XGBoost Trained on Synthetic Data to Extract Material Parameters of Organic Semiconductors

Evelyne Knapp
and Mattia Battaglia
Zürich University of Applied Sciences
Institute of Computational Phyiscs
Winterthur, Switzerland
Email: evelyne.knapp@zhaw.ch

Thilo Stadelmann
Zürich University of Applied Sciences
ZHAW Datalab
Winterthur, Switzerland
ECLT Fellow
Venice, Italy

Sandra Jenatsch and Beat Ruhstaller
Fluxim AG
Winterthur, Switzerland

***Abstract*—The optimization of organic semiconductor devices relies on the determination of material and device parameters. However, these parameters are often not directly measurable or accessible and may change depending on the neighboring materials in the layered stack. Once the parameters are known, devices can be optimized in order to maximize a certain target, e.g. the brightness of a LED. Here, we combine the use of machine learning and a semiconductor device modelling tool to extract the material parameters from measurements. Therefore, we train our machine learning model with synthetic training data originating from a semiconductor simulator. In a second step, the machine learning model is applied to a measured data set and determines the underlying material parameters. This novel and reliable method for the determination of material parameters paves the way to further device performance optimization.**

***Index Terms*—XGBoost, synthetic data, organic semiconductor, parameter extraction**

I. Introduction

Organic light-emitting diodes (OLEDs) [1] are successfully commercialized in display applications. To overcome limitations in stability and efficiency, further research efforts are crucial. A thorough understanding of the device operation is key which again requires the knowledge of material parameters. Traditionally, material parameters are determined with the aid of dedicated measurements. Some material parameters may vary depending on the sequence of the device layers or measurement techniques, others are not directly measureable. An alternative approach is fitting a device simulation to the corresponding measurement and to derive the material parameters from the simulation. Commonly, least-squares algorithms such as e.g. Levenberg–Marquardt are used to minimize the sum of squared difference between the measurement and simulation. Due to the amount of unknown material parameters and their correlation, multiple (different) experiments [2] are performed and fitted leading to a multi-modal error landscape. The error optimization between measurements and simulations has been demonstrated in various publications [3]–[5], but the process is still not fully automated and requires domain knowledge from the user to direct the search in an appropriate direction or to escape a local minimum. In this contribution, we apply machine learning to the material parameter extraction problem, namely the XGBoost algorithm which is a competitive alternative to neural networks [6]. A set of simplified single-carrier p-doped/intrinsic/p-doped devices varying in thickness is therefore measured and analyzed. The same data has already been used in combination with a manual fitting approach [3]. The production and characterization of such prototype devices is however time-consuming and involves a lot of manual steps. Therefore, the data is extremely scarce and not suited for training a machine learning model. To apply a data-driven machine learning approach a physical model is used for the training data generation.

II. Approach and Workflow

The approach taken in this contribution is visualized in the workflow in Fig. 1. The device under analysis is described in Section III.

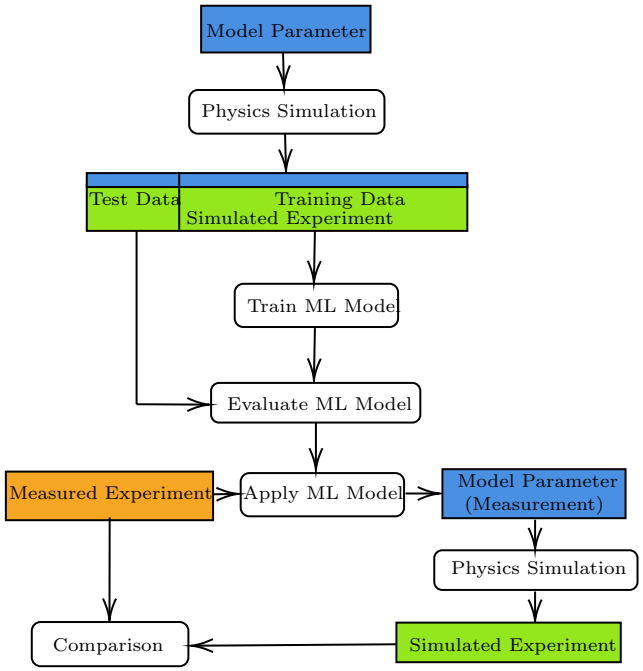


Fig. 1. Workflow of synthetic data generation and subsequent machine learning and validation on the measurement data.

As a first step, we generate synthetic data with the aid of a semiconductor simulator. The resulting simulations are current-voltage curves and electrochemical impedance spectroscopy simulations and further described in Section IV. The data set is then split into a training and test set which are firstly used to train the model as explained in Section V and secondly to evaluate the performance on unseen synthetic data as reported in Section VI-A. Once the training is terminated we present the measurements that are of the same structure as the synthetic data to the machine model and show the results in Section VI-B. In this step, the underlying physical material parameters are predicted which are then used in a semiconductor simulation to reproduce the measured experiment. Once a good agreement between the measurement and the simulation is obtained, the optimization process of the device would start and the influence of parameters on the overall performance would be investigated which is, however, not further pursued in this work. The approach introduced above assumes that the physical model captures the main features in the measurements and is an adequate description of the underlying physical processes.

III. Device under Investigation

The analysis is concerned with three hole-only devices consisting of a 100, 150, and 200 nm thick intrinsic tris[(3-phenyl-1H-benzimidazol-1-yl-2(3H)-ylidene)-1,2-phenylene]Ir (DPBIC) layer, respectively, which is sandwiched between two 30 nm thick, 10 Vol. % MoO_3 doped (p-type) DPBIC layers. The contact is made of indium tin oxide (ITO) and gold (Au) which ensure a good band alignment with the Highest Occupied Molecular Orbital (HOMO) of DPBIC (5.28 eV [3]). The device structure is shown in Fig. 2 (not to scale) with an additional external series resistance. All devices were measured at room temperature and in the dark with Paios [7]. The current–voltage measurement is the most basic characterization method for OLEDs and solar cells. A more advanced technique is the electrochemical impedance spectroscopy that determines the impedance, i.e. the AC resistance, of electrochemical systems as a function of the frequency of a small AC voltage V_{AC} that is added to an offset voltage. The oscillating current I_{AC} is measured, and the resulting impedance Z is calculated according to $Z = \frac{V_{AC}}{I_{AC}}$. As the current might be phase-shifted with respect to the AC voltage modulation the impedance Z is complex and can be represented in different ways. We will use the Nyquist representation for the impedance at a certain offset voltage while the frequency is swept. The second impedance measurement uses a fixed frequency while we sweep the voltage resulting in an impedance-frequency representation. All measurements on one device were carried out subsequently without changing the contact pins. For impedance analysis, an oscillating voltage modulation of 70mV was used.

IV. Synthetic Data Generation

Algorithms, frameworks and machine learning packages have flourished over the last decade and a wide range and variety is available. A more delicate and scarce resource is high-quality data [8]. To circumvent this problem we generate synthetic data. Synthetic data is artificially created by simulations and not collected from the real world or generated by actual events. The advantages of synthetic data generation are manifold: The amount of data can easily be increased at the cost of simulation time. The diversity of data can be chosen such that all possible scenarios are included and the data is perfectly annotated.

For our data generation case, we refer to semiconductor modelling and use a simplified and reduced OLED structure that facilitates material characterization for holes in organic semiconductors. Further, the three-dimensional OLED geometry is reduced to a one-dimensional simulation domain as shown in Fig. 2 as the red line. The material parameters that determine the behavior of the device are mostly unknown or can only be measured with great effort. The goal of this work is to determine these underlying material parameters from device characterization measurements. We display the unknown parameters in Fig. 2 in their corresponding domain. The doped

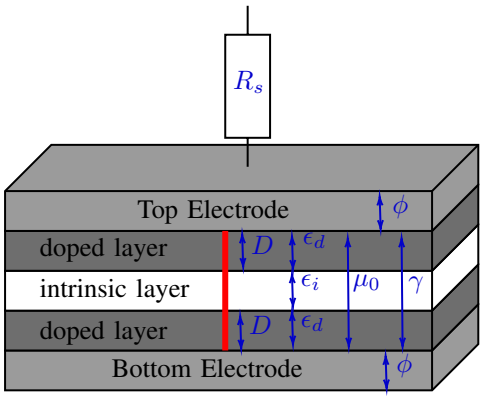


Fig. 2. Device structure and material parameters of a single-carrier device. The red line indicates the one-dimensional simulation domain for the semiconductor modelling.

TABLE I
Material Parameter Overview

Symbol	Name	Sampling Interval	Units	Position
R_s	Series Resistance	uniform [49,53]	Ω	External
ϕ	Work Function	uniform [4.8,5.28]	eV	Electrodes
μ_0	Mobility	log-uniform [1e-10,1e-5]	$cm^2/(Vs)$	Doped & Intrinsic Layers
γ	Mobility Field-Enhancement Factor	log-uniform $[10^{-6}, 10^{-3}]$	$\sqrt{m}/\sqrt{V}$	Doped & Intrinsic Layers
D	Doping Density	log-uniform $[10^{23}, 10^{29}]$	m^{-3}	Doped Layers
ϵ_d	Relative Permittivity	uniform $[3.5, 6.5]$		Doped Layers
ϵ_i	Relative Permittivity	uniform $[3.5, 6.5]$		Intrinsic Layers

layers differ in terms of the relative permittivity ϵ and the doping density D from the intrinsic layer. The field-dependent Poole–Frenkel mobility model, $\mu(E) = \mu_0 \exp\left(\gamma\sqrt{|E|}\right)$ where μ_0 is the zero-field mobility, γ the field-enhancement factor, and E the electric field is the same in all semiconductor layers. As input parameters for the simulation we vary the values within the boundaries indicated in Table 1. Depending on the parameter the values are uniformly or log-uniformly sampled. For the physical interpretation of the material parameters refer to [3]. With the simulation tool Setfos [9] we create, by randomly varying the material parameters, 100'000 sample simulations for all three thicknesses (100, 150, and 200 nm) that are used for the training. Therefore, we solve the system of coupled partial differential equations for semiconductors [10]–[12] on the one-dimensional domain and vary the seven material parameters simultaneously within prescribed boundaries from Table I. In Fig. 3 a single sample of the synthetic data set is shown with the predictor and target variables as well as their sizes. The input data of the physical model is the output data of the machine model and vice versa. In summary, the data set consists of a current-voltage simulation, impedance simulations at two bias voltages with a frequency sweep, and an impedance simulation for a fixed frequency with a voltage sweep for each device thickness.

V. Training of Machine Learning Model

With the synthetic data we proceed in the flow chart in Fig. 1 and split the data in training and test sets in a ratio of 80% to 20%. The training set is used to train the machine model while we validate the model with the test set. We deal with a multi-target regression problem that is concerned with the prediction of multiple continuous target variables using a shared set of predictors. In the following we apply XGBoost [13] which is short for eXtreme Gradient Boosting package to the regression problem. It was created by Tianqi Chen [14] and is an efficient and scalable implementation of gradient boosting framework by [15], [16] along with some regularization factors. XGBoost is considered for supervised machine learning tasks in classification and regression and has performed well for structured, tabular data in the past.

XGBoost is categorized as boosting techniques in ensemble learning. Boosting is a method that combines simpler and weaker models (trees) to make better predictions of the target variable. Models are added gradually and in a sequential manner until there are no more improvements in the predictions. Gradient boosting uses the gradient descent algorithm to add the simpler models and thereby minimizes a regularized objective function. It is a combination of a convex loss function that takes the difference between the predicted and target outputs into account and a penalty term for model complexity. The training is performed iteratively by adding a new tree to reduce the residuals of the current ensemble of trees.

We create for each target variable an XGBoost model that is individually trained. The target variables were previously transformed by logarithmizing (if necessary) and scaling each

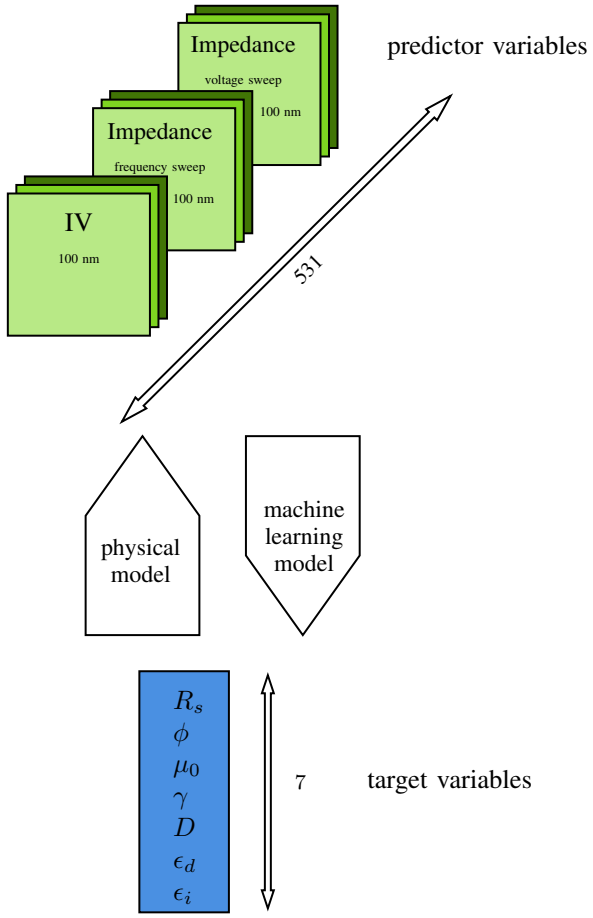


Fig. 3. Simulated predictor (top) and corresponding target (bottom) variables with their dimensions. The physical model works in the opposite direction as the machine learning model.

target between 0 and 1. The XGBoost model was trained on 80'000 training samples and validated on 20'000 test samples. For the training we use a manually tuned learning rate $\eta = 0.25$, the maximum depth of a tree is limited to 15 and the maximum number of trees to 40. As a loss function the Root Mean Square Error (RMSE) is selected. In Fig. 4 the train and validation loss is plotted against the number of trees for three exemplary parameters. The first material parameter is the zero-field mobility μ_0 and represents a successful training. The training loss function decreases fast and levels out. The loss function of the test set shows a very similar behavior with the difference that the plateau is slightly higher. The second parameter is the intrinsic relative permittivity ϵ_i. Its performance is to some extent worse since the training set reaches a clearly lower value than the test set. The third parameter, the doped relative permittivity ϵ_d, behaves differently because the training loss is still decreasing while the loss function of the test set remains on the same level. The reason for the worst performance lies in the sensitivity of the physical model to the particular parameter, i.e. the model is not very sensitive to the doped relative permittivity ϵ_d and therefore hard to predict, or in other words, a change in the

doped relative permittivity ϵ_d has little impact on the result of the semiconductor device simulation in contrast to other material parameters.
The remaining parameters act as in the first or second case in Fig. 4. Only the work function ϕ has the characteristic of the third scenario, the physical reasoning being that the work function ϕ has little impact on the device properties if the boundary layers are highly doped.

VI. Evaluation

A. Synthetic Test Data

The performance of the XGBoost model on the material parameters is analyzed again for the three exemplary parameters in Fig. 5 where an identity chart is displayed. The material parameter value predicted by the XGBoost model versus the true parameter value from the training or test set is plotted. As a guide-to-the-eye the red diagonal represents perfect prediction on the identity chart. The first parameter is the zero-field mobility μ_0 which can be very well predicted even for the test set as displayed in a). Also decreasing the size of the training set has only little effect on the prediction. As a second parameter we present in b) the intrinsic relative permittivity ϵ_i where the prediction is still valid, but worse than in the previous case. The uniform sampling of the relative permittivity was rounded to one decimal place as visible in the figure. A parameter that is more difficult to predict is the doped relative permittivity ϵ_d shown in c). The spread around the diagonal is wider for the training set and even worse for the test set. The results in Fig. 5 are very much in line with Fig. 4 and the identity charts visualize again the outcome of the training phase.

B. Measurement Data

In this section, we feed the XGBoost model with the measured current-voltage and impedance spectroscopy data. Since the measurements are reliable and repeatable, we have not performed any pre-processing on the data such as e.g. de-noising. The measured data set has exactly the same structure

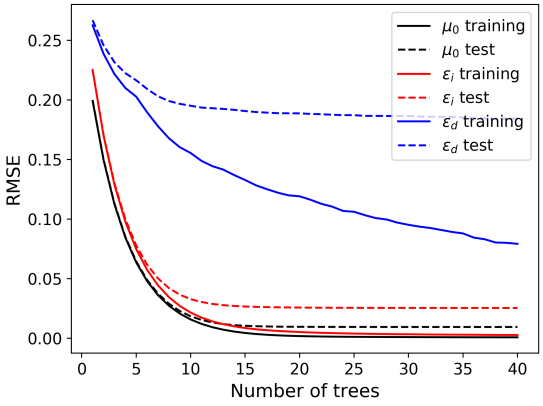
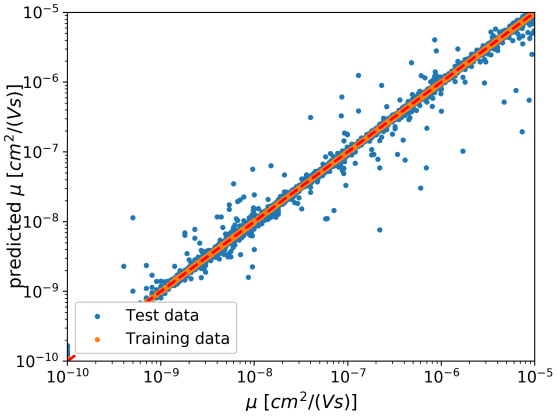
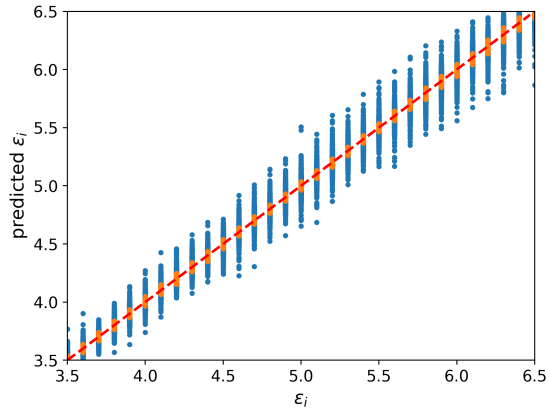


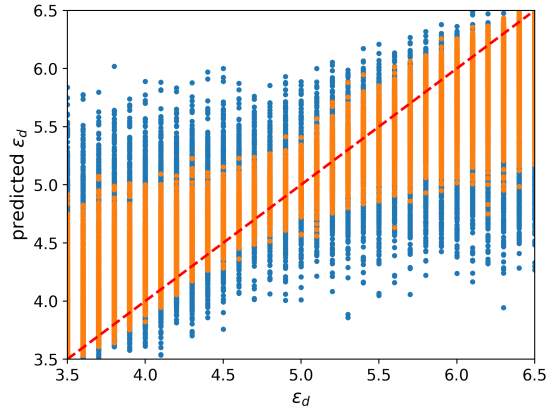
Fig. 4. RMSE loss function for the training and test sets versus number of trees.



(a) For the training set the prediction of the zero-field mobility μ_0 is accurate. A slightly wider spread in the prediction for the test set is displayed.



(b) The predicted intrinsic relative permittivity ϵ_i is shown. The training set focuses more on the diagonal than the test set.



(c) The predicted doped relative permittivity ϵ_d is displayed which is clearly more challenging to predict for the test as well as the training set.

Fig. 5. Identity charts for three exemplary parameters for the training and test sets.

TABLE II
EXTRACTED MATERIAL PARAMETERS FROM XGBOOST

Symbol	Value	Units
R_s	49	Ω
ϕ	5.00	eV
μ_0	1.1×10^{-7}	$\mathrm{cm}^2/(\mathrm{Vs})$
γ	8.5×10^{-4}	$\sqrt{\mathrm{m}}/\sqrt{\mathrm{V}}$
D	2.7×10^{26}	m^{-3}
ϵ_d	6.0	
ϵ_i	4.2	

as the synthetic data in Fig. 3 and is fed to the XGBoost model. The extracted material parameters are re-transformed and shown in Table II. We note that these material parameters obtained with the help of machine learning compare favorably with the ones obtained in a more traditional semi-automatic least square fitting approach [3].

Depending on the size of the training set and the selected hyperparameters the extracted material parameters will slightly vary. Increasing the training data set helps to reduce this variation e.g. for the doping parameter D. A hyperparameter search for each material parameter has the potential to further improve the training and find the optimum configuration of XGBoost. In order to circumvent this problem the resulting material parameter set can be further processed and serve as an initial guess to a local optimization algorithm that minimizes the error between the measurements and the simulations.
As a next step, the material parameters are fed back into the semiconductor simulator to reproduce the measurements. In Fig. 6 the measurements as well as the simulations based on the material parameters predicted by XGBoost are displayed. The first figure shows the measured and simulated current-voltage curves with a very good agreement. Also in the second plot the impedance in log-log representation for two offset voltages represents an accurate description. The plots at the bottom display the real and imaginary part, respectively, of the impedance versus the applied voltage at a constant frequency of 28.8 kHz. The characteristic features are captured in all four situations. The final agreement between measurements and simulations confirm that the semiconductor model consists of all important physical ingredients to describe the measurement with one set of material parameters.

VII. CONCLUSIONS

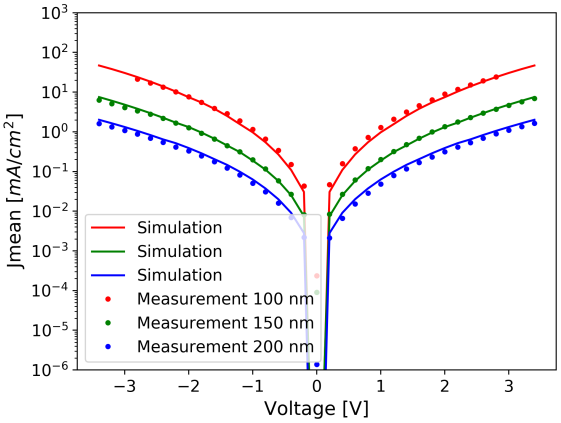
A material parameter extraction problem for single-carrier organic semiconductor devices with three different thicknesses was presented. The approach taken in this work combined a physical semiconductor model for synthetic data generation and machine learning. We successfully trained an XGBoost model on the synthetic data to a multi-target regression problem to determine underlying material parameters from current-voltage and impedance spectroscopy measurement data. The material parameters extracted from the measurement were fed to the semiconductor simulator. The simulation and the measurement are in close agreement. The concept of merging machine learning and physical modelling for data generation is a powerful alternative to classical fitting algorithms provided the simulation times for the physical modelling are short. As a next step, we will introduce a quantitative measure of the fit quality. Further, we envisage to extend the application to the determination of the underlying physical model with its key ingredients. Such a physics-informed machine learning approach can potentially be helpful for various applied physics and engineering problems.

ACKNOWLEDGMENT

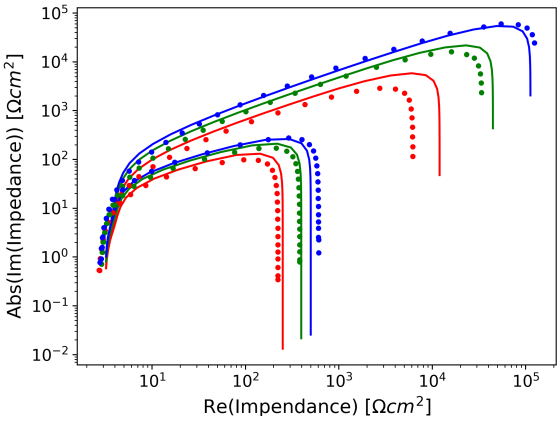
Financial support from Innosuisse (project AIPV 37304.1-IP) in the impulse programme digitalisation is acknowledged. P.-A. Will, S. Lenk and S. Reineke from TU Dresden are acknowledged for providing the organic semiconductor devices based on which the measured data was acquired.

REFERENCES

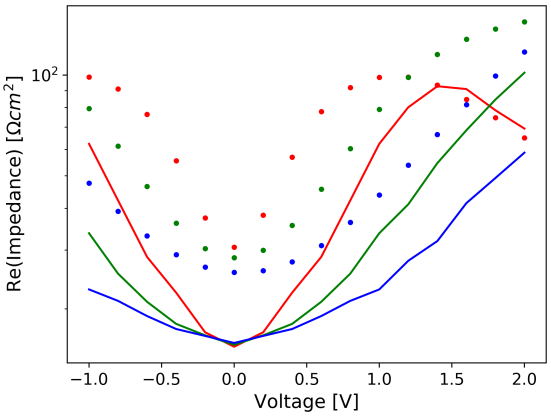
[1] C. W. Tang and S. A. VanSlyke, "Organic Electroluminescent Diodes," Applied Physics Letters, Vol. 51, No.12, 1987, pp. 913-915. DOI:https://doi.org/10.1063/1.98799
[2] M. T. Neukom, S. Züfle, and B. Ruhstaller, "Reliable Extraction of Organic Solar Cell Parameters by Combining Steady-State and Transient Techniques," Org. Electron. 2012, 13, 2910– 2916.
[3] S. Jenatsch, S. Altazin, P.-A. Will, M. T. Neukom, E. Knapp, S. Züfle, S. Lenk, S. Reineke, and B. Ruhstaller, "Quantitative analysis of charge transport in intrinsic and doped organic semiconductors combining steady-state and frequency-domain data," Journal of Applied Physics 124, 105501 (2018); DOI:https://doi.org/10.1063/1.5044494
[4] S. Jenatsch, S. Züfle, B. Blülle, and B. Ruhstaller, "Combining steady-state with frequency and time domain data to quantitatively analyze charge transport in organic light-emitting diodes", Journal of Applied Physics 127, 031102 (2020).
[5] M. T. Neukom, A. Schiller, S. Züfle, E. Knapp, J. Avila, D. Perez-del-Rey, C. Dreessen, K. Zanoni, M. Sessolo, H. J. Bolink, and B. Ruhstaller, "Consistent Device Simulation Model Describing Perovskite Solar Cells in Steady-State, Transient, and Frequency Domain," ACS Appl. Mater. Interfaces 2019, 11, 26, 23320–23328.
[6] https://mljar.com/machine-learning/neural-network-vs-xgboost/
[7] Platform for All-in-One Characterization (PAIOS); Fluxim AG, Switzerland,www.fluxim.com
[8] L. Hollenstein, L. Lichtensteiger, T. Stadelmann, M. Amirian, L. Budde, J. Meierhofer, R. Füchslin, and T. Friedli, "Unsupervised learning and simulation for complexity management in business operations," Applied Data Science, 313–331, Springer, (2019).
[9] Semiconductor Simulator (SETFOS); Fluxim AG, Switzerland,www.fluxim.com
[10] S. Selberherr, "Analysis and Simulation of Semiconductor Devices,"Springer, Vienna, (1984).
[11] E. Knapp, and B. Ruhstaller, "Numerical analysis of steady-state and transient charge transport in organic semiconductor devices,"Opt Quant Electron (2011) 42:667–677, DOI:https://doi.org/10.1007/s11082-011-9443-1
[12] E. Knapp, and B. Ruhstaller, "Numerical impedance analysis for organic semiconductors with exponential distribution of localized states," Applied Physics Letters, 99(9),(2011). DOI: https://doi.org/10.1063/1.3633109
[13] https://xgboost.ai/ or https://github.com/dmlc/xgboost
[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 2016, 785–794. DOI:https://doi.org/10.1145/2939672.2939785
[15] J. H. Friedman,"Greedy Function Approximation: A Gradient Boosting Machine,"The Annals of Statistics, 2001, pp. 1189-1232.
[16] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)," The Annals of Statistics, 2000, Vol. 28, No. 2, 337–407



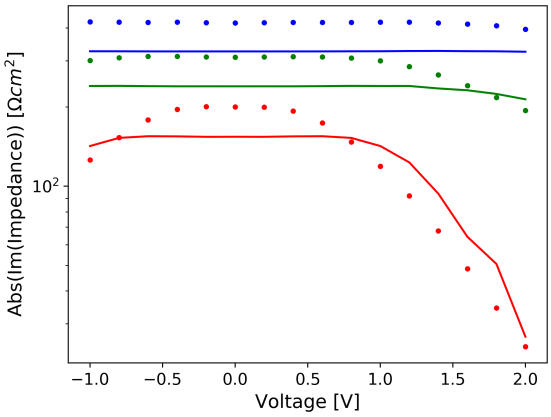
(a) The current-voltage curves for all three thicknesses are shown.



(b) The absolute imaginary part of the impedance is plotted versus the real part of the impedance.



(c) The real part of the impedance versus the applied voltage at a frequency of 28.8 kHz is displayed.



(d) The absolute imaginary part of the impedance versus the applied voltage at a frequency of 28.8 kHz is shown.

Fig. 6. Predicted simulations based on extracted material parameters by XGBoost in comparison with the measured data.