Predicting Materials Parameters in Colloidal Quantum Dot Photovoltaic Devices Using Machine Learning Models Trained On Experimental Data

Hoon Jeong Lee, Ariana B. Hofelmann, Yida Lin, and Susanna M. Thon

Department of Electrical and Computer Engineering, Johns Hopkins University, 3400 N. Charles Street, Baltimore, Maryland, 21218, USA

2022 IEEE 49th Photovoltaics Specialists Conference (PVSC) | 978-1-7281-6117-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/PVSC48317.2022.9938867

***Abstract*—Numerous characterization techniques have been developed over the last century, which have advanced progress on the development of a variety of photovoltaic technologies. However, this multitude of techniques leads to increasing experimental costs and complexity. It would be useful to have an approach that does not require the time commitment or operation costs to directly learn and implement every new measurement technique. Herein, we explore several machine learning (ML) models that output complex materials parameters, such as electronic trap state density, solely using illuminated current-voltage curves. This greatly reduces both the complexity and cost of the characterization process. Current-voltage curves were chosen as the only input to our models because this type of measurement is relatively simple to perform and most photovoltaic research labs already collect this information on all devices. We compare several different ML network architectures, all of which are trained on experimental data from PbS colloidal quantum dot thin film solar cells. We predict values for underlying materials parameters and compare them to experimentally measured results.**

***Index Terms*—thin film, lead sulfide, machine learning, colloidal quantum dots**

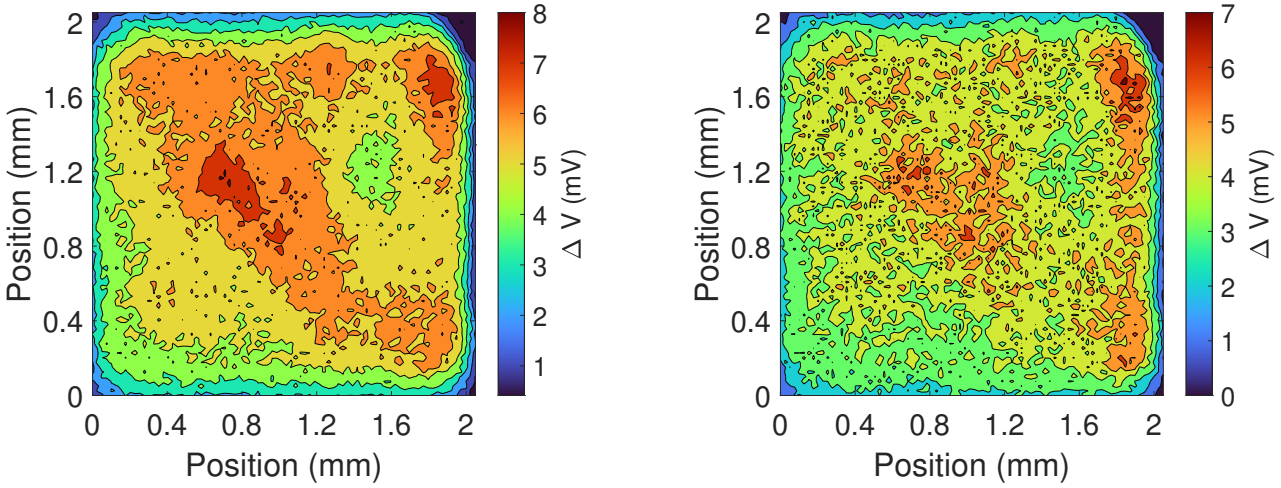
I. Introduction

The field of photovoltaics has taken advantage of numerous experimental techniques to measure the critical underlying materials parameters which determine solar cell device performance. Techniques such as charge extraction by linearly increasing voltage (CELIV), time of flight (TOF) measurements, photocurrent transient spectroscopy, and space-charge-limited current (SCLC) measurements are used to measure charge mobility while other techniques such as the transient photovoltage method and deep level transient spectroscopy (DLTS) can be used to determine mid-gap trap state densities [1]. Determining these parameters is essential in the development cycle of solar cells as they allow researchers to compare different devices and fabrication methods, as well as identify limits to and improve device performance. However, these techniques often require specialized device architectures and an experienced researcher to determine the best underlying analytical or numerical model to fit to the data. This can not only be time consuming, but complex as well. In addition, large upfront costs for equipment and apparatuses make some measurement techniques inaccessible to smaller labs. This sometimes leads to labs needing to choose between measuring one materials parameter versus another. One possible alternative to conventional methods is to leverage machine learning (ML) to obtain these materials parameters by taking advantage of existing datasets and correlating them with new, simpler measurements.

Herein, we test several machine learning models to aid in the solar cell development process. The key difference between our models and others found in the literature is that our algorithms are trained on experimental data rather than simulation data. The latter is usually preferred due to the high cost and complexity of fabricating numerous devices; however, data extracted from simulations is not as reliable, accurate, or comprehensive as data collected from real devices. This is due to errors such as the simulation of thermodynamically unstable, physically impossible, or idealized structures [2]. We leveraged our past work on development of a multi-modal optoelectronic scanning instrument for solution-processed solar cells [3] to generate massive training data sets on colloidal quantum dot photovoltaic devices that can be used to bolster and diversify existing experimental and computationally generated datasets.

In general, our models utilize supervised machine learning methods to predict materials parameters. The goal is to train an artificial neural network to predict an output vector $\mathbf{t} \equiv [t_1, t_2, \cdots, t_N]$ from a particular input vector $\mathbf{x} \equiv [x_1, x_2, \cdots, x_N]$. In the general case, these vectors could be of different dimensions. We train the model by providing numerous examples of input-output pairs, so that it can infer the underlying relationship between the two variables via back-propagation and the gradient descent method. Since this method uses statistical correlations instead of physical laws [4], it eliminates the need for complex user analysis and the need to encode data presumptions. There are several feed-forward network architectures that can be used to solve these classes of regression problems: the multilayer perceptron, autoencoders, and the convolutional variants of these networks. With new data being generated at an exponential rate, ML offers a time- and memory-efficient way of analyzing large

This research was funded by the National Science Foundation (DMR-1807342).



(a) Measured photovoltage map

(b) ML predicted photovoltage map

Fig. 1: a). Experimentally measured transient photovoltage map for a CQD solar cell. This device was used to validate the machine learning models. b). Predicted transient photovoltage map for the same cell by a convolutional autoencoder.

datasets and a promising alternative to traditional approaches for determining materials parameters in complex materials such as colloidal quantum dot thin films that may not behave like traditional semiconductors in all aspects.

In the following sections, we introduce our device structure and experimental setup. We explain in detail the materials parameters studied and how we measure them. Next, we discuss the inner workings of several ML models, compare their efficacy, and report their hyperparameters. We also discuss several ways in which model accuracy can be improved. Lastly, the predicted materials parameters from our models are compared to experimentally measured materials parameters.

II. Methods

A. Experimental Setup

We took measurements from several colloidal quantum dot (CQD) thin film solar cells. All the devices share the same structure: a glass substrate, a transparent electrode (fluorine-doped tin oxide), a n-type zinc oxide electron-extraction layer, a bulk absorbing layer made up of a PbS CQD film with PbX_2 (X=Br, I) ligands, a p-type hole extraction layer (PbS CQD thin film with ethanedithiol ligands), and lastly a top evaporated gold contact [5]. The absorbing layer of our devices is around $500\,\mathrm{nm}$ thick.

A custom optoelectronic scanning setup was used to collect data on all the devices [3]. The sample is mounted on an XYZ translation stage, which allows us to create spatially-resolved materials parameter maps. In contrast to single-point measurements, this system allows us to resolve macroscopic physical phenomena such as defect regions and film inhomogeneities. We collected illuminated current-voltage curves using a Keithley 2400 Source Measurement Unit, and we used an Ocean Optics NIRQuest512 spectrometer to collect photoluminescence (PL) data. The entire system is automated to produce parameter maps that are correlated in both space and time.

To determine the electronic trap state density n in our photovoltaic CQD thin films, we utilize the iterative transient photovoltage method [6]. A Thorlabs MCWHL5 White light emitting diode (LED) was used to provide the steady state background illumination, and a Thorlabs L520P50 Laser Diode ($\lambda = 520\,\mathrm{nm}$) was used as a perturbation source. The pulsed laser generates excess electrons which recombine shortly after each pulse. We can calculate the excess charges generated (ΔQ) by integrating the photocurrent transient while the device is under short circuit conditions:

$$\Delta Q = \int I(t)dt$$

The photovoltage transient signal can be modeled to fit a mono-exponential of the following form:

$$\Delta V_{oc}(t) = \Delta V_{oc}(0) \exp\left(\frac{-t}{\tau_s}\right)$$

where $\Delta V_{oc}(0)$ is the maximum change in the open circuit voltage caused by the perturbation source and τ_s is the small signal lifetime [7]. We perform these measurements at different light biases corresponding to different open circuit voltages. Afterwards, a differential capacitance can be calculated using the following:

$$C = \frac{\Delta Q}{\Delta V_{oc}(0)}$$

Integrating this capacitance up to a particular V_{oc} will give us an estimation of the midgap trap state density n:

$$n = \frac{1}{Aed}\int_0^{V_{oc}} C dV \tag{1}$$

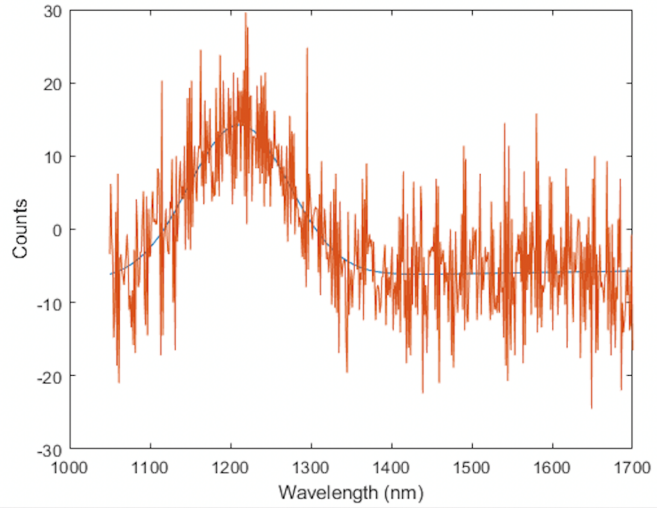


Fig. 2: Photoluminescence plot for a single point in the scan fit to a Gaussian curve.

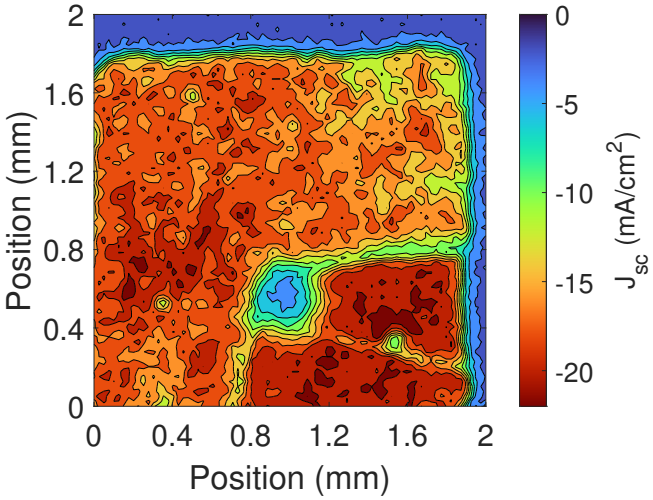


Fig. 3: Short circuit current density map of one of the CQD solar cell devices used to train the neural networks with a visible defect near the center.

where A is the area of the device, e is the electronic charge, and d is the thickness of the active layer. The spot size of the combined laser/LED beam was measured to be approximately $50\,\mu$m, but the translation stage has the capability of stepping in increments as small as $10\,\mu$m. A trade-off is made between the resolution of the transient signals and the diameter of the beam spot. Smaller spot sizes allow us to study devices in greater detail, but this may lead to longer acquisition times and lower signal-to-noise ratios.

Lastly, the collected photoluminescence data was fit to a Gaussian curve of the following form.

$$G = a \exp\left(\frac{-(x-b)}{c}\right)$$

where a, b, and c are constants. Figure 2 shows an example photoluminescence plot, from which we can obtain parameters such as the peak wavelength (λ_{peak}), peak intensity, and full width at half maximum (FWHM) of the intensity.

B. Neural Networks

We predicted materials parameters using a simple multilayer perceptron, an autoencoder (AE), and convolutional variants of each [8]. The goal of autoencoders is to learn the identity function, i.e. the desired output of the network is the input [9]. By itself, that is not useful, but if we add a bottleneck to the network, meaning if we add a layer z that has dimensions smaller than the input layer, then the autoencoder will learn a way to map the input data onto this lower dimensional space [2]. It is in this bottleneck layer that we obtain our materials parameters. We can achieve this by adding an additional term to the cost function J of the autoencoder [10]:

$$J = \frac{1}{N}\sum_{i=1}^{N}(x_i - t_i)^2 + \sum_{j=1}^{M}(k_j - z_j)^2 \qquad (2)$$

where k is the desired materials parameter that we measure and train the network with. Additional terms may be added for more materials parameters. Note that there are two parts to an autoencoder: the encoder and the decoder. The encoder performs dimensionality reduction and also learns the relationship between the input space and latent space. The decoder can be used alongside a Gaussian noise generator to create a generative adversarial network (GAN). This is one possible method of using ML to generate datasets for future use.

III. Discussion/Results

Four different networks were used in our study: a simple multilayer perceptron (MLP) with six hidden layers, one convolutional MLP network, one autoencoder (AE) network, and lastly a convolutional AE network [8]. All models were trained for 20 epochs with a learning rate of $\eta = 2.5 \times 10^{-3}$. The Rectified Linear Unit (ReLU) activation function was used between each layer. The input to each model was a 28×2 vector which is the illuminated current-voltage curve. For preprocessing, the input current-voltage curve and materials parameters were scaled to the range of 0 to 1 and zero-centered. We scale the data to make the problem bounded and because of the properties of the sigmoid and ReLU activation functions [11]. The data is zero centered in order to help convergence to a solution during gradient descent. In a typical solar cell device, the current is negative from zero volts all the way to the open-circuit voltage. This is undesirable because then the gradients will all have the same sign and be limited to either the negative or positive direction during training [11].

The MLP models are based on a network with six layers that have 200, 150, 150, 50, 50, and 50 neurons respectively [1]. Both convolutional MLP and AE models used a filter of size 7×2 with stride length of 1. AE networks had a hidden layer of size 1×1, which is equal to the materials parameter being

TABLE I: Mean Squared Errors for Predicted Materials Parameters from Different ML Networks

	MLP	**Conv. MLP**	**AE**	**Conv. AE**
Trap State Density	0.065	0.063	0.058	0.057
Peak PL Wavelength	0.026	0.022	0.019	0.018
Transient Photocurrent Decay	0.071	0.069	0.062	0.060

TABLE II: Mean Squared Errors While Predicting Multiple Materials Parameters

No. of Variables	**1**	**2**	**3**
Trap State Density	0.057	0.054	0.053
Peak PL Wavelength	-	0.017	0.017
Transient Photocurrent Decay	-	-	0.073

trained. We plotted these various hyperparameters against the mean squared error (MSE) for several materials parameters, and picked the values that minimized MSE across all variables.

To test our networks, we used data maps from four different devices. In total, we measured 18,124 unique points on the devices within the maps. Three of the devices were used for training, and the remaining device was used for validation. An image of one of the training devices is given in Figure 3, and an image of the validation device is given in Figure 1. Table I summarizes the results of the various ML models. Overall, we find that convolutional networks out perform their non-convolutional counterparts.

We took the best performing network (convolutional AE) and added additional hidden neurons to the bottleneck layer. The results are tabulated in Table II. Surprisingly, MSE decreased for both the peak PL wavelength as well as the electronic trap state density as we increased the number of latent variables. This result is counter-intuitive because the form of the cost function given in Equation 2 includes a trade-off between the optimization of different parameters.

From this fact, we point out that ML models can only find dependencies if they are actually provided in the dataset. For example, the bandgap of some semiconductors is a function of both interatomic spacing and temperature. If we train a model to find the bandgap, but only provide interatomic spacing data, then we would be worse off than if we combined both spacing and temperature data. We find that the increase or decrease of MSE can be used as a proxy to determine which variables are physically correlated and not just statistically correlated.

Lastly, qualitative results from our model training is shown in Figure 1. We plot both the measured and predicted values of $\Delta V_{oc}(0)$. There is strong agreement between the two values, and because the values are of the same order of magnitude, we conclude that the system was able to properly learn the mapping function for this particular device. These preliminary results demonstrate that this method holds promise for simplifying photovoltaic materials parameter measurements.

IV. Conclusion

We demonstrated several simple machine learning methods to approximate key materials parameters in PbS CQD solar cells. These models not only enable faster device optimization, but also shed insight on the underlying physics and relationships between materials parameters. Compared to conventional methods, ML models are time- and cost-effective. This work is not only applicable to photovoltaic devices, but could be extended to other types of optoelectronic devices such as photodetectors and light emitting diodes. Future work will incorporate unsupervised machine learning methods (e.g. self-organizing maps) to automatically characterize different regions of devices. In addition, we plan to build a GAN in conjunction with the decoder of the AE model to allow for the creation of large and physically-motivated datasets. Because our training data is spatially resolved, we will be able to simulate non-uniform devices with features and defects such as spin-casting streaks and hairline cracks and predict their effects on device performance. We plan to eventually extend this work to other photovoltaic technologies, and encourage the field to make experimentally correlated data publicly available. This work paves the way for simplifying measurements in photovoltaics and could lead to a faster development cycle for new solar cell technologies.

References

[1] N. Majeed, M. Saladina, M. Krompiec, S. Greedy, C. Deibel, and R. C. MacKenzie, "Using deep machine learning to understand the physical performance bottlenecks in novel thin-film solar cells," *Advanced Functional Materials*, vol. 30, no. 7, p. 1907259, 2020.

[2] J. Li, K. Lim, H. Yang, Z. Ren, S. Raghavan, P.-Y. Chen, T. Buonassisi, and X. Wang, "Ai applications through the whole life cycle of material discovery," *Matter*, vol. 3, no. 2, pp. 393–432, 2020.

[3] Y. Lin, T. Gao, X. Pan, M. Kamenetska, and S. M. Thon, "Local defects in colloidal quantum dot thin films measured via spatially resolved multimodal optoelectronic spectroscopy," *Advanced Materials*, vol. 32, no. 11, p. 1906602, 2020.

[4] F. Häse, L. M. Roch, P. Friederich, and A. Aspuru-Guzik, "Designing and understanding light-harvesting devices with machine learning," *Nature Communications*, vol. 11, no. 1, pp. 1–11, 2020.

[5] A. Chiu, C. Bambini, E. Rong, Y. Lin, and S. M. Thon, "New hole transport materials via stoichiometry-tuning for colloidal quantum dot photovoltaics," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, pp. 1096–1097, IEEE, 2020.

[6] C. Shuttle, B. O'Regan, A. Ballantyne, J. Nelson, D. D. Bradley, J. De Mello, and J. Durrant, "Experimental determination of the rate law for charge carrier decay in a polythiophene: Fullerene solar cell," *Applied Physics Letters*, vol. 92, no. 9, p. 80, 2008.

[7] D. Abou-Ras, T. Kirchartz, and U. Rau, *Advanced characterization techniques for thin film solar cells*. John Wiley & Sons, 2016.

[8] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[9] K. P. Murphy, *Probabilistic machine learning: an introduction*. MIT press, 2022.

[10] Z. Ren, F. Oviedo, H. Xue, M. Thway, K. Zhang, N. Li, J. D. Perea, M. Layurova, Y. Wang, S. Tian, *et al.*, "Physics-guided characterization and optimization of solar cells using surrogate machine learning model," in *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*, pp. 3054–3058, IEEE, 2019.

[11] N. Buduma and N. Locascio, “Fundamentals of deep learning: Designing next-generation machine intelligence algorithms.”