

Algorithmically Extracted Morphology Descriptions for Predicting Device Performance

Wesley K. Tatum,^a Diego Torrejon,^{b,c} Anton B. Resing,^a Jonathan Onorato,^a Christine K. Luscombe^{,a,d,e}.*

^a ~ Department of Materials Science and Engineering, University of Washington, Seattle, WA 98195, United States.

^b ~ BlackSky, 13241 Woodland Park Road, Suite 300, Herndon, VA 20171, United States.

^c ~ Department of Mathematical Sciences, George Mason University, Fairfax, VA, 22030 United States.

^d ~ Department of Chemistry, University of Washington, Seattle, WA 98195, United States.

^e ~ Molecular Engineering and Sciences Institute, University of Washington, Seattle, WA 98195, United States.

* ~ Corresponding Author. E-mail: luscombe@uw.edu

KEYWORDS: Machine Learning/ OPV/ Morphology/ QSPR/ Data Science/ Conjugated Polymers.

ABSTRACT: The device performance of thin film electronics is known to be dependent on morphological properties such as the size, shape, and orientation of aggregates and crystalline domains. So far, descriptions of morphology have been semi-quantitative and loosely linked to device performance. However, by using our recently reported quantitative analysis tool m2py, we create morphology labels and, for the first time, extract quantitative morphology information from scanning probe microscopy measurements. In this work, we use the labels and extracted morphology measurements to validate a novel approach of generating quantitative structure-property relationships through the use of machine learning and regression models; we present this generalizable approach by demonstrating it on organic photovoltaics. In our approach, the open-source toolkit, m2py, is used to label and describe a set of organic photovoltaic devices that have received different thermal annealing treatments, thereby altering their morphologies and, subsequently, their performance. Different regressors and types of morphology descriptions are used to examine the efficacy of quantitative morphology descriptors in predicting device performance. Despite the fact that scanning probe measurements only image the thin film surface, we find that the information-rich nature of morphology data enables accurate device performance predictions, even with datasets that would traditionally be considered too small to generate high-quality predictions. The implications of introducing quantitative morphology information for predictive analysis of other devices and materials is also discussed; given the general and material-agnostic nature of m2py, it is anticipated that the predictive capability of these labels will be ubiquitous to all thin film applications.

1. INTRODUCTION

The performance and properties of thin film electronic devices are strongly dependent on the morphology of the active layer. For instance, in π -conjugated polymers (CPs), it is important to develop methods for extracting quantitative morphological information on the nanometer- and micrometer-scale, as it is often features at these length scales that determine bulk performance for CP-based devices.¹⁻⁶ The current standards for quantitative measurements of morphological parameters, such as crystal structure and crystallinity, are techniques such as grazing incidence wide-angle X-ray spectroscopy or differential scanning calorimetry. These techniques are bulk-averaged measurements, meaning that they are insensitive to many nanoscopic changes in morphology. In order to progress towards truly predictive models of materials properties and device performance, it is useful 1) to be able to generate quantitative and reproducible descriptions of morphology that communicate local, nanoscopic details and heterogeneities, and 2) to integrate this quantitative morphology information into predictive models.

Our previous work addressed the first of these two goals by developing and describing a universal segmentation workflow that algorithmically identifies different material phases and morphological domains in scanning probe microscopy (SPM) datasets.⁷ Although only the surface is imaged by SPM, rather than the full morphology, surfaces imaged by SPM techniques are commonly used to represent active layer morphologies and are consistently shown to be correlated with thin film properties and performance. To access the important features comprising a thin film surface, the m2py workflow takes in SPM measurements to produce labels that describe the size, shape, and identity of every domain in the image. Such quantitative morphology reports will be important for defining and comparing differences between underperforming and optimal active layer morphologies, as well as for probing thermodynamic phase transitions and improving

simulations of materials, all of which require quantitative insight into the nanoscale structure of the materials.^{8–13}

In the current work, we address the second step – using the m2py-generated morphology descriptions to predict final device performance. Predictive models for organic photovoltaic (OPV) performance have focused primarily on chemical structure or electron density information,^{14–19} and there has been little integration of morphological information in predictive models.^{6, 20–22} To validate our approach, a series of OPV devices with bulk heterojunction (BHJ) active layers that are made from the well-understood materials poly(3-hexylthiophene) (P3HT) and phenyl-C₆₁-butyric acid methyl ester (PC₆₁BM). The OPV devices are identical in fabrication, except for a post-deposition thermal annealing treatment, which is known to affect morphology and final device performance.^{23–25} The P3HT:PC₆₁BM OPV power conversion efficiency (*PCE*), open-circuit voltage (V_{oc}), short-circuit current (J_{sc}), and fill-factor (*FF*) are measured across a range of annealing times and temperatures. For a subset of these devices, the surface morphology of the active layers was imaged using fast force-distance mapping (FDM), a nanomechanical SPM technique. While this manuscript utilizes SPM datasets as an example input of morphological data, other data formats, including scanning tunneling microscopy data, could be used as inputs following a similar methodology. The m2py toolkit was used to generate morphology-label maps and domain measurements from this surface morphology dataset. Using this well-understood OPV system, we show that incorporating quantitative morphology information drastically improves model learning and predictive accuracy, even when using datasets that are typically too small to perform well.

2. METHODS AND MATERIALS

P3HT was purchased from Rieke Metals (4002-EE). Glass/ITO substrates were purchased from WRS. All other materials were purchased and used as-received from Sigma-Aldrich.

2.1 REGRESSION AND CODE DEVELOPMENT

All modules were written in Python. NumPy²⁶ was used for the preparation and manipulation of data. PyTorch was used to create and train all artificial neural networks.²⁷ Matplotlib²⁸ and Seaborn²⁹ were used to create all data plots. Scikit-learn³⁰ was used for regression and optimization. To maximize data use and mitigate overtraining, five-fold cross validation was used for model training. Loss and accuracy were monitored during training cycles to ensure overfitting did not occur. The results of this training are shown in the supporting information, Section II. All hyperparameters were optimized to balance for maximum training stability, minimum loss, and minimum error by using Scikit-learn's GridSearchCV(), as discussed in the supporting information, Section III. All data was subjected to MinMax normalization before regression to account for differences in orders of magnitude between the different device metrics and their units. Batch normalization was used in all neural network models. Several optimization algorithms (Adam, AdamW, SGD, Adagrad) were tested for performance; however, the optimizer used in this work, Adam, resulted in the best performance.

2.2 SUBSTRATE PREPARATION

All substrates were cleaned by scrubbing with Micro90 soap and DI water, followed by sonication in deionized water, acetone, and isopropyl alcohol for 15 min each. Following sonication, the substrates were dried under a stream of air and subsequently plasma cleaned for 15 minutes using air plasma. As one of the transport layers in these P3HT:PC61BM OPV, ZnOAc was spin-coated at 2500 RPM for 60 seconds under air. After annealing at 150 °C for 15 minutes,

the ZnO transport layer was formed and the substrates were transferred to a glovebox for active layer deposition.

2.3 ACTIVE LAYER DEPOSITION

P3HT:PC₆₁BM blends (1:0.95, total concentration 30 mg mL⁻¹) were stirred in chlorobenzene overnight at 70 °C prior to spin-coating. Spin-coating was performed at 2000 RPM for 60 seconds. Solution preparation and spin-coating were all carried out in a nitrogen glovebox.

2.4 OPV DEVICES AND TESTING

Using an Ångström Engineering NextDep PVD system, 10 nm of MoO₃ were deposited, followed by 100 nm of Ag. Both films were deposited at 0.5 Å/s under no more than 10⁻⁶ torr of pressure. Through a photomask, 8 different back contacts and a single front contact was deposited. After contact deposition, the devices were annealed as prescribed by Table S1, still in the glovebox. Finally, using a VeraSol Solar Simulator, *J-V* profiles were extracted while the devices were under AM 1.5 illumination. Current–voltage characteristics were measured using a 2400 Series Keithley Source Meter and calibrated with an IR-filtered Si reference cell. The illuminated area of each pixel is 0.013 cm². A range of voltages from -0.2 to 1.0 V was used. *J-V* curve analysis was performed to extract *PCE*, *V_{oc}*, *J_{sc}*, and *FF*. The Python code associated with this GUI is available open-source and free of charge from www.github.com/wesleyktatum/OPV_analysis.

2.5 FAST FORCE-DISTANCE MAPPING ATOMIC FORCE MICROSCOPY

FDM AFM micrographs were obtained on a Bruker ICON AFM in Peak-force Quantitative Nanomechanical Mapping mode. ScanAsyst-Air tips were used for these measurements (SiN/Al, ~70kHz, and 0.4 N m⁻¹). Samples were stored in a N₂ glovebox prior to analysis.

3. RESULTS AND DISCUSSION

3.1 DATA AND SAMPLING

OPVs made from P3HT:PC₆₁BM were fabricated in an inverted architecture and thermally annealed across a range of temperatures and durations, as indicated in Figure 1. Once fabricated, the devices were tested to determine four different metrics of their performance – PCE , V_{oc} , J_{sc} , and FF . In agreement with literature, the devices with the highest performance were obtained when the active layers were annealed near 150 °C for 15-30 minutes, due to active layer morphology evolution during annealing.^{31,32} To record the active layers' morphology, three representative 1 μm^2 FDM scans were taken for a subset of the substrates, representing the morphology of all devices for that annealing procedure. Following the guidelines outlined in our previous work, thirty-six of the FDM imaged active layers were labeled with domain and phase labels by an m2py workflow.⁷ Briefly, this workflow assigns each pixel in an image with two sets of labels that designate the pixel's material phase and domain. Figure 1 shows examples of these m2py phase labels at each of the sample points across the range of annealing conditions and illustrates the thermally driven evolution in morphology and phase distribution. For the imaged OPV, three different phases are labeled: polymer-rich, fullerene-rich, and mixed. As expected, short annealing times combined with low annealing temperatures resulted in the nucleation of small, spherical aggregates of the polymer-rich phase. With more heat and time, these aggregates ripen into more elongated and co-continuous domains, which are necessary for efficient charge separation and transport. This is expected, as the highest performing P3HT:PC₆₁BM devices tend have ~150 °C annealing for 15-30 min.³¹ Through the m2py labels, these rules of thumb for active layer optimization can be quantitatively identified and described using domain shape descriptors such as eccentricity, perimeter, and orientation. The descriptors' measurements are easily plotted, as

shown in Figure 2.a, allowing for the direct viewing of morphological evolution throughout annealing conditions. For example, our visual observations in Figure 1 regarding the nucleation and growth of domains throughout annealing are clearly reflected in the surfaces shown in Figure 2.a. At lower temperatures and shorter annealing times, the major and minor axis lengths are smaller for the polymer-rich phase, as visible in Figure 2.a. As annealing continues, the major axis length increases while the minor axis length becomes smaller or stays the same. This corresponds to the observed elongation and growth of the polymer-rich phase, which corresponds to increases in the *PCE* and J_{sc} of the OPV devices, shown in Figure 2.b. This is anticipated because these elongated, anisotropic domains are generally known to improve charge separation and transport. When the devices are over-annealed, the elongated domains spheroidize or become irregularly shaped. This sub-optimal morphology is visible in Figure 1 at 175 °C, 30 min. and is reflected in Figure 2.a by the increase in the minor axis length at these conditions, during which the major axis length remains relatively static. As a result of this over-annealed morphology, there is less exciton dissociation and more recombination of charge-carriers. This, in turn, decreases both the J_{sc} and *PCE*, as is clearly seen in Figure 2.b. Also visible in Figure 2.a, as the nucleated domains grow and elongate, the perimeter of the domains concomitantly increases, until spheroidization of some domains and aggregation of others, which is reflected in the slight decrease of the perimeter metric's surface and an increase in its standard deviation. These observations are all in agreement with the literature understanding of morphological impacts on P3HT:PC₆₁BM OPV, which has been empirically developed. However, this reported workflow allows for an unprecedented degree of specificity. Not only are the average sizes extractable, as might be expected of Scherrer analysis, but the standard deviation, median, and variance can be calculated as a result of cataloging all of the individual domains with m2py. Further, the shapes, orientations, and other feature descriptions

of the domains, which are not measurable by diffraction or calorimetry, can be quantitatively reported. Additional representative surface plots that show the evolution of OPV domain descriptors throughout annealing are shown in Figures S3-7 and a comprehensive archive of the morphology and performance surfaces are available for free online at www.github.com/wesleyktatum/py-conjugated. The m2py toolkit source code is also available for free online at www.github.com/ponl/m2py. It should be noted that while the work herein focuses on a P3HT:PCBM OPV architecture, the m2py toolkit is material agnostic, and can extract morphological features from any material. From this, the algorithms presented herein could be applied to other OPV architectures and material compositions.

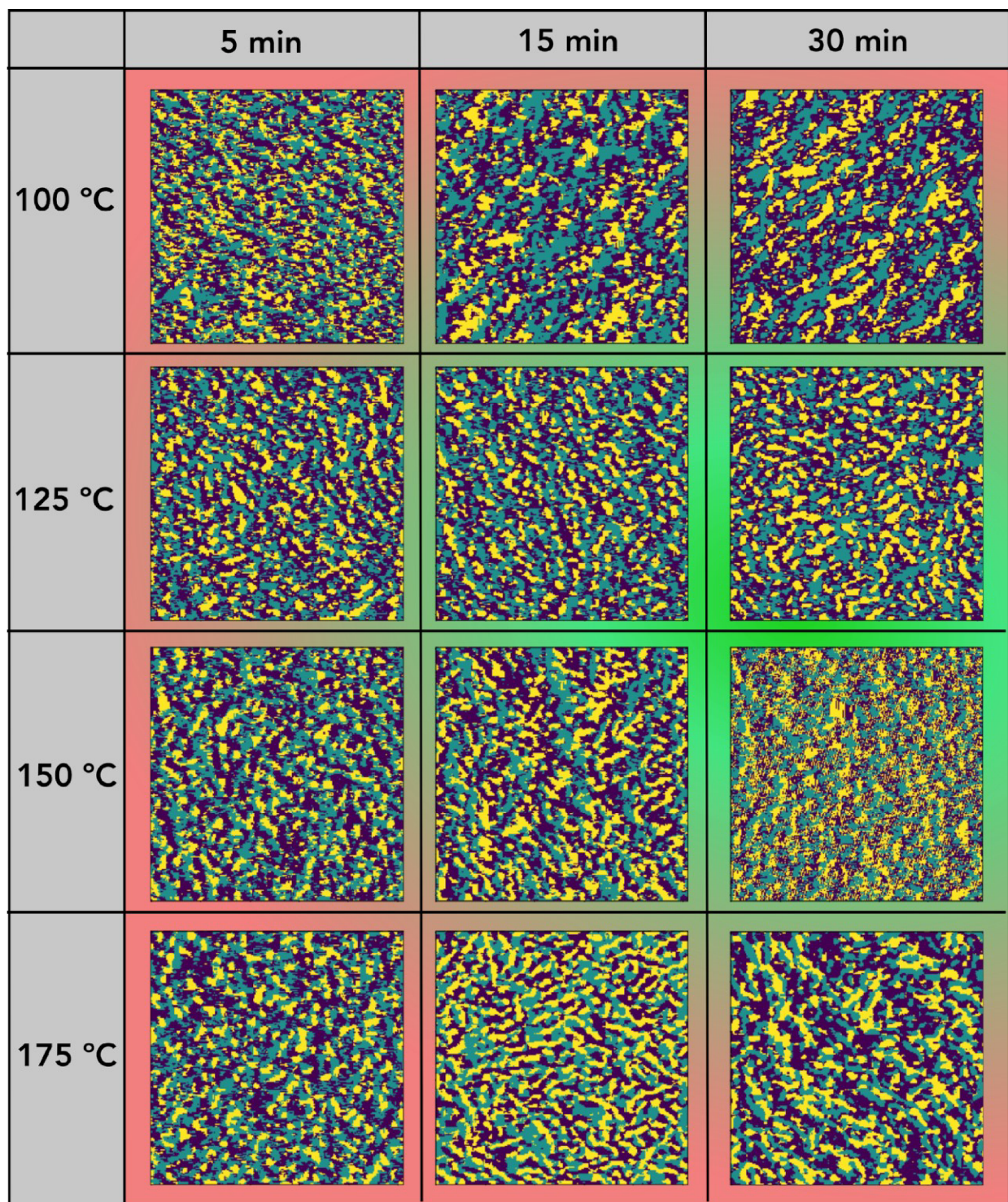


Figure 1. m2py⁷ labeled P3HT:PC₆₁BM active layers. Colors in each square correspond to the 3 expected phases: yellow – P3HT-rich, green – PC₆₁BM-rich, and purple – mixed. The color behind

the squares corresponds to observed device quality, green conditions producing better and more consistent OPV devices than red conditions.

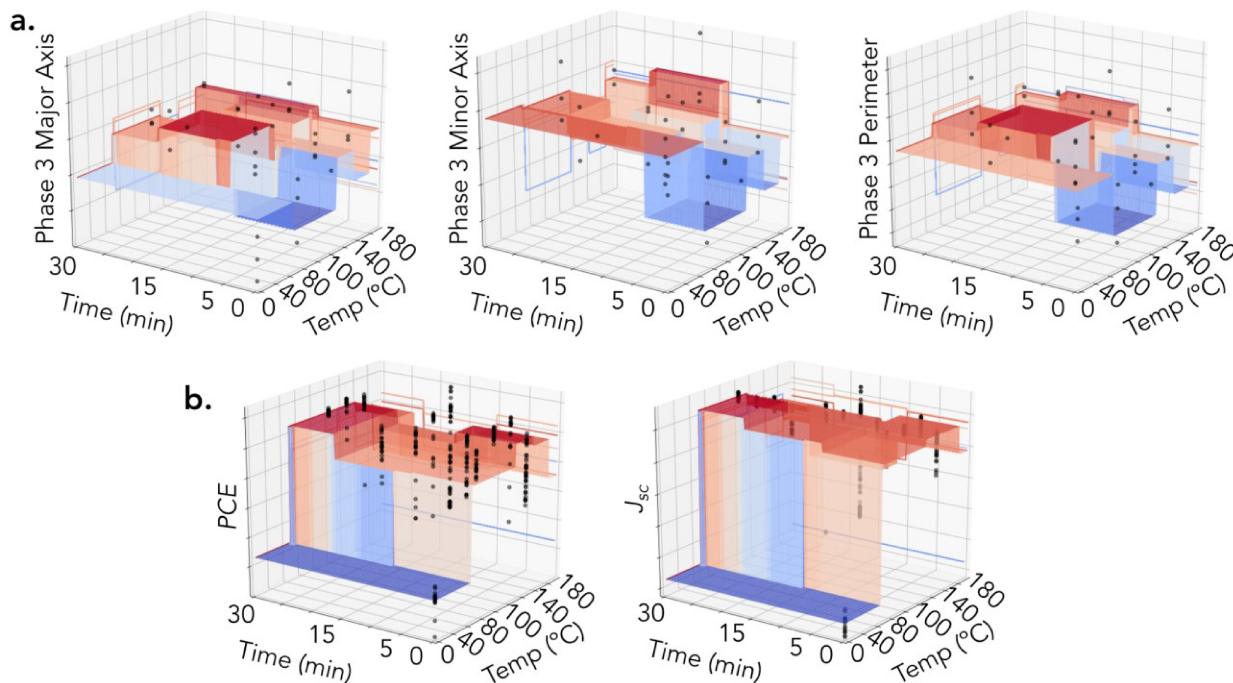


Figure 2. Representative examples of the variations in features throughout the datasets are shown. The average value for all devices at each time-temperature point were used to generate the surface, while each individual device is shown as a black circle. Z-axis units are normalized for all samples to be between 0 and 1. **a.** Three morphology descriptors for phase 3, the polymer-rich phase (yellow in Figure 1), are shown for the imaged device active layers in the complete tabular dataset (CTD). **b.** The PCE and J_{sc} for all devices in the annealing-conditions only dataset (AOD).

Having collected a detailed spread of data on the OPV device performance and active layer morphology, we can examine and compare various methods of encoding morphological information into numerical representations in order to evaluate its influence on the ability of

machine learning models to predict several solar cell performance metrics. To that end, we use different subsets of the overall dataset to isolate and understand the influence of different morphological encodings on our performance predictions. As a baseline for the level of predictivity that annealing conditions alone can have, a dataset containing only annealing conditions and device performance was created. This dataset is referred to as the Annealing-conditions Only Dataset (AOD, $N_{\text{samples}} = 366$) and representative surface plots of the device performances are shown in Figure 2.b. Although it does have the most samples of all the datasets, the AOD encodes no morphology information and is therefore expected to not have enough information to produce good predictions. The most information-rich encoding of morphology information would be to directly use the FDM images of the active layers as an input into our predictive models. This is because the FDM images capture a morphology representation that closely resembles the actual, ground truth surface morphology of the active-layer and can thereby convey many of the ground truth’s local and continuous variations. This information is also encoded across eight different imaging channels that include material properties such as Young’s modulus and deformation. These measurements are defined as the image dataset ($N_{\text{samples}} = 36$). Another substituent dataset used was the labelled images, called the m2py labels dataset ($N_{\text{samples}} = 36$). In this dataset, there are only 3 image channels of morphology information, which are derived from the material phase labels – individual channels that are used to convey a binary representation of the distribution and placement for each of the three phases identified in the OPV thin films (*i.e.* polymer-rich, fullerene-rich, and mixed-phase). Because the m2py labels are a distillation of the FDM images, there are fewer channels and the m2py dataset retains fewer local and continuous fluctuations than the raw FDM images, potentially losing predictive information. However, the information is more densely encoded and could potentially result in improved

performance due to the removal of instrumental noise and overall information distillation. The morphology encodings can be further simplified with m2py by measuring the features of the labeled morphological domains. This most distilled morphological description is the Complete Tabular Dataset (CTD, $N_{\text{samples}} = 36$) and was used to produce the descriptor surfaces in Figure 2.a. In this most simple quantitative morphology description, nine different domain measurement values (relative area, total area, filled area, extent, major axis length, minor axis length, eccentricity, orientation, and perimeter)³³ were averaged across all the domains within each of the phases labeled by m2py. The average values are stored in a table, along with their associated standard deviation. By simplifying the m2py labels into these tabular summaries, the morphology data is able to be directly incorporated into simple predictive models, like polynomial regression. An advantage of this is that the contribution of different morphological features to the final prediction can be more explicitly extrapolated, enabling an understanding of which quantitative morphology descriptors have the largest influence on device performance. In this work, we use several types of regressors to identify contributions of the aforementioned individual features. In particular, we focus on understanding the impact of different morphology encodings on performance predictions of a well-understood OPV system. We utilize traditional machine learning regressors as baseline comparisons, while using neural networks (NN) as the main regressor model because of its feasibility in creating models that accommodate different formats and combinations of data.

3.2 MODEL ACCURACY AND PERFORMANCE

Because the datasets are small, results are reported from 5-fold cross validation, with an 80:20 train-test split for each fold, through custom and Scikit-Learn functions. The regressors tested include custom NNs,²⁷ least absolute shrinkage selection operator (LASSO), ridge, support vector

machine (SVM), and random forest (RF).³⁰ For comparison of model performance, mean absolute percent error (MAPE), mean-squared error (MSE), and the coefficient of determination (R^2) are used. The best results of MAPE and R^2 reported by the cross-validation are shown in Figure 4, and MSE is shown in Figure S1. The best, average, and worst cross-validation values for each NN are reported in Table S1, and the best hyperparameters for all models are reported in Table S2, including architecture parameters for the NN.

The architecture of the NNs is a common backbone followed by multiple output branches as depicted in Figures 3 and S2. One key difference to note is that each NN model predicts all four metrics at once, while each of the non-NN regressors need four separate models to predict the metrics. Thus, the traditional regressors should have outperformed NNs since they are learning multiple models with each specializing on a single metric. However, the results in Figures 4 and

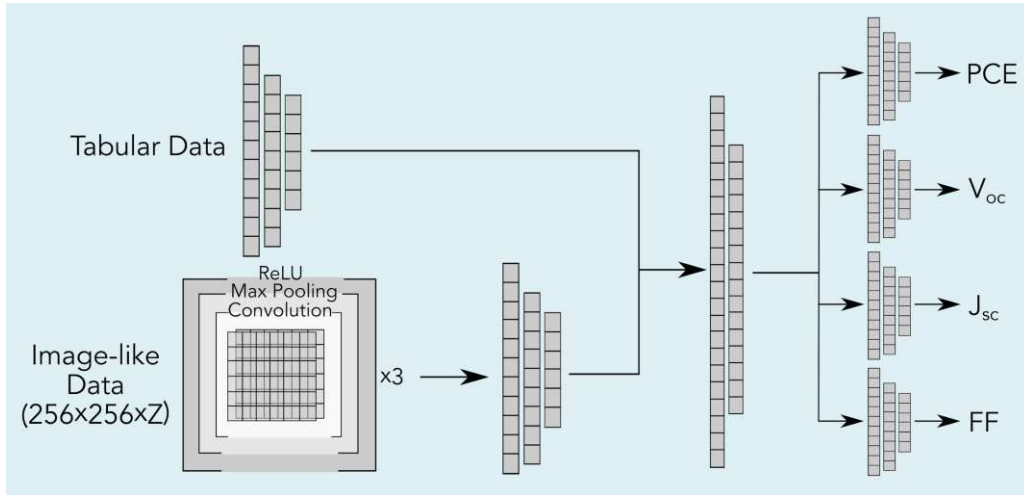


Figure 3. A representation of the structure used for NN5 and NN6. These networks take in both tabular data (the CTD) and image data, where NN5 uses m2py labels and NN6 uses FDM images. After 3 cycles of convolution, pooling, and activation, the flattened image encoding is combined with the CTD encoding. This common embedding is then sent to

each of the 4 branches that are used to predict OPV performance metrics. In each branch, further compression of the encoding allows for specialized predictions for each metric.

S1 indicate that when morphological information is included, the added complexity of the NN regressors make up for the lack of specialization, allowing them to make higher quality predictions than the traditional regressors. This is attributed to three main factors: first, NNs are feature engineers that optimize learned features based on the outputs of the model; second, the learned features from the shared backbone have extracted meaningful descriptors from the input morphology information; and lastly, the learned features for each of the output branches have extracted relevant information specialized to each of the OPV performance metrics. This specialization of branches for predicting specific values from a common input, commonly known as multi-task learning, has been shown to be, not only more effective than individually trained networks, but also quicker and simpler to use and train.³⁴ Therefore, by leveraging the advantages of multi-task learning and the interdependent nature of the physical properties being predicted by the system, the NN models are able to outperform the traditional regressors. Within this work, our goal was to explore how the level of incorporated morphological information introduced into the model influenced its performance. We anticipate that using a more complicated NN architecture could potentially improve the performance of the NNs further; such modifications represent reasonable next steps, but are beyond the scope of the current work.

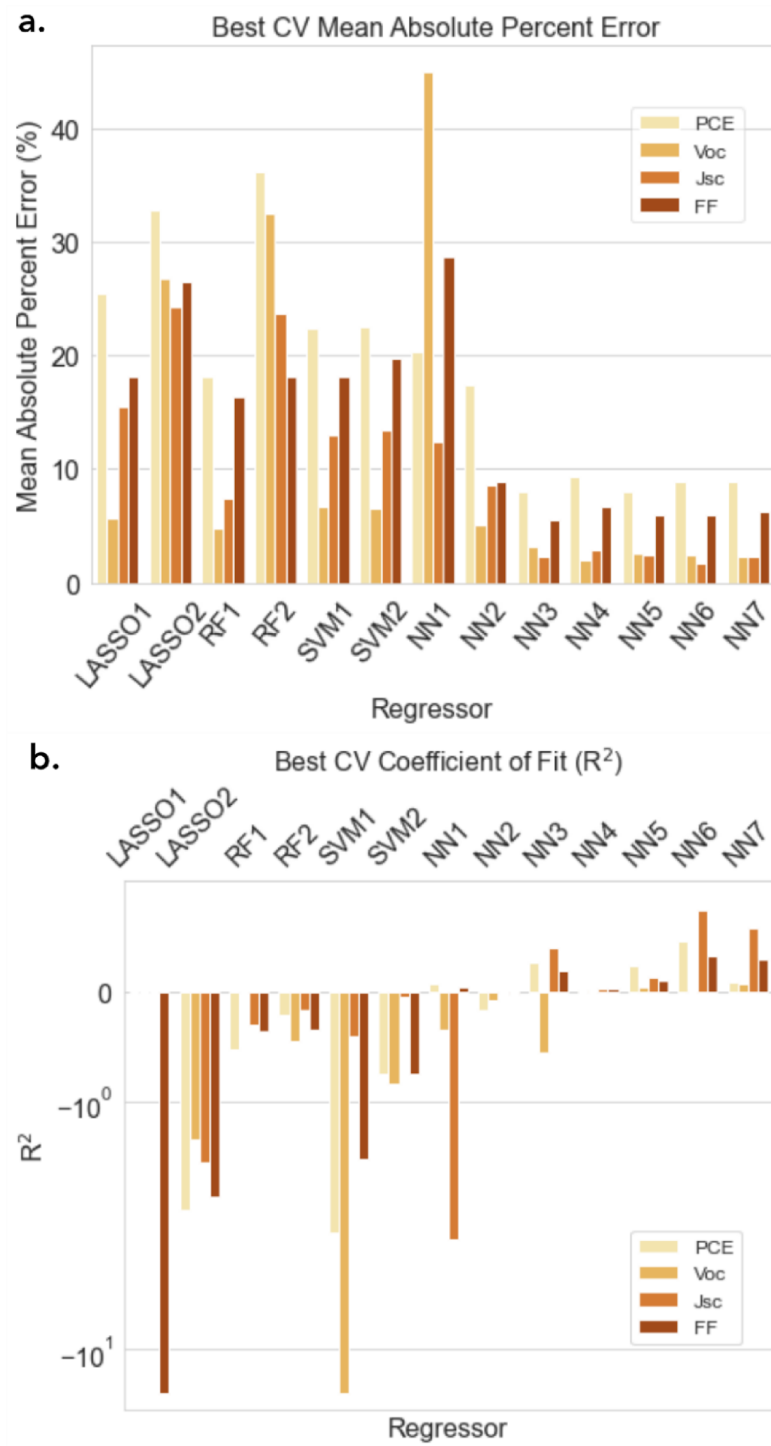


Figure 4. The best **a.** Mean Absolute Percent Error (MAPE), **b.** R^2 for each model's prediction of the four OPV device metrics (PCE , V_{oc} , J_{sc} , and FF). Models were trained on one of the following

datasets: AOD (1), the CTD (2), m2py labels (3), FDM images (4), CTD and m2py labels (5), CTD and FDM images (6), or the CTD, FDM images, and their m2py labels (7).

In general, we see that the models tend to have the easiest time predicting V_{oc} and J_{sc} , while there is consistently more error in predicting PCE and FF . In looking at the V_{oc} surfaces for both the CTD and AOD (Figures S3 and S4, respectively), we see that this device metric is the most tightly distributed and consistent across all annealing conditions. This is due to the fact that V_{oc} is not dependent so much on device morphology as it is on the energy-levels of the donor and acceptor compounds.³⁵ Since all of these devices were made with the same ratio of P3HT:PC₆₁BM, it can be assumed that only minor fluctuations will be present as a result of local fluctuations in P3HT:PC₆₁BM, and that these differences are small enough that most models will predict V_{oc} with relatively low error. J_{sc} , however, shows large distributions and variations throughout the annealing conditions. Despite this, many models were able to predict J_{sc} with relatively low error, especially as more morphological information is presented to the models. This is due to the fact that J_{sc} is strongly linked to active-layer morphology, with sub-optimal morphologies leading to charge-recombination pathways that directly decrease the overall J_{sc} possible in the thin film.^{24,36–40} This correlation is so strong that the models with more encoded morphological information consistently predict J_{sc} most accurately, surpassing even the tightly distributed V_{oc} , as seen with NN3-NN7 in Figure 4 (especially Figure 4.b). This observation confirms that the morphological information distilled by these reported methods are highly descriptive and useful, even with small datasets. Due to the number of factors outside of morphology that influence PCE and FF , such as contact resistance, it is expected that prediction accuracy will be worse for these performance metrics. In looking at the PCE and FF results of Figure 4, this trend is seen. Promisingly, though,

there are still large improvements in PCE and FF upon the incorporation of morphology information, especially when most or all of the morphology information is included (as with NN6 and NN7). In order to further improve the model predictions for PCE and FF , it is expected that adding additional training features that convey such information, such as $J-V$ curves or the measured contact resistance, would address these extra-morphological characteristics that are included in the calculation of PCE and FF .

Looking at the morphological encodings used to train these models, the CTD contains summarized domain measurements for each phase in the image. In addition to these, each phase's relative area occupied in the image and its total area in pixels were included as training features. As such, along with the annealing conditions, there were a total of 74 training features in the CTD to describe the image's morphology. Four different types of regressors were trained using the AOD and the CTD. The results of training with these datasets are shown in Figure 4 with the naming convention following the pattern "Regressor1" and "Regressor2", respectively (*e.g.* LASSO1, LASSO2). For the NN that saw more than just the AOD and CTD, the models are named similarly, as described by Table 1. When compared to the AOD, which has far more samples (366 as compared to 36), it is remarkable how well the models trained on the smaller datasets perform. The information-rich morphology summary of the CTD allows all regressors to perform similar to or better than those trained on the AOD, in spite of having an order of magnitude fewer training samples. This is true for all four regressors and each of the OPV performance metrics, although the most drastic improvements with the introduction of morphology data are seen in the SVM and NN (NN1 to NN2).

These results already point to significant improvements made to each model when they are trained on the CTD as compared to just the AOD, even though the CTD is only a high-level

summary of the OPV active layer morphology. In order to explore the extent to which more descriptive quantitative morphology data informs the prediction of device performance, NNs were trained on gradually increasing degrees of encoded morphology data, as described by Table 1. This table outlines the different NN models and structures, listing them in order of increasing extent of morphology information incorporation. The generalized architectures for each model are shown in Figure S2, as denoted in Table 1 and their hyper parameters are reported in Table S2.

Table 1. Visual architecture reference, model name, and dataset for the series of NN models evaluated in this work.

Architecture	Model	Dataset
Figure S2.a	NN1	Annealing-conditions only (AOD)
	NN2	All Tabular Data (CTD)
Figure S2.b	NN3	m2py labels only
	NN4	Image Data Only
Figure 3, Figure S2.c	NN5	m2py labels & tabular data
	NN6	Image Data & Tabular Data
Figure S2.d	NN7	All data

As anticipated, the NN models with the worst performance were NN1 and NN2, which relied only on tabular data. Although there were improvements from NN1 to NN2 with the incorporation of morphological training features, the addition of two-dimensional morphological data, either as m2py labels or as FDM images, significantly improves model accuracy beyond that of NN2. As previously stated, the densest encoding of morphological information is in FDM images and models incorporating this data are expected to demonstrate better performance. In looking at the results presented in Figure 4 and Table S1, we do see that the best reported cross-validation (CV)

metrics tend to come from models that incorporate the image data, especially NN6. However, Figure 4 shows that the models incorporating the m2py label dataset have similar, if not better performance in their best CV folds. On top of this, Table S1 shows that models that are trained on m2py labels have far higher 5-fold CV metrics for both the average and worst folds than their image dataset counterparts (e.g. the MAPE of NN3 is lower for both the average and worse CV folds than NN4). This result demonstrates that, although there are fewer continuous, local fluctuations encoded in the m2py label dataset, the captured features are more consistently representative of the active-layer's morphology; even though there is less overall data in the m2py label dataset, it contains more information that is densely represented. So, the m2py label dataset allows the NN models to learn how the distribution of the phases present in the active layer affect its device performance.

It is clear that having quantitative morphology information is important to making accurate predictions of final device performance. The inter- and intramolecular interactions of the components in the active-layer directly influence the generation and harvesting of charge-carriers in OPV and affect the electronic properties in many other thin film devices. The highly distilled CTD introduces tractable quantitative morphology features that describe these interactions and that can be used in any regression model. However, the benefits of the richly informative and distilled morphological encoding of the m2py labels seem to surpass these simplified representations by providing broader training features to the NN models, while remaining clearer than the FDM images, as evidenced in Table S1. That said, the performances of NN3 and NN5 are similar enough to imply that the distilled CTD could be sufficient for many applications. Distinguishing the extent to which the more densely encoded morphology information is beneficial will require much larger datasets than those investigated here. It is also worth noting that NN3

through NN7 all have a MAPE below 10% for all OPV metrics, reaching as low as 1.69% for NN6's J_{sc} prediction (Table S1.b). The fact that these models exhibit such high accuracy with so few training samples points to just how informative these reported quantitative morphology encodings are in determining properties and performance of thin film photovoltaics. Further, NN6 and NN7 produce the highest R^2 values of all metrics, with R^2 for NN6's J_{sc} being the highest at 0.73. This is a remarkable result, especially considering the dataset size. These results show that quantitative morphology information, as distilled by m2py labels or summarized in the CTD, are crucial to the success of projects seeking to predict final properties based on thin film morphology. Addition of the CTD descriptions significantly improved the predictive capabilities of all models, even when already training with FDM images, as evidenced by the performance of NN6 compared to NN4.

Finally, although the addition of more detailed device performance information during training, such as contact-resistance or full J-V curves, is expected to improve predictions of device metrics that are not fully dependent on active-layer morphology, such as PCE and FF , the quantitative morphology encodings are shown to already be highly predictive of thin film device performance. So, morphology datasets collected and prepared as those shown in this work can be used to train predictive models that identify the optimal processing conditions for high performing morphologies.

4. CONCLUSIONS

In this work, we outline a generalizable workflow for quantitatively examining thin film morphologies and using those descriptions to train predictive models. These results utilize labels produced by the open-source toolkit, m2py, to identify and measure the size, shape, and

distribution of domains and phases in SPM images of the thin films. Despite the fact that only the surface is imaged, these quantitative morphology summaries are shown to be highly informative in describing changes in thin film nanostructures as a result of annealing conditions. Using this workflow, varying levels of morphological information were extracted from the images and encoded for training predictive models. As more morphology information was included in the model training, the prediction accuracy also increased, in spite of a drastically reduced dataset sample size from the larger AOD to the CTD, m2py labels, and FDM images. When summary morphology descriptors were included alongside SPM image data, the percent error reached as low as 1.7% for J_{sc} , which is highly correlated to morphology, and as high as 8.8% for PCE , which is more challenging to predict. These results confirm the high correlation between morphology and thin film device performance. As such, quantitative morphology descriptions extracted using m2py are shown to be crucial in understanding morphological contributions to the properties of thin film active layers, as well as in predicting final device performance of OPVs. This workflow is expected to be generalizable to all thin film optoelectronic devices whose active layers can be measured by SPM.

CONFLICTS OF INTEREST

There are no conflicts of interest to declare.

ACKNOWLEDGEMENTS

W.K.T. and C.K.L. acknowledge NSF DMR-1708317 and NSF DESE-1633216 for funding. W.K.T. and D.T. acknowledge MATDAT18 (NSF DMR-1748198) for fostering their

collaboration. Part of this work was conducted at the Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure site at the University of Washington, which is supported in part by the National Science Foundation (Grant ECCS-1542101), the University of Washington, the Molecular Engineering & Sciences Institute, and the Clean Energy Institute.

DATA AVAILABILITY

The processed data required to reproduce these findings are available to download from <http://dx.doi.org/10.17632/9j7wm22cgn.1> and is citable as Tatum, Wesley; Torrejon, Diego; Resing, Anton; Luscombe, Christine (2021), “Py-Conjugated: Quantitative Morphology Data for Organic Photovoltaics”, Mendeley Data, V1, DOI: 10.17632/9j7wm22cgn.1

REFERENCES

- (1) Goh, T.; Huang, J.-S.; Yager, K. G.; Sfeir, M. Y.; Nam, C.-Y.; Tong, X.; Guard, L. M.; Melvin, P. R.; Antonio, F.; Bartolome, B. G.; et al. Quaternary Organic Solar Cells Enhanced by Cocrystalline Squaraines with Power Conversion Efficiencies >10%. *Adv. Energy Mater.* **2016**, 6 (21), 1600660.
- (2) Botiz, I.; Stingelin, N. Influence of Molecular Conformations and Microstructure on the Optoelectronic Properties of Conjugated Polymers. *Materials (Basel)*. **2014**, 7 (12), 2273–2300.
- (3) Yan, Y.; Liu, X.; Wang, T. Conjugated-Polymer Blends for Organic Photovoltaics: Rational Control of Vertical Stratification for High Performance. *Adv. Mater.* **2017**, 29 (20), 1601674.
- (4) Chang, M.; Lim, G.; Park, B.; Reichmanis, E. Control of Molecular Ordering, Alignment,

- and Charge Transport in Solution-Processed Conjugated Polymer Thin Films. *Polymers (Basel)*. **2017**, *9* (12), 212.
- (5) Lan, S.; Yang, H.; Zhang, G.; Wu, X.; Ning, W.; Wang, S.; Chen, H.; Guo, T. Impact of Fullerene Structure on Nanoscale Morphology and Miscibility and Correlation of Performance on Small Molecules: Fullerene Solar Cell. *J. Phys. Chem. C* **2016**, *120* (38), 21317–21324.
 - (6) Wang, T.; Kupgan, G.; Brédas, J. L. Organic Photovoltaics: Relating Chemical Structure, Local Morphology, and Electronic Properties. *Trends in Chemistry*. Cell Press June 1, 2020, pp 535–554.
 - (7) Tatum, W. K.; Torrejon, D.; O’Neil, P.; Onorato, J. W.; Resing, A. B.; Holliday, S.; Flagg, L. Q.; Ginger, D. S.; Luscombe, C. K.; O’Neil, P.; et al. A Generalizable Framework for Algorithmic Interpretation of Thin Film Morphologies in Scanning Probe Images. *J. Chem. Inf. Model.* **2020**, *16* (03).
 - (8) Zhou, K.; Liu, J.; Li, M.; Yu, X.; Xing, R.; Han, Y. Phase Diagram of Conjugated Polymer Blend P3HT/PF12TBT and the Morphology-Dependent Photovoltaic Performance. *J. Phys. Chem. C* **2015**, *119* (4), 1729–1736.
 - (9) Zhou, G.; Ding, H.; Zhu, L.; Qiu, C.; Zhang, M.; Hao, T.; Feng, W.; Zhang, Y.; Zhu, H.; Liu, F. Photophysics, Morphology and Device Performances Correlation on Non-Fullerene Acceptor Based Binary and Ternary Solar Cells. *J. Energy Chem.* **2020**, *47*, 180–187.
 - (10) Jones, M. L.; Dyer, R.; Clarke, N.; Groves, C. Are Hot Charge Transfer States the Primary Cause of Efficient Free-Charge Generation in Polymer:Fullerene Organic Photovoltaic Devices? A Kinetic Monte Carlo Study. *Phys. Chem. Chem. Phys.* **2014**, *16* (38), 20310–

20320.

- (11) Groves, C. Developing Understanding of Organic Photovoltaic Devices: Kinetic Monte Carlo Models of Geminate and Non-Geminate Recombination, Charge Transport and Charge Extraction. *Energy Environ. Sci.* **2013**, 6 (11), 3202.
- (12) Ye, L.; Xiong, Y.; Zhang, M.; Guo, X.; Guan, H.; Zou, Y.; Ade, H. Enhanced Efficiency in Nonfullerene Organic Solar Cells by Tuning Molecular Order and Domain Characteristics. *Nano Energy* **2020**, 77, 105310.
- (13) Li, Haoyuan; Sini, Gjergji; Sit, Joseph; Moulé, Adam J.; Bredas, Jean-Luc. Understanding Charge Transport in Donor/Acceptor Blends From Large-Scale Device Simulations Based on Experimental Film Morphologies. *Energy & Env. Sci.* **2020**, 13 (2), 601–615.
- (14) Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; et al. Machine Learning–Assisted Molecular Design and Efficiency Prediction for High-Performance Organic Photovoltaic Materials. *Sci. Adv.* **2019**, 5 (11).
- (15) Wu, Y.; Guo, J.; Sun, R.; Min, J. Machine Learning for Accelerating the Discovery of High-Performance Donor/Acceptor Pairs in Non-Fullerene Organic Solar Cells. *npj Comput. Mater.* **2020**, 6 (1), 1–8.
- (16) Miller, E. D.; Jones, M. L.; Henry, M. M.; Stanfill, B.; Jankowski, E. Machine Learning Predictions of Electronic Couplings for Charge Transport Calculations of P3HT. *AIChE J.* **2019**, 65 (12).
- (17) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, 10 (8), 1903242.
- (18) Nagasawa, S.; Al-Naamani, E.; Saeki, A. Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. *J. Phys. Chem. Lett.*

- 2018**, 9 (10), 2639–2646.
- (19) Sahu, H.; Ma, H. Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. *J. Phys. Chem. Lett.* **2019**, 10 (22), 7277–7284.
 - (20) Pokuri, B. S. S.; Ghosal, S.; Kokate, A.; Sarkar, S.; Ganapathysubramanian, B. Interpretable Deep Learning for Guided Microstructure-Property Explorations in Photovoltaics. *npj Comput. Mater.* **2019**, 5 (1), 1–11.
 - (21) Cao, B.; Adutwum, L. A.; Oliynyk, A. O.; Lubner, E. J.; Olsen, B. C.; Mar, A.; Buriak, J. M. How To Optimize Materials and Devices *via* Design of Experiments and Machine Learning: Demonstration Using Organic Photovoltaics. *ACS Nano* **2018**, 12 (8), 7434–7444.
 - (22) Rodríguez-Martínez, X.; Pascual-San-José, E.; Fei, Z.; Heeney, M.; Guimerà, R.; Campoy-Quiles, M. Predicting the Photocurrent–Composition Dependence in Organic Solar Cells. *Energy Environ. Sci.* **2021**, 14 (2), 986–994.
 - (23) Kohn, P.; Rong, Z.; Scherer, K. H.; Sepe, A.; Sommer, M.; Müller-Buschbaum, P.; Friend, R. H.; Steiner, U.; Hüttner, S. Crystallization-Induced 10-Nm Structure Formation in P3HT/PCBM Blends. *Macromolecules* **2013**, 46 (10), 4002–4013.
 - (24) Huang, Y.-C. C.; Cha, H.-C. C.; Chen, C.-Y. Y.; Tsao, C.-S. S. Morphological Control and Performance Improvement of Organic Photovoltaic Layer of Roll-to-Roll Coated Polymer Solar Cells. *Sol. Energy Mater. Sol. Cells* **2016**, 150, 10–17.
 - (25) Chen, D.; Nakahara, A.; Wei, D.; Nordlund, D.; Russell, T. P. P3HT/PCBM Bulk Heterojunction Organic Photovoltaics: Correlating Efficiency and Morphology. *Nano Lett.* **2011**, 11 (2), 561–567.

- (26) van der Walt, S. C. C.; Varoquaux, G. The Numpy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.
- (27) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
- (28) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.
- (29) Waskom, M.; Team, S. development. Seaborn: Statistical Data Visualization. Zenodo 2020.
- (30) Fabian Pedregosa Gael Varoquaux, A. G. et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (31) Dang, M. T.; Hirsch, L.; Wantz, G. P3HT:PCBM, Best Seller in Polymer Photovoltaic Research. *Adv. Mater.* **2011**, *23* (31), 3597–3602.
- (32) Chen, J.-Y.; Kuo, C.-C.; Lai, C.-S.; Chen, W.-C.; Chen, H.-L. Manipulation on the Morphology and Electrical Properties of Aligned Electrospun Nanofibers of Poly(3-Hexylthiophene) for Field-Effect Transistor Applications. *Macromolecules* **2011**, *44* (8), 2883–2892.
- (33) van der Walt Johannes L. Schonberger, J. N.-I. et al. Scikit-Image: Image Processing in Python. *PeerJ*. 2014, p e453.
- (34) Girshick, R. *Fast R-CNN*.
- (35) Azzouzi, M.; Kirchartz, T.; Nelson, J. Factors Controlling Open-Circuit Voltage Losses in Organic Solar Cells. *Trends in Chemistry*. Cell Press April 1, 2019, pp 49–62.
- (36) Min Nam, Y.; Huh, J.; Ho Jo, W. Optimization of Thickness and Morphology of Active Layer for High Performance of Bulk-Heterojunction Organic Solar Cells. *Sol. Energy*

- Mater. Sol. Cells* **2010**, *94* (6), 1118–1124.
- (37) Mazzio, K. A.; Luscombe, C. K. The Future of Organic Photovoltaics. *Chem. Soc. Rev.* **2015**, *44*, 78–90.
- (38) Giridharagopal, R.; Ginger, D. S. Characterizing Morphology in Bulk Heterojunction Organic Photovoltaic Systems. *J. Phys. Chem. Lett.* **2010**, *1* (7), 1160–1169.
- (39) Masters, R. C.; Wan, Q.; Zhang, Y.; Dapor, M.; Sandu, A. M.; Jiao, C.; Zhou, Y.; Zhang, H.; Lidzey, D. G.; Rodenburg, C. Novel Organic Photovoltaic Polymer Blends: A Rapid, 3-Dimensional Morphology Analysis Using Backscattered Electron Imaging in the Scanning Electron Microscope. *Sol. Energy Mater. Sol. Cells* **2017**, *160*, 182–192.
- (40) Treat, N. D.; Chabinyc, M. L. Phase Separation in Bulk Heterojunctions of Semiconducting Polymers and Fullerenes for Photovoltaics. *Annu. Rev. Phys. Chem.* **2014**, *65* (1), 4.