series of current at distinct voltages, *i.e.* $J(V,t)$, where J is the current density of the solar cell at voltage V and measurement time t. For clear visualization, we include only the forward bias JV in figures, but the DeepDeg model is trained and tested including reverse bias as well. The deep learning architecture of the forecasting model is inspired by sequence to sequence models [33–35], combining a linear auto-regressive component with a non-linear convolutional neural network architecture (Fig. 2a), as described in Methods. In contrast to classical degradation forecasting, which predicts one or a few figures of merit based on data from independent time series [36], DeepDeg

forecasts the complete JV degradation trend in a single shot. This provides a more informative picture of the degradation dynamics without accumulating auto-regressive forecasting errors, and it also elucidates expected trade-offs between multiple figures of merit. DeepDeg demonstrates good generalization across varied active layers and device architectures.

Although the forecasting model excels at predicting complex dynamics without any hard assumptions of degradation mechanisms, an experimentalist may struggle to correlate these dynamics to hypothesized degradation processes or driving factors. To address this limitation, we developed an auxiliary model (the "Explanation model", Fig. 1c) which provides explanations of the solar cell degradation with respect to any physical or physicochemical device model, agnostic to the model definition. To avoid model overfitting across OPV degradation dynamics [37], we demonstrate our approach using an analytical one-diode equivalent circuit model (ECM) [38], as defined in Eq. 1 in Methods and Fig. 2b. The *DeepDeg* explanation model attributes the time-dependent degradation trend $J(V,t)$ and any relevant figures of merit to the change of physical model parameters. For example, rapid change in the energy conversion efficiency of a particular sample might be attributed to the observed change in time of shunt resistance R_{sh}. Such explanation may inform the design or manufacturing process or may lead to hypotheses or follow-up experiments. The DeepDeg Explanation model is summarized in Fig. 2b. The model consists of an initial "physical inference stage" that fits a device model to time-varying JV characteristics ($J(V,t)$) to extract time series of interpretable parameters. Since most device models, including ECM, are not time dependent, we minimize a time-regularized loss function to estimate the time series of the model parameters in a consistent manner, as explained in Methods. Then, in a second explanation stage, the evolution of these physical parameters is correlated to the dynamics of any particular figure of merit. This *explanation surrogate* model can be an analytical function derived form the original model or an estimated ML model. For simplicity, in this work we model ΔP_{MPP}, the power loss at the maximum power point [38] as function of the one-diode ECM parameters in time. This explanation surrogate model can be used to extract driving factors of degradation using any feature attribution technique from the explainable ML or, potentially, the causal ML literature. In our case, we choose KernelSHAP attribution [39] due to its model-agnostic characteristics. Using KernelSHAP, we are able to estimate the impact of any device parameter on any relevant figure of merit. This approach is more robust to non-linear feature interactions than traditional sensitivity analysis [39]. Fig. 2b presents an example of the power loss in time according to SHAP attributions.

DeepDeg has fundamental advantages compared to previous methods for novel PV and batteries [17, 19, 28, 36]: a) DeepDeg **decouples** the forecasting and explanation problems, avoiding the common mispecification of physical or chemical device models that leads to poor predictions or incorrect explanations as model assumptions break in time or across devices, b) DeepDeg **improves forecasting** by using multivariate degradation characteristics

across samples, allowing scalable learning across different devices and degradation dynamics, c) DeepDeg **facilitates the explanation** of degradation by using time-regularized inference and by making possible the use of any device model or figure of merit, and, finally, d) when the forecasting and explanations models are run in parallel, DeepDeg is able to directly predict the future driving factors of degradation.

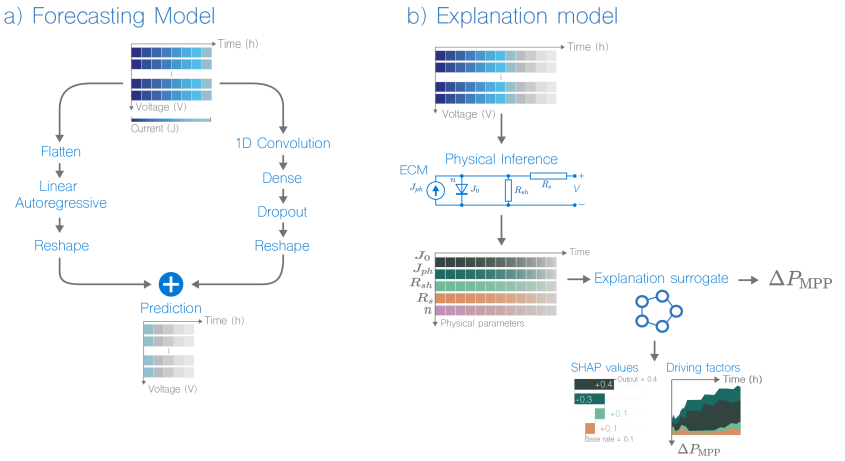

Fig. 2 DeepDeg Forecasting and Explanation Models: a) The forecasting model consists of a linear component (left branch) and a non-linear component (right branch). More information of network architecture and hyperparameters is included in SI. b) The explanation model consists of a physical inference stage based on a device model. In this case, the device model is a one-diode equivalent circuit model. Time-regularized inference is used to extract the dynamics of the physical parameters, such as $J_0(t)$, $R_s(t)$, etc. These dynamics are correlated to a figure of merit through a surrogate model. SHAP values are computed to decouple the contribution of each physical parameter to the figure of merit in time.

OPV degradation and PCE10:OIDTBR held-out datasets We develop DeepDeg using a large database of OPV degradation consisting of 789 samples of OPV solar cells fabricated and evaluated at i-MEET in the 2017-2019 period. This database includes normal and inverted architectures, more than a dozen different active layer materials (including IDTBR, ITIC, PCBM, among others as acceptors as well as various polymer donors) and various charge transport layers. In total, the dataset cover 23 different device configurations, summarized in Fig. S1 and Table S1. We use the JV degradation data of the samples agnostic of their detailed fabrication parameters and architecture to learn general trends in OPV stability. The dataset has over 230,000 hours of JV degradation under controlled conditions described in Methods. To the best of our knowledge, this dataset is the largest and broadest public dataset of degradation of novel PV. Each sample is subjected to 295 hours of controlled degradation and its JV characteristics are measured in logarithmic time intervals, which account for the *approximately exponential* nature of degradation

in non-packaged solar cells. Figure 1a presents an illustrative example of the JV degradation trend of a novel solar cell. For validation, we fabricated and measured a held-out test set with a known degradation mechanism, consisting of 31 PCE10:OIDTBR solar cells with cell architecture summarized in 5a and over 9,000 hours of measurement. ZnO interface degradation is a well-known effect in organic solar cells [5, 40]. In order to test DeepDeg's capabilities to forecast and explain degradation dynamics, the ZnO charge-transport layer was annealed at two different conditions (80°C as well as 200°C) in the test set. The first annealing condition causes faster and more severe degradation due to higher defect density and charge build-up in the ZnO/active layer interface [40]. Given that the PCE-10:O-IDTBR active layer is not present the training dataset, this represents a challenging and realistic validation system for our model.

Forecasting cross-validation We perform grouped 5-fold cross validation across the OPV degradation dataset. Often 4-6 similar but independent solar cell samples are fabricated on a shared substrate. To avoid data leakage, we group the training and test splits so the samples on a given substrate are only present in one of the splits. For an initial fraction of the JV degradation measurement up to a cutoff time τ, we train the DeepDeg and other machine learning models to forecast the future JV dynamics after time $t = \tau$. At inference time, we make predictions using exclusively the initial JV measurements up to time $t = \tau$ of a given sample. Fig. 3a illustrates the ground-truth and DeepDeg-predicted trends for a test sample based on $\tau = 30$ hours of measured degradation. We observe that the model correctly captures dynamics in the JV characteristics. In contrast, Fig. 3b presents a poor forecast by DeepDeg. In this case, the degradation trend is incorrect in terms of magnitude, JV covariance and degradation rate. For model comparison, we consider two metrics at the sample level: a) *Whole Trend RMSE*: the root mean squared error (RMSE) between ground truth and predicted $J(V, t)$ characteristics across all predicted times (Eq. 7), b) *Last Time RMSE*: the RMSE between ground truth and predicted $J(V, T)$ at the final degraded time T (Eq. 6). Fig. 3c compares both test metrics, averaged across all samples and folds, for the DeepDeg and other models. The *baseline* model is a naive model of degradation that repeats the last known measurement at time τ for all future times, while the other models are described in Table S2. Using both the initial 5% and 20% of the degradation measurement, DeepDeg outperforms all benchmark models for both whole trend and last time RMSEs. The combination of a linear term and non-linear convolutional term in DeepDeg forecasting model seems effective and sufficient to capture the degradation dynamics. More complex models (such as those developed for non-stationary multivariate time series, TCN (also known as Wave-Net) [41] and an adaption of LSTNet [34]) perform worse in this case. Fig. 3d evaluates the effect of different τ cutoff times on both mean RMSE test metrics. In general, the RMSE decreases sharply as the initial times of degradation are considered until it plateaus. This trend is explained by degradation dynamics: change occurs rapidly during the initial

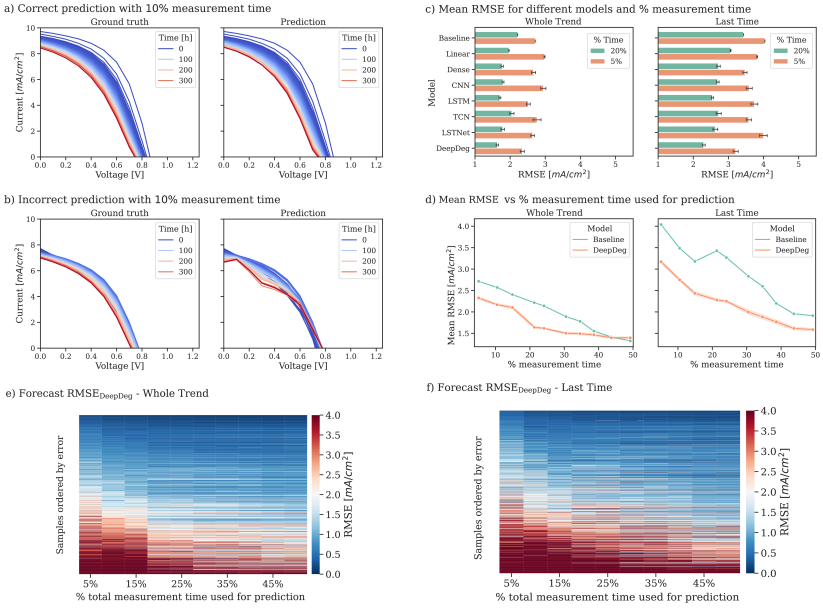

Fig. 3 Forecasting cross-validation: a) High-accuracy degradation prediction ($\text{RMSE}_{\text{Last}} = 0.67$ mA/cm^2), b) Low-accuracy prediction ($\text{RMSE}_{\text{Last}} = 2.46$ mA/cm^2). For both a) and b), the predicted trends including reverse bias are presented in Fig S4. c) Mean RMSE across cross-validation test folds for various models, as detailed in the SI. DeepDeg has consistently the lowest error for predicting the whole degradation trend and the last degraded time. d) Mean RMSE across cross-validation test folds for DeepDeg and the Baseline (no prediction) model as a function of the % of measurement time used for prediction. Error bars in bot c) and d) correspond to training initialization on 10 random seeds. e)-f) Test RMSEs for each sample in the dataset, according to grouped cross-validation splits, as a function of the measurement time used for prediction. The color scale is defined by inspecting the JV predictions and ranking them in ranges according to expert judgment: good predictions are consider such under 1.5 mA/cm^2 and acceptable predictions under 2.0 mA/cm^2.

degradation hours and eventually stabilizes. As expected, DeepDeg outperforms the naive baseline for all times, except after a substantial fraction of all the degradation has occurred. Figs. 3e-f presents the test RMSE for each sample (row) as a function of the measurement time used for prediction. Most samples are predicted correctly using only the initial hours of measurement. We observe a limited number of samples (less than 5% of all test samples) are predicted incorrectly by the model for all times, which likely have very unique degradation dynamics.

Explanation cross-validation For the explanation model, we are interested in correctly identifying and forecasting the driving factors of degradation in each test sample. As a demonstration, we use the one diode ECM to infer the

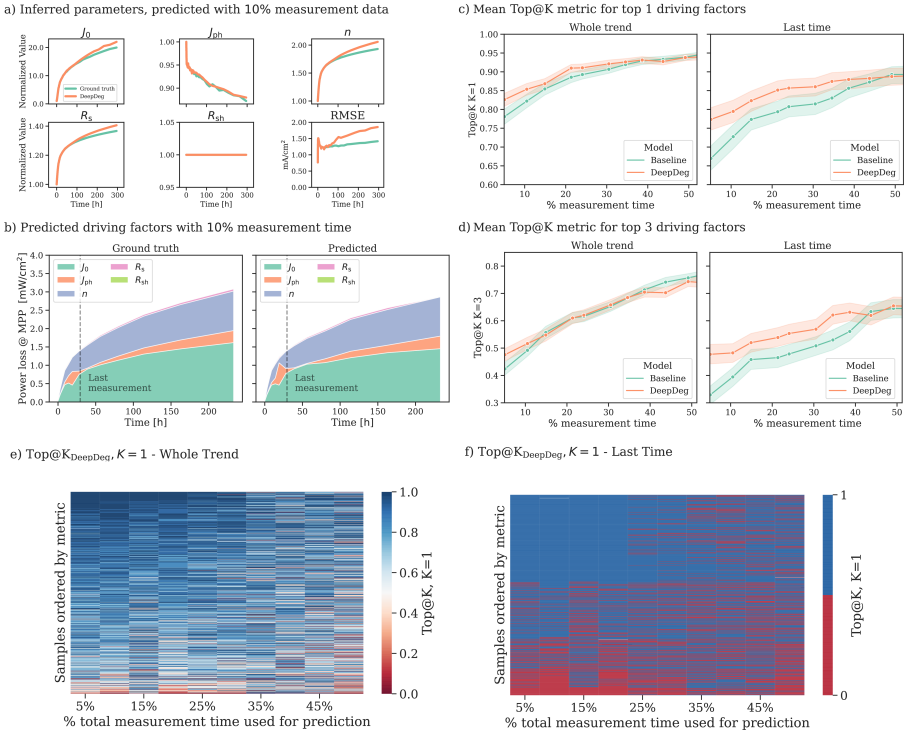

Fig. 4 Explanation cross-validation: a) Inferred parameters based on the ground truth degradation trend and the DeepDeg forecast. ECM physical parameters are normalized to their value at time $t = 0$ in the figure. RMSE is the root mean squared error of fitting JV characteristics with the one-diode ECM model. b) Explanation of driving factors of degradation on the power loss at the MPP (ΔP_{MPP}). c) Mean of Top@1 metric for whole trend and last time according to % of the measurement time used for prediction. d) Mean of Top@3 driving factors for the whole trend and the last time. e) Top@1 for the whole trend for all samples in the test folds. Figure S7 presents the Top@2 and Top@3 metrics as function of measurement time used for prediction. f) Top@1 for the last time for all samples in the test folds. Figure S6 presents the Top@3 metrics across samples.

physical parameters that determine JV characteristics at each time point during measured and predicted degradation. Then, we apply DeepDeg to attribute the impact of the physical factor dynamics on the power loss at the maximum power point, ΔP_{MPP}. Using the JV forecast in combination with the explanation model, we rank each physical parameter in the ECM according to its predicted impact and compare it to the ground truth factors. Fig. 4a illustrates the inferred dynamics of the parameters of the ECM model and the mean RMSE for each JV fitted by the ECM model, using the same sample of Fig. 3a. Using 10% of the measurement time, DeepDeg correctly predicts the dynamics of model parameters with low fitting RMSE's. The ground truth and predicted dynamics for the ECM model are used to attribute the impact on the power loss at each point in time, as presented in Fig. 4b. With a limited measurement time, DeepDeg correctly predicts the main degradation driving

factors and their contributions to the power loss at the MPP during the complete degradation process. To assess explanation performance, we compute the *Top@K* metric across test samples, which correspond to the fraction of samples for which the predicted driving factor was correctly identified in the K ordered ranking, *i.e.* Top@3 corresponds to the fraction of samples for which the top three driving factors were identified in correct order. We evaluate two metrics at the sample level: a) *Whole Trend*: the mean Top@K across all predicted times (Eq. 9 in Methods), b) *Last Time*: the Top@K metric at the final degraded time T (Eq. 8). Figs. 4c-d present the Top@1 and Top@3 for the whole trend and last time. DeepDeg performs very well in both cases, surpassing the naive baseline, particularly for the *last time* scenario. We observe the prediction of the top driving factor (K=1) is more accurate than the top 3 ranked factors (K=3), mostly due to top driving factors ranked in incorrect order (see Fig. S7 for K=2). Figs. 4e-f summarize the Top@1 metrics across all samples in the dataset, as a function of the measurement time. DeepDeg predicts correctly the top 1 driving factors of power loss at the MPP for the whole degradation trend and for the last time (Fig. S6) for most samples, using as little as 5% of the initial data. The performance for some samples seems to be hampered by increasing the measurement time available for prediction. This is explained by the Top@1 whole trend metric averaging over less prediction times as we increase measurement time, which makes explanation errors at later times more significant. Although the vast majority of samples are correctly forecast and explained, we observe that the DeepDeg is sensible to changes in the data used for prediction: at a given time the explanation is correct for a sample, but more measurement data may not necessarily improve the Top@K metrics as the degradation dynamics of ΔP_{MPP} may change substantially in time.

Test on PCE10-OIDTBR held-out dataset

As discussed, the PCE10-OIDTBR devices constitute a challenging validation dataset for our model. In this case, we designed the dataset to reproduce a known degradation pathway in OPV devices: a low annealing temperature (80°C) in the ZnO charge transport layer causes excessive recombination and charge build-up leading to fast interface degradation. A higher annealing temperature (200°C) limits this degradation pathway. Fig. 5b presents two representative samples with both annealing conditions (80°C and 200°C). In both cases, we make predictions using 15% of the total measurement time as input to the DeepDep model trained on exclusively on the legacy dataset. We observe the degradation forecast tracks closely the ground truth degradation. Fig. 5c compares the inferred ECM model parameters for both annealing conditions. We observe that the 80°C condition is dominated by substantial increase in increase in J_0 and n, indicating higher bulk or interface recombination compared to the 200°C condition. We expect DeepDeg to forecast and identify this trend. Fig. 5d compares the ΔP_{MPP} ground truth and DeepDeg-predicted trend at both conditions. For 80°C annealing, we observe the dominant power

loss is indeed predicted correctly, with J_o and n being the dominant driving factors. In contrast, the 200°C annealing conditions, present substantially lower power loss ΔP_{MPP}, and it is mostly explained by the J_{ph} and R_{sh} parameters. We observe that although the magnitude of ΔP_{MPP} is underestimated in the high temperature annealing condition, the relative driving factor attribution is the same, confirming the robustness of DeepDeg explanations to forecasting error. In this case, the forecasting error is likely caused by limited degradation during the initial hours of measurement which might confound potential degradation dynamics. Interestingly, in addition to correct physical trends, we observe the light-soaking step at around 25 hours for both conditions, which is evidenced as a sudden recovery in the power loss. We observe in Fig. 5e that the final JV characteristics of most samples are correctly predicted after the 15 % threshold, albeit with a higher variance than in the cross-validation test sets. Fig. 5f presents the Top@1 metric for the whole time series, which is correctly predicted for a substantial fraction of the test. The RMSE and Top@1 last time metrics are presented in Fig S8.

Conclusion

Compared to data-driven approaches in PV and other energy devices, DeepDeg decouples the multivariate forecasting and explanation problems. This facilitates the use of flexible physical or chemical models to explain degradation combined with scalable learning of degradation dynamics. For training, we curate the largest, to our knowledge, dataset of novel PV degradation, and DeepDeg demonstrates noteworthy performance for predicting and explaining long term degradation and good extrapolation capabilities to diverse solar cell materials and architectures. We acknowledge several limitations of the model: (1) novel degradation dynamics (out-of-distribution) might not be predicted accurately, (2) although mitigated by our contributions, the quality of degradation explanation relies on correct device model specification, (3) model predictions are limited to accelerated degradation tests under controlled conditions, underestimating the stressing effect of environmental conditions such as combined high humidity and high heat. As larger datasets and high-throughput experimental setups become accessible, we expect these limitations to be mitigated. We expect limitation (1) to be addressed by increased available data and fast model calibration from high-throughput experiments. Limitation (2) can be mitigated by improved physicochemical models, learning interpretable statistical models, or inclusion of fast secondary characterization. Finally, to address limitation (3) the DeepDeg framework can be extended to non-constant environmental conditions or potentially make use of causal inference approaches to better decouple environmental interactions. We envision DeepDeg being useful for fast experimental screening of novel energy devices and architectures. Given its potential to accelerate experiments, DeepDeg can be useful for closed-loop learning and optimization of complete device architectures. In future work, DeepDeg can be extended to other energy devices, such

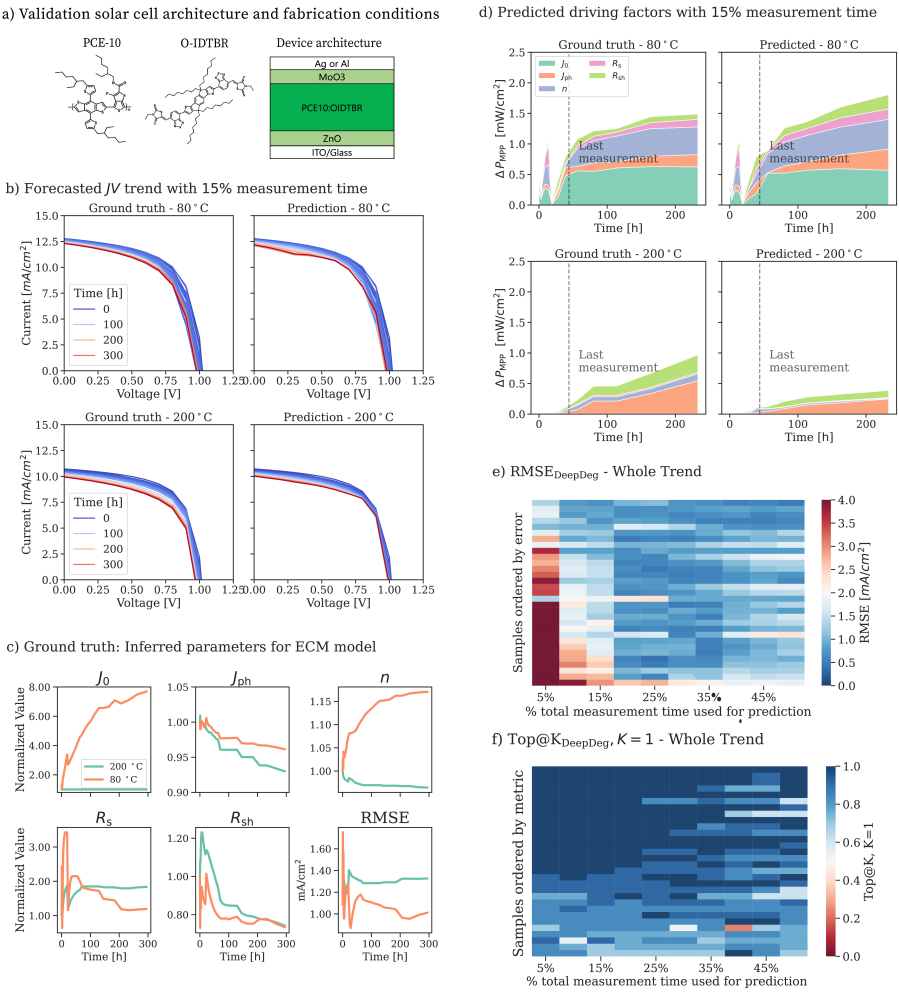

Fig. 5 Held-out PCE10:O-IDTBR test set: a) Molecular structure of PCE-10 and O-IDTBR, along with the device architecture. b) Comparison between the ground truth and DeepDeg forecast for the two ZnO annealing conditions. c) Inferred ECM parameters and fitting RMSE for the ground truth degradation trends of two annealing conditions. d) Comparison of ΔP_{MPP} explanations for two annealing conditions. Cells are the same as sub-figures b and c. e) RMSE for DeepDeg forecast for all samples in the dataset. f) Top@1 metrics across all samples in the dataset. No sample in the held-out test set was used during model training. Figure S8 presents the last time RMSE and Top@1 metrics.

as batteries and fuel cells, for which multivariate characteristics are adequate descriptors of degradation (*e.g.* discharge-voltage measurements [25] in batteries) or when physical or chemical models are used to pinpoint degradation issues.

Methods

Degradation measurements and data pre-processing All degradation measurements are performed in a custom-built degradation chamber under controlled nitrogen atmosphere ($<$ 1 ppm O_2) at 1∼sun constant illumination intensity under white light LED lamps (3000∼K, CMA3090-0000-000Q0H0A30G). The temperature is actively controlled with air flow around the chamber ensuring that the temperature of the active layer is 35°C +/- 5. The JV characteristics are measured in steady intervals (0.5 h - 1 h), re-sampled to be the same for all samples, during at least 295∼hours. We curate the dataset by dropping samples that present measurement error, including poor or non- rectifying behaviour at the start of accelerated degradation test (shunts, S-shape), sudden loss of electrical contact not explained by degradation and substantial time measurement gaps. The SI contains an overview of the 24 solar cell types and configurations in the dataset (Fig. S1, Table S1)

PCE10:O-IDTBR sample preparation Fabrication of photovoltaic devices: Pre-structured indium tin oxide substrates were cleaned with detergent followed by two 10 min ultrasonification steps in acetone and isopropanol. ZnO (N10 bought from Avantama) was ultrsonicated for 10 mins and then filtered with a 0.2 µm PTFE-filter. The 30 nm thick layer of ZnO was spin-coated at 3000 rpm in air and annealed at 80 °C or 200 °C for 5 min. PCE10 (batch MY7118CH) and O-IDTBR (batch YY13033B) was bought from 1-Material. The PCE10:OIDTBR solution was produced with a 1:2 D:A-Ratio and a concentration of 24 mg/ml. The active layer was spin-coated from a 60 °C hot solution at 2000 rpm on the substrate. To complete device fabrication 12 nm of MoO_x and 100 nm of Al/Ag were thermally evaporated through a mask (active area of 10.4 mm^2) in vacuum of $\sim 10^{-6}$ mbar. The devices were not encapsulated but were loaded into sealed degradation boxes in the glovebox. A light-soaking step was applied at 3 and 23 hours to improve device properties, evidenced by a sudden improvement of JV characteristics. To increase diversity across samples in the held-out test set, we use two different electrodes, Ag and Al, along with UV treatment in the ZnO in certain devices at the 80 °C condition. The ZnO annealing conditions produce similar effects for all electrodes and the UV-treated cells (Figs. S2 and S3).

DeepDeg forecasting model The time-resolved JV characteristics can be seen as a multivariate time series $J(V,t)$, where J correspond to the current at the measurement voltage V and time t. For convenience, we define the JV characteristics at each time t as a vector $J_{V,t}$ and frame the forecasting problem as a sequence to sequence prediction problem. Based on the initial degradation trend up to time $t = t_\tau$, we are interested in predicting the degradation to the final degraded or measurement time $t = t_T$. Thus, the DeepDeg forecasting model is defined as a sequence to sequence model such that:

$$J_{V,t_{\tau+1}}, J_{V,t_{\tau+2}}, \ldots, J_{V,t_T} = f(J_{V,t_0}, J_{V,t_1}, \ldots, J_{V,t_\tau}) \quad (1)$$

In the case of the DeepDeg forecasting model, f is composed of a linear auto-regressive component f' of order 3, *i.e.* applied only to the last three times in the sequence, combined to a neural network component f'' that considers the complete sequence:

$$f = f'(J_{V,t_{\tau-2}}, J_{V,t_{\tau-1}}, \ldots, J_{V,t_\tau}) + f''(J_{V,t_0}, J_{V,t_1}, \ldots, J_{V,t_\tau}) \tag{2}$$

The order of f' was determined by hyperparameter sweep on a 10% random sample of the training data in the range of 1 to 8. f'' in this case is a neural network composed of 1D convolutional layer, a dense layer with ReLu activation and Dropout, and Reshape linear layer. We compare DeepDeg to common compact sequence to sequence models defined in the SI.

DeepDeg explanation model The explanation model is composed of two stages: i) inference and ii) explanation. In the inference stage, we determine the physical or chemical parameters, defined by vector α, that fits the $J(V,t)$ characteristics. At a given time t, this is given according to a physical or physicochemical device model g, such that $J_{V,t_i} = g(\alpha_{t_i})$. In the inference step, we are interested in determining the α_{t_i} for each time t_i. Most device models do not consider time dynamics as assumptions tend to break with complex degradation processes, so we are forced to learn α_{t_i} from each $J_{V,t}$ characteristics at a given time. Naturally, the parameters $\alpha_{t_{i-1}}$) and α_{t_i} are not independent of each other (*i.e.* $\mathbb{P}(\alpha_{t-1} \cap \alpha_t) \neq \mathbb{P}(\alpha_{t-1})\mathbb{P}(\alpha_t)$). *A priori*, this dependency is generally unknown, except for two considerations. First, there is certain degree of *smoothness* in the transition between small time steps, which reflects the occurrence of steady physical processes. Second, certain parameters exhibit *on average* monotonic tendencies, which can be embedded into the inference procedure. In consequence, the inference step can be formulated as minimizing the time-regularized loss function of the JV characteristics:

$$\alpha^*_{t_i} = \arg\min_{\alpha} \mid g(V, \alpha_{t_i}) - J(V, t_i) \mid^2 + \lambda \mid \alpha_{t_i} - \alpha_{t_{i-1}} \mid^2 \tag{3}$$

All α_i components are normalized in the loss function. The first term controls how well the model approximates the $J(V,t_i)$ characteristics, and the λ term how consistent the parameters are with respect to the previous inference time. We choose $\lambda = 5e-06$ by hyperparameter sweep in the range of $[1e-03, 1e-06]$. Notice that $g(V, \alpha_{t_i})$ can be any device model: an analytical equivalent circuit model, a numerical PDE (Partial Differential Equation) model or an empirical statistical model. The inference step is repeated to determine the time dynamics of model parameters $\alpha(t)$. The explanation step is performed to relate the inferred $\alpha(t)$ dynamics to the change in a figure of merit of interest during degradation. As device models g are non-linear, the impact of a physical parameter α_k changes according to time, the model sensitivity and the relative value of other model parameters. In consequence, to

rank the driving factors of degradation we quantify the impact of a given $\alpha_k(t)$ on the output metric of interest Θ, such as power conversion efficiency, fill factor, etc. Traditionally, this impact is measured by varying a single parameter independently and observing the model's output [42]. This approach is correct for degradation studies that contain only a few discrete samples and linear device models. Instead, we frame this problem as a ML feature attribution problem, providing greater flexibility and accuracy. We construct an auxiliary explanation model based on the change in time of the inferred device parameters, $\Theta = h(\Delta\alpha(t))$. For the purpose of this work, we choose $\Theta = \Delta P_{\mathrm{MPP}}$, where Δ is the total change from $t = 0$ to t_τ, and P_{MPP} is the maximum power extracted from the solar cell by maximizing the $J \times V$ product. Thus, the explanation model considers the maximum power loss with respect to the initial time as a function of the change of device parameters. In our ECM demonstration, for simplicity we derive h analytically, but we note it may correspond to any numerical or statistical model. Once the explanation model is defined, we attribute the impact of each factor using a model-agnostic approach, KernelSHAP (Kernel Shapley Additive Explanations) [39]. According to it, the contribution of each feature to the model output is approximated by a local linear attribution model:

$$q(z) = \phi_0 + \sum_{j=1}^{M} \phi_j z_j \tag{4}$$

where q is the local linear model, $z \in \{0,1\}^M$ are the *simplified features*, M is the maximum number of input features, and $\phi_j \in \mathbb{R}$ is the feature attribution coefficient for feature j. For an specific data point, the simplified features z take a value of 0 if a feature is "absent" (*i.e.* represented by a random value) and a value of z if the feature is "present" (represented by the actual feature value). According to game theory, ϕ_j is a SHAP value [39], or a weighted average of a feature's marginal contributions over all possible feature combinations. To generate feature attributions, KernelSHAP minimizes the loss L of the local linear model to matched the output of the actual model [39]. In this work, we compute the impact of each k feature change in time $\Delta\alpha_k(t)$ on $\Delta P_{\mathrm{MPP}}(t)$ with the SHAP value ϕ_k, and then plot this value as a fraction of the total power loss $\Delta P_{\mathrm{MPP}}(t)$, exploiting the additive nature of SHAP values.

One-diode equivalent circuit model (ECM) We choose a one-diode model equivalent circuit model as a device model of current density for a particular time and voltage $J_{V_i,t_i} = g(\alpha_{t_i}, V_i)$. According to it:

$$J = -J_{\mathrm{ph}} + J_0 \left\{ \exp\left[\frac{V - JR_{\mathrm{s}}}{nV_T}\right] - 1 \right\} + \frac{V - JR_{\mathrm{s}}}{R_{\mathrm{sh}}} \tag{5}$$

where J is the current, V is the voltage, and the vector of parameters α is composed of J_{ph}, the photo-generated current of the solar cell, J_o a reverse diode saturation current with diode ideality factor n, a series resistance R_s and a shunt (parallel) resistance R_{sh}. Although PDE models may offer deeper

physical insights [43, 44], their inherent assumptions may break during degradation. In the case of the OPV dataset, we find the one-diode ECM to be sufficient to provide reasonable explanations for a broad range of degradation trends, which is consistent with literature [37]. This model can be easily updated within the DeepDeg explanation framework to account for physical or chemical parameters of interest, along with particular figures of merit.

Evaluation metrics The forecasting performance is assessed by root mean squared error of the last predicted JV in time or the whole trend of predicted JV characteristics. The last time metric is defined at time $t = \tau$ for n measurement voltages as:

$$\text{RMSE}_{\text{Last}} = \sqrt{\frac{1}{n}\sum_{i}^{n}(J_{V_i,t_\tau} - \hat{J_{V_i,t_\tau}})^2} \tag{6}$$

where $\hat{J_{V,t_\tau}}$ are the predicted JV characteristics at time τ and J_{V,t_τ} is the ground truth degraded measurement. In the same way, the RMSE for the predicted trend with m elements (times) is given by:

$$\text{RMSE}_{\text{Whole}} = \sqrt{\frac{1}{m+n}\sum_{i}^{m}\sum_{j}^{n}(J_{V_j,t_i} - \hat{J_{V_j,t_i}})^2} \tag{7}$$

The explanation metrics Top@K metrics are based on the k_{th} magnitude-ranked SHAP value of a given physical parameter as defined by $\phi_{k,t}$ at time t up to order of interest K. For a DeepDeg explanation at time t, the indicator function $I(\phi_{k,t}, \phi_{k+1,t}, \ \ldots, \phi_{K,t})$ equals 1 if the set of parameters $\phi_{k,t}, \phi_{k+1,t}, ..., \phi_{K,t}$ ordered according to decreasing SHAP value is equal the ground-truth degradation driving factors, and 0 elsewhere. Then, at time T, $\text{Top@K}_{\text{Last}}$ is either 0 or 1 defined by this indicator function:

$$\text{Top@K}_{\text{Last}} = I(\phi_{k,t}, \phi_{k+1,t}, \ \ldots, \phi_{K,t}) \tag{8}$$

Similarly, the $\text{Top@K}_{\text{Whole}}$ metric will be the mean of Top@K metrics across time:

$$\text{Top@K}_{\text{Whole}} = \frac{1}{m}\sum_{i}^{m} I(\phi_{k,t_i}, \phi_{k+1,t_i}, \ \ldots, \phi_{K,t_i}) \tag{9}$$

Model development The forecasting model development procedure along with hyperparameter search and ablation study is presented in Section S.3 and Fig S5. For the explanation model, we considered also a complex charge generation and transport PDE model for O-PV whose assumptions broke down in time, leading to poor fitting of JV dynamics. Thus, we decided to use a more flexible 1-diode ECM model. We explored other possibilities for the ECM model, including 2-diode and 3-diode ECM models, available in the DeepDeg code. We select the 1-diode ECM model due to its good performance for O-PV devices with no S-shapes in JV characteristics at time 0 and limited number of parameters to analyze.

References

[1] Osterwald, C. R., Anderberg, A., Rummel, S. & Ottoson, L. *Degradation analysis of weathered crystalline-silicon PV modules*, 1392–1395 (2002).

[2] Ndiaye, A. *et al.* Degradations of silicon photovoltaic modules: A literature review. *Solar Energy* **96**, 140–151 (2013). https://doi.org/10.1016/j.solener.2013.07.005 .

[3] Boyd, C. C., Cheacharoen, R., Leijtens, T. & Mcgehee, M. D. Understanding Degradation Mechanisms and Improving Stability of Perovskite Photovoltaics (2018). URL https://pubs.acs.org/sharingguidelines. https://doi.org/10.1021/acs.chemrev.8b00336 .

[4] Du, X. *et al.* Efficient Polymer Solar Cells Based on Non-fullerene Acceptors with Potential Device Lifetime Approaching 10 Years. *Joule* **3** (1), 215–226 (2019). URL https://www.sciencedirect.com/science/article/pii/S2542435118304045. https://doi.org/10.1016/J.JOULE.2018.09.001 .

[5] Grossiord, N., Kroon, J. M., Andriessen, R. & Blom, P. W. Degradation mechanisms in organic photovoltaic devices. *Organic Electronics* **13** (3), 432–456 (2012). https://doi.org/10.1016/j.orgel.2011.11.027 .

[6] Bala Subramaniyan, A., Pan, R., Kuitche, J. & TamizhMani, G. Quantification of Environmental Effects on PV Module Degradation: A Physics-Based Data-Driven Modeling Method. *IEEE Journal of Photovoltaics* **8** (5), 1289–1296 (2018). URL https://ieeexplore.ieee.org/document/8410667/. https://doi.org/10.1109/JPHOTOV.2018.2850527 .

[7] Han, X. *et al.* A review on the key issues of the lithium ion battery degradation among the whole life cycle. *ETransportation* **1**, 100005 (2019) .

[8] Yuan, X.-Z. *et al.* A review of all-vanadium redox flow battery durability: Degradation mechanisms and mitigation strategies. *International Journal of Energy Research* **43** (13), 6599–6638 (2019) .

[9] Rama, P., Chen, R. & Andrews, J. A review of performance degradation and failure modes for hydrogen-fuelled polymer electrolyte fuel cells. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy* **222** (5), 421–441 (2008) .

[10] Correa-Baena, J.-P. *et al.* Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* **2** (8), 1410–1420 (2018) .

[11] Gu, G. H., Noh, J., Kim, I. & Jung, Y. Machine learning for renewable energy materials. *Journal of Materials Chemistry A* **7** (29), 17096–17117 (2019) .

[12] Lindig, S., Kaaya, I., Weiß, K.-A., Moser, D. & Topic, M. Review of statistical and analytical degradation models for photovoltaic modules and systems as well as related improvements. *IEEE Journal of Photovoltaics* **8** (6), 1773–1786 (2018) .

[13] Akhter, M. N., Mekhilef, S., Mokhlis, H. & Mohamed Shah, N. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. *IET Renewable Power Generation* **13** (7), 1009–1023 (2019) .

[14] Cacovich, S. *et al.* Elemental Mapping of Perovskite Solar Cells by Using Multivariate Analysis: An Insight into Degradation Processes. *ChemSusChem* **9** (18), 2673–2678 (2016). URL http://doi.wiley.com/10.1002/cssc.201600913. https://doi.org/10.1002/cssc.201600913 .

[15] David, T. W., Soares, G. A., Bristow, N., Bagnis, D. & Kettle, J. Predicting diurnal outdoor performance and degradation of organic photovoltaics via machine learning; relating degradation to outdoor stress conditions. *Progress in Photovoltaics: Research and Applications* **29** (12), 1274–1284 (2021) .

[16] Howard, J. M., Tennyson, E. M., Neves, B. R. & Leite, M. S. Machine learning for perovskites' reap-rest-recovery cycle. *Joule* **3** (2), 325–337 (2019). URL https://www.sciencedirect.com/science/article/pii/S2542435118305543. https://doi.org/https://doi.org/10.1016/j.joule.2018.11.010 .

[17] Dunlap-Shohl, W. *et al.* Physiochemical machine learning models predict operational lifetimes of ch3nh3pbi3 perovskite solar cells (2022) .

[18] Peshek, T. J. *et al.* Insights into metastability of photovoltaic materials at the mesoscale through massive i–v analytics. *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* **34** (5), 050801 (2016) .

[19] Stoddard, R. J. *et al.* Forecasting the decay of hybrid perovskite performance using optical transmittance or reflected dark-field imaging. *ACS Energy Letters* **5** (3), 946–954 (2020) .

[20] Karimi, A. M. *et al.* Generalized and mechanistic pv module performance prediction from computer vision and machine learning on electroluminescence images. *IEEE Journal of Photovoltaics* **10** (3), 878–887 (2020) .

[21] Sutanto, A. A. *et al.* In situ analysis reveals the role of 2d perovskite in preventing thermal-induced degradation in 2d/3d perovskite interfaces. *Nano letters* **20** (5), 3992–3998 (2020) .

[22] Hillhouse, H. W., Dunlap-Shohl, W. A. & Stoddard, R. J. *Quantitative prediction of perovskite stability using accelerated testing and machine learning*, 2116–2119 (IEEE, 2020).

[23] Kim, D. *et al.* Exploring transport behavior in hybrid perovskites solar cells via machine learning analysis of environmental-dependent impedance spectroscopy. *Advanced Science* **8** (15), 2002510 (2021) .

[24] David, T. W. *et al.* Enhancing the stability of organic photovoltaics through machine learning. *Nano Energy* **78**, 105342 (2020) .

[25] Severson, K. A. *et al.* Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy* URL https://doi.org/10.1038/s41560-019-0356-8. https://doi.org/10.1038/s41560-019-0356-8 .

[26] Ng, M.-F., Zhao, J., Yan, Q., Conduit, G. J. & Wei Seh, Z. Predicting the state of charge and health of batteries using data-driven machine learning. *Nature Machine Intelligence* URL https://doi.org/10.1038/s42256-020-0156-7. https://doi.org/10.1038/s42256-020-0156-7 .

[27] Li, A. G., West, A. C. & Preindl, M. Towards unified machine learning characterization of lithium-ion battery degradation across multiple levels: A critical review. *Applied Energy* **316**, 119030 (2022) .

[28] Attia, P. M., Severson, K. A. & Witmer, J. D. Statistical learning for accurate and interpretable battery lifetime prediction. *Journal of The Electrochemical Society* **168** (9), 090547 (2021) .

[29] Ren, L. *et al.* Remaining useful life prediction for lithium-ion battery: A deep learning approach. *Ieee Access* **6**, 50587–50598 (2018) .

[30] Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research* **3** (6), 597–607 (2022) .

[31] Naik, R. R. *et al.* Discovering equations that govern experimental materials stability under environmental stress using scientific machine learning. *npj Computational Materials* **8** (1), 1–8 (2022) .

[32] Smestad, G. & Ries, H. Luminescence and current-voltage characteristics of solar cells and optoelectronic devices. *Solar Energy Materials and Solar Cells* **25** (1-2), 51–71 (1992). https://doi.org/10.1016/0927-0248(92)90016-I .

[33] Sutskever, I. Sequence to Sequence Learning with Neural Networks 1–9 .

[34] Lai, G., Chang, W.-C., Yang, Y. & Liu, H. *Modeling long-and short-term temporal patterns with deep neural networks*, 95–104 (2018).

[35] Shih, S.-Y., Sun, F.-K. & Lee, H.-y. Temporal pattern attention for multivariate time series forecasting. *Machine Learning* **108** (8), 1421–1441 (2019) .

[36] Ray, B. & Alam, M. A. A compact physical model for morphology induced intrinsic degradation of organic bulk heterojunction solar cell. *Cite as: Appl. Phys. Lett* **99**, 33303 (2011). URL https://doi.org/10.1063/1.3610460. https://doi.org/10.1063/1.3610460 .

[37] Sesa, E. *et al.* A building-block approach to the development of an equivalent circuit model for organic photovoltaic cells. *Organic Electronics* **58**, 207–215 (2018) .

[38] Green, M. A., Green & A., M. *Solar cells: Operating principles, technology, and system applications* (1982).

[39] Lundberg, S. M. & Lee, S.-I. in *A unified approach to interpreting model predictions* (eds Guyon, I. *et al.*) *Advances in Neural Information Processing Systems 30* 4765–4774 (Curran Associates, Inc., 2017). URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[40] Ryu, S. Y. *et al.* Effects of the wrinkle structure and flat structure formed during static low-temperature annealing of zno on the performance of inverted polymer solar cells. *The Journal of Physical Chemistry C* **121** (17), 9191–9201 (2017) .

[41] Van Den Oord, A. *et al.* WAVENET: A GENERATIVE MODEL FOR RAW AUDIO. Tech. Rep.

[42] Sun, X., Chavali, R. V. K. & Alam, M. A. In-situ self-monitoring of real-time photovoltaic degradation only using maximum power point: the suns-vmp method. *arXiv preprint arXiv:1802.01206* (2018) .

[43] Brandt, R. E. *et al.* Rapid Photovoltaic Device Characterization through Bayesian Parameter Estimation Rapid Photovoltaic Device Characterization through Bayesian Parameter Estimation. *Joule* **1** (4), 843–856 (2017). URL https://doi.org/10.1016/j.joule.2017.10.001. https://doi.org/10.1016/j.joule.2017.10.001 .

[44] Xue, H., Fu, K., Wong, L. H., Birgersson, E. & Stangl, R. Modelling and loss analysis of meso-structured perovskite solar cells. *Journal of Applied*

Physics **122** (8), 083105 (2017). URL http://aip.scitation.org/doi/10.1063/1.4986115. https://doi.org/10.1063/1.4986115 .

Data availability. The processed degradation data, along with cell identifiers and metadata, will be available at the time of publication in a Github repository.

Code availability. The DeepDeg model will be available in a public Github repository at the time of publication. A public interface for forecasting degradation using DeepDeg can will be found in https://emerging-pv.org at the time of publication.

Author contributions. FO, JL, JF, CB and TB ideated the work. FO and DH developed and implemented ML algorithms. JDP and TH curated degradation data. JW and TH designed, fabricated and measured the held-out test set. RN and HX provided useful feedback and performed exploration on device modelling. FO and TB wrote the manuscript with feedback from all co-authors.

Acknowledgments. The authors thank Prof. Rafael Gomez-Bombarelli at MIT and Dr. Liu Zhe for useful discussion and brainstorming. FO and JL acknowledge funding of Microsoft's AI for Good and AI for Earth programs. C.J.B. acknowledges the financial support through the Bavarian Initiative 'Solar Technologies Go Hybrid' (SolTech) and the SFB 953 (DFG). We further acknowledge the grants 'ELF-PV—Design and Development of Solution-Processed Functional Materials for the Next Generations of PV Technologies' (no. 44-6521a/20/4) and 'Solar Factory of the Future' (FKZ 20.2-3410.5-4-5) by the Bavarian state government and the financial support from the German Research Foundation with grant DFG INST 90/917-1 FUGG, INST 90/1093-1 FUGG, DFG International Graduate School GRK2495/E and DFG Research Group PoPular.

Declaration. The authors declare no conflict of interest.