npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

# Discovering equations that govern experimental materials stability under environmental stress using scientific machine learning

Richa Ramesh Naik[1] ✉, Armi Tiihonen[1] ✉, Janak Thapa[1], Clio Batali[1], Zhe Liu[1], Shijing Sun[1] and Tonio Buonassisi[1] ✉

While machine learning (ML) in experimental research has demonstrated impressive predictive capabilities, extracting fungible knowledge representations from experimental data remains an elusive task. In this manuscript, we use ML to infer the underlying differential equation (DE) from experimental data of degrading organic-inorganic methylammonium lead iodide (MAPI) perovskite thin films under environmental stressors (elevated temperature, humidity, and light). Using a sparse regression algorithm, we find that the underlying DE governing MAPI degradation across a broad temperature range of 35 to 85 °C is described minimally by a second-order polynomial. This DE corresponds to the Verhulst logistic function, which describes reaction kinetics analogous to self-propagating reactions. We examine the robustness of our conclusions to experimental variance and Gaussian noise and describe the experimental limits within which this methodology can be applied. Our study highlights the promise and challenges associated with ML-aided scientific discovery by demonstrating its application in experimental chemical and materials systems.

*npj Computational Materials* (2022)8:72 ; https://doi.org/10.1038/s41524-022-00751-5

## INTRODUCTION

In the traditional scientific discovery process, prior knowledge from first principles and empirical laws are combined with experimental data and intuition to yield governing equations. Newton's law of gravitation[1], Einstein's mass-energy equivalence equation[2], Kepler's laws of planetary motion[3], and other physical principles were uncovered through careful interpretation of experimental data and inductive reasoning[4]. The approach of fitting experimental data through regression is difficult with systems that are yet to be understood fully—the set of feasible equations capturing the physics is enormous[5–7].

One such area where underlying physics is often poorly understood is the study of materials under environmental stress. For example, alloys[8,9], polymers[10], doped silicon[11], and hybrid materials[12] experience changes at elevated temperatures. The degradation pathways can be complex and not directly obvious when examining the experimental data. Machine learning (ML) has been used to predict degradation[13–17] as well as to optimize process conditions to reduce material decomposition[17,18]. However, traditional data-science methods yield little insight into the underlying mechanisms. We posit that hidden in the black-box ML models is valuable scientific information on the dynamics of the system. If uncovered, the knowledge of the governing dynamics can serve as foundation for physical interpretation of phenomena and scientific discovery.

Herein, we use Scientific ML, which combines regression-based ML with sparsity generating techniques in order to automatically identify governing equations directly from data, especially when the systems being studied are too complicated to yield to traditional theoretical analysis. Not only does Scientific ML help us understand the underlying scientific phenomena better, it also has the potential to help to make simulations faster and extrapolate beyond the dataset at hand.

Recently, many approaches aiming for this target have been presented in literature. A method that we apply in this contribution is PDE-FIND by Rudy et al.[19]. This method is used for the discovery of physical laws describing dynamical systems. First, a library of potential candidate functions is built. Differentials are calculated by finite difference or polynomial interpolation. Once a large matrix with all candidate functions is composed, different sparse regression methods may be used to extract the partial differential equation (PDE) describing the system. The sparse methods implemented are sequential threshold ridge regression, lasso regression, elastic net regression, and greedy algorithm. Another sparse technique is Sparse Identification of nonlinear Dynamics (SINDy)[20]. It uses a custom deep autoencoder to find a coordinate system in which the dynamics of the system are sparse, and then uses sparse regression to find the governing equations in the associated coordinate system. Atkinson et al.[21] present a generalized method for the discovery of differential equations using genetic programming. Physics Informed Neural Networks (PINN)[22] and PDE-NET[23,24] are deep learning methodologies to extract governing partial differential equations using dynamical data. These methods have shown great promise in several applications[25–28]. The automatic discovery of scientific laws and principles is at the frontier of machine learning that awaits application to materials science[29] and other domains[30–32].

Halide perovskite materials, which have potential to provide high performing and cost-effective solar energy, degrade at elevated temperature[33–38], humidity[39–41], and illumination[42–44]. This is a major issue hindering the commercialization of perovskite photovoltaic technology. However, the degradation mechanisms affecting halide perovskites are not well understood. Discovering the underlying equations directly from perovskite degradation data could accelerate the development of stable perovskite solar cells. Herein, we apply Scientific ML to study the environmental degradation of methylammonium lead iodide (MAPI).

From prior knowledge in the literature, MAPI has multiple documented reaction pathways, including decomposition to

[1]Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. ✉email: richarnaik96@gmail.com; armi.tiihonen@gmail.com; buonassisi@mit.edu

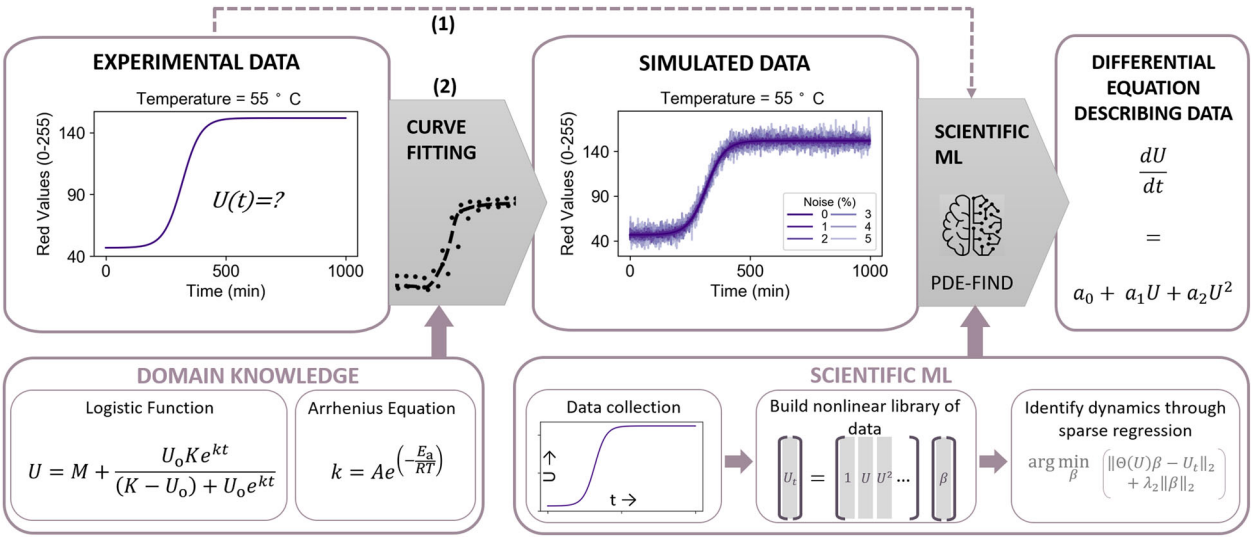


**Fig. 1 Schematic of the data-management workflow used in this study.** Workflow (1) applies PDE-FIND directly to experimental data; Workflow (2) first fits the experimental data with a logistic function to create a simulated dataset, optionally adds Gaussian noise, and then applies PDE-FIND.

$PbI_2$ via reaction[35]:

$$MAPbI_3 \rightarrow PbI_2 + \left[CH_3NH_3^+ + I^-\right] \rightarrow PbI_2 + CH_3NH_2 + HI \quad (1)$$

Smecca et al.[36] demonstrate that the rate of MAPI degradation obeys an Arrhenius-type law. Their data suggest that the degradation of MAPI follows zero-order kinetics in the presence of moisture and first-order kinetics in vacuum at temperatures ranging from 90 to 135 °C. Bastos et al.[45] hypothesize that the thermal degradation of MAPI is defined by the Avrami equation[46,47] of nucleation and growth. The Avrami equation has also been used to describe degradation kinetics in humid air[48]. Recently, studies have shown that halide perovskite degradation follows autocatalytic reaction kinetics[49] with the hypothesis that the degradation is propagated by iodine vapors[50]. The derivation of exact kinetics through first principles as well as Arrhenius-type dependence is difficult because of the complexity of MAPI decomposition, despite the availability of well-resolved dynamical data, inviting the application of Scientific ML.

In this study, we focus on the application of PDE-FIND to perovskite degradation data. We choose PDE-FIND as it is an interpretable method that provides a parsimonious description of the dynamics with the flexibility to apply domain expertise for library selection. Successfully identifying governing differential equations directly from the experimental aging test data would deepen the understanding of thermal degradation and provide tools for reliable lifetime prediction of perovskite solar cells as well as the determination of acceleration factors for long-term aging tests. These developments could spur the advancement of the perovskite photovoltaic technology and have been called for by the community[51–53]. This study provides a generalizable pathway to identify degradation modes in other materials research domains as well.

The objectives of this study are two-fold, as illustrated by the workflows shown in Fig. 1: Uncover the underlying differential equation corresponding to perovskite degradation using sparse regression methodology PDE-FIND (Workflow (1)) and quantify the effect of noise on the accuracy of extraction of differential equations by PDE-FIND by comparing noiseless and noisy simulated data (Workflow (2)).

Further details can be found in the "Methods" section; a summary is provided here. To generate the experimental data, we subjected 206 thin-film samples of methylammonium lead iodide (MAPI) to 0.15 ± 0.01 Sun illumination, 20 ± 5% relative humidity, and temperatures varying from 35 to 85 °C in our in-house environmental chamber described in detail in ref. [18] (Fig. 2b). A camera is used to monitor color changes versus time; the red color time-series is chosen for further analysis because two studies[18,54] have shown a clear correlation between film color and device performance, in the limit of MAPI composition, given one of the principal degradation products is yellow $PbI_2$. One hundred and eight samples were grown under low-variance conditions (labeled 'low-variance experimental', quantified in the "Results" section); 98 samples were grown under high-variance conditions (labeled 'high-variance experimental') (Supplementary Fig. 2), referring to the amount of variance (color change versus time) between samples synthesized and degraded under ostensibly identical conditions. Unless specified otherwise, we assume 'experimental' data in this paper refers to the low-variance sample set.

## RESULTS

### Results on experimental data

Our aim is to obtain the equation that most accurately describes the environmental degradation of methylammonium lead iodide (MAPI) as a function of time and temperature. There are two main challenges for Scientific ML in this application that are also common with many other experimental applications, especially in materials science: The function space that could in principle capture the degradation processes is enormous, complicating identification of unique equations. Furthermore, experimental data has measurement noise as well as sample-to-sample variance, making the identification of quantitative analytic descriptions even more challenging. These conditions can be optimized to some extent, but not excluded.

Our experimental setup represents a typical materials science experiment: The noise in our experimental data is of the order of 0.35% for both high-variance and low-variance experimental datasets. The low value indicates that the camera measurement of degradation is optimized. The sample-to-sample variance for the 'low variance experimental' dataset is estimated to be 20% in relative standard deviation and the maximum mean absolute deviation is 12 units (red color values vary from 0 to 255). For the 'high variance experimental' dataset, variance is estimated to be 23% in relative standard deviation and the maximum mean absolute deviation is 31 units. These values are typical for

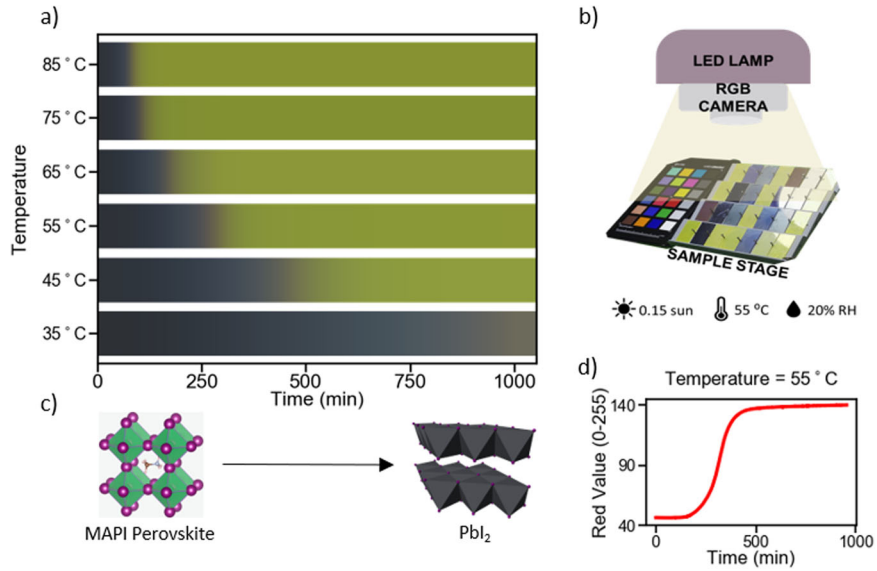


**Fig. 2 The experimental data. a** Average perovskite film color as a function of time at different temperatures. **b** A schematic diagram of the in-house accelerated degradation chamber with a superimposed camera image of degrading MAPI films. **c** The structure of MAPI perovskite (reactant, image source: ref. [60], © Christopher Eames et al./CC BY 4.0) and lead iodide (degradation product). **d** Processed average red color component of films degraded at $T = 55$ °C as a function of time.

spin-coated perovskite film samples that tend to have rather high variations, especially when aged.

First, we attempt to uncover the differential equation governing perovskite degradation directly from experimental data (Workflow (1)). A simple way to analyze reaction rate orders is to fit the data to pure 0th, 1st, and 2nd order dynamics (Supplementary Fig. 3). These equations do not fit the data, showing that the environmental degradation of MAPI does not follow a simple *n*-th order kinetics. This motivates the use of PDE-FIND. We apply sparse regression to the whole experimental dataset with a broad function library consisting of polynomials of $U$ up to order 5, sine and cosine of $U$, polynomials of $t$ up to order 3, the square root of $t$, $U$ multiplied with polynomials of $t$, temperature $T$ and adjusted negative exponent of $\frac{1}{T}\left(\exp\left(-\frac{100}{T}\right)\right)$. While we choose a broad set of candidate functions, the choice of function library is critical and determines the outcome of PDE-FIND (candidate function libraries considered are in Supplementary Table 2). We find that sine and cosine terms are not selected by PDE-FIND—indicating as a sanity check that the algorithm correctly identifies that periodicity is not a feature of the dynamics. Polynomials of $t$ and $U$ times the polynomials of $t$, which correspond to the Avrami equation, are not included in the chosen library or assigned very small weights. To understand how well the obtained DE represents our data, we compare the derivative estimated by our DE to the numerical derivative obtained from the experimental data. While certain trends in the derivative are captured, errors exist because of the variance in our experimental data (Supplementary Fig. 4). Refinements to the approach are thus needed.

We proceed to narrow the application of PDE-FIND, by applying PDE-FIND to the averaged data at each temperature individually to extract the governing ODE. Using the averaged data helps us deal with sample-to-sample variance. Since all environmental conditions were almost identical for samples degraded at a particular temperature but aging tests of each temperature were conducted one after another (introducing differences, e.g., in sample storage times and exact equipment atmosphere), we aim to reduce the influence of variance-inducing conditions by applying PDE-FIND at each temperature separately. First, we apply PDE-FIND with a large library as described in the previous paragraph. Here too, we see that sine and cosine of $U$, polynomials of $t$ and $U$ times the polynomials of $t$ are either removed from the library or have small coefficient values. We exclude these terms in further analysis. Then, we apply PDE-FIND with 1st to 5th order polynomial libraries. We find that with the 1st order polynomial library, PDE-FIND is unable to find an equation that fits the derivative of our data (Fig. 3a). All other libraries from 2nd order polynomial to 5th order polynomial appear to fit the derivative of our data with significant accuracy (Fig. 3a, b). When these differential equations are integrated, they have the same S-shape as our experimental data (Fig. 3c). The 2nd order polynomial library is the most minimal library that fits our data without high error. The functional form of this ODE is:

$$\frac{dU}{dt} = a_0 + a_1 U + a_2 U^2 \tag{2}$$

We also notice a trend in the values of the fitting coefficients with respect to temperature—especially in the case of the 2nd order polynomial library (Fig. 3d, Supplementary Fig. 5). The slope of the curve changes between 55 °C and 65 °C, the temperature at which a well-known MAPI phase transition[55,56] occurs. This may indicate that the phase transition affects the degradation mechanism, but is not experimentally confirmed in this work.

Next, we evaluate the effect of variance on PDE extraction by comparing the above results (obtained on the low-variance experimental dataset) with the same workflow applied to the high-variance data (Supplementary Fig. 6). After averaging multiple curves ($U(t)$) for each temperature, the results are qualitatively similar for a constrained function library of polynomials of 2nd order—the obtained coefficients have the same sign and order of magnitude (Supplementary Table 3). This indicates that PDE-FIND can fit even high-variance experimental data when appropriately averaging over multiple samples. To quantify the effect of sample-to-sample variance, we apply PDE-FIND to each curve individually. As expected, PDE-FIND extracts a large variance in coefficient values. The values of coefficients vary as much as 60% with the low variance dataset and up to 90% with the high variance datasets for $T = 55$ °C.

## Results on simulated data

Now, we evaluate the effect of noise on PDE extraction using simulated data. We use the non-linear least-squares method to fit

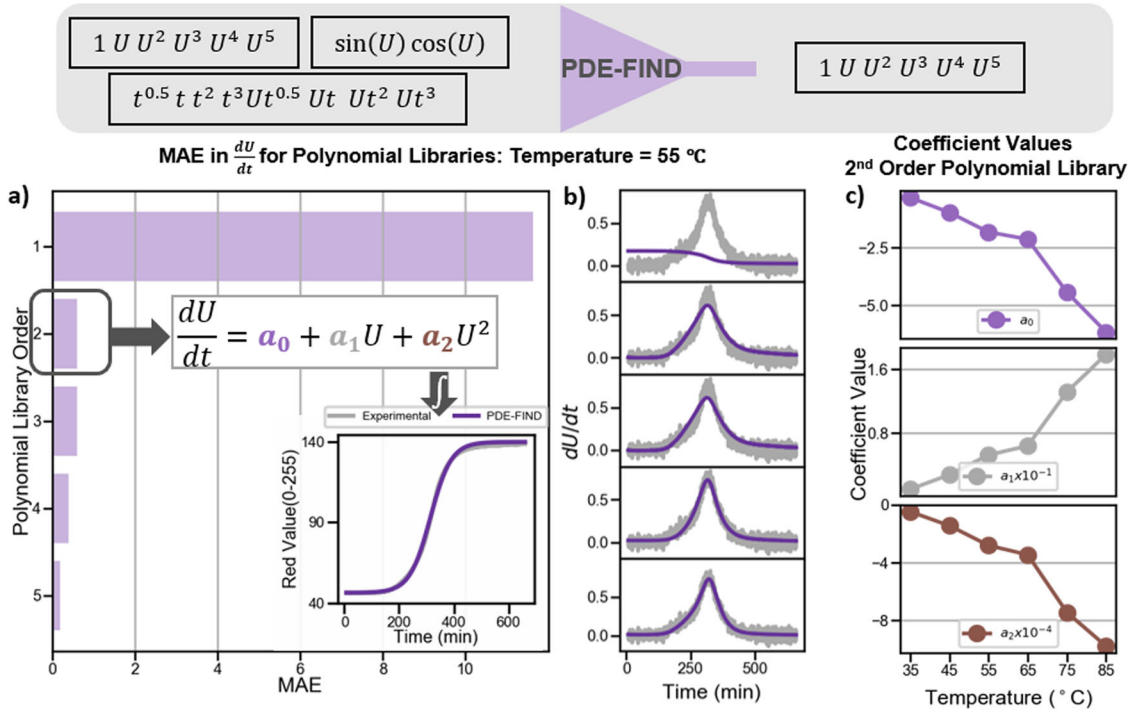


**Fig. 3 PDE-FIND results with experimental data. a** A bar plot shows the MAE between the actual experimental derivative (smoothened) and the value of the derivative estimated using the differential equation identified by PDE-FIND. Inset: Comparison of the experimental data with the curve obtained by integrating the equation identified by PDE-FIND with 2nd order polynomial library. **b** Comparison of the *dU/dt* calculated from experimental data for $T = 55$ °C and estimated from PDE-FIND for 2nd order polynomial library to 5th order polynomial libraries. Results for other temperatures are available in Supplementary Fig. 8. **c** Coefficient values estimated by PDE-FIND as a function of temperature for 2nd order polynomial library.

our experimental data to the Verhulst logistic equation[57] and the Arrhenius equation, as shown in the Methods section. We produce both noise-free simulated data and simulated data with Gaussian noise (Workflow (2)) with this model.

We apply sparse regression to the simulated dataset at each temperature individually to discover the governing ODEs with libraries ranging from 2nd to 5th order polynomials. With the noise-free data, PDE-FIND's identified DEs fit the derivative as well as the data on integration of the DE with significant accuracy for libraries from 2nd order to 5th order. In the case of the 2nd order polynomial library, both the underlying differential equation and the fitting parameters are identified with significant accuracy, as shown in Fig. 4. We know that the underlying governing equation for this dataset (which we defined as $\frac{dU}{dt} = a_0 + a_1U + a_2U^2$) does not have any terms higher than order two, thus the higher-order coefficients (e.g., $a_3, a_4, a_5, \ldots$ of functional terms $U^3, U^4, U^5 \ldots$) are equal to zero. PDE-FIND assigns small non-zero values to these functional forms, although they are not set to zero. In the case where sine and cosine are added to the library, the algorithm correctly identifies that these terms do not represent the dynamics and are set to zero exactly. The MAE between the exact numerical derivative and one estimated from the differential equation identified by PDE-FIND is of order $10^{-7}$ (when derivative varies from 0 to 1). This indicates that PDE-FIND works well for simulated curves with zero noise. Thus, with the candidate function library constrained to polynomials of *U*, PDE-FIND is able to identify the same ODE that fits the data at each temperature.

We then add varying amounts of Gaussian noise to this simulated equation at different temperatures. First, we consider the effect of varying amounts of noise at a fixed temperature of 55 °C, as indicated by the black box in Fig. 4a. We add up to 5% noise, which is typical in many experimental settings. The equation identified by PDE-FIND yields an S-shaped curve similar to the noise-free simulated curve upon integration (Fig. 4d) for up to 5% noise, after which the DE identified by PDE-FIND does not seem to model the dynamics. We compare the error of estimating the parameter values in the differential equation describing the simulated data. At 5% Gaussian noise the error of the fitting parameters increases to almost 80% (Fig. 4b). The resulting integrated curve has MAE is 6 (on a color scale of 0–255) relative to the 'ground truth' noise-free simulated curve (Fig. 4c, d). In addition, PDE-FIND is no longer able to threshold sin and cosine terms to 0, as it even fits the noise with sinusoidal pattern.

We then consider different temperatures at the same noise level. The Verhulst logistic equation model becomes increasingly steep and shifts to the left with higher temperature. PDE-FIND successfully identifies this trend. It appears that the MAE is higher for equation extraction at higher-temperature data. This could be because of noise obscuring PDE-FIND's ability to fit steeper peaks accurately.

## DISCUSSION

There remain many complex systems that have eluded quantitative analytic descriptions or even characterization of a suitable choice of variables in many disciplines such as biology, finance and materials science. With today's state-of-the art equipment, acquiring large quantities of data has never been easier. As put by Rackauckas et al.[58], *'the well-known adage 'a picture is worth a thousand words' might well be 'a model is worth a thousand datasets.'*.

Scientific ML enables unique insights into MAPI degradation in this work. It is a promising method that can be used to uncover governing equations through data, especially when the derivation of physical laws using first principles is challenging. In our study, we demonstrate that PDE-FIND identifies an underlying rate equation for the degradation of perovskite solar cells. MAPI degradation does not follow a simple single-order reaction rate

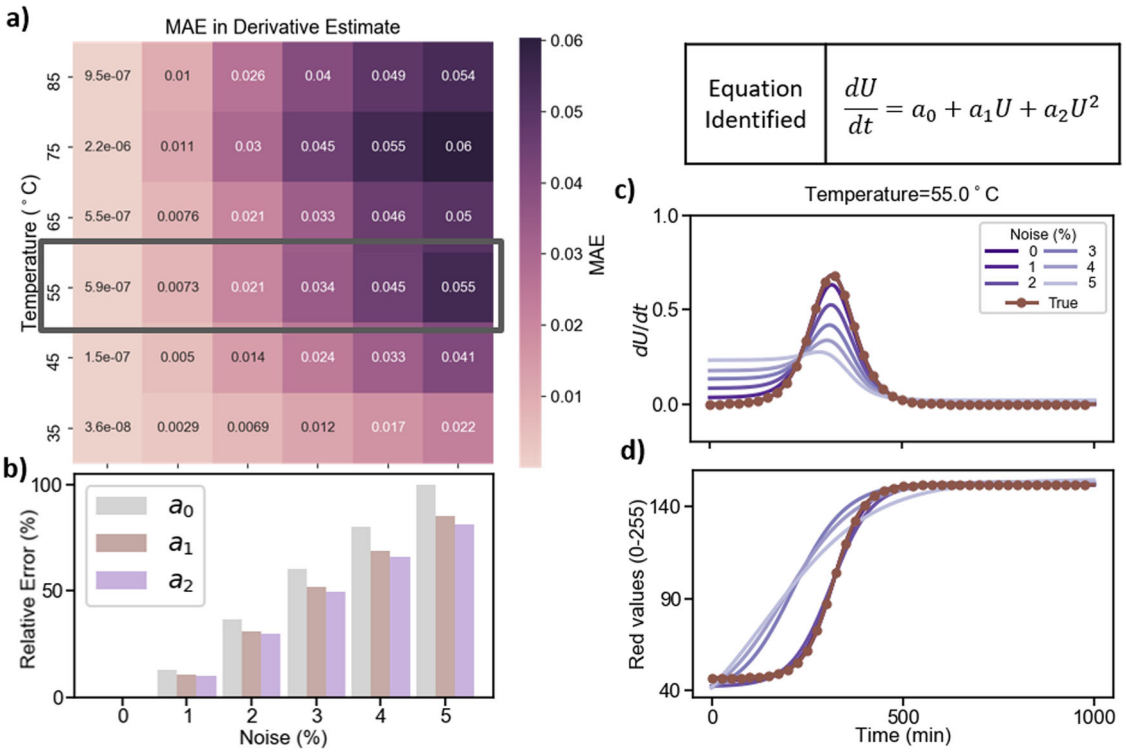


**Fig. 4 PDE-FIND results. a** Heatmap of the mean absolute error (MAE) between the exact numerical derivative and the derivative estimate obtained from the DE identified by PDE-FIND at different noise levels at different temperatures for the 2nd order polynomial library. **b** Relative error between the exact fitting parameter weight and those estimated by PDE-FIND at different noise levels for $T = 55$ °C data. **c** Comparison of the exact numerical derivative at $T = 55$ °C with derivative estimated from the DE identified by PDE-FIND at different levels of noise. **d** Comparison of the exact solution at $T = 55$ °C with solution curves obtained by integrating the DE identified by PDE-FIND at different levels of noise.

law, defined as:

$$\frac{dU}{dt} = kU^n \tag{3}$$

where, $n$ is the order of the reaction and $U$ is the concentration of the species. In our system, this equation does not yield a good fit for $n = 0$, 1 or 2. The S-shaped dynamics we see in our study have been reported in other studies involving MAPI degradation as well[45,48–50]. Some articles report that the degradation results from nucleation and growth of $PbI_2$ crystals[45,48], supporting the hypothesis that the kinetics follows the Johnson–Mehl–Avrami–Kolmorgorov or simply, the Avrami equation[46,47]:

$$\frac{\partial U'}{\partial t} = a_0 t^{n-1} - a_1 U' t^{n-1} \tag{4}$$

$$U'(t) = 1 - \exp(-kt^n) \tag{5}$$

where,

$$U' = \frac{U(t) - \min(U)}{\max(U - \min(U))}$$

And $a_0$, $a_1$, $n$, and $k$ are fitting constants.

The Avrami equation represents dynamics where degradation starts at nucleation spots on the film and these spots of degraded material grow radially, diffusion-limited. Some recent studies have presented an alternate hypothesis of self-propagating or auto-catalytic kinetics[49,50], which is described by another differential equation, the logistic function (discussed in the "Methods" section Eqs. (7), (8)). In this study, we build a large library of candidate terms for the DE– polynomials of $U$, that make up the logistic function, and polynomials of $t$ and $U$ multiplied with polynomials of $t$, which feature in the Avrami equation. PDE-FIND determines that the simplest ODE that fits our experimental dataset best is of the form (Fig. 3),

$$\frac{dU}{dt} = a_0 + a_1U + a_2U^2 \tag{6}$$

the Verhulst logistic function. This equation indicates that the reaction is first, propelled forward by the presence of the reactant as well as the product, leading to a rapid growth in the product that eventually saturates when it exhausts its reactants—a self-propagating reaction. While the Avrami equation (nucleation and growth) model is limited by the rate at which the reactant diffuses to the reaction site, the Verhulst logistic function (self-propelling kinetics) model is not limited by this because the reactant is already present at the reaction site. This is why we chose the logistic function model for the simulated dataset over the Avrami equations that has been used to model nucleation-growth reactions. The algorithm picks terms that describe self-propelling kinetics (2nd order polynomial library) as opposed to diffusion-limited nucleation and growth (Avrami equation). When visually inspecting the videos of degrading films, without the benefit of PDE-FIND, it can be hard to identify the underlying mechanism. In the example shown in Supplementary Fig. 7 [and Supplementary Video], one can see light areas of degraded material in the middle of the film degradation.

Equation (6) also offers insights that could help engineer more stable MAPI films. Once the degradation has begun, the autocatalytic nature suggests that degradation will continue, as the reaction products catalyze further MAPI degradation. Therefore, suppressing degradation means delaying the creation of the first reaction products for as long as possible. To engineer more stable MAPI films, this equation suggests that reducing MAPI degradation may be possible by reducing the density of nucleation points inside the material, including, e.g., by ensuring

that all $PbI_2$ precursors are fully converted during film formation, and possibly by using highly purified (i.e., devoid of contaminant particles) reagents in the film and adjacent layers that could nucleate $PbI_2$.

These insights bear consequence for researchers attempting to identify the underlying root cause(s) of perovskite degradation, as well as those modeling or predicting the (accelerated) degradation of these materials. If indeed this is a nucleation and growth phenomenon, little can be done to halt the growth of degraded regions once the initial nucleation event occurs. Therefore, to improve phase stability of perovskite films, an emphasis can be placed on identifying the nucleation points of these phase transformations, and inhibiting them, perhaps through improved precursor purification to remove impurities, improved control of the nucleation process, improved processing to remove growth catalysts, and improved packaging to prevent ingress of exogenous gasses. Changes to the film composition may increase the nucleation energy barrier; therefore, further investigation of stoichiometry optimization may be warranted in combination with the above.

We demonstrate the application of a Scientific ML tool, PDE-FIND on MAPI degradation data. When applied to experimental data, PDE-FIND identifies a differential equation that fits the data, when appropriate constraints are applied. In spite of the noise and variance in the dataset, only functions corresponding to the dynamics of the system are picked and the DEs show good agreement with the numerical derivatives. Our 'robustness analysis' with simulated data shows that PDE-FIND with a 2nd order polynomial library succeeds at identifying the differential equation describing the simulated data when up to 5% Gaussian noise is added. However, the error of the fitting parameters increases with noise, to almost 80%. With 5% noise, the resulting integrated curve has a 6 MAE relative to the underlying noise-free simulated curve but the coefficients differ by as much as 80%. With the addition of noise, PDE-FIND is unable to eliminate terms not in the DE (sine and cosine) and even fits the noise with these terms. Applying a de-noising filter may allow for higher levels of noise in the data, assuming that an appropriate filter is chosen.

Scientific ML methods can be immensely useful at uncovering governing equations of dynamical systems, if the data obtained has low noise or can be denoised by noise-reduction techniques. Data obtained through experiments is not devoid of measurement noise and de-noising the data adequately can be challenging. In addition, certain operating conditions cannot be fully controlled, leading to sample-to-sample variance making it hard to get rid of. Our contribution motivates the development of Scientific ML techniques that are more robust to noise as well as variance in data. Scientific ML, in its current state, is well-suited to be applied to domains where obtaining large quantities of low-noise data is possible, and will find more applications with methods that are robust to noise.

We show that Scientific ML has the potential to accelerate the understanding of materials degradation and the reliability optimization of perovskite materials. Extracting physical laws may facilitate the definition of acceleration factors for aging tests and also help in the prediction of perovskite solar cell degradation under varying environmental conditions. Not only does scientific machine learning aide us with understanding the underlying scientific phenomena better, it may also enable faster simulations and better extrapolations beyond our experimental datasets. The conclusions of any given materials study may well be rendered more generalizable by identifying underlying equations governing the observations.

## METHODS

### Data collection

For the experimental portion of our study (Workflow (1) in Fig. 1), the input is the experimental data obtained from degrading MAPI films. MAPI film synthesis conditions follow those of the Materials subsection of the Experimental procedures section of ref. [18] (More information in Supplementary Methods). Our experimental data is shown in Fig. 2.

We monitored the degradation of MAPI based on the color change of the material. As MAPI films decompose, they change their color from initial black (majority MAPI) to degraded yellow (minority MAPI). We acquired images of the degrading films with 0.5-min temporal resolution and processed them to obtain the average red, blue and green color components of the films as a function of time (Fig. 2a, Supplementary Fig. 1). There are limits to the use of film color as a proxy; for example, in mixed perovskites, degradation may proceed via phase de-mixing into pure phases that may be dark in color, and camera-based imaging technique should be modified or complemented with other metrology, e.g., X-ray diffraction[18].

For the study of noise robustness (Workflow (2) in Fig. 1), we generate simulated degradation data to analyze how noise obfuscates the identification of underlying DEs. We apply a non-linear least-squares method to fit the experimental data (e.g., those shown in Fig. 2d) to the Verhulst logistic equation[57] to model the S-shaped curve. This is a reasonable assumption because the logistic function is used to describe the thermal decomposition dynamics of several materials[49,50,59]. We obtain,

$$U = M + \frac{U_o K e^{kt}}{(K - U_o) + U_o e^{kt}}, \tag{7}$$

$$\frac{\partial U}{\partial t} = k(U - M)\left(1 - \frac{(U - M)}{K}\right) \tag{8}$$

where $U_o$ is the initial concentration, $k$ is growth rate, $K$ is the carrying capacity and $M$ is a fitting constant. In the context of MAPI degradation, $M$, $U_o$, and $K$ can be considered as fitting parameters. The growth rate $k$ varies with temperature according to the Arrhenius equation:

$$k = Ae^{\left(-\frac{E_a}{RT}\right)} \tag{9}$$

here, $E_a$ is the activation energy, $T$ is the temperature in Kelvin, $A$ is the pre-exponential factor and $R$ is the universal gas constant. We use this model to produce noise-free simulated data (labeled 'simulated') and simulated data with Gaussian noise (labeled 'simulated with Gaussian noise').

### Data analysis

First, we apply the sparse regression methodology PDE-FIND[19] to experimental data (Workflow (1)). We use the time-series from all the temperatures to infer the partial differential equation (PDE) defining the relationship between MAPI degradation, temperature, and time. Then, we apply PDE-FIND to the time-dependent degradation data at each temperature, to infer the ordinary differential equation (ODE) that describes MAPI decomposition at a particular temperature. To study the effect of noise, we apply PDE-FIND to simulated data with and without Gaussian noise (Workflow (2)).

The library of potential candidate functions consists of polynomials of $U$, polynomials of time $t$, sine and cosine of $U$, temperature $T$, and other non-linear functions of $U$, $t$, and $T$ (Supplementary Table 1). Differentials are calculated by finite difference with convolutional smoothing using a 1D Gaussian kernel. Once a large tall matrix ($\Theta(U)$) with all candidate functions is composed, we use sequential threshold ridge regression to identify which terms contribute to the dynamics described by the data as well as those terms' weights. The goal of this method is to find a sparse coefficient vector $\beta$ that only consists of the active features that best represent the time derivative $U_t$. The rest of the features are hard-thresholded to zero. The loss functions are follows ($\lambda_2$ and $\lambda_0$ are the L-2 and L-0 regularization penalties, respectively, more details can be found in the supplementary information of ref. [19]):

$$\hat{\beta} = \arg\min_{\beta}\left(\|\Theta(U)\beta - U_t\|_2 + \lambda_2\|\beta\|_2\right) \tag{10}$$

for a given $\widehat{\text{tol}}$, where $\widehat{\text{tol}}$ is:

$$\widehat{\text{tol}} = \arg\min_{\text{tol}}\left(\|\Theta(U)\beta - U_t\|_2 + \lambda_0\|\beta\|_0\right) \tag{11}$$

## DATA AVAILABILITY

The experimental dataset analyzed during the current study, the simulated data and labeled data supporting the findings of this study, and the data comprising the figures in this paper are all available in the following GitHub repository: https://github.com/PV-Lab/PDE-Extraction.

## CODE AVAILABILITY

The code used for data analysis in this study is available in the following GitHub repository: https://github.com/PV-Lab/PDE-Extraction.

Received: 22 June 2021; Accepted: 22 February 2022;
Published online: 20 April 2022

## REFERENCES

1. Newton, I. *Philosophiae Naturalis Principia Mathematica* (A. et JM Duncan, 1833).
2. Einstein, A. Does the inertia of a body depend upon its energy-content. *Ann. Phys.* **18**, 639–641 (1905).
3. Russell, J. L. Kepler's laws of planetary motion: 1609–1666. *Br. J. Hist. Sci.* **2**, 1–24 (1964).
4. Heit, E. Properties of inductive reasoning. *Psychonomic Bull. Rev.* **7**, 569–592 (2000).
5. Pitt, M. A. & Myung, I. J. When a good fit can be bad. *Trends Cogn. Sci.* **6**, 421–425 (2002).
6. Christopoulos, A. & Lew, M. J. Beyond eyeballing: fitting models to experimental data. *Crit. Rev. Biochem. Mol. Biol.* **35**, 359–391 (2000).
7. Roberts, S. & Pashler, H. How persuasive is a good fit? A comment on theory testing. *Psychol. Rev.* **107**, 358–367 (2000).
8. Otto, F. et al. Decomposition of the single-phase high-entropy alloy CrMnFeCoNi after prolonged anneals at intermediate temperatures. *Acta Materialia* **112**, 40–52 (2016).
9. Starke Jr, E. A. et al. *Accelerated Aging of Materials and Structures*. *Accelerated Aging of Materials and Structures* (National Academies Press, 1996).
10. McKeen, L. W. *The Effect of Long Term Thermal Exposure on Plastics and Elastomers*. *The Effect of Long Term Thermal Exposure on Plastics and Elastomers* (Elsevier Inc., 2013).
11. Simmons, C. B. et al. Deactivation of metastable single-crystal silicon hyperdoped with sulfur. *J. Appl. Phys.* **114**, 243514 (2013).
12. Macan, J., Brnardić, I., Orlić, S., Ivanković, H. & Ivanković, M. Thermal degradation of epoxy - Silica organic - Inorganic hybrid materials. *Polym. Degrad. Stab.* **91**, 122–127 (2006).
13. Choi, W., Huh, H., Tama, B., Park, G. & Lee, S. A neural network model for material degradation detection and diagnosis using microscopic images. *IEEE Access* **7**, 92151–92160 (2019).
14. Severson, K., Attia, P., Jin, N., Perkins, N. & Jiang, B. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**, 383–391 (2019).
15. Nash, Will, Drummond, T. & Birbilis, N. A review of deep learning in the study of materials degradation. *npj Mater. Degrad.* **2**, 1–12 (2018).
16. Entekhabi, E., Haghbin Nazarpak, M., Sedighi, M. & Kazemzadeh, A. Predicting degradation rate of genipin cross-linked gelatin scaffolds with machine learning. *Mater. Sci. Eng. C.* **107**, 110362 (2020).
17. Hartono, N. T. P. et al. How machine learning can help select capping layers to suppress perovskite degradation. *Nat. Commun.* **11**, 1–9 (2020).
18. Sun, S. et al. A data fusion approach to optimize compositional stability of halide perovskites. *Matter* **4**, 1305–1322 (2021).
19. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
20. Champion, K., Lusch, B., Nathan Kutz, J. & Brunton, S. L. Data-driven discovery of coordinates and governing equations. *Proc. Natl Acad. Sci. USA* **116**, 22445–22451 (2019).
21. Atkinson, S. et al. Data-driven discovery of free-form governing differential equations. Preprint at http://arxiv.org/abs/1910.05117 (2019).
22. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Computat. Phys.* **378**, 686–707 (2019).
23. Long, Z., Lu, Y. & Dong, B. PDE-Net 2.0: learning PDEs from data with a numeric-symbolic hybrid deep network. *J. Comput. Phys.* **399**, 108925 (2019).
24. Long, Z., Lu, Y., Ma, X. & Dong, B. PDE-Net: learning PDEs from data. in *35th International Conference on Machine Learning, ICML 2018* Vol. 7, 5067–5078 (2017).
25. Raissi, M., Yazdani, A. & Karniadakis, G. E. Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. *Science* **367**, 1026–1030 (2020).
26. Yin, M., Zheng, X., Humphrey, J. D. & Karniadakis, G. E. Non-invasive inference of thrombus material properties with physics-informed neural networks. *Computer Methods Appl. Mech. Eng.* **375**, 113603 (2021).
27. Zanna, L. & Bolton, T. Data-driven equation discovery of ocean mesoscale closures. *Geophys. Res. Lett.* **47**, e2020GL088376 (2020).
28. Schmelzer, M., Dwight, R. P. & Cinnella, P. Discovery of algebraic Reynolds-stress models using sparse symbolic regression. *Flow., Turbulence Combust.* **104**, 579–603 (2020).
29. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
30. Cichos, F., Gustavsson, K., Mehlig, B. & Volpe, G. Machine learning for active matter. *Nat. Mach. Intell.* **2**, 94–103 (2020).
31. Brunton, S. L., Noack, B. R. & Koumoutsakos, P. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech. 2020* **52**, 477–508 (2019).
32. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explainable machine learning for scientific insights and discoveries. *IEEE Access* **8**, 42200–42216 (2020).
33. Juarez-Perez, E. J. et al. Photodecomposition and thermal decomposition in methylammonium halide lead perovskites and inferred design principles to increase photovoltaic device stability. *J. Mater. Chem. A* **6**, 9604–9612 (2018).
34. Divitini, G. et al. In situ observation of heat-induced degradation of perovskite solar cells. *Nat. Energy* **1**, 1–6 (2016).
35. Fan, Z. et al. Layer-by-layer degradation of methylammonium lead tri-iodide perovskite microplates. *Joule* **1**, 548–562 (2017).
36. Smecca, E. et al. Stability of solution-processed $MAPbI_3$ and $FAPbI_3$ layers. *Phys. Chem. Chem. Phys.* **18**, 13413–13422 (2016).
37. Conings, B. et al. Intrinsic thermal instability of methylammonium lead trihalide perovskite. *Adv. Energy Mater.* **5**, 1500477 (2015).
38. Schwenzer, J. A. et al. Thermal stability and cation composition of hybrid organic–inorganic perovskites. *ACS Appl. Mater. Interfaces* **13**, 15292–15304 (2021).
39. Yang, J., Siempelkamp, B. D., Liu, D. & Kelly, T. L. Investigation of $CH_3NH_3PbI_3$ degradation rates and mechanisms in controlled humidity environments using in situ techniques. *ACS Nano* **9**, 1955–1963 (2015).
40. Kim, N. K. et al. Investigation of thermally induced degradation in $CH_3NH_3PbI_3$ perovskite solar cells using in-situ synchrotron radiation analysis. *Sci. Rep.* **7**, 1–9 (2017).
41. Han, Y. et al. Degradation observations of encapsulated planar $CH_3NH_3PbI_3$ perovskite solar cells at high temperatures and humidity. *J. Mater. Chem. A* **3**, 8139–8147 (2015).
42. Nie, W. et al. Light-activated photocurrent degradation and self-healing in perovskite solar cells. *Nat. Commun.* **7**, 1–9 (2016).
43. Lee, S. W. et al. UV degradation and recovery of perovskite solar cells. *Sci. Rep.* **6**, 1–10 (2016).
44. Abdelmageed, G. et al. Effect of temperature on light induced degradation in methylammonium lead iodide perovskite thin films and solar cells. *Sol. Energy Mater. Sol. Cells* **174**, 566–571 (2018).
45. Bastos, J. P. et al. Model for the prediction of the lifetime and energy yield of methyl ammonium lead iodide perovskite solar cells at elevated temperatures. *ACS Appl. Mater. Interfaces* **11**, 16517–16526 (2019).
46. Avrami, M. Kinetics of phase change. I: General theory. *J. Chem. Phys.* **7**, 1103–1112 (1939).
47. Fanfoni, M. & Tomellini, M. The Johnson-Mehl-Avrami-Kolmogorov model: a brief review. *Nuovo Cim. Della Soc. Ital. di Fis. D. - Condens. Matter, At., Mol. Chem. Phys., Biophys.* **20**, 1171–1182 (1998).
48. Tran, C. D. T., Liu, Y., Thibau, E. S., Llanos, A. & Lu, Z. H. Stability of organometal perovskites with organic overlayers. *AIP Adv.* **5**, 087185 (2015).
49. Ellis, C. L. C., Javaid, H., Smith, E. C. & Venkataraman, D. Hybrid perovskites with larger organic cations reveal autocatalytic degradation kinetics and increased stability under light. *Inorg. Chem.* **59**, 12176–12186 (2020).
50. Fu, F. et al. I2 vapor-induced degradation of formamidinium lead iodide based perovskite solar cells under heat-light soaking conditions. *Energy Environ. Sci.* **12**, 3074–3088 (2019).
51. Asghar, M. I., Zhang, J., Wang, H. & Lund, P. D. Device stability of perovskite solar cells—a review. *Renew. Sustain. Energy Rev.* **77**, 131–146 (2017).
52. Boyd, C. C., Cheacharoen, R., Leijtens, T. & McGehee, M. D. Understanding degradation mechanisms and improving stability of perovskite photovoltaics. *Chem. Rev.* **119**, 3418–3451 (2019).
53. Khenkin, M. V. et al. Consensus statement for stability assessment and reporting for perovskite photovoltaics based on ISOS procedures. *Nat. Energy* **5**, 35–49 (2020).

54. Hashmi, S. G. et al. Long term stability of air processed inkjet infiltrated carbon-based printed perovskite solar cells under intense ultra-violet light soaking. *J. Mater. Chem. A* **5**, 4797–4802 (2017).
55. Whitfield, P. S. et al. Structures, phase transitions and tricritical behavior of the hybrid perovskite methyl ammonium lead iodide. *Sci. Rep.* **6**, 1–16 (2016).
56. Rajendra Kumar, G. et al. Phase transition kinetics and surface binding states of methylammonium lead iodide perovskite. *Phys. Chem. Chem. Phys.* **18**, 7284–7292 (2016).
57. Tsoularis, A. & Wallace, J. Analysis of logistic growth models. *Math. Biosci.* **179**, 21–55 (2002).
58. Rackauckas, C. et al. Universal differential equations for scientific machine learning. Preprint at http://arxiv.org/abs/2001.04385 (2020).
59. Burnham, A. K. Use and misuse of logistic equations for modeling chemical kinetics. *J. Therm. Anal. Calorim.* **127**, 1107–1116 (2017).
60. Eames, C. et al. Ionic transport in hybrid lead iodide perovskite solar cells. *Nat. Commun.* **6**, 7497 (2015).

## ACKNOWLEDGEMENTS

The authors thank Kathleen Champion, Samuel Rudy, Zichao Long, and Steven Atkinson for helpful discussions regarding Scientific ML. This work was supported by Defense Advanced Research Projects Agency (DARPA) under contract no. HR001118C0036 (R.N., J.T., A.T.), TotalEnergies SE research grant funded through MITei Sustng Mbr 9/08 (A.T., S.S., Z.L.), and the U.S. Department of Energy (DOE) under Photovoltaic Research and Development (PVRD) program under Award no. DE-EE0007535 (Z.L.). This work was partially supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Advanced Manufacturing Office (AMO) Award Number DE-EE0009096 (R.N.). A.T. acknowledges the Alfred Kordelin Foundation.

## AUTHOR CONTRIBUTIONS

R.N., A.T., S.S., and T.B. conceived of and designed the study. C.B. and J.T. fabricated the samples. R.N. executed different aspects of the study such as the experiments and ML modeling. R.N., A.T., and T.B. wrote the paper while all co-authors contributed to reviewing the manuscript.

## COMPETING INTERESTS

Although our laboratory has IP filed covering photovoltaic technologies and materials informatics broadly, we do not envision a direct COI with this study, the content of which is open-sourced. Two of the authors (Z.L., T.B.) own equity in a startup company, Xinterra Pte Ltd, which applies machine learning to materials.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00751-5.

**Correspondence** and requests for materials should be addressed to Richa Ramesh Naik, Armi Tiihonen or Tonio Buonassisi.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.