Uncovering Perovskite Degradation Equations Using Scientific Machine Learning

by

Richa Ramesh Naik

Submitted to the Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computational Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Center for Computational Science and Engineering
May 21, 2021

Certified by .....
Tonio Buonassisi
Professor, Mechanical Engineering
Thesis Supervisor

Read by .....
Rafael Gomez-Bombarelli
Assistant Professor, Department of Materials Science

Accepted by .....
Youssef Marzouk and Nicolas Hadjiconstantinou
Co-Directors, Center for Computational Science and Engineering

Uncovering Perovskite Degradation Equations Using Scientific Machine Learning

by

Richa Ramesh Naik

Submitted to the Center for Computational Science and Engineering
on May 21, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Computational Science and Engineering

Abstract

Many important materials are metastable or unstable under certain operating regimes. The degradation mechanisms can be varied and complex, making the discovery of underlying differential equations (DEs) through a first-principles approach challenging. This invites the application of data-science methods to infer root causes. Traditionally, machine learning (ML) applied to materials research has focused on optimization and regression over a limited training set. Inferring physical laws directly from data may allow the extraction of more generalizable scientific information that enables one to understand underlying mechanisms. In this study, we apply scientific ML — a blend of traditional scientific mechanistic modeling (differential equations) with machine learning methodologies — to identify differential equations governing the degradation of methylammonium lead iodide perovskite (MAPI), a material with known instability under environmental stress. We explore scientific ML applied to simulated and experimental datasets, obtaining equations that describe the temperature- and time-dependencies of MAPI degradation. Our method of choice is sparse regression method PDE-FIND [57]. We find that the underlying DE governing MAPI degradation corresponds to the Verhulst logistic function, often used to describe autocatalytic or self-propagating kinetics. This thesis demonstrates the application of scientific ML in practical materials science systems, highlighting the promise and challenges associated with ML-aided scientific discovery.

Thesis Supervisor: Tonio Buonassisi
Title: Professor, Mechanical Engineering

Acknowledgments

The satiation and euphoria that accompany the successful completion of a thesis would be incomplete without the mention of the people who made it possible.

This thesis is dedicated to my parents and sisters, for their endless love, support and encouragement. I big shout-out to my friends for being there for me throughout this journey.

I am greatly indebted to my mentor Armi Tihonnen for her enthusiastic support and valuable advice which inspired me greatly. I would like to thank Shijing Sun for mentoring me and motivating me to push my boundaries. I am deeply grateful to Janak Thapa and Clio Batali for lending me a helping hand with experiments, this study would not have been possible without them. I owe my wholehearted thanks and appreciation to all the members of PVLab for their insightful conversations and advice, constructive suggestions, positive and supportive attitude and continuous encouragement.

I would like to express a deep sense of gratitude to my Thesis supervisor Tonio Buonassisi. I am grateful for the time he gave me in spite of being on the other side of the globe. His valuable contribution and guidance has been certainly indispensable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

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 [48], Einstein's mass-energy equivalence equation [43], Kepler's laws of planetary motion [58] and other physical principles were uncovered through careful interpretation of experimental data and inductive reasoning [33]. The Edisonian approach of trial and error is difficult with systems that are yet to be understood fully– the set of feasible equations capturing the physics is enormous. This is the case for many areas, such as finance, biology and materials science where there are complex systems for which we do not have quantitative analytic descriptions. A model could be fit to the experimental data through regression. However, without knowledge of the underlying physics, verifying the result of a model obtained from curve-fitting is challenging [16, 55].

Many materials are unstable under environmental stress. For example, alloys [65, 50] polymers [46], doped silicon [63] and hybrid materials [44] experience structural changes at elevated temperatures. The degradation pathways can be complex and not directly obvious when examining the experimental data. In these cases, data-science methods can facilitate the interpretation. Machine learning (ML) has been used to model and predict degradation [15, 62, 47, 29, 24] as well as to optimize

process conditions to reduce material decomposition [29, 66]. Traditional data-science methods extrapolate poorly beyond the training data, which renders them of less use when trying to identify and decouple complex degradation pathways. 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 be used to extrapolate beyond the dataset and serve as foundation for physical interpretation of phenomena and scientific discovery.

1.2 Scientific Machine Learning

Herein, we use scientific ML, which combines regression-based ML with sparsity generating techniques in order to automatically learn equations. Scientific ML methods are well-suited to 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 helps to make simulations faster and extrapolate beyond the dataset at hand. Recently, many approaches aiming for this target have been presented in literature. We review some of them below.

1.2.1 Partial Differential Equation - Functional Identification of Nonlinear Dynamics [57]

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 the greedy algorithm. Some knowledge of the system being studied is necessary to appropriately select the library of candidate functions. This is the

method that is applied in this thesis. More details can be found in the section 2.

1.2.2 Sparse Identification of Nonlinear Dynamics [14]

Sparse Identification of Nonlinear Dynamics (SINDy) uses a custom deep autoencoder to find a coordinate system in which the dynamics of the system are sparse. A library of candidate functions is built based on low-dimensional representation and and sparse regression is to find the governing equations in the associated coordinate system. The sparse regression technique used is sequential thershold least-squares algorithm. The method lays equal importance on the discovery of a low-dimensional latent space and the model describing the dynamics. SINDy finds models are based on the latent space representation, making the inference of equations based on the actual features difficult. Moreover, some knowledge of the system being studied is necessary to appropriately set the dimension of the alternate coordinate system and build the library of candidate functions.

1.2.3 Physics-Informed Research Assistant for Theory Extraction [4]

Physics-Informed Research Assistant for Theory Extraction (PIRATE) is a generalized method for the discovery of differential equations using genetic programming. It uses computations over arbitrary compositions of functions, parameters, and potentially differential operators. The dynamical dataset is preprocessed through differential function representations. The differential equation is represented using a graph, where each node represents a mathematical operator, extracted from a library of user defined operators. These operators include basic mathematical operations (+, -, *, /) as well as differential operators. The leaves of the graph represent instances of fitted models. An evolutionary algorithm initializes, mutates and mates graphs. The parameters enter the graph as constant functions which are calibrated in a Bayesian manner using black-box variational inference.

1.2.4 Physics-Informed Neural Networks [52]

Physics-Informed Neural Networks (PINN) is a deep learning methodology to incorporate scientific insights into machine learning approaches to solve and discover partial differential equations. The system of interest, U is modelled using a deep neural network. Then, a *physics-informed* neural network is obtained by applying the chain rule of differentiation to the network approximating U using automatic differentiation [7]. This gives a physics-informed estimate of U_t.

1.2.5 Partial Differential Equation Network [42, 41]

The cited papers above describe "PDE-NET," a transparent neural network capable of extracting the governing PDE from dynamical data available about the system without using any prior information about the system. It relies on the fact that certain neural networks such as ResNet [32] have a connection with ODEs/PDEs and can be merged with computational mathematical tasks. A feed forward deep neural network is used. Time derivatives are described using forward Euler method. Spatial derivatives are estimated using convolutionl layers stacked together. In order to obtain the analytical form of the equation describing data, a symbolic neural network called SymNet is designed by the authors. This method is promising however the current implementation has been designed for the use cases in the paper and does not allow for application to general problems.

These methods have shown great promise in several applications [53, 72, 73, 59]. The automatic discovery of scientific laws and principles is at the frontier of machine learning that awaits application to materials science [13] and other domains [17, 11, 56]. More articles can be found here [60, 74, 21, 45, 70].

1.3 Lead-Halide Perovskites

Halide perovskite materials have potential to provide high performing and cost-effective solar energy. In the past 10 years, the efficiencies of perovskite-based so-

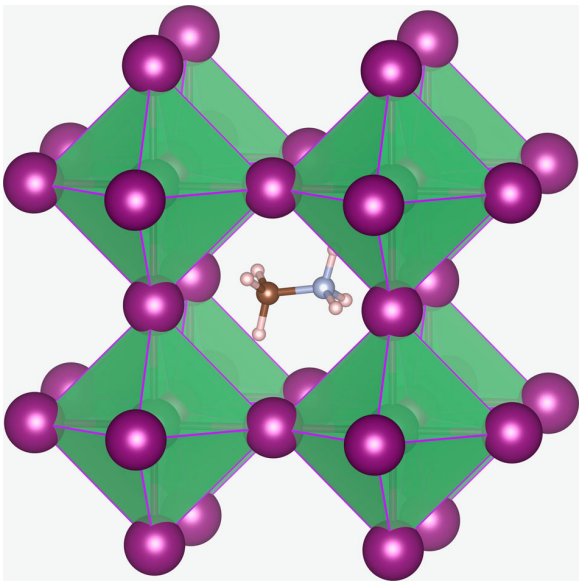

Figure 1-1: Methyl Ammonium Lead Iodide Crystal Structure

lar cells have risen from 3.8% in 2009 to 25.5% in 2020 [54, 37], competing strongly with silicon-based photovoltaic technology. Solar cells using perovskites as the light-absorbing layer could be manufactured cheaply and efficiently. Using renewable energy more extensively is the need of the hour to help mitigate climate change [19, 38]. Perovskite photovoltaic technology deployment could help produce cheap solar energy.

However, perovskites degrade at elevated temperature [34, 18, 22, 25, 61, 64], humidity [71, 36, 28], and illumination [49, 40]. 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 PbI_2 via reaction [1]:

$$MAPbI_3 \longrightarrow PbI_2 + [CH_3NH_3{}^+ + I-] \longrightarrow PbI_2 + CH_3NH_2 + HI$$

Smecca *et al.* [64] demonstrate that the rate of MAPI degradation obeys an Arrhenius-type law. Their data suggests 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℃. Bastos *et al.* [6] hypothesize that the thermal degradation of MAPI is defined by the Avrami equation [5, 26] of nucleation and growth. The Avrami equation has also been used to describe degradation kinetics in humid air [67]. Recently, studies have shown that halide perovskite degradation follows autocatalytic reaction kinetics [23] with the hypothesis that the degradation is propagated by iodine vapors [27]. The derivation of exact kinetics as well as Arrhenius-type dependence through first principles 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 thesis, 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 [3, 10, 35], and provide a generalizable pathway to identify degradation modes also in other materials research domains.

Chapter 2

Methods

2.1 Experimental Methods

Our experimental workflow is shown in Figure 2-1. 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°C to 85°C in our in-house environmental chamber (Figure 2-1 (a)). This degradation chamber simulates harsh environmental conditions of elevated temperature and humidity that solar cells can be subject to, in a controlled manner. It has been described in detail in [66]. The low value of illumination is to ensure low light-induced degradation while guaranteeing visibility within the chamber. Humidity is kept in a low range of 20±5% relative humidity, the lowest setting possible in our experimental setup. This is to ensure humidity-driven decomposition is minimized. We vary the nominal temperature from 35°C to 85°C to emulate temperature pertaining to indoors and harsh tropical outdoors. We wish to study how high temperature affects the degradation of MAPI, while other factors are held constant.

One hundred and eight samples were grown under low-variance conditions (labeled "low-variance experimental"); ninety-eight samples were grown under high-variance conditions (labeled "high-variance experimental") (Figure B-13). The red value curves from low and high variance datasets for samples degraded at 35 °C are shown in Figure 2-2. Unless specified otherwise, we assume "experimental" data in this paper refers

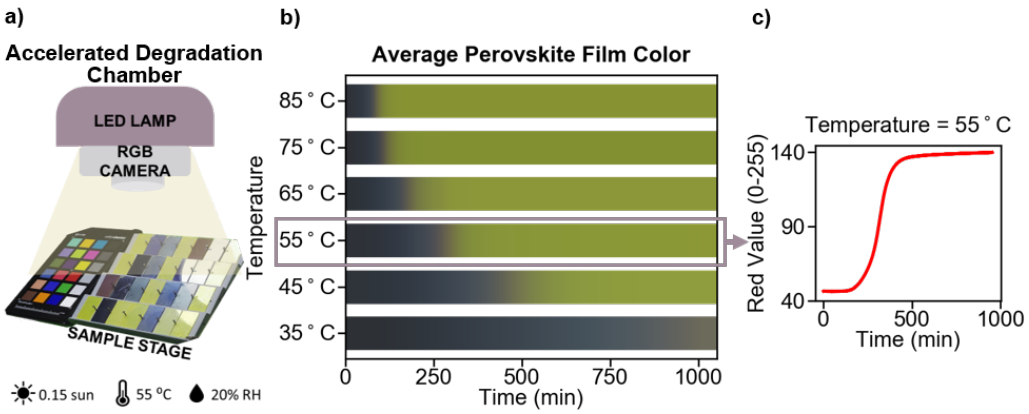

Figure 2-1: Experimental Methods

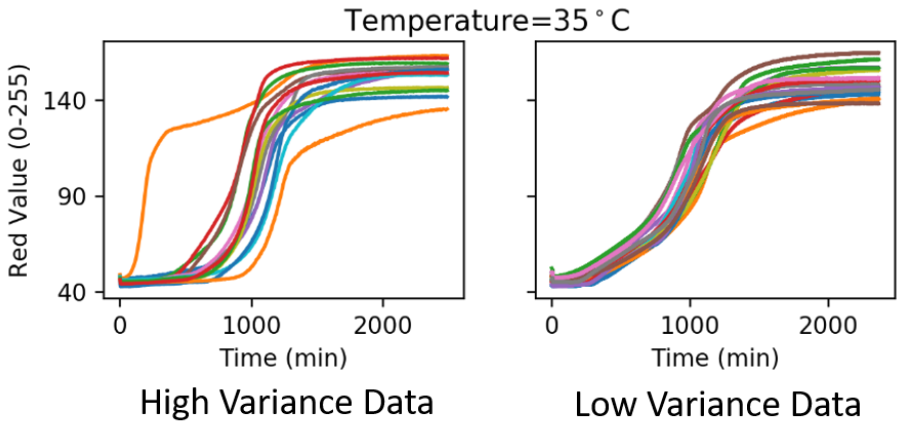

Figure 2-2: Comparing High and Low Variance Samples for 35°C

to the low-variance sample set. We quantify the noise in our experimental data to be of the order of 0.35% for both high-variance and low-variance experimental data sets. 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 value varies from 0-255). 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. The factors contributing to the variance in the samples are related to the film fabrication process.

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)

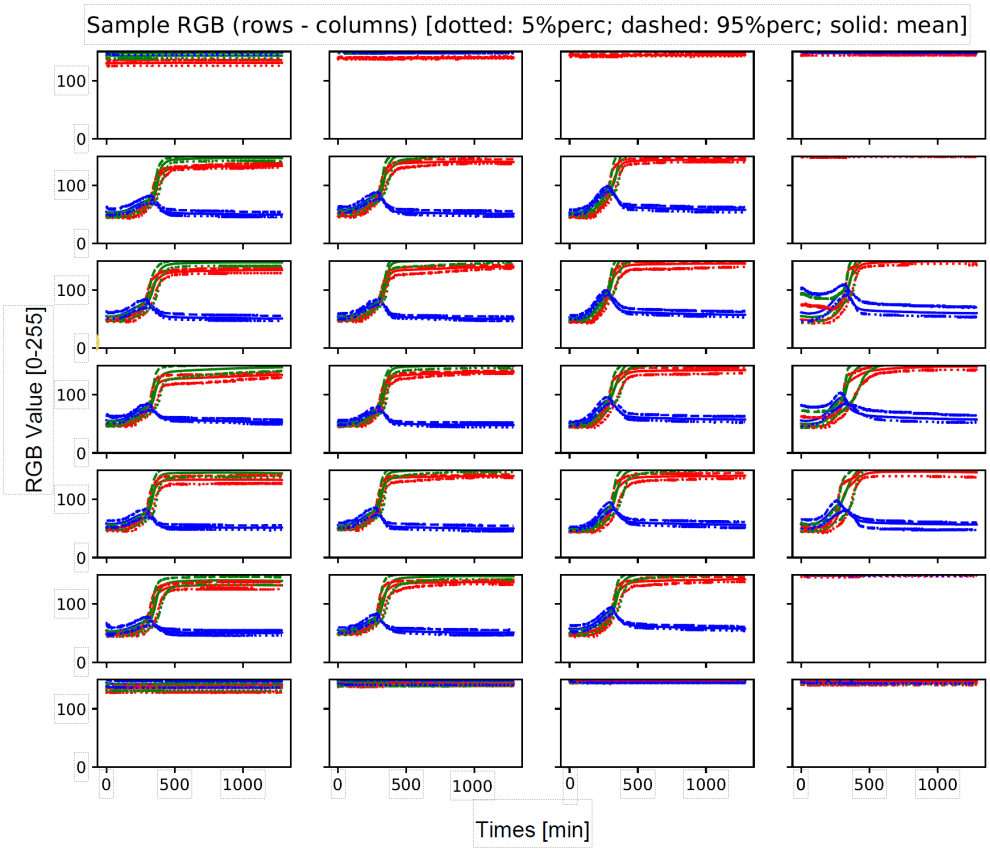

Figure 2-3: RGB Curves from a Degradation test. Flat lines mean that particular location in the sample holder was empty.

to degraded yellow (minority MAPI) (Figure 2-1 (b)). We acquired images of the degrading films with 0.5-minute temporal resolution and processed them to obtain the average red, blue and green color components of the films as a function of time (Figure 2-3). The red color time-series is chosen for further analysis because it sufficiently captures the temporal perovskite decomposition behavior at the MAPI bandgap, as shown in the Supporting Information of reference [29] (Figure 2-1 (c)).

2.2 Computational Methods

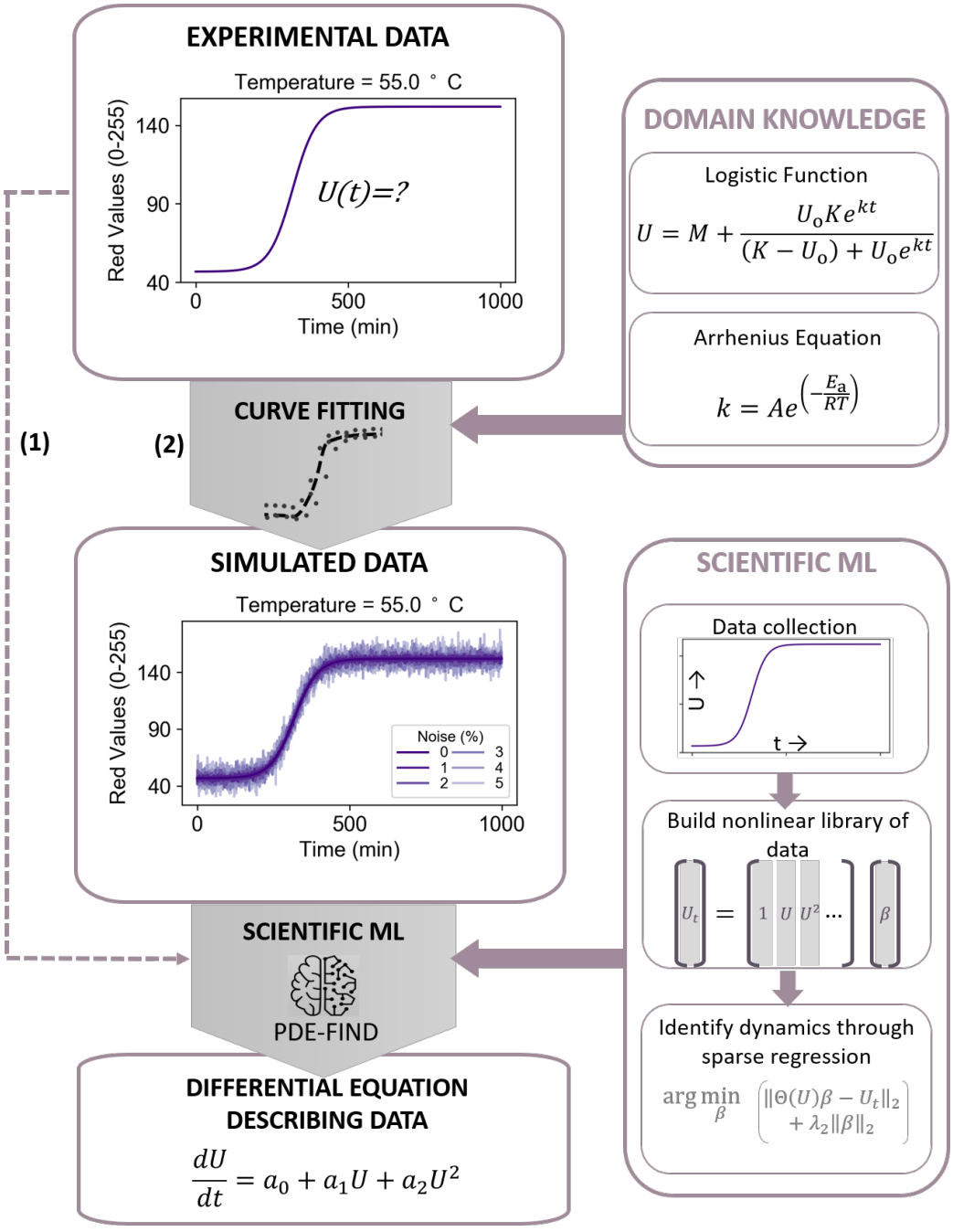

Figure 2-4: Computational Workflow

Experimental data is often noisy. In order to study how noise affects the identification of underlying differential equations, we produce a simulated dataset. Least-squares

regression is used to fit the Verhulst logistic function [68] combined with the Arrhenius equation to experimental data (e.g., those shown in Figure 2-1 (c)). This is a reasonable assumption because the logistic function is used to describe the thermal decomposition dynamics of several materials [23, 27, 12]. Analysing the accuracy of equation-identifying methods is simplified with simulated data since the underlying equation as well as the corresponding weights are known precisely. The equations are as follows,

$$U = M + \frac{U_0 K e^{kt}}{(K - U_0) + U_0 e^{kt}} \tag{2.1}$$

$$\frac{\partial U}{\partial t} = k(U - M)\left(1 - \frac{U - M}{K}\right) \tag{2.2}$$

where U_0 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_0 and K can be considered as fitting parameters. The growth rate k varies with temperature according to the Arrhenius equation:

$$k = Ae^{\frac{-E_a}{RT}} \tag{2.3}$$

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 "noiseless simulated") and simulated data with Gaussian noise (labelled "noisy simulated").

2.2.1 Data Analysis

A summary of our data-analysis workflow is shown in Figure 2-4. Our goals for this study are two-fold:

1. Uncover the underlying differential equation corresponding to perovskite degradation using sparse regression methodology PDE-FIND [57] (Workflow (1)) and

Polynomials of U	1, U, U^2, U^3, U^4, U^5
Polynomials of t	$t^{0.5}$, t, t^2, t^3
$U\times$ Polynomials of t	$Ut^{0.5}$, Ut, Ut^2, Ut^3
Sine and Cosine of U	$sin(U)$, $cos(U)$
Functions of T	T, $exp(\frac{100}{T})$

Table 2.1: Candidate Functions Used

2. 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))

This is represented by the two workflows (dashed arrows) in the figure. For the first objective, the input data is the red-color curves from degrading MAPI films. About 18 samples are degraded at six temperatures from 35°C to 85°C, giving us 96 time-series. We use these time-series data 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. 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 (Table 2.1).

Differentials are calculated by finite difference with convolutional smoothing using a 1D Gaussian kernel. The convolutional smoothing is used to reduce the amplification of noise that is often seen when numerical derivatives are computed. A large tall matrix ($\Theta(U)$) is composed, with each column representing a candidate function and each row corresponding to a particular time for a sample.

Sequential Threshold Ridge Regression (STRidge)

The sparse regression algorithm we employ is called sequential threshold ridge regression (STRidge). Our goal is to identify which terms contribute to the dynamics described by the data as well as those terms' weights. The algorithm is summarized

below:

Algorithm 1: STRidge_T - Thresholding Step

Result: Return $\hat{\beta}$ through the application of thresholding for a value of `tol`

`inc_f` = [0,1, ... # features] (included features)

while *iters* $<$ *max_iters* **do**

  $\beta = \arg\min\left(\|\Theta(U)[:, \texttt{inc\_f}]\beta - U_t\|_2 + \lambda_2\|\beta\|_2\right)$ (ridge regression)

  `inc_f` = [i if β[i] $>$ `tol`] (select contributing features)

  β[i not in `inc_f`] $= 0$ (hard-thresholding)

  iters = iters + 1

end

$\hat{\beta} = \arg\min_\beta\left(\|\Theta(U)[:, \texttt{inc\_f}]\beta - U_t\|_2 + \lambda_2\|\beta\|_2\right)$ (final ridge regression)

return $\hat{\beta}$

Algorithm 2: STRidge - Optimizing Step

Result: Optimize the value of `tol` and return corresponding $\hat{\beta}$

$\Theta(U) \rightarrow [\Theta(U)^{\text{train}} \quad \Theta(U)^{\text{test}}]$

$U_t \rightarrow [U_t^{\text{train}} \quad U_t^{\text{test}}]$ (80-20 train-test split)

$\hat{\beta} = \arg\min_\beta\left(\|\Theta(U)^{\text{train}}\beta - U_t^{\text{train}}\|_2\right)$ (baseline result)

$L_{\text{best}} = (\|\Theta(U)^{\text{test}}\beta - U_t^{\text{test}}\|_2 + \lambda_0\|\beta\|_0)$ (baseline loss)

for *iters* $= 1, \ldots$ *tol_iters* **do**

  $\beta = \texttt{STRidge\_T}(\Theta(U)^{\text{train}}, U_t^{\text{train}}, \texttt{tol})$ (thresholding step)

  $L = \|\Theta(U)^{\text{test}}\beta - U_t^{\text{test}}\|_2 + \lambda_0\|\beta\|_0$ (calculate loss)

  if $L < L_{best}$ **then**

    $L_{\text{best}} = L$

    `tol` $= tol + \delta$`tol` (increase `tol`)

    $\hat{\beta} = \beta$

  else

    `tol` $= \max([0, \texttt{tol} - 2\delta\texttt{tol}])$ (reduce `tol`)

    δ`tol` $= \frac{\delta\texttt{tol}}{\texttt{tol\_iters} - \texttt{iters}}$ (reduce δ`tol`)

    `tol` $= tol + \delta$`tol` (increase `tol` by smaller δ`tol`)

  end

end

return $\hat{\beta}$

The goal of this method is to find a sparse coefficient vector β 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. Thus, the algorithm conducts two tasks:

1. Identifying the terms contributing to the dynamics, that is, the columns of $(\Theta(U))$ that add up to give U_t; and

2. Assigning appropriate weights to these terms.

The first task is achieved by performing several rounds of ridge regression. Each time, a few terms that do not contribute significantly to the dynamics are hard-thresholded to zero. This means, the weight corresponding to these terms is set to 0, effectively eliminating them from the library for that round. This process is repeated till convergence is reached. To determine which terms "contribute significantly", a parameter called tolerance is optimized in each round. The initial value of tolerance is a hyperparameter. After the contributing terms have been identified, ridge regression is used to determine the weight or coefficient value of each contributing term.

The loss functions for the two steps are as follows (λ_2 and λ_0 are the L-2 and L-0 regularization penalties respectively):

$$\hat{\beta} = \arg\ \min_{\beta} \left(\|\Theta(U)\beta - U_t\|_2 + \lambda_2\|\beta\|_2\right) \tag{2.4}$$

$$\hat{\texttt{tol}} = \arg\ \min_{\texttt{tol}} \left(\|\Theta(U)\beta - U_t\|_2 + \lambda_0\|\beta\|_0\right) \tag{2.5}$$

More details can be found in the supplementary information of reference [57]. All hyperparameters are optimized for using the `hyperopt` package [9].

Other Sparse Regression Algorithms

Lasso Regression Solved Using Fast Iterative Shrinkage Thresholding Algorithm (FISTA)

Lasso regression uses the L-1 norm to regularize weights which is a well-known sparsity-inducing norm [30]. The objective of lasso regression is as follows (λ_1 is the

L-1 regularization penalty):

$$\hat{\beta} = \arg\ \min_{\beta} \left(\frac{1}{2} \|\Theta(U)\beta - U_t\|_2^2 + \lambda_1 \|\beta\|_1 \right) \tag{2.6}$$

Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [8] is an accelerated proximal gradient method used to solve this objective function. It is a faster version of Iterative Thresholding Algorithm (ISTA) [20] that uses a first-order approximation along with the gradient descent algorithm. FISTA uses information from previous iterations to reach faster convergence rates.

Elastic-Net Regression Solved Using Fast Iterative Shrinkage Thresholding Algorithm (FISTA)

Elastic-Net regression consists of both L-1 and L-2 regularization. It introduces sparsity through the L-1 norm and shrinks together correlated features through the L-2 norm [31]. The objective function is as follows:

$$\hat{\beta} = \arg\ \min_{\beta} \left(\frac{1}{2} \|\Theta(U)\beta - U_t\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2 \right) \tag{2.7}$$

Like lasso regression, the solution can be found using FISTA.

Adaptive Forward-Backward Greedy Algorithm

Greedy algorithms recursively pick a solution that is sparse and locally optimal, eventually hoping to get to a globally optimal solution. Adaptive Forward-Backward Greedy Algorithm [75] uses forward greedy steps to find a sparse solution and backward steps periodically in order to correct any mistakes made in the forward iterations.

All the above mentioned methods were tried with synthetic datasets (*e.g.*, Burgers' equation, reaction diffusion equation). STRidge algorithm was picked for our analysis

as it performed better empirically across the different synthetic datasets. The authors of [57] also recommend the use of STRidge in their work.

Chapter 3

Results

3.1 Experimental Data

The workflow is represented in Figure 3-1. First, we fit our experimental data to the most simple kinetic equations: 0th, 1st and 2nd order rate -kinetic equations to find that these do not fit the data. We understand that the environmental degradation of methyl ammonium lead iodide (MAPI) does not follow simple n-th order reaction kinetics (Figure 3-2). This motivates the use of Scientific Machine Learning. Thus, the PDE-FIND algorithm is employed. The full low-variance dataset containing all the samples from all the temperatures is combined together. A broad set of candidate functions is used in the library (Table 2.1).

The STRidge algorithm as described in 2.2.1 is used to obtain a differential equation that describes the dynamics of the data. PDE-FIND eliminates sine and cosine terms, correctly discovering that periodicity is not a feature in the dynamics described by our dataset. Moreover, polynomials of t and $U\times$ polynomials of t are removed from the library or set to values very close to zero (See Table A.1). This is an interesting result as polynomials of t and $U\times$ polynomials of t correspond to the terms from the Avrami equation with different orders. To better understand how the obtained equation fits the experimental data, the derivative estimated from the obtained equation is compared with the numerical derivative with respect to time (Figure B-1) at various temperatures. While the equation captures the overall trends in the data, it

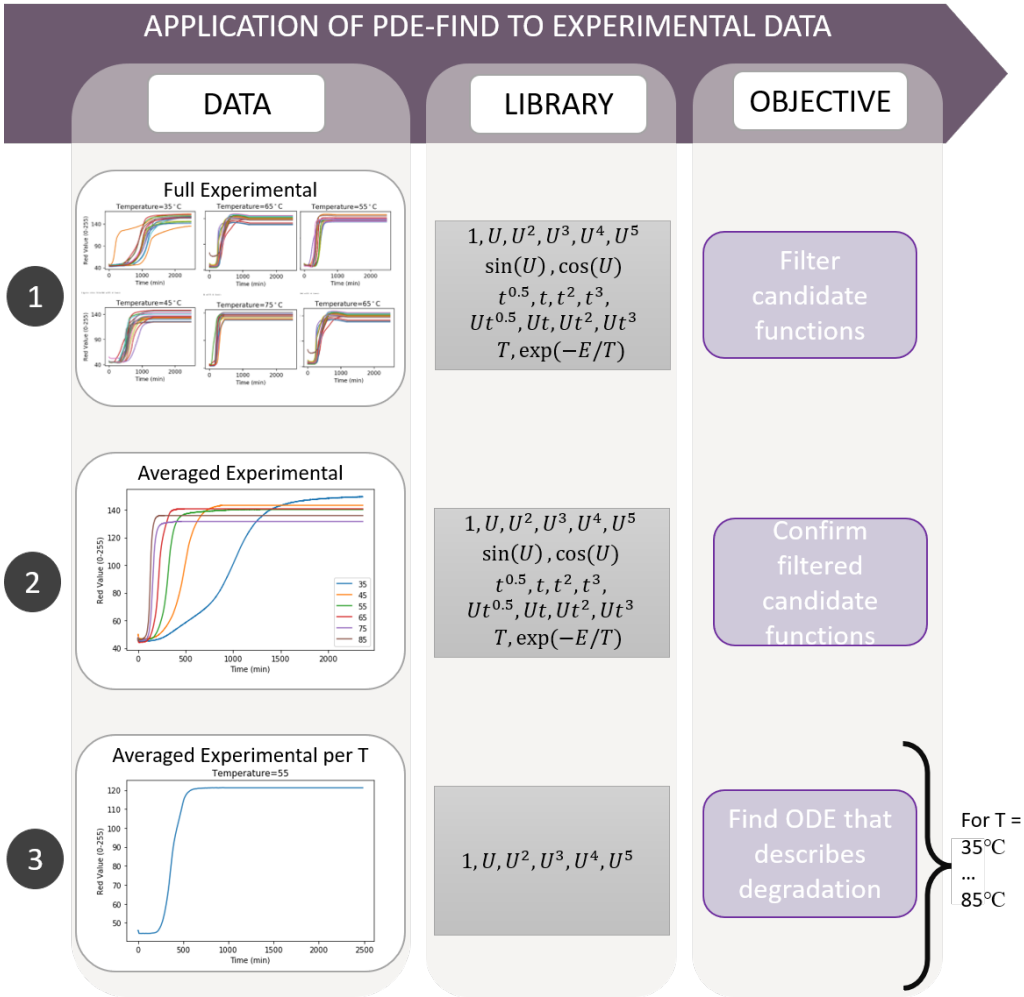

Figure 3-1: Experimental Data Analysis Workflow

can be seen in many cases that the error in fitting the derivative is quite large for certain temperatures (Figure 3-3). The averaged experimental data versus time at each temperature is also plotted with the curve obtained on integrating the equation identified by PDE-FIND (Figure B-1). We see that, when integrated, this equation does not match experimental data. The mean absolute error (MAE) bar plots (Figure 3-4) elucidate these results graphically. Refinements to the approach are thus required.

PDE-FIND is now applied to the averaged data at each temperature separately, with the aim to obtain an ordinary differential equation (ODE) that correctly describes the degradation of methyl ammonium lead iodide (MAPI) at a particular

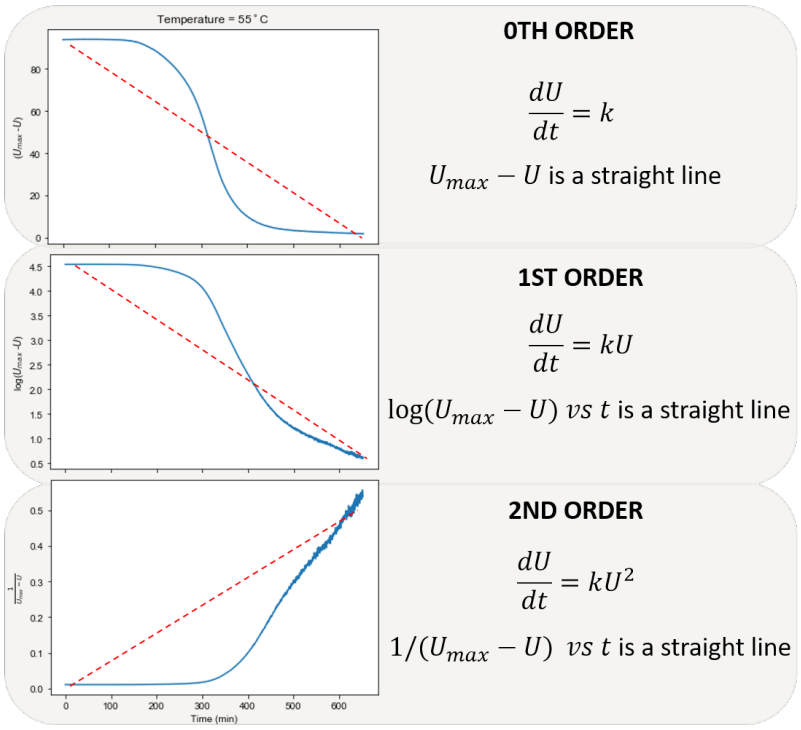

Figure 3-2: Fitting 0th, 1st and 2nd Order Kinetic Equations to Experimental Data

temperature. Using averaged data help us deal with sample-to-sample variance. Since all environmental conditions were exactly identical for samples degraded at a particular temperature, by applying PDE-FIND at each temperature separately, we hope to reduce the influence of other variance-inducing conditions. Thus PDE-FIND is applied with the broad library as described in Table 2.1, excluding candidate functions corresponding to T as they are no longer required. From this analysis, it is seen that the sine and cosine of U, polynomials of t and $U\times$ polynomials of t do not feature in the dynamics represented by the data. These results are in agreement with those obtained from the use of the full dataset, as described in the previous paragraph. Thus, the library is truncated to include polynomials of U only. PDE-FIND is run 5 times with data at each temperature, with the polynomial order of the library ranging from 1 to 5. We find that with the 1st order polynomial library, PDE-FIND is unable to find an equation that fits the derivative of our data (Figures B-2, B-3, B-4, B-5, B-6, B-7). All other libraries from 2nd order polynomial to 5th order polynomial appear to fit the derivative of our data with significant accuracy – the MAE being as low as

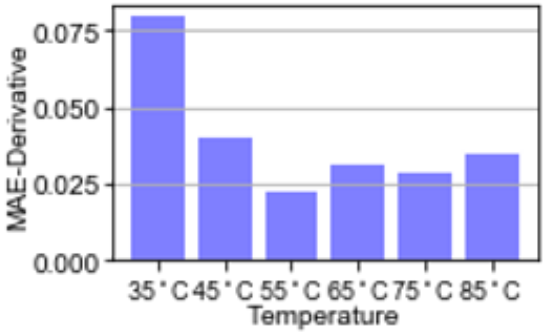

Figure 3-3: PDE-FIND Results with Full Experimental Data: MAE of Derivatives

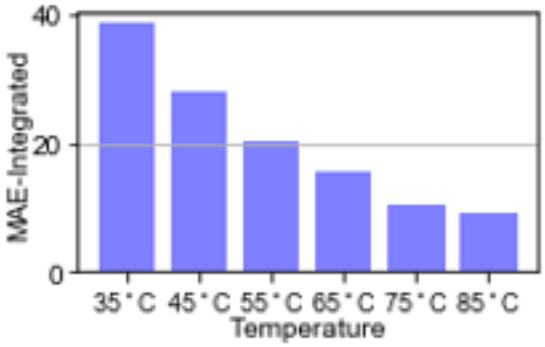

Figure 3-4: PDE-FIND Results with Full Experimental Data: MAE of Integrated Curve

10^{-3} (Figures B-2, B-3, B-4, B-5, B-6, B-7). When these differential equations are integrated, they have the same S-shape as our experimental data (Figures B-2, B-3, B-4, B-5, B-6, B-7). The 2nd order polynomial library seems to be the most minimal library that fits our data. The functional form of this ODE is:

$$\frac{dU}{dt} = a_0 + a_1 U + a_2 U^2 \tag{3.1}$$

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 (Figure 3-5). The slope of the curve changes between 55 °C and 65°C, the temperature at which a well-known MAPI phase transition [69, 39] occurs. This may indicate that the phase transition affects the degradation mechanism, but is not experimentally confirmed in this work. The trends in coefficient weights with temperature are not so evident with

other libraries (Figure B-8, B-9, B-10, B-11, B-12).

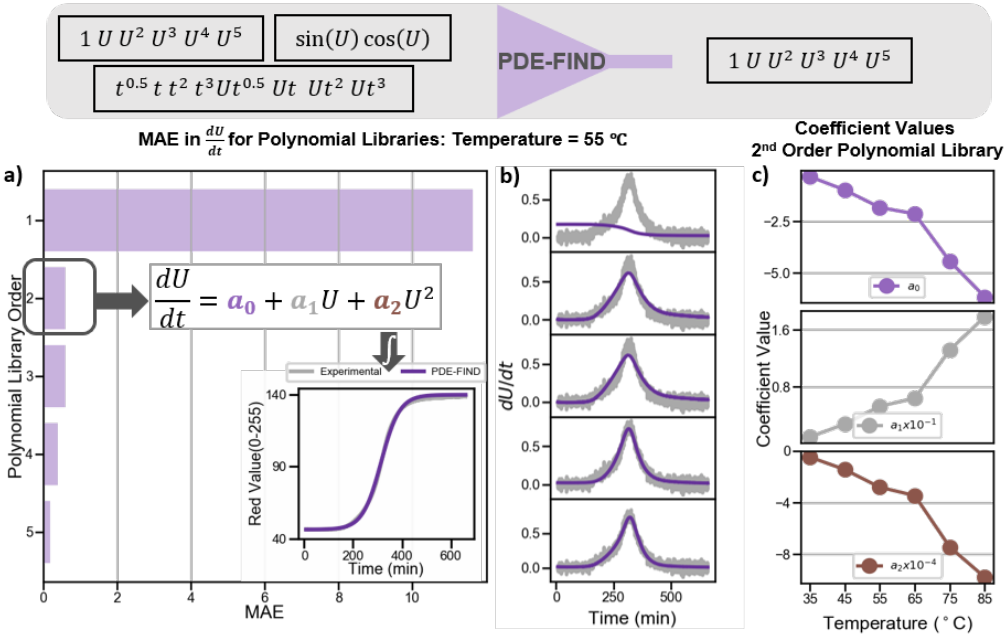

Figure 3-5: PDE-FIND results with experimental data: a) A bar plot shows the MAE between the actual experimental derivative (smoothened) and the value of the derivative estimated using the differential equation identified by PDE-FIND. Inset: Comparison of the experimental data with the curve obtained 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 c) Coefficient values estimated by PDE-FIND as a function of temperature for 2nd order polynomial library.

Next, the effect of variance on PDE extraction is evaluated by comparing the above results (obtained on the low-variance experimental dataset) with the same workflow applied to the high-variance data. 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 (Figure B-14). This indicates that PDE-FIND can fit even high-variance experimental data when appropriately averaging over multiple samples. Results are shown in Figures B-14 and B-15. 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 about as 30% with

the low variance dataset and up to 60% with the high variance datasets for T = 55 °C. This shows that one could increase the confidence in the results obtained through PDE-FIND when experimental variance is reduced.

3.2 Simulated Data

Figure 3-6 explains the workflow adapted in the analysis of simulated data.

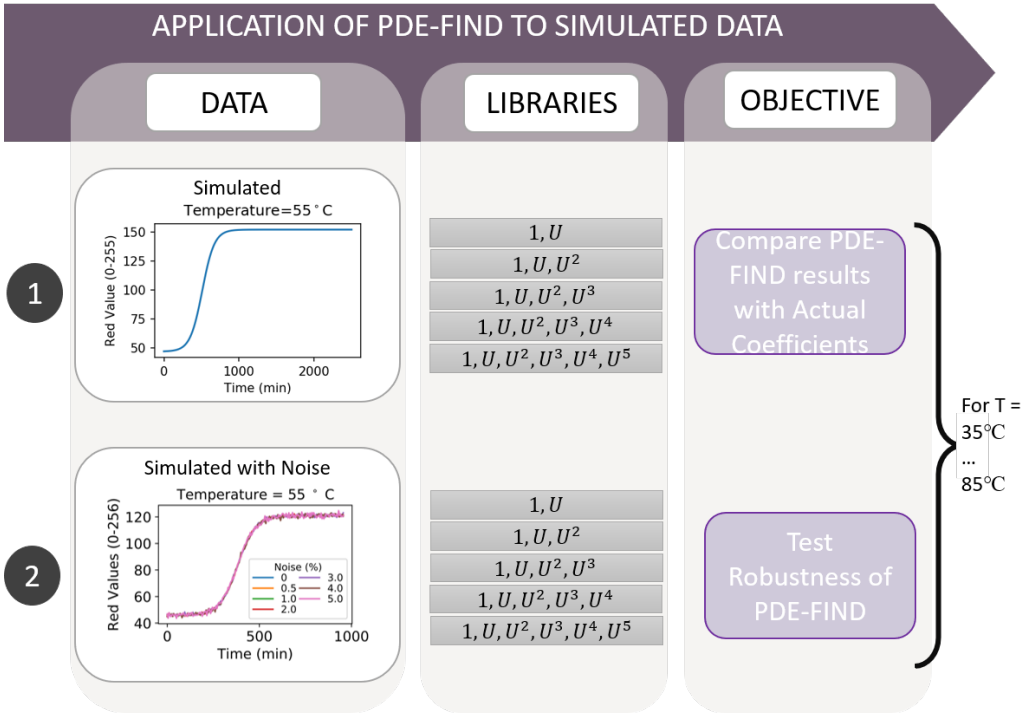

Figure 3-6: Simulated Data Analysis Workflow

3.2.1 Noiseless Data

To understand to the effect of noise on PDE-FIND's ability to identify of equations, we create a simulated dataset, by fitting the Verhulst logisitc function [68, 12] and the Arrhenius equation to experimental data as shown in the Methods section 2. We add up to 5% Gaussian noise to produce simulated data with artificial noise.

First, we apply PDE-FIND to the noiseless simulated data. The differential equation governing this dataset are known. As shown in section 2.2, the dynamics is

described by,

$$\frac{dU}{dt} = a_0 + a_1 U + a_2 U^2 \tag{3.2}$$

The values of the coefficients a_0, a_1 and a_2 are known. Sparse regression is applied to the simulated dataset with different polynomial libraries - ranging from 1st order to 5th order. It is known that the 2nd order polynomial library would fit the data best, corresponding to the differential equation obtained on differentiating the Verhulst logistic function with respect to time (2.2). As expected, we see that the equation obtained from the 1st order library does not fit the data. All the other polynomial libraries fit the simulated data with remarkable precision at all temperatures. The coefficient values of a_0, a_1 and a_2 of terms 1, U and U^2 match the true values. For 3th, 4th and 5th order libraries, functional terms apart from 1, U and U^2 are expected to be excluded from the dynamics. However, PDE-FIND does not eliminate these terms in the differential equation it outputs - they are allotted very small non-zero weights. Here, we see a small drawback of the sparse regression algorithm -its inability to assess and separate certain dynamics even when no noise is present. As a sanity check, we also add $\sin(U)$ and $\cos(U)$ to the 5th order polynomial library. In this case, PDE-FIND correctly recognises that there is no periodicity in the data and eliminates $\sin(U)$ and $\cos(U)$ from the library. The mean absolute error (MAE) between the true derivative and one estimated from PDE-FIND's differential equation when a 2nd order polynomial library is used is as low as 0.0003 (derivative varies from 0 to 1). When the differential equation identified by PDE-FIND is integrated with respect to time, the curve matches true data very accurately to yield mean absolute error less than 0.003 (on a color scale from 0 to 255). The error in fitting coefficients a_0, a_1 and a_2 is 0%. This shows that when PDE-FIND is provided with data containing little or no noise, it is able to identify the governing equation correctly. The results hold true for data from all temperatures. The same differential equation is identified across all temperatures.

3.2.2 Noisy Data

Artificial Gaussian noise is added to simulated data using `numpy`'s `random.randn()` function that draws samples from a normal distribution. 1%, 2%, 3%, 4% and 5% Gaussian noise (in proportion to the data mean) are added to the simulated curves as up to 5% noise is typical in most experimental settings. Figure 3-7 illustrates noisy data at 55°C.

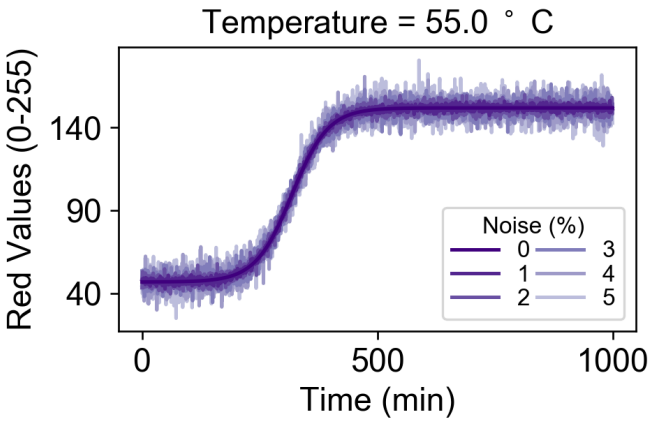

Figure 3-7: Noisy Simulated Data at 55 °C

First, we analyse the effect of varying levels of noise on a simulated curve from a particular temperature. As expected, noise obfuscates PDE-FIND's ability to accurately identify the underlying equation and the corresponding coefficient values. We apply sparse regression with the 2 candidate function libraries: one, consisting of U exponentiated up to 5th power, $\sin(U)$ and $\cos(U)$, the "expanded library". The second library consists of 1, U and U^2, the "constrained library", the Figure 3-8 summarizes of results from 55°C data with the constrained library.

Identification of Correct Contributing Functions

We apply PDE-FIND to noisy data while using the "expanded library", which is described in the preceding paragraph. The rationale behind this is to test PDE-FIND's ability to identify the terms contributing to the dynamics, even when the data is noisy. From the section 3.2 we see that PDE-FIND fails to eliminate certain extrane-

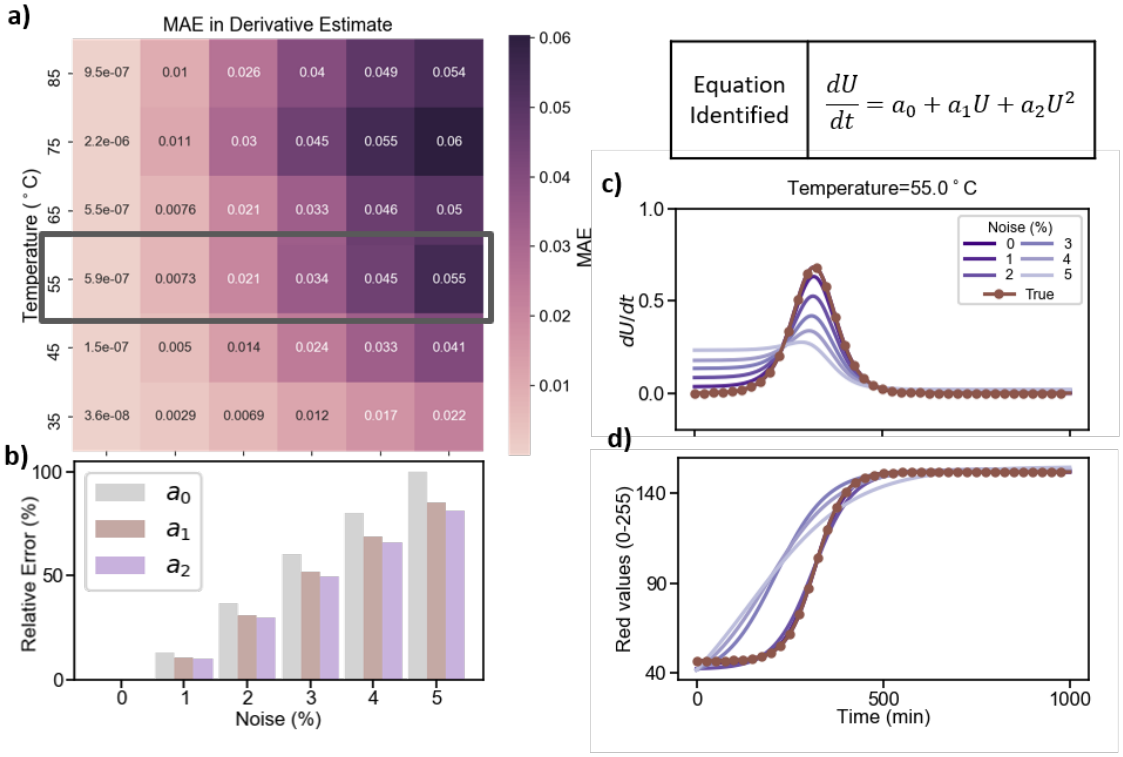

Figure 3-8: Results with Noisy Simulated Data at 55 °C

ous candidate functions in spite of being supplied with clean data. The results are not different with noisy data. Not only are the extraneous terms not eliminated, but also assigned higher weights as the amount of noise is increased. The $\sin(U)$ and $\cos(U)$ terms, which are eliminated when noiseless data is used, are now assigned sizable weights as well. Figure 3-9 shows how this occurs. Despite de-noising efforts through 1D convolutional smoothening, the noise in the numerical derivative is amplified and PDE-FIND mistakenly fits this noise to sinusoidal terms.

Identification of Coefficient Values (Weights)

With zero noise, extraneous terms are not fully eliminated. However, the weights attached to these terms are small and coefficient values of the terms that actually make up the differential equation are identified correctly with little error. With noisy data, this is not the case. The noise obfuscates the algorithm's ability to correctly identify contributing terms and thus assigns weights to each of them. With

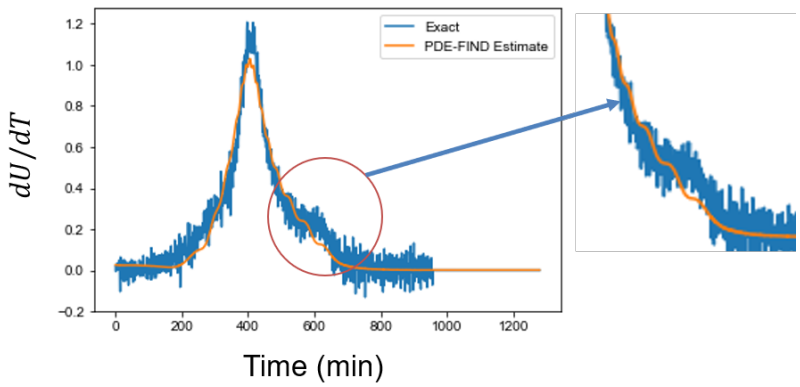

Figure 3-9: With Noisy Data, PDE-FIND Mistakenly Fits Sinusoidal Terms to Noise. The figure shows results with 65 °C data with 5% noise and the expanded library.

the extended library, the weights assigned to U^3, U^4 and U^5 increases substantially. $\sin(U)$ and $\cos(U)$ terms which are eliminated when no noise is added to the data are not removed from the final library in the presence of noise. This effect is seen even when the candidate library only consists of 1, U and U^2, the terms that are a part of the true solution. The weights of these terms are distorted by values as high as 85% with 55°C data. Figure 3-8(b) shows the relative error in the coefficients of 1, U and U^2 when up to 5% noise is added. As a result, the estimate of the derivative is affected, as shown in Figure 3-8(c) and B-16. When the equations identified by PDE-FIND are numerically integrated, they yield curves that differ from the true data (Figure B-17). For 55°C, we see that the MAE between the derivative estimate fom PDE-FIND and the true derivative is 0.055 (on the derivative scale from 0 to 1). Figure 3-8(a) depicts the full heatmap for different noise levels and temperatures. The MAE between the true curve and the one obtained by numerically integrating the differential equation identified by PDE-FIND is 15 (on a color scale from 0 to 255). Figure B-18 depicts the full heatmap for different noise levels and temperatures. These errors are much smaller compared to the amount error in the estimation of the coefficients. This shows that the weights are very sensitive to the amount of noise. This could also indicate the fact that these solutions are "sloppy" [2], meaning a vast number of combinations weights could fit the data with reasonable accuracy. This makes the collecting noiseless data extremely vital. Alternately, efficient de-noising

needs to be applied. With Gaussian noise higher than 5%, the equation does not fit the true data, as seen in Figure 3-10. The curve obtained on integrating the differential identified by PDE-FIND does not have S-shaped characteristics and does not match the true data.

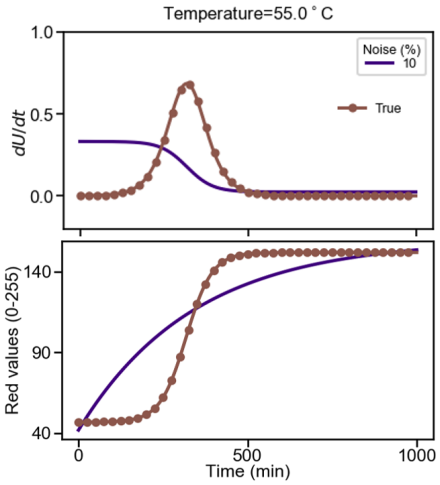

Figure 3-10: PDE-FIND Results at 10% Noise

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.

Chapter 4

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. [51], "the well-known adage 'a picture is worth a thousand words' might well be 'a model is worth a thousand datasets.'".

4.1 Scientific ML enables unique insights into MAPI degradation

Scientific ML 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 rate law, defined as:

$$\frac{dU}{dt} = kU^n \tag{4.1}$$

Here, n order of the reaction and U is the concentration of the species. In our

system, fitting with this simple reaction rate equation for $n = 0$, 1 and 2 does not yield a good fit . The S-shaped dynamics we see in our study have been reported in other studies involving MAPI degradation as well [6, 67, 23, 27]. Some articles report that the degradation results from nucleation and growth of PbI_2 crystals [6, 67], supporting the hypothesis that the kinetics follows the Johnson-Mehl-Avrami-Kolmorgorov or simply, the Avrami equation [5, 26]:

$$\frac{d\tilde{U}}{dt} = a_0 t^{n-1} - a_1 \tilde{U} t^{n-1} \tag{4.2}$$

$$\tilde{U}(t) = 1 - e^{-kt^n} \tag{4.3}$$

Where,

$$\tilde{U}(t) = \frac{U(t) - \min_t U}{\max_t U - \min_t U}$$

And a_0, a_1 and k are fitting constants. Recent studies have presented an alternate hypothesis of self-propagating or autocatalytic kinetics [23, 27], arguing that the reaction dynamics is described by the logistic function (Eqn 2.2). 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,

$$\frac{dU}{dt} = a_0 + a_1 U + a_2 U^2 \tag{4.4}$$

indicating 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. This is why we chose the logistic function model for the simulated dataset over the Avrami equations [26, 5] 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 nucleation and growth (Avrami equation). In Figure 4-1, snapshots of films degrading at 55 °C are shown where one can see light areas of degraded material in the middle of the film degradation. When we examine videos of degrading films, we often observe a nucleation and growth behavior, whereby (lighter) regions of degraded material nucleate at specific points in the films, and expand with time as shown in Figure 4-2.

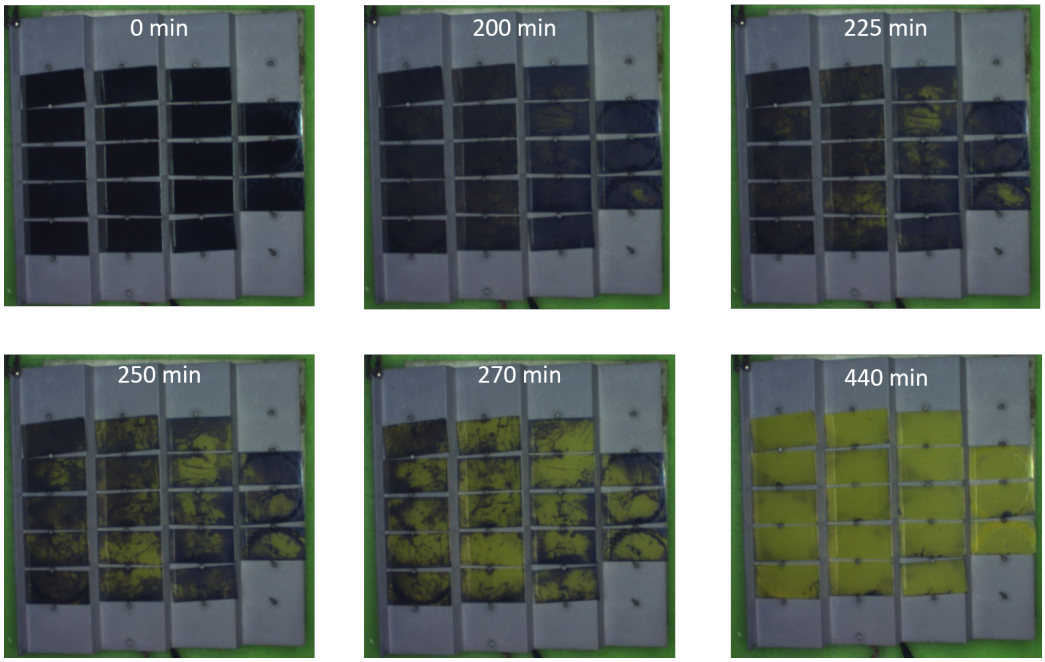

Figure 4-1: Snapshots of MAPI Films Degrading at 55 °C

Equation 2.2 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.

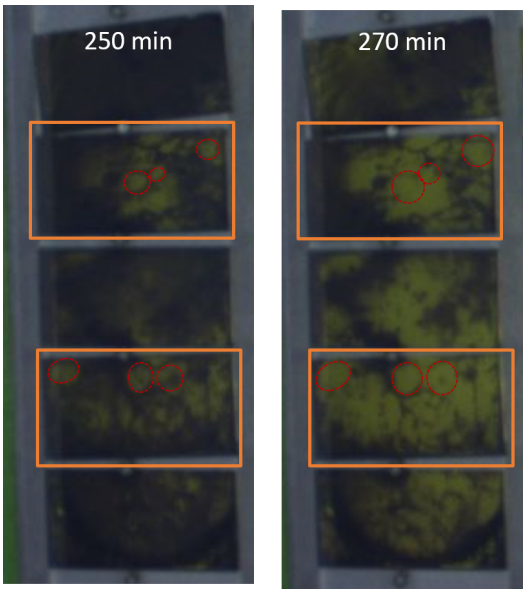

Figure 4-2: Yellow regions of PbI_2 grow radially

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 given the reaction is thereafter self-propagating. 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 packaging to prevent ingress of exogenous gasses, and improved control of the nucleation process. 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.

4.2 Evaluating Scientific ML's ability to accommodate noisy experimental data

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. With 5% noise, the resulting integrated curve has a 6.2 MAE relative to the underlying noise-free simulated curve. However, the error of the fitting parameters increases with noise, we see that 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.

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 denoising the data adequately can be challenging. Additionally, 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.

Chapter 5

Conclusions and Future Work

While machine learning (ML) has demonstrated impressive predictive capabilities, inductive reasoning and knowledge extraction remain elusive tasks. In this thesis, we use ML to infer underlying dynamical equations from experimental data of degrading organic-inorganic lead-halide perovskite thin films under environmental stressors (light, humidity, and temperature). Specifically, we apply the algorithm PDE-FIND[57], a sparse regression algorithm that automatically identifies the differential equation describing the dynamics from time-series data. PDE-FIND is an example of a broader class of algorithms called "Scientific ML," which is a blend of traditional scientific mechanistic modeling (differential equations) with machine learning methodologies.

We adapt and apply the PDE-FIND algorithm to methyl ammonium lead iodide (MAPI) experimental degradation data, to extract the underlying differential equation. We observe that MAPI degradation does not follow a simple 0th, 1st or 2nd order reaction. We see that the minimal equation that describes the degradation of MAPI is a second order polynomial, which corresponds to the Verhulst logistic function. (The algorithm chooses this equation as opposed to the Avrami equation, corresponding to diffusion-limited nucleation and growth phenomena). The Verhulst logistic function describes reaction kinetics analogous in functional form to autocatalytic or self-propagating reactions and population growth, which suggests the degradation product (PbI_2) is catalyzing further degradation of the film.

Experimental perovskite degradation data are obtained from accelerated environmental testing. From these data, we ascertain that with only three terms (specifically, a second-order polynomial equation), we can minimally describe the temporal degradation of MAPI across a broad range of temperatures. To explore the robustness of this approach, we apply the above approach to simulated data with varying degrees of Gaussian noise. PDE-FIND still succeeds at identifying the differential equation when up to 5% Gaussian noise is added to simulated data. However, at this noise level, the relative error in the identified coefficients is high (about 80%). At above 5% noise, PDE-FIND does not correctly identify the underlying equation. Thus, we illustrate the use of PDE-FIND with experimental data, and highlight the potential and limits of this approach to other fields.

In closing, we believe this work to be of broader interest to researchers who wish to infer scientific knowledge from trained machine-learning models applied to physical systems. 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. We show that Scientific ML has the potential to accelerate the understanding of materials degradation and the reliability optimization of perovskite materials. The conclusions of any given materials study may well be rendered more generalizable by identifying underlying equations governing the observations. 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.

Appendix A

Tables

1	9.158
U	-6.302×10^{-1}
U^2	1.426×10^{-2}
U^3	-1.475×10^{-4}
U^4	7.096×10^{-7}
U^5	-1.291×10^{-9}
$\sin(U)$	0
$\cos(U)$	0
$t^{0.5}$	0
$Ut^{0.5}$	0
t	0
Ut	0
t^2	0
Ut^2	0
t^3	0
Ut^3	0
T	1.294×10^{-3}
$\exp(-100/T)$	2.093

Table A.1: Coefficient Values Corresponding to Candidate Terms in the Differential Equation Identified by PDE-FIND With Full Experimental Data

Appendix B

Figures

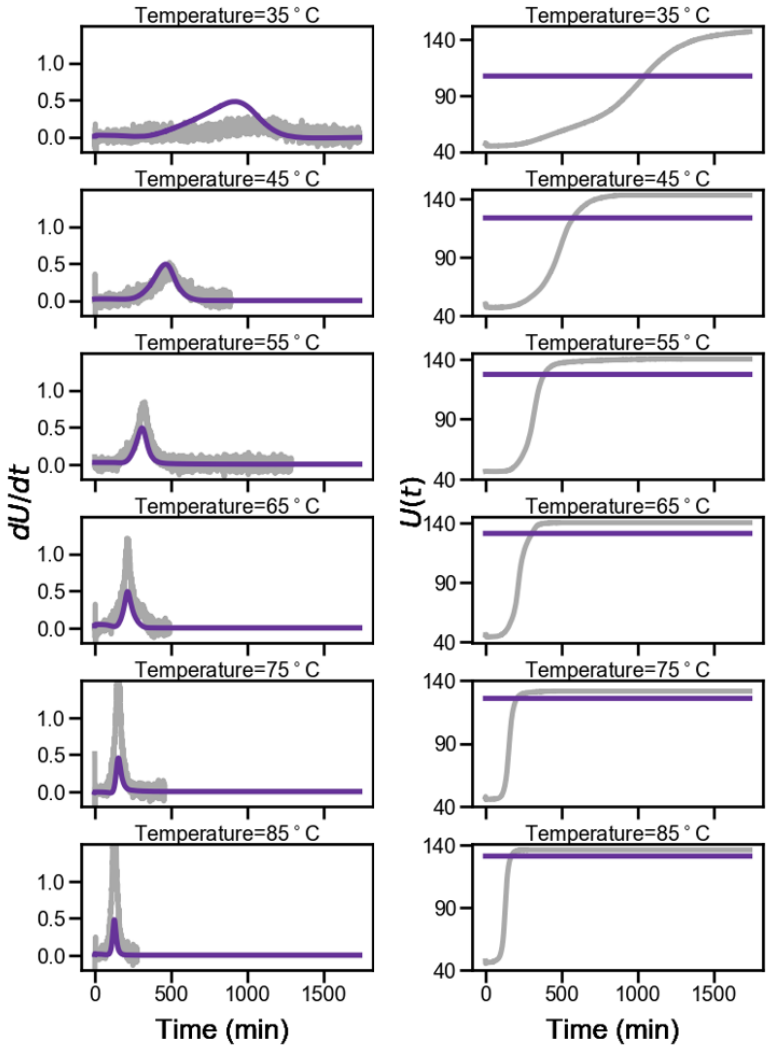

Figure B-1: PDE-FIND Results with Full Experimental Data

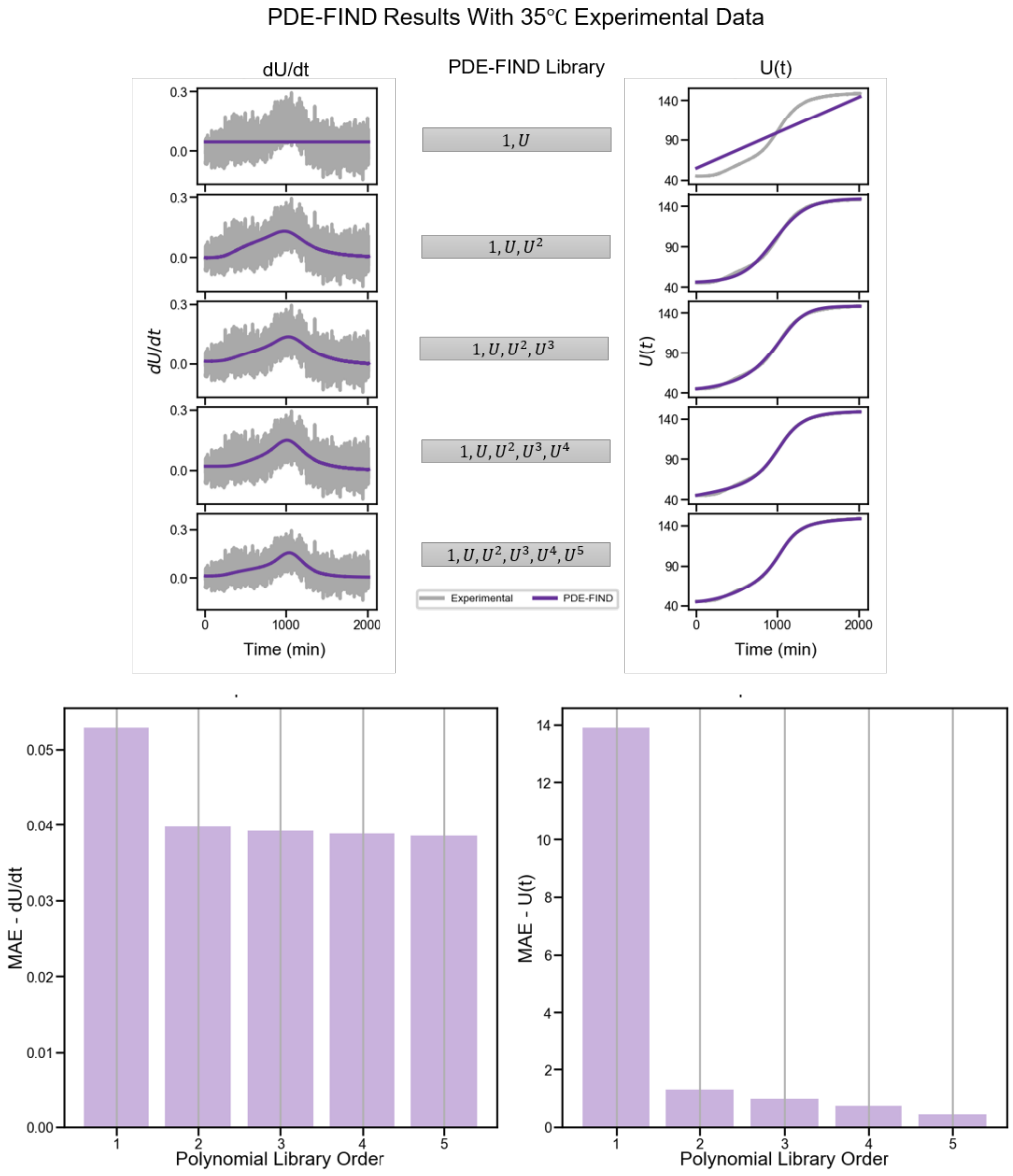

Figure B-2: PDE-FIND Results with Average Experimental Data at 35 °C

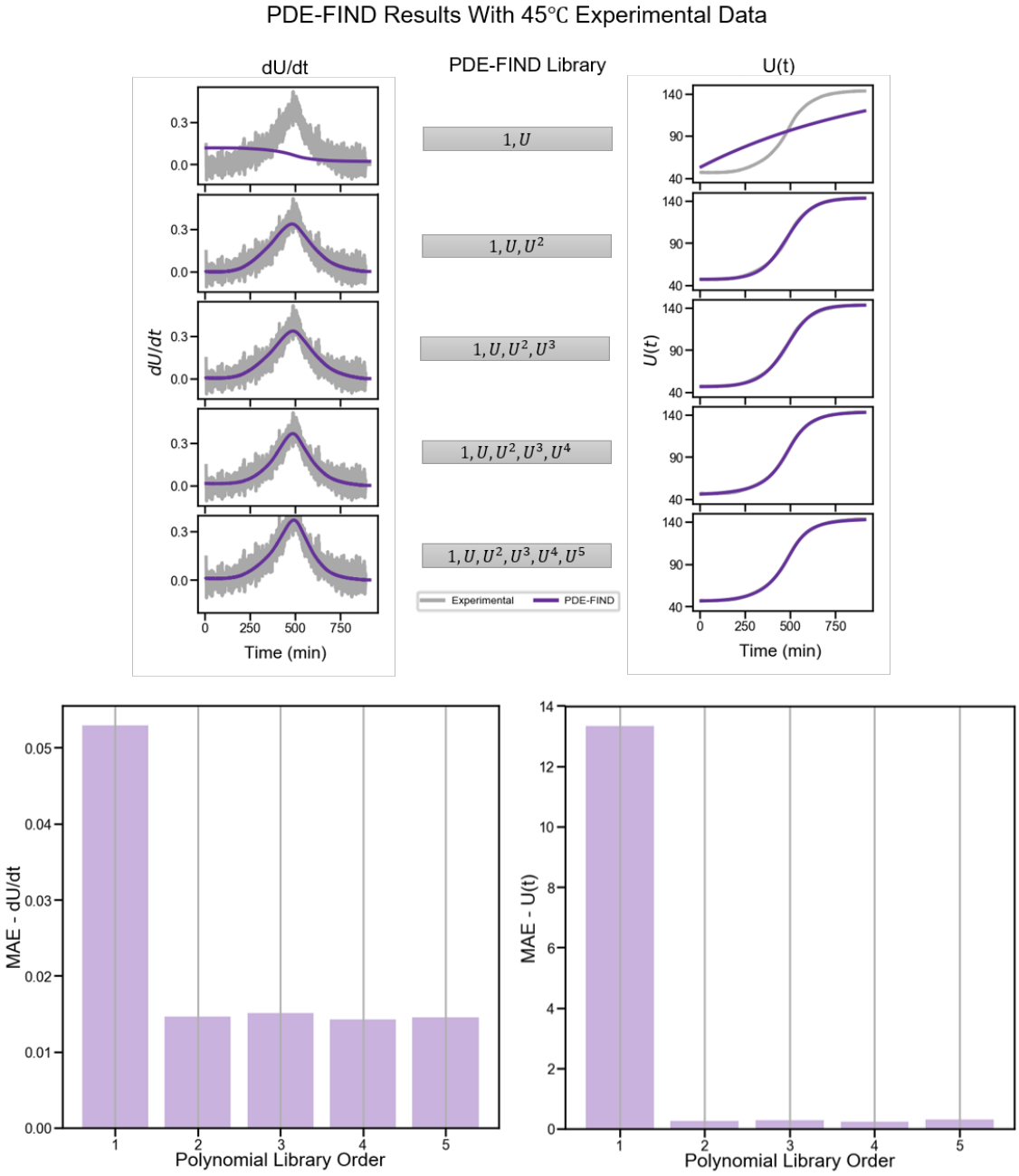

Figure B-3: PDE-FIND Results with Average Experimental Data at 45 °C

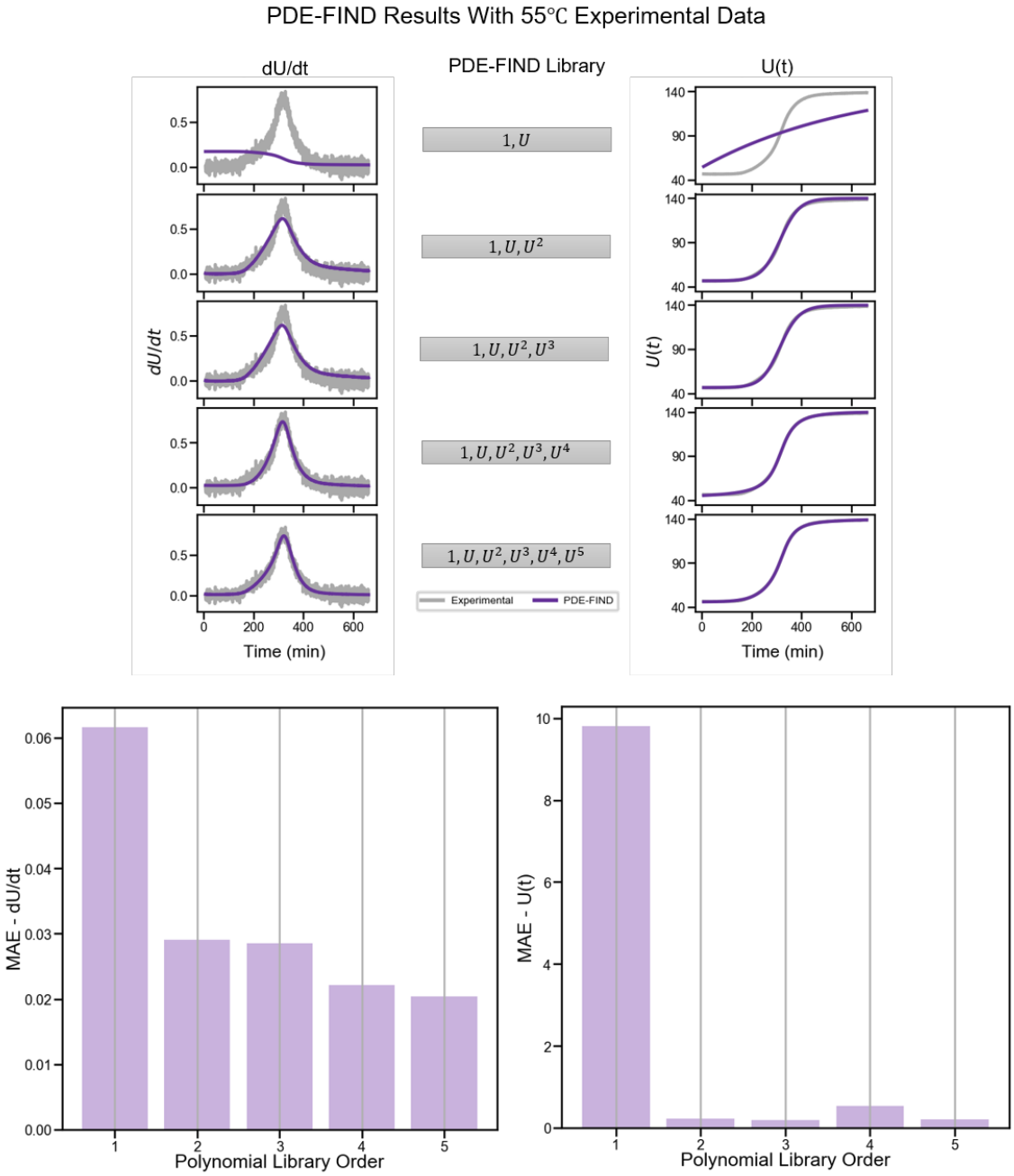

Figure B-4: PDE-FIND Results with Average Experimental Data at 55 °C

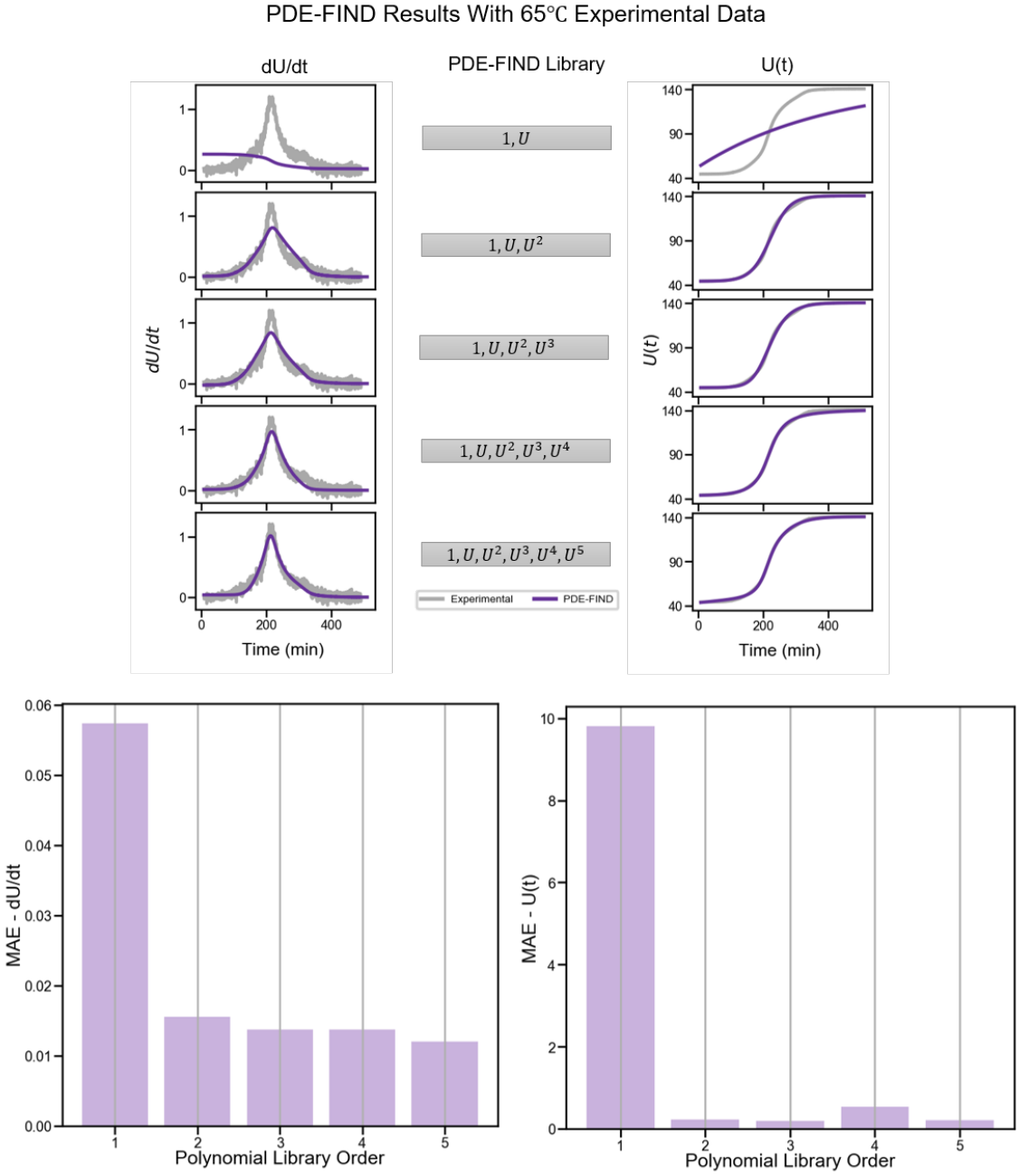

Figure B-5: PDE-FIND Results with Average Experimental Data at 65 °C

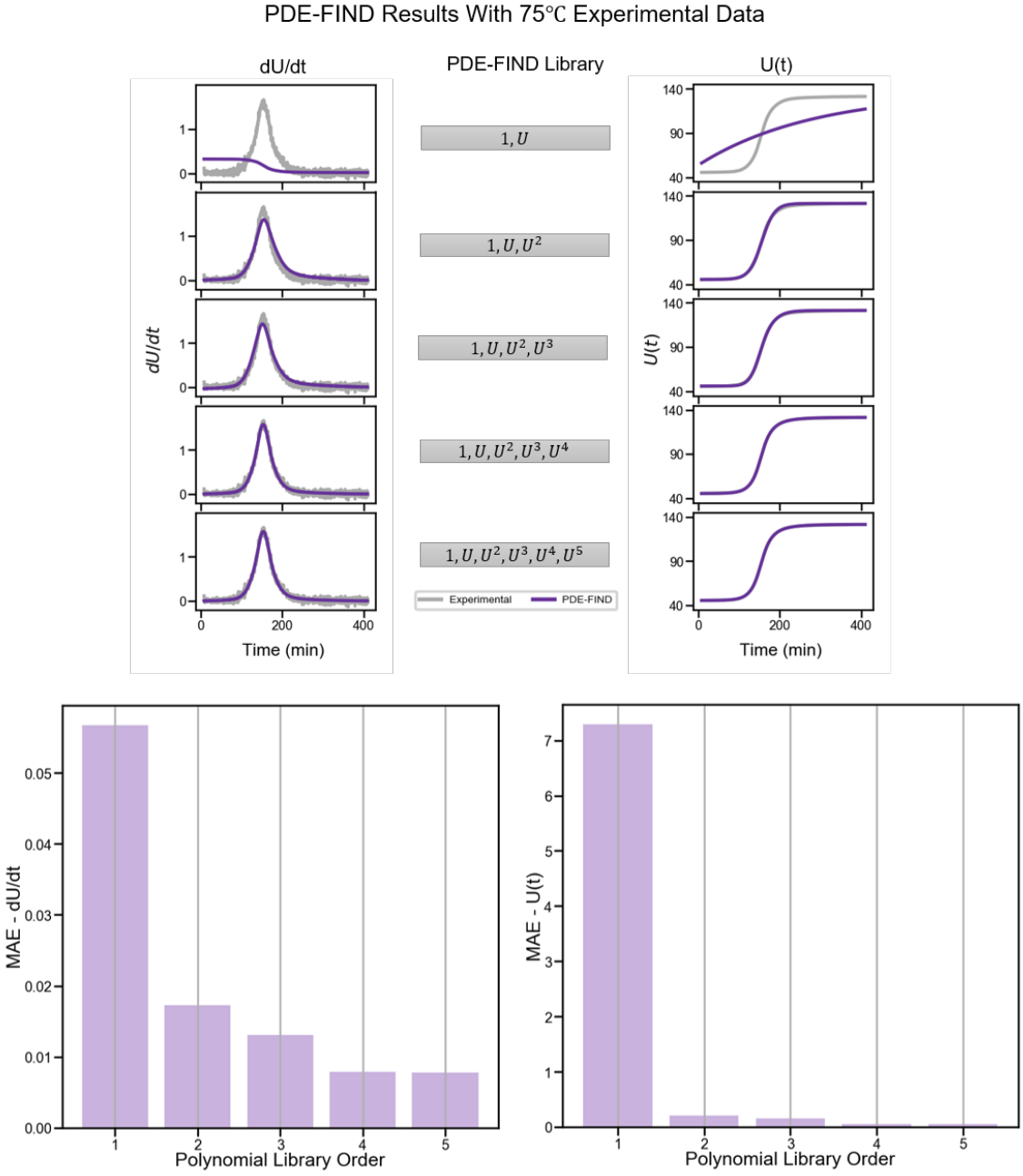

Figure B-6: PDE-FIND Results with Average Experimental Data at 75 °C

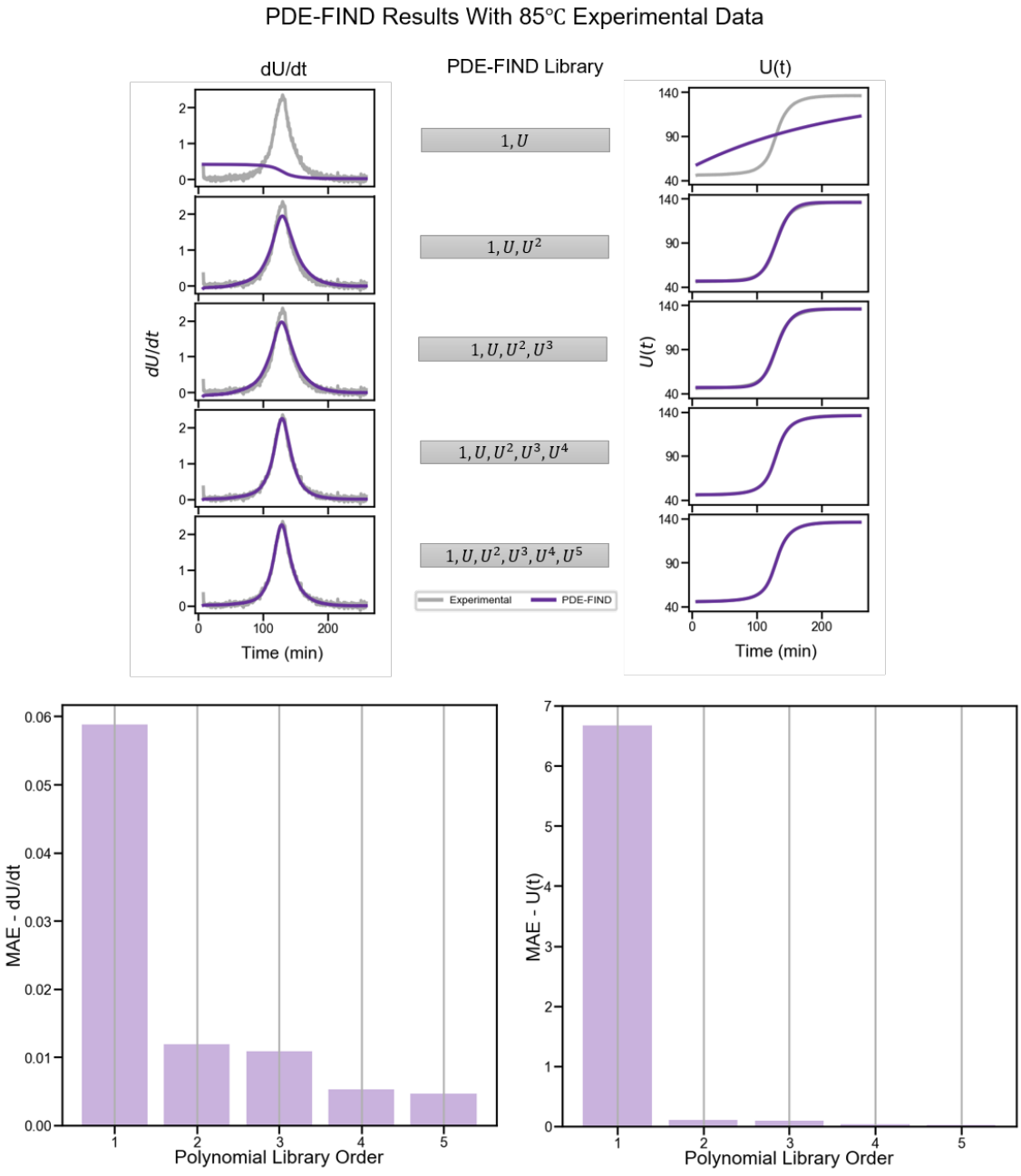

Figure B-7: PDE-FIND Results with Average Experimental Data at 85 °C

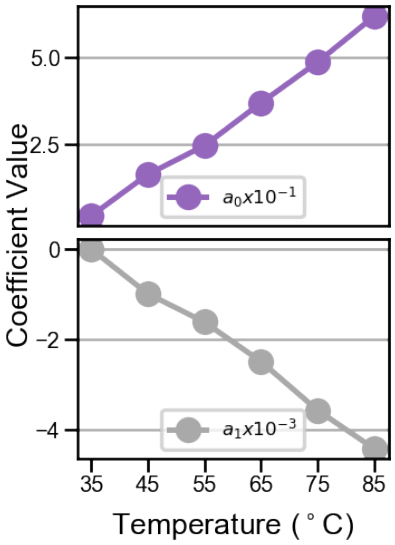

Figure B-8: Coefficient Values vs Temperature with 1st Order Library with Average Experimental Data

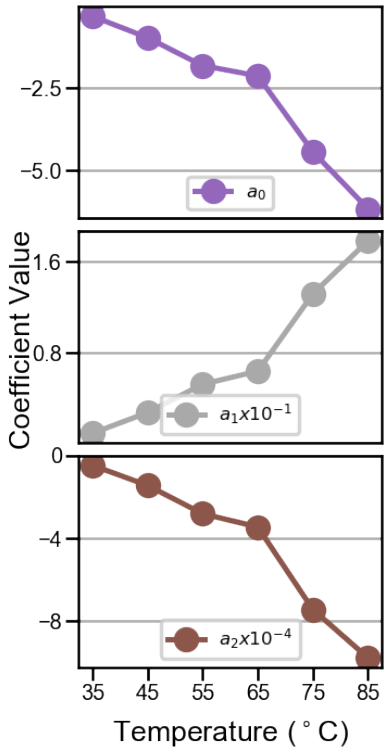

Figure B-9: Coefficient Values vs Temperature with 2nd Order Library with Average Experimental Data

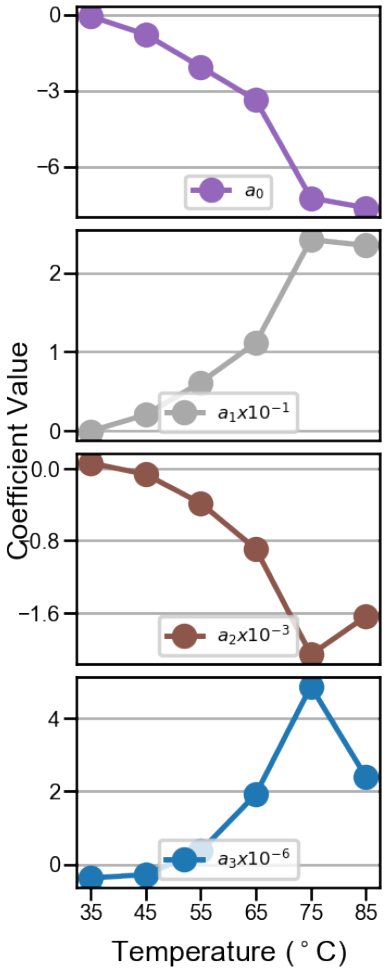

Figure B-10: Coefficient Values vs Temperature with 3rd Order Library with Average Experimental Data

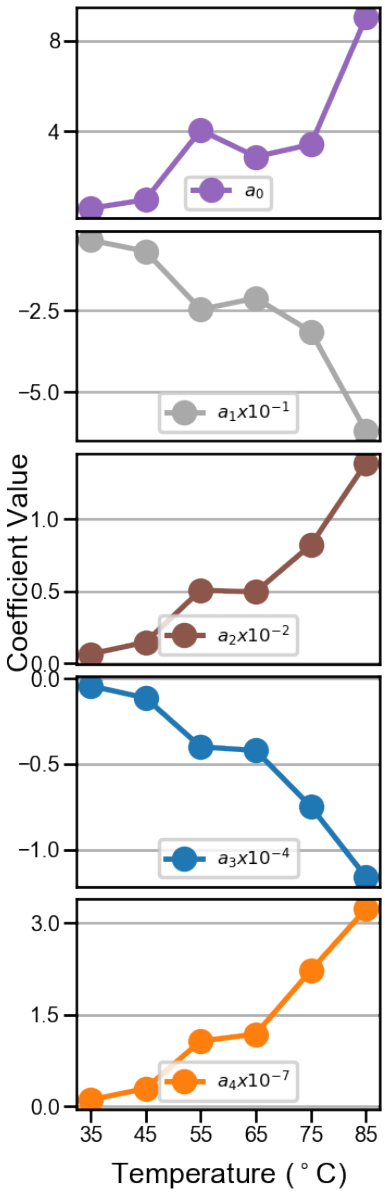

Figure B-11: Coefficient Values vs Temperature with 4th Order Library with Average Experimental Data

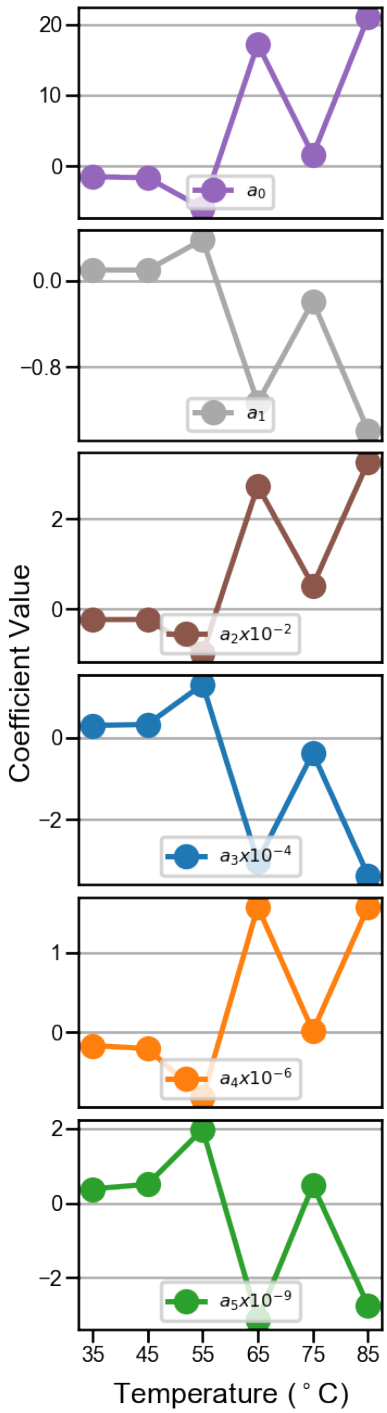

Figure B-12: Coefficient Values vs Temperature with 5th Order Library with Average Experimental Data

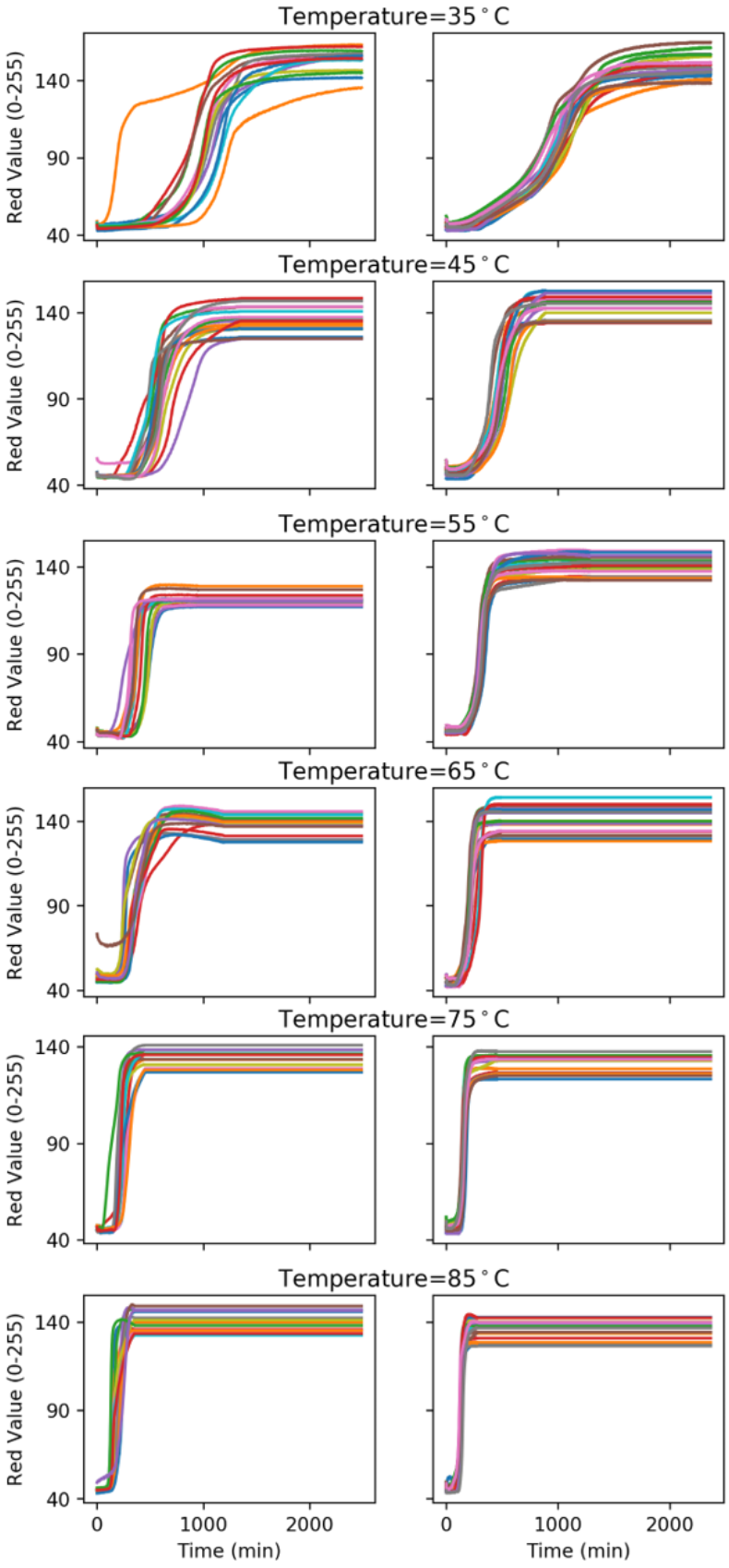

Figure B-13: Comparison Between the Red Value Plot of High and Low Variance Experimental Data

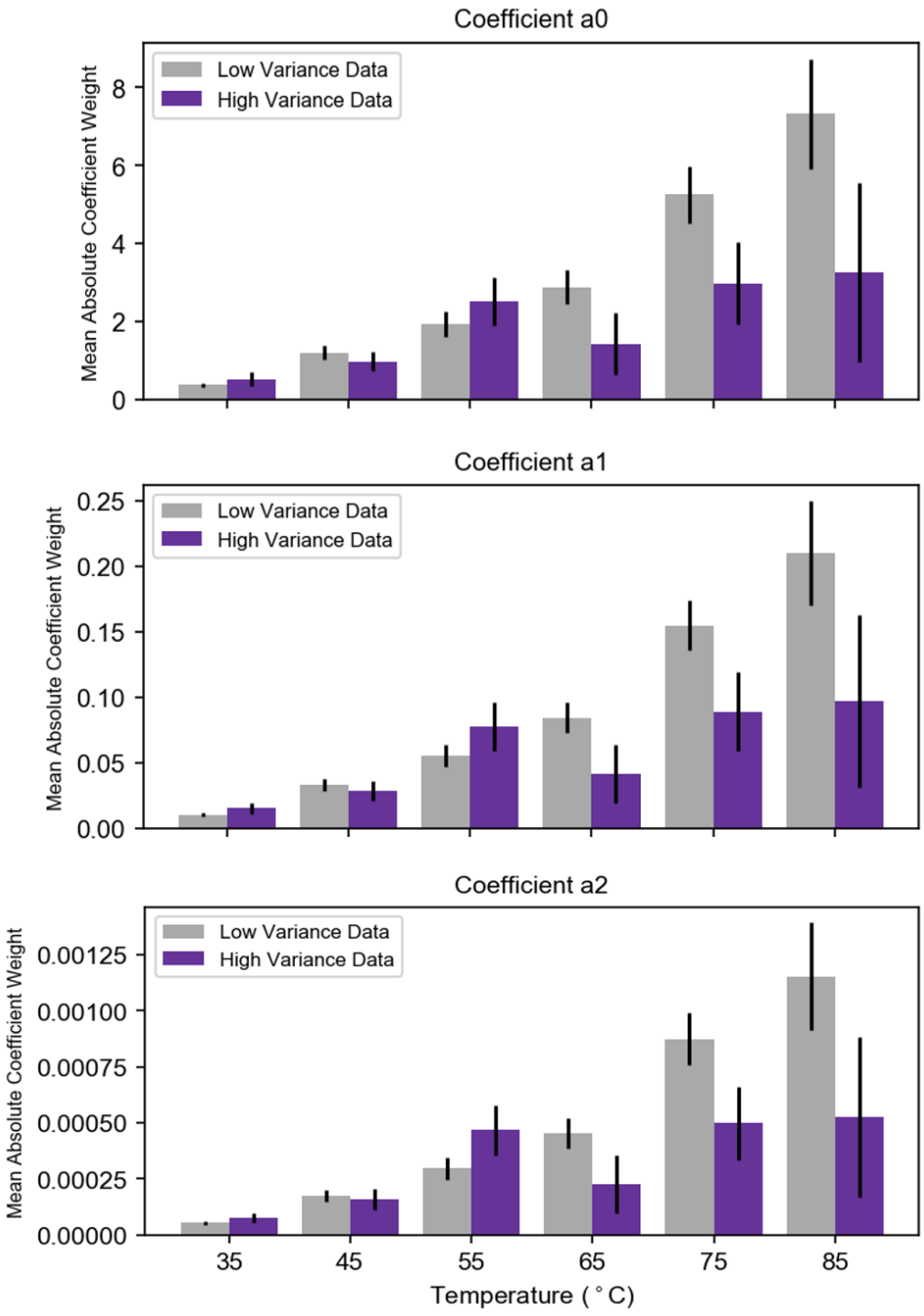

Figure B-14: Comparison of High and Low Variance Datasets: Coefficient Values Averaged Over Samples with 2nd Order Library

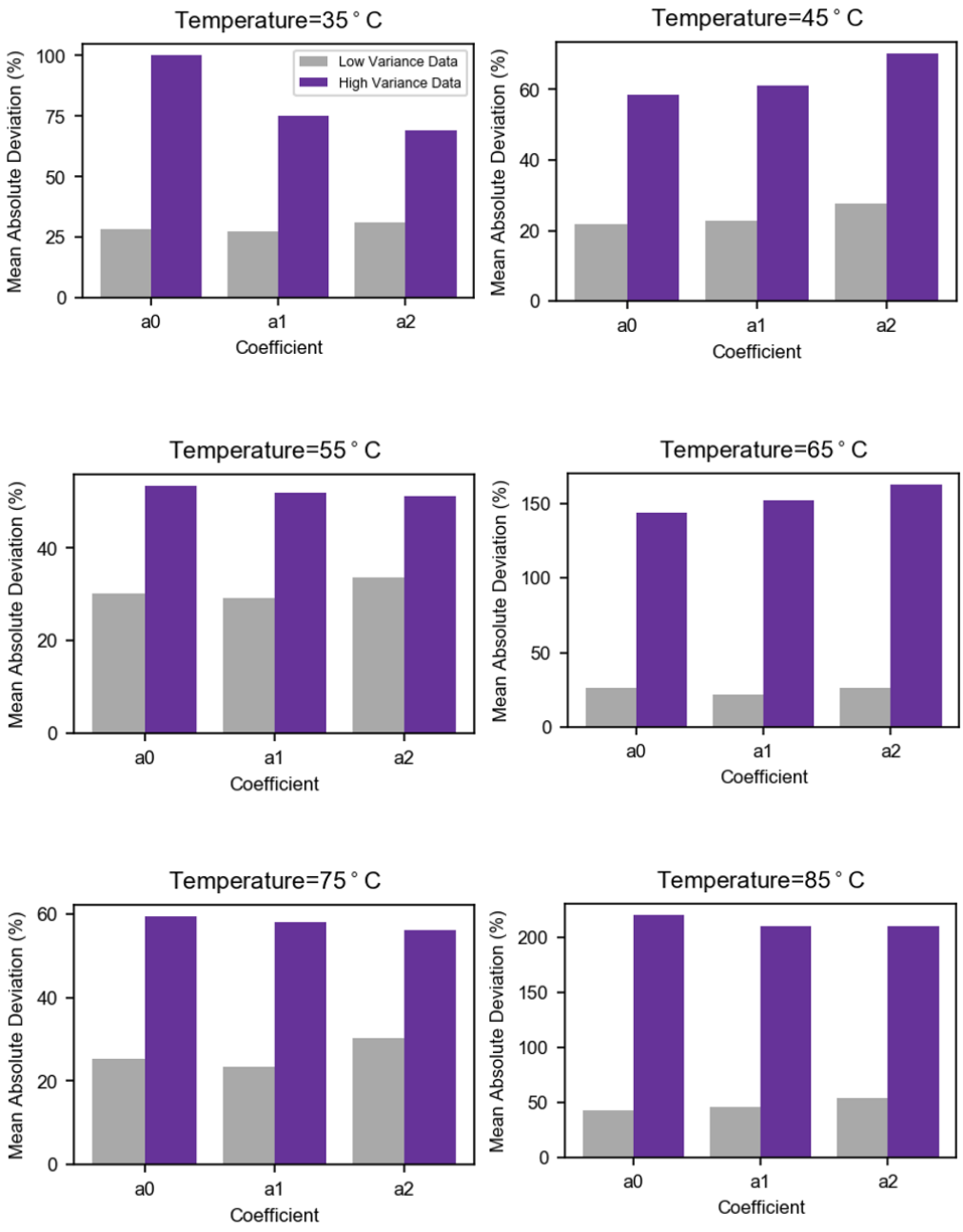

Figure B-15: Comparison of High and Low Variance Datasets: Mean Absolute Deviation (%) in Coefficient Values Samples with 2nd Order Library

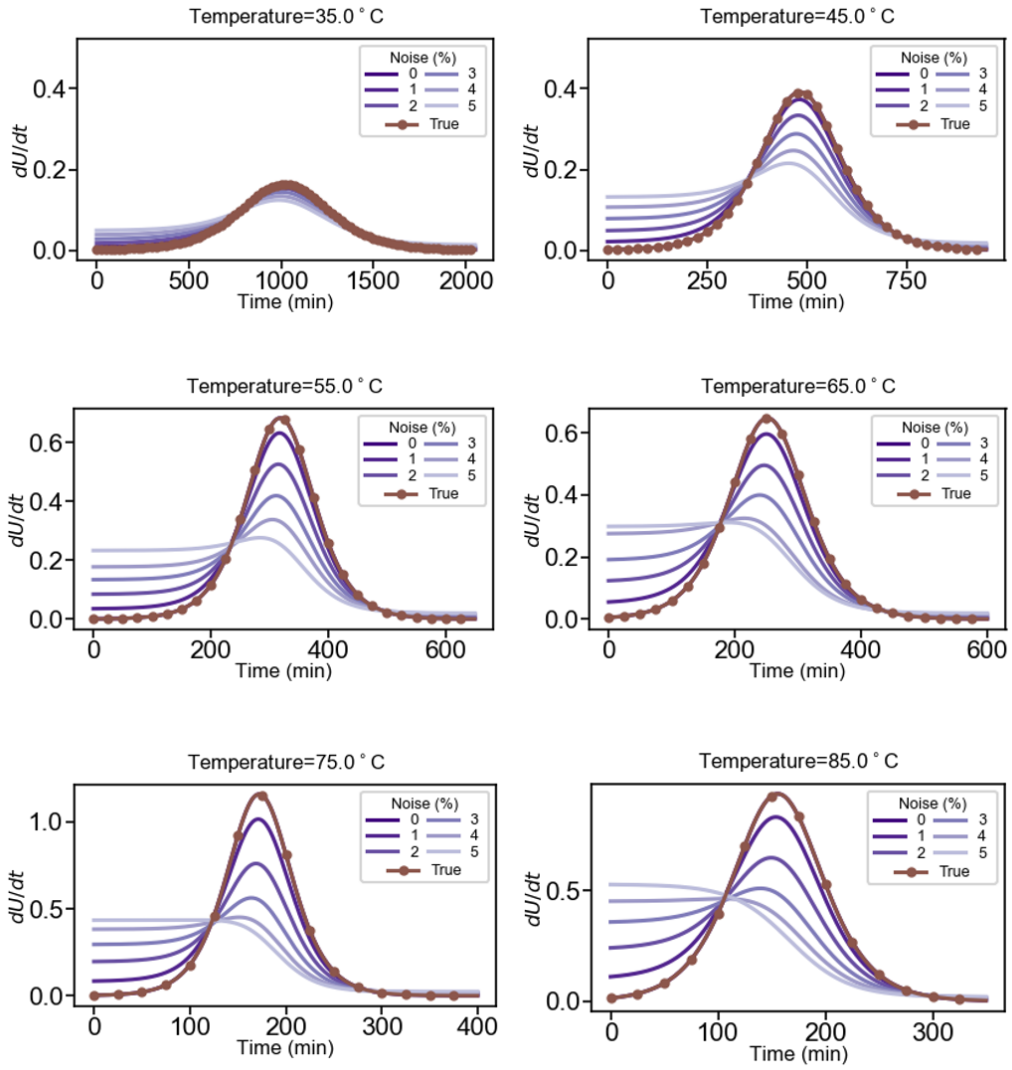

Figure B-16: PDE-FIND Results with Noisy Simulated Data - dU/dt estimated from the differential equation identified by PDE-FIND (with 2nd order library) vs ground truth

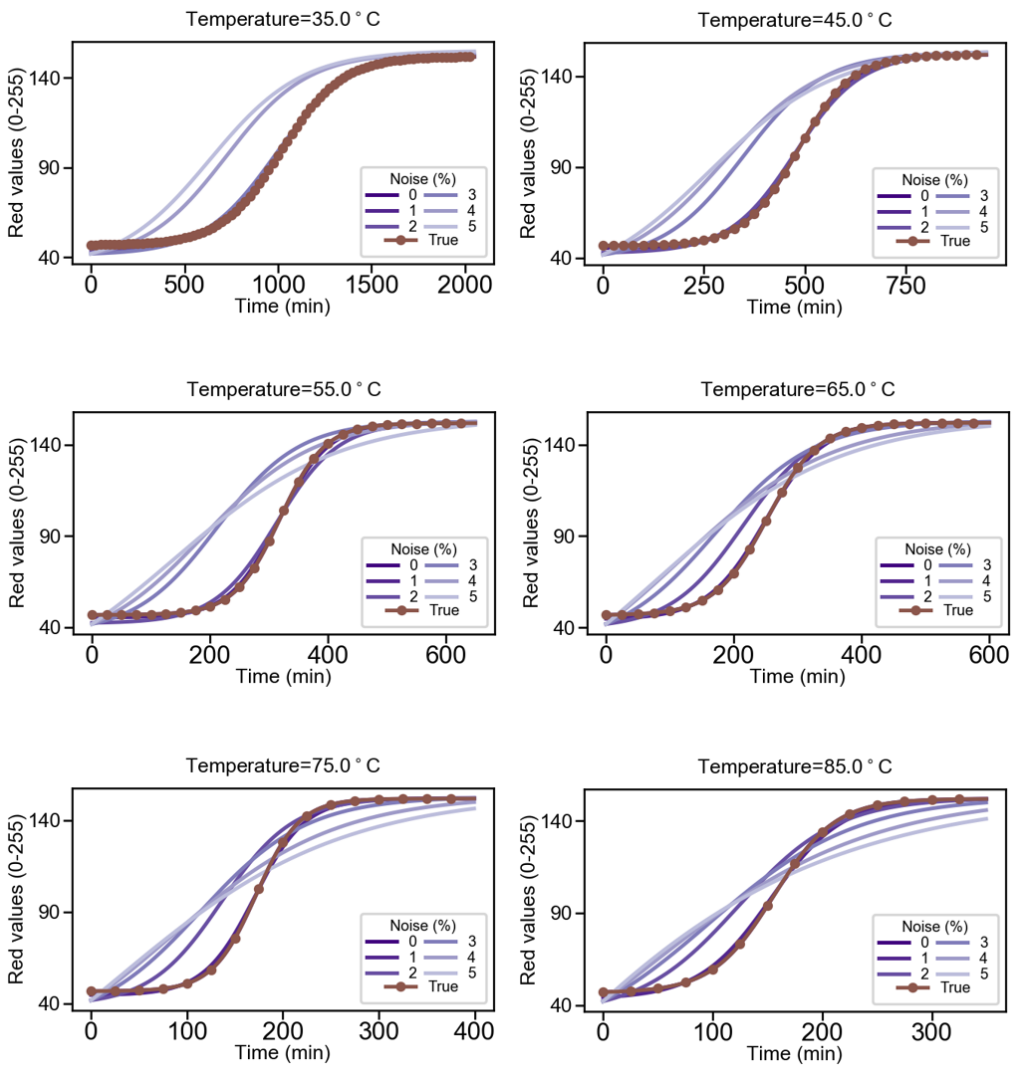

Figure B-17: PDE-FIND Results with Noisy Simulated Data - $U(t)$ estimated by integrating the differential equation identified by PDE-FIND (with 2nd order library) vs ground truth.

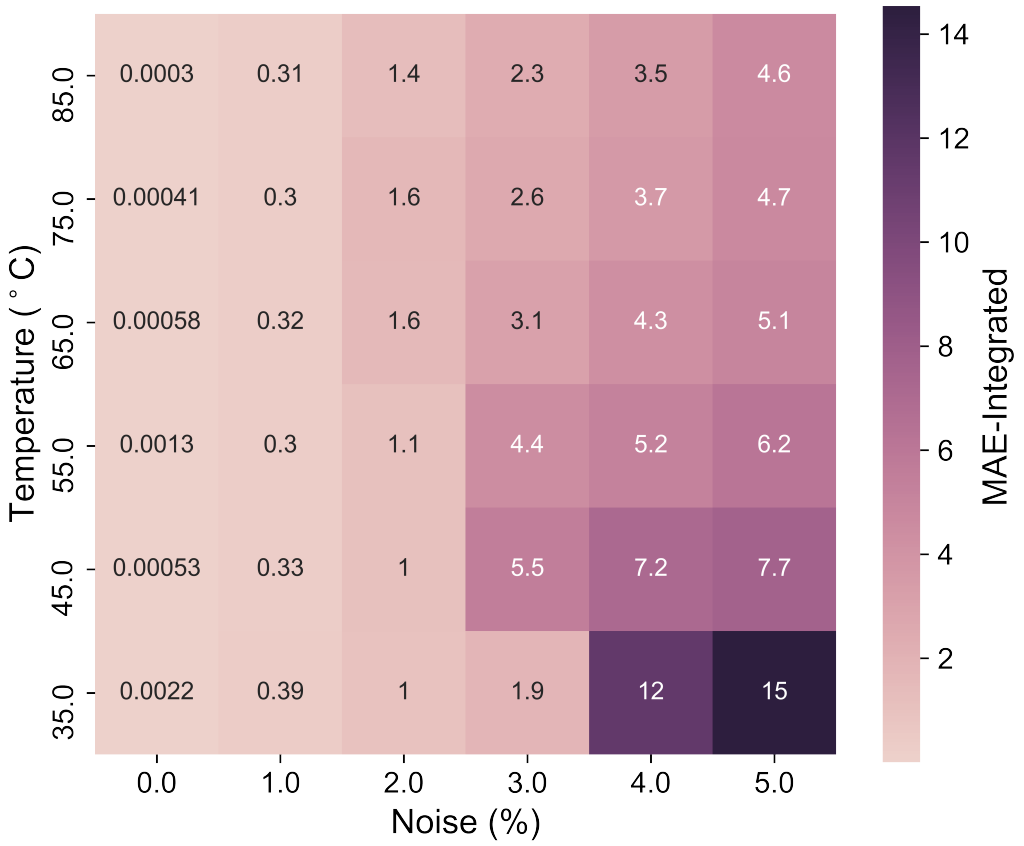

Figure B-18: Heatmap depicting the MAE between the ground truth data and the curves obtained by integrating the differential equation identified by PDE-FIND.

Appendix C

Code and Data

The code and data used to produce the results presented in this thesis will be available on Tonio Buonassisi Group's official GitHub account, `https://github.com/PV-Lab` in the `PDE-Extraction` repository: `https://github.com/PV-Lab/PDE-Extration`.

PDE-FIND is available on GitHub at `https://github.com/snagcliffs/PDE-FIND`.

Bibliography

[1] Ghada Abdelmageed, Cameron Mackeen, Kaitlin Hellier, Leila Jewell, Lydia Seymour, Mark Tingwald, Frank Bridges, Jin Z. Zhang, and Sue Carter. Effect of temperature on light induced degradation in methylammonium lead iodide perovskite thin films and solar cells. *Solar Energy Materials and Solar Cells*, 174:566–571, 1 2018.

[2] Joshua F Apgar, David K Witmer, Forest M White, and Bruce Tidor. Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*, 6(10):1890–1900, 2010.

[3] MI Asghar, J Zhang, H Wang, and PD Lund. Device stability of perovskite solar cells–a review. *Renewable and Sustainable Energy Reviews*, 77:131–146, 2017.

[4] Steven Atkinson, Waad Subber, Liping Wang, Genghis Khan, Philippe Hawi, and Roger Ghanem. Data-driven discovery of free-form governing differential equations. *arXiv*, 9 2019.

[5] Mélvin Avrami. Kinetics of phase change. i: General theory. *The Journal of Chemical Physics*, 7:1103–1112, 12 1939.

[6] João P. Bastos, Griet Uytterhoeven, Weiming Qiu, Ulrich W. Paetzold, David Cheyns, Supriya Surana, Javier Rivas, Manoj Jaysankar, Wenya Song, Tom Aernouts, Jef Poortmans, and Robert Gehlhaar. Model for the prediction of the lifetime and energy yield of methyl ammonium lead iodide perovskite solar cells at elevated temperatures. *ACS Applied Materials and Interfaces*, 11:16517–16526, 5 2019.

[7] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18, 2018.

[8] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[9] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[10] Caleb C Boyd, Rongrong Cheacharoen, Tomas Leijtens, and Michael D McGehee. Understanding degradation mechanisms and improving stability of perovskite photovoltaics. *Chemical reviews*, 119(5):3418–3451, 2018.

[11] Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech. 2020*, 52:477–508, 2019.

[12] Alan K. Burnham. Use and misuse of logistic equations for modeling chemical kinetics. *Journal of Thermal Analysis and Calorimetry*, 127:1107–1116, 1 2017.

[13] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[14] Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences of the United States of America*, 116:22445–22451, 11 2019.

[15] W Choi, H Huh, BA Tama, G Park, S Lee IEEE Access, and undefined 2019. A neural network model for material degradation detection and diagnosis using microscopic images. *ieeexplore.ieee.org*, 2019.

[16] Arthur Christopoulos and Michael J Lew. Beyond eyeballing: Fitting models to experimental data. *Critical Reviews in Biochemistry and Molecular Biology*, 35:359–391, 2000.

[17] Frank Cichos, Kristian Gustavsson, Bernhard Mehlig, and Giovanni Volpe. Machine learning for active matter. *Nature Machine Intelligence*, 2:94–103, 2 2020.

[18] Bert Conings, Jeroen Drijkoningen, Nicolas Gauquelin, Aslihan Babayigit, Jan D'Haen, Lien D'Olieslaeger, Anitha Ethirajan, Jo Verbeeck, Jean Manca, Edoardo Mosconi, Filippo De Angelis, and Hans Gerd Boyen. Intrinsic thermal instability of methylammonium lead trihalide perovskite. *Advanced Energy Materials*, 5, 8 2015.

[19] Felix Creutzig, Peter Agoston, Jan Christoph Goldschmidt, Gunnar Luderer, Gregory Nemet, and Robert C Pietzcker. The underestimated potential of solar energy to mitigate climate change. *Nature Energy*, 2(9):1–9, 2017.

[20] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[21] Brian M de Silva, David M Higdon, Steven L Brunton, and J Nathan Kutz. Discovery of physics from data: universal laws and discrepancies. *Frontiers in artificial intelligence*, 3:25, 2020.

[22] G. Divitini, S. Cacovich, F. Matteocci, L. Cinà, A. Di Carlo, and C. Ducati. In situ observation of heat-induced degradation of perovskite solar cells. *Nature Energy*, 1:1–6, 1 2016.

[23] Christie L.C. Ellis, Hamza Javaid, Emily C. Smith, and D. Venkataraman. Hybrid perovskites with larger organic cations reveal autocatalytic degradation kinetics and increased stability under light. *Inorganic Chemistry*, 59:12176–12186, 9 2020.

[24] Elahe Entekhabi, Masoumeh Haghbin Nazarpak, Mehdi Sedighi, and Arghavan Kazemzadeh. Predicting degradation rate of genipin cross-linked gelatin scaffolds with machine learning. *Materials Science and Engineering C*, 107:110362, 2 2020.

[25] Zheng Fan, Hai Xiao, Yiliu Wang, Zipeng Zhao, Zhaoyang Lin, Hung Chieh Cheng, Sung Joon Lee, Gongming Wang, Ziying Feng, William A. Goddard, Yu Huang, and Xiangfeng Duan. Layer-by-layer degradation of methylammonium lead tri-iodide perovskite microplates. *Joule*, 1:548–562, 11 2017.

[26] M. Fanfoni and M. Tomellini. The johnson-mehl-avrami-kolmogorov model: A brief review. *Nuovo Cimento della Societa Italiana di Fisica D - Condensed Matter, Atomic, Molecular and Chemical Physics, Biophysics*, 20:1171–1182, 1998.

[27] Fan Fu, Stefano Pisoni, Quentin Jeangros, Jordi Sastre-Pellicer, Maciej Kawecki, Adriana Paracchino, Thierry Moser, Jérémie Werner, Christian Andres, Léo Duchêne, Peter Fiala, Michael Rawlence, Sylvain Nicolay, Christophe Ballif, Ayodhya N. Tiwari, and Stephan Buecheler. I2 vapor-induced degradation of formamidinium lead iodide based perovskite solar cells under heat-light soaking conditions. *Energy and Environmental Science*, 12:3074–3088, 10 2019.

[28] Yu Han, Steffen Meyer, Yasmina Dkhissi, Karl Weber, Jennifer M. Pringle, Udo Bach, Leone Spiccia, and Yi Bing Cheng. Degradation observations of encapsulated planar ch3nh3pbi3 perovskite solar cells at high temperatures and humidity. *Journal of Materials Chemistry A*, 3:8139–8147, 4 2015.

[29] Noor Titan Putri Hartono, Janak Thapa, Armi Tiihonen, Felipe Oviedo, Clio Batali, Jason J Yoo, Zhe Liu, Ruipeng Li, David Fuertes Marrón, Moungi G Bawendi, et al. How machine learning can help select capping layers to suppress perovskite degradation. *Nature communications*, 11(1):1–9, 2020.

[30] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Evan Heit. Properties of inductive reasoning. *Psychonomic Bulletin and Review*, 7:569–592, 2000.

[34] Emilio J. Juarez-Perez, Luis K. Ono, Maki Maeda, Yan Jiang, Zafer Hawash, and Yabing Qi. Photodecomposition and thermal decomposition in methylammonium halide lead perovskites and inferred design principles to increase photovoltaic device stability. *Journal of Materials Chemistry A*, 6:9604–9612, 5 2018.

[35] Mark V. Khenkin, Eugene A. Katz, Antonio Abate, Giorgio Bardizza, Joseph J. Berry, Christoph Brabec, Francesca Brunetti, Vladimir Bulović, Quinn Burlingame, Aldo Di Carlo, Rongrong Cheacharoen, Yi Bing Cheng, Alexander Colsmann, Stephane Cros, Konrad Domanski, Michał Dusza, Christopher J. Fell, Stephen R. Forrest, Yulia Galagan, Diego Di Girolamo, Michael Grätzel, Anders Hagfeldt, Elizabeth von Hauff, Harald Hoppe, Jeff Kettle, Hans Köbler, Marina S. Leite, Shengzhong (Frank) Liu, Yueh Lin Loo, Joseph M. Luther, Chang Qi Ma, Morten Madsen, Matthieu Manceau, Muriel Matheron, Michael McGehee, Rico Meitzner, Mohammad Khaja Nazeeruddin, Ana Flavia Nogueira, Çağla Odabaşı, Anna Osherov, Nam Gyu Park, Matthew O. Reese, Francesca De Rossi, Michael Saliba, Ulrich S. Schubert, Henry J. Snaith, Samuel D. Stranks, Wolfgang Tress, Pavel A. Troshin, Vida Turkovic, Sjoerd Veenstra, Iris Visoly-Fisher, Aron Walsh, Trystan Watson, Haibing Xie, Ramazan Yıldırım, Shaik Mohammed Zakeeruddin, Kai Zhu, and Monica Lira-Cantu. Consensus statement for stability assessment and reporting for perovskite photovoltaics based on isos procedures. *Nature Energy*, 5:35–49, 1 2020.

[36] Nam Koo Kim, Young Hwan Min, Seokhwan Noh, Eunkyung Cho, Gitaeg Jeong, Minho Joo, Seh Won Ahn, Jeong Soo Lee, Seongtak Kim, Kyuwook Ihm, Hyungju Ahn, Yoonmook Kang, Hae Seok Lee, and Donghwan Kim. Investigation of thermally induced degradation in ch3nh3pbi3 perovskite solar cells using in-situ synchrotron radiation analysis. *Scientific Reports*, 7:1–9, 12 2017.

[37] Akihiro Kojima, Kenjiro Teshima, Yasuo Shirai, and Tsutomu Miyasaka. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *Journal of the American Chemical Society*, 131(17):6050–6051, 2009.

[38] G. Rajendra Kumar, A. Dennyson Savariraj, S. N. Karthick, S. Selvam, B. Balamuralitharan, Hee Je Kim, K. K. Viswanathan, M. Vijaykumar, and Kandasamy Prabakar. Phase transition kinetics and surface binding states of methylammonium lead iodide perovskite. *Physical Chemistry Chemical Physics*, 18:7284–7292, 3 2016.

[39] G Rajendra Kumar, A Dennyson Savariraj, SN Karthick, Samayanan Selvam, B Balamuralitharan, Hee-Je Kim, Kodakkal K Viswanathan, M Vijaykumar, and

Kandasamy Prabakar. Phase transition kinetics and surface binding states of methylammonium lead iodide perovskite. *Physical Chemistry Chemical Physics*, 18(10):7284–7292, 2016.

[40] Sang Won Lee, Seongtak Kim, Soohyun Bae, Kyungjin Cho, Taewon Chung, Laura E. Mundt, Seunghun Lee, Sungeun Park, Hyomin Park, Martin C. Schubert, Stefan W. Glunz, Yohan Ko, Yongseok Jun, Yoonmook Kang, Hae Seok Lee, and Donghwan Kim. Uv degradation and recovery of perovskite solar cells. *Scientific Reports*, 6:1–10, 12 2016.

[41] Zichao Long, Yiping Lu, and Bin Dong. Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399, 11 2018.

[42] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. Pde-net: Learning pdes from data. *35th International Conference on Machine Learning, ICML 2018*, 7:5067–5078, 10 2017.

[43] Hendrik Antoon Lorentz, Albert Einstein, Hermann Minkowski, Hermann Weyl, and Arnold Sommerfeld. *The principle of relativity: a collection of original memoirs on the special and general theory of relativity*. Courier Corporation, 1952.

[44] Jelena Macan, Ivan Brnardić, Sebastijan Orlić, Hrvoje Ivanković, and Marica Ivanković. Thermal degradation of epoxy - silica organic - inorganic hybrid materials. *Polymer Degradation and Stability*, 91:122–127, 1 2006.

[45] Niall M Mangan, J Nathan Kutz, Steven L Brunton, and Joshua L Proctor. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2204):20170009, 2017.

[46] Laurence W. McKeen. *The Effect of Long Term Thermal Exposure on Plastics and Elastomers*. Elsevier Inc., 2013.

[47] W Nash, T Drummond, N Birbilis npj Materials Degradation, and undefined 2018. A review of deep learning in the study of materials degradation. *nature.com*, 2018.

[48] Isaac Newton. Philosophiæ naturalis principia mathematica (mathematical principles of natural philosophy). *London (1687)*, 1687, 1987.

[49] Wanyi Nie, Jean Christophe Blancon, Amanda J. Neukirch, Kannatassen Appavoo, Hsinhan Tsai, Manish Chhowalla, Muhammad A. Alam, Matthew Y. Sfeir, Claudine Katan, Jacky Even, Sergei Tretiak, Jared J. Crochet, Gautam Gupta, and Aditya D. Mohite. Light-activated photocurrent degradation and self-healing in perovskite solar cells. *Nature Communications*, 7:1–9, 5 2016.

[50] F. Otto, A. Dlouhý, K. G. Pradeep, M. Kuběnová, D. Raabe, G. Eggeler, and E. P. George. Decomposition of the single-phase high-entropy alloy crmnfeconi after prolonged anneals at intermediate temperatures. *Acta Materialia*, 112:40–52, 6 2016.

[51] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv*, 1 2020.

[52] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2 2019.

[53] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367:1026–1030, 2 2020.

[54] NREL Photovoltaic Research. Best research-cell efficiency chart. `https://www.nrel.gov/pv/cell-efficiency.html`. accessed: 05.21.2021.

[55] Seth Roberts and Harold Pashler. How persuasive is a good fit? a comment on theory testing. *Psychological Review*, 107:358–367, 2000.

[56] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[57] Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3:e1602614, 4 2017.

[58] John L Russell. Kepler's laws of planetary motion: 1609-1666. *The British journal for the history of science*, pages 1–24, 1964.

[59] Martin Schmelzer, Richard P. Dwight, and Paola Cinnella. Discovery of algebraic reynolds-stress models using sparse symbolic regression. *Flow, Turbulence and Combustion*, 104:579–603, 3 2020.

[60] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324:81–85, 4 2009.

[61] Jonas A Schwenzer, Tim Hellmann, Bahram Abdollahi Nejand, Hang Hu, Tobias Abzieher, Fabian Schackmar, Ihteaz M Hossain, Paul Fassl, Thomas Mayer, Wolfram Jaegermann, Uli Lemmer, and Ulrich W Paetzold. Thermal stability and cation composition of hybrid organic–inorganic perovskites. *ACS Applied Materials & Interfaces*, page acsami.1c01547, 2021.

[62] KA Severson, PM Attia, N Jin, N Perkins, B Jiang Nature Energy, and undefined 2019. Data-driven prediction of battery cycle life before capacity degradation. *nature.com*, 2019.

[63] C. B. Simmons, Austin J. Akey, Jacob J. Krich, Joseph T. Sullivan, Daniel Recht, Michael J. Aziz, and Tonio Buonassisi. Deactivation of metastable single-crystal silicon hyperdoped with sulfur. *Journal of Applied Physics*, 114:243514, 12 2013.

[64] Emanuele Smecca, Youhei Numata, Ioannis Deretzis, Giovanna Pellegrino, Simona Boninelli, Tsutomu Miyasaka, Antonino La Magna, and Alessandra Alberti. Stability of solution-processed mapbi3 and fapbi3 layers. *Physical Chemistry Chemical Physics*, 18:13413–13422, 5 2016.

[65] EA Starke et al. *Accelerated Aging of Materials and Structures*. National Academies Press, 1 1996.

[66] Shijing Sun, Armi Tiihonen, Felipe Oviedo, Zhe Liu, Janak Thapa, Yicheng Zhao, Noor Titan P. Hartono, Anuj Goyal, Thomas Heumueller, Clio Batali, Alex Encinas, Jason J. Yoo, Ruipeng Li, Zekun Ren, I. Marius Peters, Christoph J. Brabec, Moungi G. Bawendi, Vladan Stevanovic, John Fisher, and Tonio Buonassisi. A data fusion approach to optimize compositional stability of halide perovskites. *Matter*, 4:1305–1322, 4 2021.

[67] Catherine D.T. Tran, Yi Liu, Emmanuel S. Thibau, Adrian Llanos, and Zheng Hong Lu. Stability of organometal perovskites with organic overlayers. *AIP Advances*, 5:087185, 8 2015.

[68] A. Tsoularis and J. Wallace. Analysis of logistic growth models. *Mathematical Biosciences*, 179:21–55, 7 2002.

[69] P. S. Whitfield, N. Herron, W. E. Guise, K. Page, Y. Q. Cheng, I. Milas, and M. K. Crawford. Structures, phase transitions and tricritical behavior of the hybrid perovskite methyl ammonium lead iodide. *Scientific Reports*, 6:1–16, 10 2016.

[70] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, 2020.

[71] Jinli Yang, Braden D. Siempelkamp, Dianyi Liu, and Timothy L. Kelly. Investigation of ch3nh3pbi3degradation rates and mechanisms in controlled humidity environments using in situ techniques. *ACS Nano*, 9:1955–1963, 2 2015.

[72] Minglang Yin, Xiaoning Zheng, Jay D. Humphrey, and George Em Karniadakis. Non-invasive inference of thrombus material properties with physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 375:113603, 3 2021.

[73] Laure Zanna and Thomas Bolton. Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47:e2020GL088376, 9 2020.

[74] Sheng Zhang and Guang Lin. Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217):20180305, 2018.

[75] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Advances in Neural Information Processing Systems*, 21:1921–1928, 2008.